ÉCOLE NATIONALE SUPÉRIEURE AGRONOMIQUE DE MONTPELLIER

THÈSE

Présentée pour obtenir le diplôme de doctorat

Spécialité: Biologie des Organismes et des Populations
Formation Doctorale: Biologie de l'Évolution et Écologie
École Doctorale: Biologie des Systèmes Intégrés, Agronomie, Environnement

APPROCHES ÉCOLOGIQUE ET ÉVOLUTIVE DE LA VACCINATION

par

Marc CHOISY

Soutenue le 1 décembre 2004 devant le jury composé de

Pierre AUGER	Directeur de Recherche, IRD, Paris	Examinateur
Éric DELAPORTE	Professeur, Université Montpellier I, Montpellier	Examinateur
Jean-François GUÉGAN	Directeur de Recherche, IRD, Montpellier	Directeur de Thèse
Marie-Laure NAVAS	Professeur, ENSAM, Montpellier	Examinateur
Dominique PONTIER	Professeur, Université Claude Bernard, Lyon	Rapporteur
Alain-Jacques VALLERON	Professeur, Université Pierre & Marie Curie, Paris	Rapporteur

ÉCOLE NATIONALE SUPÉRIEURE AGRONOMIQUE DE MONTPELLIER

THÈSE

Présentée pour obtenir le diplôme de doctorat

Spécialité: Biologie des Organismes et des Populations
Formation Doctorale: Biologie de l'Évolution et Écologie
École Doctorale: Biologie des Systèmes Intégrés, Agronomie, Environnement

APPROCHES ÉCOLOGIQUE ET ÉVOLUTIVE DE LA VACCINATION

par

Marc CHOISY

Soutenue le 1 décembre 2004 devant le jury composé de

Pierre AUGER	Directeur de Recherche, IRD, Paris	Examinateur
Éric DELAPORTE	Professeur, Université Montpellier I, Montpellier	Examinateur
Jean-François GUÉGAN	Directeur de Recherche, IRD, Montpellier	Directeur de Thèse
Marie-Laure NAVAS	Professeur, ENSAM, Montpellier	Examinateur
Dominique PONTIER	Professeur, Université Claude Bernard, Lyon	Rapporteur
Alain-Jacques VALLERON	Professeur, Université Pierre & Marie Curie, Paris	Rapporteur

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction

> *« Disease! That is the main force, the diligent force, the devastating force! It attacks the infant the moment it is born; it furnishes it one malady after another: croup, measles, mumps, bowel troubles, teething pains, scarlet fever, and other childhood specialties. It chases the child into youth and furnishes it some specialties for that time of life. It chases the youth into maturity, maturity into age, age into the grave. »*
>
> — Mark TWAIN, 1909

1.1 Les relations hôte-parasite

LES êtres vivants interagissent entre eux. Au sein d'une même espèce, les individus entrent en compétition, communiquent et se reproduisent [24]. Les relations entre espèces différentes font intervenir, outre la compétition, la prédation et le parasitisme au sens large [40]. Il y a une relation de parasitisme *sensu lato* lorsqu'un organisme passe une partie de sa vie dans ou sur un ou plusieurs autre(s) organisme(s). Cette relation peut être facultative ou obligatoire, selon la dépendance du parasite à son hôte. On peut distinguer trois phases clef dans le cycle de vie d'un parasite, schématisées sur la figure 1.1. La première étape est l'**infection**, *i.e.* la pénétration d'un parasite dans (ou son attachement sur) un hôte. Ensuite, le parasite **colonise** son nouvel environnement. Au cours de cette phase, le parasite peut se multiplier sexuellement ou asexuellement. Enfin, le parasite se **disperse** et colonise de nouveaux hôtes. Entre les phases intra-hôtes, le parasite peut éventuellement séjourner dans l'environnement extérieur, généralement sous une forme de résistance.

Selon les coûts et bénéfices conférés à chacun des protagonistes (hôte ou

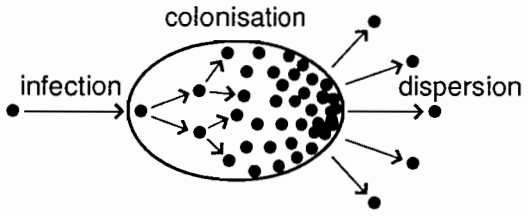


FIG. 1.1 – **Schématisation d'un cycle de vie parasitaire. L'ovale représente un hôte et les ronds noirs des parasites. Le cycle de vie commence par la pénétration du parasite dans (ou sa fixation sur) un hôte (infection). Dans ou sur cet hôte, le parasite peut se reproduire sexuellement ou asexuellement (colonisation). Certains individus parasites quittent ensuite l'hôte pour infecter d'autres hôtes de la même espèce ou d'une espèce différente (dispersion).**

parasite), on peut distinguer deux grands types de relation hôte-parasite au sens large : la **symbiose**, où hôte et parasite tirent avantage de la relation (*e.g.* lichens, symbioses entre une algue et un champignon), et le **parasitisme** *sensu stricto*, caractérisé par un avantage pour le parasite seul. Cet avantage peut être en termes de protection et/ou locomotion et/ou nourriture. Le coût subit par l'hôte (*i.e.* **virulence**) peut être d'intensité très variable. Certains parasites n'affectent que très peu leur hôte, on parle alors de **commensalisme**. Les bactéries *Escherichia coli* résidant dans le tube digestif des vertébrés en sont un bon exemple. D'autres, au contraire, diminuent très notablement la valeur sélective de leur hôte. C'est, par exemple, le cas du protozoaire *Plasmodium falciparum*, agent du paludisme. Le déterminisme du niveau de virulence n'est pas toujours très bien connu et semble en général dépendre de caractéristiques liées non seulement au parasite, mais aussi à l'hôte (niveau de santé général, *etc...*).

1.2 La lutte contre les maladies infectieuses

Un objectif pratique en santé publique ou en agronomie concerne la lutte contre les maladies parasitaires[1]. Les motivations peuvent être de l'ordre du confort, de l'économie ou de la santé humaine.

L'histoire de l'Homme a été très profondément marquée par de grandes

1. Dans ce mémoire, on utilise indifféremment les termes de maladies infectieuses, contagieuses, ou parasitaires.

épidémies dévastatrices et les pressions parasitaires ont fortement façonné l'évolution humaine [175, 78, 200]. Citons, par exemple, les pestes de l'Antiquité et du Moyen-Âge [2], les ravages de la **variole**, du **choléra** et de la **tuberculose**. Avec les progrès de la médecine et l'augmentation générale de l'hygiène, la fin du XXième siècle nous a donné l'illusion que nous ne connaîtrons plus ces plaies qui se sont abattues sur nos ancêtres [35]. Ainsi, dans les années soixante-dix, les maladies infectieuses ne constituaient plus qu'une part anecdotique de la recherche médicale, alors presque entièrement vouée aux maladies cardio-vasculaires ou à la lutte contre le cancer [68]. Toutefois, les **maladies émergentes** ou ré-émergentes qui touchent l'Homme (SIDA [3], SRAS [4], légionellose) ou son agriculture (encéphalite spongiforme bovine, fièvre aphteuse, grippe aviaire) depuis ces dernières années nous ont fait peu à peu prendre conscience que l'Homme est toujours aussi fragile qu'il l'a été, et qu'il devra sans cesse se défendre contre un environnement agressif et qui évolue [152]. Cette observation est de plus en plus d'actualité. En effet, la croissance de la population humaine ne va pas sans une augmentation du nombre de contacts avec un environnement sauvage, et donc du risque de **zoonoses** comme la grippe, le SIDA ou les fièvres causées par le virus Ebola [134, 161].

Par rapport au cycle de vie présenté sur la figure 1.1, la lutte contre les maladies parasitaires peut s'opérer à deux niveaux : (*i*) l'infection proprement dite ou (*ii*) la colonisation de l'organisme hôte par le parasite, une fois l'infection effectuée. Réduire les risques d'infections passe par la réduction des contacts entre individus sains et individus contagieux. La pratique de la **quarantaine** ou la lutte contre les insectes vecteurs en sont des exemples. Empêcher la colonisation d'un organisme hôte après infection peut nécessiter l'administration de médicaments comme les antibiotiques anti-bactériens. La vaccination se situe également à ce deuxième niveau et est basée sur les propriétés du système immunitaire des vertébrés.

2. La peste noire du haut Moyen-Âge est causée par le bacille *Yersinia pestis* identifié en 1894. Le mot « peste » désigne aussi plus généralement toute épidémie de grande envergure, les symptômes rapportés dans les écrits les plus anciens ne permettant pas toujours d'identifier la maladie avec certitude.

3. Syndrome d'immuno-déficience humaine causé par le HIV, virus d'immunodéficience humaine.

4. Syndrome respiratoire aigü sévère.

1.3 Le principe de la vaccination

1.3.1 Historique

L'invention du premier vaccin est très étroitement liée à la lutte contre la variole[5]. Avant son éradication à la fin des années soixante-dix, la variole faisait partie des six maladies infectieuses les plus meurtrières [175]. Un virus proche du virus de la variole humaine est responsable de la variole de la vache (*cow-pox* en anglais) et ne cause pas de maladie sévère chez l'être humain. À la fin du XVIIIième siècle, le médecin anglais Edward JENNER observe que les fermières effectuant la traite des vaches contractaient moins la variole que le reste de la population. Pensant qu'une infection avec la variole de la vache pouvait protéger contre la variole humaine, JENNER prélève du pus d'une pustule sur une vachère infectée par la variole de la vache et l'inocule à un jeune garçon de huit ans. Quelques mois plus tard, il lui inocule le virus de la variole humaine, sans conséquence. Le garçon était immunisé et la vaccination était née. Le mot **vaccin** tire son origine de *vaccine*, terme désignant le sérum d'une vache (du latin *vacca*, vache) infectée par le poxvirus responsable du *cow-pox*.

1.3.2 Mémoire du système immunitaire des vertébrés

Les vertébrés possèdent un système immunitaire perfectionné permettant de lutter contre tout corps étranger, parasite ou non, pénétrant à l'intérieur de l'organisme [96]. Le système immunitaire des vertébrés est un ensemble complexe de cellules et molécules aux fonctions différentes et complémentaires [189]. Une première caractéristique de ce système est son **adaptabilité**, lui permettant de produire une réponse spécifique à chaque type d'agression. Une deuxième caractéristique du système immunitaire des vertébrés est sa **mémoire** des infections passées. Cette mémoire permet de répondre à une seconde infection d'un agent pathogène plus rapidement et plus efficacement qu'à une primo-infection de ce même agent. On parle d'**immunité acquise** lorsque ce mécanisme de mémoire confère une résistance à l'agent pathogène en question. Cette résistance peut être d'une durée plus ou moins longue, de quelques semaines contre la plupart des helminthes à la vie entière contre certaines maladies virales comme la rougeole [11]. D'autre part, cette immunité peut être plus ou moins spécifique et l'on parle d'**immunité croisée** (*cross-specific immunity* en anglais) lorsque l'immunité contre un agent pathogène

5. La variole est due à un virus appartenant au genre des orthopoxvirus. Cette affection, très contagieuse, se manifeste par une éruption vésiculo-pustuleuse généralisée à l'origine de cicatrices indélébiles.

protège également contre d'autres agents plus ou moins proches [64].

La vaccination utilise les propriétés d'adaptabilité et de mémoire du système immunitaire pour protéger les individus contre les infections parasitaires. L'administration d'un agent vivant modifié, d'une suspension d'organismes tués, ou d'une toxine inactivée à un individu peut stimuler son système immunitaire, le protégeant ainsi de toute infection ultérieure. Ainsi, dans l'exemple ci-dessus, nous avons vu qu'une exposition au virus de la variole de la vache immunise contre la variole humaine.

1.4 Les objectifs de la vaccination

1.4.1 Protection individuelle

La vaccination protège une personne contre une maladie et peut être utilisée au cas par cas, selon l'état de santé ou à risque de chaque individu. Ainsi, les vétérinaires pourront se vacciner contre la rage, les habitants des régions tropicales contre la fièvre jaune, les personnes agées ou immunodéprimées contre la grippe.

1.4.2 Protection populationnelle

Pour les maladies très contagieuses et à forte mortalité et/ou morbidité et/ou coût économique, une volonté d'**éradication globale** a émergé dès le début des années cinquante. La réalisation d'une telle extinction passe par la mise en place d'une politique de vaccination à grande échelle. L'idée est d'utiliser la vaccination dans le but de réduire le nombre de susceptibles dans la population à un niveau tel que les contacts contagieux entre individus deviennent trop rares pour que la maladie puisse persister [11]. De telles politiques vaccinales ont été établies, au cours des années cinquante, dans plusieurs pays européens et nord-américains contre plusieurs maladies comme la rougeole, la coqueluche, les oreillons. Les prévalences de ces maladies ont chuté très notablement dans les pays concernés en seulement une dizaine d'années. Après une campagne mondiale de vaccination contre la variole, l'Organisation Mondiale de la Santé (OMS) a officiellement proclamé son éradication en 1980 [151]. La disparition de la variole de la surface de la Terre constitue à ce jour le plus grand succès de la vaccination. Fort de cet exemple, l'OMS lance en 1988 une campagne d'éradication mondiale de la poliomyélite. En quinze années, le nombre de cas avait déjà diminué de 99%.

1.5 Une approche écologique et évolutive de la vaccination

Les relations entre un agent pathogène et son hôte s'opèrent à différentes **échelles spatiales et temporelles** [39, 74], comme illustré sur la figure 1.2. Les échelles spatiales varient des interactions moléculaires aux populations, et même métapopulations [6]. La dimension temporelle est plus une opposition entre durées de **temps écologique**, dominées par des processus essentiellement dynamiques [7], et durées de **temps évolutif**, caractérisées par des phénomènes de coévolution entre hôte et pathogène. Ces deux dimensions ne sont pas totalement indépendantes puisque les phénomènes à grande échelle temporelle ne peuvent se produire que pour des structures de grande dimension spatiale. En effet, la durée de vie d'une molécule est plus courte que celle d'une cellule, elle-même plus courte que celle d'un organisme, elle-même plus courte que celle d'une population, elle-même plus courte que celle d'une métapopulation. Cette hiérarchisation est responsable de la zone hachurée sur la figure 1.2.

Dans ce contexte, l'**épidémiologie** s'intéresse à l'évolution du nombre de malades dans les populations. L'**immunologie** s'intéresse aux processus à petite échelle spatiale qui s'opèrent essentiellement autour de la réponse immunitaire. Ces processus font intervenir des phénomènes dynamiques entre les cellules du système immunitaire et les pathogènes, ainsi que des phénomènes d'évolution, ceux-là même responsables de l'adaptabilité du système immunitaire dont nous avons parlé ci-dessus, au paragraphe 1.3.2. Cette adaptabilité se fait, en réalité, par simple sélection sur une importante variabilité pré-existante dans le système immunitaire. Enfin, la **phylogénie** et la **génétique des populations** s'intéressent aux processus s'opérant à la fois à grande échelle spatiale et temporelle (migration, mutation, sélection). La figure 1.2 illustre les relations entre épidémiologie, immunologie et génétique des populations.

Avec l'émergence et la ré-émergence de nouvelles maladies, nous avons assisté ces dernières années à un développement des études écologiques et évolutives sur les maladies infectieuses humaines. Essentiellement, par rapport aux approches plus classiques, ces nouveaux types d'approches s'intéressent au comportement des maladies à grande échelle spatiale (écologie) et/ou temporelle (évolution). Dans le cadre de la vaccination, une approche écologique et évolutive se justifie à deux niveaux que nous détaillons ci-dessous.

6. Une métapopulation est une population de populations. Voir encadré 5.1 page 98 pour plus de détails.

7. Au sens « dynamique des populations ».

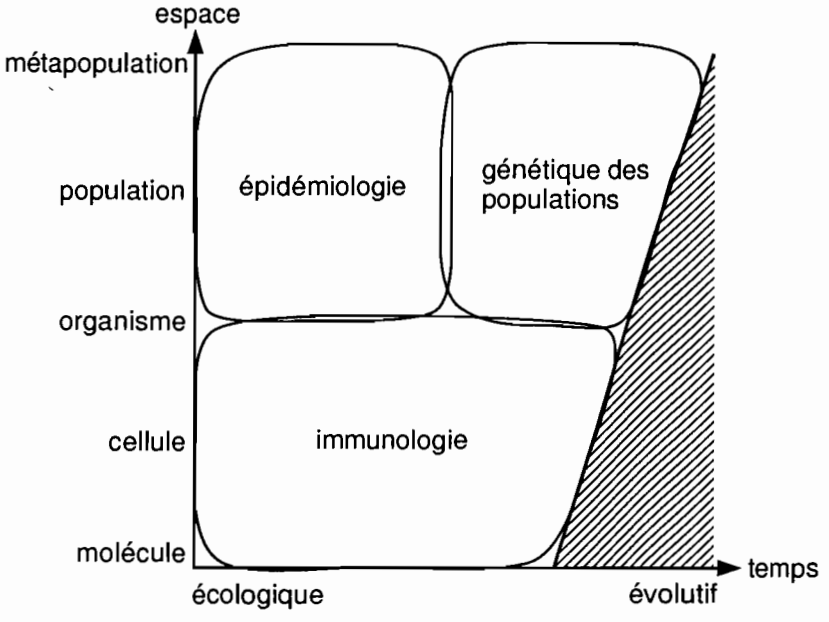


FIG. 1.2 – **Niveaux d'échelles et hiérarchies. Les relations entre un agent pathogène et son hôte s'opèrent à différentes échelles temporelles (des phénomènes écologiques aux phénomènes évolutifs) et spatiales (des interactions moléculaires aux populations et métapopulations). La partie hachurée rend compte de la structure générale du vivant impliquant que les molécules ont une durée de vie plus courte que les cellules et ainsi de suite jusqu'aux métapopulations. Les disciplines d'études des relations hôte-parasite dépendent de ces niveaux d'échelle, voir texte.**

1.5.1 Mise au point d'un vaccin

Le premier aspect passe par la mise au point du vaccin proprement dit. Plusieurs questions se posent alors, comme le choix de la particule constituant le vaccin : agent mort, vivant atténué, ou toxine désactivée. La mise au point d'un vaccin contre des agents à évolution rapide se complique davantage. C'est par exemple le cas des virus de la grippe et du SIDA.

La grippe est une infection respiratoire commune caractérisée par la régularité annuelle de ses pics épidémiques, généralement durant les mois d'hiver, de novembre à mars [207]. L'agent responsable de la grippe est un virus à ARN[8] de la famille des Orthomyxoviridae. Le virus comprend **trois types principaux** (A, B, et C) caractérisés par des différences sur deux protéines internes majeures [48, 64]. Le type A, le plus prépondérant dans les populations humaines, est lui-même divisé en plusieurs **sous-types** caractérisés par des différences sur les protéines membranaires hæmagglutinine (H) et neuraminidase (N) qui sont les cibles majeures du système immunitaire. Deux sous-types circulent actuellement dans les populations humaines : H1N1 et H3N2. Les sous-types sont ensuite divisés en plusieurs **souches**, définis comme des isolats distincts de virus. La plupart des individus en bonne santé guérissent naturellement de la grippe et semblent ensuite bénéficier d'une immunité à vie contre les souches proches de la souche avec laquelle ils ont été infectés. Comme beaucoup de virus à ARN, le virus de la grippe évolue particulièrement vite et de récentes études moléculaires et phylogénétiques mettent en évidence que chaque pic épidémique est caractérisé par un ensemble différent de souches virales [48] (voir figure 1.3). Dans un tel contexte, le vaccin contre la grippe doit changer chaque année et sa mise au point doit tenir compte de l'intensité de l'**immunité croisée** entre souches différentes, ainsi que de la direction de la **dérive génétique** d'une année à l'autre.

Le SIDA est une maladie mortelle causée par un virus à ARN double brin. La fréquence des erreurs de transcription entre ARN et ADN[9], la possibilité de recombinaison entre les deux brins d'ARN, l'importante taille des populations virales, ainsi que la rapidité du cycle viral, font de ce virus un des organismes ayant l'évolution la plus rapide du monde vivant [30, 97]. Comme pour la grippe, des études moléculaires et phylogénétiques ont mis en évidence une multitude de types, sous-types et souches différentes [90]. Il n'existe actuellement pas de vaccin contre le virus du SIDA et des essais de mise au point sont actuellement en cours. Face à cette importante diversité de HIV, une des préoccupations majeures dans la mise au point d'un vaccin potentiel concerne l'étendue de l'**immunité croisée** entre les différentes

8. Acide ribonucléique.
9. Acide désoxyribonucléique.

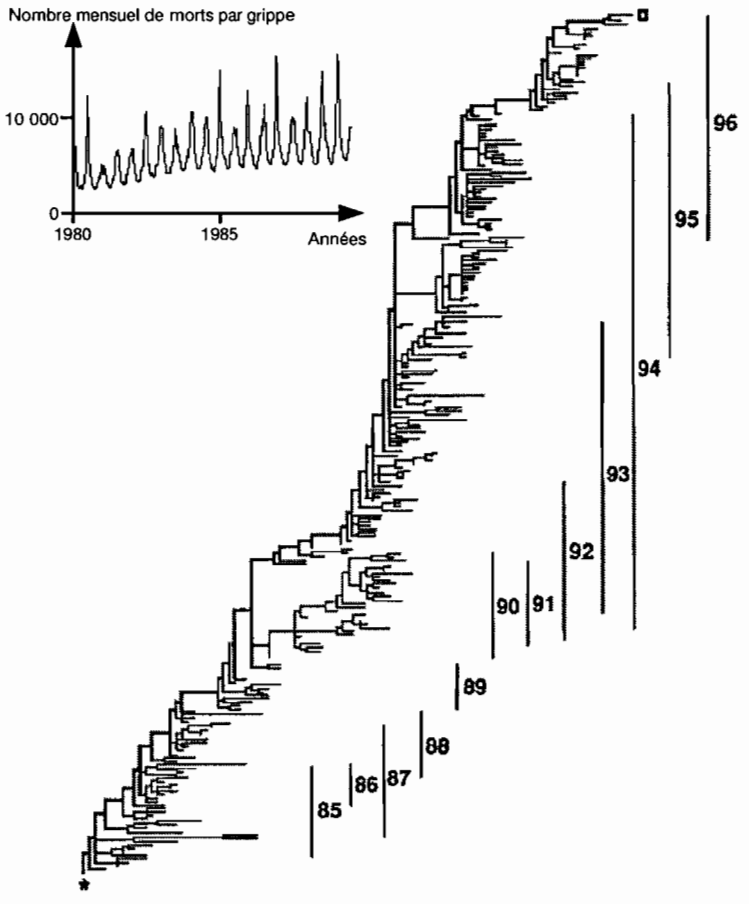


FIG. 1.3 – **Dynamique et évolution du virus de la grippe. L'arbre est une reconstruction phylogénétique par maximum de parcimonie du domaine HA1 du sous-type H3N2 du virus de la grippe de type A. Les lignes verticales représentent les années épidémiques des isolats (voir dynamique du nombre mensuel de morts par grippe ou pneumonie). Les données concernent les USA. L'arbre est issu de [63] et la série temporelle de [48]. Noter la structure déséquilibrée de l'arbre résultant de la sélection continue exercée par le système immunitaire.**

souches de virus [69]. Ce point particulier sera traité plus en détail dans le chapitre 3.

1.5.2 Effets de la vaccination

Une fois le vaccin mis au point, il faut s'intéresser aux effets de la vaccination. Idéalement, les effets de la vaccination devraient être égaux aux objectifs présentés au paragraphe 1.4, à savoir une protection à l'échelle individuelle ou populationnelle. En pratique, ces objectifs ne sont pas toujours atteints et des effets autres que ceux escomptés peuvent se produire. Au niveau individuel, ces effets peuvent se traduire par la déclaration de la maladie ou des phénomènes d'**allergie**. Au niveau populationnel, peuvent se produire des effets dynamiques ou évolutifs à plus ou moins longue échéance. Savoir prédire ces effets est un préalable fondamental à la mise en place d'une politique vaccinale.

La vaccination peut par exemple modifier la **périodicité** des épidémies de certaines maladies. Ainsi, la vaccination peut faire passer les épidémies de rougeole d'annuelles à bisannuelles ou inversement, selon le taux de natalité [49]. Certaines maladies bénignes chez l'enfant peuvent provoquer des **complications** sévères chez l'adulte. Un des effets de la vaccination est l'augmentation de l'âge moyen à l'infection, ce qui peut entraîner, dans certains cas, une augmentation du nombre de cas graves, comme dans l'exemple d'une vaccination insuffisante contre la rubéole [9] (voir chapitre 6). D'un point de vue évolutif, certains vaccins à efficacité partielle peuvent sélectionner pour une **augmentation de la virulence**. Ainsi, un modèle mathématique suggère qu'un tel vaccin contre la malaria, développé pour réduire la croissance ou la toxicité du pathogène dans son hôte, pourrait favoriser l'évolution vers une augmentation de virulence chez les individus non vaccinés [66].

1.6 L'objet de la thèse

Le travail de thèse présenté dans ce mémoire s'insère dans ce cadre général d'étude écologique et évolutive de la vaccination. Cette thèse est basée sur le développement de modèles mathématiques, ainsi que l'analyse statistique de données. Nous avons essayé d'explorer une diversité de problématiques écologiques et évolutives associées à la pratique de la vaccination. Pour ce faire, nous avons étudié plusieurs systèmes hôte-microparasite. Ces derniers ont été choisis en fonction de chaque question posée et des collaborations que nous avons pu établir au sein de mon laboratoire d'accueil (GEMI, à l'IRD Montpellier), ainsi qu'avec d'autres laboratoires. Nous parlerons donc, dans cette

thèse, de **SIDA** (travaux réalisés en collaboration avec David ROBERTSON de l'Université de Manchester – GB et Christopher WOELK de l'Université de San Diego – USA), de **leishmaniose** (travaux réalisés en collaboration avec Anne-Laure BAÑULS et Mallorie HIDE du GEMI), de **rougeole** (travaux réalisés en collaboration avec Pejman ROHANI de l'Université de Georgia – USA) et de **varicelle** (travaux réalisés en collaboration avec l'unité 444 de l'INSERM à Paris).

Par rapport aux niveaux d'échelles présentés au paragraphe 1.5, cette thèse s'intéresse à deux extrémités d'un large spectre, comme illustré sur la figure 1.4. Les parties 1 et 2 peuvent être abordées indépendamment l'une de l'autre.

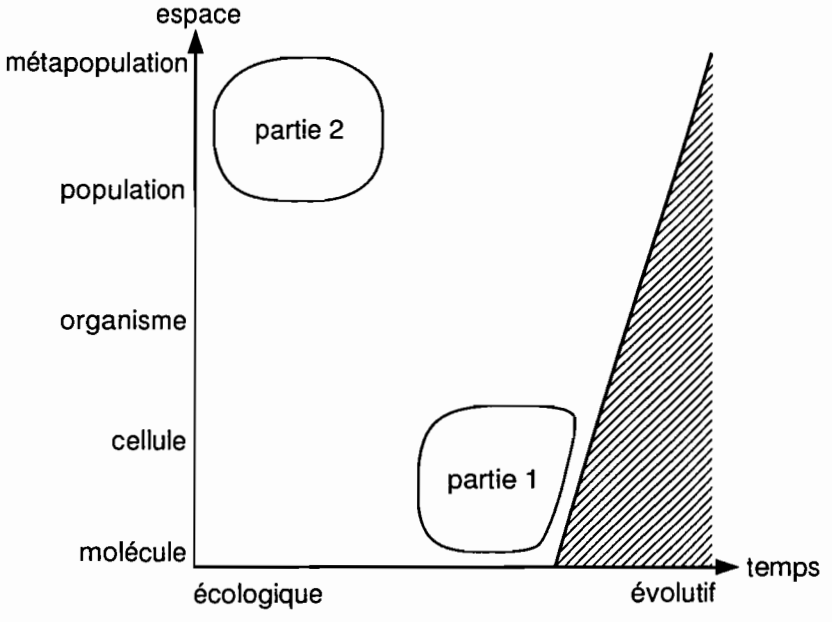


FIG. 1.4 – **Organisation du mémoire dans le contexte des multiples échelles spatio-temporelles d'interactions entre hôte et parasite présenté au paragraphe 1.5. Comparer avec la figure 1.2.**

La partie 1 est axée sur l'identification de **cibles vaccinales** potentielles. Pour cela, nous nous intéressons aux interactions moléculaires fines entre un pathogène et son hôte, en particulier les phénomènes d'**adaptation rapide** qui entrent en jeu dans les interactions avec le système immunitaire. Cette partie comprend trois chapitres. Le premier présente les méthodes statistiques employées pour détecter l'adaptation au niveau moléculaire. En particulier, il

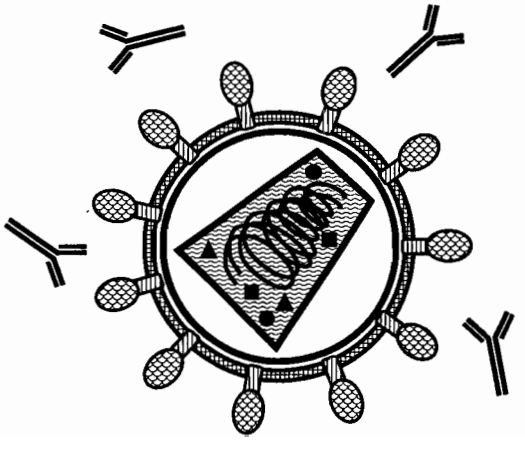
contraste les points forts et points faibles des deux grands types de méthodes actuellement utilisées, à savoir les méthodes dites intuitives ou *ad hoc* et les méthodes basées sur le maximum de vraisemblance. Ce chapitre fait également ressortir les défis qu'il reste à résoudre pour le développement futur de ces méthodes. Les deux chapitres qui suivent sont des applications de ces méthodes à des systèmes hôte-microparasite concrets pour lesquels le développement de vaccins est actuellement à l'étude. Le premier concerne l'évolution des lentivirus de primates, à savoir le HIV mais aussi ses proches parents, les divers SIV, infectant différents singes africains. Le deuxième quantifie et localise l'évolution moléculaire s'opérant sur une famille de gènes identifiés comme facteurs de virulence chez les protozoaires parasites du genre *Leishmania*.

La partie 2 s'intéresse aux conséquences dynamiques de politiques vaccinales imparfaites (*i.e.* qui ne conduisent pas à l'extinction totale de la maladie). Pour cela, nous nous focalisons sur les maladies infantiles microparasitaires dans les populations humaines. Cette partie est également constituée de trois chapitres. Le premier est consacré à l'étude des dynamiques de maladies, en l'absence de vaccination. Spécifiquement, nous nous intéressons à la varicelle en France et sa dynamique spatiale. Connaître la dynamique d'une maladie en l'absence de vaccination est important pour l'application de politiques vaccinales optimisées. Les deux chapitres suivants sont très liés l'un à l'autre et traitent des deux principales politiques vaccinales actuellement en application. Le premier est une courte revue des principaux résultats sur la vaccination de masse. Cette politique vaccinale est la plus ancienne et largement la plus utilisée. Bien qu'efficace, elle est très lourde à mettre en place. Le deuxième étudie les conséquences dynamiques de la vaccination par pulsations. Cette politique vaccinale a été développée seulement récemment dans le but de fournir aux pays en voie de développement une stratégie vaccinale économe. Ses conséquences dynamiques sont comparées avec celles de la vaccination de masse.

Le dernier chapitre offre une conclusion générale et présente plusieurs niveaux de perspectives de recherches à ce travail de thèse.

Première partie

INTERACTIONS MOLÉCULAIRES FINES



Chapitre 2

Détection de l'adaptation moléculaire[1]

UNE des difficultés majeures dans la lutte contre les agents infectieux réside dans le fait que ceux-ci, comme tous les êtres vivants, évoluent. Cette capacité leur permet de s'adapter à de nouveaux hôtes [134], de contourner leurs défenses immunitaires et d'échapper aux drogues administrées. Comprendre en détail l'évolution des pathogènes est donc cruciale pour le développement de stratégies de contrôle efficaces [161]. Du fait de leur cycle de vie, les parasites peuvent présenter deux schémas d'évolution différents : une **évolution intra-hôte** qui s'opère au cours de la colonisation de l'organisme (voir figure 1.1 page 16) et une **évolution inter-hôte**, liée au processus d'infection, et qui s'opère lors de la propagation du parasite dans la population hôte (voir figure 1.1 page 16). Distinguer ces deux niveaux d'évolution n'est pas toujours aisé dans la pratique.

A l'échelle moléculaire, l'évolution intra-hôte des pathogènes est fortement influencée par le système immunitaire. Quantifier et localiser précisément l'adaptation moléculaire des pathogènes peut ainsi être de prime importance pour l'identification de cibles vaccinales potentielles. Avec la production de données génétiques de plus en plus complètes, des méthodes statistiques de plus en plus élaborées ont été mises au point pour détecter l'adaptation moléculaire. Dans ce chapitre, nous dressons une description critique des méthodes les plus couramment utilisées. Les deux chapitres qui suivent sont basés sur l'utilisation de l'une d'entre elles.

1. Le contenu de ce chapitre fera l'objet d'un article de revue (Choisy M., Robertson D.L. & Woelk C.H.) sollicité par le journal *Biological Procedures*.

2.1 Les types d'évolution chez les êtres vivants

La diversité génétique est à la base de l'évolution des êtres vivants. Cette diversité a plusieurs sources, la plus fondamentale étant la **mutation ponctuelle**, *i.e.* le changement d'une base en une autre, sur une séquence d'ADN ou d'ARN. Les mutations ponctuelles peuvent être provoquées par des forces extérieures comme la radioactivité, être spontanées – par instabilité chimique –, ou encore être le résultat d'erreurs au niveau de la réplication ou de la traduction du matériel génétique. Si la mutation ponctuelle est la source ultime de diversité génétique, cette dernière peut être grandement amplifiée à l'échelle du génome par des phénomènes de duplication, d'insertion/délétion, de recombinaison, ou encore de réassortiments de portions de gènes. Ces deux derniers phénomènes, recombinaison et réassortiment, ont pu évoluer non seulement parce qu'ils permettent de propager efficacement des combinaisons de mutations bénéfiques [139], mais aussi parce qu'ils permettent de purger le génome de mutations délétères qui peuvent s'accumuler lorsque les populations sont de petite taille [140].

2.1.1 Sélection purifiante

Les mécanismes énumérés ci-dessus constituent un important potentiel adaptatif. Toutefois, du fait de son caractère aléatoire, une mutation sur un gène lié à de fortes contraintes structurales et/ou fonctionnelles a toutes chances d'être **délétère**, diminuant ainsi la valeur sélective des individus qui la portent. La sélection naturelle aura alors tendance à diminuer la fréquence des porteurs de telles mutations. On parle dans ce cas de **sélection purifiante** ou encore **sélection négative**.

2.1.2 Sélection positive

Même si dans un contexte de fortes contraintes structurelles et fonctionnelles les mutations ont tendance à être contre-sélectionnées, il existe des situations où les mutations sont au contraire avantageuses. C'est le cas notamment pour les sites de reconnaissance d'antigènes où l'avantage de l'hétérozygote est de mise [93]. Dans cet exemple précis, la diversité génétique fait partie intégrante des contraintes s'exerçant sur le gène. Il peut arriver aussi que les contraintes elles-mêmes changent. C'est ce qui peut se produire, par exemple, après une duplication de gènes [144, 229], ou lorsqu'une espèce étend son aire de répartition à de nouveaux environnements. Là encore, les mutations auront tendance à être sélectionnées et leurs fréquences augmenteront dans la population. On parle dans ce cas de **sélection positive**, de

sélection diversifiante, ou encore d'**adaptation**.

2.1.3 Dérive génétique

Les deux types d'évolution décrits ci-dessus résultent de sélection, positive ou négative, imposée par des contraintes structurelles ou fonctionnelles fortes. On parle d'**évolution sous sélection** ou d'**évolution darwinienne**. En l'absence de telles contraintes, le matériel génétique peut évoluer au hasard, simplement au gré des mutations se produisant dans la population. Dans des populations de petite taille, le phénomène d'échantillonnage des génotypes entre chaque génération peut fortement amplifier l'effet de ces mutations. On parle dans ce cas de **dérive génétique** ou d'**évolution neutre**. Dans les années soixante-dix, Motoo KIMURA propose que, au niveau moléculaire, la dérive génétique est la principale force évolutive [86, 108, 106]. Dès lors, d'importants efforts ont été produits pour détecter de l'évolution darwinienne au niveau moléculaire [211, 181, 220, 56], en prenant comme hypothèse nulle le modèle neutre de KIMURA. C'est ce que nous allons détailler dans la suite du chapitre.

Pour résumer, on peut illustrer les trois types d'évolution en termes de fréquences alléliques, comme sur la figure 2.1. Dans le cas d'évolution positive, la sélection favorise l'augmentation (figure 2.1A) ou la diminution (figure 2.1B) d'une fréquence allélique. Dans le cas d'évolution négative, la pression de sélection exercée par les contraintes structurelles et fonctionnelles contre-sélectionne tout changement, maintenant une fréquence allélique fixe (figure 2.1C). Enfin, en l'absence de pression de sélection, l'évolution par dérive génétique suit le hasard des mutations (figure 2.1D).

2.2 Les premières méthodes de détection de la sélection positive

2.2.1 Méthodes basées sur les distributions alléliques

Les premières données génétiques établies étant sous forme de fréquences alléliques (essentiellement d'allozymes), il n'est pas surprenant que les premières méthodes proposées pour détecter la sélection positive utilisent ce type de données. La **méthode de LEWONTIN-KRAKAUER** [115] repose sur l'analyse de la distribution de F_{ST} (voir encadré 2.1 pour une définition) de plusieurs allèles et plusieurs locus. La variation des fréquences alléliques

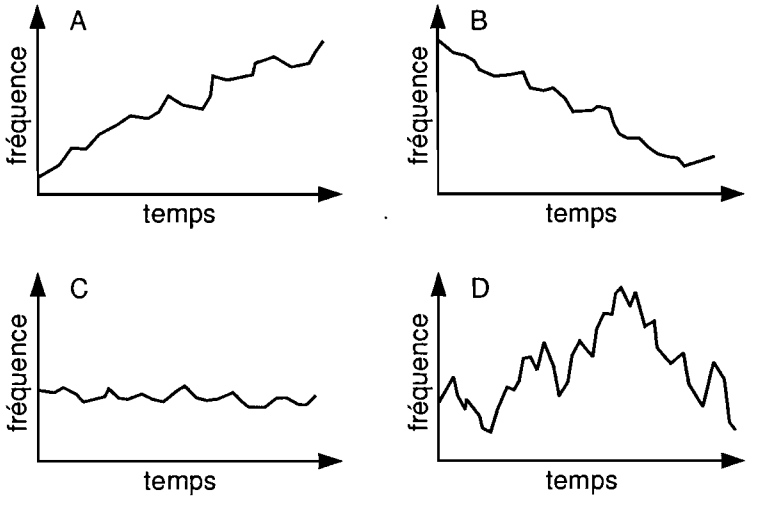


FIG. 2.1 – **Les types d'évolution chez les êtres vivants. L'exemple est pris sur l'évolution d'une fréquence allélique. Dans le cas d'évolution positive, la sélection favorise l'augmentation (A) ou la diminution (B) d'une fréquence allélique. Dans le cas d'évolution négative, la sélection exercée par les contraintes structurelles et fonctionnelles contre-sélectionne tout changement, maintenant une fréquence allélique fixe (C). Enfin, en l'absence de pression de sélection, l'évolution par dérive génétique suit le hasard des mutations (D) (mouvement brownien).**

ENCADRÉ 2.1
LA STATISTIQUE F_{ST} DE WRIGHT

Les statistiques de WRIGHT sont définies en termes de **probabilité d'identité par descendance** de deux gènes homologues. F_{IS} est un paramètre d'une population et mesure la corrélation des gènes à l'intérieur d'un individu diploïde. F_{ST} étend cette notion à une métapopulation[a] et mesure la corrélation des gènes à l'intérieur d'une sous-population[a]. C'est donc un paramètre mesurant le degré de structuration des populations. Si Q_1 est la probabilité d'identité par descendance de deux individus issus d'une même sous-population et Q_2 est cette probabilité pour deux individus issus de sous-populations différentes, alors F_{ST} s'écrit :

$$F_{ST} = \frac{Q_1 - Q_2}{1 - Q_2}$$

On peut également écrire F_{ST} en terme d'hétérozygotie :

$$F_{ST} = \frac{H_T - H_e}{H_e}$$

où H_e est l'hétérozygotie attendue dans la sous-population et H_T est l'hétérozygotie totale. Enfin, un estimateur non biaisé de F_{ST} est θ, défini en termes de composantes de la variance comme :

$$\theta = \frac{a}{a + b + c}$$

où a est la composante de la variance des fréquences alléliques entre individus de sous-populations différentes, b est la composante de la variance entre individus d'une même sous-population et c la composante de la variance entre gamètes d'un même individu. L'analyse de variance doit être pondérée par le nombre et la taille des échantillons.

[a]Voir encadré 5.1, page 98 pour des définitions de métapopulation et sous-population.

entre sous-populations peut être due à la structure de reproduction ou à la sélection. L'idée de la méthode de LEWONTIN-KRAKAUER repose sur le fait que, contrairement à la sélection, la structure de reproduction devrait influencer tous les locus et tous les allèles de la même façon. Ainsi, des différences significatives d'hétérogénéité entre locus peuvent être prises comme preuves de la sélection. En pratique, la distribution des F_{ST} sous le modèle neutre est calculée analytiquement ou par simulations de Monte Carlo[2]. La distribution de F_{ST} observée est ensuite comparée à la distribution attendue

2. Voir encadré 3.3 page 67 pour une définition.

sous le modèle neutre, et une différence statistiquement significative (test du χ^2) entre les deux distributions est indicative de sélection positive.

La **méthode de EWENS-WATTERSON** [210] est basée sur l'analyse de l'homozygotie F à un seul locus. Un test statistique est proposé, basé sur la vraisemblance des F observés, étant donné un modèle neutre à nombre d'allèles infinis (*Infinite Allele Model* en anglais).

2.2.2 Méthodes basées sur les distributions de variabilité génétique

Avec le développement des techniques d'acquisition des données génétiques (séquençages de gènes), les méthodes de détection de sélection positive se sont peu à peu basées sur l'analyse de variabilité génétique de séquences de nucléotides. La **méthode de HUDSON-KREITMAN-AGUADÉ** [91] compare le polymorphisme nucléotidique de deux locus à taux de mutation différents : un locus de référence (non-codant) et le locus d'intérêt (codant). Sous le modèle de neutralité, il est attendu que le rapport entre les deux polymorphismes soit proportionnel au rapport des taux de mutation. Cette méthode peut-être étendue à la comparaison de plusieurs locus, non seulement de la même espèce mais, aussi d'espèces différentes.

La **méthode de TAJIMA** [198] et la **méthode de FU-LI** [65] utilisent les séquences nucléotidiques d'un seul locus et sont basées sur l'estimation de la variabilité génétique $\theta = 4N_e\mu$, où N_e est la taille efficace de la population et μ le taux de mutation par génération. La variabilité génétique θ peut être estimée en utilisant le nombre moyen $\hat{k}$ de différences nucléotidiques estimé à partir de comparaisons de séquences deux à deux [197], le nombre total η de mutations le long d'une généalogie de gènes [65], ou simplement le nombre S de sites[3] polymorphes [209]. En l'absence de sélection (modèle neutre), il est attendu que les estimations de θ utilisant S, η ou $\hat{k}$ soient les mêmes. En revanche, en présence de sélection, on peut s'attendre à des différences entre les trois estimations de θ. La significativité d'une telle différence est indicative de sélection positive. La méthode de TAJIMA compare les estimations de θ obtenues en utilisant S et $\hat{k}$, tandis que la méthode de FU-LI compare les estimations de θ obtenues en utilisant S et η.

3. De façon générale, on appelle « site » l'unité évolutive qui nous intéresse. Dans le contexte présent, il s'agit d'une base nucléotidique sur une séquence d'ADN ou d'ARN. Nous verrons par la suite que nous pourrons considérer comme sites les codons d'une séquence nucléotidique ou les acides aminés d'une protéine.

2.3 Méthodes récentes

Les méthodes de détection de sélection positive les plus récentes ont été développées pour étudier les gènes codant des protéines, et sont basées sur l'étude des taux de substitution **synonyme** et **non-synonyme**. Du fait de la dégénérescence du code génétique (tableau 2.1), certaines substitutions nucléotidiques modifient l'acide aminé codé, tandis que d'autres ne le modifient pas. Les premières sont appelées non-synonymes et les secondes, synonymes.

TAB. 2.1 – **Le code génétique. Tout changement de base à la seconde position et la plupart des changements à la première position sont non-synonymes. Seulement quelques changements de base à la troisième position sont synonymes.**

Codon[a]	AA[b]	Codon[a]	AA[b]	Codon[a]	AA[b]	Codon[a]	AA[b]
UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
UUA	Leu	UCA	Ser	UAA	Stop	UGC	Stop
UUG	Leu	UCG	Ser	UAG	Stop	UGC	Trp
CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
CUC	Leu	CCC	Pro	CAC	His	CGC	Arg
CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg
GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
GUC	Val	GCG	Ala	GAC	Asp	GGC	Gly
GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

[a]Les bases sont celles de l'ARN : adénine (A), cystosine (C), guanine (G) et uracile (U).
[b]Les acides aminés (AA) sont indiqués par leur code à trois lettres. Ala : alanine ; Arg : arginine ; Asn : asparagine ; Asp : aspartate ; Cys : cystéine ; Gln : glutamate ; Gly : glycine ; His : histidine ; Ile : isoleucine ; Leu : leucine ; Lys : lysine ; Met : méthionine ; Phe : phénylalanine ; Pro : proline ; Ser : sérine ; Thr : thréonine ; Trp : tryptophane ; Tyr : tyrosine ; Val : valine ; Stop : codon stop.

La première méthode à s'être basée sur les taux de substitution synonyme et non-synonyme est la **méthode de McDonald-Kreitman** [125], directement inspirée de celle de Hudson-Kreitman-Aguadé [91]. La différence principale avec cette dernière est qu'au lieu de comparer le polymorphisme nucléotidique entre un gène codant d'intérêt et un gène non-codant de référence, ce test compare le polymorphisme génétique entre les substitutions synonymes et non-synonymes d'un même gène codant d'intérêt. En l'absence de sélection, les rapports des taux de substitution synonyme et non-synonyme devraient être les mêmes entre espèces différentes et au sein d'une même espèce. Une différence significative (par test G ou test exact de Fisher) constitue une preuve de sélection.

Les autres méthodes sont explicitement basées sur la comparaison des taux de substitution synonyme (d_S) et non-synonyme (d_N). L'idée est que les substitutions nucléotidiques s'accumulent à un taux qui est essentiellement déterminé par la dérive génétique et la sélection naturelle. L'hypothèse sous-jacente à ces méthodes est que la sélection naturelle s'opère majoritairement au niveau de la protéine, et seulement légèrement au niveau de l'ADN et de l'ARN. Ceci suppose également que le taux de substitution synonyme reflète le taux de mutation seul, sans sélection [131] (mais voir Réf. [5]) et que le taux de substitution non-synonyme rend compte de la sélection au niveau de la protéine. Sous ces hypothèses, la détection de sélection positive revient donc à estimer des taux d_N et d_S. Pour cela, il existe deux grandes familles de méthodes. La première et la plus ancienne regroupe des méthodes *ad hoc* et intuitives [155, 131, 146, 109, 94, 229, 192, 223]. Les méthodes de la deuxième famille reposent sur des bases statistiques plus solides puisqu'elles sont basées sur l'expression d'un modèle explicite d'évolution moléculaire [70, 142] et estiment les paramètres par maximum de vraisemblance [141, 147, 225].

Dans la suite, nous présentons les deux grandes familles de méthodes. Nous comparons ensuite les avantages et inconvénients des deux méthodes. Avant cela, nous introduisons le lecteur au modèle de substitution de nucléotide développé par Goldman et Yang [70]. Ceci nous servira de cadre général pour comprendre les deux méthodes. Aussi, c'est le modèle explicitement utilisé dans les méthodes de maximum de vraisemblance.

2.3.1 Un modèle markovien de substitution de codons

Les modèles d'évolution moléculaire ont été développés en phylogénie afin de corriger pour les **substitutions multiples**. Ce sont des **modèles pro-**

babilistes basés sur des **processus de Markov**[4] à états dénombrables et temps discret. Les modèles d'évolution de séquences nucléotidiques contiennent quatre états – les quatre bases A, C, G, T (ou U), voir encadré 2.2 – et les modèles d'évolution de séquences protéiques contiennent vingt états, un pour chaque acide aminé. Pour une séquence codante, un modèle d'évolution nucléotidique est inadapté car les trois bases d'un codon ne sont pas indépendantes. En revanche, l'utilisation d'un modèle d'évolution de protéine est malheureuse car l'information statistique contenue au niveau de la séquence d'acide nucléotidique est alors perdue. Le **modèle d'évolution de codons** proposé par GOLDMAN et YANG [70] permet de combiner les avantages des modèles d'évolution de séquences nucléotidique et protéique. Ce modèle comprend 61 états, qui sont les 61 codons ayant un sens (*i.e.* qui ne sont pas des codons « stop », voir tableau 2.1).

Le processus de Markov est caractérisé par une **matrice génératrice** $\mathbf{Q} = \{q_{ij}\}$, où q_{ij} est le taux de substitution d'un codon i en un codon j ($i \neq j$). Plus formellement, $q_{ij}\Delta t$ est la probabilité que le processus se trouve dans l'état j après un temps infiniment petit Δt, sachant qu'il était dans l'état i à l'instant t. La matrice génératrice est définie par

$$q_{ij} = \begin{cases} 0 & \text{si les codons } i \text{ et } j \text{ diffèrent à plus d'une position} \\ \pi_j & \text{pour les transversions synonymes} \\ \kappa\pi_j & \text{pour les transitions synonymes} \\ \omega\pi_j & \text{pour les transversions non-synonymes} \\ \omega\kappa\pi_j & \text{pour les transitions non-synonymes} \end{cases}$$

où $\kappa = T_S/T_V$ est le rapport entre les taux de transition[5] et de transversion[6], $\omega = d_N/d_S$ est le rapport des taux de substitution synonyme et non-synonyme, et π_j est la fréquence à l'équilibre du codon j, simplement calculée comme le produit des fréquences nucléotidiques à chacune des trois positions du codon. Les paramètres κ et π_j caractérisent les processus au niveau de l'ADN (mutations) tandis que le paramètre ω caractérise la sélection au niveau de la protéine. Les éléments de la diagonale sont calculés par la contrainte mathématique que les lignes de la matrice somment à 0 [77] :

$$\sum_j q_{ij} = 0, \forall i$$

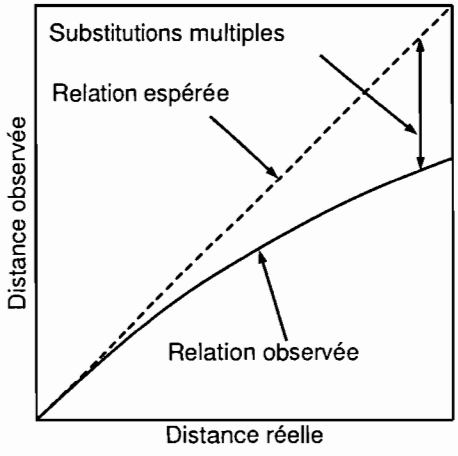
4. Un processus de Markov est un processus aléatoire dont les probabilités futures sont déterminées par les probabilités les plus récentes : $P(x(t_n) \leq x_n | x(t), \forall t \leq t_{n-1}) = P(x(t_n) \leq x_n | x(t_{n-1}))$.

5. Une transition est une substitution d'une base purique (A et G) en une base purique ou d'une base pyrimidique (C et T) en une base pyrimidique.

6. Une transversion est une substitution d'une base purique par une base pyrimidique ou *vice versa*.

ENCADRÉ 2.2
MODÈLES D'ÉVOLUTION DE SÉQUENCES NUCLÉOTIDIQUES

La première étape dans l'analyse d'un alignement de séquences nucléotidiques consiste généralement en la détermination des **distances génétiques** (ou **distances évolutives**) entre chacune de ces séquences. Une façon simple et intuitive d'exprimer une telle distance est de compter le nombre de différences nucléotidiques entre deux séquences. Cette méthode simple souffre néanmoins d'un défaut majeur : elle ne tient pas compte des substitutions multiples et sous-estime donc la vraie distance génétique.



Comme on ne peut évidemment pas connaître le nombre de substitutions multiples, l'idée est de la modéliser au moyen d'un modèle probabiliste. Les modèles proposés sont basés sur des processus de Markov à quatre états (un pour chaque nucléotide). La forme la plus générale de la matrice de substitution de tels modèles est celle du modèle GTR (*general time-reversible*) :

$$\mathbf{P}_t = \begin{bmatrix} \cdot & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & \cdot & d\pi_G & e\pi_T \\ b\pi_A & d\pi_C & \cdot & f\pi_T \\ c\pi_A & e\pi_C & f\pi_G & \cdot \end{bmatrix}$$

La première ligne contient les taux de substitution de la base A vers les bases A, C, G et T respectivement. *Idem* pour les trois autres lignes mais pour les bases C, G, et T respectivement. Les lettres a à f reflètent les probabilités de transition d'une base vers une autre et les π_i, $i \in \{A, C, G, T\}$, traduisent les biais d'utilisation des bases. Les éléments de la diagonale, non représentés, sont déterminés par la condition mathématique que les lignes doivent sommer à 0. Ce modèle a 8 degrés de liberté et les autres modèles proposés peuvent être considérés comme des cas particuliers de celui-ci.

(Encadré 2.2 suite)

Hasegawa, Kishino et Yano (1985) – Les transitions (α) et transversions (β) sont différenciées, ainsi que les biais dans l'utilisation des bases [87].

$$\mathbf{P}_t = \begin{bmatrix} \cdot & \beta\pi_C & \alpha\pi_G & \beta\pi_T \\ \beta\pi_A & \cdot & \beta\pi_G & \alpha\pi_T \\ \alpha\pi_A & \beta\pi_C & \cdot & \beta\pi_T \\ \beta\pi_A & \alpha\pi_C & \beta\pi_G & \cdot \end{bmatrix}$$

Felsenstein (1981) – Les transitions et transversions ne pas sont différenciées. Seuls les biais dans l'utilisation des bases sont pris en compte [57].

$$\mathbf{P}_t = \begin{bmatrix} \cdot & \alpha\pi_C & \alpha\pi_G & \alpha\pi_T \\ \alpha\pi_A & \cdot & \alpha\pi_G & \alpha\pi_T \\ \alpha\pi_A & \alpha\pi_C & \cdot & \alpha\pi_T \\ \beta\pi_A & \alpha\pi_C & \alpha\pi_G & \cdot \end{bmatrix}$$

Kimura (1980) – Les transitions (α) et transversions (β) sont différenciées mais les biais dans l'utilisation des bases ne sont pas pris en compte [107].

$$\mathbf{P}_t = \begin{bmatrix} \cdot & \beta & \alpha & \beta \\ \beta & \cdot & \beta & \alpha \\ \alpha & \beta & \cdot & \beta \\ \beta & \alpha & \beta & \cdot \end{bmatrix}$$

Jukes et Cantor (1969) – Toutes les substitutions sont équiprobables [99].

$$\mathbf{P}_t = \begin{bmatrix} \cdot & \alpha & \alpha & \alpha \\ \alpha & \cdot & \alpha & \alpha \\ \alpha & \alpha & \cdot & \alpha \\ \alpha & \alpha & \alpha & \cdot \end{bmatrix}$$

Les substitutions à plus d'une position dans un seul codon ne sont pas permises car, en considérant l'indépendance entre chaque nucléotide de la séquence, de tels événements sont hautement improbables. A partir de cette matrice génératrice $\mathbf{Q}$, une **matrice de transition P** est calculée comme

$$\mathbf{P}(t) = \{p_{ij}(t)\} = \mathrm{e}^{\mathbf{Q}t}$$

2.3.2 Méthodes *ad hoc*

Méthodes basées sur la comparaison de deux séquences nucléotidiques

Dans les premières méthodes développées, l'estimation des taux d_N et d_S est très simple et intuitive et implique la comparaison de séquences nucléotidiques prises deux à deux. Ces méthodes comportent toutes trois étapes successives.

1. La première étape consiste en la détermination des nombres **attendus** de sites nucléotidiques synonymes (S) et non-synonymes (N) dans les deux séquences, en l'absence de sélection. En effet, même en l'absence de sélection, on ne s'attend pas à observer des proportions équivalentes de sites synonymes et non-synonymes et ces proportions peuvent dépendre notamment des biais π_j dans l'utilisation des bases nucléotidiques et des codons, ainsi que du rapport κ entre les taux de transition et de transversion [223]. Par exemple, la troisième position du codon CTT codant pour la leucine est synonyme parce que toutes les substitutions possibles T→C, T→A, and T→G sont synonymes alors que les premières et secondes positions de CTT sont non-synonymes (voir tableau 2.1).
2. La deuxième étape détermine les nombres C_S et C_N de substitutions synonymes et non-synonymes effectivement **observés**. Ceci se fait par la comparaison des séquences prises deux à deux. Cette opération est assez aisée lorsque les codons ne diffèrent que d'une seule base. Lorsque les codons diffèrent à deux ou trois bases, il existe quatre ou six **chemins évolutifs** possibles d'un codon à l'autre. Chaque chemin doit être pondéré de façon appropriée.
3. La dernière étape applique une **correction** pour les substitutions multiples, par des modèles de substitution de nucléotides comme le modèle de JUKES et CANTOR [99] ou le modèle à deux paramètres de KIMURA [107] (voir encadré 2.2). Les taux de substitution synonymes (d_S) et non-synonymes (d_N) sont ensuite déterminés comme les rapports S/C_S et N/C_N respectivement.

La statistique

$$Z = \frac{d_N - d_S}{\sqrt{V(d_N) + V(d_S)}}$$

est ensuite considérée et sa distribution est supposée normale. V réfère à la variance. Z significativement positif constitue une preuve de sélection positive, Z significativement négatif est une preuve de sélection négative tandis

que Z non significativement différent de 0 est un cas d'évolution neutre. Parmi ces méthodes, citons par exemple celles de PERLER et collaborateurs [155], MIYATA et YASUNAGA [131], NEI et GOJOBORI [146], KONDO et collaborateurs [109], INA [94], ZHANG et collaborateurs [229], ou celle de YANG et NIELSEN [223]. Elles se distinguent essentiellement par les modèles utilisés pour estimer N et S (étape 1, notamment prise en compte des taux de transition/transversion, des biais dans l'utilisation des bases/codons, *etc...*), par la pondération de chaque chemin évolutif lors du comptage des substitutions non-synonymes et par les modèles utilisés pour corriger pour les substitutions multiples (étape 3).

Méthodes utilisant un alignement de séquences nucléotidiques

Les événements de sélection positive sont généralement très localisés dans le temps (le long de l'arbre phylogénétique) et dans l'espace (le long du gène). Or, la méthode générale présentée ci-dessus estime en réalité des moyennes à la fois temporelles et spatiales des taux de substitution synonymes et non-synonymes. Si les événements de sélection positive sont rares, ils risquent alors de ne pas être détectés. Une façon *ad hoc* de palier à ce manque de puissance est d'effectuer les estimations sur des portions de gènes et des portions de l'arbre phylogénétique [93, 130].

La **méthode de SUZUKI et GOJOBORI** [192] permet d'estimer les taux de substitution synonyme et non-synonyme à l'échelle d'un site[7] de la séquence. Ceci est rendu possible par la généralisation de la méthode présentée ci-dessus à un alignement de plusieurs séquences. Un arbre phylogénétique de l'alignement est préalablement calculé et les séquences des nœuds ancestraux sont estimées (par maximum de parcimonie ou maximum de vraisemblance). Ensuite, les trois étapes précédemment décrites sont appliquées entre chaque nœud de l'arbre, et ce en corrigeant les nombres N et S par la longueur de la branche. Enfin, les taux de substitution synonyme et non-synonyme de chaque site sont déterminés en prenant la moyenne sur tout l'arbre phylogénétique et le test statistique de neutralité est appliqué. Sous l'hypothèse de neutralité, on a $C_S/(C_S+C_N) = S/(S+N)$ et $C_N/(C_N+C_S) = N/(N+S)$. Les probabilités de changement non-synonyme et synonyme à un site particulier sont $S/(S+N)$ et $N/(N+S)$ respectivement. En faisant l'hypothèse que C_N et C_S suivent une distribution binomiale, l'hypothèse nulle de neutralité peut alors être testée à chaque site.

7. Dans ce cadre, on appelle « site » un codon, i.e. un groupe de trois bases sur une séquence d'acides nucléotidiques.

2.3.3 Méthodes basées sur le maximum de vraisemblance

Ziheng YANG et ses collaborateurs ont développé des méthodes d'estimation du rapport d_N/d_S dans un cadre général de vraisemblance. L'avantage principal de ces méthodes par rapport aux méthodes *ad hoc* est qu'elles présentent une base statistique solide. L'utilisation de la vraisemblance requiert l'expression explicite d'un modèle stochastique d'évolution. Le modèle utilisé dans ces méthodes est le modèle d'évolution de GOLDMAN et YANG [70], présenté au paragraphe **??**. L'utilisation d'un tel modèle rend aisée la gestion de l'**hétérogénéité de la sélection positive** le long d'un arbre phylogénétique ou d'une séquence nucléotidique, augmentant ainsi la puissance dans la détection de la sélection positive. Comme les méthodes *ad hoc*, les méthodes de maximum de vraisemblance se sont développées par étapes. Les principales étapes sont résumées dans le tableau 2.2.

TAB. 2.2 – **Les principales étapes dans le développement des méthodes de détection de sélection positive basées sur le maximum de vraisemblance.**

Auteurs	Année	Description	Réf.
Goldman & Yang	1994	Modèle d'évolution des codons.	[70]
Yang & Nielsen	1998	Test de rapport de vraisemblance pour détecter la sélection positive à l'échelle d'un gène entier. Procède par comparaison de séquences deux à deux.	[224]
Nielsen & Yang	1998	Généralisation de Yang & Nielsen 1998 à un alignement de séquences. Le Bayesien empirique permet d'identifier les sites sous sélection positive.	[147]
Yang	1998	Modèle rendant compte de l'hétérogénéité de la sélection positive le long d'un arbre phylogénétique.	[218]
Yang, Nielsen, Goldman & Pedersen	2000	14 modèles rendant compte de l'hétérogénéité de la sélection positive le long de la séquence nucléotidique.	[225]
Yang & Nielsen	2002	Prise en compte de l'hétérogénéité de la sélection positive à la fois le long de l'arbre phylogénétique et le long de la séquence nucléotidique.	[220]

Le noyau dur de la méthode est traité dans l'article de YANG, NIELSEN, GOLDMAN et PEDERSEN de 2000 [225]. Il existe également de bonnes revues exposant clairement le principe général de la méthode [222, 219, 221]. Cette

méthode comprend trois étapes successives comme expliqué dans [222]. La première étape consiste en l'estimation de la proportion de sites sous sélection positive, ainsi que de l'intensité de la sélection positive. La deuxième étape teste statistiquement la significativité d'une éventuelle sélection positive à l'échelle du gène étudié, pris en entier. La troisième étape permet de localiser les sites du gène sous sélection positive.

Première étape : estimations à l'échelle du gène entier

Puisque l'unité évolutive est le codon, l'idée est d'utiliser un modèle probabiliste de substitution de codon. C'est le modèle de GOLDMAN et YANG [70], présenté au paragraphe 2.3.1, qui est utilisé. Les paramètres π_j, κ et ω du modèle d'évolution de codons sont estimés par **maximum de vraisemblance** en utilisant l'algorithme général de FELSENSTEIN [57]. Ceci permet donc d'estimer le rapport $d_N/d_S = \omega$ que nous cherchons. Estimer un rapport d_N/d_S pour chaque codon serait excessivement long et conduirait à une surparamétrisation. L'approche adoptée consiste en fait à considérer un nombre fixe de classes de valeurs de d_N/d_S. On doit alors estimer la valeur du rapport d_N/d_S de chaque classe ainsi que la proportion de sites dans chaque classe.

Au début, YANG et collaborateurs ont proposé pas moins de 14 modèles différents et de complexité[8] croissante (de M0 à M13) pour rendre compte de la distribution des d_N/d_S le long de l'alignement de séquences [225]. En pratique, estimer les paramètres de chacun de ces 14 modèles est extrêmement long. Fort heureusement, il s'avère que 6 (M0, M1, M2, M3, M7 et M8, tous de complexité raisonnable) de ces 14 modèles suffisent pour estimer efficacement la sélection positive sur un alignement de séquences nucléotidiques [225]. Les modèles M0, M1 et M7 sont des modèles nuls dans le sens où leurs valeurs de d_N/d_S dans chacune des classes ne permettent pas de rendre compte d'une sélection positive (toutes les classes de d_N/d_S sont inférieures ou égales à 1). Au contraire, les modèles M2, M3 et M8 peuvent rendre compte de sélection positive en permettant une classe de d_N/d_S d'être supérieure à 1.

Le modèle M0 estime une seule valeur de d_N/d_S (entre 0 et 1) pour tous les sites de l'alignement de séquences. Le modèle M1 fixe 2 valeurs de $\omega = d_N/d_S$: $\omega_0 = 0$ et $\omega_1 = 1$ et estime la proportion de sites ayant chacune de ces deux valeurs. Par rapport au modèle M1, le modèle M2 rajoute une classe de valeur pour d_N/d_S, ω_2, dont la valeur, estimée, peut être supérieure à 1. Par rapport au modèle M2, le modèle M3 à 3 classes estime les valeurs de d_N/d_S dans chacune des classes de valeurs, avec comme contrainte que seule la dernière permet à d_N/d_S d'être supérieure à 1. Ce modèle M3 peut

8. La complexité d'un modèle se traduit par son nombre de degrés de liberté.

être généralisé à K classes. Le modèle M7 considère que les valeurs de d_N/d_S le long de la séquence suivent une distribution beta[9]. Cette distribution est discrétisée en 10 classes, chacune ayant une probabilité de 1/10. Les paramètres du modèle M7 sont donc simplement les deux paramètres de forme de la distribution beta. Enfin, par rapport au modèle M7, le modèle M8 rajoute une onzième classe dont la valeur de d_N/d_S, ω_{11}, estimée, peut être supérieure à 1. Le tableau 2.3 résume ces informations, ainsi que le nombre de paramètres à estimer associés à chacun des 6 modèles.

TAB. 2.3 – **Caractéristiques des modèles les plus utilisés pour la détection de sélection positive. D'après [225].**

Modèle	p[a]	Paramètres	Commentaires
M0	1	ω	une seule valeur de d_N/d_S pour tous les sites
M1	1	p_0	$p_1 = 1 - p_0$, $\omega_0 = 0$, $\omega_1 = 1$
M2	3	p_0, p_1, ω_2	$p_2 = 1 - p_0 - p_1$, $\omega_0 = 0$, $\omega_1 = 1$
M3	$2K - 1$ $(K = 3)$	$p_0, p_1, \ldots, p_{K-2}$ $\omega_0, \omega_1, \ldots, \omega_{K-2}$	$p_{K-1} = 1 - p_0 - p_1 - \ldots - p_{K-2}$
M7	2	p, q	les valeurs de d_N/d_S suivent une distribution beta $\mathcal{B}(p, q)$
M8	4	p_0, p, q, ω_{11}	p_0 est la proportion de sites dans les 10 premières classes, modélisées par $\mathcal{B}(p, q)$. $1 - p_0$ est la proportion de sites avec ω_{11} comme valeur de d_N/d_S

[a]**Nombre de paramètres du modèle.**

Deuxième étape : test statistique de la sélection positive

On peut comparer statistiquement deux modèles par analyse de leur vraisemblance. La procédure classique, lorsque les modèles sont emboîtés, est le **test de rapport de vraisemblance**. Deux modèles sont dits **emboîtés** lorsque l'un est un cas particulier de l'autre. Par exemple, le modèle M2 est un cas particulier du modèle M3 où $p_2 = 0$. Dans ce cas, on peut montrer que le rapport des vraisemblances des deux modèles suit une loi de χ^2 dont le nombre de degrés de liberté est égal à la différence des nombres de paramètres des deux modèles. Ici, en comparant à chaque fois un modèle permettant la sélection positive avec un modèle nul ne permettant pas la sélection positive,

9. La distribution beta est une distribution de probabilité à deux paramètres et définie sur le segment [0 ; 1]. Cette distribution peut prendre une importante variété de formes. La forme dépend des valeurs des deux paramètres.

on peut tester la significativité de la sélection positive sur le gène étudié. Parmi les 6 modèles présentés dans le tableau 2.3, M0 et M1 sont des cas particuliers de M2 et M3, M2 est un cas particulier de M3, et M7 est un cas particulier de M8. Ceci nous définit donc six tests de rapport de vraisemblance : M0 *vs* M2, M0 *vs* M3, M1 *vs* M2, M1 *vs* M3, M2 *vs* M3 et M7 *vs* M8.

Troisième étape : Identification des sites sous sélection positive

Une fois qu'un test de rapport de vraisemblance conclut à la présence de sélection positive sur un gène, la dernière étape consiste en l'identification précise des sites de la séquence sous sélection positive. Cette étape fait appel aux **statistiques bayesiennes**. Ces statistiques sont basées sur la formule de Bayes des probabilités conditionnelles (voir [184] et [22] pour des introductions et [166] pour plus de détails techniques) :

$$p(\theta|X) = \frac{p(\theta)p(X|\theta)}{p(X)}$$

Si θ est un vecteur de paramètres et X un vecteur de données, on voit que données et paramètres ont un rôle exactement symétrique[10]. La puissance des statistiques bayesiennes réside dans le fait qu'il est possible d'exprimer la distribution d'un paramètre $p(\theta|X)$ en tenant compte des informations que l'on peut avoir *a priori* sur ce paramètre. Cette information *a priori* est contenue dans la distribution *a priori* $p(\theta)$ du paramètre. $p(X|\theta)$ est homogène à la vraisemblance et $p(X)$ est la probabilité des données. Cette dernière quantité peut sembler difficile à exprimer. En fait il suffit de remarquer que $p(X) = p(X|\theta)p(\theta) + p(X|\overline{\theta})p(\overline{\theta})$[11]. Donc, de façon générale, on a

$$p(\theta_i|X) = \frac{p(\theta_i)p(X|\theta_i)}{\sum_{j=1}^{n} p(X|\theta_j)p(\theta_j)}$$

L'identification des sites sous sélection positive se fait par l'utilisation d'un **bayesien empirique**. Les modèles de sélection positive de l'étape précédente nous donnent la proportion de sites de la séquence présents dans chacune des classes de valeur de d_N/d_S. Le bayesien empirique utilise ces proportions comme distribution *a priori* : chaque site de la séquence a donc la même probabilité d'appartenir à chacune des classes de valeur de d_N/d_S. La formule de Bayes nous permet de « redistribuer » ces probabilités sur les différents sites de la séquence pour en déduire la distribution *a posteriori*

10. Ce qui n'est pas le cas dans les statistiques classiques de Neyman-Pearson.
11. La barre horizontale signifie l'événement complémentaire.

du rapport d_N/d_S. Avec cette distribution, tous les sites n'ont pas la même probabilité d'appartenir aux différentes classes de valeur de d_N/d_S. La figure 2.2 illustre ce passage de la probabilité *a priori* à la probabilité *a posteriori* sur un exemple avec un modèle M3 à trois classes.

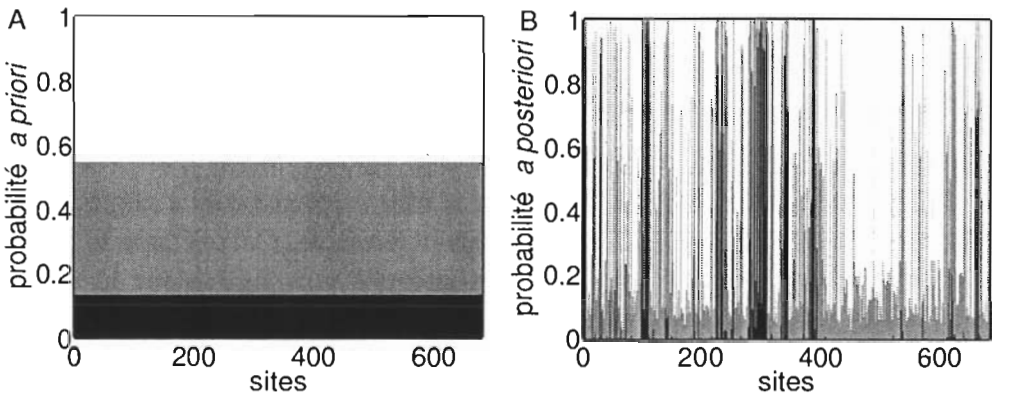


FIG. 2.2 – **Application de la formule de Bayes sur un modèle M3 à trois classes. Le modèle M3 à trois classes considère trois classes de valeurs du rapport d_N/d_S et estime la proportion de sites dans chacune de ces trois classes (voir tableau 2.3). La première classe ($\omega_0 = 0.0165$, blanc) contient 45.18% des sites, la deuxième classe ($\omega_1 = 0.2061$, gris) contient 41.43% des sites et la troisième classe ($\omega_2 = 2.7559$, noir) contient 13.39% des sites. (A) : la distribution *a priori* considère que les 684 sites de la séquence ont les mêmes probabilités d'appartenir à chacune des trois classes, à savoir 0.4518, 0.4143 et 0.1339 pour ω_0, ω_1 et ω_2 respectivement. (B) : La formule de Bayes permet de « redistribuer » ces probabilités parmi les 684 sites de la séquence, mettant en évidence des sites ayant une forte probabilité *a posteriori* d'appartenir à la classe $\omega_2 > 1$ et donc d'être sous sélection positive (en noir).**

2.3.4 Comparaison des méthodes

Les méthodes de détection de sélection positive considérant un d_N/d_S moyen le long de l'arbre phylogénétique ou le long de la séquence nucléotidique sont très peu puissantes, et si les événements de sélection positive sont rares, ils risquent fort de ne pas être détectés [181]. Les méthodes les plus utilisées actuellement sont donc celles gérant une hétérogénéité sur d_N/d_S, c'est à dire la méthode de SUZUKI et GOJOBORI [192] pour les méthodes *ad hoc* et la méthode de YANG et collaborateurs [225] pour les méthodes de maximum de vraisemblance. Cependant, le choix entre l'une ou l'autre n'est pas univoque, chacune semble avoir ses avantages et inconvénients, et

les critiques de l'une ou de l'autre sont encore vives dans la littérature. Dans ce paragraphe, nous allons essayer de comparer les points forts et les points faibles de chacune des deux méthodes.

Avantages des méthodes de maximum de vraisemblance

- Bases statistiques solides. Le test de rapport de vraisemblance rend aisé le test de très nombreuses hypothèses [219].
- Formulation explicite d'un modèle d'évolution. On a ainsi une idée claire des mécanismes évolutifs sous-jacents [219].
- Test de rapport de vraisemblance performant, même avec des séquences courtes, de l'ordre de 100 ou 200 nucléotides seulement. Pour des séquences plus courtes, la distribution de la statistique peut être simulée par Monte Carlo [219].
- Pas d'estimation des séquences ancestrales requise. Évite les erreurs dues aux imprécisions sur les séquences ancestrales [219].
- Gestion de plusieurs problèmes en une seule fois : biais dans l'utilisation des bases/codons, taux de transition/transversion, chemins évolutifs multiples, substitutions multiples, *etc...* [219].
- Réversibilité temporelle du processus de Markov assurée par le théorème de Chapman-Kolmogorov. Ainsi, les temps de divergence estimés par la fonction de vraisemblance rendent compte de tous les chemins évolutifs possibles en les pondérant de façon appropriée, selon leurs probabilités respectives [219].
- Prise en compte explicite de la phylogénie entre les séquences. Les résultats sont assez robustes par rapport à la forme exacte de la phylogénie [225].
- Gestion de toutes les séquences dans leur ensemble. Pas de comparaison de séquences deux à deux [219].
- Possibilité de détection des branches de l'arbre phylogénétique sous sélection positive, et ce conjointement avec la détection des sites de la séquence sous sélection positive [218, 220].
- Puissance dans la détection de la sélection positive, même avec relativement peu de séquence et peu de nucléotides dans les séquences. La puissance et la précision augmentent toutefois avec le nombre et la longueur des séquences [12, 13].

Inconvénients des méthodes de maximum de vraisemblance

- Lourd et long à implémenter.
- Augmentation du nombre de pics de vraisemblance avec le nombre de

paramètres du modèle. Parfois difficulté de savoir si l'algorithme d'optimisation a convergé sur un maximum absolu. Nécessité donc souvent de relancer plusieurs fois l'algorithme avec des valeurs initiales différentes [193]. Rend donc la méthode encore plus lourde et plus longue à implémenter.
- Formulation d'hypothèses liées au modèle. Il faut avoir une certaine confiance dans le modèle ou que les estimations soient assez robustes par rapport aux hypothèses du modèle. En particulier, rien ne semble justifier telle ou telle forme pour la distribution des rapports d_N/d_S [227] (mais voir [12, 13]).
- Pas d'estimation des taux d_N et d_S séparément, mais seulement de leur rapport. Ceci fait l'hypothèse implicite que les d_S sont constants le long de l'arbre phylogénétique et le long de la séquence nucléotidique [219].
- Tendance à être trop libéral et à identifier des « faux positifs », *i.e.* des sites identifiés comme positifs alors qu'ils ne le sont pas [193, 194, 195].

Avantages des méthodes *ad hoc*

- Aucune formulation d'hypothèses. Les résultats ne risquent donc pas d'être dus à des hypothèses fausses [227, 193, 194, 195].
- Très rapide à implémenter.
- Assez conservateur. On peut donc avoir une bonne confiance dans la sélection positive détectée par cette méthode [193, 194, 195].
- Distances phylogénétiques entre les différentes séquences prises en compte [192].
- Estimation des taux d_N et d_S séparément. Ne fait donc pas l'hypothèse que les d_S sont constants le long de l'arbre phylogénétique et le long de la séquence nucléotidique [146].

Inconvénients des méthodes *ad hoc*

- Pas de bases statistiques solides. Les méthodes utilisées sont intuitives et n'ont aucune justification théorique [219].
- Nécessite beaucoup de séquences dans l'alignement pour diminuer la variance sur les estimations [192, 195].
- Nécessite des longueurs de branches courtes entre les diverses séquences de l'alignement [192, 195].
- Nécessité d'inférer les séquences ancestrales. Une incertitude réside à ce niveau-là [219]. Les séquences de deux nœuds adjacents sont comparées deux à deux. Cette approche est moins performante que la comparaison de toutes les séquences simultanément, comme fait par la méthode du

maximum de vraisemblance [219].
- Mauvaise gestion des biais dans l'utilisation des bases/codons, les taux de transition/transversion, *etc...* La mauvaise gestion de ces paramètres peut grandement influencer les résultats [219].
- Détecter les branches de l'arbre phylogénétique sous sélection positive n'est pas possible, contrairement à la méthode du maximum de vraisemblance [218, 220].
- Manque de puissance dans la détection de sélection positive [223].

Ajoutons à ceci que, en présence de recombinaison, différentes portions d'une séquence évoluent selon différents arbres phylogénétiques. Le phénomène de recombinaison est donc susceptible d'affecter les résultats des deux méthodes détaillées ici puisqu'elles sont basées sur l'existence d'une seule phylogénie. L'effet de la recombinaison sur la détection de sélection positive n'a pas été étudié pour les méthodes *ad hoc*. Pour les méthodes de vraisemblance il a été montré [14] que la recombinaison pouvait effectivement fausser les résultats, celle-ci étant souvent interprétée comme de la sélection positive. Des travaux de simulations montrent que cet effet reste toutefois modéré pour des niveaux de recombinaison faibles (moins de trois événements de recombinaison dans l'histoire phylogénétique de dix séquences), et le risque de détection de faux positifs diminue lorsque le réalisme (et donc la complexité) des modèles utilisés augmente (par exemple les modèles M7 et M8 sont assez performants). Le problème de la recombinaison devient un problème sérieux surtout pour les gènes de virus pour lesquels la divergence entre séquences peut être importante et la recombinaison fréquente (voir chapitre suivant).

Les modèles et algorithmes d'estimation des méthodes de maximum de vraisemblance ont été codés par YANG dans le module CODEML du logiciel PAML [12] [217]. Ce sont ces programmes qui ont été utilisés pour les études présentées dans les deux chapitres suivants. Ces mêmes modèles ont été récemment recodés, de façon plus optimisée, par POND, MUSE et FROST dans leur logiciel HyPhy [13]. Outre sa convivialité et son ergonomie incomparable à celle de CODEML, Hyphy effectue les estimations de maximum de vraisemblance beaucoup plus rapidement que CODEML. De plus, Hyphy code également les méthodes *ad hoc*. Dans cette thèse, la méthode de maximum de vraisemblance et le programme CODEML ont été choisis pour des raisons historiques d'une part, mais aussi pour ses bases statistiques solides. Au moment où ces travaux ont été engagés, la méthode de maximum de vraisemblance semblait de loin la meilleure. Avec du recul, comme nous l'avons

12. Disponible gratuitement sur `http://abacus.gene.ucl.ac.uk/software/paml.html`
13. Disponible gratuitement sur `http://www.hyphy.org`

vu dans ce paragraphe, chacune des deux méthodes semble avoir ses points forts et ses points faibles. Un projet futur est de refaire les analyses présentées dans les deux chapitres suivants avec les méthodes *ad hoc* codées dans Hyphy, afin de tester la fiabilité des résultats.

Le développement des méthodes de détection de sélection positive est actuellement très rapide et, d'ici quelques années, des méthodes bien plus performantes seront proposées. L'utilisation de ces méthodes requiert une bonne compréhension de leur fonctionnement. C'était l'objet de ce chapitre que d'exposer les principes de bases des méthodes de détection de sélection positive, en retraçant leur développement historique. Les principes expliqués dans ce chapitre seront également d'une grande utilité pour la compréhension des méthodes qui vont être développées dans le futur.

Chapitre 3

Adaptation moléculaire des lentivirus de primates[1]

L'INTÉRÊT porté depuis quelques années à l'étude des lentivirus de primates est étroitement lié au fait que certains d'entre eux, les HIV, sont responsables de la **pandémie** actuelle de SIDA, aux effets dévastateurs sans précédent [156]. Depuis le premier diagnostique en 1981, plus de 25 millions de personnes sont mortes du SIDA [213] et aujourd'hui, 42 millions d'individus sont porteurs du HIV [161]. La seule année 2002 a connu 5 millions de contaminations, dont 70% en Afrique sub-saharienne [161]. Le virus est responsable d'une dégénérescence progressive du système immunitaire, et l'apparition de maladies opportunistes (tuberculose, *etc...*) se solde par la mort dans presque 100% des cas [129]. Des applications combinées de différents agents antiviraux (AZT[2], *etc...*) permettent de ralentir la progression de la maladie chez certains patients [165]. Toutefois, le coût élevé de telles thérapies réserve leur usage aux pays industrialisés [165]. De plus, du fait d'une **évolution extrêmement rapide** [116, 97], de nombreuses souches résistantes à plusieurs drogues sont apparues [138, 113, 153]. Ce même potentiel évolutif – parmi les plus forts connus dans le monde vivant – rend également difficile la mise au point d'un vaccin efficace [199, 143]. Comprendre l'évolution du virus du SIDA est aujourd'hui d'un intérêt pratique qui n'est plus à démontrer. En particulier, un domaine d'investigation important concerne son **évolution intra-hôte** et sa relation avec le système immunitaire. Comment le système immunitaire agit sur le virus ?, Comment en détermine-t-il l'évolution ?, Est-ce que ces mécanismes sont les mêmes pour toutes les souches

1. Une partie de ce chapitre a fait l'objet d'une publication (Choisy M., Woelk C.H., Guégan J.-F. & Robertson D.L. **2004** *Journal of Virology* **78**(4) : 1962-1970) présentée en annexe A.

2. Azidothymidine.

de HIV ?, *etc...* sont aujourd'hui des questions cruciales pour la mise au point d'un vaccin contre le SIDA. Par ailleurs, la relative simplicité du génome des lentivirus de primates, leur évolution rapide et l'importante quantité de données génétiques produites ces dernières années font de ces virus un modèle de choix pour l'étude de questions fondamentales sur l'évolution des relations hôte-parasite en général. Beaucoup de maladies émergentes sont des **zoonoses**[3]. L'acquisition d'une nouvelle espèce hôte peut être le résultat de changements écologiques de l'hôte (augmentation de densité, augmentation des mouvements, contacts étroits avec la faune sauvage, *etc...*), mais également d'une évolution du pathogène [134, 204, 15, 161]. Il semble aujourd'hui entendu que le HIV est le résultat du passage de lentivirus de singes (SIV[4]) à l'Homme. Une observation importante est que les SIV ne semblent pas causer d'importantes pathologies chez leurs hôtes [90, 79, 182]. La quantité des données génétiques accumulées sur les lentivirus de primates offre une opportunité sans précédent pour rechercher les évolutions génétiques du virus associées à son changement d'hôte. On peut ainsi espérer mettre en évidence des **facteurs de virulence**, voire même comprendre les mécanismes liés à la virulence.

Après une brève présentation de l'histoire naturelle des lentivirus de primates (paragraphe 3.1), nous décrivons le matériel et les méthodes générales utilisés pour l'étude de l'adaptation moléculaire des lentivirus de primates (paragraphe 3.2). Nous réalisons ensuite une étude comparative très générale de l'adaptation moléculaire sur différents gènes du génome de plusieurs lentivirus de primates (paragraphe 3.3). Enfin, nous focalisons notre attention sur le gène *env* des HIV (paragraphe 3.4).

3.1 Présentation des lentivirus de primates

Les **rétrovirus** constituent une famille de virus à ARN infectant une très grande variété d'espèces animales. Ce sont des virus enveloppés[5], caractérisés par la possession d'une enzyme, la **reverse transcriptase**, leur permettant de traduire leur matériel génétique en ADN et de l'intégrer dans le génome de leur cellule hôte. Cette famille est divisée en trois sous-familles : les **oncovirus**, pouvant causer des tumeurs et des leucémies, les **spumavirus**, sans pathologie connue, et les **lentivirus** dont le développement des pathologies est très lent, d'où leur nom. Quatre rétrovirus humains ont été

3. Passage à l'Homme d'un agent pathogène infectant habituellement des animaux.
4. Nommés SIV par analogie au HIV : *simian immunodeficiency virus.*
5. *i.e.* entourés d'une membrane lipidique.

décrits. Les HTLV[6] 1 et 2 de la sous-famille des oncovirus et les HIV-1 et 2 de la sous-famille des lentivirus.

Les lentivirus sont des virus de mammifères et infectent une grande variété d'espèces de primates, de périssodactyles[7], de carnivores et d'artiodactyles[8]. Chez les primates, outre les HIV-1 et HIV-2, on trouve de nombreux SIV infectant des espèces de singes africains différentes (*Cercopithecidae* et *Hominidae*).

3.1.1 Structure

Le matériel génétique des lentivirus est constitué de deux brins d'ARN monomères (**pseudodiploïdie**) d'environ 10 000 bases chacun, voir figure 3.1. Il est entouré d'une capside protéique contenant également un certain nombre d'enzymes indispensables et spécifiques au virus, comme la reverse transcriptase. Le virus est un virus enveloppé, ce qui signifie que la capside est entourée d'une membrane lipidique, ou enveloppe. Cette dernière est issue de la membrane de la cellule hôte, lors du processus de bourgeonnement (voir figure 3.3). Insérées dans l'enveloppe, se trouvent des **glycoprotéines membranaires**[9] dont le rôle est crucial dans l'infection d'une cellule hôte (voir figure 3.3). Ces glycoprotéines membranaires (gp160) sont formées de deux sous-unités : une sous-unité intra-membranaire (gp41) et une sous-unité extra-membranaire (gp120).

Le génome du virus est assez court et compact (figure 3.2). Outre les extrémités terminales répétées (LTR[10]), il contient **neuf gènes** dont trois principaux : *gag* code pour les protéines de la capside, *pol* code pour les enzymes, dont la réverse-transcriptase, et *env* code pour les deux sous-unités des glycoprotéines membranaires. Ces dernières contiennent, chez les HIV, cinq régions immunogènes hyper-variables nommées V1 à V5.

3.1.2 Cycle viral

Les glycoprotéines membranaires du virus peuvent s'arrimer aux récepteurs membranaires CD4 (voir figure 3.3). Le virus infecte donc préférentiellement les cellules arborant ces antigènes, c'est à dire essentiellement les macrophages, les cellules dendritiques et les lymphocytes T4 (voir encadré 3.1 pour les notions de bases d'immunologie). L'arrimage à un récepteur CD4

6. *Human T-cell lymphotrophic viruses.*
7. Comprend les chevaux, tapirs, rhinocéroces *etc...*
8. Comprend les porcs, bovidés, cervidés, *etc...*
9. Protéines arborant à leur surface des résidus de sucre.
10. *Long terminal repeats.*

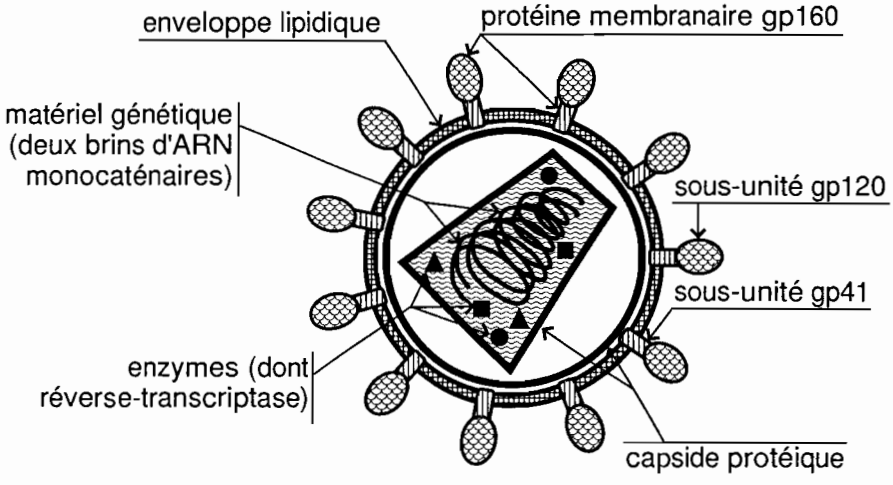


FIG. 3.1 – **Structure typique d'un lentivirus de primate. La glycoprotéine membranaire gp160 est constituée de deux sous-unités. Une sous-unité membranaire gp41 et une sous-unité extra-membranaire gp120. Ce sont des virus de petite taille, d'environ 100 nm de diamètre.**

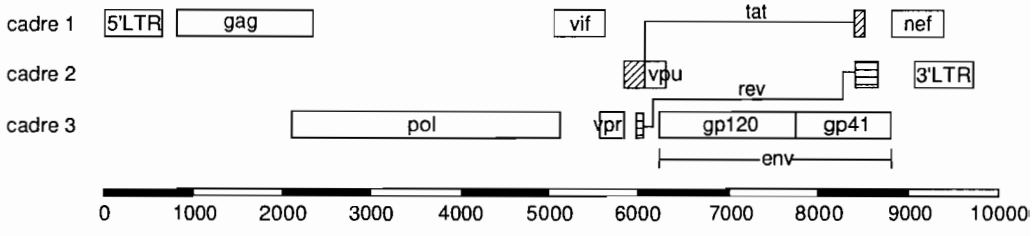


FIG. 3.2 – **Le génome des lentivirus de primates. La figure illustre la carte du génome du virus de référence HIV-1 HXB2. La structure du génome est globalement la même pour tous les lentivirus de primates, à quelques détails près. Le génome contient neuf gènes, plus des extrémités terminales répétées (LTR, *long terminal repeat*). Les trois cadres de lecture sont utilisés et chaque ligne correspond à un cadre. Les trois gènes majeurs sont *gag* codant pour les protéines de la capside, *pol* codant pour les enzymes, et *env* codant pour les glycoprotéines membranaires. L'échelle au bas de la figure indique le nombre de bases nucléotidiques, de l'extrémité 5' à l'extrémité 3'.**

ENCADRÉ 3.1
NOTIONS D'IMMUNOLOGIE

Dans cet encadré, nous présentons les principales notions d'immunologie. Le système immunitaire des vertébrés est constitué d'une multitude de molécules et cellules en interaction. On distingue généralement la **réponse non-spécifique** de la **réponse spécifique**. La première fait intervenir essentiellement les **protéines du complément** (qui ont pour effet de trouer la surface cellulaire des pathogènes, les faisant ainsi éclater) et les **macrophages**, de grosses cellules qui phagocytent les agents pathogènes.

Plus élaborée et efficace, est la réponse immunitaire spécifique. Celle-ci est basée sur la reconnaissance de molécules parasitaires spécifiques, les **antigènes**. Les portions de ces molécules directement impliquées dans la reconnaissance sont appelées **épitopes**. Cette réponse implique essentiellement deux acteurs. Les premiers sont les **lymphocytes B**, cellules caractérisées par leurs **immunoglobulines** membranaires. Sous forme excrétée et libre, on les appelle des **anticorps**. Les anticorps se fixent sur les épitopes, ce qui a pour effet de concentrer et de neutraliser les cellules parasitaires avant leur phagocytose par les macrophages. L'autre lignée d'acteurs est celle des lymphocytes T. Ces cellules reconnaissent les cellules de l'organisme parasitées par un agent intracellulaire. On distingue deux types de lymphocytes T, selon leur récepteurs membranaires : les **lymphocytes T4** dits aussi *helper*, arborant les récepteurs CD4 et les **lymphocytes T8** dits aussi *killer*, arborant les récepteurs CD8. Ces derniers ont pour effet de détruire directement la cellule infectée, limitant ainsi la propagation du parasite vers les cellules saines de l'organisme. Les lymphocytes T4 ont un rôle central dans la réponse immunitaire puisqu'ils coordonnent l'ensemble de la réponse en stimulant les autres acteurs de la réponse. Les **chimiokines** sont des protéines servant à la communication entre les différentes cellules de la réponse immunitaire.

est nécessaire mais pas suffisant et le HIV utilise d'autres récepteurs membranaires de la cellule hôte tels que les récepteurs de chimiokines des classes CC (*e.g.* CCR5) et CXC (*e.g.* CXCR4) (voir encadré 3.2). Cet arrimage de la glycoprotéine membranaire du virus aux récepteurs et corécepteurs membranaires de la cellule hôte permet la fusion des membranes lipidiques de la cellule hôte et du virus, et l'introduction de la capside dans le cytoplasme de la cellule hôte (figure 3.3). La capside se désagrège et la réverse-transcriptase traduit le matériel génétique en ADN. Cette réverse-transcription de l'ARN en ADN se fait avec **recombinaison** entre les deux « chromosomes » du virus. Ce phénomène de recombinaison est extrêmement important dans la biologie du virus et permet d'amplifier la diversité génétique produite par la mutation. Dans le cas de coinfections, si une même cellule est infectée par

deux souches différentes, il peut y avoir une recombinaison inter-souches. Il semble que ce phénomène soit assez fréquent [98, 45, 167]. L'ADN migre ensuite dans le noyau et s'insère dans le génome de la cellule hôte. De là, il est exprimé par la cellule hôte. Des molécules d'ARN et des protéines virales sont ainsi produites et assemblées dans le cytoplasme (figure 3.3). Les glycoprotéines membranaires sont insérées dans la membrane de la cellule hôte qui produit de nouveaux virions[11] par simple bourgeonnement de sa membrane cytoplasmique (figure 3.3). Chaque jour, quelques 10^{10} virions sont produits de cette façon par un hôte infecté [154].

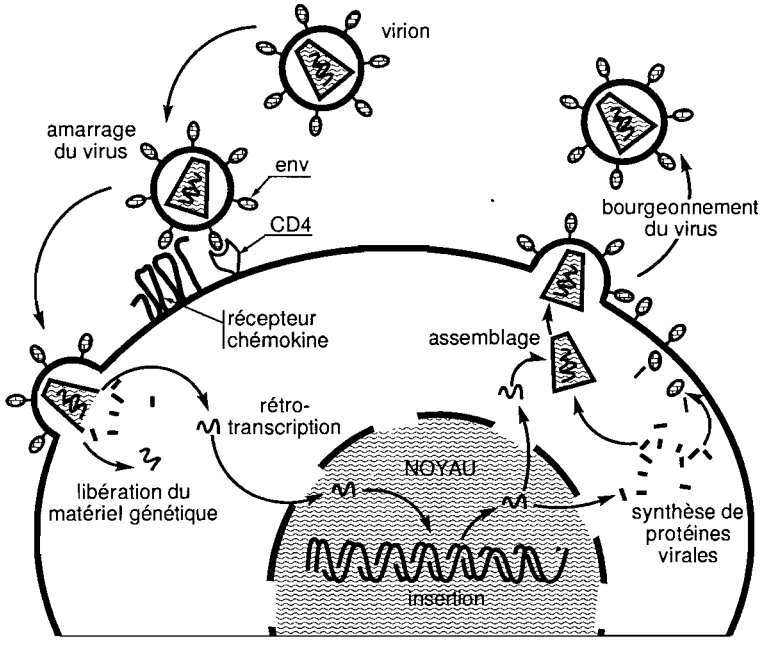


FIG. 3.3 – **Le cycle de vie du HIV. Les glycoprotéines membranaires du virion s'amarrent sur les CD4 et les récepteurs membranaires à chimiokines de la cellule hôte. Ceci permet la fusion des membranes cytoplasmiques et la pénétration de la capside à l'intérieur de la cellule hôte. La capside se désagrège, l'ARN viral est rétro-transcrit (avec recombinaison) en ADN qui migre dans le noyau où il s'intègre dans le génome de la cellule hôte. De là, le matériel génétique est exprimé par la cellule hôte, produisant des ARN et protéines viraux. Ces derniers sont assemblés dans le cytoplasme, les glycoprotéines membranaires du virus sont insérées dans la membrane cytoplasmique de la cellule hôte et de nouveaux virions sont produits par bourgeonnement.**

11. Particules virales libres.

ENCADRÉ 3.2
TROPHISME CELLULAIRE DES HIV

Certaines souches de HIV ont une préférence pour les lymphocytes T4 et ne parviennent pas à infecter les macrophages. Dans ce cas, on observe que ce sont les corécepteurs CXCR4 qui sont utilisés [174]. D'autres souches de HIV, au contraire, infectent les macrophages et pas les lymphocytes T4. Dans ce cas, ce sont les corécepteurs CCR5 qui sont utilisés [174]. La principale différence phénotypique entre ces deux types de HIV est que la souche à CCR5 domine pendant les premières années de l'infection et n'induit pas de **syncitium**[a] *in vitro*. Un changement d'utilisation des corécepteurs de CCR5 à CXCR4 apparaît plus tard et coïncide généralement avec la diminution de la quantité de lymphocytes T4 et le début des premiers symptômes du SIDA [174]. De plus, les HIV à CXCR4 induisent la formation de syncitium *in vitro*.

[a]Cellule à plusieurs noyaux.

3.1.3 Diversité, distribution et origine des HIV

On distingue deux lentivirus humains, le HIV-1 et le HIV-2, et plusieurs lentivirus de singes, nommés d'après l'espèce qu'ils infectent (*e.g.* SIV_{cpz} infectant les chimpanzés). L'accumulation, depuis une quinzaine d'années, de séquences de lentivirus de primates a permis une appréciation de leur extra-ordinaire diversité. Les analyses phylogénétiques des HIV ont révélé une organisation **agrégée** des différentes souches le long de l'arbre phylogénétique. Ainsi, suivant le niveau de divergence, on distingue chez les HIV-1 trois **groupes** : le groupe M (comme *main* en anglais, le principal), O (comme *outgroup* en anglais, dont la première souche a été découverte en 1991 et les suivantes après 1995), et le groupe N (comme *new* en anglais, récemment découvert). Le groupe M, le plus important, est ensuite subdivisé en neuf **sous-types** différents (nommés de A–D, F–H, J et K) et au moins quatorze **recombinants inter-sous-types** (nommés CRF01–CRF14 pour *circulating recombinant forms*). Des études récentes ont identifié une diversité équivalente au sein du group O, bien qu'il ne semble pas y avoir de structure en sous-type comme dans le groupe M. En ce qui concerne le groupe N, très peu de séquences ont été décrites. De façon identique, les HIV-2 ont été divisés en sept sous-types différents, nommés de A à G.

Ces différents virus, groupes et sous-types se caractérisent par des régions d'endémicité géographique (et semble-t-il aussi des virulences) différentes (figure 3.4). La très grande majorité des cas de SIDA sont dus au HIV-1. En Amérique du Nord et en Europe, ces cas sont essentiellement liés au HIV-1 groupe M sous-type B. Le sous-type C est très répandu en Inde et dans

le sud de l'Afrique, et le recombinant entre A et E en Asie du sud-est. La plus grande diversité est présente en **Afrique sub-saharienne** où toutes les souches de virus sont observées, conformément à l'hypothèse selon laquelle les HIV seraient issus de cette région du monde. Les groupes O et N, ainsi que le HIV-2, sont restreints au Cameroun et régions avoisinantes. Il semblerait que les HIV-2 soient moins virulents que les HIV-1 [100].

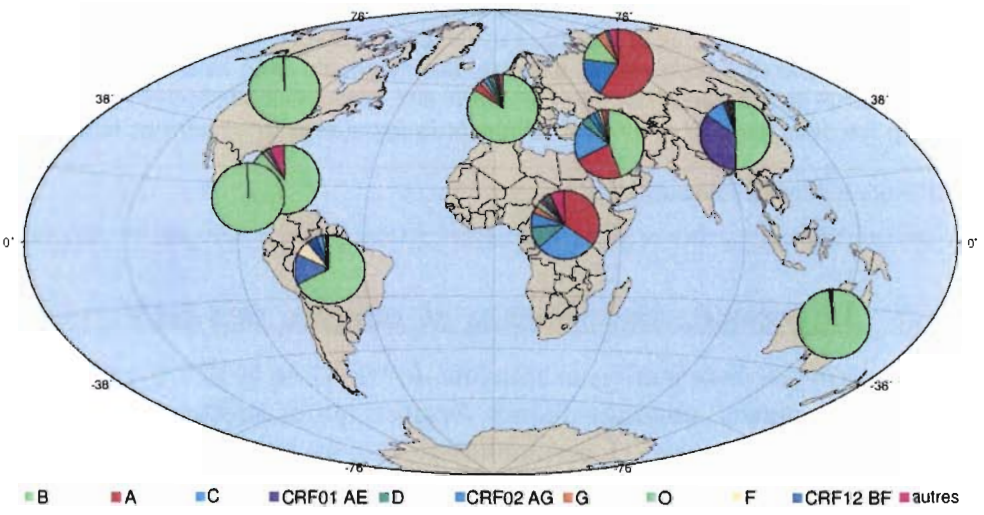


FIG. 3.4 – **Prévalences des sous-types et recombinants principaux de HIV-1 groupe M. Les camemberts représentent des pourcentages locaux et en aucun cas des nombres absolus. Rappelons en effet que plus de 70% des porteurs de HIV se trouvent en Afrique sub-saharienne. Extrait de *Los Alamos National Laboratory*,** `http://www.hiv.lanl.gov/content/index`.

Les analyses phylogénétiques suggèrent que les HIV sont issus des SIV (figure 3.5) [67, 79]. La première observation faite est que HIV-1 et HIV-2 sont moins proches l'un de l'autre que de virus infectant des espèces de singes africains. De plus, HIV-1 et HIV-2 ne sont pas des groupes **monophylétiques**, suggérant plusieurs passages inter-espèces indépendants [79]. Ainsi, HIV-1 semble issu du passage du SIV_{cpz} (infectant les chimpanzés) à l'homme et nous sommes obligés d'envisager au moins trois contaminations inter-espèces indépendantes afin de rendre compte des trois groupes M, N, et O. De même, chacun des sept sous-types de HIV-2 devrait correspondre à un passage du SIV_{smm} (infectant le grivet) à l'homme.

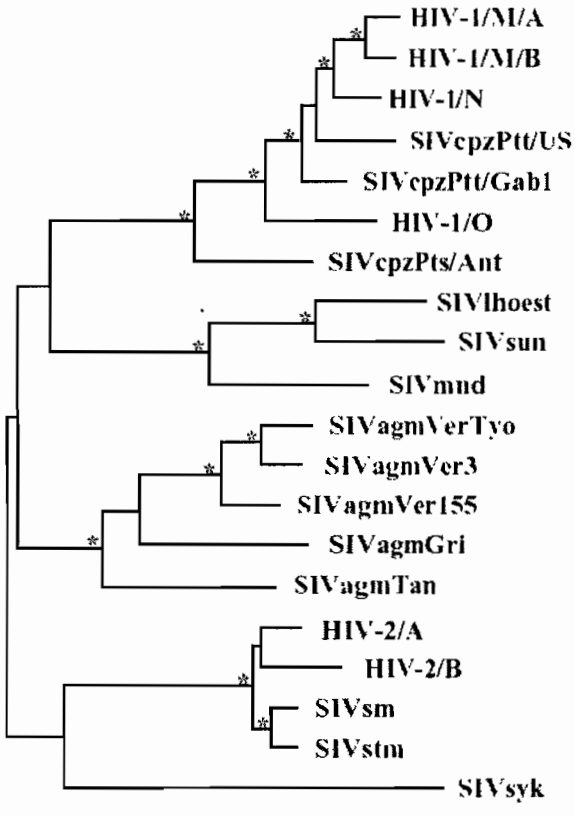


FIG. 3.5 – **Relations phylogénétiques entre lentivirus de primates. Les longueurs de branches sont proportionnelles au temps évolutif. L'arbre a été estimé par maximum de vraisemblance. Les astérisques indiquent une confiance d'au moins 80% par *bootstrap*. D'après la référence [23].**

3.2 Matériels et méthodes

3.2.1 Préparation des jeux de données

Nous avons utilisé la méthode du maximum de vraisemblance développée par YANG et collaborateurs (voir chapitre 2) pour analyser la sélection positive sur les trois gènes principaux du génome de plusieurs lentivirus de primates. Une analyse de sélection positive concerne donc un gène donné, pour une lignée de virus donnée (par exemple le gène *env* du HIV-1 groupe M sous-type B, voir tableau 3.1). Chaque analyse de sélection positive nécessite un **alignement de séquences nucléotidiques**. La puissance des méthodes de détection augmente avec le nombre de codons dans chaque séquence (idéalement au moins une centaine) et le nombre de séquences dans chaque alignement (idéalement au moins une dizaine). Ces contraintes ont donc restreints les analyses au trois gènes principaux, *gag*, *pol* et *env*.

TAB. 3.1 – **Les alignements de séquences nucléotidiques analysés.**

	gag		*pol*		*env*		*gp*120		*gp*41	
	cod[a]	séq[b]	cod[a]	séq[b]	cod[a]	séq[b]	cod[a]	séq[b]	cod[a]	séq[b]
HIV-1 M A	404	11	838	13	578	16	415	20	232	19
HIV-1 M B	425	35	913	33	578	30	433	20	233	30
HIV-1 M C	418	17	911	16	578	30	423	20	273	30
HIV-1 M D	—	—	—	—	578	15	436	15	236	15
HIV-1 M F	—	—	—	—	—	—	448	9	237	9
HIV-1 M G	—	—	—	—	—	—	449	9	237	9
HIV-1 O	—	—	—	—	621	30	454	10	232	10
HIV-2 A1	286	12	916	12	679	22	460	20	193	22
SIV_{smm}	269	7	911	6	699	10	496	10	167	13
SIV_{agm}	415	12	917	6	—	—	285	17	—	—
SIV_{cpz}	—	—	—	—	684	5	430	5	236	5

[a]Nombre de codons par séquence.
[b]Nombre de séquences dans l'alignement.

Les alignements de séquences nucléotidiques ont été téléchargés à partir du site de *Los Alamos National Laboratory*[12], à l'exception des séquences de HIV-1 groupe O téléchargées à partir du site de GenBank[13] et alignées avec CLUSTALW[14]. L'alignement a été vérifié manuellement au moyen de l'éditeur

12. `http://www.hiv.lanl.gov/content/index`
13. `http://www.ncbi.nih.gov`
14. Disponible gratuitement à `http://www.ebi.ac.uk/clustalw`

de séquence Se-Al [15]. Comme les méthodes de détection de sélection positive ne gèrent pas la recombinaison, les recombinants connus ont été exclus des analyses. Les sites contenant des codons stop ou des délétions dans au moins une séquence de l'alignement ont été exclus pour toutes les séquences de l'alignement. De plus, puisque les méthodes de détection de sélection positive de YANG et collaborateurs utilisent des modèles d'évolution des codons (voir chapitre 2), tous les sites appartenant à plusieurs cadres de lecture différents ont également été exclus pour toutes les séquences de l'alignement. Enfin, il existe énormément de séquences disponibles pour HIV-1 M B et il était informatiquement impossible d'effectuer les analyses sur toutes ces séquences à la fois. Nous avons dû en sélectionner une trentaine. Le choix des séquences a été fait de telle sorte qu'elles soient représentatives du groupe (*i.e.* ni trop proches ni trop divergentes les unes des autres). Afin de tester un possible effet de l'échantillonnage, une dizaine [16] d'alignements de trente séquences ont ainsi été constitués pour HIV-1 M B et analysés séparément. Malgré ce faible nombre de répétitions, les résultats se sont avérés constants. Le tableau 3.1 liste les alignements de séquences analysés et dont les résultats sont présentés dans ce chapitre.

3.2.2 Phylogénies des alignements de séquences

Les méthodes de détection de sélection positive de YANG et collaborateurs tiennent compte de la phylogénie des séquences de l'alignement analysé. Cette phylogénie doit être estimée au préalable. Nous avons utilisé le logiciel PAUP* 4.0b6 [196]. Les arbres phylogénétiques ont été estimés par **maximum de vraisemblance** en utilisant le modèle d'évolution de nucléotides de HASEGAWA, KISHINO et YANO [87] (voir chapitre 2, encadré 2.2, page 38), avec une distribution gamma discrétisée (en huit classes) pour tenir compte des variations des taux de substitution le long de la séquence. Les valeurs du rapport des taux de transition/transversion $\kappa = T_S/T_V$ ainsi que du paramètre α de forme de la distribution gamma ont été estimés au cours de la construction de l'arbre.

3.2.3 Détection de sélection positive

Les modèles de YANG et collaborateurs ont été estimés par le module CODEML du programme PAML 3.1 [17]. Quatorze modèles ont été proposés par YANG et collaborateurs [225]. Étant donnés les problèmes liés à la

15. Disponible gratuitement à `http://evolve.zoo.ox.ac.uk/software.html`
16. Etant donné la longueur des analyses, nous n'avons pas pu réaliser plus de répétitions.
17. Disponible gratuitement à `http://abacus.gene.ucl.ac.uk/software/paml.html`

convergence sur un pic de maximum de vraisemblance, il est nécessaire de réaliser les analyses avec plusieurs modèles différents et avec plusieurs valeurs initiales différentes lorsque le nombre de paramètres devient important (*e.g.* modèles M7 et M8). Toutefois, estimer les valeurs des quatorze modèles pour tous les alignements de séquences serait extrêmement long, et YANG et collaborateurs ont montré que six d'entre eux (M0, M1, M2, M3, M7 et M8) suffisaient pour détecter efficacement la sélection positive [225]. Nous nous sommes donc restreint à ces six modèles. Parmi ces modèles, M0, M1 et M7 sont des modèles nuls tandis que M2, M3 et M8 permettent la sélection positive. Les résultats des six tests de rapport de vraisemblance (voir chapitre 2, page 44) étaient cohérents, ainsi que les probabilités *a posteriori* des modèles M2, M3 et M8. En considérant comme positivement sélectionnés les sites avec une probabilité *a posteriori* supérieure à 0.95, les modèles M2 et M8 identifiaient les mêmes sites sous sélection positive. M3 identifiait les mêmes sites que M2 et M8, plus d'autres, M3 étant connu pour être un peu libéral dans la détection de la sélection positive [225, 12]. Par soucis de simplicité, les résultats présentés dans le texte principal ne concernent que les modèles M7 et M8, les résultats des autres modèles étant présentés en annexe B.

3.3 Analyse de l'adaptation chez les lentivirus de primates

Les analyses phylogénétiques semblent indiquer que les HIV seraient issus du passage de SIV à l'homme (voir paragraphe 3.1.3). Même si les datations moléculaires sont entachées de beaucoup d'incertitudes, il semblerait toutefois que les SIV aient coévolué avec leur hôte depuis beaucoup plus longtemps que les HIV, ce qui serait cohérent avec l'hypothèse du transfert des HIV de singes à l'homme.

Une observation importante est que la virulence du virus d'immunodéficience semble d'autant plus forte que l'histoire de coévolution avec son hôte est courte. En effet, la maladie causée par le virus d'immunodéficience n'est mortelle que chez l'Homme et le macaque [18], alors qu'elle ne présente pas de symptômes pathologiques chez les singes africains. Une conclusion de cette observation pourrait être que les virus d'immunodéficience, en s'adaptant à leur hôte, verraient leur virulence diminuer et que le SIDA correspondrait à une période transitoire de **maladaptation** [54]. Aussi séduisante que cette

18. Le macaque est le seul singe non-africain connu a être infecté par un lentivirus. Aucun macaque infecté n'a été observé dans la nature et il est fort possible que son infection se soit produite artificiellement par des contacts en captivité avec des grivets infectés.

hypothèse puisse être, la réalité n'est certainement pas aussi simple, et une série d'autres hypothèses ont été formulées dans la littérature pour expliquer la forte virulence des HIV et la faible virulence des SIV [31, 114].

Une façon de mieux comprendre la virulence des HIV est de comparer leurs traits d'histoire de vie et leurs évolutions avec leurs relatifs SIV non virulents. Les HIV et les SIV ne se différencient pas en termes de taux de mutation, de taux de réplication, ni de charge virale [163]. Une différence majeure entre les SIV et les HIV est le fait que les premiers induisent une réponse immunitaire beaucoup plus faible que les derniers et qu'ils sembleraient avoir une évolution moléculaire beaucoup plus lente [185, 163]. Ceci est particulièrement marqué pour la région immunogène V3 du gène *env* qui est conservée chez les SIV et fortement variable chez les HIV [79]. Ceci semblerait donc suggérer que, ayant coévolué depuis très longtemps avec leurs hôtes, les SIV auraient atteint une valeur sélective optimale et seraient donc essentiellement sujet à une forme de sélection conservatrice. Au contraire, les HIV, en interaction récente avec leur hôte, seraient encore dans une phase d'adaptation caractérisée par une forme de sélection positive diversifiante. Nous avons essayé de vérifier cette hypothèse en analysant et comparant l'intensité et la localité de la sélection positive chez différents SIV et HIV.

3.3.1 Résultats

La comparaison des rapports d_N/d_S moyens le long de la séquence révèle de **grandes différences**, à la fois **entre gènes**, mais aussi **entre virus** (voir figure 3.6). Pour tous les virus analysés, les gènes *pol* et *gag* sont fortement sous sélection négative, comme attendu pour des gènes soumis à de fortes contraintes fonctionnelles. Seul le gène *env* présente de forts niveaux de sélection positive. Au sein de ce gène, la sous-unité extra-membranaire *gp*120 présente des rapports d_N/d_S légèrement plus élevés que la sous-unité intra-membranaire *gp*41. Pour un même gène, la valeur du rapport d_N/d_S est toujours plus élevée chez les HIV que chez les SIV. Le seul alignement de SIV à présenter un rapport d_N/d_S supérieur à 0.3 est celui de la sous-unité *gp*120 du $\mathrm{SIV_{smm}}$ (figure 3.6).

Pour tous les alignements de séquences analysés, il apparaît clairement que la sous-unité extra-membranaire *gp*120 est la région du génome soumise à la plus forte sélection diversifiante. En focalisant sur cette portion de gène, nous avons pu comparer des alignements de séquences d'une plus grande diversité de virus (figure 3.7). Il ressort que la valeur du rapport d_N/d_S augmente des SIV aux HIV et que les HIV-2 occupent une position intermédiaire entre les SIV et les HIV-1.

Nous avons également analysé les régions variables immunogènes V1 à

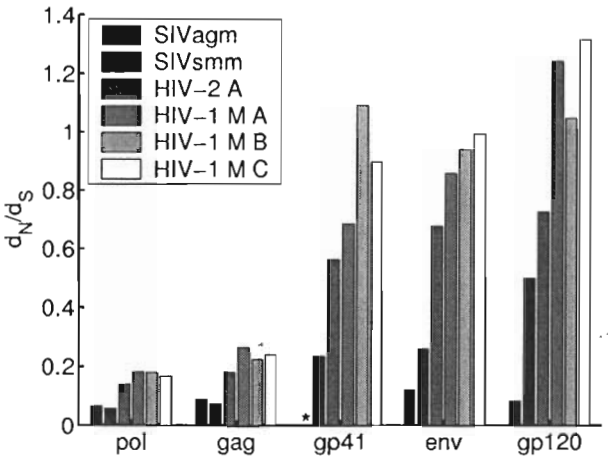


FIG. 3.6 – **Comparaison des rapports d_N/d_S moyens pour 5 gènes ou portions de gènes et 6 lignées de lentivirus de primates. Le gène *env* est analysé dans son ensemble ainsi que chaque sous-unité séparément (*gp*41 et *gp*120). Aucun alignement de séquence n'a pu être constitué pour la sous-unité *gp*41 du SIV_{agm} (marqué par une étoile).**

V5 de la sous-unité *gp*120, séparément et toutes ensembles. Les résultats sont tout à fait concordants avec ceux présentés ci-dessus pour les autres portions du génome, à savoir que les HIV présentent un rapport d_N/d_S plus important que les SIV.

3.3.2 Discussion

Les gènes *gag* et *pol* sont fortement conservés. Seul le gène codant pour la glycoprotéine de l'enveloppe contient des sites sous évolution diversifiante, reflétant probablement la pression de sélection exercée par la réponse humorale du système immunitaire (*i.e.* essentiellement les anticorps). Le niveau de sélection positive apparaît bien plus élevé chez les HIV que chez les SIV. Ceci semble en accord avec les observations et hypothèses précédemment faites sur la coévolution entre hôte et parasite. Il serait intéressant à cet égard d'essayer de relier les valeurs du rapport d_N/d_S observées sur la figure 3.7 avec les temps de divergence par méthodes phylogénétiques [110]. Le travail de SHARP et collaborateurs [182] semble un premier pas dans cette direction.

En résumé, ces résultats confirmeraient l'idée selon laquelle les virus ayant une histoire ancienne avec leur hôte seraient moins virulents et présenteraient des niveaux de sélection positive plus faibles que leurs cousins HIV et SIV_{mac}, récemment introduits chez l'Homme et le macaque respectivement. Certains auteurs ont interprété cette différence de sélection positive liée au niveau

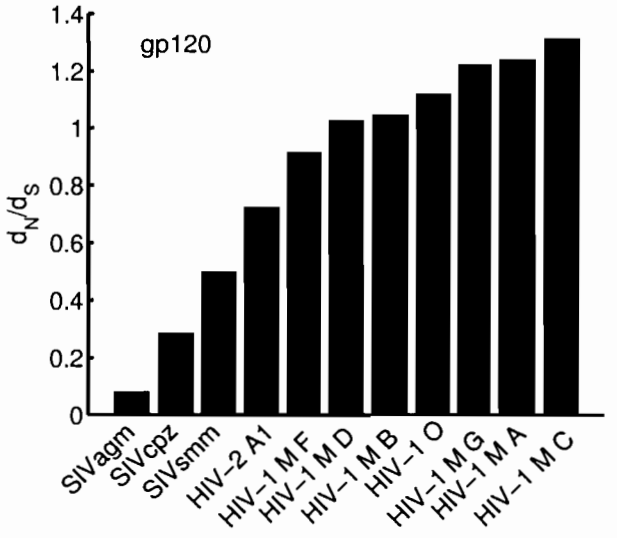


FIG. 3.7 – **Comparaison des rapports d_N/d_S moyens de la partie codant la sous-unité extra-membranaire *gp*120 de la glycoprotéine de l'enveloppe, et ce pour onze lignées de lentivirus de primates.**

de virulence en termes d'**utilisation des corécepteurs**. En effet, une différence importante entre les HIV et les SIV non-pathogènes est le fait que ces derniers utilisent uniquement le corécepteur CCR5 et jamais le CXCR4 [37], malgré le fait que la capacité à utiliser le récepteur CXCR4 ne requiert que quelques changements d'acides aminés dans la région V3. Cette différence dans l'usage des co-récepteurs pourrait expliquer les différences de **virulence** (voir encadré 3.2). Le fait que l'utilisation du corécepteur CXCR4 ne se fait pas chez les SIV, malgré un nombre de changements d'acides aminés requis très faible, suggère que la région V3 des SIV est soumise à des contraintes sélectives fortes empêchant l'utilisation des corécepteurs CXCR4 [90]. Le fait aussi que l'utilisation des CXCR4 par les HIV apparaisse tard dans l'infection (voir encadré 3.2), suggère également que les changements d'acides aminés requis ont tendance à être contre-sélectionnés, peut-être parce qu'ils induisent une réponse immunitaire forte. Ainsi, ce ne serait que lorsque le niveau de lymphocytes T4 est suffisamment faible que la substitution d'acides aminés deviendrait évolutivement avantageuse [36].

3.4 Localisation de l'adaptation moléculaire chez les HIV

Du paragraphe précédent il ressort que l'essentiel de la sélection positive chez les lentivirus de primates semble localisée au gène *env* des HIV. Dans ce paragraphe, nous nous concentrons donc sur ce gène dans le but de localiser précisément les sites sous sélection positive. L'**évolution intra-hôte** des HIV apparaît à la fois extrêmement rapide et essentiellement due à la pression de sélection exercée par le **système immunitaire** [128, 159, 132, 226]. Localiser précisément les sites sous sélection positive peut permettre d'identifier les forces sélectives en action pour ainsi mieux cibler les interventions thérapeutiques ou vaccinales (voir chapitre 1, page 22). De plus, une comparaison de l'intensité et de la localisation de la sélection positive entre plusieurs groupes et sous-types de HIV permet de mesurer l'**immunité croisée** entre souches. Ce type d'information est capital pour la mise au point d'un vaccin efficace, surtout dans certaines régions du monde comme l'Afrique sub-saharienne où plusieurs sous-types différents circulent [69].

Les données disponibles nous ont permis d'étudier le gène *env* pour les HIV-1, groupe M, sous-types A, B, C, D, groupe O et les HIV-2, sous-type A. Nous localisons d'abord les sites sous sélection positive en utilisant le bayesien empirique présenté au chapitre 2, page 45. Nous explorons ensuite dans quelle mesure les sites sous sélection positive apparaissent aux mêmes endroits dans les six lignées phylogénétiques étudiées. Enfin, nous essayons de mettre ces sites sous sélection positive en relation avec la pression exercée par le système immunitaire. En particulier, nous étudions la correspondance entre sites sous sélection positive et **zones épitopiques** ainsi que **sites de glycosylation** N et O [112, 212]. Les sites de glycosylation sont les acides aminés qui fixent les molécules de sucre à la surface de la glycoprotéine.

3.4.1 Matériels et méthodes

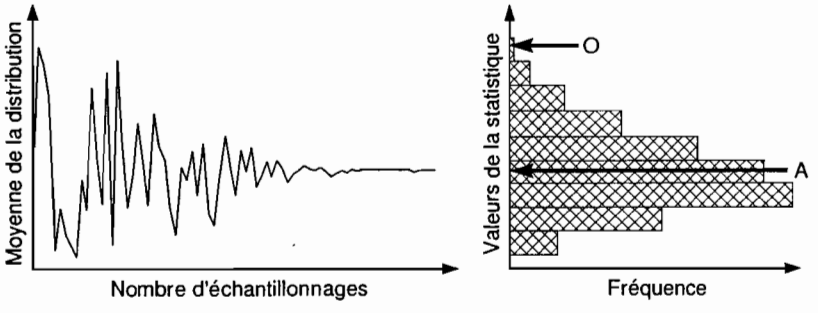
Les sites ont été considérés sous sélection positive lorsque leur probabilité *a posteriori* d'appartenir à la classe de $\omega > 1$ (*i.e.* la onzième classe, ω_{11}, pour le modèle M8 [19]) était supérieure à 0.95. La correspondance entre les positions des sites sous sélection positive de deux lignées phylogénétiques différentes a été testée par simulations de Monte Carlo (voir encadré 3.3), la statistique considérée étant le nombre de fois où deux sites apparaissent à la même

19. Rappelons que le modèle M8 contient onze classes de valeurs de ω : $\omega_1 < \omega_2 < \ldots < \omega_{10} < 1 < \omega_{11}$. Seule la dernière, supérieure à 1, rend compte d'une éventuelle sélection positive.

ENCADRÉ 3.3
LES TESTS DE MONTE CARLO

Un test statistique cherche à déterminer – en termes de probabilités – si la valeur d'une **statistique** est due au hasard ou non. Ceci est chose relativement aisée dès lors que l'on connaît la **distribution de probabilités** de la statistique en question. Dans le cadre des tests paramétriques classiques, cette dernière est déduite de considérations théoriques. Par exemple, on peut montrer que la différence entre les moyennes de deux distributions normales suit une distribution de Student. Cette distribution est donc utilisée dans le cadre des tests t pour tester l'égalité de deux moyennes. Toutefois, dans certaines situations, la distribution de la statistique ne peut être prédite théoriquement.

Lorsqu'on ne connaît pas la distribution de la statistique qui nous intéresse, une idée consiste à la simuler, c'est ce que l'on fait avec les méthodes de Monte Carlo. Une de ces méthodes est le ***bootstrap* non paramétrique** qui consiste tout simplement à rééchantillonner les données avec remise, un certain nombre de fois. Dans chaque échantillon, certaines données du jeu original seront absentes, tandis que d'autres seront dupliquées. La statistique est calculée sur chaque échantillon. L'histogramme des valeurs de la statistique sur chaque échantillon est supposé représenter une approximation de sa distribution de probabilité. La confrontation de la valeur de la statistique observée dans le jeu de données original avec sa distribution simulée constitue le test statistique à proprement parlé.



Une question importante avec ces méthodes est de savoir combien de fois on doit rééchantillonner le jeu de données pour construire la distribution de la statistique. Idéalement, le plus grand nombre de fois est le mieux mais nous sommes vite limité par le temps et la capacité des ordinateurs. Le tout est donc de déterminer le nombre minimum de répétitions au bout duquel il est raisonnable de s'arrêter. La distribution de la statistique est très dépendante du nombre d'échantillonnages. Lorsque ce dernier augmente, la distribution

(Encadré 3.3 suite)

tend vers un **état stationnaire**, alors indépendant du nombre d'échantillonnages. C'est cet état qu'il faut atteindre. Le comportement de la distribution de la statistique en fonction du nombre de rééchantillonnages peut être suivi par le comportement d'un indice comme la moyenne (voir figure de gauche) ou la médiane. Dès que les variations de cet indice sont très petites (inférieures à un seuil ε fixé), on peut considérer que le régime stationnaire est atteint. La figure de droite montre l'histogramme de la distribution simulée de la statistique. A est la valeur attendue moyenne (d'après la distribution simulée) tandis que O est la valeur effectivement observée de la statistique.

Le programme MCSites 0.9, écrit en Delphi dans le cadre de cette thèse, permet de tester si certains sites d'intérêt (*e.g.* sites sous sélection positive, sites de glycosylation, *etc...*) apparaissent aux mêmes endroits, dans deux séquences **homologues**. La statistique considérée est le nombre de fois où deux sites d'intérêt apparaissent à la même position dans les deux séquences. Le programme MCSites 0.9 permet un suivi efficace de la convergence vers un régime stationnaire de la distribution simulée.

position dans deux alignements de séquences différents [120]. Puisque ce type de test dépend de la qualité de l'alignement entre les deux jeux de données, on peut considérer que les résultats sont conservateurs.

Plus quantitativement, nous nous sommes également intéressés à ce que nous appelons la **force de la sélection**, calculée simplement par *model averaging* [208], comme précédemment fait par Gaschen et collaborateurs [69]. La méthode consiste à définir pour chaque site i un $\overline{\omega}_i$ défini, pour le modèle M8, comme

$$\overline{\omega}_i = \sum_{k=1}^{11} f_k \omega_k$$

où f_k est la probabilité *a posteriori* d'appartenir à la classe de valeur ω_k. L'égalité de la force de sélection entre deux alignements de séquences a été testée par un test de Wilcoxon apparié avec une correction de continuité appliquée à l'approximation normale des risques d'erreur de première espèce [188]. L'analyse a été réalisée sur deux types de données différents :

1. Les sites ayant un $\overline{\omega}_i$ supérieur à 1 ;
2. Les sites ayant une probabilité *a posteriori* d'appartenir à la onzième classe de ω supérieure à 0.95.

Notons ici que ces deux types de données ne sont pas exactement équivalents, le deuxième étant un sous-ensemble du premier.

Enfin, la correspondance entre les sites sous sélection positive et (i) les zones épitopiques d'une part et (ii) les sites de glycosylation d'autre part a

été testée par simulations de Monte Carlo de la même façon qu'expliqué précédemment. Les zones épitopiques considérées sont celles mises en évidence expérimentalement sur le HIV-1 groupe M sous-type B et correspondent à des régions reconnues par les anticorps, les lymphocytes T4 et les lymphocytes T8 [110]. Naturellement, comme les épitopes n'ont été définis que pour le sous-type B, seule la correspondance avec les sites sous sélection positive de ce sous-type a été testée. Pour chaque jeu de données, les sites de glycosylation N et O ont été prédits par programmes NetNGlyc 1.0 [20] (Gupta, Jung et Brunak, en préparation) et NetOGlyc 3.1 [21] [82] respectivement.

Pour toutes les simulations de Monte Carlo réalisées, nous avons utilisé 9 999 répétitions, ce qui s'est avéré largement suffisant pour atteindre un régime stationnaire. Le programme MCSites 0.9 utilisé pour ces tests de correspondance de sites entre deux séquences alignées a été écrit pour l'occasion (voir encadré 3.3). Comme c'est actuellement une version très rudimentaire et pas parfaitement stable, il n'est pas encore disponible en ligne. Toutefois, il m'a déjà été demandé par d'autres laboratoires effectuant ce même type d'analyses.

3.4.2 Résultats

Le nombre de sites sous sélection positive identifiés sur le gène *env* est de 22 pour HIV-2 sous-type A, entre 30 et 35 pour les sous-types de HIV-1 groupe M et de 40 pour HIV-1 groupe O (tableau 3.2). Les sites sous sélection positive ne se cantonnent pas aux régions variables V1 à V5. De plus, il apparaît que les sites d'amarrage sur le récepteur CD4 et les sites impliqués dans le changement d'utilisation du corécepteur CCR5 vers le corécepteur CXCR4 ne sont pas des sites sous sélection positive.

Les sites sous sélection positive ont tendance à être les mêmes pour différentes lignées de HIV

L'hypothèse nulle H_0 selon laquelle il n'y a aucune correspondance entre les positions des sites sous sélection positive dans deux lignées différentes a été rejetée dans la majorité des comparaisons effectuées, l'exception étant HIV-2 sous-type A (tableau 3.3). Ces résultats suggèrent donc que les glycoprotéines membranaires de différentes lignées de HIV subissent des pressions de sélection similaires.

20. Gratuit en ligne à `http://www.cbs.dtu.dk/services/NetNGlyc`
21. Gratuit en ligne à `http://www.cbs.dtu.dk/services/NetOGlyc`

TAB. 3.2 – **Sites sous sélection positive sur le gène *env* de HIV.**

Lignée	ω moyen[a]	ω_{11}^{b}	Nombre des sites[c]	Probabilité[d]
HIV-1 M A	0.690	4.702	33	**<0.001**
HIV-1 M B	0.623	4.009	35	**<0.001**
HIV-1 M C	0.610	4.463	33	**<0.001**
HIV-1 M D	0.568	3.821	30	**<0.001**
HIV-1 O	0.590	3.992	40	**<0.001**
HIV-2 A	0.444	3.568	25	**<0.001**

[a]Moyenne sur tous les sites de la séquence.
[b]Valeur du ω dans la onzième classe du modèle M8, *i.e.* la classe pour laquelle ω est supérieur à 1.
[c]Sites considérés sous sélection positive si la probabilité *a posteriori* d'appartenir à la onzième classe de ω est supérieure à 0.95.
[d]Probabilité résultant du test de rapport de vraisemblance entre M7 et M8. C'est donc la probabilité que le modèle M7 soit meilleur que le modèle M8 ou, dit autrement, le niveau de significativité de la sélection positive. Les probabilités significatives (P<0.05) sont en gras.

TAB. 3.3 – **Tests de Monte Carlo sur l'association des sites sous sélection positive dans deux lignées différentes de HIV.**

Lignées		HIV-1 M A	HIV-1 M B	HIV-1 M C	HIV-1 M D	HIV-1O
	A[a]	2.018				
HIV-1 M B	O[b]	13				
	P[c]	**0.001**				
	A[a]	1.990	2.015			
HIV-1 M C	O[b]	16	15			
	P[c]	**0.001**	**0.001**			
	A[a]	1.760	1.793	1.741		
HIV-1 M D	O[b]	10	14	15		
	P[c]	**0.001**	**0.001**	**0.001**		
	A[a]	1.016	0.996	0.889	0.848	
HIV-1 O	O[b]	7	7	8	6	
	P[c]	**0.001**	**0.001**	**0.001**	**0.001**	
	A[a]	0.633	0.722	0.571	0.511	0.729
HIV-2 A	O[b]	3	1	2	2	1
	P[c]	**0.024**	0.535	0.104	0.091	0.539

[a]Valeur moyenne attendue d'après la distribution simulée de la statistique.
[b]Valeur observée.
[c]Niveau de significativité avec lequel la valeur observée O est différente de la valeur attendue A. Les probabilités significatives (P<0.05) sont en gras.

Les intensités de sélection positive ont tendance à être les mêmes pour différentes lignées de HIV

Les seules différences significatives d'intensité de sélection positive le sont entre (*i*) les sous-types A et B du HIV-1 groupe M, (*ii*) les sous-types B et C du HIV-1 groupe M et (*iii*) HIV-1 groupe O et HIV-2 groupe A (tableau 3.4).

TAB. 3.4 – **Tests de Wilcoxon appariés sur les différences d'intensité de sélection entre deux lignées différentes de HIV. Les tests ne sont réalisés que sur les sites pour lesquels ω est supérieur à 1 dans les deux lignées.**

Lignées		HIV-1 M A	HIV-1 M B	HIV-1 M C	HIV-1 M D	HIV-1O
	Z[a]	3.827				
HIV-1 M B	N[b]	69				
	P[c]	**<0.001**				
	Z[a]	1.215	-2.195			
HIV-1 M C	N[b]	67	62			
	P[c]	0.224	**0.028**			
	Z[a]	0.601	-1.102	-0.408		
HIV-1 M D	N[b]	46	54	51		
	P[c]	0.548	0.270	0.684		
	Z[a]	0.365	-1.853	-1.093	0.000	
HIV-1 O	N[b]	23	22	26	18	
	P[c]	0.715	0.064	0.274	1.000	
	Z[a]	-0.400	-1.019	1.835	0.829	2.282
HIV-2 A	N[b]	11	10	10	9	7
	P[c]	0.689	0.308	0.067	0.407	**0.023**

[a]Valeur de la statistique de Wilcoxon.
[b]Nombre de sites avec ω supérieurs à 1.
[c]Niveau de significativité des différences. Les probabilités significatives (P<0.05) sont en gras.

La sélection positive ne semble pas liée aux zones épitopiques

Notre hypothèse nulle H_0 était que les sites sous sélection positive ne sont pas liés aux zones épitopiques. Contre cette hypothèse nulle, nous avons utilisé les simulations de Monte Carlo pour tester deux hypothèses alternatives :

- H_1 : les sites sous sélection positive ont tendance à se trouver sur les zones épitopiques ;
- H_2 : les sites sous sélection positive ont tendance à se trouver hors des zones épitopiques.

Comme zones épitopiques, nous avons considéré les épitopes à anticorps, à lymphocytes T4 et à lymphocytes T8 séparément, ainsi que toutes les combinaisons de deux. Les résultats suggèrent qu'il n'y a pas de relation particulière entre les sites sous sélection positive et les zones épitopiques (tableau 3.5).

TAB. 3.5 – **Tests de Monte Carlo sur l'association entre les sites sous sélection positive et les zones épitopiques pour HIV-1 groupe M sous-type B.**

Epitope(s)	N_{IN}[a]	A_{IN}[b]	O_{IN}[c]	P_{IN}[d]	N_{OUT}[e]	A_{OUT}[f]	O_{OUT}[g]	P_{OUT}[h]
Ac	370	22.17	18	0.946	208	12.53	17	0.072
T4	499	30.16	26	0.989	79	4.78	9	**0.028**
T8	394	23.82	19	0.976	184	11.12	16	0.053
Ac+T4	537	32.40	33	0.496	41	1.90	2	0.612
Ac+T8	507	30.64	30	0.726	71	4.17	5	0.401
T4+T8	524	31.67	27	0.999	54	2.92	8	**0.018**

[a]Nombre de sites appartenant aux épitopes.
[b]Nombre moyen attendu de sites sous sélection positive appartenant à une région épitopique. Déduit de la distribution simulée de la statistique.
[c]Nombre observé de sites sous sélection positive appartenant à une région épitopique.
[d]Niveau de significativité avec lequel les nombres observés O_{IN} sont différents des nombres moyens attendus A_{IN}.
[e]Nombre de sites n'appartenant pas aux épitopes.
[f]Nombre moyen attendu de sites sous sélection positive n'appartenant pas à une région épitopique. Déduit de la distribution simulée de la statistique.
[g]Nombre observé de sites sous sélection positive n'appartenant pas à une région épitopique.
[h]Niveau de significativité avec lequel les nombres observés O_{OUT} sont différents des nombres moyens attendus A_{OUT}. Les probabilités significatives ($P<0.05$) sont en gras.

La sélection positive semble être liée aux sites de glycosylation N

Aucun site de glycosylation N n'a été identifié pour le HIV-2 A. Pour les autres lignées, entre 22 et 39 sites de glycosylation N ont été identifiés et leur association avec les sites sous sélection positive s'est avérée forte (tableau 3.6). Le nombre de sites de glycosylation O est beaucoup plus faible (entre 0 et 8) et aucune association n'a été détectée avec les sites sous sélection positive.

TAB. 3.6 – **Tests de Monte Carlo sur l'association entre les sites sous sélection positive et les sites de glycosylation N pour les 5 lignées de HIV-1.**

Lignée	Glyc. N[a]	Glyc. N conservés[b]	A[c]	O[d]	P[e]
HIV-1 M A	28	5	1.64	11	**0.001**
HIV-1 M B	27	2	1.62	5	**0.019**
HIV-1 M C	30	2	1.69	7	**0.002**
HIV-1 M D	22	5	1.17	4	**0.023**
HIV-1 O	39	6	2.46	13	**0.001**

[a]Nombre total de sites de glycosylation N détectés dans l'alignement.
[b]Nombre de sites de glycosylation N situés sur des sites conservés de l'alignement.
[c]Nombre moyen attendu de sites sous sélection positive coïncidant avec un site de glycosylation N. Déduit de la distribution simulée de la statistique.
[d]Nombre observé de sites sous sélection positive coïncidant avec un site de glycosylation N.
[e]Niveau de significativité avec lequel les nombres observés O et attendus A sont différents. Les probabilités significatives ($P<0.05$) sont en gras.

3.4.3 Discussion

Le fait que les sites sous sélection positive ne soient pas confinés aux zones variables V1 à V3 confirme des résultats obtenus sur HIV-1 groupe M sous-type B en utilisant des méthodes *ad hoc* de détection de sélection positive [216]. Les sites impliqués dans l'attachement sur les récepteurs membranaires CD4 apparaissent conservés, ce qui est attendu étant donné l'importance de leur rôle fonctionnel. En revanche, il est surprenant que les sites impliqués dans le changement d'utilisation du corécepteur, du CCR5 au CXCR4, ne soient pas sous sélection positive. Il est fort possible que ce résultat soit dû au fait que nous comparons des séquences issues de plusieurs individus différents, plutôt que plusieurs séquences d'un même individu. Il est donc plus difficile d'appréhender l'**évolution intra-hôte**.

Les comparaisons d'intensités et de localisations de sélection positive entre les différentes lignées de HIV-1 suggèrent une certaine forme d'**immunité croisée inter-sous-type**. Ce résultat nuance donc les conclusions de GASCHEN et collaborateurs [69], et un vaccin « cocktail » contre plusieurs sous-types de HIV-1 apparaît comme une possibilité à envisager. Ceci est d'une importance cruciale, notamment pour les régions du monde, comme l'Afrique sub-saharienne, où plusieurs sous-types co-circulent.

Aucune association forte n'a été mise en évidence entre les sites sous sélection positive et les zones épitopiques. Ce résultat peut sembler surprenant,

mais on peut lui apporter deux types d'explications. Le premier est que le nombre de sites épitopiques de la glycoprotéine de l'enveloppe est particulièrement important (voir tableau 3.5). Or, plus le nombre de sites épitopiques est important, moins les tests de Monte Carlo pour détecter une association avec les sites sous sélection positive sont puissants. En effet, dans le cas extrême où tous les sites de la séquence appartiennent à des épitopes, les tests de Monte Carlo deviennent alors d'aucune utilité puisque la statistique devient égale au nombre de sites sous sélection positive. La deuxième explication serait que les sites de zones épitopiques sont effectivement sous sélection conservatrice, et ce à cause de contraintes fortes liées à son rôle d'attachement sur le récepteur membranaire CD4.

Cette deuxième explication, couplée avec l'observation que les sites sous sélection positive sont très significativement associés aux sites de glycosylation N, est intéressante puisqu'elle permet de confirmer un modèle conceptuel formulé par KWONG et collaborateurs [112, 212, 121]. Ces derniers ont observés que les virus ayant beaucoup de sites de glycosylation sur leur glycoprotéine membranaire échappaient beaucoup plus efficacement à l'action des anticorps que les virus ayant peu de sites de glycosylation. Les sites de glycosylation attachent des molécules de sucre à la surface de la glycoprotéine. Ces molécules de sucres sont particulièrement grosses et peuvent représenter jusqu'à 50% du poids moléculaire de la glycoprotéine [32]. Kwong et collaborateurs ont imaginé que le simple **encombrement stérique** pourrait empêcher les anticorps d'accéder aux épitopes, permettant ainsi au virus d'échapper à la réponse immunitaire humorale. Un tel modèle est tout à fait cohérent avec nos observations selon lesquelles la sélection positive est située sur les sites de glycosylation, indépendamment des zones épitopiques. En effet, l'idée est que la glycoprotéine membranaire du HIV est soumise à deux pressions de sélection opposées. La première pression est **conservatrice** et liée à son rôle d'attachement sur le récepteur membranaire CD4 de la cellule hôte. Cet attachement s'opère par complémentarité stérique fine entre les deux molécules. Un changement d'acide aminé n'importe où sur la glycoprotéine membranaire du virus est susceptible de changer la forme de la molécule et de rendre l'attachement au récepteur CD4 impossible. La deuxième pression est au contraire **diversifiante**. En effet, les anticorps du système immunitaire sont en perpétuelle évolution, ce qui leur permet de s'adapter à la diversité antigénique des agents pathogènes. Pour échapper à la réponse immunitaire, les virus ont besoin également d'un potentiel adaptatif. On a ainsi, au court d'une seule infection, une **coévolution permanente** entre le système immunitaire et l'agent pathogène. Dans ce contexte, la restriction des sites sous sélection positive aux sites de glycosylation permet de gérer ces deux pressions de sélection opposées.

Pour résumer, la majorité des sites de la glycoprotéine membranaire sont sous conservation, lui permettant de maintenir sa fonction d'attachement au récepteur CD4. Des molécules de sucres accrochées à la surface de la glycoprotéine empêchent l'accès des anticorps aux épitopes, constituant une sorte de **bouclier protectif** (figure 3.8). Le fait que les quelques rares sites attachant ces molécules de sucre soient sous forte évolution positive fait de ce bouclier protecteur une structure mouvante, lui permettant de répondre efficacement à l'évolution des anticorps. En effet si ce bouclier était une structure figée, les anticorps finiraient pas « trouver la faille » et neutraliser le virus.

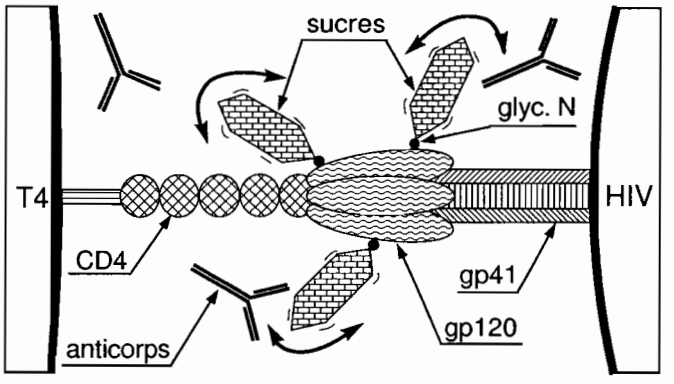


FIG. 3.8 – **Le modèle du bouclier de sucre de Kwong et collaborateurs [112, 212, 121]. La glycoprotéine membranaire du virus (gp120, gp41) est soumise à deux pressions de sélection opposées. La première, conservatrice, résulte de sa fonction d'attachement au récepteur membranaire CD4 de la cellule hôte. La deuxième, diversifiante, est imposée par l'action des anticorps. Dans ce contexte, la stratégie optimale pour le virus semble être la suivante. La quasi-totalité des sites de la glycoprotéine membranaire du virus sont conservés, y compris les épitopes, reconnus par les anticorps. Ceci permet de maintenir la fonction vitale d'attachement au récepteur CD4 de la cellule hôte. Des molécules de sucre à la surface de la glycoprotéine membranaire du virus gênent l'accès des anticorps aux épitopes, formant ainsi une véritable structure protectrice. Les quelques rares sites d'attachement des sucres à la glycoprotéine membranaire du virus sont sous forte sélection positive. Ceci permet au bouclier de sucre d'être une structure mouvante plutôt que figée, ce qui permet de s'adapter continuellement à l'évolution des anticorps.**

Le fait que nous n'ayons pas trouvé de sites de glycosylation N chez les HIV-2 est surprenant. Il est possible que ceci soit dû au fait que les HIV-2 ont beaucoup moins de sites de glycosylation N que les HIV-1 [117]. Les HIV-2

sont apparemment moins virulents que les HIV-1, certainement parce que moins antigéniques [79]. Cette plus faible virulence pourrait peut-être due à ce manque de sites de glycosylation N.

Chapitre 4

Adaptation moléculaire des cystéine protéinases de leishmanies[1]

Les Leishmanioses font partie des six maladies parasitaires considérées comme majeures par l'Organisation Mondiale de la Santé, et sont endémiques de la plupart des pays en voie de développement. Avec une prévalence (essentiellement dans les régions tropicales) de plus de 12 millions de personnes et une incidence annuelle de plus de 400 000 cas, quelques 350 millions de personnes sont actuellement considérées comme étant à risque [214]. Les leishmanioses se caractérisent par la **diversité de leurs formes cliniques**, pouvant aller de formes relativement bénignes, comme la leishmaniose cutanée, aux formes mortelles, comme la leishmaniose viscérale. Assez aisément soignées avec un traitement approprié[2], les leishmanioses constituent toutefois un problème de santé publique majeur, essentiellement dans les pays en voie de développement. Les leishmanioses sont aussi souvent observées comme maladies associées au SIDA. L'établissement d'une immunité permanente (mais spécifique à la souche) succédant à une primo-infection laisse beaucoup d'espoir pour le développement d'un **vaccin**. Si un vaccin contre la leishmaniose canine a été récemment mis au point[3] et est actuellement en cours de commercialisation, aucun vaccin n'existe aujourd'hui contre les leishmanioses humaines.

1. Ce travail a fait l'objet d'un manuscrit (Hide B., Choisy M. & Bañuls A.L.) soumis à *International Journal for Parasitology* et présenté en annexe C.

2. Lorsque le traitement est appliqué à temps, le taux de guérison est élevé. Toutefois, dans tous les cas, le traitement est extrêmement lourd et laisse des séquelles sur le patient.

3. Equipe de Jean-Loup Lemesre, laboratoire « Pathogénie des Trypanosomatidae », centre IRD de Montpellier.

Chez les leishmanies (ou *Leishmania*), agents responsables des leishmanioses, quelques molécules ont été identifiées comme **facteurs de virulence**, parmi lesquelles les cystéines protéinases (notées CP dans la suite). En outre, la **forte immunogénicité** des CP fait de ces molécules des cibles vaccinales de choix [160]. Les *Leishmania* possèdent une importante variété de CP, classées en trois types. Les CP les plus abondantes (Type I) sont particulièrement intéressantes d'un point de vue évolutif puisqu'elles sont codées par un gène **multi-copies** (pas moins de 19 copies chez *Leishmania mexicana*). Les gènes multi-copies (ou familles de gènes) sont produits par **duplication de gènes**. Même si l'évolution par duplication de gènes a été proposée dès 1970, ce n'est qu'à la fin des années quatre-vingt-dix, avec le développement de la génomique et du séquençage de gènes, que l'importance de ce mécanisme évolutif a été réellement appréciée, notamment dans les relations hôte-parasite [228]. En effet, de part leur diversité, les familles de gènes constituent un potentiel adaptatif exceptionnel. Ainsi, c'est la recombinaison entre gènes d'une même famille qui est à l'origine de l'importante diversité des immunoglobulines. De même, chez les parasites à grands génomes[4], les familles de gènes ont été identifiées comme constituant de véritables **bibliothèques d'archives**, à l'origine de la **diversité antigénique** [64].

Dans ce chapitre nous nous proposons d'étudier l'évolution moléculaire des CP de leishmanies. En particulier, nous nous proposons de comprendre l'histoire évolutive des trois types de CP de *Leishmania* ainsi que de quantifier, localiser et comparer l'adaptation moléculaire sur ces différentes CP, dans le but d'identifier des cibles vaccinales potentielles. Nous commençons par présenter les leishmanies, agents aetiologiques des leishmanioses, leur diversité et leur cycle de vie (paragraphe 4.1). Nous voyons ensuite les CP et l'intérêt de l'étude de leur évolution moléculaire en vaccinologie (paragraphe 4.2). Nous présentons les méthodes utilisées pour les analyses phylogénétiques et de détection de sélection positive (paragraphe 4.3), puis les principaux résultats (paragraphe 4.4) et enfin leur discussion (paragraphe 4.5).

4.1 Les leishmanies

Les *Leishmania* sont des parasites protozoaires[5] flagellés à deux hôtes : un insecte vecteur de la famille des **phlébotomes** et un hôte définitif **mammifère** pouvant être l'Homme. Il existe 24 espèces de leishmanies occupant surtout les régions tropicales. Chez l'Homme, les leishmanies peuvent provoquer une variété de maladies (les leishmanioses) dont la gravité peut aller de

4. Excluant essentiellement les virus chez qui le génome est extrêmement compact.
5. Organismes animaux unicellulaires.

petites lésions cutanées à guérison naturelle aux formes viscérales, mortelles dans 100% des cas non traités. Le déterminisme de la maladie est encore peu connu et dépendrait d'une multitude de facteurs incluant l'hôte mammifère (génétique, histoire, état de santé général, *etc...*), le vecteur (espèce), le parasite (espèce), l'environnement, *etc...*

4.1.1 Classification

Les *Leishmania* font partie de l'ordre des Kinétoplastidae, protozoaires caractérisés par la présence d'un organite spécifique, le **kinétoplaste**, aux fonctions proches de la mitochondrie. La famille la plus importante de cet ordre est celle des Trypanosomatidae (figure 4.1). Cette famille comprend neuf genres dont le genre *Trypanosoma*[6] et le genre *Leishmania*. Ce dernier comprend deux sous-genres, *Leishmania* et *Viannia*, eux-même divisés en complexes d'espèces à répartitions géographiques différentes.

4.1.2 Cycle de vie

Les leishmanies ont un cycle de vie complexe faisant intervenir deux hôtes et plusieurs formes cellulaires [176] (figure 4.2). Lors de son repas sanguin, une femelle de phlébotome infectée peut libérer dans le derme du mammifère des leishmanies sous forme **promastigote** mobile. Chez l'hôte vertébré, les leishmanies parasitent les macrophages. Les leishmanies font en effet partie des rares parasites à survivre dans la **vacuole parasitophore** des macrophages. Dans cette vacuole, les leishmanies perdent leur flagelle (et deviennent non-mobiles), s'arrondissent et entrent en phase de multiplication intense (forme **amastigote**). Lorsque la charge en amastigotes devient trop importante, le macrophage éclate et les amastigotes libérés infectent de nouveaux macrophages, poursuivant ainsi le processus de multiplication. Le cycle est complété lorsqu'un phlébotome femelle prend son repas sanguin à l'endroit de l'infection et ingère des macrophages parasités. Dans le tube digestif du phlébotome les leishmanies se différencient en **promastigotes procycliques**. Ces derniers, flagellés et mobiles, se divisent activement mais ne sont pas infectieux. Après migration jusqu'à la valve du *stomodaeum*, les promastigotes procycliques se transforment en **promastigotes métacycliques** très mobiles et ne se divisent plus. C'est cette forme qui est infectieuse pour les mammifères [176].

6. Ce genre comprend les agents responsables de la maladie du sommeil en Afrique et de la maladie de Chagas en Amérique latine.

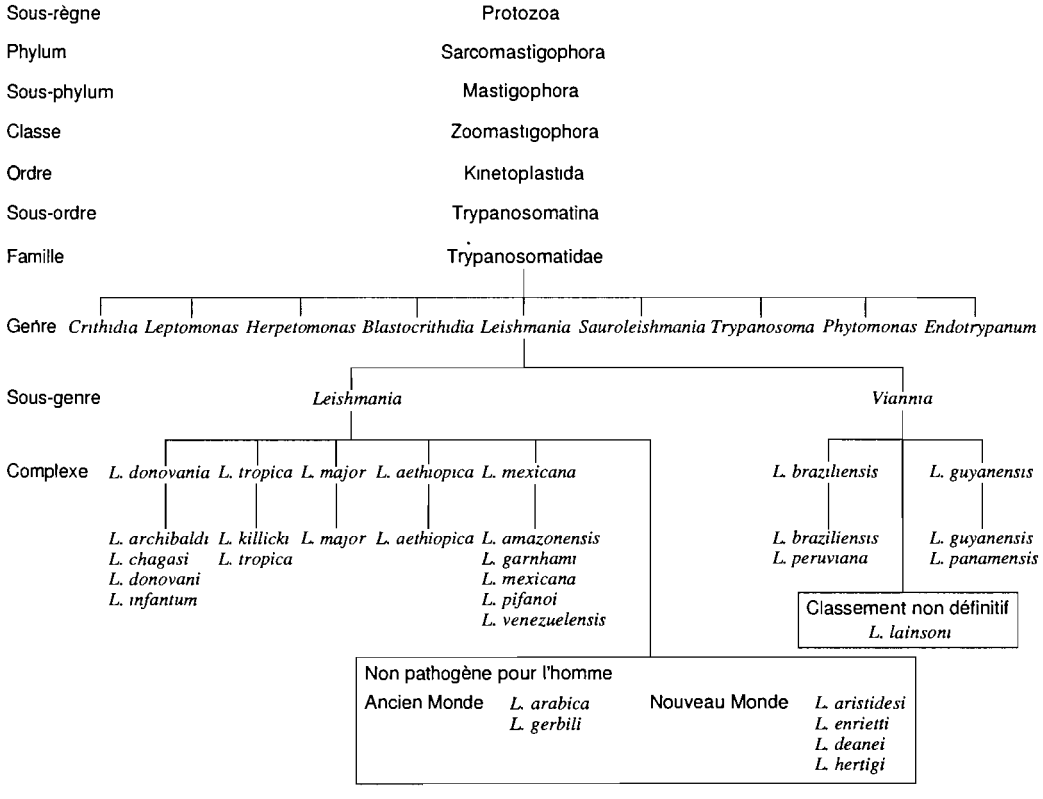


FIG. 4.1 – **Taxonomie des leishmanies. D'après [21].**

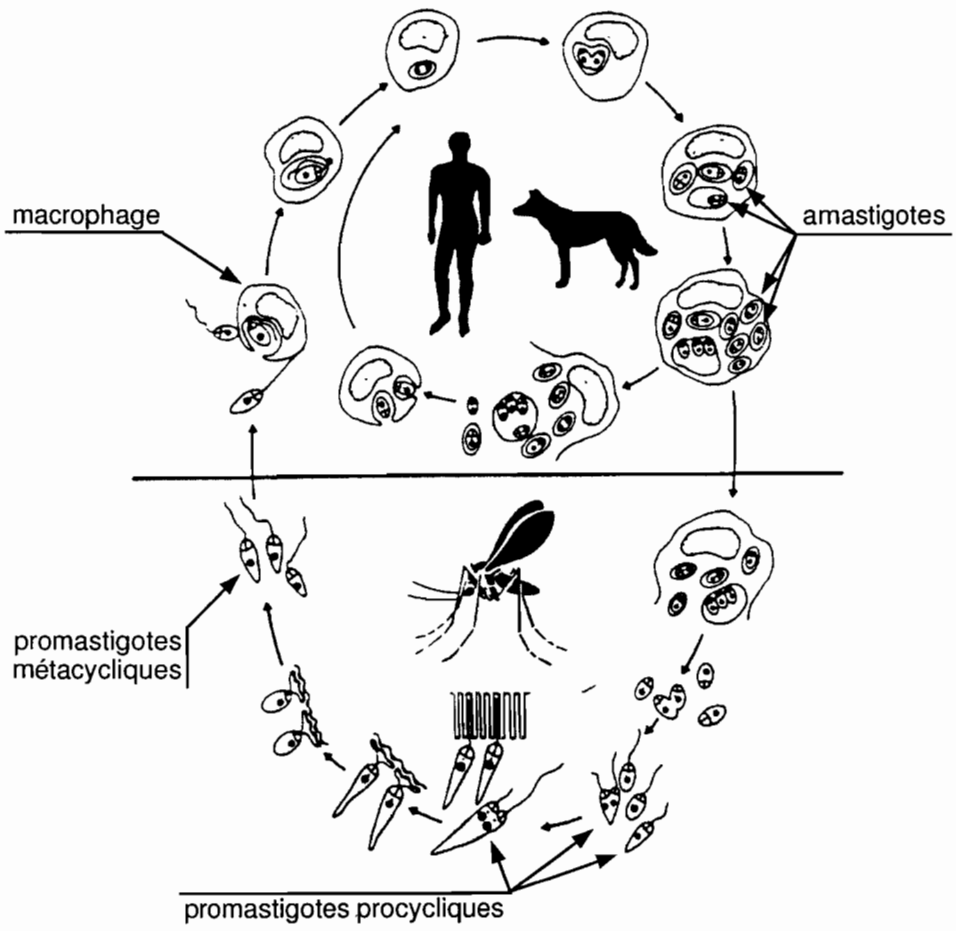


FIG. 4.2 – **Cycle cellulaire des leishmanies. D'après [55].**

4.1.3 Relations avec le système immunitaire de l'hôte mammifère

L'acquisition d'une immunité protectrice contre les *Leishmania* dépend de la capacité de l'hôte mammifère à monter une réponse immunitaire de type Th1 (*T helper 1*) [1]. Cette réponse immunitaire, dite à **médiation cellulaire**, est stimulée par les interleukines 12 (IL-12) et produit des interférons γ (IFN-γ) (voir encadré 4.1). Ces derniers augmentent la production, par les macrophages, d'acide nitrique, un puissant microbicide [183]. La réponse immunitaire de type Th1 est en concurrence avec la réponse immunitaire de type Th2, dite à **médiation humorale**. Cette dernière est stimulée par les IL-4 et produit des interleukines et des anticorps de classe E (IgE) mais, contrairement à la réponse Th1, est incapable de neutraliser la prolifération des *Leishmania*. On ne connaît pas encore très bien ce qui favorise une voie plutôt qu'une autre. L'étude de la différenciation Th1/Th2 est actuellement un domaine de recherche actif en immunologie et la majorité de nos connaissances sur le sujet sont issues d'infections expérimentales de souches de souris génétiquement sensibles comme la BALBc par *Leishmania major* [59]. De façon générale, le système BALBc / *Leishmania major* est très utilisé en immunologie [59].

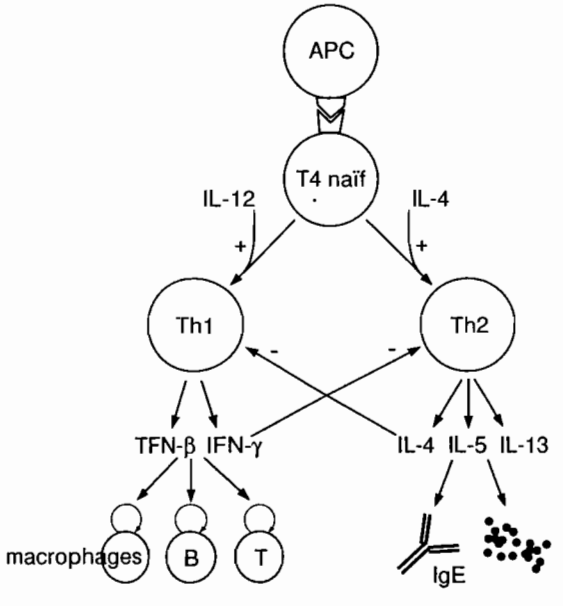
4.2 Les cystéines protéinases

La plupart des parasites protozoaires pathogènes expriment une importante variété de CP [126, 135], voir encadré 4.2. Les Trypanosomatidae contiennent de nombreuses CP de la superfamille des papaïnes. Parmi celles-ci, celles qui ont de loin la plus forte activité sont nommées CP de Type I. Ces CP de Type I sont des cathepsines L-*like*[7] et sont codées par une **famille de gènes répétés en tandem**. De plus, elles sont caractérisées par la présence d'une **extension** d'une centaine d'acides aminés à l'extrémité C de la protéine qui, dans certains cas, est très fortement glycosylée et dont la fonction est actuellement inconnue. Chez les *Leishmania*, ces CP de Type I sont nommées CPB et sont codées par le gène multi-copies *cpb* [190]. L'expression de ces gènes répétés en tandem dépend du stade de développement, avec une activité beaucoup plus forte au stade amastigote qu'aux stades promastigotes [190]. Chez les *Leishmania*, en plus de ce gène *cpb*, il existe deux autres gènes simple-copie codant pour deux autres types de CP (Type II et Type III) et dont on ne connaît pas d'homologue chez les autres Trypanosomatidae. Le

7. On nomme ainsi les CP ayant une structure et une fonction proches des cathepsines L de mammifères.

ENCADRÉ 4.1
LA DIFFÉRENCIATION DES LYMPHOCYTES T4 NAÏFS

Dans l'encadré 3.1, nous avons vu le rôle central que jouent les lymphocytes T4 dans l'organisation de la **réponse immunitaire spécifique**. Nous présentons dans cet encadré les mécanismes de différenciation qui s'opèrent sur les lymphocytes T4, après stimulation par les cellules présentatrices d'antigènes (APC, *Antigens Presenting Cells*, *i.e.* macrophages, cellules dendritiques, *etc...*).



Après exposition à des antigènes parasitaires, les lymphocytes T4 naïfs peuvent se différencier en cellules de type Th1 (*T helper* 1, **réponse à médiation cellulaire**), ou en cellules de type Th2 (**réponse à médiation humorale**) [59]. La première réponse est stimulée par les interleukines 12 (IL-12) et produit des interférons γ (IFN-γ) qui sont des stimulateurs cellulaires (macrophages, lymphocytes B et T). La deuxième est stimulée par les IL-4 et produit des interleukines (IL-3, IL-4, IL-13) favorisant notamment la synthèse d'immunoglobulines (anticorps) E (IgE). Il semble y avoir une sorte de compétition entre les deux types de réponses puisque la différenciation en Th1 tend à inhiber la différenciation en Th2 et *vice versa*. Les causes de spécialisation vers la voie Th1 plutôt que Th2 sont encore mal connues et semblent dépendre, outre du contexte cytokinique, de facteurs liés au parasite et de facteurs (génétiques) liés à l'individu hôte. L'intérêt porté sur l'étude de la dichotomie Th1/Th2 est surtout lié au fait que les IgE produits par les Th2 sont les anticorps responsables des phénomènes d'allergie.

premier, *cpa*, code également pour une cathepsine L-*like*, la CPA (Type II), mais est beaucoup moins abondante que les CPB [136]. Le deuxième, *cpc*, code pour une cathepsine B-*like*[8], la CPC (Type III) [19].

Les CP apparaissent comme indispensables à la survie des leishmanies dans leur hôte mammifère et pourraient entre autres être responsables de la capacité à survivre dans la vacuole parasitophore des macrophages [180]. De plus, il a été reconnu que les cystéines protéinases (CP) sont des **facteurs de virulence** [137, 6]. Il semblerait en effet que les CP de *Leishmania* interféreraient avec le système immunitaire de l'hôte [202]. Il a été proposé que les CP seraient responsables de la destruction des molécules du MCH-II des macrophages, se protégeant ainsi du système immunitaire [43]. Un autre mode d'action des CP serait la stimulation de la production d'IL-4, favorisant ainsi la réponse immunitaire Th2 au détriment de la Th1 [62, 34]. En plus d'être un facteur de virulence, les CP semblent également être **immunogènes** [160]. Ces deux propriétés font de ces molécules des cibles de choix pour la mise au point de médicaments curatifs [135, 172] aussi bien que de vaccins préventifs [81, 127] et sont responsables du récent intérêt porté sur ces enzymes [177].

4.3 Matériels et méthodes

4.3.1 Séquences nucléotidiques

Nous avons travaillé sur des séquences nucléotidiques des gènes codant pour les CP cathepsine-*like*. Les séquences de *Leishmania* publiées ont été collectées sur le site de l'Institut Européen de Bioinformatique[9]. Les séquences de Trypanosomatidae autres que les *Leishmania* ont été recherchées avec WU-blastN[10] et sur le site du *National Center for Biotechnology Information*[11]. Ceci nous a permis de collecter un total de 47 séquences de *Trypanosomatidae*, dont 29 de différents complexes d'espèces de *Leishmania*. En plus de ces séquences, nous avons utilisé une séquence de *Cryptobia salmositica* – famille des Bodonidae, seule séquence trouvée de Kinetoplastidae n'appartenant pas à la famille Trypanosomatidae –, et une séquence de *Plasmodium falciparum* utilisée comme groupe externe dans les phylogénies.

Dans le but de tester la validité des analyses de sélection positive, nous avons également considéré un alignement de 11 séquences du gène *nsnh* de

8. On nomme ainsi les CP ayant une structure et une fonction proches des cathepsines B de mammifères.
9. `http://www.ebi.ac.uk`
10. `http://www.ebi.ac.uk/blast2/parasites.html`
11. `http://www.ncbi.nlm.nih.gov`

ENCADRÉ 4.2
LES CYSTÉINES PROTÉINASES

Les peptides hydrolases (ou **protéases**) sont des enzymes qui catalysent les ponts peptides d'une protéine. Ces enzymes sont indispensables au bon fonctionnement de toute forme de vie, des virus aux vertébrés. Il a été estimé, avec peu de variance inter-spécifique, que les protéases représentent 2% de tous les gènes exprimés [177]. Les **protéinases** sont les protéases qui hydrolysent les ponts peptidiques **à l'intérieur** des protéines. En fonction de leur mécanisme catalytique on distingue les sérine, les cystéine, les aspartique et les métallo-protéinases. Ainsi, les **cystéine** (ou thiol, ou sulfhydryl) protéinases sont celles qui contiennent un résidu cystéine dans leur site actif.

On connaît aujourd'hui plus de 400 cystéines protéinases (CP) dans l'ensemble du monde vivant et dont beaucoup ont des fonctions vitales ou sont impliquées dans des processus pathogéniques. Les CP de parasites protozoaires sont impliquées dans la colonisation des cellules hôtes, l'évitement du système immunitaire, la pathogénicité et la virulence, et sont donc ainsi considérées comme des cibles de choix pour les médicaments ou les vaccins [172, 177]. BARRETT et RAWLINGS ont proposé une classification évolutive des CP [162] et répartissent les 400 CP connues en 6 **super-familles** (ou **clan**) et 30 **familles**. La plupart des CP de protozoaires parasites étudiés à ce jour appartiennent à la **famille C1** (ou famille des papaïnes) du **clan CA** (ou super-famille des papaïnes), un groupe qui contient également des CP de plantes comme la **papaïne** mais aussi des CP de mammifères comme les **cathepsines** B, C, K, L et S qui, en excès, peuvent provoquer chez l'Homme des maladies auto-immunes ou des cancers. Ces multiples fonctions des cystéines protéinases sont responsables du récent intérêt qu'elles ont suscité de la part des chercheurs [177].

Leishmania codant pour la nucléoside hydrolase (EC 3.2.2.1), attendue sous forte sélection purifiante. Ces gènes ont été séquencés par Mallorie HIDE et Anne-Laure BAÑULS au GEMI (IRD Montpellier).

4.3.2 Alignement de séquences et phylogénies

Les séquences ont été alignées avec le programme ClustalX 1.81 [12] [201]. Les terminaisons N et C répétées ont été supprimées des séquences car non-alignables.

Les analyses phylogénétiques ont été réalisées avec le logiciel PHYLIP 3.5c [13] [58]. Les arbres ont été estimés avec les méthodes de parcimonie (al-

12. Gratuit sur `http://www-igbmc.u-strasbg.fr/BioInfo/ClustalX/Top.html`
13. Gratuit sur `http://evolution.genetics.washington.edu/phylip.html`

gorithme de Wagner), de distance (algorithme NJ, *neighbor joining*) et de vraisemblance. Pour les deux dernières méthodes, les modèles d'évolution moléculaires de JUKES et CANTOR [99] et de KIMURA [107] ont été utilisés (voir encadré 2.2 page 38). La séquence de *Plasmodium falciparum* a été utilisée comme groupe externe dans toutes les analyses. Afin de tester la robustesse des estimations, les analyses ont été faites également sur les séquences protéiques.

4.3.3 Analyse de sélection positive

Les analyses de sélection positive ont été réalisées sur quatre alignements de séquences de *Leishmania* correspondant aux gènes *cpa*, *cpb*, *cpc* et *nsnh* (voir tableau 4.1 pour plus d'information sur les séquences). La méthode utilisée est la même que celle décrite au chapitre 3, page 61. Comme dans le chapitre 3, nous avons estimé les modèles M0, M1, M2, M3, M7 et M8. Les résultats des six tests de rapport de vraisemblance étaient cohérents. Les sites sous sélection positive détectés par le bayesien empirique sur les modèles M2 et M8 étaient les mêmes. Le modèle M3, connu pour être plus libéral [225, 12], identifiait les mêmes sites que M2 et M8, plus quelques autres. Pour simplifier, nous ne présenterons que les résultats issus du modèle M8. Pour étudier la sélection positive et effectuer des comparaisons entre les quatre alignements nous nous sommes intéressés à

1. la valeur du $\omega = d_N/d_S$ moyen sur toute la séquence ;
2. la probabilité *a posteriori* d'appartenir à la onzième classe du modèle M8 [14] ;
3. la valeur du $\omega_{11} = d_N/d_S$ dans la onzième classe du modèle M8 ;
4. le nombre de sites considérés sous sélection positive, *i.e.* ceux ayant une probabilité *a posteriori* d'appartenir à la onzième classe du modèle M8 supérieure à 0.95 ;
5. la force de la sélection que nous avons défini précédemment (chapitre 3 page 66) comme étant, pour chaque site, une moyenne pondérée sur toutes les valeurs de ω_i, $i = 1, \ldots, 11$.

Les modèles ont été estimés avec le module CODEML du logiciel PAML 3.1 [15].

14. Rappelons que le modèle M8 contient onze classes de valeurs de ω : $\omega_1 < \omega_2 < \ldots < \omega_{10} < 1 < \omega_{11}$. Seule la dernière, supérieure à 1, rend compte d'une éventuelle sélection positive.

15. Disponible gratuitement à `http://abacus.gene.ucl.ac.uk/software/paml.html`

4.4 Résultats

4.4.1 Phylogénie

L'arbre phylogénétique, présenté en figure 4.3, discrimine clairement les gènes codant pour les cathepsines L-*like* (*i.e.* *cpa* et *cpb*) des gènes codant pour les cathepsines B-*like* (*i.e.* *cpc*), mettant ainsi en évidence que les gènes *cpa* sont plus proches des *cpb* que des *cpc*. La séquence de *Cryptobia salmositica* apparaît proche des *cpa* et *cpb*, au-dessus de la famille des Trypanosomatidae. *cpa* est le seul gène présent uniquement chez les *Leishmania* et les *cpb* de *Leishmania* sont plus proches des *cpa* que des *cpb* de *Trypanosoma*. Les *cpc* de *Leishmania* sont groupés avec les *cpc* de *Trypanosoma*. Au sein de chacun des quatre grands groupes (*cpa*, *cpb* de *Leishmania*, *cpb* de *Trypanosoma* et *cpc*), les séquences s'organisent selon la classification consensuelle de la famille des Trypanosomatidae (voir figure 4.1).

4.4.2 Sélection positive

Les résultats de sélection positive sont reportés dans le tableau 4.1. La figure 4.4 montre, pour chaque site des quatre gènes étudiés, les probabilités *a posteriori* d'appartenir à la onzième classe du modèle M8 (première colonne) ainsi que l'intensité de la sélection (deuxième colonne).

TAB. 4.1 – **Résultats du modèle M8 sur les quatre alignements de séquences.**

Gène	Cod[a]	Seq[b]	$\overline{\omega}^c$	ω_{11}^d	N[e]	P[f]
nsnh	291	11	0.2084	0.0001	0	0.884
cpc	340	6	0.4504	15.8400	1	**0.018**
cpa	353	7	0.5265	5.8909	2	**0.001**
cpb	443	8	0.8840	5.8122	3	**<0.001**

[a]Nombre de codons dans chaque séquences de l'alignement.
[b]Nombre de séquences dans l'alignement.
[c]Valeur moyenne, sur l'ensemble du gène, du rapport d_N/d_S.
[d]Valeur du rapport d_N/d_S dans la onzième classe du modèle M8.
[e]Nombre de sites identifiés comme positivement sélectionnés au seuil de 5%.
[f]Probabilité associée au test de rapport de vraisemblance entre les modèles M7 et M8. C'est un indice de la significativité de la sélection positive dans l'alignement de séquences considéré. Les résultats significatifs (au seuil de 5%) sont en gras.

Le gène *nsnh* ne semble pas sous sélection positive. En effet, il ressort des tests de rapport de vraisemblance qu'aucun des modèles permettant la

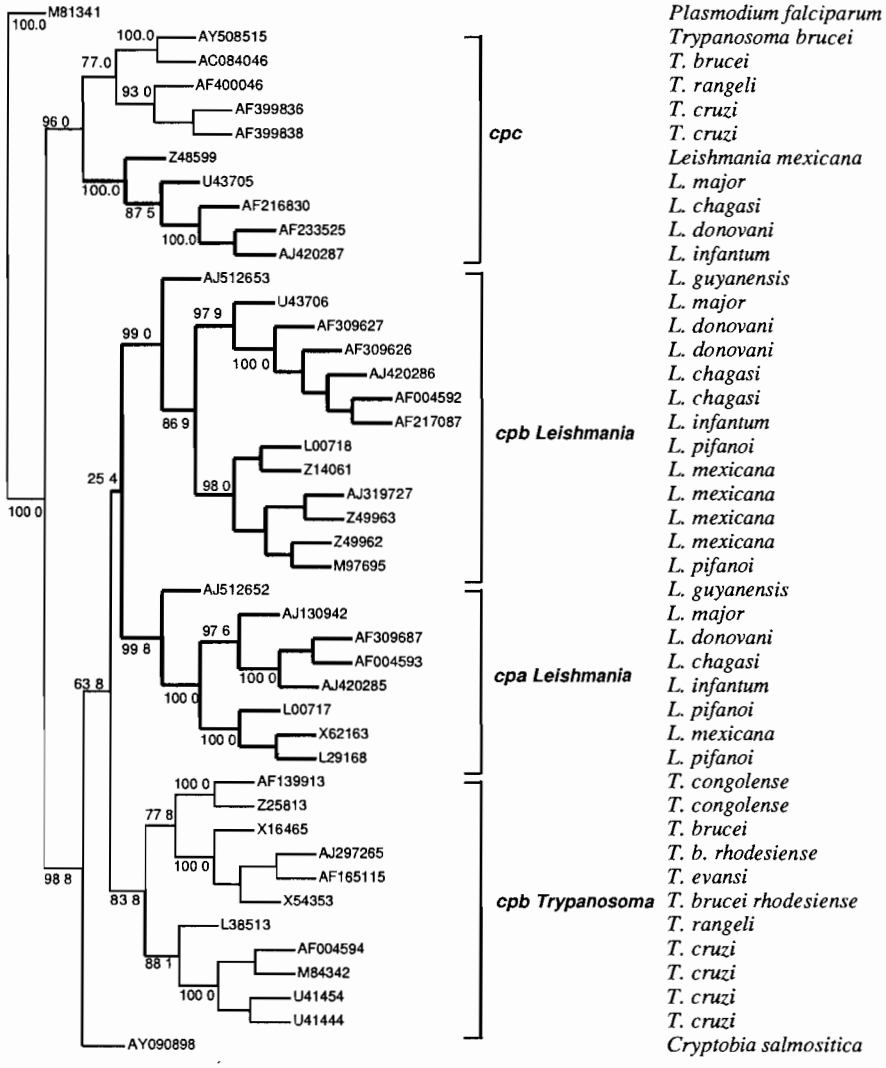


FIG. 4.3 – **Phylogénie des gènes codant pour les CP cathepsines *like* de Trypanosomatidae. L'arbre a été construit par parcimonie après un *bootstrap* à 100 répétitions. La séquence de *Plasmodium falciparum* a été utilisée comme groupe externe. Les séquences de *Leishmania* sont représentées par les branches en gras. Les longueurs de branches sont proportionnelles aux valeurs de *bootstrap*.**

sélection positive (M2, M3, M8) n'a pu rejeter ceux qui ne permettent pas la sélection positive (M0, M1, M7), voir tableau 4.1. Il est également caractéristique sur la figure 4.4 que l'intensité de la sélection est homogène le long de la séquence. Au contraire, les gènes *cpa*, *cpb* et *cpc* présentent tous de la sélection positive (tableau 4.1) avec une intensité croissante de *cpc* à *cpa* et de *cpa* à *cpb*. De plus, le niveau de sélection positive est très hétérogène le long de la séquence avec la plupart des sites sous sélection positive situés dans la seconde moitié du gène, notamment sur la partie terminale C du *cpb* (figure 4.4) [16].

4.5 Discussion

Comme remarqué par Hughes [92], l'histoire évolutive des cystéine protéinases s'est faite essentiellement par **duplication de gènes** (voir encadré 4.3). L'analyse de notre arbre phylogénétique (figure 4.3) suggère au moins trois événements de duplication. Le premier, antérieur à la divergence des familles de Trypanosomatidae, a généré la séparation entre les cathepsines L-*like* et les cathepsines B-*like*. Ensuite, au sein du genre *Leishmania*, un second événement de duplication serait à l'origine des gènes **paralogues** [17] *cpa* et *cpb* codant pour les cathepsines L-*like*. Un troisième événement de duplication aurait généré les différentes copies des *cpb*. La datation de cet événement reste incertaine. L'arbre phylogénétique pourrait laisser croire que ces événements ont eu lieu indépendamment dans les différentes espèces de Trypanosomatidae. Toutefois, une hypothèse plus parcimonieuse est que ces duplications de gènes ont eu lieu avant la divergence des espèces et qu'un phénomène d'**évolution concertée** [18] serait responsable des similarités observées entre les différents exemplaires d'une famille de gène. En effet, puisqu'il a été montré que ces gènes sont fortement immunogènes [160], il est possible que cette famille joue le rôle d'une **bibliothèque d'archives** [64] comme il a été décrit pour le gène *var* de *Plasmodium falciparum* [187].

Plus que la simple identification de la sélection positive sur un alignement de séquence particulier, ce qui nous intéresse ici est la **comparaison** de la forme de sélection qui s'opère sur quatre gènes différents. Parmi ces quatre gènes, *nsnh* sert de référence pour les comparaisons puisqu'il est attendu être

16. Remarquer les différences d'échelles sur les axes des abscisses : le gène *nsnh* est plus court que les autres (291 codons). Les gènes *cpc* et *cpa* sont de taille comparable (environ 350 codons). Par rapport aux gènes *cpc* et *cpa*, le gène *cpb* est caractérisé par une extension terminale d'une centaine de codons à l'extrémité C.

17. Voir encadré 4.3 pour une définition.

18. Voir encadré 4.3 pour une définition.

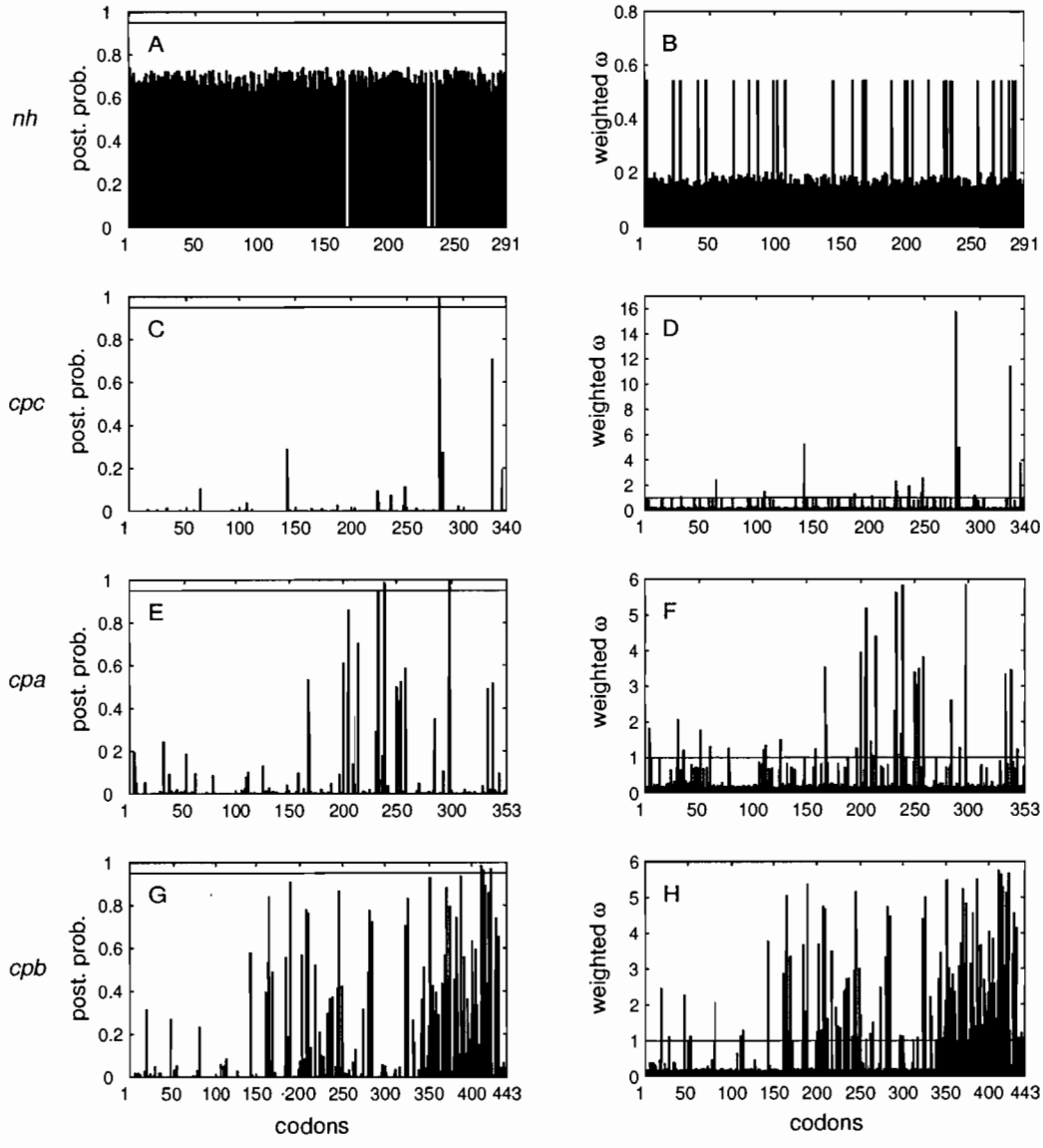
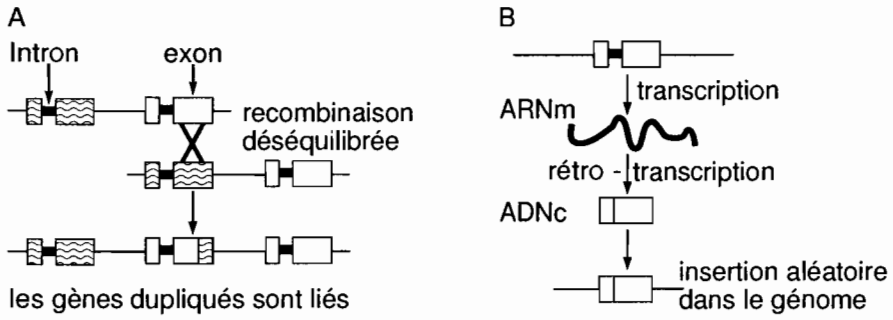


FIG. 4.4 – **Résultats du modèle M8. La première colonne (A, C, E, G) présente les probabilités *a posteriori* d'appartenir à la onzième classe de M8 et la seconde colonne (B, D, F, G) montre les intensités de sélection pour chaque site des gènes *nsnh* (première ligne), *cpc* (deuxième ligne), *cpa* (troisième ligne) et *cpb* (quatrième ligne). Les lignes horizontales représentent le niveau de significativité à 0.95 dans la première colonne et la valeur seuil du niveau de sélection positive (*i.e.* 1) dans le deuxième colonne. Noter les différences d'échelles sur les axes des ordonnées dans la seconde colonne.**

ENCADRÉ 4.3
ÉVOLUTION PAR DUPLICATION DE GÈNES

Au niveau moléculaire, on peut considérer que l'évolution fait intervenir (*i*) une augmentation de la quantité d'ADN et (*ii*) une augmentation de la quantité d'information par substitution de nucléotides [145]. L'augmentation de la quantité d'ADN se produit essentiellement par duplication de gène [145]. Le phénomène de duplication de gènes fournirait en quelque sorte la matière première à l'évolution biologique [228].



La duplication de gènes peut se produire par **recombinaison déséquilibrée** (A) ou **rétro-transposition**, *i.e.* rétro-transcription d'un ARN messager en ADN (B). Les gènes dupliqués sont dits **paralogues** et forment une **famille de gènes**. La duplication produit de la **redondance fonctionnelle**. Cette redondance peut être sélectionnée si elle permet d'amplifier l'expression d'un gène très sollicité, comme le gène codant pour l'ARN ribosomial. Si ce n'est pas le cas, la sélection purifiante sur l'une des copies est relâchée et cette dernière peut évoluer par dérive. Le plus souvent, le gène perd sa fonction sans en acquérir d'autre et devient un **pseudogène**.

Les différents variants d'un gène peuvent être conservés si des phénomènes de recombinaison entre eux permettent de générer une diversité avantageuse. C'est par ce mécanisme que la diversité des immunoglobulines est générée. Ce phénomène est appelé **conversion de gènes**. En fonction de sa fréquence, il peut conduire à un degré de similarité entre les différents gènes paralogues plus fort qu'attendu en cas d'évolution indépendante. On parle dans ce cas d'**évolution concertée**.

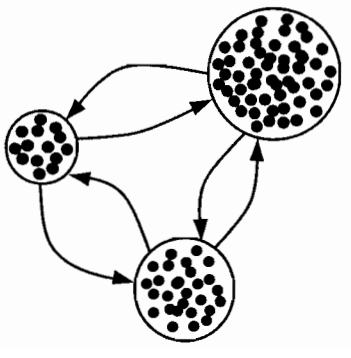
Enfin, les gènes dupliqués peuvent se **spécialiser** dans certaines fonctions ou certains compartiments et même évoluer vers de **nouvelles fonctions**.

sous forte sélection de conservation. Les trois autres gènes (*cpa*, *cpb* et *cpc*) codent pour les trois types de CP répertoriés chez les *Leishmania*. Tous les gènes *cp* montrent des niveaux significatifs de sélection positive. Le gène *cpc*

semble avoir une valeur exceptionnellement forte de ω_{11}. Ceci doit toutefois être interprété avec prudence et peut être un *artefact* de la distribution beta. En effet, le fait que la très grande majorité des sites soit sous sélection fortement purifiante peut être responsable de la forte valeur du rapport d_N/d_S allouée aux rares sites sous sélection positive. De façon générale, une grande prudence doit être adoptée dans la comparaison des ω_{11} obtenus sur différents alignements de séquences [38]. Cette remarque étant faite, les résultats suggèrent que *cpb* a une évolution diversifiante plus forte que *cpa* et que lui-même a une évolution diversifiante plus forte que *cpc*. Ceci peut être expliqué par différentes fonctions cellulaires des trois types de CP et donc par différentes pressions de sélection. Le gène *cpb*, qui joue un rôle important dans les **interactions hôte-parasite**, serait ainsi sous forte pression de sélection et évoluerait plus rapidement que les autres *cp*. Enfin, il est intéressant de noter que la répartition de l'intensité de sélection sur les 350 premiers codons de *cpa* est très similaire à celle observée sur les 350 premiers codons de *cpb* (figure 4.4F et H) : la sélection positive semble localisée sur des régions **homologues** des deux gènes. De plus, sur *cpb*, la sélection positive apparaît particulièrement intense sur les 100 derniers sites, correspondant à la longue extension terminale de l'extrémité C, absente chez les autres *cp*. Bien que l'on ne connaisse pas encore la fonction précise de cette partie terminale, ces résultats semblent suggérer une fonction dans les relations hôte-parasite. Si cette hypothèse s'avérait exacte, cette région des CPB pourrait alors constituer une cible vaccinale de choix.

Deuxième partie

INTERACTIONS POPULATIONNELLES



Chapitre 5

Dynamiques de maladies[1]

COMPRENDRE les variations du nombre de malades dans les populations a motivé les premières recherches en épidémiologie. Aujourd'hui, l'épidémiologie est devenue une discipline qui a étendu ses domaines d'expertises en dehors du simple champ de la dynamique des populations. On parle par exemple d'**épidémiologie moléculaire** pour désigner les recherches utilisant les outils moléculaires afin de comprendre le comportement des agents aetiologiques dans les populations (voir partie 1). Nous parlons, dans cette partie 2, d'épidémiologie au sens originel et restreint du terme, c'est à dire de dynamique de maladies, dans l'espace et dans le temps.

Les données épidémiologiques s'accumulent depuis les premiers écrits. Les Egyptiens du IIième millénaire avant notre ère rapportent déjà des épidémies dévastant de grandes populations [175]. On retrouve de tels témoignages chez les premières civilisations chinoises, chez les auteurs et historiens de la Grèce antique et de l'empire romain où les pestes déciment des populations entières [175]. Les maladies contagieuses ont effrayé les populations, non seulement pour leur caractère dévastateur, mais aussi pour leur aspect **imprévisible**[2]. Cette imprévisibilité des épidémies a été à l'origine de vastes mouvements de panique et de comportements humains qui nous paraissent aujourd'hui complètement absurdes et irrationnels.

Les premières tentatives de compréhension du comportement dynamique des maladies se font avec le développement des **modèles mathématiques**. Le tout premier est celui développé en 1760 par Daniel Bernoulli [25] dans le but de comprendre l'effet de la variolisation sur l'espérance de vie moyenne d'une population (voir introduction générale, chapitre 1). Il a fallu ensuite

1. Une partie de ce chapitre a donné lieu à un manuscrit en préparation (Roche B., Choisy M., Dorléans Y., Flahault A. & Guégan J.F.).

2. Avant la découverte du microscope à la fin du XVIIième siècle, la plupart des agents aetiologiques des maladies contagieuses étaient inconnus.

attendre le XXième siècle pour assister au développement de l'épidémiologie théorique avec les travaux pionniers de Hamer [80] et Ross [173], suivis du théorème de Kermack et McKendrick [105] et des modèles mathématiques de Bailey [16].

L'étude de la dynamique temporelle et spatiale des maladies infectieuses a connu un intérêt particulièrement marqué ces toutes dernières années et ce, pour deux raisons. La première est d'ordre **pratique** et liée à l'émergence et la réémergence de nombreuses maladies infectieuses ces deux dernières décennies (voir introduction générale, chapitre 1). Dans un tel contexte, la compréhension de la propagation d'une maladie dans le temps et l'espace est cruciale pour une lutte efficace [17, 29, 118, 50, 169, 72, 102]. La deuxième raison est plus d'ordre **fondamental**. En effet, l'accumulation, depuis une quarantaine d'années, de données de notifications en de nombreuses localités offre une opportunité sans précédent pour les théoriciens en dynamique des populations de tester leurs modèles [178, 28, 164, 51, 203, 104, 26]. Ceci est d'autant plus vrai que ces données épidémiologiques fines sont très souvent couplées avec des données démographiques (taux de natalité, taille de la population, *etc...*) de qualité comparable. Dans ce contexte, la mise en place des politiques vaccinales joue, pour les théoriciens, le rôle d'expérience grandeur nature.

Parmi les grandes questions fondamentales actuellement posées aux systèmes dynamiques biologiques, citons la compréhension de la variabilité observée dans les séries temporelles. En particulier, quelles sont les parts respectives du **chaos** (source de variation intrinsèque) et de la **stochasticité** (source de variation extrinsèque) responsables de la complexité observée [28, 150, 171] ? En effet, depuis la première proposition théorique de chaos en biologie par Robert MAY dans les années soixante-dix [122, 123], la question de l'existence effective de chaos dans les systèmes biologiques réels est encore très largement débattue [41, 158, 157, 88, 52].

Les préoccupations plus pratiques gravitent autour de la **vaccination** et de ses conséquences. Dans cette partie 2, nous parlons donc des aspects dynamiques des politiques de vaccination. En effet, si le but ultime d'une politique vaccinale est d'éradiquer une maladie d'une population, ceci est rarement atteint en pratique. Tout au plus arrive-t-on à diminuer fortement le nombre de cas et la maladie persiste dans la population à bas bruit. On parle dans ce cas de **vaccination imparfaite**. De telles vaccinations imparfaites, si elles ne parviennent pas à éradiquer une maladie d'une population, peuvent en modifier notablement le comportement (voir introduction générale, chapitre 1). C'est ce à quoi nous nous intéressons dans cette partie 2.

Les chapitres 6 et 7 traitent respectivement des deux types de politique vaccinale majeurs : la vaccination classique de masse et la vaccination par

pulsations, récemment proposée et en cours d'application. Avant cela, le présent chapitre est consacré à la compréhension de la dynamique temporelle et spatiale des maladies infectieuses, en l'absence de vaccination. Connaître la dynamique d'une maladie en l'absence de vaccination est important pour l'application de politiques vaccinales optimisées. Nous présentons les résultats issus de la littérature sur la dynamique spatiale de la rougeole en Grande-Bretagne que nous comparons ensuite à nos résultats obtenus sur la varicelle en France.

5.1 Pourquoi la dynamique spatiale ?

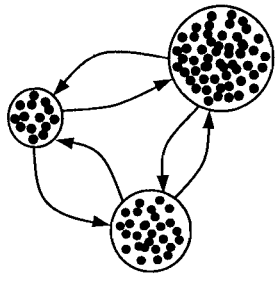
Les individus ne sont pas fixes, ils se déplacent. La conséquence est que, à une extinction locale, peut succéder une recolonisation par migration. D'un point de vue pratique, l'étude des dynamiques spatiales devient primordial dès lors que l'on s'intéresse aux phénomènes d'**extinction/recolonisation** [84, 85, 83]. Ceci concerne principalement la biologie de la conservation et l'épidémiologie. La première cherche à éviter les extinctions tandis que la seconde a pour but de les favoriser [50].

Plusieurs types de modèles ont été proposés pour décrire la dynamique spatiale. Parmi ceux-là, les modèles de **réaction-diffusion** considèrent une dimension spatiale continue [111] tandis que les modèles **métapopulationnels** décrivent une population comme étant composée de sous-populations (voir encadré 5.1). La répartition des êtres humains n'est pas homogène, les individus étant clairement agrégés dans les villes. Dans ce contexte, le formalisme de la théorie métapopulationnelle est particulièrement bien adapté pour l'étude de la dynamique spatiale des maladies contagieuses [75].

Comprendre la dynamique spatiale des maladies contagieuses est fondamental pour la mise en place d'une politique vaccinale efficace à large échelle. L'étude de la dynamique spatiale d'une maladie consiste essentiellement à déterminer ce qui gouverne (*i*) les **extinctions locales** (taille de sous-population, taux de migration, distance aux autres sous-populations) et (*ii*) la **propagation** des maladies dans l'espace. Ce dernier point est important pour son caractère prédictif. Nous commençons par présenter ces différents concepts à partir de l'exemple de la rougeole en Angleterre. Dans le paragraphe suivant, nous voyons une application à la varicelle en France, basée sur notre propre étude.

ENCADRÉ 5.1
LE MODÈLE DE MÉTAPOPULATION

Une métapopulation est une population de populations d'individus (ronds noirs sur le dessin ci-dessous). Ces dernières sont appelées **sous-populations** (ou dèmes, cercles sur le dessin ci-dessous) et sont connectées entre elles par des migrations d'individus. Ainsi, un individu entretient la plupart de ses contacts avec les individus de sa propre sous-population et quelques rares contacts avec les individus des autres sous-populations.



On parle de **population structurée** et, selon l'intensité des migrations, la structure des populations est plus ou moins marquée. Cette structure de population a des conséquences mesurables en terme de génétique et de dynamique des populations. Pour le premier point nous renvoyons le lecteur à l'encadré 2.1, page 33. Le deuxième point concerne la dynamique des **extinctions/recolonisations** qui dépend fortement des **tailles** des sous-populations et des intensités des **migrations**. Un exemple de schéma de dynamique métapopulationnelle est le schéma **source/puits** dans lequel les individus migrent des grandes populations (sources) vers les petites populations (puits), trop petites pour persister dans le temps. Il existe une variété de modèles métapopulationnels, le plus simple étant le **modèle en îles** qui considère une équiprobabilité de migration entre toutes les sous-populations. Un modèle plus complexe et plus réaliste est le **modèle en *stepping-stones*** dans lequel les populations sont organisées sur un réseau, les individus de chaque population ne pouvant migrer que vers les populations adjacentes.

5.2 Dynamique spatiale de la rougeole

5.2.1 Présentation de la rougeole

La rougeole est une maladie **virale**[3] à **transmission directe** et à **immunité permanente**. Cette maladie est très étudiée, essentiellement pour la simplicité de son cycle de vie, facile à modéliser, ainsi que l'abondance et la qualité des données liées à sa forte prévalence (en l'absence de vaccination) et son diagnostic inconfondable. De plus, la rougeole est encore aujourd'hui une importante **cause de mortalité** dans les pays en voie de développement[4]. La forte transmission (par aérosols) couplée avec la permanence de l'immunité sont responsables d'un âge moyen à l'infection extrêmement bas (environ 5 ans). De plus, au-delà de 10 ans, 90% des individus ont contracté la rougeole. On nomme habituellement **maladies infantiles** les maladies microparasitaires dont l'âge moyen à l'infection est bas. L'infection est suivie d'une phase asymptomatique d'environ une semaine, au cours de laquelle la population virale se développe et colonise l'organisme. Cette phase de latence est suivie d'une phase symptomatique d'environ une semaine également, au cours de laquelle l'individu devient contagieux. Après élimination du virus par le système immunitaire, l'immunité conférée est permanente. Cette succession de phases rend appropriée une **modélisation en compartiments** de type SEIR dans laquelle la population hôte est divisée selon le stade clinique des individus : susceptibles (S), contaminés mais non encore contagieux (E), contagieux (I), et définitivement guéris (R)[5]. Les individus passent du compartiment S au compartiment E, puis I et enfin R.

Plusieurs méthodes ont été utilisées pour étudier la dynamique spatiale de la rougeole. Nous nous intéressons d'abord à la persistance des maladies en fonction de la taille de la sous-population hôte. Nous nous concentrons ensuite sur les déterminants de la propagation de la maladie dans l'espace.

5.2.2 Persistance

Lorsque l'on s'intéresse à la persistance locale ou globale d'une maladie, deux concepts distincts sont d'importance [101, 103]. Le premier concerne le **taux de reproduction basal**[6] R_0. Cette quantité définit le nombre moyen

3. Morbillivirus (virus à ARN) de la famille des Paramyxoviridae. C'est un virus enveloppé de 100 à 250 nm de diamètre.

4. En 1995, la rougeole a causé pas moins de 500 000 décès en Afrique centrale (source : OMS).

5. Les lettres correspondent aux noms anglais : S (*susceptible*), E (*exposed*), I (*infectious*), R (*recovered*).

6. Noté à tort « taux » alors que c'est un nombre sans dimension.

d'infections causées par l'introduction d'un individu susceptible dans une population constituée entièrement d'individus sains [11]. De sa définition on déduit que R_0 présente une valeur seuil : pour $R_0 < 1$ la maladie ne peut pas persister dans la population tandis que pour $R_0 \geq 1$ la maladie peut persister dans la population. Rappelons que R_0 est un nombre **moyen** et que la réalité est variable par rapport à la moyenne. Ainsi, $R_0 \geq 1$ ne garantit pas une persistance dans la population et la simple **stochasticité démographique** peut conduire à une extinction. Ceci nous mène à la deuxième notion : même lorsque $R_0 \geq 1$, la persistance de la maladie sur le long terme dépend d'un seuil lié à la **taille de la population**, au **taux de contact** entre individus et au **taux de contagion** [101, 103]. Ce seuil est généralement déterminé empiriquement en termes de taille de population au-dessous de laquelle la maladie ne peut pas persister sur le long terme. On appelle cette taille de population seuil CCS (de l'anglais *critical community size*).

La CCS peut être déterminée à partir de séries temporelles de maladies dans plusieurs localités de tailles différentes. La procédure classique consiste à représenter sur un graphique la durée moyenne annuelle des extinctions en fonction de la taille de la population. Une extinction est définie comme période sans cas d'une durée supérieure ou égale au **temps de génération de la maladie** (*i.e.* durée de la période de latence [E] + durée de la période de contagion [I]). La CCS est alors définie comme la taille de la population au-dessous de laquelle la durée moyenne annuelle sans cas est supérieure au temps de génération de la maladie, voir figure 5.1 [11].

Une CSS de 115 000 habitants signifie que la rougeole ne peut pas persister sur le long terme dans une population isolée de moins de 115 000 habitants. C'est essentiellement pour cette raison que l'on ne trouve pas de maladies à forte contagion et immunité permanente, comme la rougeole, dans les petites communautés isolées d'indiens d'Amazonie.

Une autre méthode consiste à représenter le nombre annuel moyen d'extinctions en fonction de la taille de la population, une extinction étant alors définie comme une période sans cas, d'une durée supérieure au temps de génération de la maladie. Dans le cas de cette deuxième méthode, la CCS est déterminée par l'intersection de la courbe de régression avec la droite d'ordonnée 1. Les deux méthodes donnent généralement des résultats très similaires. La figure 5.4 a été réalisée avec cette deuxième méthode.

5.2.3 Diffusion de la maladie dans la population

L'étude de la diffusion d'une maladie dans une population est basée sur la **corrélation** entre les dynamiques de la maladie dans plusieurs localités géographiques différentes. Deux types d'hypothèses intéressantes et commu-

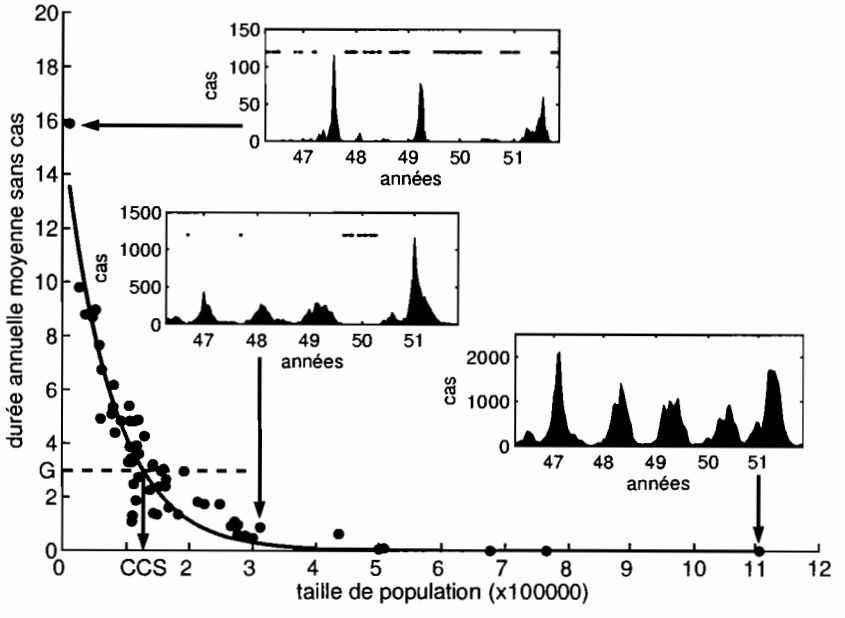


FIG. 5.1 – **Détermination graphique de la CCS de la rougeole en Grande-Bretagne durant la période pré-vaccinale 1944-1966. La durée moyenne annuelle sans cas est représentée en fonction de la taille de la population pour 59 villes. Les trois séries temporelles en encart illustrent les trois niveaux de persistance identifiés par Bartlett [20]. Les dynamiques de Type I (encart du bas, Birmingham, population de 1,1 million d'habitants) sont régulières, endémiques, sans extinction. Les dynamiques de Type II (encart du milieu, Nottingham, population de 300 000 habitants) sont régulières, mais avec quelques extinctions (représentées par des points noirs) dans les creux. Enfin, les dynamiques de Type III (encart du haut, Teignmouth, 11 000 habitants) sont irrégulières, avec de longues périodes sans cas entre les épidémies. La courbe représente la régression non-linéaire ($y \simeq 16 \times \exp[-10^{-5}x]$) et son intersection avec le temps de génération de la maladie (G, ici estimée à 3 semaines) donne la CCS à environ 115 000 individus. Les données sont issues de** `http ://www.zoo.cam.ac.uk/zoostaff/grenfell/measles.html`.

nément testées sont (*i*) le modèle de diffusion **source/puits** (voir encadré 5.1) et (*ii*) les **vagues de propagation**. Il existe plusieurs méthodes statistiques pour l'étude de la corrélation entre séries temporelles. Des plus simples aux plus complexes, celles-ci font intervenir les autocorrélations croisées, les spectres croisés de Fourier et les décalages de phase estimés à partir de décomposition en ondelettes. Ces méthodes appliquées aux données de rougeole en Angleterre ont mis en évidence une spectaculaire hiérarchisation de type source-puits entre zones urbaines et zones rurales, ainsi qu'une propagation dans l'espace d'ondes de prévalence [73, 72]. Il semble donc qu'au Royaume-Uni les grands pôles urbains jouent le rôle de foyers où sont initiées et d'où se propagent les épidémies de rougeole. Cette observation est d'importance en termes de santé publique puisqu'elle permet non seulement de prévoir dans le temps et l'espace l'apparition des épidémies, mais aussi de cibler les campagnes vaccinales.

5.3 La varicelle en France

Nous avons appliqué les méthodes de détection de dynamique spatiale présentée ci-dessus à la varicelle en France. La varicelle se distingue de la rougeole, d'une part par une moindre contagiosité, et d'autre part par sa relation avec une autre maladie, le **zona** [215, 44, 191]. Il est possible que cette dernière particularité influence la dynamique spatiale. C'est ce que nous nous proposons d'explorer sur les données françaises. Un modèle stochastique a été développé nous permettant de tester les hypothèses sur l'influence du zona.

5.3.1 Cycle de la varicelle

Comme la rougeole, la varicelle est une maladie virale infantile[7] assez contagieuse et se transmet par aérosols [215, 44, 191]. Bénigne chez l'enfant, elle peut entraîner des **complications** neurologiques et pneumoniques graves chez l'adulte [18, 124]. En France, il n'existe actuellement pas de politique vaccinale contre la varicelle. La mise en place d'une telle politique est envisagée et actuellement à l'étude [44]. L'agent aetiologique est le VZV (*Varicella-Zoster Virus*), un herpesviridae strictement humain[8] [215]. La contamination est suivie d'une période de latence asymptomatique, d'environ deux semaines, et d'une période symptomatique de contagion, d'environ une semaine [191].

7. 90% des cas sont d'âge inférieur à 14 ans.

8. Virus enveloppé à ADN d'environ 120-200 nm de diamètre. Le génome contient entre 120 000 et 220 000 nucléotides.

Après guérison, l'immunité contre la varicelle est permanente [191]. Toutefois, le virus n'est pas éliminé de l'organisme et devient quiescent dans les ganglions sensitifs des racines postérieures de la moelle osseuse et dans les ganglions de certains nerfs crâniens [191]. Le zona est une récurrence plus ou moins tardive du virus qui affecte 15 à 30% des cas selon les estimations [33, 191]. Des individus susceptibles au contact de personnes avec zona peuvent contracter la varicelle.

5.3.2 Matériels et méthodes

Les données

Les données sont issues du **réseau Sentinelles**, organisé depuis 1984 par l'unité 444 de l'INSERM [9], et qui assure la surveillance de plusieurs maladies transmissibles sur l'ensemble du territoire. L'information est recueillie auprès de plus de 1 200 médecins généralistes volontaires [10], recrutés de telles sorte que leur répartition géographique régionale soit proche de celle de l'ensemble des médecins généralistes français [205]. Afin de distinguer les périodes sans cas des périodes sans déclaration, les médecins en activité doivent se connecter sur le réseau au moins une fois tous les douze jours (période d'exclusion), quel que soit le nombre de cas.

La varicelle est ainsi surveillée par le réseau depuis 1991. Les symptômes caractéristiques de la varicelle rendent son diagnostic relativement facile. Chaque médecin sentinelle déclare sur le réseau le nombre de cas de varicelle, de complications, ainsi que l'âge et le sexe du patient. Des mesures de **redressement de données** ont été appliquées dans le but d'éliminer les éventuelles distorsions de représentativité temporelles et spatiales [206]. Nous avons pu ainsi produire des séries temporelles d'incidence hebdomadaire de varicelle pour 753 communes françaises. Toutefois, à cause de la disponibilité en médecins sentinelles, beaucoup de ces localités ne contiennent que des séries temporelles courtes ou partielles (voir figure 5.2). En conséquence, en fonction de la qualité des données requise, seul un sous-échantillon de ces 753 communes a été parfois utilisé dans les analyses présentées ci-dessous. La taille de la population de chaque commune est issue de l'INSEE.

9. Dans le cadre d'une convention avec l'Institut National de Veille Sanitaire et la Direction Générale de la Santé.

10. Soit environ 1% de la population totale de médecins généralistes.

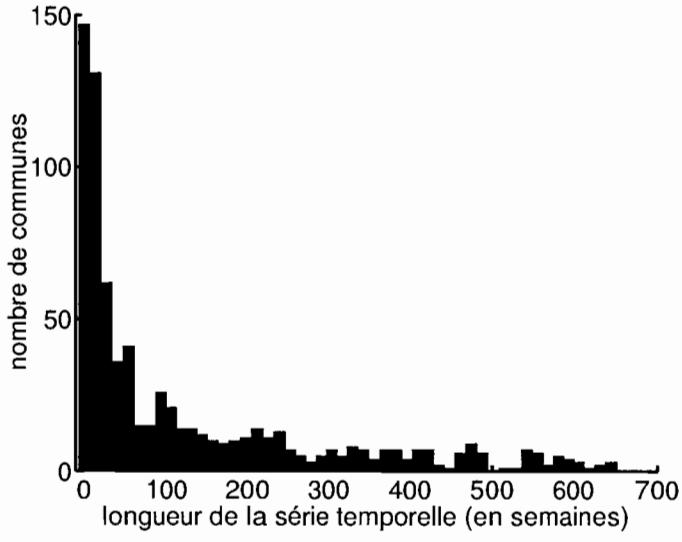


FIG. 5.2 – **Histogramme de la longueur des séries temporelles de varicelle dans les 753 communes étudiées. Noter que la plupart sont très courtes et donc inutilisables dans les analyses.**

Persistance locale

L'influence de la taille de la sous-population sur la persistance de la varicelle a été explorée par la détermination de la CCS (voir paragraphe 5.2.2). Seules les 60 communes avec plus de 10 années de notification ont été considérées pour le calcul de CCS.

Dynamique source/puits

L'éventualité d'une hiérarchisation entre communes a été explorée par l'étude de la synchronie entre les dynamiques moyennes de deux groupes de séries temporelles. Le premier groupe contient les communes de grandes tailles de population tandis que le deuxième groupe contient les communes de petites tailles de population. La limite entre grande taille et petite taille est fixée arbitrairement et plusieurs valeurs ont été essayées pour tester la robustesse des résultats. Chacune des deux séries moyennes a été décomposée en ondelettes puis filtrée sur la fréquence principale (*i.e.* annuelle), et la phase de la série filtrée extraite (voir encadré 5.2). La différence de phase entre les deux séries moyennes a ensuite été calculée pour chaque semaine, produisant ainsi une série temporelle de différence de phase entre communes de grande taille et communes de petite taille.

ENCADRÉ 5.2
DÉCOMPOSITION EN ONDELETTES

La décomposition en ondelettes est inspirée de la décomposition en somme de Fourier. Le théorème de Fourier stipule que tout signal **périodique** de fréquence F_0 peut être décomposé en une somme de sinusoïdes :

$$s(t) = a_0 + \sum_{n=1}^{+\infty} [a_n \cos(2\pi F_0 nt) + b_n \sin(2\pi F_0 nt)]$$

où a_0 est la moyenne du signal et a_n et b_n ($n \in \mathbb{N}$) sont les coefficients de Fourier. Ces derniers représentent la part de variance totale du signal $s(t)$ expliquée par chaque harmonique de fréquence nF_0. En d'autres termes, c'est une mesure de la corrélation entre le signal $s(t)$ et la sinusoïde de fréquence nF_0. L'utilisation de la relation $\mathrm{e}^{j\theta} = \cos\theta + j\sin\theta$ nous permet de reécrire l'équation ci-dessus sous la forme complexe plus synthétique :

$$s(t) = \sum_{n=-\infty}^{+\infty} c_n \mathrm{e}^{jn(2\pi F_0)t}$$

Cette fois-ci, les coefficients c_n sont naturellement complexes. Décomposer le signal en somme de Fourier consiste à estimer les coefficient c_n :

$$c_n = \frac{1}{T_0} \int_0^{T_0} s(t) \mathrm{e}^{-j2\pi F_0 nt} dt$$

Les signaux $s(t)$ et c_n représentent exactement la même réalité physique, le premier dans le **domaine temporel** et le deuxième dans le **domaine fréquentiel**. En fonction du problème traité, il peut être plus aisé de travailler dans l'un ou l'autre des domaines. Par exemple, **filtrer** un signal consiste à sélectionner un des coefficients c_n. Voyons maintenant deux généralisations de la décomposition en somme de Fourier.

La première généralisation concerne le traitement des **signaux non périodiques**, mais toujours stationnaires (*i.e.* à moyenne et variance indépendantes du temps). On peut considérer qu'un signal non périodique est un signal périodique de période égale à l'infini. L'équation ci-dessus se généralise donc en l'équation de la **transformation de Fourier** :

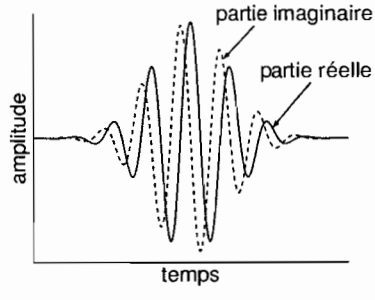
$$S(f) = \int_{-\infty}^{+\infty} s(t) \mathrm{e}^{-j2\pi ft} dt$$

La seconde généralisation concerne le traitement des **signaux non stationnaires** (*i.e.* où la fréquence et/ou l'amplitude varient en fonction du temps). La première tentative dans cette direction a été l'utilisation de transformée

(Encadré 5.2 suite)

de Fourier en **fenêtres coulissantes**. L'idée est de réaliser une transformée de Fourier localement sur une portion de la série seulement, de faire coulisser la fenêtre pour passer à une autre portion de la série, et ainsi de suite. Ceci est donc une façon de visualiser l'information d'un signal dans les domaines temporel et fréquentiel en même temps. Le problème majeur avec ces analyses est que la taille de la fenêtre est fixe, impliquant une sur-représentation des grandes fréquences par rapport aux petites. Ce problème a été résolu avec l'invention des ondelettes.

Une transformée en ondelettes est comme une transformée de Fourier mais basée sur l'utilisation d'ondelettes au lieu de sinusoïdes. Les ondelettes sont des **dilatations** dans le domaine fréquentiel et des **translations** dans le domaine temporel d'une **ondelette mère**. Il existe une variété d'ondelettes décrites dans la littérature. La plus utilisée en écologie et celle que nous avons choisie ici est l'**ondelette de Morlet**. Cette dernière est en fait simplement le produit d'une sinusoïde complexe et d'une enveloppe gaussienne (voir figure ci-dessous).



Le fait que cette ondelette soit complexe nous permet de séparer facilement la phase et l'amplitude du signal étudié.

La significativité de la différence de phase a été testée par simulations de Monte Carlo (voir encadré 3.3 page 67). Soit x le nombre de communes de petite taille et y le nombre de communes de grande taille. A chaque répétition des simulations de Monte Carlo, les séries temporelles ont été réechantillonnées au hasard et avec remise de façon à constituer deux groupes de x et y séries temporelles respectivement et ce, sans tenir compte des tailles des communes. Les séries moyennes des deux groupes ont été calculées et leur différence de phase estimée. Quatre cent répétitions se sont avérées être suffisantes pour atteindre une distribution stationnaire de la différence de phase attendue sous l'hypothèse nulle, et ainsi tester la significativité de la différence de phase observée.

Ces analyses ont été réalisées sur les 753 communes en essayant les tailles

seuils entre petites et grandes communes suivantes : 10 000, 50 000, 100 000, 150 000 et 200 000 habitants.

Modèle stochastique

Le modèle stochastique construit est basé sur un modèle SEIR classique [11] auquel nous avons rajouté un compartiment rendant compte des cas de zona. La figure 5.3 décrit le modèle pour une sous-population de la métapopulation. Le modèle de métapopulation considéré est le simple modèle en îles (voir encadré 5.1). Une version stochastique du modèle est décrite en termes de **processus de Markov** (voir encadré 2.2, page 38). L'espace des états de ce processus est défini par les nombres d'individus dans chacun des 6 compartiments S, E, I, X, Z et R (figure 5.3). Les changements dans l'espace des états sont caractérisés par les événements de transitions répertoriés dans le tableau 5.1. Chaque événement de transition se produit avec une probabilité définie à partir des taux du tableau 5.1, voir encadré 2.2 page 38. Par exemple, la probabilité correspondant à un événement de naissance s'exprime

$$\mathrm{P}\{1 \text{ naissance dans } (t, t+\Delta t] | S(t) = n\} = \mu n \Delta t + \mathrm{o}(\Delta t)$$

où $\lim_{\Delta t \to 0} \frac{\mathrm{o}(\Delta t)}{t} = 0$.

TAB. 5.1 – **Évènements de transition du modèle stochastique de la figure 5.3.**

Évènements	Taux	Types de transition
Naissance	$\mu(S_i + E_i + I_i + X_i + Z_i + R_i)$	$S_i \to S_i + 1$
Décès	μS_i	$S_i \to S_i - 1$
Décès	μE_i	$E_i \to E_i - 1$
Décès	μI_i	$I_i \to I_i - 1$
Décès	μX_i	$X_i \to X_i - 1$
Décès	μZ_i	$Z_i \to Z_i - 1$
Décès	μR_i	$R_i \to R_i - 1$
Infection	$\beta_1(I_i + \xi \Sigma_{i \neq j} I_j) + \beta_2(Z_i + \xi \Sigma_{i \neq j} Z_j)$	$S_i \to S_i - 1$; $E_i \to E_i + 1$
Contagieux	σE_i	$E_i \to E_i - 1$; $I_i \to I_i + 1$
Guérison totale	$(1 - \varepsilon)\gamma I_i$	$I_i \to I_i - 1$; $R_i \to R_i + 1$
Guérison partielle	$\varepsilon \gamma I_i$	$I_i \to I_i - 1$; $X_i \to X_i + 1$
Zona	λX_i	$X_i \to X_i - 1$; $Z_i \to Z_i + 1$
Guérison zona	θZ_i	$Z_i \to Z_i - 1$; $X_i \to X_i + 1$

Pour les simulations numériques de la dynamique, nous procédons en deux étapes [11]. D'abord nous cherchons la date du prochain événement de transition, quelle que soit sa nature. Ensuite, nous déterminons la nature de cet événement. Puisque, par définition, tous les événements sont **indépendants**, la probabilité qu'un événement se produise, quelle que soit sa nature,

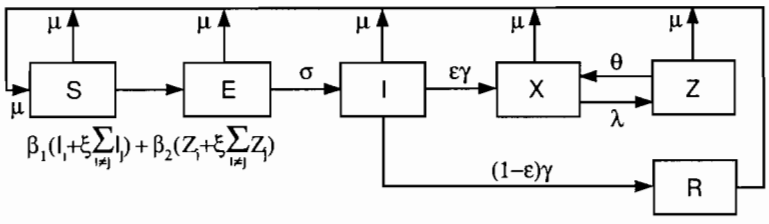


FIG. 5.3 – **Structure du modèle stochastique pour une sous-population de la métapopulation. La taille de la sous-population est constante, les taux de mortalité et de natalité (μ) étant égaux et les migrations d'individus entrant et sortant étant supposées se compenser. Les individus susceptibles S sont contaminés avec une force d'infection proportionnelle au nombre d'infectieux I et de zona Z de la sous-population et, dans une moindre mesure (ξ) des autres sous-populations. Les individus contaminés E deviennent contagieux (I) avec un taux σ. Ces derniers guérissent avec un taux γ, et une proportion ε manifeste et guérit du zona aux taux λ et θ respectivement.**

est simplement la somme des probabilités de tous les événements : $r = \sum_i r_i$. Puisque les événements futurs sont indépendants des événements passés, la durée qui sépare deux événements suit une **distribution exponentielle** de paramètre r. Ainsi, en pratique, pour connaître la date du prochain événement, il suffit de tirer un nombre aléatoire dans une distribution exponentielle de paramètre r. Ensuite, la détermination de la nature du prochain événement se fait par tirage dans une **loi multinomiale** de paramètres r_1/r, r_2/r, *etc...* Ce processus est réitéré pour la durée souhaitée de la dynamique à simuler.

Les simulations ont été réalisées avec les valeurs de paramètres suivantes : $\beta_1 = 1.82 \times 10^{-5}$ an^{-1}ind.$^{-1}$, $\beta_2 = 1.25 \times 10^{-6}$ an^{-1}ind.$^{-1}$, $\gamma = 1/14$ jour^{-1}, $\sigma = 1/5$ jour^{-1}, $\mu = 1/75$ an^{-1}, $\theta = 1/60$ an^{-1}, $\lambda = 1/30$ jour^{-1}. La métapopulation est formée de 30 sous-populations de 60 000 individus. Ces valeurs de paramètres donnent un R_0 dans chaque sous-population légèrement supérieure à 1 (variant de 1.14 à 1.92 selon la prévalence du zona), assurant ainsi l'importance des phénomènes d'extinction/recolonisation. Le modèle présenté a été utilisé pour explorer l'influence du zona sur la dynamique spatiale de la varicelle. Pour se faire, le modèle a été simulé pour des valeurs de ε de 0 à 1 par pas de 0.05 et, à chaque simulation, les décalages de phase entre tous les couples de sous-populations ont été calculés, soit $30 \times 30 - 30 = 870$ valeurs. De la même façon, nous avons utilisé ce modèle stochastique pour explorer l'influence du zona sur la forme de la courbe de CCS.

5.3.3 Résultats

Le graphique du nombre moyen annuel d'extinctions en fonction de la taille de la sous-population ne permet pas de détecter de CCS (figure 5.4). En effet, pour les 60 communes considérées, le nombre moyen annuel d'extinctions est supérieur à 1. De plus, les points ne sont pas répartis uniformément le long de l'axe des abscisses : on trouve beaucoup de points pour les communes de petite taille, où la variance est particulièrement forte. Ceci nous empêche d'estimer une courbe de tendance.

Les calculs de différences de phases ne mettent en évidence aucune hiérarchie de type source/puits significative, quelle que soit la valeur du seuil entre petites et grandes communes choisie (voir figure 5.5 pour les seuils à 10 000 individus (A) et 100 000 individus (B)).

La figure 5.6 représente l'influence de la prévalence du zona sur la forme de la courbe de CCS. On observe que le zona a tendance à diminuer la CCS et ceci est sensible, même pour des prévalences de zona assez faible (15 à 30%).

La figure 5.7 illustre l'influence de la proportion de cas de zonas sur la

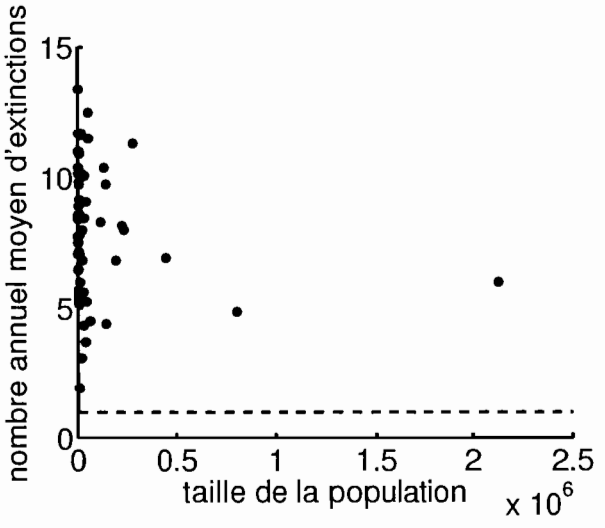


FIG. 5.4 – **Détermination de la CCS pour la varicelle en France. Au lieu de représenter la durée annuelle moyenne sans cas (comme sur la figure 5.1), nous avons choisi ici de représenter le nombre annuel moyen d'extinctions, une extinction étant définie comme une période sans cas d'une durée supérieure au temps de génération (*i.e.* 3 semaines pour la varicelle). La ligne horizontale en pointillés représente une extinction.**

distribution des décalages de phases, tels qu'estimés à partir des simulations sur le modèle stochastique. Dans le cas d'une parfaite synchronie entre toutes les sous-populations, on s'attend à ce que tous les décalages de phases soient égaux à 0. Dans le cas d'une totale asynchronie entre les dynamiques de toutes les sous-populations, on s'attend au contraire à observer une distribution quasi-uniforme, *i.e.* une équiprobabilité de chaque valeur de décalage de phase. La figure 5.7 montre une diminution de la synchronie entre les dynamiques de chaque sous-population lorsque la proportion de zonas augmente.

5.3.4 Discussion

Contrairement à la rougeole en Grande Bretagne, aucune CCS n'a pu être détectée pour la varicelle en France. Nous voyons à cela deux causes principales. La première, d'ordre biologique, est due à la contagiosité de la varicelle qui est approximativement quatre fois plus faible que celle de la rougeole. Dans un tel contexte, on s'attend donc *a priori* à observer une CCS plus élevée pour la varicelle que pour la rougeole (même si au contraire le zona a tendance à diminuer la CCS). Ceci nous mène à la deuxième cause, d'ordre statistique. Cette dernière est liée au fait que les distributions des

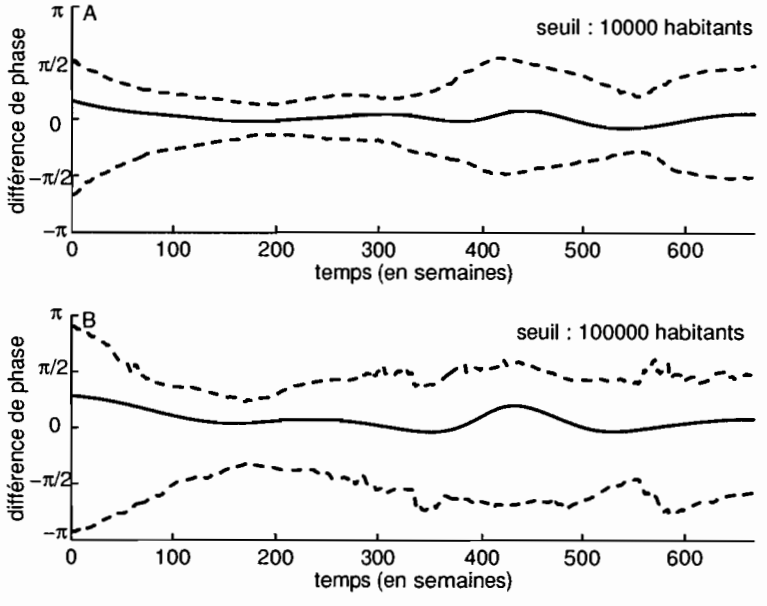


FIG. 5.5 – **Différence de phase entre grandes communes et petites commues. La taille de population seuil entre grandes et petites communes est de 10 000 habitants pour A et 100 000 habitants pour B. Les traits pleins représentent les différences de phases observées sur les données et les traits en pointillés représentent l'intervalle de confiance à 95%, tel qu'attendu d'après la distribution générée par simulations de Monte Carlo.**

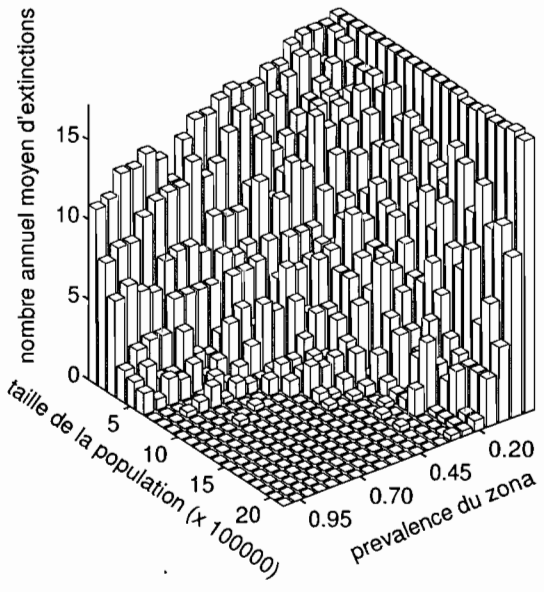


FIG. 5.6 – **Influence de la prévalence du zona sur la CCS de la varicelle. Chaque barre est calculée sur une série temporelle simulée de 12 années.**

tailles de communes en France et en Grande Bretagne sont très différentes. Notamment, on compte seulement 9 villes françaises de plus de 200 000 habitants alors que le Royaume-Uni en dénombre plus de 70. Ceci rend donc plus difficile la détection de grandes CCS en France qu'au Royaume-Uni.

L'analyse des décalages de phases sur les données françaises de varicelle ne suggère aucune hiérarchisation de type source/puits, comme observée pour la rougeole en Grande-Bretagne. Une différence majeure entre la rougeole et la varicelle est la relation de cette dernière avec une seconde maladie, le zona. Parmi les cas guéris de varicelle, une certaine proportion (entre 15 et 30% selon les estimations) développe un zona plusieurs années plus tard. De plus, un susceptible au contact d'une personne présentant un zona peut contracter la varicelle. Par rapport à ce schéma il est donc possible que la relation avec le zona brouille la dynamique spatiale de la varicelle, le zona jouant en quelque sorte le rôle de « réservoir local » à virus. Afin de tester cette hypothèse, nous avons développé un modèle stochastique couplant épidémiologie de la varicelle et du zona dans une structure métapopulationnelle en îles. Nos résultats de simulations indiquent qu'une augmentation de la prévalence de zona se traduit effectivement par une diminution de la synchronie entre les dynamiques de chaque sous-population.

En termes de stratégie de vaccination, ces résultats sont d'importance. Il

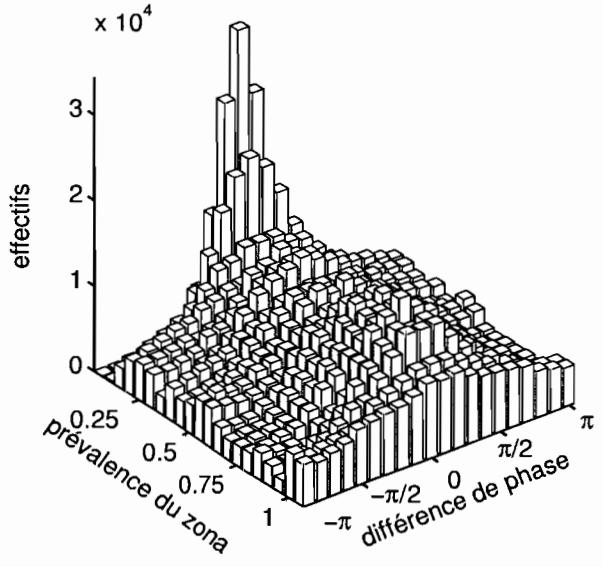


FIG. 5.7 – **Distribution des différences de phases entre les différentes sous-populations et ce pour chaque semaine d'une série temporelle simulée de 12 années. Les prévalences de zonas rapportées dans la littérature varient entre 0.15 et 0.30.**

n'existe actuellement aucune politique vaccinale contre la varicelle en France et un projet est en cours d'étude [44]. Le fait que la varicelle soit nettement moins contagieuse que la rougeole pourrait laisser croire, *a priori*, que la vaccination pourrait facilement conduire à une éradication globale de la varicelle. En réalité, il pourrait en être tout autrement, du fait justement de la relation entre varicelle et zona. Envisager une politique vaccinale contre la varicelle ne peut donc se faire sans se poser un certain nombre de questions comme : Dans quelle mesure le zona rend l'éradication globale de la varicelle plus difficile ? Est-ce que les individus vaccinés, même s'ils ne contractent pas la varicelle, peuvent développer le zona ? Si le zona est susceptible de conduire à une vaccination imparfaite, il devient alors important de se poser la question de l'effet de la vaccination sur le nombre de cas graves. Rappelons en effet que la varicelle, bénigne chez l'enfant, peut entraîner des complications graves chez l'adulte. Un des effets d'une vaccination imparfaite est une augmentation de l'âge moyen à l'infection [11].. Il apparaît alors important de pouvoir prédire quel peut être l'effet d'une vaccination imparfaite sur le nombre absolu de cas graves.

Chapitre 6

Vaccination de masse[1]

La vaccination de masse est la stratégie vaccinale la plus ancienne et encore largement la plus utilisée de part le monde [11]. Les premières applications remontent au début des années quarante, avec la mise en place de politiques vaccinales contre les maladies infantiles majeures dans la plupart des pays européens et nord-américains. En France, l'application du vaccin Rougeole-Oreillons-Rubéole (ROR) est un exemple. Le fondement théorique de la vaccination de masse est basé sur les **propriétés statiques** des maladies [11]. Le caractère dynamique, en particulier la succession des pics épidémiques, n'est pas pris en compte. Ce court chapitre présente une revue de la littérature sur les principes et conséquences de ce type de vaccination. Sa fonction est de permettre de mieux comprendre le chapitre qui suit, dédié à la vaccination par pulsations. Nous commençons par présenter le formalisme théorique associé aux modèles en compartiments de type SEIR (à la base de tous les modèles d'infections microparasitaires) et en déduisons les propriétés statiques des maladies en termes de taux de reproduction et d'âge moyen à l'infection. Le deuxième paragraphe dérive le théorème fondamental du seuil épidémique sur lequel repose le principe de la vaccination de masse. Enfin, nous étudions les conséquences de la vaccination de masse en termes d'âge moyen à l'infection et de dynamique de maladie.

6.1 Aspects statiques des maladies

La dynamique des maladies **microparasitaires** (virus, bactéries, protozoaires) est généralement modélisée en considérant l'état clinique de l'hôte.

1. Les éléments de ce chapitre sont issus d'une revue rédigée pour le chapitre 22 de *Encyclopedia of Infectious Diseases - Modern Methods* (Tibayrenc M. Ed.) John Wiley & Sons, Chichester, USA (Choisy M. & Guégan J.F. 2005) présenté en annexe D.

Dans un tel contexte, une façon simple de représenter la population totale est d'utiliser un **modèle en compartiments** de type SEIR [11] comme représenté sur la figure 6.1.

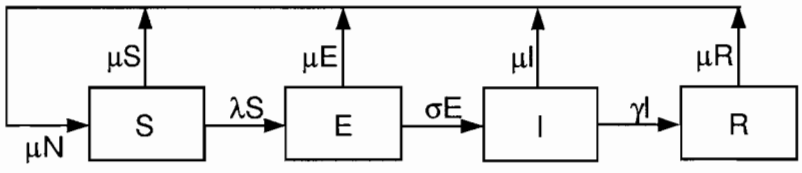


FIG. 6.1 – **Un modèle SEIR. Les rectangles représentent les compartiments de la population hôte et les flèches symbolisent les flux d'individus d'un compartiment à l'autre. Ici, les individus naissent susceptibles (S) au taux de natalité μ et se font contaminer (E) avec la force d'infection λ puis deviennent contagieux (I) au taux σ avant de devenir guéris (R) au taux γ. Les individus de chaque compartiment meurent tous au même taux de mortalité, égal au taux de natalité μ. La taille de la population totale N est donc constante.**

Pour les maladies à **transmission directe** (comme la plupart des maladies infantiles : rougeole, coqueluche, oreillons, varicelle, *etc...*), on peut faire l'hypothèse que la force d'infection λ (définie comme la probabilité pour un susceptible de contracter la maladie) est proportionnelle à la fréquence de contagieux dans la population [46] : $\lambda = \beta \mathrm{I}/\mathrm{N}$, où β est le nombre moyen de contacts par individu par unité de temps, I est le nombre de contagieux et N est la taille de la population totale.

S'intéresser aux aspects statiques d'une maladie revient à faire l'hypothèse que le nombre d'individus dans chaque état clinique (susceptibles, latents, contagieux, et guéris) **ne dépend pas du temps**. Ceci suppose en particulier que les naissances compensent exactement les décès (taille de la population totale constante) et qu'aucune mortalité n'est associée à la maladie [11]. On considère également que la dynamique de la maladie est dans un état d'**équilibre endémique**.

Pour les populations de **grande taille**, les variations par rapport à la tendance moyenne sont négligeables et la structure déterministe du modèle SEIR peut être décrite mathématiquement avec le formalisme des **équations différentielles**. Ainsi, les variations des nombres d'individus dans chacun des

quatre compartiments de la figure 6.1 peuvent s'écrire :

$$\begin{aligned} dS/dt &= \mu(N - S) - \beta SI/N \\ dE/dt &= \beta SI/N - (\sigma + \mu)E \\ dI/dt &= \sigma E - (\gamma + \mu)I \\ dR/dt &= \gamma I - \mu R \end{aligned}$$

Dans les deux sous-paragraphes suivants, nous utilisons ces équations différentielles pour en déduire une expression des taux de reproduction et de l'âge moyen à l'infection.

6.1.1 Les taux de reproduction R et R_0

A partir des équations différentielles du modèle SEIR, on peut exprimer le taux de reproduction R, défini comme le nombre moyen de nouveaux cas produits par chaque infectieux :

$$R = \frac{S\beta\sigma}{N(\sigma + \mu)(\gamma + \mu)}$$

Lorsque $R < 1$, le nombre moyen de malades dans la population diminue et, lorsque $R > 1$, le nombre moyen de malades augmente. Par définition, la valeur initiale (*i.e.* lorsque toute la population est susceptible) de R est R_0 (taux de reproduction basal ou nombre moyen de nouveaux cas issus de l'introduction d'un individu infectieux dans une population entièrement susceptible). En égalisant donc S à N dans l'équation ci-dessus, on obtient une expression de R_0 :

$$R_0 = \frac{\beta\sigma}{(\sigma + \mu)(\gamma + \mu)}$$

Lorsque $R_0 < 1$, la dynamique converge vers un état d'équilibre sans maladie, tandis que lorsque $R_0 \geq 1$, la dynamique converge vers un état d'équilibre **endémique**. La combinaison des deux équations ci-dessus nous fournit une expression de R en fonction de R_0 :

$$R = \frac{S}{N} R_0$$

Par définition, à l'équilibre endémique, $R = 1$ (*i.e.* chaque cas produit en moyenne un nouveau cas), soit

$$\frac{S^*}{N} R_0 = 1$$

où S* est le nombre de susceptibles dans la population à l'équilibre endémique. Cette dernière relation permet de déterminer empiriquement les taux de reproduction basaux à partir d'une simple sérologie déterminant la proportion S*/N de susceptibles dans la population.

6.1.2 L'âge moyen à l'infection A

L'âge moyen à l'infection se définit naturellement comme

$$A \equiv \int_0^\infty a \frac{\lambda S(a)}{\int_0^\infty \lambda S(a) da} da$$

qui est simplement la somme de tous les âge a, pondérés par la proportion de nouveaux contaminés d'âge a. Le calcul de cette intégrale avec un taux de mortalité μ indépendant de l'âge conduit à la relation intuitive $A = 1/(\lambda + \mu)$. Ceci signifie que, plus la force d'infection est élevée, plus l'âge moyen à l'infection est bas.

6.2 Principe de la vaccination de masse

Si p est la **couverture vaccinale** (*i.e.* la proportion d'individus protégés par vaccination), alors la proportion de susceptibles dans la population ne peut être qu'inférieure à $(1-p)$:

$$\frac{S}{N} \leq 1 - p$$

D'après les équations du paragraphe 6.1.1, $R = R_0 S/N$. Donc

$$R \leq (1-p)R_0$$

La condition nécessaire à l'éradication d'une maladie est un taux de reproduction inférieur à 1 (voir paragraphe 6.1.1). Une condition alors suffisante est que $R_0(1-p)$ soit inférieur à 1 :

$$R \leq (1-p)R_0 < 1$$

En réarrangeant la deuxième condition, on obtient :

$$p > 1 - \frac{1}{R_0}$$

On peut donc en déduire que la couverture vaccinale minimale p_c à appliquer pour éradiquer une maladie est d'autant plus forte que le taux de reproduction basal de la maladie est fort :

$$p_c = 1 - \frac{1}{R_0}$$

Une interprétation importante de ce résultat est qu'il n'est pas nécessaire de vacciner toute la population pour espérer éradiquer une maladie. Ce résultat fondamental de l'épidémiologie théorique est connu sous le nom de **théorème du seuil** (établit en 1927 par KERMACK et MCKENDRICK [105]) et la propriété émergente sous le nom d'**immunité de groupe**. Toutefois, les R_0 de la plupart des maladies imposent une couverture vaccinale extrêmement élevée, souvent irréalisable en pratique (voir tableau 6.1).

TAB. 6.1 – Quelques valeurs de paramètres épidémiques de maladies issus de la littérature. Tous les paramètres sont estimés dans les pays développés sauf contre-indication.

Maladies	A^a	R_0^b	p_c^c
Rougeole	4-6^d, 1-3^e	16-17	90-95%
Oreillons	6-7	7-8^d, 11-14^e	85-90%
Coqueluche	4-5	16-17	90-95%
Rubéole	9-10^d, 2-3^e	6-7^d, 15-16^e	82-87%
Varicelle	6-8	7-8^d, 10-12^e	85-90%
Variole	—	—	70-80%
Malaria	—	—	99%

^{a}Âge moyen à l'infection. Données issues de [10].
^{b}Taux de reproduction basal. Données issues de [7, 10, 148].
^{c}Couverture vaccinale minimale. Données issues de [11].
^{d}Pays développés.
^{e}Pays en voie de développement.

Les valeurs présentées dans le tableau 6.1 nous permettent de comprendre pourquoi la variole est actuellement la seule maladie que nous ayons éradiquée avec succès. De plus, les valeurs présentées dans le tableau 6.1 ont été estimées en considérant une **efficacité de vaccination** parfaite, ce qui n'existe pas. Si nous tenons explicitement compte des efficacités de vaccination réelles, les valeurs de p_c sont encore plus fortes.

Pour être efficace, le vaccin doit être appliqué avant l'âge moyen à l'infection. En France, le vaccin ROR est appliqué de façon aussi systématique que possible aux enfants, avant l'âge de 2 ans.

6.3 Conséquences de la vaccination de masse

6.3.1 Conséquences statiques

L'effet direct de la vaccination est une diminution du nombre de malades. Toutefois, cette diminution du nombre de malades conduit également à une diminution de la force d'infection. C'est l'effet indirect de la vaccination. Les résultats du paragraphe 6.1.2 nous permettent alors de prédire qu'une conséquence statique de la vaccination est une **augmentation de l'âge moyen à l'infection**. Ceci peut paraître anodin, mais c'est en réalité un effet très important dans le cas des maladies dont la gravité augmente avec l'âge, comme la varicelle (voir chapitre 5) ou la rubéole. Comme discuté à la fin du chapitre 5, il convient donc, avant de mettre en place une politique vaccinale contre de telles maladies, d'explorer, à l'aide de **modèles**, l'effet de la vaccination sur le nombre absolu de cas graves. Un exemple classique d'une telle exploration est l'étude d'ANDERSON et MAY sur la rubéole [9]. La rubéole est une maladie infantile virale[2] conférant une immunité permanente. La maladie est bénigne chez l'enfant. Toutefois, une contamination pendant la grossesse est grave en raison d'un risque élevé de malformations fœtales. ANDERSON et MAY ont utilisé un modèle SEIR pour explorer l'influence de la couverture vaccinale sur le nombre de malformations fœtales (figure 6.2). Il ressort de leur modèle que, si une politique vaccinale doit être appliquée contre la rubéole, il faut que la couverture vaccinale soit supérieure à 50% sinon, le nombre de malformations fœtales risque d'augmenter.

6.3.2 Conséquences dynamiques

Comme souligné dans le chapitre précédent, l'étude de la dynamique spatio-temporelle d'une maladie est primordiale pour la mise en place d'une politique vaccinale efficace. Nous avons vu en particulier que la dynamique spatiale de la rougeole en Angleterre est très hiérarchisée selon un modèle de diffusion source-puits, alors que le zona semble brouiller toute trace de cohérence spatiale dans la dynamique de la varicelle en France (voir chapitre 5).

Dans une métapopulation, la synchronie entre les dynamiques des différentes sous-populations est une donnée clef pour la compréhension de la **persistance globale**. En effet, une éradication globale sera beaucoup plus facile dans un système où toutes les dynamiques sont synchrones que dans un système où les dynamiques sont asynchrones. Dans la première situation,

2. Virus à ARN de la famille des Togaviridae, genre Rubivirus. C'est un virus enveloppé de petite taille (60 à 70 nm).

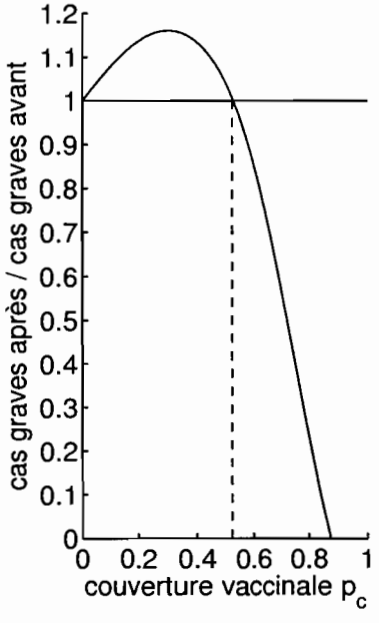


FIG. 6.2 – **Prédictions théoriques de l'influence de la vaccination contre la rubéole sur le nombre de malformations fœtales. Le graphique représente le rapport entre les nombres de cas graves après et avant vaccination, en fonction de la couverture vaccinale. Modèle issu de [9].**

le nombre de cas est au plus bas dans toutes les sous-populations en même temps. En conséquence, une éradication dans une localité pourra être difficilement suivie par une colonisation venant des localités avoisinantes où le nombre de cas est également bas. Au contraire, dans la deuxième situation, une extinction locale pourra être suivie d'une recolonisation venant d'une localité voisine où le nombre de cas est encore élevé.

La synchronie entre les dynamiques de maladies dans différentes sous-populations est assurée par la migration d'individus contagieux. L'effet direct de la vaccination est, nous l'avons vu, une diminution du nombre de malades. On peut donc s'attendre à ce que la force de couplage entre les dynamiques des différentes sous-populations soit réduite par la vaccination. Cet attendu théorique a été effectivement vérifié par ROHANI, EARN et GRENFELL sur les dynamiques de rougeole de soixante villes anglaises [169] (voir figure 6.3).

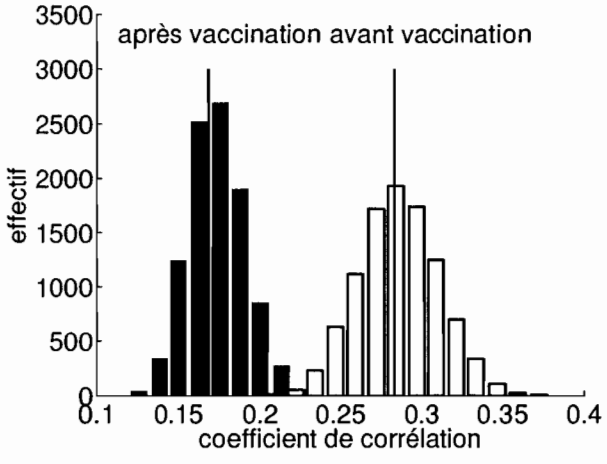


FIG. 6.3 – **Effet de la vaccination sur la synchronie des dynamiques de rougeole. Les données sont issues de 60 villes anglaises. La période de pré-vaccination comprend les années 1944–1968 et la période vaccinale comprend les années 1968–1988. Les deux lignes verticales représentent les moyennes des $60 \times 60 - 60 = 3540$ coefficients de corrélation avant et après vaccination. Les histogrammes représentent les distributions établies par simulations de Monte Carlo avec 10 000 répétitions. Le coefficient de corrélation moyen est de 0.29 (intervalle de confiance à 95% : [0.24, 0.34]) avant vaccination, et de 0.17 (intervalle de confiance à 95% : [0.14, 0.21]) après vaccination. Figure reprise de [169].**

On observe que la vaccination tend à désynchroniser les dynamiques. Or, plus les dynamiques locales sont asynchrones, plus une éradication globale devient difficile (voir discussion ci-dessus). Ces deux effets conjugués peuvent

expliquer pourquoi une éradication globale est très difficile à atteindre en pratique, même avec des couvertures vaccinales élevées.

Chapitre 7

Vaccination par pulsations[1]

La vaccination de masse est une politique relativement efficace, à en juger par la diminution impressionnante du nombre de cas de rougeole dans la plupart des pays européens et nord américains. Toutefois, comme mis en évidence dans le chapitre précédent, c'est une politique extrêmement lourde à mettre en place et à suivre, non seulement en termes financiers mais également en termes logistiques. Par exemple, le vaccin ROR[2] est appliqué en France de façon aussi systématique que possible aux enfants, avant l'âge de deux ans. Ce genre de politique est clairement inabordable pour les pays en voie de développement. En conséquence, des maladies infantiles comme la rougeole, devenues rares en Europe ou en Amérique du Nord, sont encore très fréquentes en Afrique ou en Amérique Latine. Associées à des conditions de vie précaires, ces maladies constituent encore aujourd'hui une cause de mortalité importante dans ces régions du monde. Ce constat a motivé des recherches vers des politiques vaccinales **plus économe**s et moins lourdes à mettre en place. Ainsi, un schéma de vaccination par pulsations a été proposé au début des années quatre-vingt-dix, et appliqué dans un nombre de pays croissant quelques années plus tard [4, 149]. Dans ce chapitre, nous nous intéressons à ce nouveau type de vaccination encore peu connu. Après avoir présenté les principes théoriques de la vaccination par pulsations, nous nous concentrons sur les conséquences dynamiques, avant de conclure sur une discussion et des perspectives de recherche relatives à cette politique vaccinale.

0. Les résultats de ce chapitre ont donné lieu à la rédaction d'un manuscrit (Choisy M., Rohani P. & Guégan J.F.) soumis à *Theoretical Population Biology* et présenté en annexe E.

2. Contre la rougeole, les oreillons, la rubéole.

7.1 Fondements théoriques

La vaccination de masse est basée sur les propriétés statiques des maladies, telles que résumées par des statistiques comme le R_0. Même si l'état d'endémie existe dans la nature, il est plutôt rare pour les maladies infantiles. En effet, les dynamiques de ces dernières sont généralement marquées par des successions, plus ou moins régulières, de pics épidémiques (voir figure 7.1). Considérer que les maladies infantiles sont dans un état d'endémie est donc une approximation très grossière. L'idée de la vaccination par pulsations est d'essayer de prendre en compte explicitement la **dynamique du système hôte-parasite**, *i.e.* les variations autour de la moyenne endémique.

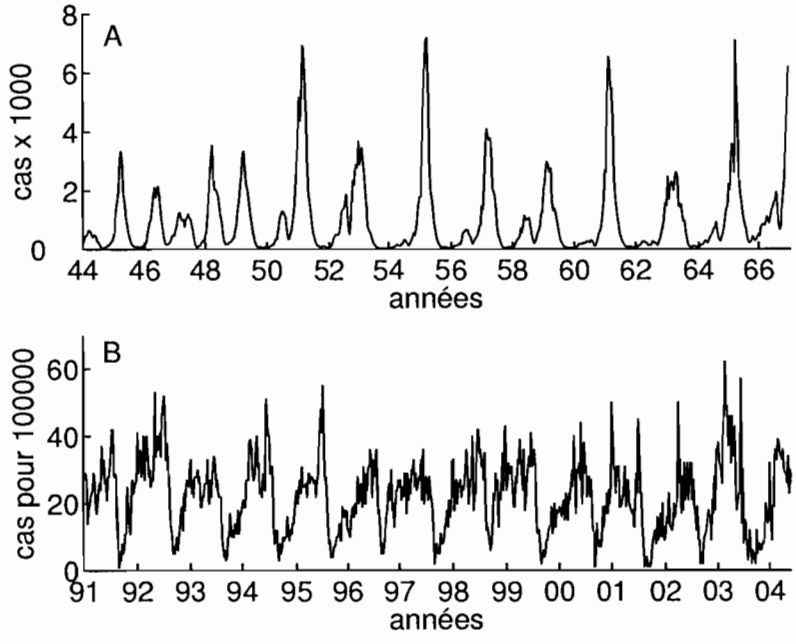


FIG. 7.1 – **Exemples de séries temporelles de maladies infectieuses. A : incidence bi-hebdomadaire de rougeole pour la ville de Londres dans la période pré-vaccinale. Données issues de** `http://www.zoo.cam.ac.uk/zoostaff/grenfell/measles.html`**. B : incidence (rapportée à 100 000 habitants) hebdomadaire de varicelle en France. Données issues de** `http://rhone.b3e.jussieu.fr/senti/php/navigation/accueil`.

Le principe de la vaccination par pulsations repose sur des résultats théoriques concernant les dynamiques de populations en environnements variables [3]. Plusieurs développements théoriques ont été proposés depuis le début des années quatre-vingt-dix, mais tous reposent sur le même schéma de base tel qu'illustré sur la figure 7.2.

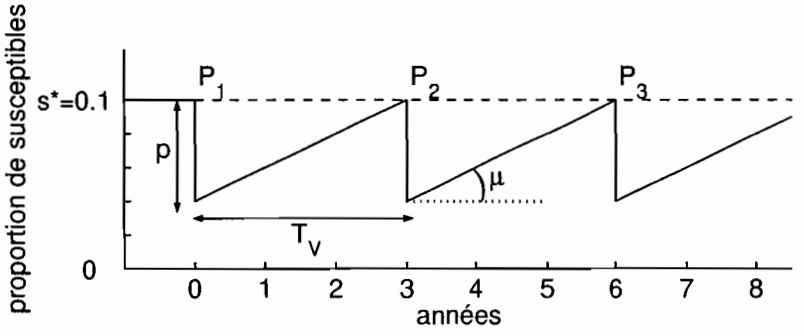


FIG. 7.2 – **Schéma de vaccination par pulsations. Le graphique représente la proportion de susceptibles en fonction du temps. Le but est de déterminer une couverture vaccinale p et une fréquence $1/T_V$ des événements de vaccination P_1, P_2, P_3, *etc...* telle que la proportion de susceptibles soit maintenue en dessous du seuil critique s^* représenté par la ligne en tirets. Les naissances approvisionnent le stock de susceptibles au taux constant μ.**

Comme la vaccination de masse, la vaccination par pulsation est basée sur le **théorème du seuil** présenté au chapitre 6, page 118. La différence majeure entre le principe de la vaccination de masse et celui de la vaccination par pulsations est que le premier ne considère que les propriétés statiques du système hôte-parasite alors que le deuxième tient compte explicitement de la **dynamique de susceptibles**, à travers le **taux de naissance**. Le théorème du seuil met en évidence qu'une épidémie ne peut démarrer que si la proportion de susceptibles $s = \mathrm{S/N}$ dans la population est au-dessus d'un certain seuil $s^* = 1/R_0$. Au-dessous, nous observons un phénomène d'immunité de groupe empêchant tout démarrage d'épidémie. Approximativement, la dynamique de susceptibles est contrôlée par deux facteurs : (i) le taux de natalité qui augmente la proportion de susceptibles s de façon continue et régulière, et (ii) les événements de vaccination P_n qui diminuent la proportion de susceptibles s de façon discontinue et instantanée (voir figure 7.2). Dans ce contexte, le principe de la vaccination par pulsations consiste à appliquer des événements ponctuels de vaccination (avec une couverture p) suffisamment fréquemment (à une fréquence $1/T_V$) pour maintenir la proportion de susceptibles au-dessous du seuil critique s^*.

Les travaux théoriques sur la vaccination par pulsations se sont intéressés essentiellement à la détermination analytique d'un couple (T_V, p) en fonction d'un taux de natalité μ [4, 186, 47]. En fonction du réalisme des hypothèses faites sur les modèles utilisés, nous obtenons des estimations plus ou moins efficaces et économes. Par exemple, un des premiers résultats établis sur un

modèle SIR a été que la période des événements de vaccination ne doit pas être supérieure à l'**âge moyen à l'infection** : $T_V < A$ (rappelons que A dépend de la couverture vaccinale p, voir chapitre 6, page 120) [4]. Des analyses plus détaillées ont ensuite mis en évidence que pour empêcher une épidémie il suffit en fait que la valeur **moyenne** de la proportion de susceptibles soit inférieure au seuil critique s^*. Toujours dans le cadre d'un modèle SIR, SHULGIN et collaborateurs [186] ont montré que cette condition mène à la relation suivante :

$$T_V < T_V^{max} = \frac{p\gamma}{\beta\mu(1 - p/2 - \gamma/\beta)}$$

où γ et β sont respectivement les taux de guérison et de contact entre individus (voir présentation du modèle SEIR au chapitre 6, page 116). Ainsi, la valeur de T_V peut être cette fois-ci plus élevée que l'âge moyen à l'infection. D'ONOFRIO [47] a ensuite montré que cette relation restait une bonne approximation pour les **modèles de type SEIR**, plus couramment utilisés pour les maladies infantiles. De plus, lorsque le taux de contact entre individus β n'est pas constant en fonction du temps, la relation ci-dessus reste valide si β est remplacé par sa **moyenne temporelle** [186, 47].

Ces résultats théoriques prédisent que, lorsque $T_V < T_V^{max}$, la maladie devrait disparaître de la population assez rapidement. Nous nous intéressons dans ce chapitre aux cas de vaccinations par pulsations imparfaites *i.e.* où l'incidence de la maladie décroît sans toutefois s'annuler complètement. Dans un premier temps nous étudions l'effet de la vaccination par pulsations sur la dynamique spatiale de la maladie (afin de pouvoir comparer avec l'effet de la vaccination de masse vu au chapitre précédent). Dans un deuxième temps, nous nous focalisons spécifiquement sur les phénomènes de résonance propres aux forçages périodiques. Le dernier paragraphe est une discussion sur les recherches futures à mener sur la vaccination par pulsations.

7.2 Dynamique spatiale

Ce qui distingue fondamentalement la vaccination de masse de la vaccination par pulsations est le mode d'application. Alors que la première est appliquée de façon continue, la seconde est au contraire appliquée **périodiquement**. Dans le cas d'une vaccination par pulsations imparfaite, on peut donc s'attendre à ce que les événements de vaccination constituent un forçage pour la dynamique de la maladie, c'est à dire que la périodicité de la vaccination influence celle de la dynamique de la maladie. EARN, ROHANI et GRENFELL [50] ont exploré cette possibilité à partir de l'étude d'un simple modèle SEIR et mettent en évidence qu'une vaccination par pulsations im-

parfaite peut changer la périodicité de la dynamique (voir figure 7.3).

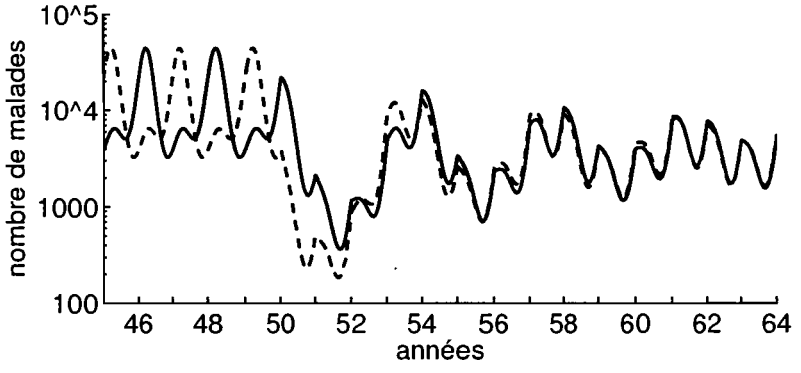


FIG. 7.3 – **Effet de la vaccination par pulsations sur la dynamique spatiale de la maladie. Les simulations sont initiées de telle sorte à ce que les dynamiques dans deux populations indépendantes soient en opposition de phase. A partir de l'année 50, une vaccination par pulsations est appliquée annuellement avec une couverture de 20%. Appliquer la même forme de vaccination sur les deux populations indépendantes a pour effet de les synchroniser.**

Plus important, des événements de vaccination synchrones dans différentes localités ont pour effet de synchroniser les dynamiques (voir figure 7.3). Cette observation est un cas d'**effet MORAN** où une même contrainte environnementale (ici la vaccination) produit des effets similaires dans des localités indépendantes [133, 73, 27].

Les effets de la vaccination par pulsations sur la dynamique spatiale de la maladie sont donc opposés à ceux de la vaccination de masse. Ceci est d'importance pratique puisque la synchronie entre sous-populations augmente la probabilité d'extinction globale [102] (voir chapitre 6, page 120). Observons toutefois que ces résultats sont issus d'un modèle théorique. Ils n'ont pas encore été vérifiés sur des données réelles. En effet, l'application de la vaccination par pulsations sur le terrain est encore trop récente et nous ne disposons pas encore de séries temporelles suffisamment longues pour tester ce type de prédictions.

7.3 Résonance

Nous venons de voir que la vaccination par pulsations, de part son caractère périodique pouvait agir comme un **forçage** et influencer la dynamique de la maladie. Or, dans bien des cas, la dynamique de la maladie est elle-même

naturellement périodique [8, 89, 169, 49, 71]. Ainsi, la varicelle est caractérisée par des épidémies annuelles, la rougeole par des épidémies annuelles ou bisannuelles, et la coqueluche par des épidémies tri- ou quadri-annuelles (voir par exemple les séries temporelles de la figure 7.1). Les modèles théoriques utilisés pour décrire la dynamique des maladies semblent indiquer également que les dynamiques de nombreuses maladies infantiles sont intrinsèquement oscillatoires.

La théorie des systèmes oscillatoires prédit qu'une dynamique oscillatoire soumise à un forçage lui-même périodique peut exhiber des phénomènes de résonance [95]. Dans ce paragraphe, nous explorons théoriquement la possibilité de phénomènes de résonance associés à la vaccination par pulsations.

7.3.1 Qu'est-ce que la résonance ?

On parle de résonance lorsque l'amplitude d'un système dynamique oscillant dépend de la période d'un forçage périodique agissant sur le système [95]. A la fréquence dite de résonance du forçage, l'amplitude du système dynamique est à son maximum que l'on appelle **pic de résonance**. Il existe plusieurs formes de résonance, stochastique ou déterministe. Pour les formes déterministes, le degré de non-linéarité du système dynamique est un important déterminant du type de résonance. On distingue en particulier le cas général de résonance paramétrique du cas plus particulier de résonance harmonique.

Résonance paramétrique

Le cas le plus général et le plus complexe est le phénomène de résonance non-linéaire (ou paramétrique) qui se produit sur des systèmes dynamiques fortement **non-linéaires**. Dans ce cas, la théorie prédit plusieurs fréquences de résonance dont les valeurs sont des fractions entières de la **période propre** du système [3]. Une deuxième caractéristique de la résonance paramétrique est la dépendance entre l'amplitude et la fréquence de résonance. Cette dépendance croît avec le degré de non-linéarité du système. Enfin, une troisième caractéristique est le phénomène de seuil sur les valeurs des paramètres : un pic de résonance donné n'apparaît que lorsque les paramètres du système dynamique sont au-dessus d'une valeur seuil. Ceci est en fait lié au degré de non-linéarité du système qui dépend des valeurs des paramètres. Plus la non-linéarité du système est forte, plus le nombre de pics de résonance est

3. On appelle période propre d'un système la période de ses oscillations en absence de forçage.

élevé, et inversement. Un cas limite est lorsque le système est parfaitement linéaire. On a alors un seul et unique pic de résonance, et on parle de résonance linéraire, ou harmonique.

Résonance harmonique

Les systèmes dynamiques parfaitement linéaires (dit harmoniques) présentent des phénomènes de résonance linéaire ou harmonique. Ce phénomène peut être considéré comme un cas limite du phénomène général de résonance paramétrique. Par rapport à ce dernier, la résonance harmonique est caractérisée par une seule fréquence de résonance, égale à la période propre du système. De plus, la fréquence de résonance est parfaitement indépendante de l'amplitude.

La résonance dans la nature

Il apparaît que la résonance est un phénomène très général dans la nature. Il est notamment à la base des technologies d'imagerie par résonance magnétique, de filtres et d'amplificateurs optiques, électriques ou électromagnétiques. De plus, nous avons tous en tête l'image du verre à pied que l'on fait éclater en sifflant. C'est un cas de résonance où le système oscillant est ici le verre à pied et le forçage le son. Le son fait vibrer le verre à pied. Lorsque la fréquence sonore est proche de la fréquence de résonance du verre à pied, l'amplitude de vibration de ce dernier augmente. Comme l'élasticité du verre est réduite, ce dernier éclate quand son amplitude de vibration devient trop élevée.

Malgré la généralité du phénomène, la résonance n'a été que très peu étudiée en dynamique des populations, et encore moins en dynamiques épidémiques. C'est ce que nous nous proposons d'explorer dans le présent chapitre.

7.3.2 Méthodes

Les effets de résonance associés aux dynamiques de maladies ont été recherchés théoriquement par l'exploration numérique d'un modèle mathématique.

Le modèle

Nous considérons une **maladie infantile** ayant les traits d'histoire de vie de la rougeole, modélisée par un simple modèle SEIR, comme présenté au

chapitre précédent :

$$\begin{aligned} dS/dt &= \mu(N - S) - \beta SI/N \\ dE/dt &= \beta SI/N - (\sigma + \mu)E \\ dI/dt &= \sigma E - (\gamma + \mu)I \\ dR/dt &= \gamma I - \mu R \end{aligned}$$

Au chapitre précédent, nous nous sommes intéressés aux propriétés statiques de ce modèle. Ici, nous explorons son comportement dynamique. Rappelons que lorsque $R_0 < 1$ la dynamique converge vers un état d'équilibre libre de maladie. Lorsque $R_0 \geq 1$ la dynamique présente des oscillations amorties qui convergent vers un état d'équilibre endémique. La lenteur de la convergence traduit une **faible stabilité** de l'équilibre endémique. En conséquence, de faibles perturbations suffiront à entretenir les oscillations, produisant ainsi des successions d'épidémies, comme observées dans la nature (voir figure 7.1).

Plusieurs sources de perturbations ont été proposées dans la littérature, la plus simple étant une **variation sur le taux de contact** β, censée représenter l'alternance des périodes scolaires et des périodes de vacances [119, 60, 61, 179, 11]. En effet, pour les maladies infantiles, l'essentiel des transmissions se produit dans les cours d'écoles. On peut donc imaginer que les taux de contacts entre enfants sont plus forts pendant les périodes d'écoles que pendant les vacances scolaires. Plusieurs formes de variation différentes de β ont été envisagées, la plus simple étant une sinusoïde :

$$\beta(t) = \beta_0 \left(1 + \beta_1 \cos\left(\frac{2\pi}{T_S} t\right)\right), \qquad 0 \leq \beta_1 < 1$$

où β_0 est la moyenne temporelle du taux de contact, β_1 l'amplitude de ses oscillations et T_S la période des oscillations (*i.e.* $T_S = 1$ an, représentant une année scolaire). Il est à noter toutefois que le comportement dynamique du système est très **robuste** par rapport à la forme exacte de la variation de β [49]. Nous nous limitons donc ici, pour cette étude théorique, à la forme la plus simple qui est la sinusoïde.

Ajoutons à présent la vaccination par pulsations : une proportion p des susceptibles est vaccinée tous les T_V ans et passe donc directement dans le compartiment R :

$$S(kT_v^+) = (1 - p) \cdot S(kT_v^-), \qquad k \in \mathbb{N}$$

où T_V^- et T_V^+ représentent respectivement les instants immédiatement avant et immédiatement après l'événement de vaccination.

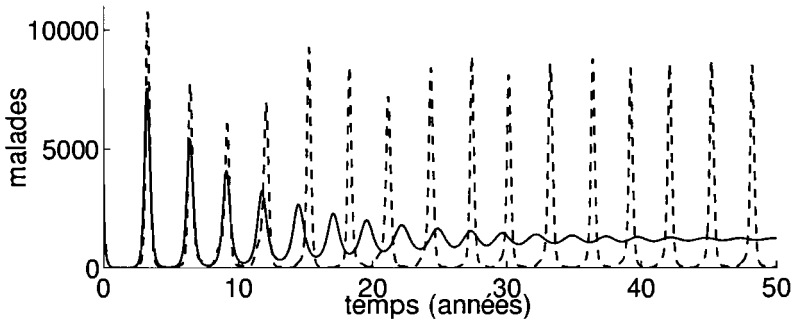


FIG. 7.4 – **Oscillations amorties et entretenues par forçage saisonnier. La courbe en trait plein représente les oscillations amorties d'un système SEIR avec taux de contact constant ($\beta_1 = 0$). La courbe en tirets représente les oscillations entretenues par des variations saisonnières du taux de contact ($\beta_1 = 0.1$). Les autres valeurs de paramètres sont $\mu = 1/70$ ans, $1/\sigma = 7.5$ jours, $1/\gamma = 6.5$ jours, $\mathbf{N} = 5 \times 10^6$, $\beta_0 = 2 \times 10^{-3}$/an/individu. Les conditions initiales sont $(\mathbf{S}_0, \mathbf{E}_0, \mathbf{I}_0) = (2 \times 10^5, 5000, 500)$.**

Étude de la dynamique

Le comportement dynamique du modèle est exploré par l'analyse de **diagrammes de bifurcations** (voir encadré 7.1). Nous avons un système dynamique **non-linéaire** (voir encadré 7.2), intrinsèquement **oscillant** (voir figure 7.4) et soumis à deux **forçages périodiques**. Le premier, forçage saisonnier, est dû aux variations du taux de contact β. Le deuxième est dû à la vaccination par pulsations. Afin de distinguer les effets de la non-linéarité, du forçage saisonnier et de la vaccination par pulsations sur la complexité de la dynamique, nous avons procédé par étapes.

Dans un premier temps, nous nous sommes intéressés uniquement au **forçage saisonnier**, sans vaccination. L'effet de la non-linéarité sur ce système a été exploré en faisant varier le taux de natalité sur une large gamme de valeurs (des plus réalistes au plus irréalistes), voir encadré 7.2. La résonance a été cherchée en examinant l'influence de T_S sur l'amplitude des oscillations. Les prédictions théoriques ont ensuite été vérifiées sur des dynamiques de rougeole de 60 villes anglaises de la période pré-vaccinale (1944–1966).

Dans un deuxième temps, nous nous sommes intéressés uniquement à la **vaccination par pulsations**, sans forçage saisonnier. Comme précédem-

ENCADRÉ 7.1
DIAGRAMMES DE BIFURCATION

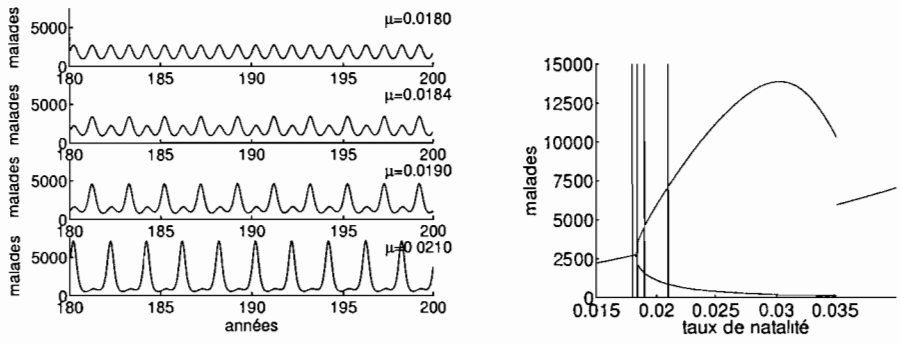
Les diagrammes de bifurcations sont des figures permettant d'appréhender la **complexité** des dynamiques de façon simple. Considérons l'exemple d'un simple modèle SIR avec taux de contact variable :

$$ds/dt = \mu - \beta is - \mu s \qquad s(0) = s_0 \geq 0$$

$$di/dt = \beta is - \gamma i - \mu i \qquad i(0) = i_0 \geq 0$$

$$\beta(t) = \beta_0(1 + \beta_1 \cos(2\pi t)) \qquad 0 \leq \beta_1 \leq 1$$

La résolution de ce modèle avec des valeurs du taux de natalité μ différentes produit des dynamiques **qualitativement** différentes. Sur la figure de gauche la dynamique passe d'annuelle à bisannuelle lorsque μ augmente de 0.018 à 0.021.



Ces changements qualitatifs sont appelés **bifurcations** et le paramètre dont on explore l'influence est appelé **paramètre de contrôle**. Les diagrammes de bifurcation (figure de droite) permettent de visualiser l'effet du paramètre de contrôle (en abscisse) sur la complexité de la dynamique. Imaginons par exemple que nous échantillonnions la dynamique tous les ans. Une dynamique annuelle produira ainsi un seul point sur le diagramme de bifurcation (chaque année la dynamique retrouve la même valeur), alors qu'une dynamique bisannuelle produira deux points (un pour les années paires et un autre pour les années impaires). Les quatre lignes verticales sur le diagramme de bifurcation représentent les valeurs des taux de natalité correspondant aux quatre séries temporelles de la figure de gauche.

ment, l'effet de la non-linéarité sur ce système a été exploré en faisant varier le taux de natalité sur une large gamme de valeurs (des plus réalistes aux plus irréalistes). La résonance a été cherchée en examinant l'influence de T_V sur l'amplitude des oscillations.

Enfin, nous avons considéré les **deux forçages à la fois** : (i) pour des valeurs du taux de natalité irréalistes mais où le système se comporte linéairement, et (ii) des valeurs du taux de natalité réalistes mais où le système se comporte non-linéairement.

7.3.3 Résultats

L'effet du forçage saisonnier associé à divers degrés de non-linéarité illustre parfaitement bien le **continuum** entre résonance harmonique et résonance paramétrique (figure 7.5). Lorsque le taux de natalité est fort, le système présente un unique pic de résonance dont la fréquence, indépendante de l'amplitude, est proche de la période propre du système. Lorsque le taux de natalité diminue, la dépendance entre amplitude et fréquence de résonance augmente et plusieurs autres pics de résonance paramétrique apparaissent pour des valeurs seuils du taux de natalité différentes. La **courbure** de la figure 7.5 implique que le phénomène de résonance peut être détecté à taux de natalité constant, pour différentes valeurs du taux de contact, ou à taux de contact constant, pour différentes valeurs du taux de natalité.

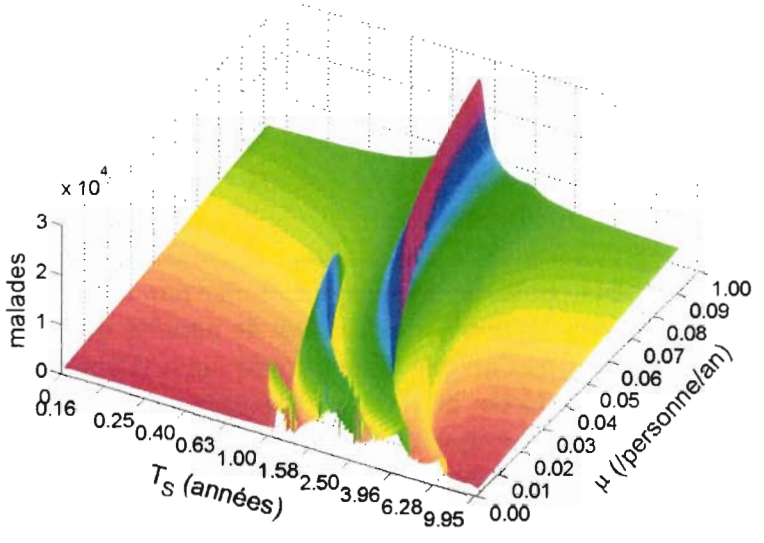


FIG. 7.5 – **Diagramme de bifurcation illustrant les effets de la saisonalité dans la transmission (T_S) et du degré de non-linéarité (μ) sur l'amplitude des épidémies.**

Nous avons utilisé cette propriété pour détecter un phénomène de résonance sur des données de dynamiques de rougeole pour lesquelles les taux de

ENCADRÉ 7.2
LA SOURCE DE NON-LINÉARITÉ DANS LE MODÈLE SEIR

Une fonction f est dite linéaire (ou proportionnelle) lorsqu'elle vérifie la relation suivante :

$$f(ax + b) = af(x) + b$$

Dans tous les autres cas, la fonction est dite non-linéaire. Les fonctions linéaires sont basées sur des **relations de proportionnalité** et sont donc relativement faciles à comprendre. En revanche, les fonctions non-linéaires ont souvent des comportements très **complexes**.

Le processus de transmission est au cœur de tous les modèles épidémiologiques. On définit la **force d'infection** λ comme la probabilité, pour un susceptible, de contracter la maladie. Pour les maladies à transmission directe comme la rougeole et autres maladies infantiles, on considère que la force d'infection est proportionnelle à la proportion de contagieux dans la population : $\lambda = \beta \mathrm{I}/\mathrm{N}$. L'équation différentielle du nombre de susceptibles dans la population s'écrit donc :

$$d\mathrm{S}/dt = \mu(\mathrm{N} - \mathrm{S}) - \beta \mathrm{SI}/\mathrm{N}$$

La **bi-linéarité** $\beta \mathrm{SI}/\mathrm{N}$ est la source de non-linéarité (et donc de complexité) des systèmes SEIR de maladies infantiles. Lorsque μ augmente, $\beta \mathrm{SI}/\mathrm{N}$ devient négligeable devant $\mu(\mathrm{N} - \mathrm{S})$, diminuant ainsi la non-linéarité du système. Ainsi, la non-linéarité d'un modèle SEIR de maladie infantile est assez bien contrôlée par l'intensité du taux de natalité : plus μ est élevé, plus le système et linéaire et inversement.

natalité sont différents. Pour un forçage saisonnier annuel ($T_S = 1$ an), la figure 7.5 prédit qu'une augmentation du taux de natalité de 0.01 à 0.05/individu/an se manifeste par un pic d'amplitude suivi d'une augmentation régulière de l'amplitude. Cette prédiction théorique a été vérifiée avec succès sur les données de dynamique de rougeole de 60 villes anglaises (figure 7.6), nous confortant ainsi dans la possibilité d'existence de résonance en dynamique de maladies infectieuses.

L'effet de la vaccination par pulsations sur la dynamique de la maladie se traduit également par des phénomènes de résonance paramétriques, comme décrit ci-dessus pour l'effet du forçage saisonnier. L'introduction du forçage saisonnier en même temps que la vaccination par pulsations augmente considérablement la **complexité** de la dynamique qui passe en **mode chaotique**. Malgré cette augmentation de complexité, le phénomène de résonance est toujours présent, même si plus difficile à détecter (figure 7.7A).

Une façon de s'affranchir de cette complexité est de considérer des statis-

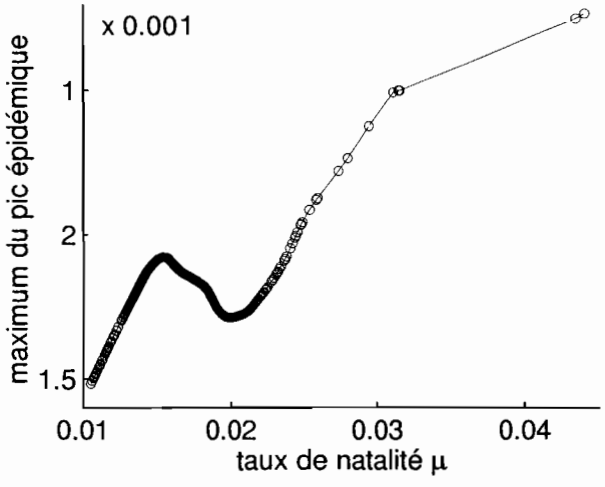


FIG. 7.6 – **Relation entre le maximum épidémique et le taux de natalité pour 60 villes anglaises dans la période pré-vaccinale (1994–1966). Les ronds représentent la tendance lissée par régression lowess, avec un paramètre de tension de 0.45. Données issues de** `http ://www.zoo.cam.ac.uk/zoostaff/grenfell/measles.html`.

tiques résumé. Par exemple, une grandeur épidémiologique intéressante est l'**incidence annuelle** (*i.e.* le nombre de malades par an). La figure 7.7B illustre l'évolution du nombre annuel moyen de malades en fonction de T_V. La tendance générale est attendue : plus la fréquence de vaccination ($1/T_V$) est forte, plus le nombre annuel moyen de malades est bas. Toutefois, localement nous observons l'inverse, et ce à cause du pic de résonance. Ainsi, sur la figure 7.7B, on observe qu'une vaccination tous les 25 mois produit un nombre annuel moyen de malades (1 315 000) 15% plus élevé qu'une vaccination tous les 24 mois (1 145 000).

Enfin, la figure 7.7C compare le nombre d'individus effectivement vaccinés dans les simulations avec le nombre d'individus à vacciner pour atteindre l'éradication de la maladie, selon les prédictions théoriques du paragraphe 7.1, page 128. Il apparaît que la condition $T_V < T_V^{max}$ ne suffit pas pour l'éradication de la maladie.

7.4 Discussion

Par sa nature périodique, la vaccination par pulsations, lorsqu'imparfaite, influence profondément la dynamique de la maladie. En **synchronisant** les dynamiques épidémiques dans différentes localités, elle augmente la

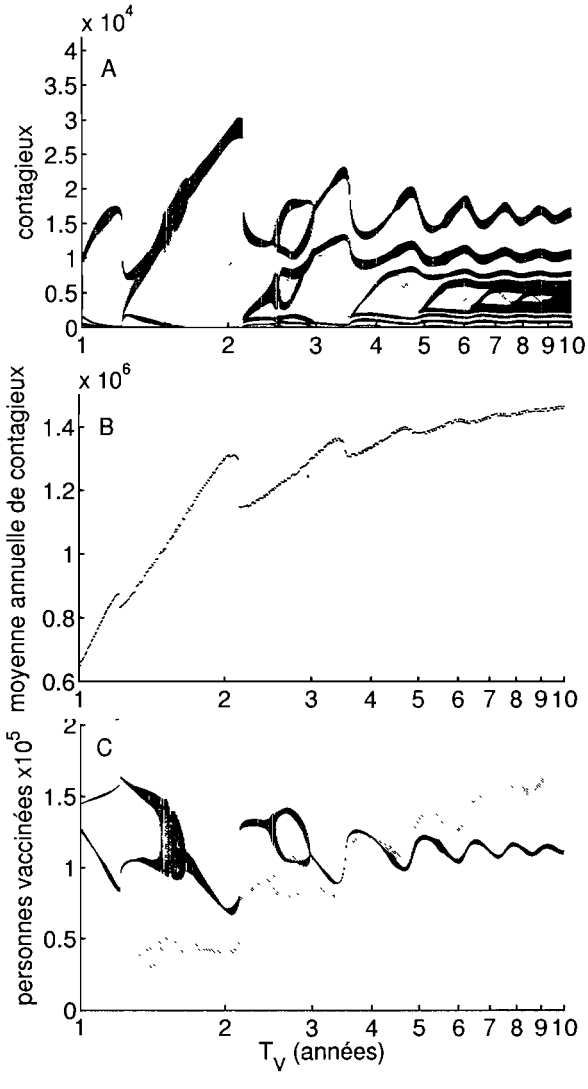


FIG. 7.7 – **Effet de la résonance paramétrique associée à une politique de vaccination par pulsations. A : diagramme de bifurcations de la dynamique. La pic de résonance principal est à $T_V = 2$ ans. Les autres pics se produisent aux fréquences harmoniques de cette fréquence de résonance. B : incidence annuelle. C : nombre d'individus effectivement vaccinés dans les simulations (en noir) et nombre théoriquement requis pour atteindre l'éradication (en gris).**

probabilité d'extinction globale de la maladie. Toutefois, les phénomènes de **résonance** qui lui sont associés peuvent produire des effets tout à fait inattendus, comme une augmentation de l'incidence annuelle lorsque le fréquence de vaccination augmente. Ces résultats sont basés sur l'étude numérique du comportement dynamique d'un modèle mathématique. Nos prédictions théoriques sur l'effet de la résonance associée au forçage saisonnier ont pu être vérifiées avec succès sur des données réelles. Malheureusement, du fait de l'application très récente de la vaccination par pulsations, nous ne disposons pas encore de données sur suffisamment d'années pour nous permettre de vérifier nos prédictions théoriques sur la résonance associée à la vaccination par pulsations.

Le principe de la vaccination de masse est basé uniquement sur les propriétés **statiques** du système hôte-parasite. Ce qui fait de la vaccination par pulsations une stratégie plus efficace et plus économique est que ses fondements théoriques tiennent explicitement compte de la **dynamique** de susceptibles, à travers le **taux de natalité**.

Toutefois, il est à noter que la dynamique du nombre de **contagieux** eux-mêmes n'est pas plus prise en compte dans le cadre de la vaccination par pulsations que dans le cadre de la vaccination de masse. En effet, les deux stratégies vaccinales reposent sur un seuil $1/R_0$ qui, par définition, est statique. Nos résultats montrent que la dynamique de la maladie elle-même peut fortement interférer avec l'application de la vaccination, produisant des phénomènes de résonance aux effets parfois inattendus, comme une augmentation du nombre de malades lorsque la fréquence de vaccination augmente. De plus, il apparaît que la prédiction théorique $T_V < T_V^{max}$ ne suffit pas pour l'éradication de la maladie. Il est très probable que ceci soit lié également à un phénomène de résonance proche de la période T_V^{max}.

Les théoriciens s'intéressant actuellement à la vaccination par pulsations ont privilégié l'**approche analytique**. L'avantage d'une telle approche est qu'elle fournit des formules générales, faciles d'application. L'inconvénient est que la résolution mathématique impose souvent des **hypothèses extrêmement simplificatrices**. C'est essentiellement pour ces raisons que la dynamique de la maladie elle-même n'est pas considérée dans les modèles actuellement développés. Nos résultats de **simulations** suggèrent que prendre en compte la dynamique de la maladie (à travers des paramètres comme sa période propre par exemple) pourrait encore améliorer très notablement l'efficacité de la vaccination par pulsations. Nous voyons pour cela deux directions. La première implique une recherche vers le développement de modèles analytiques rendant compte explicitement de la dynamique de la maladie. La deuxième, plus pragmatique, serait basée sur des simulations, comme ici, paramétrées au cas par cas, sur des systèmes spécifiques. Cette deuxième

approche ne fournit pas de résultats généraux mais est d'application pratique plus aisée. Dans tous les cas de figures, la confrontation des modèles analytiques et des simulations s'avère indispensable.

Chapitre 8

Conclusions et perspectives

LES relations hôte-parasite sont des processus durables au cours desquels les protagonistes s'altèrent mutuellement. Ces relations font intervenir une large gamme d'échelles spatiales (du moléculaire au populationnel) et temporelles (des processus dynamiques aux phénomènes évolutifs). La vaccination est une forme de protection artificielle de l'hôte basée sur la propriété de mémoire du système immunitaire des vertébrés. Le principe consiste à stimuler le système immunitaire en l'exposant à un agent (ou des parties d'agents) pathogène mort ou suffisamment affaibli pour ne pas provoquer la maladie. Simple dans le principe, la mise au point d'un vaccin efficace n'est pas toujours tâche facile[1]. La compréhension des relations moléculaires fines entre le parasite et le système immunitaire de l'hôte est un préalable à l'identification de cibles vaccinales potentiellement efficaces. Si le vaccin est efficace, il protège l'hôte contre les colonisations futures du parasite. De plus, si le vaccin protège suffisamment de personnes dans la population (immunité de groupe), il empêche la propagation de la maladie dans la population. Idéalement, ceci conduit à l'éradication totale de la maladie. En pratique, une telle éradication est rarement atteinte (le cas de la variole est le seul connu à ce jour) et, au mieux, le nombre de malades est réduit. Une vaccination imparfaite peut alors perturber les relations hôte-parasite dans un sens qui est ni toujours intuitif, ni toujours souhaitable. Savoir prédire les conséquences d'une vaccination imparfaite apparaît alors de première nécessité.

8.1 Les conclusions de la thèse

Dans ce contexte général, ce travail de thèse s'est focalisé sur quelques aspects particuliers. Une première partie a été consacrée à l'étude de la micro-

1. Comme en témoigne les difficultés rencontrées dans la lutte contre le SIDA.

coévolution entre le pathogène et le système immunitaire de l'hôte, dans le but d'identifier des cibles vaccinales potentiellement efficaces. L'étude de l'évolution moléculaire fine est aujourd'hui possible avec les capacités modernes de séquençage de gènes ainsi que le développement d'outils d'analyses statistiques puissants. Ces derniers nous permettent de quantifier et de localiser l'évolution sur les protéines, à l'échelle de l'acide aminé (chapitre 2). Nous avons appliqué ces méthodes à l'étude de deux microparasites, les HIV, virus responsables du SIDA et les *Leishmania*, protozoaires responsables des leishmanioses. Ces deux maladies sont des problèmes de santé publique majeurs à l'échelle mondiale et pour lesquelles le développement de vaccins est actuellement à l'étude.

La comparaison de l'évolution moléculaire des HIV (très virulents) et des SIV (apparemment peu virulents) met en évidence le rôle potentiellement important de la glycoprotéine membranaire dans les interactions avec le système immunitaire. Plus précisément, l'identification des acides aminés sous sélection permet de confirmer un modèle d'échappement des anticorps par les virus libres (chapitre 3).

Le développement récent de la génomique et du séquençage de gènes a permis de mettre en évidence l'importance des familles de gènes dans les relations hôte-parasite. Une famille de gène codant pour des cystéine protéases a été identifiée comme facteur de virulence chez les *Leishmania*. Des analyses phylogénétiques nous ont permis de comprendre l'histoire évolutive de cette famille de gènes. Des analyses de sélection positive ont identifié une portion de ces gènes à considérer comme cible vaccinale potentielle (chapitre 4).

Une deuxième partie s'est intéressée aux conséquences des politiques vaccinales sur la dynamique des maladies infectieuses dans les populations humaines. L'application d'une politique vaccinale vise à l'éradication globale d'une maladie. Une éradication globale ne peut être envisagée sans la compréhension de la dynamique spatiale de la maladie. Ainsi, bien que moins contagieuse que la rougeole, il est possible que la varicelle soit plus difficile à éradiquer que la rougeole (chapitre 5). Nos résultats de simulations indiquent que ceci est lié à la relation que la varicelle entretient avec une autre maladie, le zona, ce dernier jouant le rôle de réservoir de virus.

Il existe actuellement deux politiques principales de vaccination. La plus ancienne, la plus utilisée et la plus étudiée est la vaccination de masse dont nous présentons les principales propriétés dans le chapitre 6. Bien qu'efficace, cette politique est extrêmement lourde à mettre en place. La deuxième politique vaccinale, la vaccination par pulsations, est une alternative (récemment proposée) à la première. Le but était de proposer une politique vaccinale plus économique que la vaccination de masse, et qui soit abordable pour les pays en voie de développement. La nature intrinsèquement périodique de

cette deuxième politique vaccinale peut fortement influencer la dynamique des maladies dans un sens qui n'est pas toujours intuitif (chapitre 7).

8.2 Perspectives

Les suites directes de ce travail sont nombreuses. Parmi les thèmes qui n'ont pu être concrétisés durant la thèse, il y en a deux que nous aimerions explorer rapidement.

Evolution moléculaire et mode de transmission

Certains parasites ont plusieurs modes de transmission possible. Les contraintes évolutives sur le parasite dépendent fortement du mode de transmission et peuvent se traduire par des différences épidémiologiques. Par exemple, des modèles mathématiques théoriques prédisent qu'un mode de transmission vertical (de mère à enfant) devrait conduire à une moindre virulence qu'un mode de transmission horizontal (entre individus d'une même cohorte) [42, 2, 53]. L'impressionnante base de données de séquences de HIV offre une opportunité de choix pour tester de telles prédictions théoriques. En effet, l'origine de chaque séquence est très bien documentée et on peut distinguer quatre modes majeurs de transmission : hétérosexuelle, homosexuelle, intraveineuse et mère-enfant. Face à ce type d'information, deux questions apparaissent intéressantes :

- Y-a-t-il un signal phylogénétique du mode de transmission. Autrement dit, est-ce que certaines souches sont préférentiellement propagées par un mode de transmission plutôt qu'un autre ?
- Est-ce que le mode de transmission est responsable d'intensités d'évolution différentielles ?

Résonance et interactions

Nous avons mis en évidence que les dynamiques de maladies ont un fort potentiel à la résonance. En particulier, nous avons montré qu'un phénomène de résonance pouvait être associé à une saisonnalité sur le taux de contact, ainsi qu'à une vaccination par pulsations. Par ailleurs, des travaux récents ont mis en évidence des phénomènes d'interactions écologiques entre maladies infectieuses [170, 168]. L'idée est qu'un individu malade entre en convalescence et que son isolement à domicile le rend non susceptible pour d'autres maladies. Dans ce contexte, on peut s'intéresser à deux types de questions :

- Est-ce que la dynamique d'une maladie peut faire résonner celle d'une autre maladie avec laquelle elle est en compétition ?

- Est-ce que la vaccination d'une maladie peut influer sur la dynamique d'une autre maladie avec laquelle elle est en compétition ?

De nombreuses autres continuations de ce travail sont potentiellement envisageables, ne serait-ce que parce que nous ne nous sommes intéressés, dans cette thèse, qu'à deux extrêmes d'un vaste continuum d'échelles spatio-temporelles : la micro-coévolution moléculaire et la dynamique de population. L'étude de la vaccination peut faire intervenir toute une gamme d'échelles spatio-temporelles qui n'ont pas été abordées dans le présent travail : macro-coévolution entre hôte et pathogène, adaptabilité et dynamique de la réaction immunitaire, dynamique de la variabilité antigénique, *etc...* Plutôt que d'explorer ainsi plusieurs autres domaines, une perspective qui nous semble particulièrement intéressante serait d'essayer d'intégrer ces différents niveaux d'échelles sur un même modèle biologique. Chacun de ces niveaux d'échelles sont aujourd'hui étudiés séparément (et c'est essentiellement ce qui a été fait dans cette thèse) de façon souvent très complète. En revanche, les liens entre dynamique et évolution intra-hôte et inter-hôte sont encore très peu compris et étudiés. Comprendre ces liens entre différents niveaux d'échelle est d'importance pratique majeure, tant pour l'étude de l'évolution de la virulence / résistance, que pour l'étude de la vaccination [76]. Un de mes projets de post-doctorat concerne l'évolution et la dynamique de différentes souches d'une même maladie dans une population.

Bibliographie

[1] L. C. C. Afonso, T. M. Scharton, L. Q. Vieira, M. Wysocka, G. Trinchieri, and P. Scott. The adjuvant effect of interleukin-12 in a vaccine against *Leishmania major*. *Science*, 263 :235–237, 1994.

[2] P. Agnew and J. C. Koella. Virulence, parasite mode of transmission, and host fluctuation asymmetry. *Proceedings of the Royal Society of London*, B264 :9–15, 1997.

[3] Z. Agur. Randomness synchrony population persistence. *Journal of Theoretical Biology*, 112 :677–693, 1985.

[4] Z. Agur, L. Cojocaru, R. M. Anderson, and Y. L. Danon. Pulse mass measles vaccination across age cohorts. *Proceedings of the National Academy of Sciences of the USA*, 90 :11698–11702, 1993.

[5] H. Akashi. Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in drosophila DNA. *Genetics*, 139 :1067–1076, 1995.

[6] J. Alexander, G. H. Coombs, and J. C. Mottram. *Leishmania mexicana* cysteine proteinase-deficient mutants have attenuated virulence for mice and potentiate a Th1 response. *Journal of Immunology*, 161 :6794–6801, 1998.

[7] R. M. Anderson. *Transmission dynamics and control of infectious disease agents*, pages 149–176. Springer-Verlag, Berlin, 1982.

[8] R. M. Anderson, B. T. Grenfell, and R. M. May. Oscillatory fluctuations in the incidence of infectious diseases and the impact of vaccination : time series analysis. *Journal of Hygiene, Cambridge*, 93 :587–608, 1984.

[9] R. M. Anderson and R. M. May. Vaccination against rubella and measles : quantitative investigations of different policies. *Journal of Hygiene, Cambridge*, 90 :259–325, 1983.

[10] R. M. Anderson and R. M. May. Vaccination and herd immunity to infectious diseases. *Nature*, 318 :323–329, 1985.

[11] R. M. Anderson and R. M. May. *Infectious diseases of humans. Dynamics and control.* Oxford University Press, Oxford, 1991.

[12] M. Anisimova, J. P. Bielawski, and Z. Yang. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Molecular Biology and Evolution*, 18 :1585–1592, 2001.

[13] M. Anisimova, J. P. Bielawski, and Z. Yang. Accuracy and power of bayes prediction of amino acid sites under positive selection. *Molecular Biology and Evolution*, 19 :950–958, 2002.

[14] M. Anisimova, R. Nielsen, and Z. Yang. Effect of recombinaison on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics*, 164 :1229–1236, 2003.

[15] R. Antia, R. R. Regoes, J. C. Koella, and C. T. Bergstrom. The role of evolution in the emergence of infectious diseases. *Nature*, 426 :658–661, 2003.

[16] N. J. T. Bailey. *The Mathematical Theory of Infectious Diseases and its Application.* Griffin, London, 1957.

[17] N. J. T. Bailey. *The Mathematical Theory of Infectious Diseases and its Application.* Griffin, London, 1975.

[18] R. S. Baltimore. Why is chickenpox more serious in adults than in children ? *Health News*, 6 :10, 2000.

[19] G. Bart, M. J. Frame, R. Carter, G. H. Coombs, and J. C. Mottram. Cathepsin B-like cysteine proteinase-deficient mutants of *Leishmania mexicana*. *Molecular Biochemistry and Parasitology*, 88 :53–61, 1997.

[20] M. S. Bartlett. The critical community size for measles in the United States. *Journal of the Royal Statistical Society*, 123 :37–44, 1960.

[21] A. L. Bañuls. *Apport de la génétique évolutive à l'épidémiologie et à la taxonomie du genre Leishmania.* PhD thesis, 1998.

[22] M. A. Beaumont and B. Rannala. The bayesian revolution in genetics. *Nature Reviews Genetics*, 5 :251–261, 2004.

[23] B. E. Beer, E. Bailes, P. M. Sharp, and V. M. Hirsch. Diversity and evolution of primate lentiviruses. In C. L. Kuiken, B. Foley, B. Hahn, B. Korber, F. McCutchan, P. A. Marx, J. W. Mellors, J. I. Mullins, J. Sodroski, and S. Wolinksy, editors, *Human Retroviruses and AIDS : A Compilation and Analysis of Nucleic Acid and Amino Acid Sequences*, pages 460–474. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, 1999.

[24] M. Begon, J. L. Harper, and C. R. Townsend. *Ecology.* Blackwell Science, Oxford, 1996.

[25] D. Bernoulli. Essai d'une nouvelle analyse de la mortalité causée par la petite vérole et des avantages de l'inoculum pour la prévenir. *Mém. Math. Phys. Acad. Roy. Sci., Paris*, pages 1–45, 1760.

[26] L. Billings and I. B. Schwartz. Exciting chaos with noise : unexpected dynamics in epidemic outbreaks. *Journal of Mathematical Biology*, 44 :31–48, 2002.

[27] B. Blasius, A. Huppert, and L. Stone. Complex dynamics and phase synchronization in spatially extended ecological systems. *Nature*, 399 :354–359, 1999.

[28] B. M. Bolker and B. T. Grenfell. Chaos and biological complexity in measles dynamics. *Proceedings of the Royal Society of London*, B251 :75–81, 1993.

[29] B. M. Bolker and B. T. Grenfell. Impact of vaccination on the spatial correlation and persistence of measles dynamics. *Proceedings of the National Academy of Sciences of the USA*, 93 :12648–12653, 1996.

[30] S. Bonhoeffer, E. C. Holmes, and M. A. Nowak. Causes of HIV diversity. *Nature*, 376 :125, 1995.

[31] S. Bonhoeffer and M. A. Nowak. Intra-host versus inter-host selection : viral strategies of immune function impairment. *Proceedings of the National Academy of Sciences of the USA*, 91 :8062–8066, 1994.

[32] P. Botarelli, B. A. Houlden, N. L. Heigwood, C. Servis, D. Montagna, and S. Abrignani. N-glycosylation of HIV-1 *gp*120 may constrain recognition by T lymphocytes. *Journal of Immunology*, 147 :3128–3132, 1991.

[33] M. Brisson and W. J. Edmunds. Epidemiology of varicella-zoster virus in England and Wales. *Journal of Medicine and Virology*, 70 :S9–S14, 2003.

[34] L. U. Buxbaum, H. Denise, G. H. Coombs, J. Alexander, J. C. Mottram, and P. Scott. Cysteine protease B of *Leishmania mexicana* inhibits host Th1 responses and protective immunity. *Journal of Immunology*, 171 :3711–3717, 2003.

[35] J. Cairns. *Cancer, Sciences and Society.* W. H. Freeman, San Francisco, 1975.

[36] D. S. Callaway, R. M. Ribiero, and M. A. Nowak. Virus phenotype switching and disease progression in HIV-1 infection. *Proceedings of the Royal Society of London series*, B266 :2523–2530, 2000.

[37] Z. Chen, P. Telfier, A. Gettie, P. Reed, L. Q. Zhang, D. D. Ho, and P. A. Marx. Genetic characterisation of new West African simian immunodeficiency virus SIV_{sm} : geographic clustering of household-derived

SIV strains with human immunodeficiency virus type 2 subtypes and genetically diverse viruses from a single feral sooty mangabey troop. *Journal of Virology*, 70 :3617–3627, 2000.

[38] M. Choisy, C. H. Woelk, J. F. Guégan, and D. L. Robertson. Comparative study of molecular evolution in different HIV clades. *Journal of Virology*, 78 :1962–1970, 2004.

[39] D. H. Clayton and J. Moore, editors. *Host-parasite evolution.* Oxford University Press, Oxford, 1997.

[40] C. Combes. *Interactions durables. Écologie et évolution du parasitisme.* Masson, Paris, 1995.

[41] T. Coulson, P. Rohani, and M. Pascual. Skeletons, noise and population growth : the end of an old debate ? *Trends in Ecology and Evolution*, 19 :359–364, 2004.

[42] T. Day. Parasite transmission mode and the evolution of virulence. *Evolution*, 55 :2389–2400, 2001.

[43] S. De Souza Leao, T. Lang, E. Prina, R. Hellio, and J. C. Antoine. Intracellular *Leishmania amazonensis* amastigotes internalize and degrate MHC class II molecules of their host cells. *Journal of Cell Science*, 108 :3219–3231, 1995.

[44] S. Deguen. *Modélisation de l'épidémie de la varicelle en France sur la base des données (1991-1998) du Réseau Sentinelle de surveillance des maladies transmissibles.* PhD thesis, 1999.

[45] R. S. Diaz, E. C. Sabino, A. Mayer, J. W. Mosley, and M. P. Busch. Dual human immunodeficiency virus type 1 infection in a dually exposed transfusion recipient. *Journal of Virology*, 69 :3273–3281, 1995.

[46] K. Dietz. The incidence of infectious diseases under the influence of seasonal fluctuations. *Lecture Notes in Biomathematics*, 11 :1–5, 1976.

[47] A. d'Onofrio. Stability properties of pulses vaccination strategy in SEIR epidemic model. *Mathematical Biosciences*, 179 :57–72, 2002.

[48] D. J. D. Earn, J. Dushoff, and S. A. Levin. Ecology and evolution of the flu. *Trends in Ecology and Evolution*, 17(7) :334–340, 2002.

[49] D. J. D. Earn, P. Rohani, B. M. Bolker, and B. T. Grenfell. A simple model for complex dynamical transitions in epidemics. *Science*, 287 :667–670, 2000.

[50] D. J. D. Earn, P. Rohani, and B. T. Grenfell. Persistence, chaos and synchrony in ecology and epidemiology. *Proceedings of the Royal Society of London*, B265 :7–10, 1998.

[51] S. Ellner, B. A. Bailey, G. V. Bobashev, A. R. Gallant, B. T. Grenfell, and D. W. Nychka. Noise and nonlinearity in measles epidemics : combining mechanistic and statistical approaches to population modeling. *American Naturalist*, 151 :425–440, 1998.

[52] S. Ellner and P. Turchin. Chaos in a noisy world : new methods and evidence from time-series analysis. *American Naturalist*, 145 :343–375, 1995.

[53] P. W. Ewald. *Evolution of Infectious Diseases.* Oxford University Press, Oxford, 1994.

[54] P. W. Ewald. Evolutionary control of HIV and other sexually transmitted diseases. In R. Trvathan, E. O. Smith, and J. J. McKenna, editors, *Evolutionary Medicine*, pages 271–311. Oxford University Press, New York, 1999.

[55] J. P. A. Ezquerra. *Las Leishmaniasis : de la biología al control.* Junta de Castilla y León, Santa Clara, 1997.

[56] J. C. Fay and C. I. Wu. The neutral theory in the genomic era. *Current Opinion in Genetics and Development*, 11 :642–646, 2001.

[57] J. Felsenstein. Evolutionary trees from DNA sequences : a maximum likelihood approach. *Journal of Molecular Evolution*, 17 :368–376, 1981.

[58] J. Felsenstein. *PHYLIP (phylogeny inference package). Version 3.5c.* Department of Genetics, University of Washington, Seattle, 1994.

[59] C. Filippi, L. Malherbe, V. Julia, and N. Glaichenhaus. L'immunité contre les leishmanies. *Médecine/Sciences*, 17 :1120–1128, 2001.

[60] P. E. M. Fine and J. A. Clarkson. Measles in England and Wales. 1. An analysis of factors underlying seasonal patterns. *International Journal of Epidemiology*, 11 :5–14, 1982.

[61] P. E. M. Fine and J. A. Clarkson. Seasonal influences on pertussis. *International Journal of Epidemiology*, 15 :237–247, 1986.

[62] F. D. Finkelman and J. F. Urban. Cytokines making the right choice. *Parasitology Today*, 8 :311–314, 1992.

[63] W. M. Fitch, R. M. Bush, C. A. Bender, and N. J. Cox. Long term trends in the evolution of H(3) HA1 human influenza type A. *Proceedings of the National Academy of Sciences of the USA*, 94 :7712–7718, 1997.

[64] S. A. Frank. *Immunology and evolution of infectious diseases.* Princeton University Press, Princeton, 2002.

[65] Y. X. Fu and W. H. Li. Statistical tests of neutrality of mutations. *Genetics*, 133 :693–709, 1993.

[66] S. Gandon, M. J. Mackinnon, S. Nee, and A. F. Read. Imperfect vaccines and the evolution of pathogen virulence. *Nature*, 414(6865) :751–756, 2001.

[67] F. Gao, E. Bailes, D. L. Robertson, Y. Chen, C. M. Rodenburg, S. F. Michael, L. B. Cummins, L. O. Arthur, M. Peeters, G. M. Shaw, P. M. Sharp, and B. H. Hahn. Origin of HIV-1 in the chipanzee *Pan troglodytes troglodytes*. *Nature*, 397 :436–441, 1999.

[68] L. Garret. The next epidemic. In J. L. Mann, D. Tarantola, and T. W. Netter, editors, *AIDS in the world*, pages 825–839. Harvard University Press, Cambridge, 1993.

[69] B. Gaschen, J. Taylor, K. Yusim, B. Foley, F. Gao, D. Lang, V. Novitsky, B. Haynes, B. Hahn, T. Bhattacharya, and B. Korber. Diversity considerations in HIV-1 vaccine selection. *Science*, 296 :2354–2360, 2002.

[70] N. Goldman and Z. Yang. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution*, 11 :725–736, 1994.

[71] B. T. Grenfell, O. N. Bjørnstad, and B. Finkenstädt. Dynamics of measles epidemics : scaling noise, determinism and predictability with the TSIR model. *Ecological Monographs*, 72 :185–202, 2002.

[72] B. T. Grenfell, O. N. Bjørnstad, and J. Kappey. Travelling waves and spatial hierarchies in measles epidemics. *Nature*, 414 :716–723, 2001.

[73] B. T. Grenfell and B. M. Bolker. Cities and villages : infection hierarchies in a measles metapopulation. *Ecology Letters*, 1 :63–70, 1998.

[74] B. T. Grenfell and A. P. Dobson, editors. *Ecology of infectious diseases in natural populations*. Cambridge University Press, Cambridge, 1995.

[75] B. T. Grenfell and J. Harwood. (Meta)population dynamics of infectious diseases. *Trends in Ecology and Evolution*, 148 :317–335, 1997.

[76] B. T Grenfell, O. G. Pybus, J. R. Gog, J. L. N. Wood, J. M. Daly, J. A. Mumford, and E. C. Holmes. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science*, 303 :327–332, 2004.

[77] G. R. Grimmett and D. R. Stirzaker. *Probability and random processes*. Clarendon Press, Oxford, 1992.

[78] J. F. Guégan, F. Thomas, M. E. Hochberg, T. De Meeûs, and F. Renaud. Disease diversity and human fertility. *Evolution*, 55 :1308–1314, 2001.

[79] B. H. Hahn, G. M. Shaw, K. M. De Cock, and P. M. Sharp. AIDS as a zoonosis : scientific and public health implications. *Science*, 287 :607–614, 2000.

[80] W. H. Hamer. Epidemic disease in England. *The Lancet*, i :733–739, 1906.

[81] E. Handman. *Leishmania* vaccines : old and new. *Parasitology Today*, 13 :236–238, 1997.

[82] J. E. Hansen, O. Lund, N. Tolstrup, A. A. Gooley, K. L. Williams, and S. Brunak. NetOGlyc : prediction of mucin type O-glycosylation sites based on sequence context and surface accessibility. *Glycoconjugate Journal*, 15 :115–130, 1998.

[83] I. A. Hanski. *Metapopulation Ecology*. Oxford University Press, Oxford, 1999.

[84] I. A. Hanski and O. E. Gaggiotti, editors. *Ecology, Genetics and Evolution of Metapopulations. Standard Methods for Inventory and Monitoring*. Elsevier, London, 2004.

[85] I. A. Hanski and M. E. Gilpin, editors. *Metapopulation Biology. Ecology, Genetics and Evolution*. Academic Press, London, 1997.

[86] D. L. Hartl and A. G. Clark. *Principles of population genetics*. Sinauer, Sunderland, 1989.

[87] M. Hasegawa, H. Kishino, and T. Yano. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 21 :160–174, 1985.

[88] A. Hastings, C. Hom, S. Ellner, P. Turchin, and H. C. J. Godfray. Chaos in ecology : is mother nature a strange attractor ? *Annual Reviews of Ecology and Systematics*, 24 :1–33, 1993.

[89] H. W. Hethcote. Oscillations in an endemic model for pertussis. *Canadian Applied Mathematics Quarterly*, 6 :61–88, 1998.

[90] E. C. Holmes. On the origin and evolution of the human immunodeficiency virus (HIV). *Biological Review*, 76 :239–254, 2001.

[91] R. R. Hudson, M. Kreitman, and M. Aguadé. A test of neutral molecular evolution based on nucleotide data. *Genetics*, 116 :153–159, 1987.

[92] A. L. Hughes. Evolution of cysteine proteinases in eukaryotes. *Molecular Phylogenetics and Evolution*, 3 :310–321, 1994.

[93] A. L. Hughes and M. Nei. Patterns of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature*, 335(6186) :167–170, 1988.

[94] Y. Ina. New methods for estimating the number of synonymous and nonsynonymous substitutions. *Journal of Molecular Evolution*, 40 :190–226, 1995.

[95] E. A. Jackson. *Perspectives of Nonlinear Dynamics : Volume 1.* Cambridge University Press, Cambridge, 1992.

[96] C. A. Janeway, P. Travers, J. D. Capra, and M. J. Walport. *Immunobiology : the immune system in health and diseases.* Garland Publishers, New York, 1999.

[97] G. M. Jenkins, A. Rambaut, O. G. Pybus, and E. C. Holmes. Rates of molecular evolution in RNA viruses : a quantitative phylogenetic analysis. *Journal of Molecular Evolution*, 54 :156–165, 2002.

[98] M. J. Jin, H. Hui, D. L. Robertson, M. C. Müller, F. Barré-Sinoussi, V. M. Hirsch, J. S. Allan, G. M. Shaw, P. M. Sharp, and B. H. Hahn. Mosaic genome structure of simian immunodeficiency virus from West African green monkeys. *EMBO Journal*, 13 :2935–2947, 1994.

[99] T. H. Jukes and C. R. Cantor. Evolution of protein molecules. In H. N. Munro, editor, *Mamalian protein metabolism III*, pages 21–132. Academic Press, New-York, 1969.

[100] P. J. Kanki, K. U. Travers, S. M'boup, C. C. Hsiej, R. G. Marlink, A. Gueye-Ndiaye, T. Siby, I. Thior, R. G. Hernandez-Avila, M. Sankalé, J. L. Ndoye, and M. Essex. Slower heterosexual spread of HIV-2 than HIV-1. *Lancet*, 343 :943–946, 1994.

[101] M. Keeling. Modelling the persistence of measles. *Trends in Microbiology*, 5 :513–518, 1997.

[102] M. J. Keeling, O. N. Bjørnstad, and B. T. Grenfell. *Metapopulation Dynamics of Infectious Diseases*, pages 415–445. Elsevier, Amsterdam, 2004.

[103] M. J. Keeling and B. T. Grenfell. Disease extinction and community size : modeling the persistence of measles. *Science*, 275 :65–67, 1997.

[104] M. J. Keeling, P. Rohani, and B. T. Grenfell. Seasonally forced diseases dynamics explored as switching between attractors. *Physica D*, 148 :317–335, 2001.

[105] W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London*, A115 :700–721, 1927.

[106] M. Kimura. Evolutionary rate at the molecular level. *Nature*, 217 :624–626, 1968.

[107] M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16 :111–120, 1980.

[108] M. Kimura. *The neutral theory of molecular evolution.* Cambridge University Press, Cambridge, 1983.

[109] R. Kondo, S. Horai, Y. Satta, and N. Tkahata. Evolution of hominoid mitochondrial DNA with special reference to the silent substitution rate over the genome. *Journal of Molecular Evolution*, 36 :517–531, 1993.

[110] B. Korber, M. Muldoon, J. Theiler, F. Gao, R. Gupta, A. Lapedes, B. H. Hahn, S. Wolinsky, and T. Bhattacharya. Timing the ancestor of the HIV-1 pandemic strains. *Science*, 288 :1789–96, 2000.

[111] M. Kot. *Elements of Mathematical Ecology.* Cambridge University Press, Cambridge, 2001.

[112] P. D. Kwong, M. Doyle, D. J. Casper, C. Cicala, S. A. Leavitt, S. Majeed, T. D. Steenbeke, M. Venturi, I. Chaiken, M. Fung, H. Katinger, Parren P. W., J. Robinson, D. Van Ryk, L. Wang, D. R. Burton, Wyatt R. Freire, E., J. Sodroski, W. A. Hendrickson, and J. Arthos. HIV-1 evades antibody-mediated neutralization through conformational masking of receptor-binding sites. *Nature*, 420 :678–682, 2002.

[113] A. J. Leigh Brown and D. D. Richman. Gambling on the evolution of drug resistance ? *Nature Medicine*, 3 :268–271, 1997.

[114] B. R. Levin and Bull J. J. Short-sighted evolution and the virulence of pathogenic micro-organisms. *Trends in Microbiology*, 2 :76–81, 1994.

[115] R. C. Lewontin and J. Krakauer. Distribution of gene frequency as a test of the theory of selective neutrality of polymorphisms. *Genetics*, 74 :175–195, 1973.

[116] W. H. Li, M. Tanimura, and P. Sharp. Rates and dates of divergence between AIDS virus nucleotide sequences. *Molecular Biology and Evolution*, 5 :313–330, 1988.

[117] S. Liedtke, R. Geyer, and H. Geyer. Host-cell-specific glycosylation of HIV-2 envelope glycoprotein. *Glycoconjugate Journal*, 14 :785–793, 1997.

[118] A. L. Lloyd and R. M. May. Spatial heterogeneity in epidemic models. *Journal of Theoretical Biology*, 179 :1–11, 1996.

[119] W. P. London and J. A. Yorke. Recurrent outbreaks of measles, chickenpox and mumps. I Seasonal variation in contact rates. *American Journal of Epidemiology*, 98 :453–468, 1973.

[120] B. F. J. Manly. *Randomization, Bootstrap and Monte Carlo Methods in Biology.* Chapman & Hall, New-York, 1997.

[121] J. R. Mascola and D. C. Montferiori. HIV-1 : nature's master of disguise. *Nature Medicine*, 9 :393–394, 2003.

[122] R. M. May. Biological populations with nonoverlapping generations : stable points, stable cycles and chaos. *Science*, 186 :645–647, 1974.

[123] R. M. May. Simple mathematical models with very complicated dynamics. *Nature*, 261 :459–467, 1976.

[124] L. McCoy, F. Sorvillo, and P. Simon. Varicelle-related mortality in California, 1988-2000. *Pediatrics and Infectious Diseases Journal*, 23 :498–503, 2004.

[125] J. H. McDonald and M. Kreitman. Adaptative protein evolution at the Adh locus in Drosophila. *Nature*, 351 :652–654, 1991.

[126] J. H. McKerrow, E. Sun, P. J. Rosenthal, and J. Bouvier. The proteases and pathogenicity of parasitic protozoa. *Annual Review of Microbiology*, 47 :821–853, 1993.

[127] D. McMahon-Pratt, P. E. Kima, and L. Soong. *Leishmania* amastigote target antigens : the challenge of a stealthy intracellular parasite. *Parasitology Today*, 14 :31–34, 1998.

[128] A. J. McMichael and R. E. Phillips. Escape of Human immunodeficiency virus from immune control. *Annual Reviews of Immunology*, 15 :271–196, 1997.

[129] A. J. McMichael and S. L. Rowland-Jones. Cellular immune reponses to HIV. *Nature*, 410 :980–987, 2001.

[130] W. Messier and C. B. Stewart. Episodic adaptive evolution of primate lysozymes. *Nature*, 385 :151–154, 1997.

[131] T. Miyata and T. Yasunaga. Molecular evolution of mRNA : a method for estimating rates of synonymous and amino-acid substitutions from homologous nucleotide sequences and its application. *Journal of Molecular Evolution*, 16 :23–36, 1980.

[132] C. B. Moore, M. John, I. R. James, F. T. Christiansen, C. S. Witt, and S. A. Mallal. Evidence of HIV-1 adaptation to HLA-restricted immune responses at a population level. *Science*, 296 :1439–1443, 2002.

[133] P. A. P. Moran. The statistical analysis of the canadian lynx cycle. II. Synchronization and meteorology. *Australian Journal of Zoology*, 1 :291–298, 1953.

[134] S. S. Morse. Factors in the emergence of infectious diseases. *Emerging Infectious Diseases*, 1(1) :7–15, 1995.

[135] J. C. Mottram, D. R. Brooks, and G. H. Coombs. Roles of cysteine proteinases of trypanosomes and *Leishmania* in host-parasite interactions. *Current Opinion in Microbiology*, 1 :455–460, 1998.

[136] J. C. Mottram, Coombs G. H. Robertson, D. C., and J. D. Barry. A developmentally regulated cysteine proteinase gene of *Leishmania mexicana*. *Molecular Microbiology*, 6 :1925–1932, 1992.

[137] J. C. Mottram, A. E. Souza, J. E. Hutchison, R. Carter, M. J. Frame, and G. H. Coombs. Evidence from disruption of the *lmcpb* gene array of *Leishmania mexicana* that cysteine proteinases are virulence factors. *Proceedings of the National Academy of Sciences of the USA*, 93 :6008–6013, 1996.

[138] L. Moutouh, J. Corbeil, and D. D. Richman. Recombination leads to the rapide emergence of HIV-1 dually resistant mutants under selective drug presure. *Proceedings of the National Academy of Sciences of the USA*, 93 :6106–6111, 1996.

[139] H. J. Muller. Some genetic aspects of sex. *American Naturalist*, 66 :118–138, 1932.

[140] H. J. Muller. The relation of recombinaison to mutational advance. *Mutation Research*, 1 :2–9, 1964.

[141] S. V. Muse. Estimating synonymous and nonsynonymous substitution rates. *Molecular Biology and Evolution*, 13 :105–114, 1996.

[142] S. V. Muse and B. Gaut. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution*, 11 :715–724, 1994.

[143] G.J. Nabel. Challenges and opportunities for the development of an AIDS vaccine. *Nature*, 410 :1002–1007, 2001.

[144] M. Nei. Gene duplication and nucleotide substitution in evolution. *Nature*, 221(5175) :40–42, 1969.

[145] M. Nei. Gene duplication and nucleotide substitution in evolution. *Nature*, 221 :40–42, 1969.

[146] M. Nei and T. Gojobori. Simple methods for estimating the number of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution*, 3 :418–426, 1986.

[147] R. Nielsen and Z Yang. Likelihood models for detecting positively selected amino-acid sites and applications to the HIV-1 envelope gene. *Genetics*, 148 :929–936, 1998.

[148] D. J. Nokes and R. M. Anderson. The use of mathematical models in the epidemiology study of infectious diseases and in the design of mass vaccination programmes. *Epidemiology and Infection*, 101 :1–20, 1988.

[149] D. J. Nokes and J. Swinton. Vaccination in pulses : a strategy for global eradication of measles and polio ? *Trends in Microbiology*, 5(1) :14–19, 1997.

[150] L. F. Olsen and W. M. Schaffer. Chaos vs noisy periodicity : alternative hypotheses for childhood epidemics. *Science*, 249 :499–504, 1990.

[151] World Health Organisation. Final report of the global commission for the certification of smallpox eradication. Technical report, Geneva, 1980.

[152] M. T. Osterholm. Emerging infectious diseases. A real public health crisis ? *Postgraduate Medicine*, 100 :15–16, 1996.

[153] U. Parikh, C. Calef, B. Larder, and R. Schinazi. Mutations in retroviral genes associated with drug resistance. In C. L. Kuiken, B. Foley, E. Freed, B. Hahn, B. Korber, P. A. Marx, F. McCutchan, J. W. Mellors, and S. Wolinksy, editors, *HIV Sequence Compendium*, pages 94–183. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, 2002.

[154] A. S. Perelson, A. U. Neumann, M. Markowitz, J. M. Leonard, and D. D. Ho. HIV-1 dynamics in vivo : virion clearance rate, infected cell life-span, and viral generation time. *Science*, 271 :1582–1586, 1996.

[155] F. Perler, A. Efstradiadis, P. Lomedico, W. Gilbert, R. Kolodner, and J. Dodgson. The evolution of genes : the chicken preproinsulin gene. *Cell*, 20 :555–556, 1980.

[156] P. Piot, M. Bartos, P. D. Ghys, N. Walker, and Schwartländer. The global impact of HIV/AIDS. *Nature*, 410 :968–973, 2001.

[157] R. Pool. Ecologists flirt with chaos. *Nature*, 243 :310–313, 1989.

[158] R. Pool. Is it chaos, or is it just noise ? *Science*, 243 :25–28, 1989.

[159] D. A. Price, P. J. R. Gouldner, P. Klernerman, A. K. Sewell, P. J. Easterbrook, M Troop, C. R. M. Bangham, and R. E. Phillips. Positive selection of HIV-1 cytotoxic T lymphocyte escape variants during primary infection. *Proceedings of the National Academy of Sciences of the USA*, 94 :1890–1895, 1997.

[160] S. Rafati, A. H. Salmanian, K. Hashemi, C. Schaff, S. Belli, and N. Fasel. Identification of *Leishmania major* cysteine proteinases as targets of the immune response in Humans. *Molecular and Biochemical Parasitology*, 113 :35–43, 2001.

[161] A. Rambaut, D. Posada, K. A. Crandall, and E. C. Holmes. The causes and consequences of HIV evolution. *Nature Reviews Genetics*, 5 :52–61, 2004.

[162] N. D. Rawlings and A. J. Barrett. Evolutionary families of peptidases. *Biochemical Journal*, 290 :205–218, 1993.

[163] M. A. Rey-Cuillé, J. L. Berthier, M. C. Bomsel-Demontoy, Y. Chaduc, L. Montagnier, A. G. Hovanessian, and L. A. Chakrabarti. Simian immunodeficiency virus replicates to high levels in sooty mangabeys without inducing disease. *Journal of Virology*, 72 :3872–3886, 1998.

[164] C. J. Rhodes, H. J. Jensen, and R. M. Anderson. On the critical behaviour of simple epidemics. *Proceedings of the Royal Society of London*, B264 :1639–1646, 1976.

[165] D. D. Richman. HIV chemotherapy. *Nature*, 410 :995–1001, 2001.

[166] C. Robert. *L'analyse statistique bayesienne.* Economica, Paris, 1992.

[167] D. L. Robertson, P. M. Sharp, F. E. McCuthan, and B. H. Hahn. Recombination in HIV-1. *Nature*, 374 :124–126, 1995.

[168] P. Rohani, D. J. D. Earn, B. F. Finkenstadt, and B. T. Grenfell. Population dynamics interference among childhood diseases. *Proceedings of the Royal Society of London*, B265 :2033–2041, 1998.

[169] P. Rohani, D. J. D. Earn, and B. T. Grenfell. Opposite patterns of synchrony in sympatric diseases metapopulations. *Science*, 286 :968–971, 1999.

[170] P. Rohani, C. J. Green, N. B. Mantilla-Beniers, and B. T. Grenfell. Ecological interference among fatal infections. *Nature*, 422 :885–888, 2003.

[171] P. Rohani, M. J. Keeling, and Grenfell B. T. The interplay between determinism and stochasticity in childhood diseases. *American Naturalist*, 159 :569–481, 2002.

[172] P. J. Rosenthal. Proteases of protozoan parasites. *Advances in Parasitology*, 43 :106–159, 1999.

[173] R. Ross. *The Prevention of Malaria.* Murray, London, 1911.

[174] S. L. Rowland-Jones. Survival with HIV infection : good luck or good breeding ? *Trends in Genetics*, 14 :343–345, 1998.

[175] J. Ruffié and J.-C. Sournia. *Les épidémies dans l'histoire de l'homme. De la peste au SIDA.* Flammarion, Paris, 1995.

[176] D. Sacks and S. Kamhawi. Molecular aspects of parasite-vector and vector-host interactions in leishmaniasis. *Annual Review of Microbiology*, 55 :453–483, 2001.

[177] M. Sajid and J. H. McKerrow. Cysteine proteases of parasitic organisms. *Molecular and Biochemical Parasitology*, 120 :1–21, 2002.

[178] W. M. Schaffer and M. Kot. Nearly one dimensional dynamics in an epidemic. *Journal of Theoretical Biology*, 112 :403–427, 1985.

[179] D. Schenzle. An age-structured model of pre- and post-vaccination measles transmission. *IMA Journal of Mathematics Applied in Medicine and Biology*, 1 :169–191, 1984.

[180] P. M. Selzer, S. Pingel, I. Hsieh, B. Ugele, V. J. Chan, J. C. Engel, M. Bogyo, D. G. Russell, J. A. Sakanari, and J. H. McKerrow. Cysteine protease inhibitors as chemotherapy : lessons from a parasite target. *Proceedings of the National Academy of Sciences of the USA*, 96 :11015–11022, 1999.

[181] P. Sharp. In search of molecular darwinism. *Nature*, 385 :111–112, 1997. .

[182] P. M. Sharp, E. Bailes, R. R. Chaudhuri, C. M. Rodenburg, M. O. Santiago, and B. H. Hahn. The origins of acquired immune deficiency syndrome viruses : where and when ? *Phylosophical Transcations of the Royal Society of London series*, B356 :867–876, 2001.

[183] T. Sharton-Kersten, L. C. C. Afonso, M. Wysocka, G. Trinchieri, and P. Scott. IL-12 is required for natural killer cell activation and sbsequent T-helper 1 cell development in experimental leishmaniasis. *Journal of Immunology*, 154 :5320–5330, 1995.

[184] J. S. Shoemaker, I. S. Painter, and B. S. Weir. Bayesian statistics in genetics : a guide for the uninitiated. *Trends in Genetics*, 15(9) :345–358, 1999.

[185] E. G. Shpaer and J. I. Mullins. Rates of amino acid changes in the envelope protein correlate with pathogenicity of primate lentiviruses. *Journal of Molecular Evolution*, 28 :275–282, 1993.

[186] B. Shulgin, L. Stone, and Z. Agur. Pulse vaccination strategy in the SIR epidemic model. *Bulletin of Mathematical Biology*, 60 :1123–1148, 1998.

[187] J. D. Smith, C. E. Chitnis, A. G. Craig, D. J. Roberts, D. E. Hudson-Taylor, D. S. Peterson, R. Pinches, C. I. Newbold, and L. H. Miller. Switches in expression of *Plasmodium falciparum var* genes correlates with changes in antigenic and cytoadherent phenotypes of infected erythrocytes. *Cell*, 82 :101–110, 1995.

[188] R. R. Sokal and F. J. Rohlf. *Biometry*. W. H. Freeman and Company, New-York, 1981.

[189] L. M. Sompayrac. *How the immune system works.* Blackwell Science, Oxford, 1999.

[190] A. E. Souza, S. Waugh, G. H. Coombs, and J. C. Mottram. Characterization of a multicopy gene for a major stage-specific proteinase of *Leishmania mexicana. FEBS Letters*, 311 :124–127, 1992.

[191] S. E. Strauss, J. K. Ostrove, G. Inchauspé, J. M. Felser, A. Freifeld, K. D. Croen, and M. H. Sawyer. Varicella-zoster virus infections. *Annals of Internal Medicine*, 108 :221–237, 1988.

[192] Y. Suzuki and T. Gojobori. A method for detecting positive selection at single amino acid sites. *Molecular Biology and Evolution*, 16 :1315–1328, 1999.

[193] Y. Suzuki and M. Nei. Reliabilities of parsimony-based and likelihood-based methods for detecting selection at single amino acid sites. *Molecular Biology and Evolution*, 18 :2179–2185, 2001.

[194] Y. Suzuki and M. Nei. Simulation study of the reliability and robustness of the statistical methods for detecting positive selection at single amino acid sites. *Molecular Biology and Evolution*, 19 :1865–1869, 2002.

[195] Y. Suzuki and M. Nei. False-positive selection identified by ML-based methods : examples from the *Sig1* gene of the diatom *Thalassiosira weissflogii* and the *tax* gene of a human T-cell lymphotropic virus. *Molecular Biology and Evolution*, 21 :914–921, 2004.

[196] D. L. Swofford. *PAUP* : phylogenetic analysis using parsimony (* and other methods). Version 4.0b6.* Sinauer Associates, Sunderland, 2000.

[197] F. Tajima. Evolutionary relationship of dna sequences in finite populations. *Genetics*, 105 :437–460, 1983.

[198] F. Tajima. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123 :585–595, 1989.

[199] R. Takallapally, P. Rosé, S. Vasil, S. Pillai, and C. Kuiken. Reagents for HIV/SIV vaccine studies. In C. L. Kuiken, B. Foley, B. Hahn, B. Korber, F. McCutchan, P. A. Marx, J. W. Mellors, J. I. Mullins, J. Sodroski, and S. Wolinksy, editors, *Human Retroviruses and AIDS : A Compilation and Analysis of Nucleic Acid and Amino Acid Sequences*, pages 506–516. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, 1999.

[200] F. Thomas, A. T. Teriokhin, E. V. Budilova, S. P. Brown, F. Renaud, and J. F. Guégan. Human birthweight evolution across contrasting environments. *Journal of Evolutionary Biology*, 17 :542–553, 2004.

[201] J. D. Thompson, T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins. The clustalX windows interface : flexible strategies for multiple sequences alignment aided by quality analysis tools. *Nucleic Acids Research*, 25 :4876–4882, 1997.

[202] J. Travis, J. Potempa, and H. Maeda. Are bacterial proteinases pathogenic factors ? *Trends in Microbiology*, 3 :405–407, 1995.

[203] H. C. Tuckwell, L. Toubiana, and J. F. Vibert. Enhancement of epidemic spread by noise and stochastic resonance in spatial network models with viral dynamics. *Physical Review E*, 61 :5611–5619, 2000.

[204] S. S. Twiddy, C. H. Woelk, and E. C. Holmes. Phylogenetic evidence for adaptive evolution of dengue viruses in nature. *Journal of General Virology*, 83 :1679–1689, 2002.

[205] A. J. Valleron and P. Garnerin. Computer networking as a tool for public health surveillance : the French experience. *Morbidity and Mortality Weekly Report*, 41 :101–110, 1993.

[206] A. J. Valleron and P. Garnerin. Computerised surveillance of communicable diseases in France. *CDR Review*, 3 :82–87, 1994.

[207] C. Viboud. *Prédiction épidémiologiques de la grippe en zones tempérées.* PhD thesis, 2003.

[208] L. Wasserman. Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 44 :92–107, 2000.

[209] G. Watterson. On the number of segregating sites. *Theoretical Population Biology*, 7 :256–276, 1975.

[210] G. A. Watterson. Heterosis or neutrality ? *Genetics*, 85 :789–814, 1977.

[211] M. L. Wayne and K. L. Simonsen. Statistical tests of neutrality in the age of weak selection. *Trends in Ecology and Evolution*, 13 :6, 1998.

[212] X. Wei, J. M. Decker, S. Wang, H. Hui, J. C. Kappes, X. Wu, J. F. Salazar-Gonzalez, M. G. Salazar, J. M. Kilby, M. S. Saag, N. L. Komarova, M. A. Nowak, B. H. Hahn, P. D. Kwong, and G. M. Shaw. Antibody neutralization and escape by HIV-1. *Nature*, 422 :307–312, 2003.

[213] R. A. Weiss. Gulliver's travel in HIVland. *Nature*, 410 :963–967, 2001.

[214] WHO. *Control of leishmaniasis.* World Health Organ Tech Rep Parasitol, Genève, 1990.

[215] M. J. Wood. History of varicella zoster virus. *Herpes*, 7 :60–65, 2000.

[216] Y. Yamaguchi-Kabata and T. Gojobori. Reevaluation of amino acid variability of the human immunodeficiency virus type 1 *gp*120 envelope

glycoprotein and prediction of new discontinuous epidtopes. *Journal of Virology*, 74 :4335–4350, 2000.

[217] Z. Yang. PAML : a program package for the phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences*, 13 :555–556, 1997.

[218] Z. Yang. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Molecular Biology and Evolution*, 15 :568–573, 1998.

[219] Z. Yang. Adaptive molecular evolution. In D. Balding, M. Bishop, and C. Cannings, editors, *Handbook of Statistical Genetics*, pages 237–350. John Wiley & Sons, Chichester, 2001.

[220] Z. Yang. Inference of selection from multiple species alignments. *Current Opinion in Genetics and Development*, 12 :688–694, 2002.

[221] Z. Yang. Inference of selection from multiple species alignments. *Current Opinion in Genetics and Development*, 12 :688–694, 2002.

[222] Z. Yang and J. P. Bielawski. Statistical methods for detecting molecular adaptation. *Trends in Ecology and Evolution*, 15 :496–503, 2000.

[223] Z. Yang and R. Nielsen. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular Biology and Evolution*, 17 :32–43, 2000.

[224] Z. Yang and Nielsen R. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *Journal of Molecular Evolution*, 46 :409–418, 1998.

[225] Z. H. Yang, R. Nielsen, N. Goldman, and A. M. K. Pedersen. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 155 :431–449, 2000.

[226] K. Yusim, C. Kesmir, B. Gaschen, M. M. Addo, M. Altfeld, S. Brunak, A. Chigaev, V. Detours, and B. T. Korber. Clustering patterns of cytotoxic T-lymphocyte epitopes in human immunodeficiency virus type 1 (HIV-1) proteins reveal imprints of immune evasion on HIV-1 global variation. *Journal of Virology*, 76 :2523–2530, 2000.

[227] J. Zhang. Performance of likelihood ratio tests of evolutionary hypothesis under inadequate substitution models. *Molecular Biology and Evolution*, 16 :868–875, 1999.

[228] J. Zhang. Evolution by gene duplication : an update. *Trends in Ecology and Evolution*, 18 :292–298, 2003.

[229] J. Zhang, H. F. Rosenberg, and M. Nei. Positive darwinian selection after gene duplication in primate ribonuclease genes. *Proceedings of the National Academy of Sciences of the USA*, 95 :3708–3713, 1998.

ANNEXES

Annexe A

CHOISY M., WOELK C.H., GUÉGAN J.-F. & ROBERTSON D. (2004) Comparative study of adaptive molecular evolution in different HIV clades. *Journal of Virology* **78**(4) : 1962-1970

JOURNAL OF VIROLOGY, Feb. 2004, p. 1962–1970 Vol. 78, No. 4
0022-538X/04/$08.00+0 DOI: 10.1128/JVI.78.4.1962–1970.2004

Comparative Study of Adaptive Molecular Evolution in Different Human Immunodeficiency Virus Groups and Subtypes

Marc Choisy,[1] Christopher H. Woelk,[2] Jean-François Guégan,[1] and David L. Robertson[3]*

CEPM, UMR CNRS-IRD 9926, Montpellier, France[1]; Department of Pathology, University of California—San Diego, La Jolla, California 92093[2]; and School of Biological Sciences, University of Manchester, Manchester, United Kingdom[3]

Received 10 July 2003/Accepted 28 October 2003

Molecular adaptation, as characterized by the detection of positive selection, was quantified in a number of genes from different human immunodeficiency virus type 1 (HIV-1) group M subtypes, group O, and an HIV-2 subtype using the codon-based maximum-likelihood method of Yang and coworkers (Z. H. Yang, R. Nielsen, N. Goldman, and A. M. K. Pedersen, Genetics 155:431–449, 2000). The *env* gene was investigated further since it exhibited the strongest signal for positive selection compared to those of the other two major HIV genes (*gag* and *pol*). In order to investigate the pattern of adaptive evolution across *env*, the location and strength of positive selection in different HIV-1 sequence alignments was compared. The number of sites having a significant probability of being positively selected varied among these different alignment data sets, ranging from 25 in HIV-1 group M subtype A to 40 in HIV-1 group O. Strikingly, there was a significant tendency for positively selected sites to be located at the same position in different HIV-1 alignments, ranging from 10 to 16 shared sites for the group M intersubtype comparisons and from 6 to 8 for the group O to M comparisons, suggesting that all HIV-1 variants are subject to similar selective forces. As the host immune response is believed to be the dominant driving force of adaptive evolution in HIV, this result would suggest that the same sites are contributing to viral persistence in diverse HIV infections. Thus, the positions of the positively selected sites were investigated in reference to the inferred locations of different epitope types (antibody, T helper, and cytotoxic T lymphocytes) and the positions of N and O glycosylation sites. We found a significant tendency for positively selected sites to fall outside T-helper epitopes and for positively selected sites to be strongly associated with N glycosylation sites.

A detailed appreciation of the extremely high diversity of human immunodeficiency virus (HIV), the causative agent of AIDS, has resulted from the extensive sequencing and phylogenetic analysis of viral genes and gene fragments over the last decade and a half (12). In addition, phylogenetic analysis of HIV and related simian immunodeficiency virus (SIV) strains has revealed a relatively recent simian origin for HIV (HIV type 1 [HIV-1] and HIV-2) from SIV-infected primates (6, 8). More specifically, the origin of HIV-2 is linked to SIVsm-infected sooty mangabeys in West Africa, and the origin of HIV-1 is linked to SIVcpz-infected chimpanzees in Central Africa. In the case of HIV-1, at least three independent cross-species transmission events need to be postulated to account for the three most divergent HIV-1 lineages (designated groups M, N, and O), whereas seven independent events are required to account for the seven HIV-2 lineages (designated subtypes A to G) (8).

Within HIV-1 group M, nine major subtypes (A to D, F to H, J, and K) have been designated, as have 14 circulating recombinant forms (CRF01 to CRF14) (12, 24). Interestingly, recent studies have identified diversity within HIV-1 group O equivalent to that exhibited by group M (25, 33), despite the fact that almost all group O infections are restricted to Cameroon or to individuals with strong links to that region. Although there is phylogenetic substructure within group O phylogenies, distinct group M-like subtypes are not apparent (25). This is not too surprising, given that the prominence of the group M subtypes is strongly linked to founder events in the course of the HIV-AIDS pandemic that occurred outside the Democratic Republic Congo region (23). Analogous founder events have not occurred in the case of group O, as these types of infection have remained strongly associated with one geographic location, Cameroon. The third HIV-1 group, N, also remains restricted to Cameroonian residents, and to date only five infections have been conclusively documented (3).

The development of candidate vaccines specific to different HIV lineages (7) demands a thorough investigation of the consistency of the selective environment, which is presumed to be due primarily to the host immune responses (15, 22, 39) to divergent HIVs. Evidence for adaptive evolution has been found previously among HIV sequences from intra- and interpatient studies (4, 29, 30, 38, 40). Early studies involved the pairwise comparison of synonymous (silent, d_S) and nonsynonymous (amino acid changing, d_N) substitutions between protein-coding DNA sequences. The d_N/d_S ratio, ω, was then used to measure the difference between these two rates of substitution such that an ω value less than 1 corresponds to purifying (negative) selection, an ω value of 1 corresponds to neutral evolution (absence of selection), and an ω value greater than 1 indicates adaptive evolution (positive selection) (reviewed in reference 37). The pairwise approach to quantifying adaptive evolution assumes that all sites are prone to the same selective pressure, making such tests very conservative. In reality, positively selected sites normally occur in a background of negatively selected sites within a functional protein.

The problem of resolving positively selected sites against this background of negative selection has been solved in a maxi-

* Corresponding author. Mailing address: School of Biological Sciences, University of Manchester, 2.205 Stopford Building, Oxford Road, Manchester M13 9PT, United Kingdom. Phone: 44-161-275-5089. Fax: 44-161-275-5082. E-mail: david.robertson@man.ac.uk.

mum-likelihood (ML) and Bayesian statistical framework (for a review, see reference 37). First, the ML method determines whether positive selection is present by evaluating a series of models with or without a class of positively selected sites. Second, if the favored model includes positive selection, a Bayesian analysis assigns each amino acid site a "posterior probability" of being conserved, neutral, or positively selected.

Here, we focus on positively selected sites that were inferred by using the codon-based method (38), and we determine the extent to which their locations and the intensity of their selection overlap among different HIV lineages. We first quantified positive selection in the major HIV genes (*gag*, *pol*, and *env*) for the three HIV-1 group M subtypes (A, B, and C) and for HIV-2 subtype A. Since *env* exhibited the strongest signal for positive selection, the location of sites in *env* with a high probability of being under positive selection was compared across different HIV data sets corresponding to sequence alignments of HIV-1 group M subtypes A through D, group O, and an HIV-2 subtype. The hypothesis that phylogenetically divergent HIV lineages are subject to similar selective pressures was tested by determining whether the occurrence of positively selected sites at the same locations was statistically significant and whether the strength of selection was similar. On the assumption that sites are positively selected primarily as a consequence of pressure from the immune system (15, 22, 39), our results have some interesting consequences for vaccine design, as they suggest the possibility of cross-subtype and -group immunogenicity. We investigated whether the immune response, as represented by experimentally defined epitopes or the positions of N and O glycosylation (13, 28), could account for the observed distribution of the positively selected sites. We found a significant tendency for positively selected sites to fall outside T-helper epitope regions and for positively selected sites to be strongly associated with N glycosylation sites.

MATERIALS AND METHODS

Data sets. The data sets used in this computer-based study each correspond to a sequence alignment for a given genomic region (*gag*, *pol*, or *env*) and HIV group or subtype. A total of 22 data sets were analyzed and named A through V (Table 1) for convenience. Most of the data sets were retrieved as an alignment of sequences from the 2000 release of the Los Alamos National Laboratory HIV Sequence Database (12), except for the group O sequences composing data set M, which was retrieved directly from GenBank (33) and aligned with CLUSTALW (http://www.ebi.ac.uk/clustalw). Known intersubtype recombinants, gap-containing sites, and stop codons were excluded (17) from each data set. Moreover, since the models used for positive selection analysis are codon based and assume that a synonymous substitution is always synonymous, all portions of the data set consisting of overlapping reading frames were excluded. The 22 data sets used in this study (Table 1) are the data sets for which enough sequences and sites were available for effective selection analysis (1, 2).

Selection analyses. Positive selection analysis was performed on each of the 22 data sets in Table 1. For each data set, the PAUP* package (27) was first used to build an ML tree for selection analysis using the HKY85+Γ model of nucleotide substitution with optimal values for the T_S/T_V rate ratio and the shape parameter (α) of a gamma distribution (with eight categories) of rate variation among sites, both determined during tree construction. The ML method of Yang and coworkers (38) utilized codon-based models that incorporate statistical distributions to account for variable ω ratios among codons. Efficient determination of sites under positive selection requires implementation of only six models of codon substitution (M0, M1, M2, M3, M7, and M8) out of the original 14 models (for further details, see reference 38 and http://www.bioinf.man.ac.uk/~robertson/supplementary-material [appendix A]). Briefly, null models M0, M1, and M7 do not allow for the existence of positively selected sites because ω ratios are fixed or estimated between the bounds 0 and 1, whereas models M2, M3, and M8 account for positive selection by using parameters that estimate ω to be greater than 1. The significance of positive selection can be confirmed with a likelihood ratio test (LRT) between null models and those able to account for positive selection. An LRT is performed by taking twice the difference in log likelihood between nested models and comparing the result to a χ^2 distribution with degrees of freedom equivalent to the difference in the number of parameters between the models. Models M0 and M1 are both nested with M2 and M3, M2 is nested with M3, and M7 is nested with M8. All the model comparisons (M0 versus M2, M1 versus M2, M0 versus M3, M1 versus M3, M2 versus M3, and M7 versus M8) gave similar results, and for the sake of simplicity we focus on the results of models M7 and M8. M7 uses a discrete (10 classes) beta distribution to model sites with ω ratios between the bounds 0 and 1. For each class i ($1 \leq i \leq 10$) of the beta distribution, the value of the ω_i ratio and the proportion (p_i) of sites belonging to this class are estimated by maximizing the likelihood. M8 adds two additional parameters to model M7 such that p_{11} can account for a positively selected class of sites where ω_{11} is not constrained by the beta distribution and is allowed to be greater than 1. Once positively selected sites have been shown to exist, i.e., if model M7 is rejected in favor of M8 by the LRT, a Bayesian approach (for which the p_1 to p_{11} values are used as a prior distribution) is used to infer the posterior probability that site i belongs to one of the 11 ω classes: $f_1^i, f_2^i, \ldots, f_{11}^i$. Models were implemented using the CODEML program of the PAML package, version 3.1 (36).

Statistical analysis of sites identified as positively selected. A "shared-position" statistic and Monte Carlo simulations were used to test whether putative positively selected sites (defined as those having a p_{11} value of greater than 0.95 when ω_{11} is greater than 1 for model M8) tend to occur at the same positions in data sets 1 to N (H_1) more often than would be expeeted by chance (H_0). The shared-position statistic used is the count of the match between the positions of positively selected sites in one data set and the positions of positively selected sites in another data set. As this test depends on the quality of the alignment among the diverse data sets, the result should be conservative.

To study the "strength" of positive selection, we defined for each site, i, the weighted mean ω value as $\bar{\omega}_i = \sum_{k=1}^{11} f_k^i \omega_k$ as previously implemented (7). For each pair of data sets, we tested whether the strength of positive selection was significantly different (H_1), as opposed to being equivalent (H_0), by using a paired Wilcoxon rank sum test with a continuity correction applied to the normal approximation for the P values (26). Only shared sites having a weighted mean

TABLE 1. Data sets used in this study[a]

Data set	Lineage	No. of seq.	No. of codons	Gene	Source
A	HIV-1 M:A	11	404	*gag*	LANL
B	HIV-1 M:B	35	425	*gag*	LANL
C	HIV-1 M:C	17	418	*gag*	LANL
D	HIV-2 A	12	386	*gag*	LANL
E	HIV-1 M:A	13	838	*pol*	LANL
F	HIV-1 M:B	33	913	*pol*	LANL
G	HIV-1 M:C	16	911	*pol*	LANL
H	HIV-2 A	12	916	*pol*	LANL
I	HIV-1 M:A	16	578	*env*	LANL
J	HIV-1 M:B	30	578	*env*	LANL
K	HIV-1 M:C	30	578	*env*	LANL
L	HIV-1 M:D	15	578	*env*	LANL
M	HIV-1 O	30	621	*env*	GenBank
N	HIV-2 A	22	679	*env*	LANL
O	HIV-1 M:A	20	415	*env-gp120*	LANL
P	HIV-1 M:B	20	433	*env-gp120*	LANL
Q	HIV-1 M:C	20	423	*env-gp120*	LANL
R	HIV-2 A	20	460	*env-gp120*	LANL
S	HIV-1 M:A	19	232	*env-gp41*	LANL
T	HIV-1 M:B	30	233	*env-gp41*	LANL
U	HIV-1 M:C	30	237	*env-gp41*	LANL
V	HIV-2 A	22	193	*env-gp41*	LANL

[a] Each data set is an alignment of nucleotide sequences of a given HIV subtype or group and a given gene. The number of sequences (No. of seq.) and sites (No. of codons) in each alignment are indicated as well as the source: the 2000 release of the Los Alamos National Laboratory (LANL) HIV Sequence Database (12) and GenBank (33). Positive selection was analyzed for each of the data sets. Statistical analyses on the positively selected sites were performed for the *env* data sets (I to N).

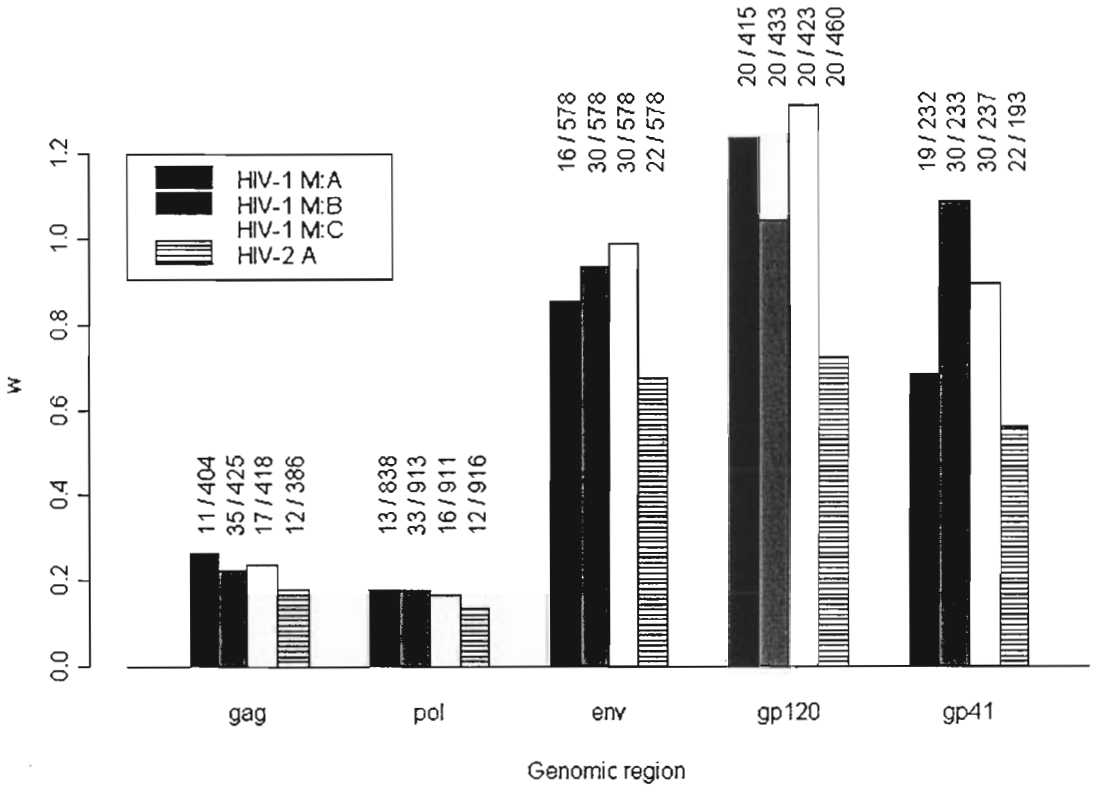


FIG. 1. Mean ω ratios in *gag*, *pol*, *env*, *env-gp120* and *env-gp41* for HIV-1 group M subtypes A, B, and C, and HIV-2 subtype A (data sets A to K and N to V in Table 1). The mean ω ratios are calculated by averaging the results over all of the sites and are obtained from model M0. The numbers above the bars indicate the number of sequences and the number of codons in each data set. For example, "11/404" above the first *gag* bar indicates that there were 11 sequences and 404 codons in the *gag* HIV-1 group M subtype A data set (called data set A in Table 1).

ω value greater than 1 in the two data sets being compared were included. Note that the positively selected sites with a weighted ω value greater than 1 are not necessarily identified as positively selected by model M8 at the 95% level. The latter sites identified at the 95% level by M8 will be a subset of the former weighted sites. The paired Wilcoxon rank sum test was repeated only for those shared sites identified by M8 at the 95% level.

Finally, Monte Carlo simulations were again used to test a null hypothesis (H_0) that sites of positive selection are not associated with the positions of epitope regions, or sites of glycosylation, against the alternative hypotheses (H_1) that the positively selected sites are associated with the location of the epitope regions (or various combinations of the three types of regions) or the positions of the glycosylation sites in the different data sets. An additional hypothesis (H_2) that the positive selected sites tend to fall outside the defined epitope regions (or various combinations of the three types of regions) was also tested against H_0. The epitope regions are experimentally defined and correspond to antibody (Ab), cytotoxic T-cell (CTL), and helper T-cell immune response data available from the Los Alamos National Laboratory HIV Immunology Database (11). As the majority of epitope mapping has focused on subtype B-infected individuals (11), only the positively selected sites identified in data set J were tested. For each data set, the positions of the N and O glycosylation sites were predicted using the NetNGlyc (R. Gupta, E. Jung, and S. Brunak, unpublished data) and NetOGlyc (9) programs, respectively.

For all Monte Carlo simulations, 9,999 repetitions proved to be enough to reach an asymptotic state. The programs used to implement the Monte Carlo simulations are available upon request from M. Choisy.

RESULTS

Mean ω values for *gag*, *pol*, and *env*. The results for the mean ω values (assuming the same value for ω at all sites) for the genes *gag*, *pol*, and *env*, and for the individual subunits of *env* (gp120 and gp41), are shown for HIV-1 group M subtypes A, B, and C and for HIV-2 subtype A in Fig. 1. Except for the group M subtype A, B, and C results for gp120 and subtype B for gp41, all ω values are less than 1, indicating that the majority of sites are subject to purifying selection. The effect of purifying selection is particularly strong in the *gag* and *pol* genes but is much weaker in the envelope region, which is not surprising given that *env* codes for the envelope surface proteins, which are the most exposed to the immune system. Note that despite the low mean ω values in the *gag* and *pol* genes, positive selection can still occur at a minority of sites, but this signal can be averaged out by M0 and pairwise methods. For example, others have previously found a comparable ω value (0.196) for the *pol* gene of a subtype B alignment as well as strong evidence for adaptive evolution (38). The contrast in mean ω ratios between *gag* and *pol* compared to that of the *env* regions indicates that the *env* region contains more positively selected sites than do the other genes. Within the *env* region, positive selection appears to be particularly strongly associated with the gp120 subunit, coding for the extramembrane envelope protein.

Identification of positively selected sites across *env*. A comparative analysis of HIV-1 group M subtypes A, B, C, and D; group O; and HIV-2 subtype A in the envelope region (data sets I to N in Table 1) was carried out in order to identify specific positively selected sites. All models that were able to

TABLE 2. Positive selection in the *env* gene[a]

Data set	Lineage	Mean ω	11th class	No. of sites	*P*
I	HIV-1 M:A	0.690	4.702	33	**<0.001**
J	HIV-1 M:B	0.623	4.009	35	**<0.001**
K	HIV-1 M:C	0.610	4.463	33	**<0.001**
L	HIV-1 M:D	0.568	3.821	30	**<0.001**
M	HIV-1 O	0.590	3.992	40	**<0.001**
N	HIV-2 A	0.444	3.568	25	**<0.001**

[a] Mean ω was calculated by averaging over all the sites. The 11th class is from model M8, and the number of sites refers to those found to be under positive selection over the 95% level. *P* is the probability resulting from the likelihood ratio test between M7 and M8. Significant results ($P < 0.05$) are indicated in boldface type.

detect positive selection (M2, M3, and M8) identified a positively selected class ($\omega > 1$) and rejected those models that were unable to account for positive selection (M0, M1, and M7). For the sake of clarity, only results for M8 are presented in Table 2 (results for the other models are available from http://www.bioinf.man.ac.uk/~robertson/supplementary-material (appendixes B and C). M2 and M8 identified the same positively selected sites when taking posterior probabilities greater than the 95% level using the Bayesian approach. M3 identified all of the sites identified by M2 and M8 and several more. We consider the sites identified by M8 only, as M3 has the potential to overestimate the number of positively selected sites (2, 38). The number of positively selected sites identified by model M8 was 22 for HIV-2 subtype A, between 30 and 35 for HIV-1 M subtypes, and 40 for HIV-1 O (Table 2). Figure 2 shows the location of these putative positively selected sites across the multiple alignment of data sets I to N. Positively selected sites are not restricted to the variable regions (V1 to V5) of *env*, a finding that supports previous work that used a maximum-parsimony-based method to identify amino acid sites that were potentially under the influence of positive selection in an HIV-1 subtype B alignment of sequences (35).

Comparison of the locations of positively selected sites. The null hypothesis that there is no association of the position of sites of positive selection among data sets I to N was rejected by using the shared-position statistic and Monte Carlo simulations for the majority of pairwise comparisons (Table 3). The exception was HIV-2 subtype A, which showed only a significant association with the position of positively selected sites with the HIV-1 group M subtype A data set. This result indicates that different HIV-1 group M subtypes and group O contain sites in *env* that are under similar selective pressures.

Comparison of the strength of positive selection. The result just described seemingly contradicts the finding by Gaschen and coworkers (7) that group M subtypes B and C undergo different evolutionary pressures in the C2V3 region of *env*; this result is presumed to be due to different antigenic exposure patterns being exhibited by different subtypes. To investigate this possibility further, we plotted for each of the I to N data sets the weighted ω ratio for each site (see Materials and Methods) that had a value greater than 1 (Fig. 3). When sites that had a weighted ω value greater than 1 were tested among different data sets with a paired Wilcoxon ranked sum test (Table 4), the comparison was significant for the strength of selection differing between HIV-1 subtypes B and C, a result that is in agreement with that of Gaschen and coworkers (7). This was also the case for the comparison between HIV-1 subtypes A and B and between HIV-1 group O and HIV-2 subtype A. However, no other comparisons were significant, indicating that generalizations about differences in the strength of selection between diverse HIV data sets should not be made based on the available data. For the subset of sites with a weighted ω value greater than 1 and that were identified as positively selected by model M8 at the 95% level, the comparisons between HIV-1 subtypes A and B, B and C, and between HIV-1 group O and HIV-2 were significant ($P = 0.0001$, 0.0282, and 0.0156, respectively; data available at http://www.bioinf.man.ac.uk/~robertson/supplementary-material [appendix D]).

Association of positively selected sites with epitope regions and glycosylation sites. None of the Monte Carlo tests used to investigate whether the positively selected sites of the *env* gene of subtype B (data set J in Table 1) had a tendency to be associated with experimentally defined Ab, CTL, T-helper epitopes, or combinations of these were significant (Table 5). However, the reciprocal investigation, which tested whether positively selected sites had a tendency to fall between the different epitope regions, was significant for the T-helper and CTL–T-helper combination ($P < 0.05$), while the significance of CTL alone was marginal ($P = 0.053$) (Table 5).

For the test of the association of N glycosylation sites identified in each HIV-1 data set (ranging from 22 to 39) with the identified positively selected sites, a significant association ($P < 0.05$) was found for all comparisons (Table 6). Note that N glycosylation sites were not identified in the HIV-2 data set. Between two and six of the N glycosylation sites were conserved in all sequences of the HIV-1 data sets. The number of O glycosylation sites for the HIV-1 M:A, HIV-1 M:B, HIV-1 M:C, HIV-1 M:D, group O, and HIV-2 A data sets was 2, 2, 4, 1, 8, and 0, respectively. No associations ($P \geq 0.05$) were found for the test of the association of O glycosylation sites identified in each HIV-1 data set with the putative positively selected sites (results not shown).

Finally, the locations of sites implicated in the binding of the envelope glycoprotein to the CD4 receptor molecules (18, 32) and of sites implicated in the receptor switch from CCR5 to CxCR4 tropism (20) are indicated in Fig. 2, and neither associates with any of the positively selected sites identified. The finding that sites involved in chemokine binding are apparently not under the influence of positive selection may be due to our comparison of viral sequence data from several different infected individuals rather than from the viral population of one individual, while CD4 binding sites are presumably under the influence of purifying selection.

DISCUSSION

As vaccine candidates are being designed to target different HIV-1 group M subtypes, it is important to investigate how the immune system responds to different HIV-1 strains. Assuming that the immune response is providing the evolutionary pressure for the majority of adaptive evolution observed in the HIV genome (15, 22, 39), we have quantified positive selection in HIV-1 group O, different group M subtypes, and HIV-2 subtype A sequence alignments. The majority of positive selection

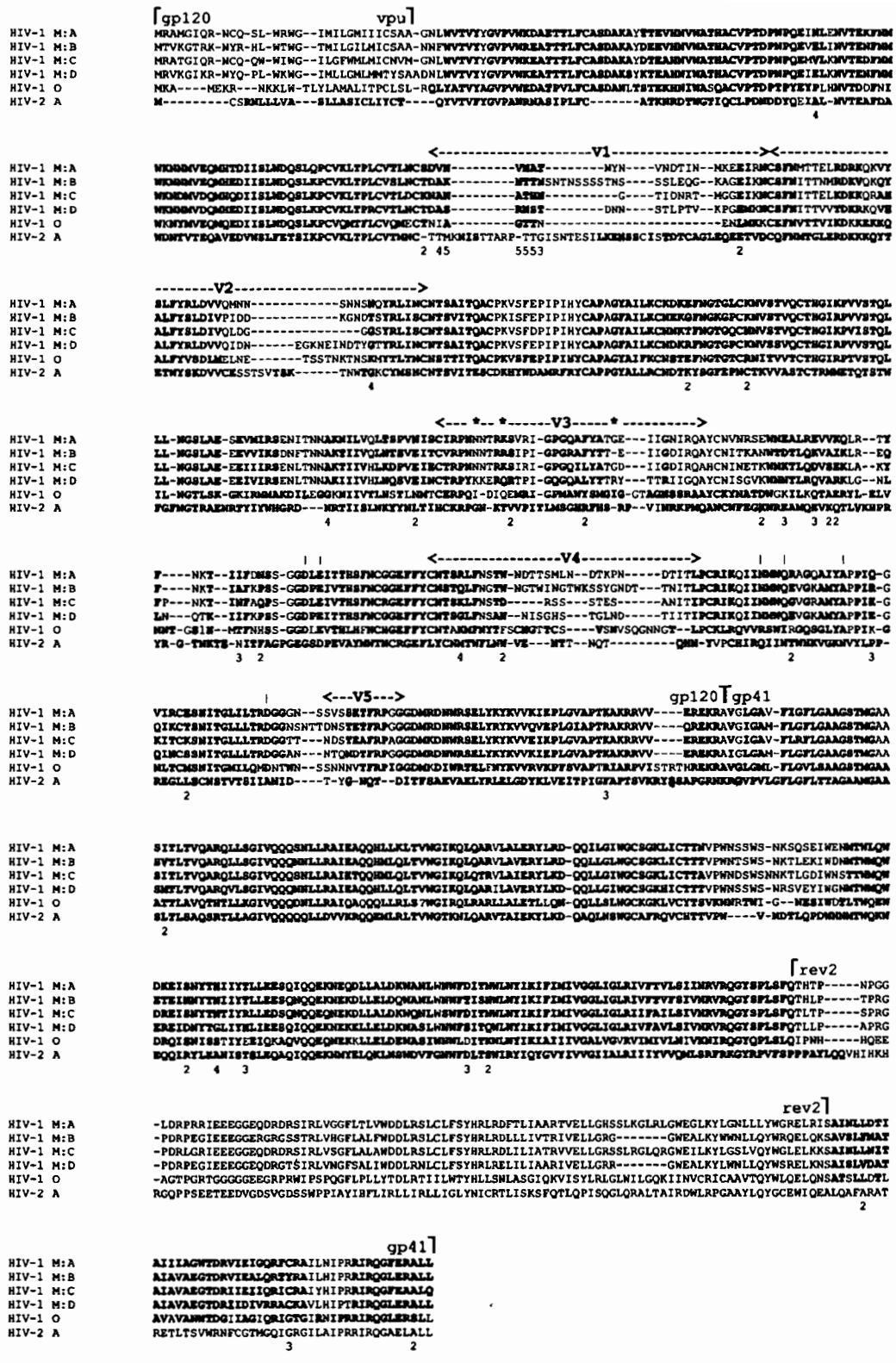


FIG. 2. Positions of positively selected sites across *env* for HIV-1 group M subtypes A, B, C, and D; group O; and HIV-2 subtype A (data sets I to N in Table 1). Each data set analyzed by CODEML is represented by one sequence, with the sites included in the analysis indicated with boldface type. Sites identified as being positively selected with a posterior probability of more than 95% are shaded. Notations above the sequences divide *env* into the gp120 and gp41 subunits and show the position of *vpu* and the second *rev* exon with the beginning and end of regions. The positions of the five variable regions V1 to V5 are indicated. Sites critical for CD4 binding are identified by a vertical bar, and sites implicated in the CXCR4 to CCR5 receptor switch are indicated with an * above the sequences. The numbers 2, 3, and 4 below the sequences indicate the number of data sets for which positive selection was identified at that site. The representative sequences for HIV-1 group M subtypes A, B, C, and D; group O; and HIV-2 subtype A are MA246, MBC18, BU910112, 84ZR085, ANT70, and CBL21, respectively.

TABLE 3. Monte Carlo simulations testing the association of sites of positive selection between data sets[a]

Data set	Lineage and value type	Values for data set and lineage				
		I	J	K	L	M
		HIV-1 M:A	HIV-1 M:B	HIV-1 M:C	HIV-1 M:D	HIV-1 O
J	HIV-1 M:B					
	E	2.018				
	O	13				
	P	**0.001**				
K	HIV-1 M:C					
	E	1.990	2.015			
	O	16	15			
	P	**0.001**	**0.001**			
L	HIV-1 M:D					
	E	1.760	1.793	1.741		
	O	10	14	15		
	P	**0.001**	**0.001**	**0.001**		
M	HIV O					
	E	1.016	0.996	0.889	0.848	
	O	7	7	8	6	
	P	**0.001**	**0.001**	**0.001**	**0.001**	
N	HIV-2 A					
	E	0.633	0.722	0.571	0.511	0.729
	O	3	1	2	2	1
	P	**0.024**	0.535	0.104	0.091	0.539

[a] E, expected value from a random distribution, O, observed value; and *P*, level of significance at which E is different from O. Significant results ($P < 0.05$) are indicated in boldface type.

was found to occur in the envelope region of the genome as opposed to the *gag* or *pol* (Fig. 1) region, thereby confirming the results of previous studies (4, 30, 34, 35).

Further analysis of *env* revealed that a proportion of the sites that were identified as positively selected (ranging from 25 in the HIV-1 M group subtype A data set to 40 in the HIV-1 group O data set) were at the same positions in the different data sets (Fig. 2). We believe that only the immune response could be driving this propensity of HIV to exhibit adaptive molecular evolution to such an extent. Furthermore, for the HIV-1 group M comparisons, between 10 and 16 sites were shared depending on the subtypes compared (Table 3). On the assumption that the immune response provides the evolutionary pressure for amino acid change at these sites (15, 22, 39), the finding that positively selected sites are shared between divergent HIV-1 lineages suggests that the immune response may be targeting the same viral regions in the different groups and subtypes, thus raising the possibility of cross-subtype or -group immunogenicity.

However, it has been reported previously (7) that the strength of selection at positively selected sites in the C2V3 region of group M subtypes B and C is different. We made the same observation here, not only for the C2V3 region but also for the entire *env* gene, and we moreover show that the finding is statistically significant (Table 4). Importantly, we found that (i) there is a tendency for the position of positively selected sites to be correlated among different HIV-1 data sets and that (ii) there can be, at the same time, a difference in the strength of selection for some comparisons; these findings are not mutually exclusive. This is true because selection may be acting at the same sites but to differing extents, or, alternatively, the different strengths of selection might be predominantly at the sites that are not correlated among the different data sets. Nevertheless, as most comparisons of the strength of selection are not significant, generalizations about differences among diverse HIV data sets should not be made based on the available data. Indeed, differences in the strength of selection may be due to other factors such as the predominant form of transmission in a given subtype or the amount of diversity in that subtype.

To explicitly test the assumption that the majority of adaptive evolution observed in the HIV envelope is due to the immune response, we then investigated the potential of the different types of immune response (Ab, CTL, and T helper) to account for the location of the positively selected sites. This comparison is relatively crude because the epitopes that can be recognized in different individuals and their frequencies in different populations will vary. Ideally, viral sequences would be analyzed that are from a distinct population in conjunction with information concerning the types of epitope that could be recognized, as has been done for comparisons between HLA haplotypes and polymorphisms present in viral sequences (15). Also, some of the positively selected sites may be relevant to nonlinear epitopes, which are difficult to detect since they are formed by protein tertiary structure bringing distant sites into proximity. For example, a "glycan shield" model (28) has been proposed, which suggests that linear epitopes in regions essential for viral fitness and that are unable to tolerate mutation can be protected from neutralizing antibodies by the bound carbohydrates. Mutations of gp120 would cause the permanent rearrangement of the carbohydrates, thus creating a moving protective shield around epitopes that are unable to tolerate

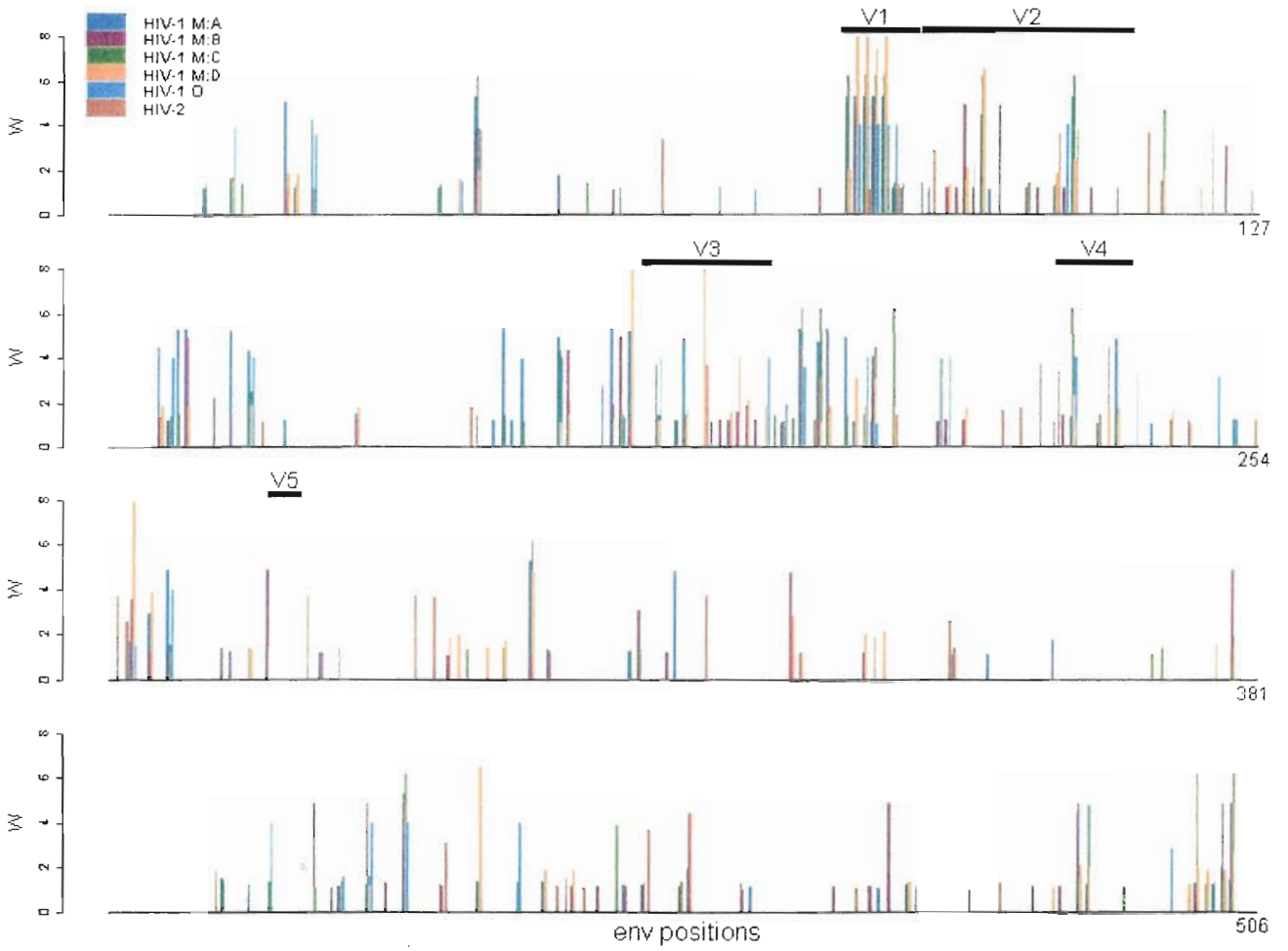


FIG. 3. The weighted mean ω ratio greater than 1 at each codon position in the *env* data sets comparing HIV-1 group M subtypes A, B, C, and D; group O; and HIV-2 subtype A (data sets I to N in Table 1). The weighted mean ω value for each site is calculated by multiplying ω by the posterior probability for each class under M8 and summing the results (see Materials and Methods). The positions of the five variable regions V1 to V5 are indicated.

mutation, as they are in functionally conserved regions. Despite these limitations, a significant result was found (Table 5) for the tendency of the T-helper and possibly for the CTL epitope regions to not include positively selected sites. This result might be explained by a finding that CTL epitopes are more concentrated in relatively conserved regions across the HIV genome, whereas positively selected sites will have a tendency to be detected in the more variable regions (39). Alternatively, positively selected sites may correspond to proteolytic cleavage sites such that mutation in the epitope-flanking residues alters intracellular processing, thereby permitting CTL escape (39).

We also investigated the predicted positions of the N glycosylation sites with respect to the positions of the identified positively selected sites. The significance of N glycosylation sites is that they allow the binding of carbohydrates to the viral envelope to mask viral protein epitopes from the immune response (5, 19, 28, 31). The bound carbohydrates are large molecules contributing to half of the molecular mass of gp120 (5) and that are linked to the gp120 protein on N glycosylation sites, and, to a lesser extent, sites of O glycosylation. They are thought to play an important role for the stability of the gp120 molecule (19), for CCR5 and CXCR4 coreceptor utilization (21), and for escape from the immune defense (5, 10). The relatively rapid turnover of mutations of the gp120 protein may also induce continual conformational changes, and such a constantly moving structure may help to distort epitopes and prevent antibody binding (13). In accordance with previous reports (5), between 22 and 39 N glycosylation sites were predicted (Table 6) for the different HIV-1 data sets. When the N glycosylation sites present in the HIV-1 data sets were considered, Monte Carlo simulations indicated that these sites are significantly associated with putative positively selected sites. These findings of a correlation among positively selected sites but not of the location of Ab epitope regions is consistent with the glycan shield model of viral escape (28). Interestingly, no N glycosylation sites were detected in the HIV-2 data set, despite glycosylation for HIV-2 being previously reported (14). This finding seems to reflect the very low number of such sites in HIV-2 strains. HIV-2 is apparently less virulent than HIV-1 (8), possibly due to HIV-2 being less antigenic, thus accounting for the lack of N glycosylation sites.

TABLE 4. Paired Wilcoxon ranked sum test to determine differences in the strength of positive selection between data sets[a]

Data set	Lineage	Values for data set and lineage				
		I	J	K	L	M
		HIV-1 M:A	HIV-1 M:B	HIV-1 M.C	HIV-1 M.D	HIV-1 O
J	HIV-1 M:B					
	Z	3.8266				
	N	69				
	P	**0.0001**				
K	HIV-1 M:C					
	Z	1.2150	−2.1945			
	N	67	62			
	P	0.2244	**0.0282**			
L	HIV-1 M:D					
	Z	0.6009	−1.1021	−0.4077		
	N	46	54	51		
	P	0.5479	0.2704	0.6835		
M	HIV-1 O					
	Z	0.3652	−1.853	−1.0934	0.0000	
	N	23	22	26	18	
	P	0.7149	0.0639	0.2742	1.0000	
N	HIV-2 A					
	Z	−0.4001	−1.0193	1.8347	0.8293	2.2819
	N	11	10	10	9	7
	P	0.6891	0.3081	0.0665	0.4069	**0.0225**

[a] Z is the statistic and N is the number of sites with ω greater than one. Significant differences ($P < 0.05$) are indicated in boldface type. A continuity correction is applied to the normal approximation for the *P* values

In our opinion, potential correlations of adaptive molecular evolution among divergent HIVs, such as those we have detected here, warrant further investigation, as they are indicative of possible shared antigenicity. Given the unquestionable need for an HIV-AIDS vaccine to elicit an immune response against multiple group M subtypes, specifically in Africa where multiple subtypes frequently cocirculate, a hypothetical vaccine cocktail that would include antigens from a number of genomic regions is clearly worth investigating. The present preoccupation with consensus and ancestral sequences targeted at an individual subtype as optimal immunogens (for an example, see references 7 and 16) make limited or no specific attempts to elicit immune responses that may be cross-reactive to different subtypes. In addition, the most antigenic viral regions will be embedded in sequence of differing immunogenic potential. Thus, there is a possibility that constructing a consensus sequence from the genetic material of circulating viruses would result in the least optimal antigenic regions being included in the vaccine. This result would occur because the consensus sequence would represent optimal genetic material

TABLE 5. Correlation of positively selected sites with epitopes in the *env* gene of HIV-1 group M subtype B (data set J)[a]

Epitope(s)	N_{IN}[b]	O_{IN}[c]	E_{IN}[d]	P_{IN}[e]	N_{OUT}[f]	O_{OUT}[g]	E_{OUT}[h]	P_{OUT}[i]
Ab	370	18	22.17	0.946	208	17	12 53	0.072
CTL	394	19	23.82	0.976	184	16	11.12	0.053
Th	499	26	30.16	0 989	79	9	4.78	**0.028**
Ab and CTL	507	30	30.64	0.726	71	5	4.17	0.401
Ab and Th	537	33	32.40	0.496	41	2	1 90	0 612
CTL and Th	524	27	31 67	0 999	54	8	2 92	**0.018**

[a] The first three rows correspond to the three epitope types analyzed separately (Ab, antibody; CTL, cytotoxic T-cell and Th, T-helper responses), and the remaining rows refer to combinations of these epitope types analyzed together.
[b] Number of sites targeted by epitopes from the HIV Immunology Database.
[c] Observed number of identified positively selected sites that fall inside the epitope regions.
[d] Expected number of positively selected sites in the epitope regions as calculated by Monte Carlo simulations.
[e] Significance level at which O_{IN} differs from E_{IN}.
[f] Number of sites that have not been identified in the HIV Immunology Database to be targeted by an epitope.
[g] Observed number of identified positively selected sites that fall outside the epitope regions.
[h] Expected number of positively selected sites that fall outside the epitope region as calculated by Monte Carlo simulations.
[i] Significance level at which O_{OUT} differs from E_{OUT}, significant values ($P < 0.05$) are indicated in boldface type

TABLE 6. Association of positively selected sites with sites of N glycosylation in *env*

Data set	Lineage	No of sites of N glyc[a]	No. of conserved N glyc[b]	No observed[c]	No. expected[c]	*P* value[d]
I	HIV-1 M:A	28	5	11	1.64	**0.001**
J	HIV-1 M:B	27	2	5	1.62	**0.019**
K	HIV-1 M:C	30	2	7	1.69	**0.002**
L	HIV-1 M:D	22	5	4	1.17	**0.023**
M	HIV-1 O	39	6	13	2.46	**0.001**

[a] Total number of N glycosylation sites in the data set.
[b] Number of N glycosylation sites that are conserved across all sequences of the data set
[c] The observed number of associations between positively selected sites and sites of N glycosylation is compared to the expected number of associations between positively selected sites and those of N glycosylation (as calculated from the mean of the Monte Carlo simulated distribution).
[d] Significant results ($P < 0.05$) are indicated in boldface type

for immune escape because it would have sequences from multiple viruses that have successfully escaped the immune response. Furthermore, neither a consensus sequence nor a reconstructed ancestral sequence (due to ongoing recombination within individuals [with or without superinfection] and positive selection resulting in escape mutants with the same convergent amino acid changes) can represent any virus that has ever existed and so may lack important properties that could be of immunogenic importance in a potential vaccine (for example, folding). In conclusion, if we are to control HIV, we must understand its evolution and conceive appropriate intervention strategies accordingly.

ACKNOWLEDGMENTS

We thank Bénédicte Lafay, Andrew Rambaut, Simon Lovell, Jay Taylor, Mike Worobey, and Eddie Holmes for helpful comments and discussion.

We also thank the Wellcome Trust (which provided assistance through their Biodiversity program while D.L.R. was at the Department of Zoology, University of Oxford, where this work was begun), the National Institutes of Health (AIDS training grant number AI07384), and the CNRS for funding (M.C. is supported by a Bourse Docteur Ingénieur from the CNRS-Région Languedoc Roussillon).

REFERENCES

1. **Anisimova, M., J. P. Bielawski, and Z. Yang.** 2002. Accuracy and power of the Bayes prediction of amino acid sites under positive selection. Mol. Biol. Evol. **19:**950–958.
2. **Anisimova, M., J. P. Bielawski, and Z. Yang.** 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol. Biol. Evol **18:**1585–1592.
3. **Ayouba, A., S. Souquieres, B. Njinku, P. M. Martin, M. C. Muller-Trutwin, P. Roques, F. Barre-Sinoussi, P. Mauclere, F. Simon, and E. Nerrienet.** 2000. HIV-1 group N among HIV-1-seropositive individuals in Cameroon. AIDS **14:**2623–2625.
4. **Bonhoeffer, S., E. C. Holmes, and M. A. Nowak.** 1995. Causes of HIV diversity. Nature **376:**125.
5. **Botarelli, P., B. A. Houlden, N. L. Haigwood, C. Servis, D. Montagna, and S. Abrignani.** 1991. N-glycosylation of HIV-gp120 may constrain recognition by T lymphocytes. J. Immunol. **147:**3128–3132.
6. **Gao, F., E. Bailes, D. L. Robertson, Y. Chen, C. M. Rodenburg, S. F. Michael, L. B. Cummins, L. O. Arthur, M. Peeters, G. M. Shaw, P. M. Sharp, and B. H. Hahn.** 1999. Origin of HIV-1 in the chimpanzee *Pan troglodytes troglodytes.* Nature **397:**436–441
7. **Gaschen, B., J. Taylor, K. Yusim, B. Foley, F. Gao, D. Lang, V. Novitsky, B. Haynes, B. Hahn, T. Bhattacharya, and B. Korber.** 2002. Diversity considerations in HIV-1 vaccine selection. Science **296:**2354–2360.
8. **Hahn, B. H., G. M. Shaw, K. M. De Cock, and P. M. Sharp.** 2000. AIDS as a zoonosis: scientific and public health implications. Science **287:**607–614.
9. **Hansen, J. E., O. Lund, N. Tolstrup, A. A. Gooley, K. L. Williams, and S. Brunak.** 1998. NetOglyc: prediction of mucin type O-glycosylation sites based on sequence context and surface accessibility. Glycoconj. J. **15:**115–130.
10. **Huang, X. L., J. J. Barchi, F. D. T. Lung, P. P. Roller, P. L. Nara, J. Muschik, and R. R. Garrity.** 1997. Glycosylation affects both the three-dimensional structure and antibody binding properties of the HIV-1_{IIIB} BP120 peptide RP135. Biochemistry **36:**10846–10856.
11. **Korber, B., C. Brander, B. Haynes, R. Koup, C. Kuiken, J. P. Moore, B. D. Walker, and D. I. Watkins.** 2000. HIV molecular immunology. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, N.Mex
12. **Kuiken, C., B. Foley, B. Hahn, P. A. Marx, F. McCutchan, J. W. Mellors, J. L. Mullins, S. Wolinsky, and B. Korber.** 2000. HIV sequence compendium. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, N.Mex.
13. **Kwong, P. D., M. L. Doyle, D. J. Casper, C. Cicala, S. A. Leavitt, S. Majeed, T. D. Steenbeke, M. Venturi, I. Chaiken, M. Fung, H. Katinger, P. W. Parren, J. Robinson, D. Van Ryk, L. Wang, D. R. Burton, E. Freire, R. Wyatt, J. Sodroski, W. A. Hendrickson, and J. Arthos.** 2002. HIV-1 evades antibody-mediated neutralization through conformational masking of receptor-binding sites. Nature **420:**678–682.
14. **Liedtke, S., R. Geyer, and H. Geyer.** 1997. Host-cell-specific glycosylation of HIV-2 envelope glycoprotein. Glycoconj. J. **14:**785–793.
15. **Moore, C. B., M. John, I. R. James, F. T. Christiansen, C. S. Witt, and S. A. Mallal.** 2002. Evidence of HIV-1 adaptation to HLA-restricted immune responses at a population level. Science **296:**1439–1443
16. **Nickle, D. C., M. A. Jensen, G. S. Gottlieb, D. Shriner, G. H. Learn, A. G. Rodrigo, and J. I. Mullins.** 2003. Consensus and ancestral state HIV vaccines. Science **299:**1515–1518.
17. **Nielsen, R., and Z. Yang.** 1998. Likelihood models for detecting positively selected amino-acid sites and applications to the HIV-1 envelope gene. Genetics **148:**929–936.
18. **Pantophlet, R., E. Ollmann Saphire, P. Poignard, P. W. Parren, I. A. Wilson, and D. R. Burton.** 2003. Fine mapping of the interaction of neutralizing and nonneutralizing monoclonal antibodies with the CD4 binding site of human immunodeficiency virus type 1 gp120. J. Virol. 77:642–658.
19. **Papandreou, M. J., T. Idziorek, R. Miquelis, and E. Fenouillet.** 1996 Glycosylation and stability of mature HIV envelope glycoprotein conformation under various conditions FEBS Lett. **379:**171–176.
20. **Pillai, S., B. Good, D. D. Richman, and J. Corbeil.** 2003. A new perspective on V3 phenotype prediction. AIDS Res. Hum. Retrovir. **19:**145–149.
21. **Pollakis, G., S. Kang, A. Kliphuis, M. I. M. Chalaby, J. Goudsmit, and W. A. Paxton.** 2001. N-linked glycosylation of the HIV type 1 gp120 envelope glycoprotein as a major determinant of CCR5 and CXCR4 coreceptor utilization. J. Biol Chem. **276:**13433–13441.
22. **Price, D. A., P. J. R. Goulder, P. Klernerman, A. K. Sewell, P. J. Easterbrook, M. Troop, C. R. M. Bangham, and R. E. Phillips.** 1997. Positive selection of HIV-1 cytotoxic T lymphocyte escape variants during primary infection Proc Natl. Acad Sci. USA **94:**1890–1895.
23. **Rambaut, A., D. L. Robertson, O. G. Pybus, M. Peeters, and E. C. Holmes.** 2001. Phylogeny and the origin of HIV-1. Nature **410:**1047–1048.
24. **Robertson, D. L., J. P. Anderson, J. A. Bradac, J. K. Carr, B. Foley, R. K. Funkhouser, F. Gao, B. H. Hahn, M. L. Kalish, C. Kuiken, G. H. Learn, T. Leitner, F. E. McCutchan, S. Osmanov, M. Peeters, D. Pieniazek, M. Salminen, P. M. Sharp, S. Wolinsky, and B. Korber.** 2000 HIV-1 nomenclature proposal. Science **288:**55–57.
25. **Roques, P., D. L. Robertson, S. Souquiere, F. Damond, A. Ayouba, I. Farfara, C. Depienne, E. Nerrienet, D. Dormont, F. Brun-Vezinet, F. Simon, and P. Mauclere.** 2002. Phylogenetic analysis of 49 newly derived HIV-1 group O strains: high viral diversity but no group M-like subtype structure. Virology **302:**259–273.
26. **Sokal, R. R., and F. J. Rohlf.** 1981. Biometry. W. H Freeman and Company, New York, N.Y
27. **Swofford, D. L.** 2000 PAUP*: phylogenetic analysis using parsimony (* and other methods). Version 4.0b6. Sinauer Associates, Sunderland, Mass.
28. **Wei, X., J. M. Decker, S. Wang, H. Hui, J. C. Kappes, X. Wu, J. F. Salazar-Gonzalez, M. G. Salazar, J. M. Kilby, M. S. Saag, N. L. Komarova, M. A. Nowak, B. H. Hahn, P. D. Kwong, and G. M. Shaw.** 2003. Antibody neutralization and escape by HIV-1. Nature **422:**307–312.
29. **Woelk, C. H., and E. C. Holmes.** 2002. Reduced positive selection in vector-borne RNA viruses. Mol. Biol. Evol. **19:**2333–2336.
30. **Wolinsky, S. M., B. T. Korber, A. U. Neumann, M. Daniels, K. J. Kunstman, A. J. Whetsell, M. R. Furtado, Y. Cao, D. D. Ho, and J. T. Safrit.** 1996. Adaptive evolution of human immunodeficiency virus-type 1 during the natural course of infection. Science **272:**537–542
31. **Wu, L., N. P. Gerard, R. Wyatt, H. Choe, C. Parolin, N. Ruffing, A. Borsetti, A. A. Cardoso, E. Desjardin, W. Newman, C. Gerard, and J. Sodroski.** 1996. CD4-induced interaction of primary HIV-1 gp120 glycoproteins with the chemokine receptor CCR-5. Nature **384:**179–183.
32. **Wyatt, R., P. D. Kwong, E. Desjardins, R. W. Sweet, J. Robinson, W. A. Hendrickson, and J. G. Sodroski.** 1998. The antigenic structure of the HIV gp120 envelope glycoprotein. Nature **393:**705–711.
33. **Yamaguchi, J., A. S. Vallari, P. Swanson, P. Bodelle, L. Kaptue, C. Ngansop, L. Zekeng, L. G. Gurtler, S. G. Devare, and C. A. Brennan.** 2002. Evaluation of HIV type 1 group O isolates: identification of five phylogenetic clusters. AIDS Res. Hum. Retrovir. **18:**269–282.
34. **Yamaguchi, Y., and T. Gojobori.** 1997. Evolutionary mechanisms and population dynamics of the third variable envelope region HIV within single hosts. Proc. Natl. Acad. Sci. USA **94:**1264–1269.
35. **Yamaguchi-Kabata, Y., and T. Gojobori.** 2000. Reevaluation of amino acid variability of the human immunodeficiency virus type 1 gp120 envelope glycoprotein and prediction of new discontinuous epitopes. J. Virol. **74:** 4335–4350.
36. **Yang, Z.** 1997. PAML: a program package for the phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **13:**555–556.
37. **Yang, Z., and J. P. Bielawski.** 2000. Statistical methods for detecting molecular adaptation. Trends Ecol. Evol. **15:**496–503.
38. **Yang, Z. H., R. Nielsen, N. Goldman, and A. M. K. Pedersen.** 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155:**431–449.
39. **Yusim, K., C. Kesmir, B. Gaschen, M. M. Addo, M. Altfeld, S. Brunak, A. Chigaev, V. Detours, and B. T. Korber.** 2002. Clustering patterns of cytotoxic T-lymphocyte epitopes in human immunodeficiency virus type 1 (HIV-1) proteins reveal imprints of immune evasion on HIV-1 global variation. J. Virol. **76:**8757–8768.
40. **Zanotto, P. M., E. G. Kallas, R. F. de Souza, and E. C. Holmes.** 1999. Genealogical evidence for positive selection in the *nef* gene of HIV-1. Genetics **153:**1077–1089.

JOURNAL OF VIROLOGY, Apr. 2004, p. 4381–4382 Vol. 78, No. 8
0022-538X/04/$08.00+0 DOI: 10.1128/JVI.78.8.4381–4382.2004

ERRATUM

Comparative Study of Adaptive Molecular Evolution in Different Human Immunodeficiency Virus Groups and Subtypes

Marc Choisy, Christopher H. Woelk, Jean-François Guégan, and David L. Robertson

CEPM, UMR CNRS-IRD 9926, Montpellier, France; Department of Pathology, University of California—San Diego, La Jolla, California 92093; and School of Biological Sciences, University of Manchester, Manchester, United Kingdom

Volume 78, no. 4, p. 1962–1970, 2004. Page 1965, column 2: The final paragraph of Results should read "Finally, the locations of sites implicated in the binding of the envelope glycoprotein to the CD4 receptor molecules (18, 32) and of sites implicated in the receptor switch from CCR5 to CxCR4 tropism (20) are indicated in Fig. 2. Two of the three sites involved in chemokine binding associate with identified positively selected sites, suggesting an expected adaptive evolution at these sites. The finding that some of these sites are apparently not under the influence of positive selection may be due to our comparison of viral sequence data being from several different infected individuals rather than from the viral population of one individual. The CD4 binding sites are not associated with any of the positively selected sites identified as they are presumably under the influence of purifying selection."

Page 1966: Figure 2 should appear as shown on the following page.

```
                 [gp120                    vpu]
HIV-1 M:A        MRAMGIQR-NCQ-SL-WRWG--IMILGMIIICSAA-GNLWVTVYYGVPVWKDAETTLFCASDAKAYTTEVHNVWATHACVPTDPNPQEINLENVTEKFNM
HIV-1 M:B        MTVKGTRK-NYR-HL-WTWG--TMILGILMICSAA-NNFWVTVYYGVPVWREATTTLFCASDAKAYDEEVHNVWATHACVPTDPNPQEVELINVTENFNM
HIV-1 M:C        MRATGIQR-NCQ-QW-WIWG--ILGFWMLMICNVM-GNLWVTVYYGVPVWKEATTTLFCASDAKAYDTEAHNVWATHACVPTDPNPQEMYLKNVTEDFNM
HIV-1 M:D        MRVKGIKR-NYQ-PL-WKWG--IMLLGMLMMTYSAADNLWVTVYYGVPVWKEATTTLFCASDAKSYKTEAHNIWATHACVPTDPNPQEIELKNVTENFNM
HIV-1 O          MRA----MEKR---NKKLW-TLYLAMALITPCLSL-RQLYATVYAGVPVWEDATFVLFCASDANLTSTEKHNIWASQACVPTDPTPIEYPLHNVTDDFNI
HIV-2 A          M---------CSRNLLLVA---SLLASICLIYCT----QYVTVFYGVPAWRNASIPLFC-------ATKNRDTWGTIQCLPDNDDYQEIKL-NVTEAFDA
                                                                                                                 4

                                                <---------------------V1--------------------><---------------
HIV-1 M:A        WKNNMVEQMHTDIISLWDQSLQPCVKLTPLCVTLECSDVN---------VNAT--------NYN----VNDTIN--MKEEIRNCSFNMTTELRDRKQKVY
HIV-1 M:B        WKNNMVEQMHEDIISLWDQSLKPCVKLTPLCVSLNCTDAK---------NTTNSNTNSSSSTNS----SSLEQG--KAGEIKNCSFNITTNMRDKVQKQY
HIV-1 M:C        WKNDMVDQMHQDIISLWDQSLKPCVKLTPLCVTLDCDNAN---------ATEN----------G----TIDNRT--MGGEIKNCSFNITTELKDKKQRAY
HIV-1 M:D        WKNNMVDQMHEDIISLWDQSLKPCVKLTPRCVTLNCTDAS---------NNST--------DNN----STLPTV--KPGEMKNCSFNITTVVTDKRKQVH
HIV-1 O          WKNYMVEQMQEDIISLWDQSLKPCVQMTPLCVQMECTNIA---------GTTN-----------------------ENIMKKCEFNVTTVIKDKKEKKQ
HIV-2 A          WDNTVTEQAVEDVWSLFETSIKPCVKLTPLCVTMNC-TTMKNISTTARP-TTGISNTESILKENSSCISTDTCAGLEQEETVDCQFNMTGLERDKKKQYT
                                                  2 45         5553                                  2

                 --------V2------------------------>
HIV-1 M:A        SLFYRLDVVQMNN-----------SNNSNQYRLINCNTSAITQACPKVSFEPIPIHYCAPAGYAILKCKDKRFNGTGECKNVSTVQCTHGIKPVVSTQL
HIV-1 M:B        ALFYSLDIVPIDD-----------KGNDTSYRLISCNTSVITQACPKISFEPIPIHYCAPAGFAILKCNEKGFNGKGPCKNVSTVQCTHGIRPVVSTQL
HIV-1 M:C        ALFYSLDIVQLDG---------------GGSYRLISCNTSAITQACPKVSFDPIPIHYCAPAGYAILKCNNKTFNGTGQCKNVSTVQCTHGIKPVISTQL
HIV-1 M:D        ALFYRLDVVQIDN------EGKNEINDTYGTYRLINCNTSAITQACPKVSFEPIPIHYCAPAGFAILKCNDKRFNGTGPCKNVSSVQCTHGIRPVVSTQL
HIV-1 O          ALFYVSDIMELNE-------TSSTNKTNSKMYTLTNCNSTTITQACPKVSFEPIPIHYCAPAGYAIFKCNSTEFNGTGTCRNITVVTCTHGIRPTVSTQL
HIV-2 A          ETWYSKDVVCESSTSVTSK------TNWTQKCYMSHCNTSVITESCDKHYWDAMRFRYCAPPGYALLRCNDTKYSGFEPNCTKVVASTCTRMMETQTSTW
                                                4                                                  2       2

                                                <---- *-- *------V3-----* ---------->
HIV-1 M:A        LL-NGSLAE-SKVMIRSENITNNAKNILVQLTSPVHISCIRPNNNTRKSVRI-GPGQAFYATGE---IIGNIRQAYCNVNRSEWNEALREVVKQLR--TK
HIV-1 M:B        LL-NGSLAE-EEVVIKSDNFTNNAKTIIVQLNTEVEITCVRPNNNTRRSIPI-GPGRAFYTT-E---IIGDIRQAYCNITKANWTDTLQKVAIKLR--EQ
HIV-1 M:C        LL-NGSLAE-EEIIIRSENLTNNAKTIIVHLKDPVEIECTRPNNNTRKSIRI-GPGQILYATGD---IIGDIRQAHCNINETKWNKTLQDVSEKLA--KY
HIV-1 M:D        LL-NGSLAE-EEIVIRSENLTNNAKIIIVHLNQSVEINCTRPYKKERQRTPI-GPGQALYTTRY---TTRIIGQAYCNISGVKWNNTLRQVARKLG--NL
HIV-1 O          IL-NGTLSK-GKIRMMAKDILEGGKNIIVTLNSTLNMTCERPQI-DIQEMQI-GPMAMYSMGIG-GTAGNSSRAAYCKYNATDWGKILKQTAERYL-ELV
HIV-2 A          FGFNGTRAENRTYIYWHGRD---NRTIISLNKYYNLTIHCKRPGN-KTVVPITLMSGHRFHS-RP--VINRKPMQAWCWFEGKWREAMQEVKQTLVKHPR
                                          4           2            2         2                        2  3   3 22

                                  | |                <-----------------V4---------------->                | |       |
HIV-1 M:A        F----NKT--IIFDNSS-GGDLEITTHSFNCGGEFFYCNTSRLFNSTW-NDTTSMLN--DTKPN-----DTITLPCRIKQIINMWQRAGQAIYAPPIQ-G
HIV-1 M:B        F----NKT--IIFKPSS-GGDPEIVTHSFNCGGEFFYCNSTQLFNGTW-NGTWINGTWKSSYGNDT---TNITLPCRIKQIINMWQEVGKAMYAPPIR-G
HIV-1 M:C        FP---NKT--INFAQPS-GGDLEIVTHSFNCRGEFFYCNTSKLFNSTD-----RSS---STES------ANITIPCRIKQIINMWQQVGRAMYAPPIK-G
HIV-1 M:D        LN---QTK--IIFKPSS-GGDPEITTHSFNCGGEFFYCNTSGLFNSAW--NISGHS---TGLND-----TIITIPCRIKQIINMWQEVGKAMYAPPIE-G
HIV-1 O          NNT-GSIM--MTFNHSS-GGDLEVTRLHFNCHGEFFYCNTAKMFNYTFSCHGTTCS----YSNVSQGNNGT--LPCKLRQVVRSWIRGQSGLYAPPIK-G
HIV-2 A          YR-G-TNKTE-NITFAGPGEGSDPEVAYMWTNCRGEFLYCNMTWFLNW-VE---NTT--NQT---------QHN-YVPCHIRQIINTWHKVGKNVYLPP-
                                 3  2                          4     2                                          2          3

                                  |       <---V5--->                                          gp120]gp41
HIV-1 M:A        VIRCKSNITGLILLTRDGGGN--SSVSSETFRPGGGDMRDNWRSELYKYKVVKIEPLGVAPTKAKRRVV----EREKRAVGLGAV-FIGFLGAAGSTMGAA
HIV-1 M:B        QIKCTSNITGLLLLTRDGGNSNTTDNSTETFRPGGGDMRDNWRSELYRYKVVQVEPLGIAPTRAKRRVV----QREKRAVGIGAM-FLGFLGAAGSTMGAA
HIV-1 M:C        KITCKSNITGLLLTRDGGTT---NDSTEAFRPAGGDMKDNWRSELYKYKVVEIKPLGVAPTKAKRRVV----EREKRAVGIGAV-FLRFLGAAGSTMGAA
HIV-1 M:D        QIKCSSNITGLLLLTRDGGAN---NTQNDTFRPGGGDMRDNWRSELYKYKVVKIEPLGVAPTKAKRRVV----EREKRAIGLGAM-FLGFLGAAGSTMGAA
HIV-1 O          NLTCMSNITGMILQMDNTWN--SSNNNVTFRPIGGDMKDIWRTELFNYKVVRVKPFSVAPTRIARPVISTRTHREKRAVGLGML-FLGVLSAAGSTMGAA
HIV-2 A          REGLLSCNSTVTSIIANID----T-YQ-NQT--DITFSAEVAELYRLELGDYKLVEITPIGFAPTSVKRYSSAPGRNKRQVFVLGFLGFLTTAGAAMGAA
                     2                                                                  3

HIV-1 M.A        SITLTVQARQLLSGIVQQQSNLLRAIEAQQHLLKLTVWGIKQLQARVLALERYLKD-QQLLGIWGCSGKLICTTNVPWNSSWS-NKSQSEIWENMTWLQW
HIV-1 M:B        SVTLTVQARQLLSGIVQQQNNLLRAIEAQQHMLQLTVWGIKQLQARVLAVERYLKD-QQLLGLWGCSGKLICTTTVPWNTSWS-NKTLEKIWDNMTWMQW
HIV-1 M:C        SITLTVQARQLLSGIVQQQSNLLRAIETQQHMLQLTVWGIKQLQTRVLAIERYLKD-QQLLGIWGCSGKLICTTAVPWNDSWSNNKTLGDIWNSTTWMQW
HIV-1 M:D        SMTLTVQARQVLSGIVQQQNNLLRAIEAQQHLLQLTVWGIKQLQARILAVERYLKD-QQLLGIWGCSGKHICTTTVPWNSSWS-NRSVEYIWGNMTWMQW
HIV-1 O          ATTLAVQTHTLLKGIVQQQDNLLRAIQAQQQLLRLS?WGIRQLRARLLALETLLQN-QQLLSLWGCKGKLVCYTSVKWNRTWI-G--NRRIWDTLTWQEW
HIV-2 A          SLTLSAQSRTLLAGIVQQQQQLLDVVKRQQEMLRLTVWGTKNLQARVTAIEKYLKD-QAQLNSNGCAFRQVCHTTVPW----V-NDTLQPDWNNMTWQKW
                  2

                                                                                                                 [rev2
HIV-1 M:A        DKEISNYTHIIYTLLEESQIQQEKNEQDLLALDKWANLWNWFDITNWLWYIKIFIMIVGGLIGLRIVFTVLSIINRVRQGYSPLSFQTHTP-----NPGG
HIV-1 M:B        ETEINNYTNIIYTLLEESQNQQEKNEKDLLELDQWANLWNWFISNWLWYIKIFIMIVGGLIGLRIVFTVFSIVNRVRQGYSPLSFQTHLP-----TPRG
HIV-1 M:C        DREISNYTNTIYRLLEDSQNQQEQNEKDLLALDKWQNLWSWFDITNWLWYIKIFIMIVGGLIGLRIIFAILSIVNRVRQGYSPLSFQTLTP-----SPRG
HIV-1 M:D        EREIDNYTGLIYNLIEESQIQQEKNEKELLELDKWASLWNWFSITQWLWYIKIFIMIVGGLIGLRIVFAVLSIVNRVRQGYSPLSFQTLLP-----APRG
HIV-1 O          DRQIRNISSTIYEEIQKAQVQQEQNEKKLLELDEWASIWNWLDITKWLWYIKIAIIIVGALVGVRVIMIVLNIVKNIRQGYQPLSLQIPNH-----HQEE
HIV-2 A          EQQIRYLEANISTSLEQAQIQQEKNMYELQKLNSWDVFGNWFDLTSWIRYIQYGVYIVVGIIALRIIIYVVQMLSRFRKGYRPVFSPPPAYLQQVHIHKH
                     2   4   3                                  3  2

                                                                                                          rev2]
HIV-1 M:A        -LDRPRRIEEEGGEQDRDRSIRLVGGFLTLVWDDLRSLCLFSYHRLRDFTLIAARTVELLGHSSLKGLRLGWEGLKYLGNLLLYWGRELRISAINLLDTI
HIV-1 M:B        -PDRPEGIEEEGGERGRGSSTRLVHGFLALFWDDLRSLCLFSYHRLRDLLLIVTRIVELLGRG-------GWEALKYWWNLLQYWRQELQKSAVSLFNAT
HIV-1 M:C        -PDRLGRIEEEGGEQDRDRSIRLVSGFLALAWDDLRSLCLFSYHRLRDLILIATRVVELLGRSSLRGLQRGWEILKYLGSLVQYWGLELKKSAINLLNIT
HIV-1 M:D        -PDRPEGIEEEGGEQDRGTSIRLVNGFSALIWDDLRNLCLFSYHRLRELILIAARIVELLGRR-------GWEALKYLWNLLQYWSRELKNSAISLVDAT
HIV-1 O          -AGTPGRTGGGGGEEGRPRWIPSPQGFLPLLYTDLRTIILWTYHLLSNLASGIQKVISYLRLGLWILGQKIINVCRICAAVTQYWLQELQNSATSLLDTL
HIV-2 A          RGQPPSEETEEDVGDSVGDSSWPPIAYIHFLIRLLIRLLIGLYNICRTLISKSFQTLQPISQGLQRALTAIRDWLRPGAAYLQYGCEWIQEALQAFARAT
                                                                                                                 2

                                          gp41]
HIV-1 M:A        AIIIAGWTDRVIEIGQRFCRAILNIPRRIRQGFERALL
HIV-1 M.B        AIAVAEGTDRVIEALQRTTRAILHIPRRIRQGLERALL
HIV-1 M:C        AIAVAEGTDRIIEIIQRICRAIYHIPRRIRQGFEAALQ
HIV-1 M:D        AIAVAEGTDRIIDIVRRACKAVLHIPTRIRQGLERALL
HIV-1 O          AVAVANWTDGIIAGIQRIGTGIRNIPRRIRQGLERSLL
HIV-2 A          RETLTSVWRNFCGTMGQIGRGIIAIPRRIRQGAELALL
                                  3                2
```

Annexe B

Annexes de l'article

CHOISY M., WOELK C.H., GUÉGAN J.-F. & ROBERTSON D. (2004) Comparative study of adaptive molecular evolution in different HIV clades. *Journal of Virology* **78**(4) : 1962-1970

Supplementary material

A comparative study of adaptive molecular evolution in different HIV groups and subtypes

Marc Choisy[1], Christopher H. Woelk[2], Jean-François Guégan[1], and David L. Robertson[3*]

[1] CEPM, UMR CNRS-IRD 9926, Montpellier, France

[2]University of California San Diego, Department of Pathology, 9500 Gilman Dr., La Jolla, CA, 92093, USA

[3]School of Biological Sciences, University of Manchester, Manchester, UK

Running title: Detection and quantification of positively selected sites in HIV gene sequence alignments

*Corresponding author. Mailing address: University of Manchester, 2.205 Stopford Building, Oxford Road, Manchester, M13 9PT. Phone: +44 |(0)161 275 5089. Fax: 0161 275 5082. E-mail: david.robertson@man.ac.uk.

Appendix A. Models M0, M1, M2, M3, M7, and M8.

Yang and coworkers (4) originally proposed 14 models (M0 through M14) for their ML analysis but it became evident from the analysis of biological sequence data that a subset of these models (M0, M1, M2, M3, M7 and M8) was sufficient for detecting positive selection. The M0 (one-ratio) model assumes a single ω for all sites. M1 (neutral) assumes a proportion p_0 of conserved sites with $\omega_0 = 0$ and a proportion $p_1 = 1\text{-}p_0$ of neutral sites with $\omega_1 = 1$. M2 (selection) adds an additional class of sites to M1 ($p_2 = 1\text{-}p_0\text{-}p_1$) for which ω_2 can be estimated from the data. M3 (discrete) estimates ω for a predetermined number of classes (in this case three). Model M7 (beta) uses a discrete beta distribution with ten categories to model different ω ratios (between 0 and 1) among sites. The shape of this beta distribution is governed by the two parameters p and q. Model M8 (beta&ω) adds an additional class of sites to model M7 whereby a proportion of sites (p_1) can have an ω_1 above 1. These models are fully described in the literature (1, 2, 4). M2, M3 and M8 are able to account for positive selection whereas M0, M1 and M7 are not. M0 and M1 are both nested with M2 and M3, M2 is nested with M3, and M7 is nested with M8. Thus the following LRTs were performed in this paper: M0 *vs* M2, M1 *vs* M2, M0 *vs* M3, M1 *vs* M3, M2 *vs* M3 and M7 *vs* M8. Models were implemented using the CODEML program of the PAML package, version 3.1(3).

1. **Goldman, N., and Z. Yang.** 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Molecular Biology and Evolution **11:**725-736.
2. **Nielsen, R., and Z. Yang.** 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics **148:**929-936.
3. **Yang, Z. H.** 1997. PAML: a program package for the phylogenetic analysis by maximum likelihood. Computer Applications in the Biosciences **13:**555-556.
4. **Yang, Z. H., R. Nielsen, N. Goldman, and A. M. K. Pedersen.** 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155:**431-449.

Appendix B. Likelihood and parameter estimates for selection analysis.

The first column lists the data sets and the models used. The second and third columns show the log likelihood of the model (ln L) and the average ω ratio (d_N/d_S) respectively. The last column contains the parameter estimates.

Data set/model	ln L	d_N/d_S	Parameter estimates
HIV-1 M A1			
M0	-8834.570	0.551	$\omega = 0.5508$
M1	-8444.831	0.421	$p_0 = 0.57860, p_1 = 0.42140$
M2	-8262.656	0.857	$p_0 = 0.55626, p_1 = 0.35631, p_2 = 0.08743$
			$\omega_2 = 5.72956$
M3	-8239.691	0.725	$p_0 = 0.69914, p_1 = 0.22231, p_2 = 0.07854$
			$\omega_0 = 0.06691, \omega_1 = 1.18505, \omega_2 = 5.27749$
M7	-8414.049	0.417	$p = 0.05740, q = 0.08560$
M8	-8244.451	0.690	$p = 0.13812, q = 0.34700$
			$p_0 = 0.90827, p_1 = 0.09173, \omega_1 = 4.70204$
HIV-1 M B			
M0	-13679.987	0.543	$\omega = 0.5433$
M1	-13164.540	0.514	$p_0 = 0.48620, p_1 = 0.51380$
M2	-12822.132	0.938	$p_0 = 0.46914, p_1 = 0.43930, p_2 = 0.09156$
			$\omega_2 = 5.44728$
M3	-12720.594	0.664	$p_0 = 0.70048, p_1 = 0.23205, p_2 = 0.06747$
			$\omega_0 = 0.09516, \omega_1 = 1.15949, \omega_2 = 4.86597$
M7	-12979.832	0.323	$p = 0.16979, q = 0.35580$
M8	-12726.897	0.623	$p = 0.22378, q = 0.53990$
			$p_0 = 0.91119, p_1 = 0.08881, \omega_1 = 4.00858$
HIV-1 M C			
M0	-13672.725	0.539	$\omega = 0.5389$
M1	-13175.912	0.528	$p_0 = 0.47250, p_1 = 0.52750$
M2	-12781.804	0.991	$p_0 = 0.45177, p_1 = 0.45985, p_2 = 0.08837$
			$\omega_2 = 6.01076$
M3	-12655.155	0.694	$p_0 = 0.74515, p_1 = 0.20237, p_2 = 0.05248$
			$\omega_0 = 0.12213, \omega_1 = 1.37975, \omega_2 = 6.16796$
M7	-12952.180	0.313	$p = 0.18750, q = 0.41193$
M8	-12668.849	0.610	$p = 0.24681, q = 0.58871$
	-13672.725	0.539	$p_0 = 0.92450, p_1 = 0.07550, \omega_1 = 4.46262$
HIV-1 M D			
M0	-7977.420	0.462	$\omega = 0.4620$

M1	-7732.625	0.437	$p_0 = 0.56301, p_1 = 0.43699$
M2	-7614.904	0.775	$p_0 = 0.54055, p_1 = 0.39073, p_2 = 0.06872$
			$\omega_2 = 5.58635$
M3	-7580.839	0.611	$p_0 = 0.81027, p_1 = 0.16476, p_2 = 0.02497$
			$\omega_0 = 0.13383, \omega_1 = 1.84093, \omega_2 = 7.96518$
M7	-7694.235	0.315	$p = 0.14035, q = 0.30567$
M8	-7583.151	0.568	$p = 0.30053, q = 0.91637$
			$p_0 = 0.90998, p_1 = 0.09002, \omega_1 = 3.82134$
HIV-1 O			
M0	-20702.779	0.494	$\omega = 0.4939$
M1	-19784.103	0.526	$p_0 = 0.47404, p_1 = 0.52596$
M2	-19352.047	0.854	$p_0 = 0.46756, p_1 = 0.45527, p_2 = 0.07717$
			$\omega_2 = 5.16582$
M3	-19245.742	0.626	$p_0 = 0.57335, p_1 = 0.34941, p_2 = 0.07724$
			$\omega_0 = 0.03775, \omega_1 = 0.83717, \omega_2 = 4.04237$
M7	-19528.141	0.341	$p = 0.15265, q = 0.29462$
M8	-19220.479	0.590	$p = 0.16001, q = 0.32942$
			$p_0 = 0.92833, p_1 = 0.07167, \omega_1 = 3.99248$
HIV-2 A1			
M0	-14763.981	0.364	$\omega = 0.3644$
M1	-14091.437	0.433	$p_0 = 0.56751, p_1 = 0.43249$
M2	-13882.169	0.676	$p_0 = 0.56062, p_1 = 0.37798, p_2 = 0.06140$
			$\omega_2 = 4.84529$
M3	-13768.586	0.463	$p_0 = 0.70010, p_1 = 0.23928, p_2 = 0.06062$
			$\omega_0 = 0.04271, \omega_1 = 0.86680, \omega_2 = 3.71720$
M7	-13924.707	0.276	$p = 0.12446, q = 0.32596$
M8	-13765.788	0.444	$p = 0.14980, q = 0.47151$
			$p_0 = 0.97065, p_1 = 0.02935, \omega_1 = 3.56379$

Appendix C. Likelihood ratio test (LRTs) between models to test the significance of results obtained through selection analysis. LRTs are performed by taking twice the difference in log likelihood between two models and comparing the value obtained with a χ^2 distribution (degrees of freedom equal to the difference in the number of parameters between the models). *p*-values in bold indicate comparisons where the null hypothesis (no positive selection) can be rejected in favour of the alternative hypothesis (positive selection) such that the model on the left is rejected in favour of the one on the right.

LRT	M0 vs M2		M1 vs M2		M0 vs M3		M1 vs M3		M2 vs M3		M7 vs M8	
df*	2		2		4		4		2		2	
	χ^2	*p*-value	χ^2	*p*-value	χ^2	*p*-value	χ^2	*p*-value	χ^2	*p*-value	χ^2	*p*-valu
HIV-1 M A1	1143.828	**<0.001**	364.349	**<0.001**	1189.758	**<0.001**	410.279	**<0.001**	45.930	**<0.001**	339.197	**<0.00**
HIV-1 M B	1715.710	**<0.001**	684.818	**<0.001**	1918.785	**<0.001**	887.892	**<0.001**	203.075	**<0.001**	505.870	**<0.00**
HIV-1 M C	1781.843	**<0.001**	788.216	**<0.001**	2035.141	**<0.001**	1041.514	**<0.001**	253.298	**<0.001**	566.663	**<0.00**
HIV-1 M D	725.032	**<0.001**	235.441	**<0.001**	793.162	**<0.001**	303.571	**<0.001**	68.130	**<0.001**	222.168	**<0.00**
HIV-1 O	2701.464	**<0.001**	864.112	**<0.001**	2914.074	**<0.001**	1076.722	**<0.001**	212.610	**<0.001**	615.324	**<0.00**
HIV-2 A1	1763.624	**<0.001**	418.536	**<0.001**	1990.790	**<0.001**	645.702	**<0.001**	227.166	**<0.001**	317.837	**<0.00**

*df, degree of freedom between the respective models.

Appendix D. Paired Wilcoxon ranked sum test to determine differences in the strength of positive selection between different HIV data sets. Z refers to the statistic and N refers to the number of sites with posterior probabilities of being in the positively selected class of M8 above the 0.95 level. Significant differences ($P < 0.05$) are indicated in bold. A continuity correction was applied to the normal approximation for the P-values.

HIV data set	HIV-1 M:A	HIV-1 M:B	HIV-1 M:C	HIV-1 M:D	HIV-1 O
HIV-1 M:B	Z =2.1567				
	N = 6				
	P = **0.0310**				
HIV-1 M:C	Z = -2.772	Z = -2.1567			
	N = 10	N = 6			
	P = **0.0056**	P = **0.0310**			
HIV-1 M:D	Z = -2.772	Z = -1.6432	Z = -1.3363		
	N = 4	N = 4	N = 3		
	P = 0.1003	P = 0.1003	P = 0.1814		
HIV-1 O	Z = 1.9799	Z =2.2222	Z =2.2222	Z = 1.3363	
	N = 5	N = 6	N = 6	N = 3	
	P = **0.0477**	P = **0.0263**	P = **0.0263**	P = 0.1814	
HIV-2 A	Z = 1.3363	Z = 0	Z =1.3363	Z = 0.8944	Z =0.8944
	N = 3	N = 1	N = 3	N = 2	N = 2
	P = 0.1814	P = 1.000	P = 0.1814	P = 0.3711	P = 0.3711

Annexe C

HIDE M., CHOISY, M. & BAÑULS A.-L. Molecular evolution of cathepsin like proteases within *Leishmania* genus and other Trypanosomatids. Soumis à *International Journal for Parasitology*

Molecular evolution of cathepsin like cysteine proteases within *Leishmania* genus and other trypanosomatids

Running title: Molecular evolution of *Leishmania* cysteine proteases

Mallorie Hide, Marc Choisy and Anne-Laure Bañuls

Génétique et Evolution des Maladies Infectieuses (ex-CEPM), centre IRD

911, avenue Agropolis, BP 64501, 34394 Montpellier cedex 5, France.

Corresponding author:

Anne-Laure Bañuls

Génétique et Evolution des Maladies Infectieuses (ex-CEPM), centre IRD

911, avenue Agropolis, BP 64501, 34394 Montpellier cedex 5, France.

Tel: +33 (0)4 67 41 61 80; Fax: +33(0)4 67 41 62 99

E-mail: banuls@mpl.ird.fr

Key words: molecular evolution, positive selection, cathepsin like, *Leishmania*, Trypanosomatidae, archival antigenic library.

Abstract
Within the *Leishmania* genus, researchers have showed an enhanced interest for cysteine proteases and especially for the cathepsin like because of their key roles in infection and expression of the disease, making them potential drug targets or vaccinal antigen. Despite a non consensual nomenclature in the literature, three groups of cathepsin like (*cpa* and *cpb* identified as cathepsin L-like and *cpc* as cathepsin B-like) have been evidenced which are different in structure and function. However, very little is know on the evolutionary processes underlying the evolution of these proteins in the *Leishmania* genus. In the present work, we propose a study of these proteins in the *Leishmania* genus and the Trypanosomatidae family through phylogenetic analyses and search for molecular adaptation. Our results help to clarify the confusing classification of *cp* reported in the literature for the *Leishmania* genus and to demonstrate, contrary to *cpb* and *cpc,* the *Leishmania* specificity of *cpa*. Several duplication events are at the origin of these three *Leishmania cp*. The positive selection and phylogenetic analysis revealed diversifying evolution in the three *Leishmania cp* genes: *cpb* presents a faster diversifying evolution than *cpa* and *cpb*. Furthermore, the strong positive selection detected in *cpb* genes and different properties already described by several authors, allow to suggest, for the first time, this gene as an archival antigenic library. This designates this gene as a good target to understand how Trypanosomatidae are able to escape the host immune system.

Introduction

Cysteine proteases, also called thiol proteinases were divided in several clans and families depending on a number of characteristics including sequence similarity, possession of inserted loops, and biochemical specificity to small peptide substrates (Sajid and McKerrow, 2002). They have been characterized for a wide range of living organisms as well within eukaryotes as within prokaryotes (Berti and Storer 1995) and even virus (Bazan and Fletterick 1988). The extraordinary ubiquity of these proteins underlines their fundamental roles in the survival of species. They have attracted considerable attention, especially on behalf of infectious diseases researchers because of their evidenced fundamental role in various aspects related to virulence (Alexander et al. 1998; Rosenthal 1999; Stanley et al. 1995). Most of the studies suggest that cysteine proteases play a key role in the pathogenesis of parasitic protozoa infections (McClelland et al. 1994; McKerrow et al. 1993; North et al. 1990), making them potential drug targets against numerous parasites (Alves et al. 2001; Lalmanach et al. 2002; Pollock et al. 2003; Selzer et al. 1999).

The leishmaniases, caused by protozoan parasites of the *Leishmania* genus (Kinetoplastida order, Trypanosomatidae family), are an important health problem in many regions of the world. According to the estimation of World Health Organization (WHO), over 12 million people are infected with *Leishmania* worldwide and about 350 million people are at risk of contracting leishmaniases. Within the *Leishmania* genus, great interest have been brought to the cysteine proteases too, since these proteins have proved to be implied in the infection and expression of leishmaniases (Denise et al. 2003; Frame et al. 2000; Pascalis et al. 2003; Pollock et al. 2003; Rafati et al. 2001). Two subfamilies of cathepsin like, cathepsin L-like and cathepsin B-like, pertaining to the clan CA (papain-like) and to the papain-family (family C1) (Sajid and McKerrow 2002) have been evidenced in *Leishmania*. The cathepsin L-like cysteine proteases of protozoan parasites would have a role in the destruction of host proteins (Stanley et al. 1995), nutrition (Rosenthal 1999), and evasion of host immune response (Alexander et al. 1998). The cathepsin B-like are suspected to play an important role in *Leishmania* survival within the host macrophages (Mundodi et al. 2002). These *Leishmania* cathepsins like comprise three groups of cysteine proteases which are different in structure and function (Robertson and Coombs 1990) but the bibliography reveals a non consensual nomenclature for these cysteine proteases with several different names recorded in the literature: Type I, Type II, and Type III (Alexander et al. 1998; Rafati et al. 2001); *cpa*, *cpb*, and *cpc*; or two cathepsin L-like and one cathepsin B-like (Mottram et al. 1992; Rafati et al. 2001; Souza et al. 1992). For the sake of clarity, we will thereafter use the *cpa*/*cpb*/*cpc* nomenclature, not only for the *Leishmania* genus but also for the whole Trypanosomatidae family. Both *cpa* and *cpb* belong to the cathepsin L-like subfamily and *cpc* to the cathepsin B-like subfamily. Moreover, *cpa* and *cpc* correspond to unique genes (Bart et al. 1995), whereas *cpb* are multicopy genes (Mottram et al. 1996; Robertson and Coombs 1994; Souza et al. 1992). The three genes comprise four regions: a pre-region, a pro-region, a core (or protease or mature) region and a C-terminal extension (Mottram et al. 1997; Mundodi et al. 2002; Omara-Opyene and Gedamu 1997). *Cpb* is characterized by the presence of an unusual long C-terminal extension compared to *cpa* and *cpc* (Mottram et al. 1997; Omara-Opyene and Gedamu 1997). It has been demonstrated in *Leishmania* that the three genes are localized on different chromosomes (Omara-Opyene and Gedamu 1997; Sakanari et al. 1997).

To date, the evolutionary history of these related but different proteins has never been studied in details in *Leishmania* and a clear understanding of the processes underlying the evolution of these proteins is an increasing need for the fighting of leishmaniases and the designing of potential vaccine or antiparasitic drug. This paper proposes a first step towards this aim in analyzing gene phylogenies and positive selection of *Leishmania* cathepsin like cysteine proteases. We studied the molecular phylogeny of the different *Leishmania* cysteine

protease gene sequences already published. In addition to these sequences, we also considered *cp* sequences from other Trypanosomatidae, available on web databases. These analyses allowed us to investigate the evolutionary history of *Leishmania* cysteine proteases, to classify the different *Leishmania* sequences published, and to reach a global view of these proteins in the Trypanosomatidae family. From a more mechanistic point of view, the study of molecular adaptation on the *Leishmania* genes was performed through the comparison of synonymous (*i.e.* silent, d_S) and nonsynonymous (*i.e.* amino acid changing, d_N) substitutions. The powerful maximum likelihood method (ML) and Bayesian statistical framework recently developed by Yang et al. (2000) allowed us to quantify the level of selection of the coding gene sequence.

Results

Blast processes and clustal alignments

The blast processes and clustal alignments allowed us to identify the 3 genes of cysteine protease, *cpa*, *cpb* and *cpc*, already described within *Leishmania,* with *cpa* and *cpb* coding for cathepsins L-like and *cpc* coding for a cathepsin B-like .

For the *Trypanosoma* genus and for the *C. salmositica* species, the blast analysis gave only one kind of cathespin L-like. These sequences presented a long C-terminal extension and were repeated in tandem in *Trypanosoma* genus as do *Leishmania cpb*. The cysteine protease structure of *C. salmositica* remains unknown because the sequence was directly submitted to GenBank without having been published (Hontzeas et al. 2002, unpublished data, GenBank n°AY090898).

The blast analysis together with a literature review allowed us to find sequences corresponding to *Leishmania cpc* within *T. cruzi* (Nobrega et al. 1998) and *T. rangeli* (Nobrega et al. unpublished data, GenBank submission n°AF400046, 2001). These sequences were also identified as cathepsins B-like. Moreover, the blast analysis allowed to align *Leishmania cpc* with a recently submitted cathepsin B-like sequence of *T. brucei* (AY508515, Mackey and McKerrow, unpublished). This last sequence could be aligned with the *T. brucei* chromosome VI), for which the sequencing is in progress, from the base number 127,701 to the base number 128,723 (Accession number: AC084046). By sequence comparison, we obtained 62.1% DNA sequence homology between *T. brucei cpc* sequence and *T. cruzi cpc* sequence and 57.6% homology with *L. infantum cpc* sequence. The translation allowed also to confirm the correspondence of this *T. brucei* protein sequence with *cpc*, *i.e.* cathepsin B-like (50.7% homology with *L. infantum* and 64.3% with *T. cruzi*).

DNA sequence homology

The comparison of the different homology percentages obtained between the *Leishmania* species for each *cp* evidenced that *cpb* presented the highest genetic divergence (table 2). The data revealed that *Leishmania cpa* was closer to *Leishmania cpb* than to *Trypanosoma cpb*. This was observed as well when *cpb* was considered with or without C-terminal extension (see table 2). Moreover, the percentages indicated that the *C. salmositica cp* sequence is genetically closer to *cpa* and *cpb* than to *cpc*. This result was robust respective to the inclusion or not of the C-terminal extension of *C. salmositica cpb*.

Phylogenetic analyses

All genetic distances used for phylogenetic analyses detailed in materials and methods gave congruent results with the parsimony method. Furthermore, the analyses performed on the corresponding protein sequences gave exactly the same results. For the sake of simplicity, we will present hereafter only the Wagner trees of gene sequences.

Figure 1 was built with all the Trypanosomatidae *cp* sequences noted F in table 1, the *C. salmositica* sequence and *P. falciparum* sequence, used as outgroup. We could clearly distinguish the two subfamilies of cathepsin like with cathepsin B-like represented by the *cpc* cysteine proteases and cathepsin L-like represented by the *cpa* and *cpb* cysteine proteases. Thus, the tree shows that *cpa* was more closely related to *cpb* than to the *cpc* genes. Figure 1 evidenced that the sequence of *C. salmositica* strain is phylogenetically related to the trypanosomatids cathepsin L-like with a bootstrap value of 98.8. The sequence of this species, which belongs to the Bodonidae family, was localized on the Wagner tree above the branch of the Trypanosomatidae family. Moreover, the node grouping all the cathepsin L-like (*cpa* and *cpb*) of the Trypanosomatidae family was sustained by a bootstrap value equal to 63.8.

For the *Leishmania* genus, the same phylogenetic hierarchization of species was found for each gene of cysteine protease: all the species of each *Leishmania* complex used were clustered together with the sequences of the *L. donovani*, *L. infantum* and *L. chagasi* species in one cluster and the sequences of the *L. mexicana*, *L. amazonensis* and *L. pifanoi* in an other one, whatever the *cp* gene considered. The stocks belonging to *L. major* species are phylogenetically closer to the *donovani* complex than to the *mexicana* complex for the three *cp* genes, consistently with the classic taxonomy. These clusterings were sustained by strong bootstrap values (≥ 87.5). The *L. guyanensis* sequences belonging to the *Viannia* subgenus fell apart from the *Leishmania* subgenus for the *cpa* and *cpb* groups. Note also that the different *cpb* sequence repeats of *L. mexicana or T. cruzi* were phylogenetically closer than the *cpb* sequences of the other species. In agreement with the percentage homology data, *Leishmania cpb* was more closely related to *Leishmania cpa* than to *Trypanosoma cpb*, but the branch presented a low bootstrap value equal to 25.4. For the other taxa, first, we could observe that all the *Trypanosoma cpc* sequences were phylogenetically related to the *Leishmania cpc* sequences. Moreover, the tree of figure 1 confirmed the phylogenetic link of the two *T. brucei* sequences with the *cpc* group. Second, the *Trypanosoma cpb* appeared hanged with the *Leishmania cpa* and *cpb*. In *cpb* and *cpc,* all the species belonging to the *Trypanosoma* genus were grouped together. This structuration is relatively well sustained for the the two *cp* genes (bootstrap value equal to 77 for the *cpc* branch and 83.8 for the *cpb* one. Moreover, the *cpb* of American species (*i.e. T. cruzi* and *T. rangeli*) fell in one cluster (bootstrap value equal to 88.1), and the *cpb* of African species (*i.e. T. brucei*, *T. congolense*, *T. b. rhodesiense* and *T. evansi*) fell in one another cluster (bootstrap value equal to 77.8).

Positive selection

All models able to detect positive selection (M2, M3, and M8) identified a positively selected class ($\omega > 1$) and could reject those models unable to account for positive selection (M0, M1, and M7). For the sake of clarity, only results for M8 are presented in table 3. M2 and M3 identified the same positively selected sites when considering posterior probabilities greater than the 95% level using the Bayesian approach. M3 identified all the sites identified by M2 and M8 and several more. We consider the sites identified by M8 only (table 3), as M3 has the potential to overestimate the number of positively selected sites (Anisimova et al. 2001; Yang et al. 2000). Figure 2 shows the posterior probabilities to belong to the selected (*i.e.* 11th) class (first column) as well as the weighted mean ω value (second column) for each site of the nucleoside hydrolase or *nsnh* (first row), *cpc* (second row), *cpa* (third row) and *cpb* (fourth row) gene sequence alignments. The two columns of figure 2 thus show substantially the same information but from different perspectives, hence their likeness. The *nsnh* gene sequence alignment did not show positive selection as none of the models allowing positive selection (M2, M3, M8) could reject those which did not (M0, M1, M7), see table 3, and it is characteristic that the level of selection is homogeneous along the sequence (first row of figure 2). On the contrary, the *cpa*, *cpb*, and *cpc* gene sequence alignments did all exhibit

significant positive selection as all the models allowing positive selection could reject those which did not (table 3). The level of selection along these three genes was quite heterogeneous with the majority of sites under a strong conservative selection and the sites under positive selection tending to occur on the second half of the sequence (figure 3). Both the mean level of positive selection (mean ω) and the number of positively selected sites increased from *cpc* to *cpa* to *cpb* gene sequence alignments (figure 2, table 3).

Discussion

The published data are confusing because of the numerous names and codes given to each proteins or genes identified in *Leishmania*. The obtained phylogenetic structuration revealed that for a same gene, the names or symbols used could be different in the literature. For example, the sequences called *Leishmania cpa* in figure 1, are also found in the bibliography under the names of cathepsin L-like, cysteine protease, or cysteine protease A and symbolically, *cpa*, *cys1* (*lpcys1* for *L. pifanoi)* or *cys2* (*ldccys2* for *L. chagasi*) (Mottram et al. 1992; Mundodi et al. 2002; Omara-Opyene and Gedamu 1997; Rafati et al. 2001; Traub-Cseko et al. 1993). Moreover, the sequences grouped as *Leishmania cpb* in figure 1 were designed as cysteine protease or cathepsin L-like and symbolically *cpb* or *cys1* (*ldccys1* for *L.chagasi)* or *cys2* (*lpcys2* for *L.pifanoi*) (Brooks et al. 2001; Mottram et al. 1996; Mundodi et al. 2002; Omara-Opyene and Gedamu 1997; Sakanari et al. 1997; Souza et al. 1992; Traub-Cseko et al. 1993) (table 1). This work helps clarifying the classification of the three different cysteine protease genes in *Leishmania* with *cpa* and *cpb* belonging to the cathepsin L-like subfamily (Alexander et al. 1998; Mottram et al. 1992; Omara-Opyene and Gedamu 1997; Traub-Cseko et al. 1993) and *cpc* belonging to the cathepsin B-like subfamily (Sajid and McKerrow 2002) (table 1, figure 1).

Although homologous *Leishmania cpc* has been evidenced in American species of the *Trypanosoma* genus, its existence within African species was only suspected by Okenu et al. in 1999. It is only recently that one sequence was submitted by Mackey and McKerrow (December 2003, unpublished). The present work confirms the presence of *cpc* gene and allows to localize it on the chromosome VI of *T. brucei*. To date, no homologous *cpa* gene has been described in the *Trypanosoma* genus. As *cpb*, this protein belongs to the cathepsin L-like subfamily but it seems to be specific to the *Leishmania* genus.

The tree obtained with *cpb* and *cpc* genes follows exactly the consensual phylogenetic trees for the *Leishmania* and *Trypanosoma* genus. The *cpa* sequences are also in agreement with the *Leishmania* species phylogeny (Cupolillo et al. 1994; Lainson and Shaw 1987; Lanotte et al. 1984; Thomaz-Soccol et al. 1993b). In the *Trypanosoma* genus, the American species (*T. rangeli* and *T. cruzi*) and the African species (*T. brucei, T. congolense, T. evansi*) fell in different clusters, in agreement with the data published by Stevens et al. (2001). The phylogenetic results evidenced the monophyletic status of both the *Leishmania* and *Trypanosoma* genus. For *Leishmania*, these results concur with the already published data (Thomaz-Soccol et al. 1993a). Concerning *Trypanosoma*, its phylogenetic status is still under debate: Alvarez et al. (1996) and Stevens et al. (2001) suggest that the *Trypanosoma* genus is monophyletic, whereas a recent phylogenetic study based on 18S rRNA (Hughes and Piontkivska 2003) does not support the monophyly of this genus.

Because *cpb* and *cpc* showed here a high level of phylogenetic signal with respect to the *Leishmania* and *Trypanosoma* taxa, these two *cysteine proteases* seem to be very useful to investigate the phylogenetic relationships among a large sample of trypanosomatid species as suggested for cathepsin L-like locus by Sakanari et al. (1997). *Cpa* appears to be phylogenetically informative specifically for the *Leishmania* genus.

Evolution of Leishmania *cysteine proteases*

As described by Hughes (1994) and Berti and Storer (1995), the evolutionary history of cysteine proteases is essentially made of gene duplication events. Our phylogenetic analyses suggest that the *Leishmania* cysteine proteases have been characterized by three gene duplication events. The first one would have generated the separation between cathepsin B-like (*cpc*) and cathepsin L-like (*cpa* and *cpb*). According to Berti and Storer (1995), this event would have occurred before the vertebrate/invertebrate divergence. This is confirmed by the structuration of a phylogenetic tree built with human cathepsin sequences (data not shown). Indeed, as reported by Sajid and McKerrow (2002), *Leishmania cpa* and *cpb* and all the *Trypanosoma cpb* were localized with human cathepsin L and *Leishmania* and *Trypanosoma cpc* were clustered with the human cathepsin B (data not shown). Within the *Leishmania* genus, a second duplication event would have separate the *cpb* and *cpa* genes. However, our results show that the *cpa* sequences were phylogenetically slightly closer to *Leishmania cpb* than to *Trypanosoma cpb* illustrated by a low bootstrap value (25.4) and the percent homologies data. This suggests that the duplication event might have occurred only in the *Leishmania* genus, just after the *Trypanosoma/Leishmania* divergence. The third duplication event would originate the multiple copies of *cpb* repeated in tandem. The phylogenetic distances between the *cpb* copies being smaller than the distances between the different species *cpb* suggests that these events would have occurred independently in each *Leishmania* species. This is demonstrated for the species *L. mexicana* and for *T. cruzi* in the present work, and in the *donovani* complex by Hide et al. (unpublished data). Zhang (2003) explains that the presence of duplicated genes is sometimes beneficial simply because extra amounts of protein or RNA products are provided. This applies mainly to strongly expressed genes, the products of which are in high demand, such as rRNAs and histones. However, many different kinds of parasites escape host immunity by switching gene expression between variants stored within each genome (Frank 2002). The functional and structural characteristics of *cpb* strongly suggest an archival library of variant genes such as *var* gene within *P. falciparum* (Smith et al. 1995). Indeed, the multicopy *cpb* gene presents the typical properties of this peculiar gene organization: structure of tandemly repeated *cpb* gene, differential expression between the copies (Mundodi et al. 2002, Denise et al. 2003; Mottram et al. 1998; Mottram et al. 1997) and implication in the host parasite interactions. Concerning the last point, as described by Frame et al. (1999), several authors demonstrated that *cp* parasites have a direct actions on components of the mammalian immune system, such as immunoglobulins and complement factors (Bontempi et al, 1990; Reed et al., 1989). Rafati et al. (2001) have also clearly showed that these proteinases are targets of the immune response in *Leishmania*. Furthermore, Thomas et al. (1997) indicated that the *cpb* has been shown to be localized to the surface of *T. cruzi* amastigotes using immuno-electronmicroscopy. All these data suggest and are direct evidences, that cpb is directly exposed to the immune system and that this protein is immunogen *in vivo*.

As suggested by Berti and Storer (1995), the Trypanosomatidae taxa would be a mixture of paralogous (*i.e.* derived from a duplication event and existing in the same organism) and orthologous sequences (*i.e.* derived from a speciation event and homologous in different organisms). Paralogous sequences have different functions in the organism and may (and often do, see Wagner 2000) have different evolutionary rates. In this context, *cpa*, *cpb* and *cpc* appear as paralogous sequences with different functions and different evolutionary mutation rates in *Leishmania* genus and as orthologous sequences since they exist within different species and genus.

Selective forces within Leishmania *cysteine proteases*

Some authors have raised the possibility that Yang's ML approach to test for positive selection may be too liberal and score false positive (Suzuki and Nei, 2002). However, it is

important to note that in this study, we are less interested in the precise identification of sites under positive selection than in the comparison of the intensity of selection along the genes as documented by the weighted mean ω value $\overline{\omega} = \sum_{k=1}^{11} f_k \cdot \omega_k$ (see materials and methods).

Moreover, in order to minimize the possibility of false positive bias raised by Susuki and Nei (2002), we considered several different models and we focused especially on the M8 Model after conducting the LRT. The bias can be produce either by a high number of nucleotide substitutions or by a high sample size (Suzuki and Nei 2000). However, risks for such biases are limited in our analysis as we did not work with a high sample size and the tests were performed only in the *Leishmania* subgenus showing a reasonable level of substitutions. Lastly, the validity of our results is supported by the strong negative selection obtained with the nucleoside hydrolase gene sequence alignment. This was expected given the key role of this protein in the metabolism of *Leishmania*.

All the *cp* gene sequences alignments exhibit significant levels of positive selection. The *cpa* and *cpb* gene sequence alignments showed similar ω value of the positively selected class (table 3). The *cpc* gene sequence alignment exhibited an exceptionally high ω value of the positively selected class (three times higher than the value observed for *cpa* and *cpb*, table 3). This is not to say however that the level of positive selection in *cpc* is effectively three times higher than in *cpa* and *cpb*. Indeed, the fact that the great majority of sites in the *cpc* are not under positive selection is certainly responsible for the fact that the only site under positive selection is attached to such a high ω value. Very high caution should be therefore taken in the comparison of the ω values of the positively selected class in different gene sequences alignments (Choisy et al. 2004). This being said, it appears from the results of positive selection analysis that, from *cpc* to *cpa* to *cpb*, the mean level of positive selection is increasing (table 3: 0.4505, 0.5265, and 0.8840) the number of weighted mean ω values above 1 is increasing (figures 3d, f, and h), and the number of sites with a posterior probability to belong to the positively selected class above a significance threshold (here 0.95) is increasing (table 3 and figure 3c, e, g: 1, 2, 3). All these observations converge toward the same conclusion that *cpb* has a faster diversifying evolution than *cpa* and that *cpa* has a faster diversifying evolution than *cpc* and is thus in perfect agreement with the results obtained on the comparison of percentages of homology within each gene. Indeed this analysis showed that *cpb* multicopy gene presents the higher genetic divergences compared to *cpa* and *cpc* genes, suggesting that *cpb* evolves at a higher rate than *cpa* and *cpc*. Sajid and McKerrow (2002) have also observed that cathepsins L-like (*i.e. cpa* and *cpb*) shared less similarity than the cathepsins B (*i.e. cpc*) considering different organisms. This could be explained by the different cellular functions of these three cysteine proteases and thus by different selective pressure. *Cpb* which plays a role in the host-parasite interactions and evasion from the host immune system, is submitted to a strong selective pressure, and thus would evolve more rapidly with a faster mutation rate. However, the *cpa* and *cpb* functions are not yet clearly distinguished (Brooks et al. 2000) despite their fundamental differences in substrate specificity and structure (Bart et al. 1995; Mottram et al. 1992; Rafati et al. 2001; Robertson and Coombs 1990). They must have distinct but complementary protease activities as theoretical population genetics predicates that both duplicates sequences can be stably maintained when they differ in some aspect of their functions (Zhang 2003). It is interesting to note that the level and location of positive selection observed in the 350 site-long *cpa* gene (figure 3e and f) is very similar to those observed in the first 350 sites of the *cpb* gene (figure 3g and h) with positive selection localized in the core domain of the two genes. Moreover, in *cpb*, positive selection appears particularly intense in the last 100 sites, which correspond to the particularly long C-terminal extension. Despite not knowing the real function of the *cpb*

C-terminal extension (Mottram et al. 1998; Rafati et al. 2001), this observation would suggest that this protein domain might play a role in the host/parasite interactions.

In conclusion, this work has helped to clarify the confusing classification of *cp* reported in the literature for the *Leishmania* genus and the Trypanosomatidae family. *Cpa* seems to be a cysteine protease gene specific to *Leishmania* as we found no homologous gene in *Trypanosoma*, whereas *cpb* and *cpc* appear ubiquitous not only in Trypanosomatidae but also reported in many eukaryotes. Our phylogeny was consistent with the systematics of the Trypanosomatidae reported in the literature, whatever the *cp* gene considered. Phylogenetic analysis brought new elements in the understanding of the evolution of *cp* genes in the Trypanosomatidae family and *Leishmania* genus. Investigation into the molecular adaptation of the *Leishmania cp* genes revealed significant positive selection for the three of them. In those three genes, positive selection was located in the core domain of the protein and its intensity was increasing from *cpc* to *cpa* to *cpb*, probably reflecting the different functions of these genes. In *cpb*, which is suspected to be the most implied in the host/parasite interactions, we noted particularly intense positive selection in the long C-terminal extension of this specific gene. This would suggest a potential implication of this domain in the host-parasite interactions. The structure of *cpb* in multicopy gene repeated in tandem furthermore makes it likely to stand as an archival antigenic library, designating this gene as a good target to understand how *Leishmania* are able to escape the host immune system.

Experimental procedures

Data sets

Published cysteine protease sequences of *Leishmania* were recovered from the European Bioinformatics Institute website (http://www.ebi.ac.uk). Additionally, we used a sample of cysteine protease gene sequences of other Trypanosomatidae in order to get a representative picture of cathepsin like cysteine proteases in the whole Trypanosomatidae family. These sequences were retrieved using WU-blastN from http://www.ebi.ac.uk/blast2/parasites.html and the NCBI website http://www.ncbi.nlm.nih.gov/entrez/query.fcgi. Accession numbers, names of organisms, and references of all the sequences used in the present study are detailed in table 1. The blast analysis allowed including 47 sequences belonging to the Trypanosomatidae family. Out of these 47 Trypanosomatidae sequences, 29 were drawn from three different complexes of the *Leishmania* subgenus of the *Leishmania* genus: *L. major*, *L. donovani* and *L. mexicana* complexes. Two sequences pertained to the *L. guyanensis* species of the *Viannia* subgenus of the *Leishmania* genus. Sixteen sequences came from different species of the *Trypanosoma* genus: *Trypanosoma cruzi*, *T. rangeli*, *T. brucei*, *T. congolense*, *T. brucei rodesiense* and *T. evansi*.

Additionally to these 47 Trypanosomatidae sequences, we considered one sequence from the *Cryptobia salmositica* species belonging to the Bodonidae family of the Kinetoplastida order (no cysteine protease sequence of Kinetoplastida species other than those of the Bodonidae and Trypanosomatidae families was available in GenBank), and one cysteine protease sequence of *Plasmodium falciparum*. Furthermore, to check the validity of the results of the positive selection analysis (see below), we also considered 11 sequences of different *Leishmania* species of the nucleoside hydrolase enzyme (EC 3.2.2.1). This enzyme is expected to be under strong conservation, because of its key metabolic role. This will act as a point of comparison when interpreting the results obtained on the *cp* genes. The accession numbers of these sequences named *nsnh* are: AY033633, AY533486, AY533489, AY533490, AY533491, AY533492, AY533493, AY533494, AY533501, AY533502, AY533503.

Sequence homology calculation

Percentages of sequence homology were calculated within the Kinetoplastida taxon, taking *L. infantum* as reference (see table 2). The results were obtained using the ALIGN program of the FASTA package (Myers and Miller 1988) available on http://www.infobiogen.fr/services/analyseq/cgi-bin/alignn_in.pl. As *cpb* displayed substantially longer C-terminal than *cpa* and *cpc*, all sequence homologies with a *cpb* were calculated both with and without the long *cpb* C-terminal extension.

Sequences alignments and phylogenetic analyses

Forty four nucleotide sequences noted F in the column "Analysis" of table 1 were aligned using Multiple Sequence Alignment Program, ClustalX, version 1.81 (Thompson et al. 1997). The N- and C-terminal sites were excluded because unalignable. Phylogenetic analyses were performed with the PHYLIP package (Felsenstein 1994) using the distance and parsimony methods with the SEQBOOT, DNADIST, DNAPARS, NEIGHBOR and CONSENSE programs. Various distance types were considered in the construction of trees: Kimura 2-parameter (Kimura 1980), Jukes-Cantor (Jukes and Cantor, 1969) and Maximum Likelihood (Felsenstein 1981). Editing of trees was done with Treedyn (Chevenet et al. 2000) and the cysteine protease sequence of *Plasmodium falciparum* was used as outgroup in all the phylogenetic analyses. In order to check the robustness of results, all the analyses were also performed on corresponding proteins using the adequate program in the PHYLIP package (Felsenstein 1994): SEQBOOT, PROTDIST, PROTPARS, NEIGHBOR and CONSENSE programs.

Analysis of positive selection

Analyses of positive selection were performed on alignments of *cpa*, *cpb* and *cpc* *Leishmania* gene sequences. The sequences used are noted PS in the column "Analysis" of the table 1. Additionally to these 3 genes, we considered the 11 *Leishmania* sequences of the nucleoside hydrolase, *nsnh* (see the Dataset subsection), a metabolism gene implied in the purine salvage and thus expected to be under strong conservation. Table 3 presents the number of sequences included in each alignment together with the number of codons in each sequence of the alignment. Positive selection is inferred through the comparison of the nonsynonymous and synonymous mutation rates d_N and d_S, with $\omega = d_N/d_S < 1$ corresponding to purifying (negative) selection, $\omega = 1$ corresponding to neutral evolution (absence of selection), and $\omega > 1$ indicating adaptive evolution (positive selection) (see review by Yang and Bielawski 2000). Estimations of ω ratios were performed in the ML framework developed by Yang et al. (2000) and which uses codon-based models of sequence evolution that account for phylogenetic structure, biases in codon usage and the transition/transversion (T_S/T_V) rate ratio. As this method is fundamentally codon-based, all portions of the gene consisting of overlapping reading frames were deleted before the analyses. Furthermore, in each alignment, gap-containing sites were excluded and the stop codon deleted (Yang et al, 2000). The PAUP package (Swofford 2000) was used to build ML trees for selection analysis using the HKY85+Γ model of nucleotide substitution with optimal values for the T_S/T_V rate ratio and the shape parameter (α) of a gamma distribution (with 8 categories) of rate variation among sites, both determined during tree construction. The ML method of Yang et al. (2000) utilizes codon-based models, which incorporate statistical distributions to account for variable ω ratios among codons. Efficient determination of sites under positive selection requires implementation of only six models of codon substitution (M0, M1, M2, M3, M7 and M8) out of the original 14 models (Yang et al. 2000). Null models M0, M1 and M7 do not allow for the existence of positively selected sites because ω is fixed or estimated between the bounds 0 and 1, whereas models M2, M3 and M8 do account for positive selection using parameters

that estimate $\omega > 1$. See Yang et al. (2000) for further details on the models. The significance of positive selection can be confirmed with a likelihood ratio test (LRT) between null models and those able to account for positive selection. A LRT is performed by taking twice the difference in log likelihood between nested models and comparing the result to a χ^2 distribution with degrees of freedom equal to the difference in the number of parameters between the models. Models M0 and M1 are both nested with M2 and M3, M2 is nested with M3, and M7 is nested with M8. All the model comparisons (M0 versus M2, M1 versus M2, M0 versus M3, M1 versus M3, M2 versus M3, and M7 versus M8) gave similar results and, for the sake of simplicity, we will focus on the results of models M7 and M8. M7 uses a discrete (10 classes) beta distribution to model sites with ω ratios between the bounds 0 and 1. For each class i ($1 \leq i \leq 10$) of the beta distribution, the value of the ω_i ratio, and the proportion p_i of sites belonging to this class, are estimated by maximizing the likelihood. M8 adds two additional parameters to model M7 such that p_{11} can account for a positively selected class of sites where ω_{11} is not constrained by the beta distribution and allowed to be greater than 1. Once positively selected sites have been shown to exist, *i.e.* if model M7 is rejected in favor of M8 by the LRT, a Bayesian approach (where the p_1 to p_{11} values are used as a prior distribution) is used to infer the posterior probabilities of each site i to belong to each of the eleven ω classes: $f_1^i, f_2^i, \ldots, f_{11}^i$. This further allowed us to define for each site i a weighted mean ω value as $\overline{\omega} = \sum_{k=1}^{11} f_k \cdot \omega_k$ as done in Choisy et al. (2004). Models were implemented using the CODEML program of the PAML package, version 3.1 (Yang 1997).

Acknowledgments
We would like to thank the discussion group on Steve Frank (2002)'s book organized by Sylvain Gandon as well as François Chevenet and Thierry De Meeûs for their useful discussions. Thanks to Philip Agnew for checking the English language. MC is supported by a BDI co-financed by CNRS and Région Languedoc-Roussillon. MH and ALB are supported by CNRS and IRD respectively.

Table 1
Cathepsin like cysteine protease DNA sequences used in analyses.

Accession number	Analysis*	Organism	Protein	Gene	Cp Group	Source
M81341	F	*Plasmodium falciparum*	cysteine proteinase	*tcp*	?	Rosenthal and Nelson. (1992)
AY090898	F	*Cryptobia salmositica*	cysteine proteinase	*Cysp1*	cpb	Hontzeas et al , 2002, Unpublished
AF309687	F, PS	*Leishmania donovani*	cysteine protease		cpa	Mundodi et al., 2000, Unpublished
AJ420285	F, PS	*L. infantum*	cathepsin L-like	*cpa*	cpa	Jimenez et al., 2001, Unpublished
AX002867	PS	*L. infantum*	unassigned		cpa	Coombs,.Mottram, 1998, Unpublished
AF004593	F, PS	*L. chagasi*	cysteine protease	*ldccys2*	cpa	Omara-Opyene and Gedamu (1997)
AJ130942	F	*L. major*	cysteine proteinase A	*cpa*	cpa	Rafati et al. (2001)
LMFL4766	PS	*L. major*	cysteine proteinase		cpa	Ivens et al. (1998)
X62163	F	*L. mexicana*	cysteine proteinase	*lmcpa*	cpa	Mottram et al. (1992)
L00717	F	*L. pifanoi*	cysteine proteinase	*lpcys1*	cpa	Traub-Cseko et al. (1993)
L29168	F, PS	*L. pifanoi*	cysteine proteinase	*cys1*	cpa	Almeida et al., 2000, Unpublished
AY141758	F, PS	*L. amazonensis*	cysteine proteinase	*Llacys1*	cpa	Lasakosvitsch et al. (2003)
AJ512652	F	*L. guyanensis*	cysteine protease a	*cpa*	cpa	Pascalis et al. (2003)
X16465	F	*Trypanosoma brucei*	cysteine proteinase		cpb	Mottram et al. (1989)
X54353	F	*T. b. rhodesiense*	cysteine protease		cpb	Parner et al. (1990)
AJ297265	F	*T. b. rhodesiense*	rhodesain		cpb	Caffrey et al. (2001)
AF165115	F	*T. evansi*	evansain		cpb	Gonzatti et al. 1999, Unpublished
AF139913	F	*T. congolense*	cysteine protease		cpb	Downey and Donelson (1999)
Z25813	F	*T. congolense*	cysteine protease		cpb	Fish et al. (1995)
L38513	F	*T. rangeli*	rangelipain		cpb	Martinez et al. (1995)
M84342	F	*T. cruzi*	cruzain		cpb	Eakin et al. (1992)
U41444	F	*T. cruzi*	cruzipain		cpb	Tomas and Kelly (1996)
U41454	F	*T. cruzi*	cruzipain		cpb	Tomas and Kelly (1996)
AF004594	F	*T. cruzi*	cysteine protease		cpb	Omara-Opyene et al.,1997,Unpublished
AF309627	F	*L. donovani*	cysteine protease		cpb	Mundodi et al., 2000, Unpublished
AF309626	F, PS	*L. donovani*	cysteine protease		cpb	Mundodi et al., 2000, Unpublished
AJ420286	F, PS	*L. infantum*	cathepsin L-like	*cpb*	cpb	Jimenez et al., 2001, Unpublished
AF004592	F, PS	*L. chagasi*	cysteine protease	*ldccys1*	cpb	Omara-Opyene and Gedamu (1997)
AF217087	F	*L. chagasi*	cathepsin L-like		cpb	Mundodi et al., 1999, Unpublished
U43706	F, PS	*L. major*	cathepsin L-like		cpb	Sakanari et al. (1997)
Z49963	F	*L. mexicana*	cysteine proteinase	*lmcpb1*	cpb	Mottram et al. (1997)
Z14061	F, PS	*L. mexicana*	cysteine proteinase	*lmcpb*	cpb	Souza et al. (1992)
AJ319727	F	*L. mexicana*	cathepsin L-like	*cpb2*	cpb	Brooks et al. (2001)
Z49962	F, PS	*L. mexicana*	cathepsin L-like	*lmcpb2.8*	cpb	Mottram et al. (1996)

Table 1 Continued

Y09958	F, PS	*L. mexicana*	cysteine proteinase	*cpb18*	cpb	Mottram et al. (1997)
M97695	F, PS	*L. pifanoi*	cysteine proteinase	*lpcys2*	cpb	Traub-Cseko et al. (1993)
L00718	F	*L. pifanoi*	cysteine proteinase	*lpcys2*	cpb	Traub-Cseko et al (1993)

AJ512653	F	*L. guyanensis*	cysteine protease b	*cpb*	cpb	Pascalis et al. (2003)
AC084046	F	*T. brucei*	chromosome VI		cpc	El-Sayed et al, 2000, Unpublished
AY508515	F	*T. brucei*	cathepsin B-like	*TbcatB*	cpc	Mackey and McKerrow, Unpublished
AF400046	F	*T. rangeli*	cathepsin B-like		cpc	Nobrega et al., 2001, Unpublished
AF399836	F	*T. cruzi*	cathepsin B-like	*tccb*	cpc	Nobrega et al. (1998)
AF399838	F	*T. cruzi*	cathepsin B-like	*tccb*	cpc	Nobrega et al. (1998)
AF233525	F, PS	*L. donovani*	cathepsin B-like	*cpc*	cpc	Mundodi et al. (2002)
AJ420287	F, PS	*L. infantum*	cathepsin-B-like	*cpc*	cpc	Jimenez et al., 2001, Unpublished
AF216830	F, PS	*L. chagasi*	cathepsin B-like		cpc	Mundodi et al. (2002)
U43705	F, PS	*L. major*	cathepsin B-like		cpc	Sakanari et al. (1997)
AC138434	PS	*L. major*	unassigned		cpc	Myler et al. 2002, Unpublished
Z48599	F, PS	*L. mexicana*	cathepsin B-like	*lmcpc*	cpc	Robertson and Coombs (1993)

* F = Sequences used for building the phylogram presented in figure 1
PS = Sequences used for the positive selection analyses presented in figure 2

Table 2

Percentages of sequence homology between *cp* genes and species in Kinetoplastida, taking *L. infantum* as reference.

	L. infantum cpa	*L. infantum cpb*	*L. infantum cpc*
~~*Leishmania infantum cpb* with C-terminal extension~~	~~49.4~~		~~47.2~~
L. infantum cpb without C-terminal extension	59.7	-	50.7
L. infantum cpc	48.7	-	-
L. chagasi cpa	99.0	-	-
L. chagasi cpb	-	90.0	-
L. chagasi cpc	-	-	99.4
L. major cpa	90.0	-	-
L. major cpb	-	86.6	-
L. major cpc	-	-	93.9
L. mexicana cpa	88.0	-	-
L. mexicana cpb	-	72.9	-
L. mexicana cpc	-	-	89.6
Trypanosoma cruzi cpb with C-terminal extension	46.7	58.0	-
T. cruzi cpb without C-terminal extension	59.0	-	-
T. cruzi cpc	-	-	61.7
Cryptobia salmositica with C-terminal extension	45.6	52.2	44.6
C. salmositica without C-terminal extension	53.6	-	48.2

Table 3

Positive selection in *Leishmania cp* genes : results of M8 model. Nucleoside hydrolase gene (*nsnh*) is used as reference gene under strong conservation. N° of sequences is the number of sequences in the alignment analyzed. N° of codons is the number of codons in each sequence of the alignment, Mean ω is the mean ω value as calculated from model M0, 11th class is the ω value in the positively selected (*i.e.* 11th) class of the M8 model, N° of sites is the number of sites whose posterior probability to belong to the positively (i.e. 11th) class is above the 0.95 threshold level, and P-value is the probability resulting from the likelihood ratio test between M7 and M8.

Gene	N° of sequences	N° of codons	Mean *w*	11th class	N° of sites	P-value
nsnh	11	291	0.2084	0.0001	0	0.884
cpc	6	340	0.4504	15.8500	1	0.018
cpa	7	353	0.5265	5.8909	2	0.001
cpb	8	443	0.8840	5.8122	3	<0.001

Figure legends

FIG.1.— Phylogeny of the *Leishmania* genus and other Trypanosomatidae based on cathepsin like cysteine protease DNA sequences. The tree was built by Wagner analysis after bootstrapping (with 100 repetitions) and the *Plasmodium falciparum* sequence was used as outgroup. Only relevant bootstrap values are shown. *Leishmania cp* gene sequences appear in dotted line. The units of the scale bar correspond to bootstrap value.

FIG.2. — Model M8 results for the *nsnh*, *cpc, cpa, cpb,* and *Leishmania* gene sequence alignments. The first column (a, c, e, g) shows the posterior probabilities of each site of the gene sequence to belong to the positively selected (*i.e.* 11th) class and the second column (b, d, f, h) shows the weighted mean ω values of each site (calculated as explained in the text) for *nsnh* (first row, a, b), *cpc* (second row, c, d), *cpa* (third row, e, f), and *cpb* (fourth row, g, h) gene sequence alignments. The horizontal lines show the 0.95 significant posterior probability threshold in the first column (a, c, e, g) and the threshold value of 1 for ω in the second column (b, d, f, h). Note the different y-axis scales in the subfigures of the second column (b, d, f, h).

References

Alexander J, Coombs G.H., Mottram J.C. (1998) *Leishmania mexicana* cysteine proteinase-deficient mutants have attenuated virulence for mice and potentiate a Th1 response. J Immunol **161**:6794-801.

Alvarez F, Cortinas M.N., Musto H. (1996) The analysis of protein coding genes suggests monophyly of *Trypanosoma*. Mol Phylogenet Evol **5**:333-43

Alves L.C., Judice W.A., St Hilaire P.M., Meldal M., Sanderson S.J., Mottram J.C., Coombs G.H., Juliano L., Juliano M.A. (2001) Substrate specificity of recombinant cysteine proteinase, CPB, of *Leishmania mexicana*. Mol Biochem Parasitol **116**:1-9.

Anisimova M, Bielawski J.P., Yang Z. (2001) Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol Biol Evol **18:**1585-92

Bart G, Coombs G.H., Mottram J.C. (1995) Isolation of *lmcpc*, a gene encoding a *Leishmania mexicana* cathepsin-B-like cysteine proteinase. Mol Biochem Parasitol **73**:271-4.

Bazan J.F., Fletterick R.J. (1988) Viral cysteine proteases are homologous to the trypsin-like family of serine proteases: structural and functional implications. Proc Natl Acad Sci U S A **85**:7872-6

Berti P.J., Storer A.C. (1995) Alignment/phylogeny of the papain superfamily of cysteine proteases. J Mol Biol **246**:273-83

Brooks D.R., Denise H., Westrop G.D., Coombs G.H., Mottram J.C. (2001) The stage-regulated expression of *Leishmania mexicana* CPB cysteine proteases is mediated by an intercistronic sequence element. J Biol Chem **276**:47061-9

Brooks D.R., Tetley L., Coombs G.H., Mottram J.C. (2000) Processing and trafficking of cysteine proteases in *Leishmania mexicana*. J Cell Sci **113** (Pt 22):4035-41

Caffrey C.R., Hansell E., Lucas K.D., Brinen L.S., Alvarez Hernandez A., Cheng J., Gwaltney S.L., 2nd, Roush W.R., Stierhof Y.D., Bogyo M., Steverding D., McKerrow J.H. (2001) Active site mapping, biochemical properties and subcellular localization of rhodesain, the major cysteine protease of *Trypanosoma brucei rhodesiense*. Mol Biochem Parasitol **118**:61-73

Chevenet F., Bañuls, A. L. ,Barnabé, C. (2000) TreeDyn: un éditeur interactif d'arbres phylogénétiques. In: G. Caraux OG, and M. F. Sagot (eds.) Actes des Premières Journées Ouvertes Biologie, Informatique et Mathématiques. ENSAM/LIRMM, Montpellier: 87-90

Choisy M, Woelk CH, Guégan J-F, Robertson DL (2004) Comparative study of adaptive molecular evolution in different human immunodeficiency virus groups and subtypes. J Virol **78**:1962-1970

Cupolillo E., Grimaldi G., Jr., Momen H. (1994) A general classification of New World *Leishmania* using numerical zymotaxonomy. Am J Trop Med Hyg **50**:296-311

Denise H., McNeil K., Brooks D.R., Alexander J., Coombs G.H., Mottram J.C. (2003) Expression of multiple CPB genes encoding cysteine proteases is required for *Leishmania mexicana* virulence in vivo. Infect Immun **71**:3190-5

Downey N., Donelson J.E. (1999) Expression of foreign proteins in *Trypanosoma congolense*. Mol Biochem Parasitol **104**:39-53

Eakin A.E., Mills A.A., Harth G., McKerrow J.H., Craik C.S. (1992) The sequence, organization, and expression of the major cysteine protease (cruzain) from *Trypanosoma cruzi*. J Biol Chem **267**:7411-20

Felsenstein J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol **17**:368-76

Felsenstein J. (1994) PHYLIP (phylogeny inference package). Version 3.5c. Distributed by the author, Department of Genetics, University of Washington, Seattle.

Fish W.R., Nkhungulu Z.M., Muriuki C.W., Ndegwa D.M., Lonsdale-Eccles J.D., Steyaert J. (1995) Primary structure and partial characterization of a life-cycle-regulated cysteine protease from *Trypanosoma* (*Nannomonas*) *congolense*. Gene **161**:125-8

Frame M.J., Mottram J.C., Coombs G.H. (2000) Analysis of the roles of cysteine proteinases of *Leishmania mexicana* in the host-parasite interaction. Parasitology **121**:367-77

Frank S.A. (2002) Immunology and Evolution of Infectious Disease. Princeton University Press, Princeton

Hughes A.L. (1994) Evolution of cysteine proteinases in eukaryotes. Mol Phylogenet Evol **3**:310-21

Hughes A.L., Piontkivska H. (2003) Phylogeny of Trypanosomatidae and Bodonidae (Kinetoplastida) Based on 18S rRNA: Evidence for Paraphyly of *Trypanosoma* and Six Other Genera. Mol Biol Evol **20**:644-52

Ivens A.C., Lewis S.M., Bagherzadeh A., Zhang L., Chan H.M., Smith D.F. (1998) A physical map of the *Leishmania major* Friedlin genome.Genome Res. **8**:135-45

Jukes T.H., Cantor, C. R. (1969) Evolution of protein molecules. In: Munro HN (ed) Mammalian Protein Metabolism. Academic Press, New York, p 21-132

Kimura M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol **16**:111-20

Lainson R., Shaw J.J. (1987) Evolution, classification and geographical distribution. In: R K-K (ed) The leishmaniases in biology and Medicine. Academic Press, London, p 1-120

Lalmanach G., Boulange A., Serveau C., Lecaille F., Scharfstein J., Gauthier F., Authie E. (2002) Congopain from *Trypanosoma congolense*: drug target and vaccine candidate. Biol Chem **383**:739-49

Lanotte G., Rioux J.A., Lepart J., Maazoun R., Pasteur N., Pratlong F. (1984) Numerical cladistics of the phylogeny of the genus *Leishmania* Ross, 1903 (Kinetoplastida-Trypanosomatidae). Use of enzyme characteristics. C R Acad Sci III **299**:769-72

Lasakosvitsch F., Gentil L.G., dos Santos M.R., da Silveira J.F., Barbieri C.L. (2003). Cloning and characterisation of a cysteine proteinase gene expressed in amastigotes of *Leishmania* (*L.*) *amazonensis*. Int J Parasitol **33**:445-54

Martinez J., Henriksson J., Rydaker M., Cazzulo J.J., Pettersson U. (1995) Genes for cysteine proteinases from *Trypanosoma rangeli*. FEMS Microbiol Lett **129**:135-41

McClelland M., Arensdorf H., Cheng R., Welsh J. (1994) Arbitrarily primed PCR fingerprints resolved on SSCP gels. Nucleic Acids Res **22**:1770-1

McKerrow J.H., Sun E., Rosenthal P.J., Bouvier J. (1993) The proteases and pathogenicity of parasitic protozoa. Annu Rev Microbiol **47**:821-53

Mottram J.C., Brooks D.R., Coombs G.H. (1998) Roles of cysteine proteinases of trypanosomes and *Leishmania* in host- parasite interactions. Curr Opin Microbiol **1**:455-60

Mottram J.C., Frame M.J., Brooks D.R., Tetley L., Hutchison JE., Souza A.E., Coombs G.. (1997) The multiple *cpb* cysteine proteinase genes of *Leishmania mexicana* encode isoenzymes that differ in their stage regulation and substrate preferences. J Biol Chem **272**:14285-93

Mottram J.C., North M.J., Barry J.D., Coombs G.H. (1989) A cysteine proteinase cDNA from *Trypanosoma brucei* predicts an enzyme with an unusual C-terminal extension. FEBS Lett **258**:211-5

Mottram J.C., Robertson C.D., Coombs G.H., Barry J.D. (1992) A developmentally regulated cysteine proteinase gene of *Leishmania mexicana*. Mol Microbiol **6**:1925-32

Mottram J.C., Souza A.E., Hutchison J.E., Carter R., Frame M.J., Coombs G.H. (1996) Evidence from disruption of the *lmcpb* gene array of *Leishmania mexicana* that cysteine proteinases are virulence factors. Proc Natl Acad Sci U S A **93**:6008-13.

Mundodi V., Somanna A., Farrell P.J., Gedamu L. (2002) Genomic organization and functional expression of differentially regulated cysteine protease genes of *Leishmania donovani* complex. Gene **282**:257-65.

Myers E.W., Miller W. (1988) Optimal alignments in linear space. Comput Appl Biosci **4**:11-7

Nobrega O.T., Santos Silva M.A., Teixeira A.R., Santana J.M. (1998) Cloning and sequencing of tccb, a gene encoding a *Trypanosoma cruzi* cathepsin B-like protease. Mol Biochem Parasitol **97**:235-40

North M.J., Robertson C.D., Coombs G.H. (1990) The specificity of trichomonad cysteine proteinases analysed using fluorogenic substrates and specific inhibitors. Mol Biochem Parasitol **39**:183-93

Okenu D.M., Opara K.N., Nwuba R.I. Nwagwu M. (1999) Purification and characterisation of an extracellularly released protease of *Trypanosoma brucei*. Parasitol Res **85**:424-8

Omara-Opyene A.L., Gedamu L. (1997) Molecular cloning, characterization and overexpression of two distinct cysteine protease cDNAs from *Leishmania donovani chagasi*. Mol Biochem Parasitol **90**:247-67.

Pamer E.G., Davis C.E., Eakin A., So M. (1990) Cloning and sequencing of the cysteine protease cDNA from *Trypanosoma brucei rhodesiense*. Nucleic Acids Res **18**:6141

Pascalis H., Lavergne A., Bourreau E., Prevot-Linguet G., Kariminia A., Pradinaud R., Rafati S., Launois P. (2003) Th1 cell development induced by cysteine proteinases A and B in localized cutaneous leishmaniases due to *Leishmania guyanensis*. Infect Immun **71**:2924-6

Pollock K.G., McNeil K.S., Mottram J.C., Lyons R.E., Brewer J.M., Scott P., Coombs G.H., Alexander J. (2003) The *Leishmania mexicana* cysteine protease, CPB2.8, induces potent Th2 responses. J Immunol **170**:1746-53

Rafati S., Salmanian A., Hashemi K., Schaff C., Belli S., Fasel N. (2001) Identification of *Leishmania major* cysteine proteinases as targets of the immune response in humans. Mol Biochem Parasitol **113**:35-43

Robertson C.D., Coombs G.H. (1990) Characterisation of three groups of cysteine proteinases in the amastigotes of *Leishmania mexicana mexicana*. Mol Biochem Parasitol **42**:269-76

Robertson C.D., Coombs G.H. (1993) Cathepsin B-like cysteine proteases of *Leishmania mexicana*. Mol Biochem Parasitol **62**:271-9

Robertson C.D., Coombs G.H. (1994) Multiple high activity cysteine proteases of *Leishmania mexicana* are encoded by the *Imcpb* gene array. Microbiology **140**:417-24.

Rosenthal P.J. (1999) Proteases of protozoan parasites. Adv Parasitol **43**:105-59

Rosenthal P.J., Nelson R.G. (1992). Isolation and characterization of a cysteine proteinase gene of *Plasmodium falciparum*. Mol Biochem Parasitol. **51**:143-52

Sajid M., McKerrow J.H. (2002) Cysteine proteases of parasitic organisms. Mol Biochem Parasitol **120**:1-21

Sakanari J.A., Nadler S.A., Chan V.J., Engel J.C., Leptak C., Bouvier J. (1997) *Leishmania major*: comparison of the cathepsin L- and B-like cysteine protease genes with those of other trypanosomatids. Exp Parasitol **85**:63-76

Selzer P.M., Pingel S., Hsieh I., Ugele B., Chan V.J., Engel J.C., Bogyo M., Russell D.G., Sakanari J.A., McKerrow J.H. (1999) Cysteine protease inhibitors as chemotherapy: lessons from a parasite target. Proc Natl Acad Sci U S A **96**:11015-22

Smith J.D., Chitnis C.E., Craig A.G., Roberts D.J., Hudson-Taylor D.E., Peterson D.S., Pinches R., Newbold C.I., Miller L.H. (1995) Switches in expression of *Plasmodium falciparum var* genes correlate with changes in antigenic and cytoadherent phenotypes of infected erythrocytes. Cell **82**:101-10

Souza A.E., Waugh S., Coombs G.H., Mottram J.C. (1992) Characterization of a multi-copy gene for a major stage-specific cysteine proteinase of *Leishmania mexicana*. FEBS Lett **311**:124-7

Stanley S.L., Jr., Zhang T., Rubin D., Li E. (1995) Role of the *Entamoeba histolytica* cysteine proteinase in amebic liver abscess formation in severe combined immunodeficient mice. Infect Immun **63**:1587-90

Stevens J.R., Noyes H.A., Schofield C.J., Gibson W (2001) The molecular evolution of Trypanosomatidae. Adv Parasitol **48**:1-56

Suzuki Y., Nei M. (2002) Simulation study of the reliability and robustness of the statistical methods for detecting positive selection at single amino acid sites. Mol Biol Evol **19**:1865-9

Swofford D.L. (2000) Phylogenetic analysis using parsimony (and other methods). Sinauer Associates, Sunderland

Thomaz-Soccol V., Lanotte G., Rioux J.A., Pratlong F., Martini-Dumas A, Serres E (1993a) Monophyletic origin of the genus *Leishmania* Ross, 1903. Ann Parasitol Hum Comp **68**:107-8

Thomaz-Soccol V., Lanotte G,. Rioux J.A., Pratlong F., Martini-Dumas A., Serres E. (1993b) Phylogenetic taxonomy of New World *Leishmania*. Ann Parasitol Hum Comp **68**:104-106

Thompson J.D., Gibson T.J., Plewniak F., Jeanmougin F., Higgins D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res **25**:4876-82

Tomas A.M., Kelly J.M. (1996) Stage-regulated expression of cruzipain, the major cysteine protease of *Trypanosoma cruzi* is independent of the level of RNA1. Mol Biochem Parasitol **76**:91-103

Traub-Cseko Y.M., Duboise M., Boukai L.K., McMahon-Pratt D. (1993) Identification of two distinct cysteine proteinase genes of *Leishmania pifanoi* axenic amastigotes using the polymerase chain reaction. Mol Biochem Parasitol **57**:101-15.

Wagner A. (2000) Decoupled evolution of coding region and mRNA expression patterns after gene duplication: implications for the neutralist-selectionist debate. Proc Natl Acad Sci USA **97**:6579-84

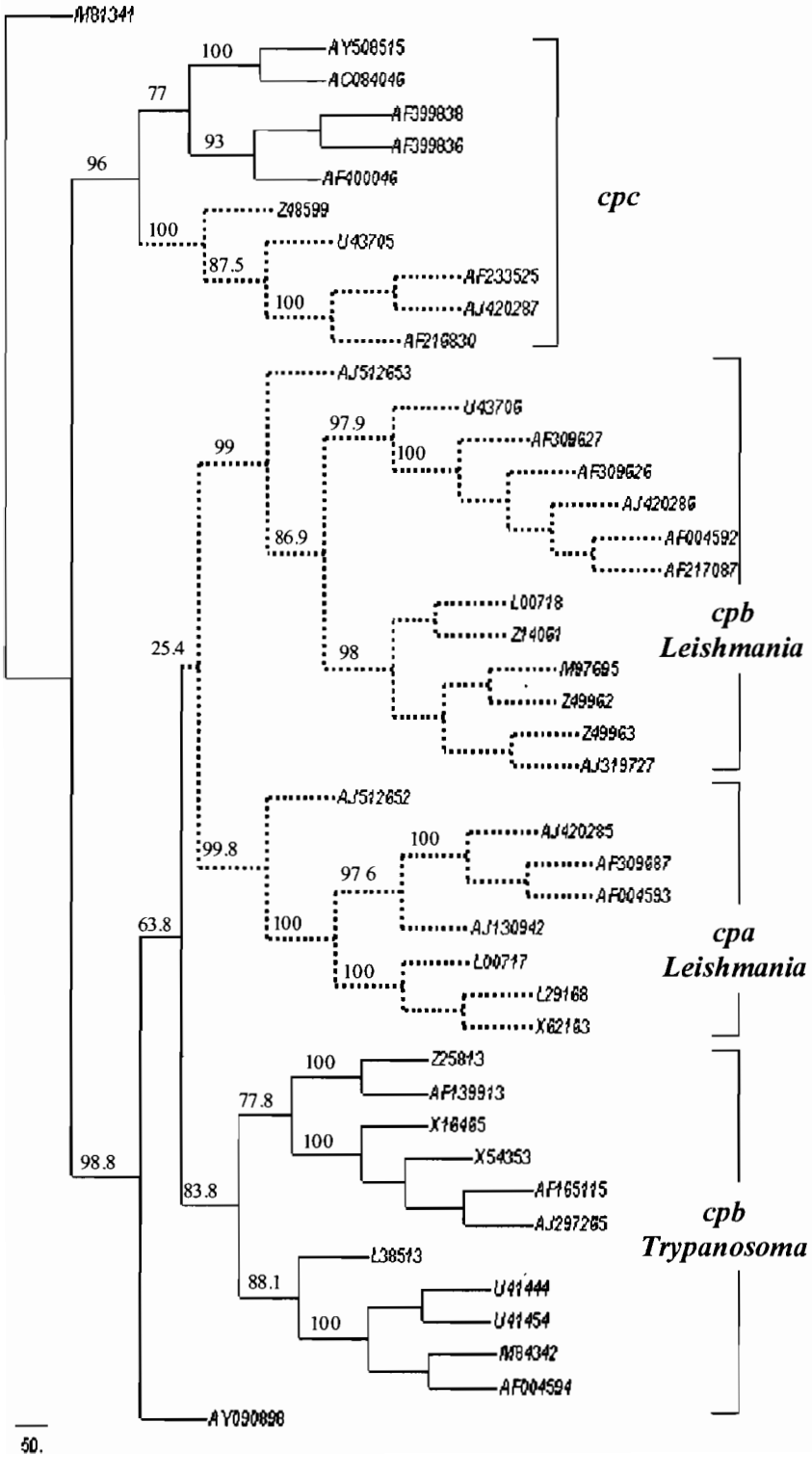
Yang Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci **13**:555-6

Yang Z., Bielawski J.P. (2000) Statistical methods for detecting molecular adaptation. TREE **15**:496-503

Yang Z., Nielsen R., Goldman N., Pedersen A.M. (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155**:431-49

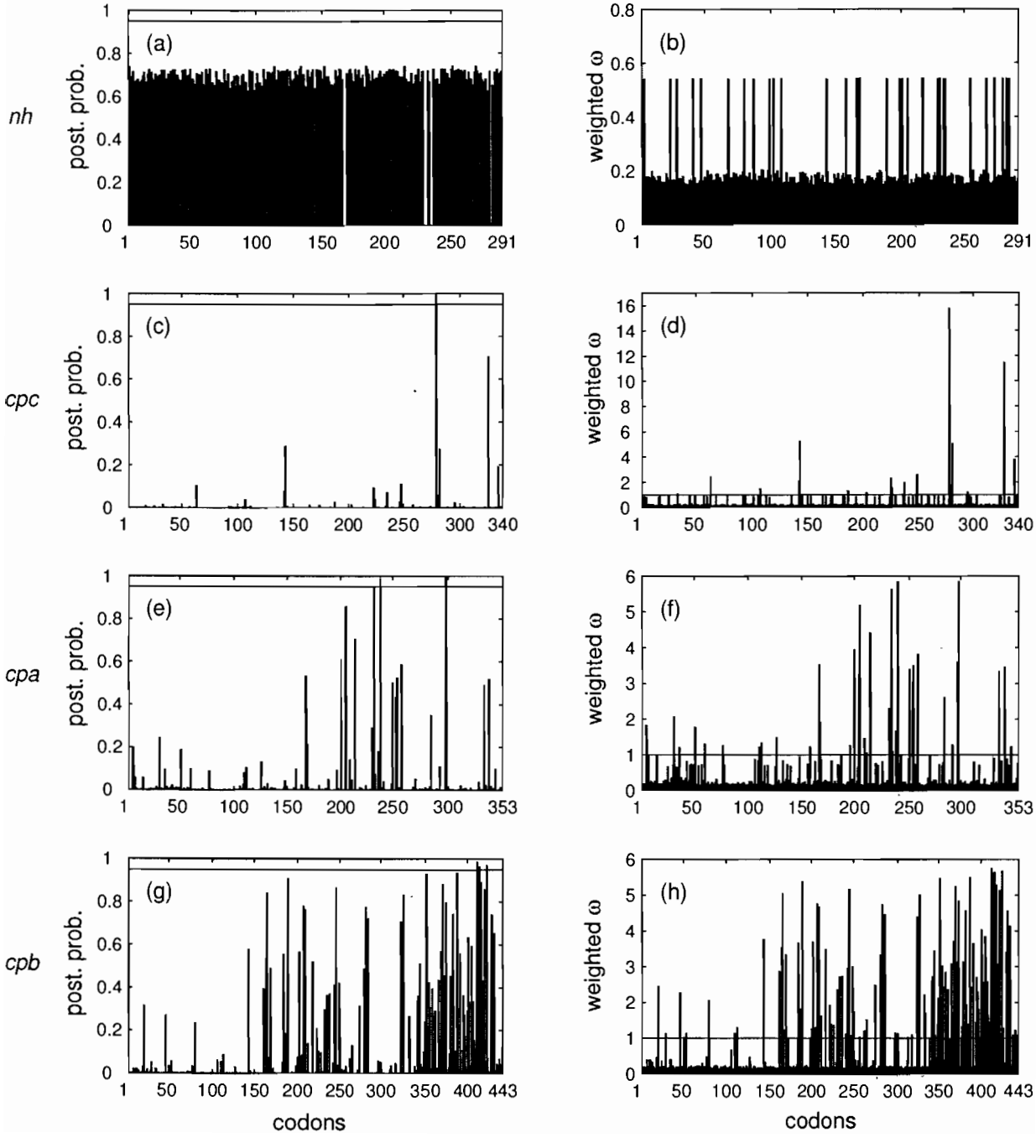
Zhang J.1. (2003) Evolution by gene duplication: an update. TREE **18**:292-298

Figure 1



Organism	Product and symbol
Plasmodium falciparum	cysteine proteinase, *tcp*
Trypanosoma brucei	cathepsin B-like, *tbcatB*
T. brucei	chromosome VI sequence
T. cruzi	cathepsin B-like, *tccb*
T. cruzi	cathepsin B-like, *tccb*
T. rangeli	cathepsin B-like
Leishmania mexicana	cathepsin B-like, *lmcpc*
L. major	cathepsin B-like
L. donovani	cathepsin B-like
L. infantum	cathepsine-B-like, *cpc*
L. chagasi	cathepsin B-like
L. guyanensis	cysteine proteinase, *cpb*
L. major	cathepsin L-like
L donovani	cysteine protease
L donovani	cysteine protease
L chagasi	cathepsin L-like
L chagasi	cysteine protease, *ldccys1*
L infantum	cathepsin L-like, *cpb*
L. pifanoi	cysteine proteinase, *lpcys2*
L. mexicana	cysteine proteinase, *lmcpb*
L pifanoi	cysteine proteinase, *lpcys2*
L. mexicana	cathepsin L-like, *lmcpb2.8*
L. mexicana	cysteine proteinase, *lmcpb1*
L. mexicana	cathepsin L-like, *cpb2*
L. guyanensis	cysteine proteinase, *cpa*
L infantum	cathepsin L-like, *cpa*
L donovani	cysteine protease
L chagasi	cysteine protease, *ldccys2*
L. major	cysteine proteinase A, *cpa*
L pifanoi	cysteine proteinase, *lpcys1*
L pifanoi	cysteine proteinase, *cys1*
L. mexicana	cysteine proteinase, *lmcpa*
T. congolense	cysteine protease
T. congolense	cysteine protease
T. brucei	cysteine proteinase
T. brucei rhodesiense	cysteine protease, rhodesain
T. evansi	cysteine protease, evansain
T. b. rhodesiense	cysteine protease
T. rangeli	cysteine proteinase, rangelipain
T. cruzi	cysteine protease, cruzipain
T. cruzi	cysteine protease, cruzipain
T. cruzi	cysteine protease, cruzain
T. cruzi	cysteine protease
Cryptobia salmositica	cysteine proteinase, *Cysp1*

Figure 2



Annexe D

CHOISY M., & GUÉGAN J.-F. (2005) Modeling of infectious diseases : when theory meets reality. In *Encylcopedia of Infectious Diseases– Modern Methods* (TIBAYRENC, M. Ed.) John Wiley & Sons, Chichester, USA

20 Modeling of infectious diseases: when theory meets reality

M. CHOISY

Génétique et Evolution des Maladies Infectieuses
UMR CNRS - IRD, Montpellier, France

"As a matter of fact all epidemiology, concerned as it is with variation of disease from time to time or from place to place, must be considered mathematically (...), if it is to be considered scientifically at all. (...) And the mathematical method of treatment is really nothing but the application of careful reasoning to the problems at hand."

—Sir Ronald Ross MD, 1911

20.1 INTRODUCTION

The concealed and apparently unpredictable nature of infectious diseases has been a source of fear and superstition since the first ages of human civilizations (see chapter 26). The worldwide panic consecutive to the SARS and avian flu emergences in the south-east Asia are recent examples that our feeling of dread increases with our ignorance of the disease [45]. One of the primary aims of epidemic modeling is helping to understand the spread of diseases in host populations, both in time and space. Indeed, in clarifying rigorously the assumptions, the variables and the parameters, the process of model formulation allows identifying precisely the underlying causes of the observed patterns. The very first epidemiological model was formulated by Daniel Bernoulli in 1760 [11] in order to evaluate the impact of variolation on human life expectancy. However, this was not followed by others until the beginning of the 20th century[1] with the pioneering works of Hamer [29] on measles and Ross [51] on malaria. Since then, the last century has witnessed the emergence and development of a real theory of epidemics. In 1927 Kermack and McKendrick [38] obtained the threshold theorem, one of the key results in epidemiology, stating that the number

[1]During the 19th century research activity on infectious diseases was dominated by the clinical studies of the Pasteur school.

of susceptibles need to exceed a critical value for an epidemic to occur. Bailey [8] published its landmark book in 1957 and the efficacy of epidemic modeling in public health concretized at the end of the seventies with the worldwide eradication of smallpox [7]. Given the diversity of infectious diseases studied since the middle of the fifties, an impressive variety of epidemiological models have been developed. A comprehensive review of them would be both beyond the scope of the present chapter and of limited interest. Rather, the idea here is to introduce the reader to the most important notions of epidemic modeling based on the presentation of the classic SIR models.

After presenting general notions of mathematical modeling (section 20.2 and the nature of epidemiological data available to the modeler (section 20.3), we detail the very basic SIR epidemiological model (section 20.5). We insist on the assumptions made on the biological processes and their consequences from an epidemiological point of view. We then review more complex models that allows to study endemic diseases (section 20.6) and the recurrence of epidemic outbreaks (section 20.7). Section 20.8 then focuses of the analysis of epidemiological data and the estimation of model parameters. The chapter ends with some examples of practical uses of models for the development of public health policies (section 20.9). Technical aspects are treated in boxes.

20.2 THE PHILOSOPHY OF MATHEMATICAL MODELING

Epidemiology is essentially a population biology discipline related to public health preoccupations. As a population biology subject, epidemiology is thus dominated by a mathematical theory. The reason is that most phenomena observed at a population level are often complex and difficult to deduce from the characteristics of an isolated member. For example, the prevalence of a disease in a population is only indirectly connected to the course of disease in an individual. In this context, the use of mathematical models is to splice what is going on into a whole picture.

20.2.1 Model complexity

A model is a caricature of reality as represented by the data. Models help understanding reality because they simplify it. Consequently, a model is by essence false. There are no good or bad models. There are simply models which are better representing the reality than others – we usually say better fitted to the data. What is essential to understand here is that a model does not need to be totally correct to be useful. A model should be just exact enough for the purpose under study.

We now proceed to introduce some of the modeler's vocabulary. A state variable is a changing quantity that characterizes the state of the system. For example, the number of infectives and susceptibles in the population are state variables of an epidemiological system. The modeler is interested in the behavior of the state variable. A parameter is an user-defined quantity that influences the value of the

state variables. For example, the average duration an individual stays infective is a parameter of an epidemiological system. The fit of a model to a data set is basically influenced by two aspects [32]. The first one is related to the complexity of the model as given by the number of parameters. Complicated models with more parameters will usually give better fits to data than simpler models. However, simpler models often provide insight that is more valuable and influential in guiding thought. The choice of the optimal level of complexity obeys a trade-off between bias and variance [14] (see box 20.1). The second aspect is related to the exact relationship between the parameters. For example, should the transmission process be linear or nonlinear? Again, it is important to realize that the nature of such a relation does not need to be totally correct for the model to be useful. Modelers speak of structural stability to designate little changes in the model prediction when the model itself changes a little bit.

20.2.2 Model formulation and hypothesis testing

A mathematical model is a set of equations. Equations are the mathematical translation of hypotheses (or assumptions). When interpreting the model predictions it is thus important to bear in mind the underlying assumptions. By definition, an assumption is an unverified proposition, tentatively accepted to explain certain facts or to provide a basis for further investigation. For example one can construct a model 1 assuming that the probability a susceptible get infected is proportional to the number of infectives and a model 2 assuming that this probability is independent of the number of infectives. In such a case of competing hypotheses, the data can act as an arbitrator by telling which model fit the data the better [56]. In modern statistics the fit of a model to a data set is measured by its likelihood (see box 20.2). Comparison of models is thus based on the comparison of their likelihoods. Since the likelihood of a model naturally increases as the number of its parameters increases (see above section 20.2.1), it is necessary that the likelihood comparisons correct for the complexities of the models. There exist two major procedures for model likelihood comparison [34].

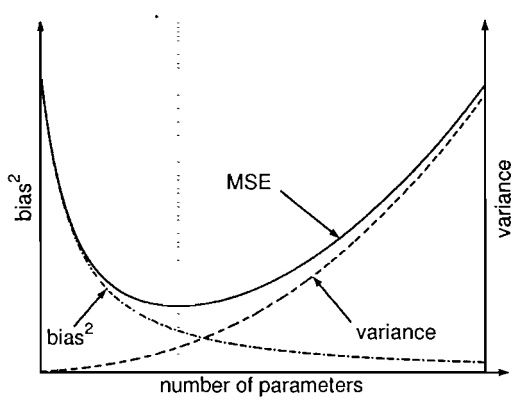
In a classic null hypothesis approach, the likelihood ratio test (LRT) is the most commonly used procedure. Two models are said nested when one is a particular case of the other. Twice the difference between log-likelihoods of two nested models follows a χ^2 distribution with degree of freedom equal to the difference between the numbers of parameters of the two models. A more complex model is thus retained if its likelihood is significantly higher than the one of a simpler model, as judged from the χ^2 statistic.

We can alternatively use model selection criteria to rank any (nested or not) competitive models. These criteria are basically constructed as a likelihood value corrected for the complexity of the model. The most used of these criteria is the Akaike information criterion (AIC) [2] defined as AIC = $2(p\text{-LL})$ where p is the number of parameters and LL is the logarithm of the likelihood.

BOX 20.1

HOW COMPLEX SHOULD A MODEL BE?

With the current power of desktop microcomputers it is tempting to build very complex models in order to fit the data the most. However, fitting the most complex model is not necessary always the best solution. Indeed, the more complex a model, the more difficult the interpretation of its outputs. Also, if a model is too complex, the modeler may not have sufficient information in the data to distinguish between the possible parameter values of the model. As said in the main text, the best-sized model depends on the purpose of the model. Given this objective, there exist quantitative methods for determining the optimal size of a model. These approaches are based on a trade-off between prediction error due to approximation (i.e. bias) which decreases as model complexity increases, and prediction error due to estimation (i.e. variance) which increases as model complexity increases as shown on the figure below [14]. The consequence is that for any model and amount of data, the total prediction error (proportional to the mean squared error) will decrease and then increase as model complexity increases, thus evidencing an optimal level of model complexity.



The mean squared error (MSE) is equal to MSE = variance+bias^2. As the number of parameters increases the bias decreases and the variances increases, defining an optimum number of parameters corresponding to the minimum of the MSE, as materialized by the vertical dotted line on the above figure.

20.2.3 Stochastic versus deterministic models

Deterministic models account for the mean trend only and thus have no components that are inherently uncertain. Stochastic models account not only for the mean trend but also for the variance structure around the mean trend. Consequently, some parameters of stochastic models are uncertain and characterized by a probability

BOX 20.2

LIKELIHOOD FUNCTIONS

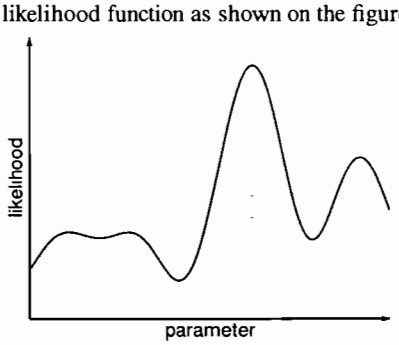
The likelihood of a model is a measure of the probability that the model is the appropriate description of the reality, given the data: L(model | data) = Pr(model | data). One powerful point of the likelihood function is that the term "model" includes not only the mean trend but also the variance, *i.e.* the distribution of the errors around the mean trend. Whereas the classical least square method implicitly assumes a normal distribution of errors, the likelihood methods allow considering any error distribution. For example suppose that **d** is a vector of data and **m** a vector of model predictions with a mean trend depending on one parameter x. Assuming now that the errors are normally distributed with a variance σ^2 then the likelihood of one prediction of the model reads

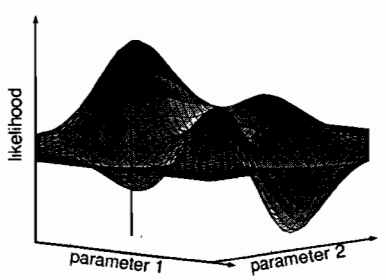
$$\mathrm{L}(\mathbf{m}_i(x), \sigma^2 | \mathbf{d}_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{(\mathbf{d}_i - \mathbf{m}_i)^2}{2\sigma^2}\right]$$

If the vector of data is a time series – as often the case for epidemiological data – then the data points are not independent. However, if the noise has a large magnitude – as often the case for epidemiological data too – the approximation of independency between the data points becomes acceptable. In that case the likelihood function of the model reads

$$\mathrm{L}(\mathbf{m}(x), \sigma^2 | \mathbf{d}) = \prod_i \mathrm{L}(\mathbf{m}_i(x), \sigma^2 | \mathbf{d}_i) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{(\mathbf{d}_i - \mathbf{m}_i)^2}{2\sigma^2}\right]$$

We thus end with a function which depends on two parameters x and σ^2. This likelihood function can be used for two different purposes. First, this function can be used to estimate parameters x and σ^2, good estimations of them being values that maximize the likelihood function as shown on the figures below with one and two parameters.





If the search of the maximum of the likelihood is straightforward when the function depends on one parameter, it becomes more complicated when the number of parameters increases. Microcomputers now allow the use of efficient numerical algorithm to find the maximum of such multiple dimensional surfaces. Among the most popular are the Newton and the Nelder-Mead algorithms [48]. Second, the expression of a likelihood function allows the comparison of different competing models, using either the likelihood ratio test or the Akaike information criterion (see main text).

distribution instead of a constant value. For fixed starting values, a deterministic model will always produce the same result whereas a stochastic model will produce many different outputs, depending on the actual values the random variables take. When a phenomenon is the sum of a large number of small individual effects (as disease propagation in large population), the weak law of large numbers reduce the effect of stochasticity and a deterministic model becomes appropriate. On the contrary, when the population is of small size, random events cannot be neglected and a stochastic model is necessary.

20.3 THE NATURE OF EPIDEMIOLOGICAL DATA

Epidemiology is fundamentally a data-driven discipline and a key element in this research field is being able to link mathematical models to data. Epidemiological data are generally based on the disease notifications reported by medical doctors. Epidemiologists usually consider incidences defined as the number of new cases per unit of time and prevalences referring to the number of diseased people, ideally at one instant, and in practice over a short period of time. Incidences thus reflect the dynamics of the disease whereas prevalence is more related to the static properties of the disease. Epidemiological data may further be stratified by age, sex, social status, geographical location, *etc*... We will in section 20.6.2 see that stratification by age is of particular interest since age reflects time [7]. Moreover the survey can be carried out longitudinally (*i.e.* through time) or horizontally (*i.e.* at one instant or over a short period of time). In the first case where the data are in form of a time series it is important to realize that the data of the series are not independent. Indeed, the number of infectives for a given week is likely to be close to the number of infectives the week before. In consequence, the statistical analysis of time series requires the use of specific tools presented in section 20.8.2. Epidemiological data sets are often accompanied by demographic data such as the population size and birth rate in different localities and at different dates. This is of primary interest as the endemic state of an infectious disease is often dependent on host social and demographic factors.

Such data sets currently exit for a variety of diseases, in different locations and over several decades. Some of these data bases are available from the internet, as the one used to draw figures 20.6, 20.10, and 20.12. Other data sets can easily be requested from governmental health services. The quality of the data set is often related to its accuracy in terms of disease diagnose, spatial location and notification frequency (weekly, monthly, or yearly).

20.4 CHILDHOOD MICRO-PARASITIC INFECTIONS

There exist a variety of epidemiological models and an exhaustive review of them cannot be performed in one single chapter. In consequence, we will focus our attention

on some of the most used ones in order to highlight the approach and the main results. We will thus be specifically interested here in childhood micro-parasitic infections. The distinction between micro-parasites and their counterpart macro-parasites is not clear-cut and actually reflects more the way they are modeled than biological realities [7]. However, micro-parasites tend to refer to small-size parasites (viruses, bacteria, or protozoan) with fast and direct reproduction within the host. The disease transmits by direct contact through the spit of aerial droplets and the infectiousness is generally high. The host usually recovers from the infection and acquires immunity for some time (and often for life). The disease generation length (*i.e.* the duration between the infection and the clearance by the host immune system) is generally short relative to the host life expectancy. Because of the fast and direct reproduction within the host, it makes sense to model the dynamics of micro-parasitic diseases according to the host clinical status with compartmental models. We call childhood diseases those diseases which confer a life-long immunity. As the infectiousness of micro-parasitic diseases is usually high, the life-long immunity makes the mean age at infection generally low, hence the name. Common childhood micro-parasitic infectious diseases include measles, rubella, chickenpox, mumps, whooping cough, *etc*...

20.5 A SIMPLE EPIDEMIC MODEL

The idea behind compartmental models is to divide the host population into a set of distinct classes, according to its epidemiological status. One simple such model is the SIR one which classifies individuals as susceptibles to the disease (S ind.), infectious (I ind.), and recovered (R ind.). The total size of the host population is then $N = S + I + R$. For childhood diseases there is no vertical transmission and thus individuals are born in the susceptible class. Upon contact with an infectious individual susceptible people may get infected and move into the infectious class. Once the immune system clears the infectious agents from the organism, infected people get immune and move to the recovered class (figure 20.1).

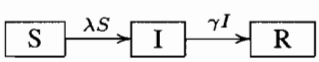


Fig. 20.1 A simple SIR epidemic model. The host population is divided into three compartments, according to their epidemiological status: susceptibles (S ind.), infectives, (I ind.), and recovered (R ind.). Individuals move to the susceptible class to the infective class, to the recovered class according to the arrows. λ is the force of infection, *i.e.* the probability that an susceptible individual get infected, and γ is the recovery rate.

20.5.1 Transmission process

The transmission process is at the heart of any epidemiological model. To describe it, epidemiologists usually consider the force of infection λ defined as the per capita

rate of acquisition of the infection. More precisely, $\lambda(t)\Delta t$ is the probability that a given susceptible individual will acquire the infection in the small interval of time Δt [31].

For airborne disease, the tradition has long been to consider the force of infection proportional to the number of infectious individual: $\lambda = \alpha I$. There is thus an analogy with the concentration of two chemical agents to which the law of mass action applies. However, humans obviously do not behave in exactly the same way as molecules in solutions since the daily contact patterns of people are often similar in large and small communities [31, 43].

Consider instead that the average number of contacts of a person per unit time is the constant β combining a multitude of epidemiological, environmental, and social factors that affect transmission rates [7]. Among these contacts, the number of contact with infectives is thus $\beta I/N$. Assuming that contacts are sufficient for transmission, the number of new cases per unit time is then $S\beta I/N$. Thus, in this case $\lambda = \beta I/N$, instead of $\lambda = \alpha I$. Fits to real data have proved that the frequency-dependant transmission process $\lambda = \beta I/N$ is more appropriate for human airborne diseases than the density-dependant one $\lambda = \alpha I$ (Anderson & May 1991). The parameter α has no clear epidemiological interpretation but can be related to β as $\alpha = \beta/N$. McCallum *et al.* [43] explored other forms of the transmission process, including nonlinear ones, and studied their influence on the epidemiological conclusions.

20.5.2 Between-compartment flux of individuals

A common assumption is that the movements out of one compartment into the next one are governed by constant rates [7]. For each time unit a constant number of individuals leave one compartment to the next, regardless to the time they spent in their compartment. The choice of this assumption is essentially motivated by an ease of mathematical tractability in a deterministic setup with ordinary differential equations. However the assumption of a flux of individual at a constant rate r corresponds to exponentially distributed waiting times in the compartments. The parameter of the negative exponential distribution is r and thus the mean of the distribution $1/r$ (see box 20.3). Analysis of real data reveals instead that each individual tends to spend a constant duration in each compartment [40, 36]. Models accounting for such realistic distributions of waiting times would implies the use of more sophisticated mathematics such as integro- or delay-differential equations. For didactic reasons we will here restrict our attention to the simplest and most used models based on simple ordinary differential equations and refer the reader interested in more realistic ones to [40] and [36]. Keeling [36] showed that the assumptions on the waiting times can strongly influence the model outputs.

BOX 20.3

MODELING THE INFECTIOUS PERIOD

In the classic SIR model it is usually assumed that the individuals leave the infectious class at a constant rate. Even if this assumption seems the most intuitive, it is not always the most realistic in terms of the duration individuals stay infective. In this box we detail the consequences of the constant recovery rate assumption on the distribution of the infective periods, and propose an alternative which yields more realistic distributions [40, 36].

Our random variable is the time of recovery since the infection. For discrete random variables (*e.g.* number of individuals) it is easy to define a probability distribution $\Pr\{Z = k\} = f_k$ (as in section 20.5.5) and then to define a cumulative distribution function $F(z) = \Pr\{Z \leq z\}$. For continuous variables like here the time of recovery since infection, it is impossible to define a probability of each time as there is an infinity of such times. The approach is then to first define a cumulative distribution and then express a probability density function from this cumulative distribution. The idea is to consider the probability associated with a short interval Δz of the random variable z.

$$\begin{aligned} \Pr\{z \leq Z \leq z + \Delta z\} &= F(z + \Delta z) - F(z) \\ &= F'(z)\Delta z + \mathrm{o}(\Delta z) \end{aligned}$$

The derivative $F'(z)$ of the cumulative distribution $F(z)$ is by definition the probability density function.

Let us now apply this method to the time of recovery since the infection. As done is section 20.6.3, we can express the probability of an infective to recover in the time interval Δt as

$$\Pr\{\text{recovery in } (t, t + \Delta t] | \text{no recovery in } (0, t]\} = \gamma \Delta t + \mathrm{o}(\Delta t)$$

where t is the time since infection and γ is a fixed constant. The cumulative distribution is defined as $F(t) = \Pr\{\text{no recovery in } (0, t]\}$. For an infective not to recover in the interval $(0, t + \Delta t]$, he must first not recover in the interval $(0, t]$ and then not recover in the next Δt. Assuming that these events are independent gives

$$\begin{aligned} F(t + \Delta t) &= F(t)[1 - \gamma \Delta t + \mathrm{o}(\Delta t)] \\ \Leftrightarrow \frac{F(f + \Delta t) - F(t)}{\Delta t} &= -\gamma F(f) + \frac{\mathrm{o}(\Delta t)}{\Delta t} \end{aligned}$$

Taking the limit as $\Delta t \to 0$ gives

$$\frac{dF}{dt} = -\gamma F(t)$$

which, after integration and setting $F(0) = 1$ (*i.e.* no recovery before the infection), yields

$$F(t) = -\mathrm{e}^{-\gamma t}$$

Thus, infectious periods are exponentially distributed with a mean infectious duration equal to $1/\gamma$, see dashed curve on the figure below. Inspecting real data, it seems that the

(Box 20.3 continued)

infectious period does not follows an exponential distribution but rather seems to be of constant duration. To account for such more realistic distributions, we need to relax the assumption that the probability of recovery does not depend on the time since infection. There are several ways to do that, including integro-differential and partial differential formulations, but the simplest one is certainly the method of stages.

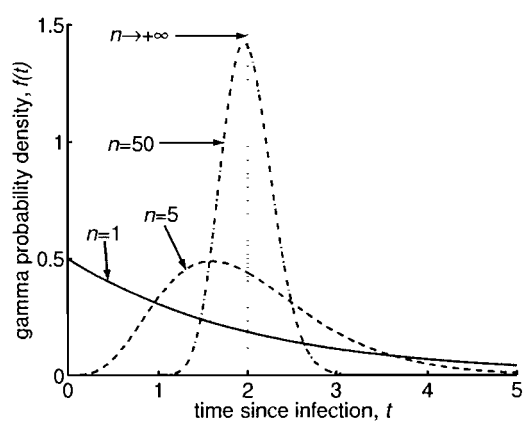
The basic idea of the method of stages is to replace the infective compartment by a series of n successive infective compartments, each with an exponential distribution of the same parameter:

$$\boxed{S} \xrightarrow{\lambda S} \boxed{I_1} \xrightarrow{\gamma I_1} \boxed{I_2} \xrightarrow{\gamma I_2} \cdots\cdots \xrightarrow{\gamma I_{n-1}} \boxed{I_n} \xrightarrow{\gamma I_n} \boxed{R}$$

The total duration of the infectious period is thus the sum of n identical and independent exponential distributions, which leads to a gamma distribution of the infectious durations:

$$f(t) = \frac{(\gamma n)^n}{\Gamma(n)} t^{n-1} e^{-\gamma n t}$$

where $\Gamma(n)$ is the gamma function. The variance of sucha distribution is $1/(n\gamma^2)$. Notice that when $n = 1$ we find back the above-presented exponential distribution, when n gets large, the gamma distribution tends towards a normal one and when $n \to \infty$ we have the delta (fixed duration with no variance) distribution, see figure below.



The above figure shows gamma distribution for $\gamma = 0.5$ and various values of the number n of classes. When $n = 1$ we have the exponential distribution and when n increases the distribution tends towards a normal one. Ultimately, when $n \to \infty$ the gamma distribution converge toward the delta distribution with a variance equal to zero. Note that for all values of n the mean is equal to $1/\gamma = 2$.

20.5.3 Basic reproduction number and threshold effects

One of the most fundamental quantities used by epidemiologists is certainly the basic reproduction number R_0. For micro-parasites it is defined as the expected number of secondary cases following the introduction of one infectious individual into a fully susceptible population [7]. We understand from here that R_0 has a threshold value in the sense that a disease must have $R_0 > 1$ to invade a host population otherwise it disappears right after its introduction. The replacement number R is the average number of secondary infections produced by a typical infective during its entire period of infectiousness. At the introduction of one infective into a fully susceptible population $R = R_0$ and then R decreases. At endemic equilibrium we will have, by definition, $R = 1$, see sections 20.6.1 and 20.9.1.1.

20.5.4 Deterministic setup and dynamics analysis

For large populations, deterministic models with continuous variations of population sizes provide a good description of the disease behavior. Epidemic models are used to describe rapid outbreaks that occur in very short periods of time, during which the host population can be assumed to be in a constant state [17, 18]. A mathematical description of the fluxes of individuals of figure 20.1 is given by the following set of differential equations:

$$\frac{dS}{dt} = -\beta S \frac{I}{N} \qquad S(0) = S_0 \geq 0 \tag{20.1}$$

$$\frac{dI}{dt} = \beta S \frac{I}{N} - \gamma I \qquad I(0) = I_0 \geq 0 \tag{20.2}$$

$$\frac{dR}{dt} = \gamma I \qquad R(0) = R_0 \geq 0 \tag{20.3}$$

where γ is the recovery rate. Since the duration of the epidemic is short, this model has no host vital rate. In consequence, the total host population size $N = S + I + R$ is constant and only two of the above equations are necessary to totally account for the disease behavior. Dividing the first two equations 20.1 and 20.2 by the constant host population size N yields

$$\frac{ds}{dt} = -\beta i s \qquad s(0) = s_0 \geq 0 \tag{20.4}$$

$$\frac{di}{dt} = \beta i s - \gamma i \qquad i(0) = i_0 \geq 0 \tag{20.5}$$

where $s(t) = S(t)/N$ and $i(t) = I(t)/N$. The basic reproduction number then reads $R_0 = s_0\beta/\gamma$. Thus we can express the threshold on R_0 as follows. When $s_0 < \gamma/\beta$, on average each infective produces less than one infective and thus the number of infectives diminishes to reach 0 as time passes on. When $s_0 > \gamma/\beta$, the number of infectives first increases to then decrease towards 0, producing this characteristic

epidemic peak. This threshold effect is illustrated on the phase plane (see box 20.4) of figure 20.2. We can see on this figure that when $s_0 < \gamma/\beta$ the proportion of infectives decreases towards 0, and when $s_0 > \gamma/\beta$ the proportion of infectives first increases to then decrease toward zero. In any case the proportion of infectives ends at zero whereas the ultimate value s_∞ of the proportion of susceptibles depends on the initial proportions s_0 and i_0 of susceptibles and infectives respectively as expressed by the following implicit equation [18]:

$$i_0 + s_0 - s_\infty + \log(s_\infty/s_0)/\sigma = 0 \tag{20.6}$$

We can see from figure 20.2 that the higher the initial proportion of susceptibles s_0, the lower the proportion of individuals who do not get diseased during the epidemic. This is known as overshot phenomenon (Diekmann & Heesterbeek 2000).

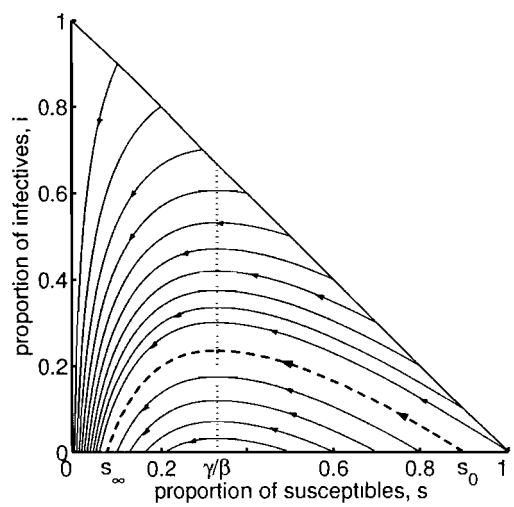


Fig. 20.2 Phase Plane of the SIR epidemic model of equations 20.4 and 20.5. The arrowed lines show the trajectories of the dynamics. The dashed arrowed line shows one particular trajectory with its initial and final proportions of suscpetible s_0 and s_∞ respectively. The vertical dotted line is the threshold γ/β on the value of the proportion of susceptibles, see main text. There is an epidemic only when s_0 is above this threshold.

20.5.5 Stochastic dynamics and probability of an epidemic in a small population

The deterministic SIR model presented above highlights a threshold value on the basic reproduction number with an epidemic when $R_0 > 1$ and no epidemic when $R_0 < 1$.

BOX 20.4

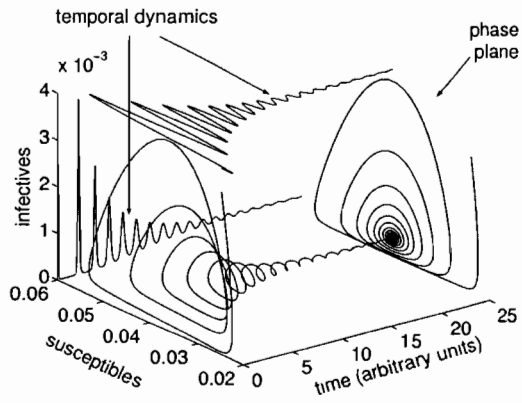
GRAPHICAL TOOLS TO STUDY DYNAMICAL SYSTEMS

In this box we present two graphical tools facilitating the study dynamical systems. The first one is the phase plane. Consider for example the endemo-epidemic SIR model of the main text

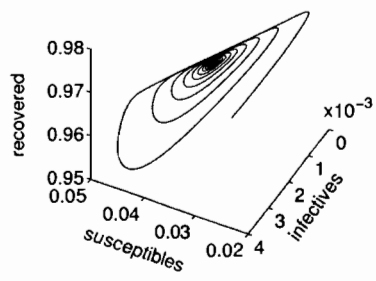
$$ds/dt = \mu - \beta is - \mu s \qquad s(0) = s_0 \geq 0$$

$$di/dt = \beta is - \gamma i - \mu i \qquad i(0) = i_0 \geq 0$$

We can solve this system and draw the temporal dynamics of each of the state variable, $s(t)$, $i(t)$, and $r(t) = 1 - s(t) - i(t)$. A phase plane plots the behavior of one state variable as a function of another state variable. Temporal dynamics and phase plane are thus two different ways of visualizing the same reality as exemplified on the figure below.



The phase plane of the above figure is the same as the one of figure 20.7. There is no time dimension on phase planes but the trajectory of the dynamics is usually indicated by arrows (see figures 20.2, 20.5 and 20.7). A phase plane can also be drawn for three state variables like on the figure below.

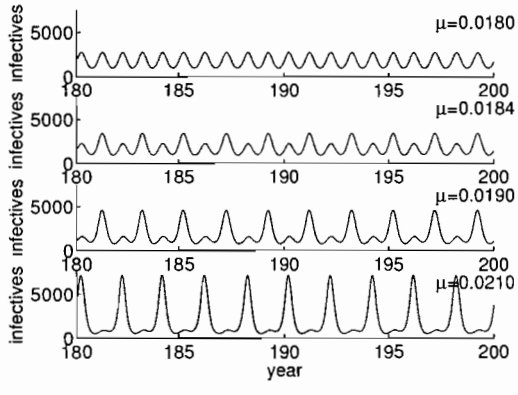


(Box 20.4 continued)

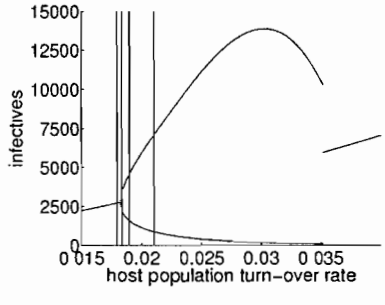
The second tool is relative to the complexity of a dynamics. Consider the same SIR model but now with a varying contact rate to sustain the oscillations (see section 20.7.1):

$$\beta(t) = \beta_0(1 + \beta_1 \cos(2\pi t)) \qquad 0 \leq \beta_1 \leq 1$$

Running this model with different values of the host population turn-over rate μ yields qualitatively different disease dynamics: on the figure below the dynamics change from annual to bi-annual when μ increases from 0.0180 to 0.0210.



These qualitative changes on the dynamics are called bifurcations and the parameter we explore the influence (here μ) is called the control parameter. Bifurcation diagrams allow to visualize the effect of a control parameter on the complexity of the dynamics. For each value of the control parameter the simulated dynamics is sampled at regular time intervals. Imagine for example that we sample the dynamics every year. Then, an annual dynamics will give one point on the bifurcation diagram (each year the dynamics recover the same value), whereas a biannual dynamics will give two points (one for the odd years and the other for the even years).



The above figure shows the bifurcation diagram of the disease dynamics with μ as the control parameter. The four vertical lines materialize the μ values corresponding to the above four time series. This diagram predicts that the disease dynamics is bi-annual for μ between 0.0183 and 0.0351 and annual for μ between 0.0150 and 0.0183 and between 0.0351 and 0.0400. At $\mu = 0.0183$ the switch from annual to biannual is progressive whereas at $\mu = 0.0351$ the switch from biannual to annual is sharp. Between 0.0183 and 0.0351 the disease oscillations reach their maximum at $\mu = 0.0303$.

However, observations on real data reveal that $R_0 > 1$ does not guarantee an epidemic in the population [9]. The cause of this discrepancy between model prediction and observed data is that the deterministic SIR model is a good approximation of the epidemic dynamics only when dealing with large populations (see above section 20.2.3), which is clearly not the case when we are interested in the initial epidemic growth following the introduction of one infective into a fully susceptible population. As the initial number of infectives during the initial epidemic growth is by definition very small, demographic stochasticity may play an important role in the start of an epidemic. The theory of branching processes is a useful framework to derive the probability that an epidemic starts [18].

Consider that the number of people infected by one infective follows a given probability distribution $\{q_k\}_{k=0}^{\infty}$. Thus, any infective infects k individuals with the probability q_k and $\sum_{k=0}^{\infty} q_k = 1$. The basic reproduction ratio R_0 can then be expressed simply as the expected number of individuals infected by one infective: $R_0 = \sum_{k=1}^{\infty} kq_k$. Then, we need to introduce the reader to one fundamental tool of branching processes: the generation function defined as

$$g(z) = \sum_{k=0}^{\infty} q_k z^k \qquad 0 \leq z \leq 1 \tag{20.7}$$

Among the interesting properties of the generating function are $g(0) = 0$, $g(1) = 1$, and $g'(1) = R_0$ [18]. Let z_n be the probability that the disease disappears from the population after n generations of transmission events. It can be shown that $z_n = g(z_n - 1)$ [30]. Since the function g is increasing, the sequence z_n is increasing and tends towards a limit z_∞. By definition z_∞ is the probability that the disease introduced by one individual into a fully susceptible population will go extinct. Thus z_∞ is the solution of the equation $z = g(z)$. It can be shown that $z_\infty = 1$ for $R_0 \leq 1$ and $0 < z_\infty < 1$ for $R_0 > 1$ [18]. Depending on the exact form of the infectious process, the solution z_∞ of the equation $z = g(z)$ can not always be expressed explicitly. For example, assuming that the number of infections during a constant time interval is according to a Poisson process, we end up with the following implicit expression of z_∞ [18]:

$$z = z\exp(R_0(z-1)) \tag{20.8}$$

which can be easily solved graphically, see figure 20.3.

20.6 A SIMPLE ENDEMIC MODEL

20.6.1 Deterministic dynamics

Epidemic models presented in the above section are used to describe rapid outbreaks that occur in very short period of time, during which the host population can be assumed to be in a constant state. Such models thus do not need to account for the

host population dynamics as governed by births and deaths. On longer period of times individuals will die and births will feed the population with new susceptibles, possibly allowing the disease to persist in the population at a low and constant prevalence. We then say that the disease is in an endemic state in the population [7]. If we are to study the endemic state of a disease, we need to construct a model that accounts for the birth and death rate of the host population. In the case of a non-fatal disease like most childhood ones in developing countries, a good approximation is to consider that the host population size $N = S + I + R$ is constant. The dynamics of the disease can then be described by the following differential equations with correspond to the flow diagram of figure 20.4:

$$\frac{ds}{dt} = \mu - \beta is - \mu s \qquad s(0) = s_0 \geq 0 \tag{20.9}$$

$$\frac{di}{dt} = \beta is - \gamma i - \mu i \qquad i(0) = i_0 \geq 0 \tag{20.10}$$

where μ is the host population turn-over rate, *i.e.* the birth rate equal to the death rate. Again, for the simplicity of the mathematical analysis, assume that this rate has a constant value. As explained above and in box 20.3, the consequence of this assumption is that the age distribution follows a negative exponential distribution. The mean of this distribution (*i.e.* the host life expectancy L) is equal to $L = 1/\mu$. This is a rather good approximation for the developing countries where the harshness of the environment imposes a similar death pressure on all the age classes [7]. However, in western countries, medical care allows most of people to reach the natural age limit, yielding this characteristics square-shape age distribution. Nevertheless, the exact form of the age pyramid does not have substantial influence on the dynamics of the disease [7].

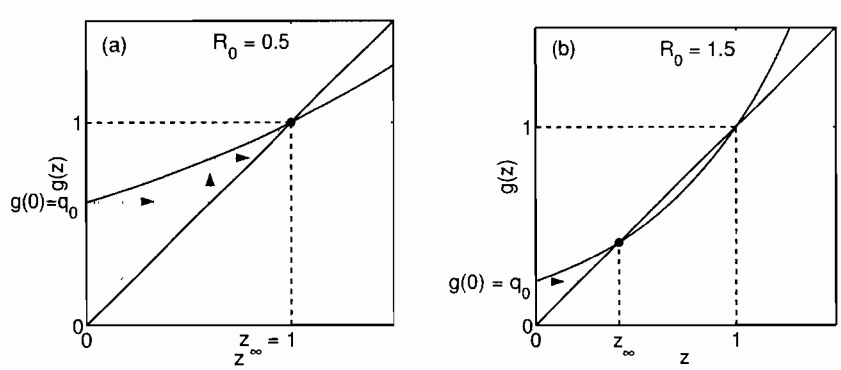


Fig. 20.3 Graphical resolution of the implicit equation 20.8. Solutions of equation 20.8 are the intersections between the first bissectrice and the curve which are respectively the l.h.s. and the r.h.s. of equation 20.8, in the domain of definition [0,1]. $z_\infty = 1$ when $R_0 < 1$ (a) and $0 < z_\infty < 1$ when $R_0 > 1$ (b), see main text.

The basic reproduction ratio now reads $R_0 = \beta/(\gamma + \mu)$. By definition, at equilibrium the system is in a constant state. Thus the differentials of equations 20.9 and 20.10 should be equated to 0: $ds/dt = di/dt = 0$ which yields the following system of equations:

$$\mu - \beta is - \mu s = 0 \tag{20.11}$$

$$\beta is - \gamma i - \mu i = 0 \tag{20.12}$$

Solving this system produces two equilibrium points: (i) the disease free scenario $(s_1^*, i_1^*, r_1^*) = (1, 0, 0)$ and (ii) the endemic case $(s_2^*, i_2^*, r_2^*) = (1/R_0, \mu(R_0 - 1)/\beta, 1 - s_2^*\text{-}i_2^*)$. The stability of these two equilibria depends solely on the value of the basic reproduction number, and not on the initial values of the proportions of susceptibles and infectives as in the above epidemic model. If R_0 is less than unity, then the disease-free equilibrium is stable [18] and the phase plane of figure 20.5 shows that the proportion of susceptibles increases towards 1 whereas the proportion of infectives decreases towards 0. When $R_0 > 1$ means that the endemic equilibrium is stable [18] and figure 20.5 shows that the proportions of susceptibles and infectives produce damped oscillations that converge towards their endemic values s^* and i^*. Linear stability analysis (see box 20.5) reveals the natural period T and the damping time D of these damped oscillations to be approximated by

$$\widehat{T} = 2\pi\sqrt{AG} \tag{20.13}$$

and

$$\widehat{D} = 2A \tag{20.14}$$

respectively where A represents the mean age at infection, $A \simeq 1/\mu(R_0 - 1)$ (see below section 20.6.2), and G gives the ecological generation length of the infection, *i.e.* the sum of the latent and infectious periods, $G = 1/(\mu + \gamma)$ [7, 50].

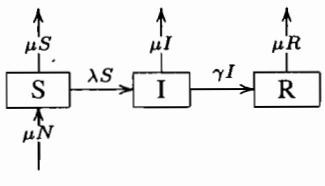


Fig. 20.4 A simple SIR endemic model. Same as figure 20.1 except that now deaths remove individual from each compartment at a constant rate μ. Also, births feed the susceptible compartment with new individuals at the same rate μ. As the birth and death rates are equal, the total size N of the whole population remains constant, see main text.

20.6.2 Statics and the average age at infection

Once the endemic equilibrium is reached we may be interested in the statics of the disease such as the mean age at infection. This is of importance as, first, numerous diseases are in endemic state in human populations and, second, the study of static properties of a disease allows the estimation of key epidemiologic parameter without requiring the long series of longitudinal notifications, often difficult to obtain in practice. The idea behind studies on the statics of diseases is that the age of the individuals reflects, in some way, time [7]. What we simply need here is horizontal data stratified by age.

Considering the age as a continuous variable, the mean age at infection is simply expressed as [7]

$$A \equiv \int_0^\infty a \frac{\lambda s(a)}{\int_0^\infty \lambda s(a) da} da \tag{20.15}$$

which is the integral sum of the age values a, weighted by the proportion of infectives of age a. Calculating this integral for a constant host population turn-over rate μ yields the intuitive relationship $A = 1/(\lambda + \mu)$. This means that the higher the force of infection (*i.e.* the probability that a susceptible gets infected), the lower the mean age at infection. Moreover, recall from above (section 20.6.1) that $i^* = \mu(R_0 - 1)/\beta$. Thus, $\lambda = \beta i^* = \mu(R_0 - 1)$. Rearranging this equation we get $\lambda + \mu = \mu R_0$. An expression of the mean age at infection then becomes $A = 1/(\mu R_0)$. This last expression allows estimating the basic reproduction number R_0 in a rather simple way as

$$R_0 = \frac{L}{A} \tag{20.16}$$

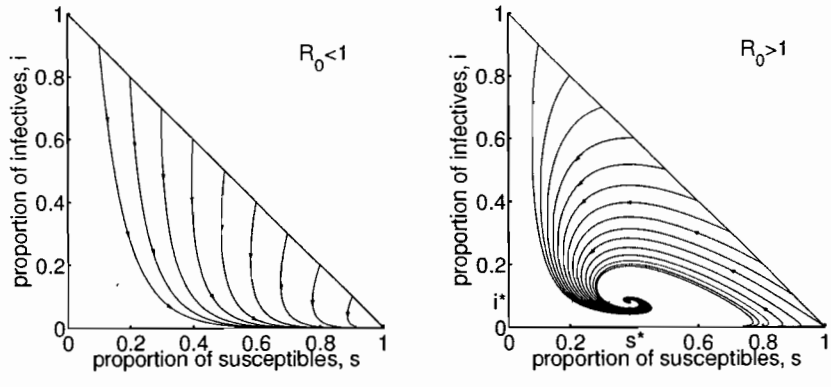


Fig. 20.5 Phase Plane of the SIR endemic model of equations 20.9 and 20.10. The arrowed lines show the trajectories of the dynamics. When the basic reproduction number $R_0 < 1$ the dynamics converges towards the stable disease-free equilibrium (left). When $R_0 > 1$ the disease dynamics converges towards endemic equilibrium (i^*, s^*) (right).

where $L = 1\mu$ is the host life expectancy (see section 20.6.1).

20.6.3 Stochastic dynamics and disease persistence

The above study of the deterministic dynamics of diseases has revealed a threshold on the value of the basic reproduction number. The disease immediately disappears after its introduction as soon as $R_0 < 1$ and persists at an endemic level in the host population when $R_0 \geq 1$. However, by inspection of real data, it appears that the condition $R_0 \geq 1$ does not guarantee the disease persistence [9]. As already mentioned about the epidemic model (see section 20.6.3), such persistence is dependent on the magnitude of the stochastic fluctuations around the endemic equilibrium.

In a meta-population context, the probability of disease extinction in one sub-population depends on both the size of the sub-population and the fluxes of infectives from neighbor sub-populations. Bartlett [9] has thus evidenced that there is a community size above which the disease can be maintained in population by itself and below which the disease cannot persist in the population without regular fluxes of infectives from neighbor populations. The determination of this critical community size is performed empirically by plotting the mean annual duration of periods with no cases against the size of the sub-population. By definition, we have a period of disease fade-out if the duration of the disease extinction is longer than the disease generation length [9]. Figure 20.6 shows an example for measles in 59 cities of England and Wales in the pre-vaccine era (1944–1966). For this disease the generation length is around 3 weeks and the critical community size is estimated here at about 115,000 individuals. The critical community size is thus a quantity very easily calculated from disease notifications and which gives a good idea of the population size required for disease persistence. Intuitively we expect that the more contagious a disease, the lower the critical community size. This explains why highly contagious diseases cannot persist in small isolated communities such as island populations or primitive Amazonian tribes.

20.7 ENDEMO-EPIDEMIC MODELS

So far we have seen simple models that allow the study of one isolated epidemic (section 20.5) or of diseases in an endemic state (section 20.6). However it appears that numerous diseases are characterized by an endemic background with regular epidemics as visible on subplots of figure 20.6 or on figure 20.10. We say that these diseases are in an endemo-epidemic state in the population. In this section we propose some complications of the basic endemic model that allow producing recurrent outbreaks as observed on many longitudinal surveys.

In section 20.6.1 we have shown that the endemic model exhibits damped oscillations which converge towards an endemic equilibrium. Linear stability analysis

BOX 20.5

LINEAR STABILITY ANALYSIS BASED ON EIGENVALUES

Linearization approximation is a standard phase-plane technique used to analyze system dynamics [39]. For an SIR system with a constant host population size we have the following system of two independent nonlinear differential equations:

$$\frac{ds}{dt} = \mu - \beta i s - \mu s \qquad s(0) = s_0 \geq 0$$

$$\frac{di}{dt} = \beta i s - \gamma i - \mu i \qquad i(0) = i_0 \geq 0$$

As found in the main text, the endemic equilibrium of this system is $(s^*, i^*) = (1/R_0, \mu(R_0 - 1)/\beta)$. Close to the endemic equilibrium, the above system can then be rewritten into the following form:

$$\begin{aligned} s(t) &= s^* + \xi(t) \\ i(t) &= i^* + \zeta(t) \end{aligned}$$

Where $\xi(t)$ and $\zeta(t)$ are the deviations from the equilibrium. In order to study the stability of the equilibrium, we then need to focus on the dynamics of the deviations $\xi(t)$ and $\zeta(t)$ [39]. Combining the above two systems, developing, and keeping only the terms which are linear in ξ and ζ, we get

$$\begin{aligned} \frac{d\xi}{dt} &= -(\beta i^* + \mu)\xi - \beta s^* \zeta + \text{NL}(\xi, \zeta) \\ \frac{d\zeta}{dt} &= \beta i^* \zeta + \text{NL}(\xi, \zeta) \end{aligned}$$

where $\text{NL}(\xi, \zeta)$ contains all the nonlinear terms in ξ and ζ. Replacing s^* and i^* by their value gives

$$\begin{aligned} \frac{d\xi}{dt} &= -[\mu(R_0 - 1) + \mu]\xi - \frac{\beta}{R_0}\zeta + \text{NL}(\xi, \zeta) \\ \frac{d\zeta}{dt} &= \mu(R_0 - 1)\xi + \text{NL}(\xi, \zeta) \end{aligned}$$

Written in matrix form, the above system becomes

$$\begin{bmatrix} d\xi/dt \\ d\zeta/dt \end{bmatrix} = \underbrace{\begin{bmatrix} -\mu(R_0 - 1) - \mu & -\beta/R_0 \\ \mu(R_0 - 1) & 0 \end{bmatrix}}_{\mathbf{J}} \cdot \begin{bmatrix} d\xi \\ d\zeta \end{bmatrix} + \text{NL}(\xi, \zeta)$$

The Jacobian matrix $\mathbf{J}$ is called the community matrix in ecology and its eigenvalues are indicative of the dynamics of the system [39]. The eigenvalues of the community matrix are solutions of the characteristic equation

$$\Lambda^2 - \text{Tr}(\mathbf{J})\Lambda + \det(\mathbf{J}) = 0$$

(Box 20.5 continued)

Where Tr and det refer to the trace and the determinant of the matrix respectively. Replacing the trace and determinant by their values gives

$$\Lambda^2 - \mu R_0 \Lambda + \mu(\mu + \gamma)(R_0 - 1) = 0$$

With the approximation pertaining to the fact that $\gamma \gg \mu$, we end with

$$\Lambda \simeq -\frac{1}{2A} \pm j\frac{1}{\sqrt{AG}}$$

where A represents the mean age at infection, $A \simeq 1/\mu R_0$, and G gives the ecological generation length of the infection, *i.e.* the sum of the latent and infectious periods, $G = 1/(\mu + \gamma) \simeq 1/\gamma$ [7, 50]. The system oscillates with a period equal to 2π time the inverse of the imaginary part of the eigenvalue, $\widehat{T} = 2\pi\sqrt{AG}$, and a damping time equal to the inverse of the real part of the eigenvalue, $\widehat{D} = 2A$ [39].

further revealed the natural period and the damping time of these oscillations to be approximated by $T = 2\pi\sqrt{AG}$ and $D = 2A$ respectively where A and G are the mean age at infection and the disease generation length respectively (see equations 20.13 and 20.14). Importantly, for most epidemiologically reasonable parameter values, the damping time is typically much longer than the natural period: $2A/T \gg 1$. This renders the endemic equilibrium weakly stable, with relatively small perturbations (intrinsic or extrinsic) exciting and sustaining the inherent oscillation behavior [28]. Alternative mechanisms for this phenomenon have been proposed in the literature and all are based on the inclusion of some heterogeneity in the endemic model. Heterogeneity can be added temporally on the coefficient of transmission, spatially in the context of meta-populations, or by cohorts for age-structured models. Lastly, heterogeneity can be added statistically for full stochastic versions of the endemic model.

20.7.1 Varying contact rate

Temporal heterogeneity on the transmission rate has been first proposed by London and Yorke [41] to explain the regular outbreaks of childhood diseases. The rationale is that for childhood diseases – for which the majority of contagions occur on the school playgrounds – the alternation of holidays and school terms is responsible for temporal variations on the coefficient of transmission. There has been a variety of forms proposed for the coefficient of transmission in the literature [19, 23, 21, 37]. Certainly the most realistic one takes the form of a plateau function, with two different values of the coefficient of transmission, one for the school terms, and one for the holidays. This necessitates the knowledge of the school holidays calendar which is not always evident, particularly for historical data. A simpler form of the coefficient

of variation would simply take the form of a sinusoidal wave:

$$\beta(t) = \beta_0(1 + \beta_1 \cos(2\pi t)) \qquad 0 \leq \beta_1 \leq 1 \tag{20.17}$$

where the strength of seasonality β_1 measures the amplitude of the oscillations around the baseline coefficient of transmission β_0. Although less realistic, this form of the coefficient of transmission produces results which are qualitatively very close to the ones obtained with a coefficient of transmission in plateau [21]. Figure 20.7 shows that even small strengths of seasonality are able to produce sustained oscillations.

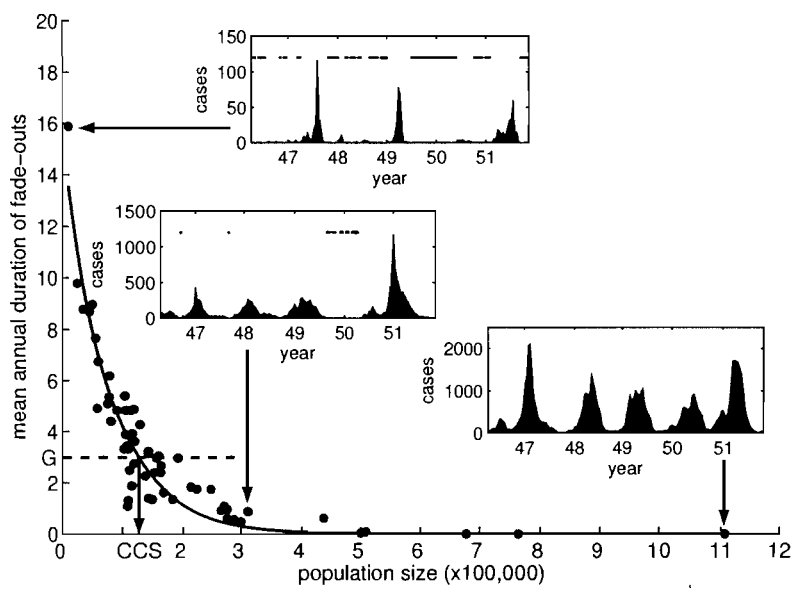


Fig. 20.6 Mean annual duration of fade-outs (*i.e.* local extinction) of measles against population size for 59 cities in England and Wales in the pre-vaccine era (1944–1966). Subplots show portions of time series illustrating the three levels of persistence identified by Bartlett [9]. Type I dynamics (bottom subplot: Birmingham, population of 1.1 millon ind.) are regular, endemic, with no fade out. Type II dynamics (middle subplot: Nottingham, population of 300,000 ind.) are regular but with some fade-outs (represented by black dots) in the troughs. Type III dynamics (top subplot: Teignmouth, population of 11,000 ind.) are irregular with long fade-out between the epidemics. The curve is the nonlinear regression ($y \simeq 16 \times \exp[-10^{-5}x]$) and its intersection with the disease generation length (G, represented by the horizontal dotted line) gives the critical community size (CCS) of the disease of around 115,000 ind. Data downloaded from `http://www.zoo.cam.ac.uk/zoostaff/grenfell/measles.htm` [54].

20.7.2 Age structured models

When the time is considered as a continuous variable, the most general form of the force of infection is actually the following [7]:

$$\lambda(a,t) = \int_0^\infty \beta(a_s, a_i, t) i(a_i, t) da_i \tag{20.18}$$

where $\beta(a_s, a_i, t)$ is the coefficient of transmission between a susceptible of age a_s and an infective of age a_i at time t. In sections 20.5 and 20.6 we averaged this relation over both time and ages. In the above section 20.7.1 we averaged over ages only and we defined the coefficient of transmission as a function of time (see equation 20.17):

$$\overline{\lambda}(t) = \overline{\beta}(t)\overline{i}(t) \tag{20.19}$$

where bars refer to age average. In the present section we average the relation 20.18 over time only. We thus have to define a coefficient of transmission as a function of age. Contrary to time, it does not really make biological sense to consider age as a continuous variable since human populations are usually aggregated by cohorts defined as the primary school children, the intermediate and high school teenagers, the college young adults, and the adults [7]. When averaging over time and considering the age variable as a discrete variable, equation 20.18 becomes:

$$\widehat{\lambda}_i = \sum_{j=1}^{n} \widehat{\beta}_{i,j} \widehat{i}_j \tag{20.20}$$

where hats refer to time average and n is the number of distinct cohorts. We thus need to define a matrix of transmission $[\widehat{\beta}_{i,j}]$.

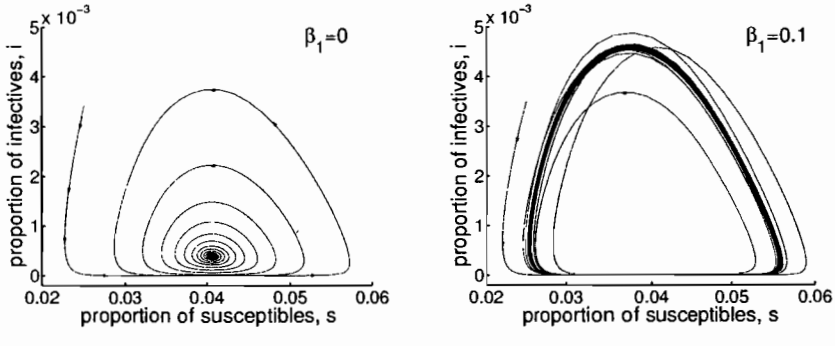


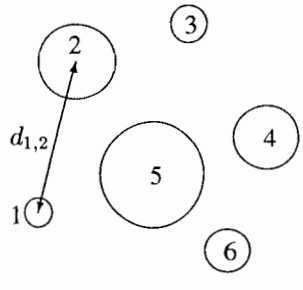
Fig. 20.7 Phase Plane of the SIR endemo-epidemic model with a constact rate varying according to equation 20.17 and $R_0 > 1$. When $\beta_1 = 0$, the dynamics converge towards an endemic equilibrium point (left). When $\beta_1 = 0.1$, the equilibrium point is destabilized and the dynamics produce sustained oscillations (right).

20.7.3 Spatially structured models

The organization of human populations in distinct cities interconnected by fluxes of individuals makes the theory of meta-populations an appropriate framework to study the spatial dynamics of infectious diseases [26, 25]. In this context the definition of a spatially structured model is pretty close to an age-structured one. We thus need to define a matrix of transmission. Two common assumptions are that this matrix is symmetric and the values $\beta_{i,j}$ are related to the geographic distance between the cities i and j (figure 20.8). However, given the speed of communication networks at a regional scale, a simple and widely used approximation of this meta-population model is the island model in which all the sub-populations are linked the ones to the others by the same coupling coefficient ε:

$$\lambda_i = \beta \times \left(i_j + \varepsilon \sum_{k \neq j} i_k \right) \tag{20.21}$$

Other spatial models are not based on the theory of meta-populations and instead consider the spatial dimension as a continuous variable. Those models are based on the reaction-diffusion equations that, for simplicity, we will not treat in the present chapter.



$$[\beta_{i,j}] \propto \begin{bmatrix} 1 & 1/d_{1,2} & \cdots & 1/d_{1,6} \\ 1/d_{1,2} & 1 & \cdots & 1/d_{2,6} \\ \vdots & \vdots & \ddots & \vdots \\ 1/d_{1,6} & 1/d_{2,6} & \cdots & 1 \end{bmatrix}$$

Fig. 20.8 In a meta-population context the matrix of contact can be modeled as inversely proportional to the distance between the communities represented represented by circles.

20.7.4 Stochastic endemic models

Sections 20.5 and 20.6 have primarily focused on deterministic models, *i.e.* models in which nothing is random. These models produce pretty good predictions as long as the population is large enough for the stochasticity to have little influence. However, in sections 20.5.5 and 20.6.3 we highlighted that, in small populations, deterministic model predictions become unreliable. To study disease dynamics in small populations, one thus often need a stochastic instead a deterministic model [6]. In this section we present an easy way to construct a stochastic version of the SIR

endemic model and we will show that stochasticity introduces enough heterogeneity in the model to produce sustained oscillations.

A stochastic version of the endemic SIR model passes through the definition of a Markov process, *i.e.* a process in which the future is independent of the past, given the present. The state space of this process is defined by the number of individuals in each of the three classes susceptibles (S), infectious (I), recovered (R). Changes in the state space are characterized by transition events which are listed in table 20.1. Each transition events occurs with a probabilistic rate derived from the rates of the deterministic model. For example, the probabilistic rate corresponding to an event of birth is defined as follows:

$$\mathrm{P}\{1 \text{ birth in } (t, t+\Delta t] | S(t) = n\} = \mu n \Delta t + \mathrm{o}(\Delta t) \tag{20.22}$$

with $\lim_{\Delta t \to 0} \frac{\mathrm{o}(\Delta t)}{t} = 0$.

Table 20.1 Transition events, and corresponding rates, for a simple stochastic SIR model.

	Type of transition event	Rate	Event
1	$S \to S+1, I \to I, R \to R$	$r_1 = \mu N$	Birth
2	$S \to S-1, I \to I, R \to R$	$r_2 = \mu S$	Death
3	$S \to S, I \to I-1, R \to R$	$r_3 = \mu I$	Death
4	$S \to S, I \to I, R \to R-1$	$r_4 = \mu R$	Death
5	$S \to S-1, I \to I+1, R \to R$	$r_5 = \beta IS/N$	Infection
6	$S \to S, I \to I-1, R \to R+1$	$r_6 = \gamma I$	Recovery
7	$S \to S, I \to I+1, R \to R$	$r_7 = \delta$	Immigration of infectives

For numerical simulation, the basic procedure consists in, first, searching the time of the next event (whatever its nature) and, second, determine the nature of this event. As all events are independent, the probabilistic rate that an event occurs, whatever its nature, is simply equal to the sum of the probabilistic rates of all the possible events $r = \sum_i r_i$. As future events are independent on past events, the time to the next event follows a negative exponential distribution of parameter r (see box 20.3). Thus, the time to the next event can simply be determined by a random realization of a negative exponential probability distribution of parameter r. Then, the nature of this event is simply determined by a random realization of a multinomial probability distribution of parameters r_1/r, r_2/r, *etc...* This process is reiterated for the duration desired. Figure 20.9 shows results of numerical simulations. Note the resemblance with real time series (compare with subplots of figure 20.6 and figure 20.10).

20.8 DATA ANALYSIS

So far we have presented a variety of disease models, each with its advantages and disadvantages. It should be clear now that there is no one model which is, in absolute, better than the others. The best model depends on the question under investigation. This section is more oriented towards the epidemiologic data. We will first see how model parameters can be estimated from the data and then focus specifically on the analysis of longitudinal data by presenting the basic tools of time series analysis.

20.8.1 Parameter estimations

We have seen that mathematical models are characterized by a certain combination of parameters, each with a biological significance such as the force of infection, the birth rate, *etc*... We have also seen that mathematical models allow the derivation of concepts which are not directly perceptible on the data, such as the basic reproduction number. In this section we are interested in trying to evaluate the numerical values of these quantities.

All the model parameters can be estimated by maximizing a model likelihood on real data. This procedure is largely used in modeling – not only in epidemiology – and its basic principles are presented in box 20.2. Parameter estimation by maximization of the likelihood takes into account an error structure and thus allows giving a

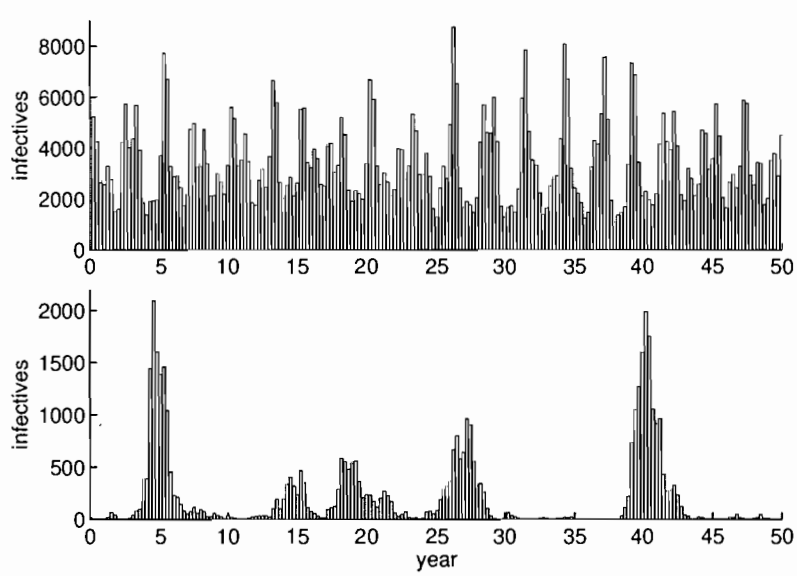


Fig. 20.9 Stochastic realizations of an SIR model in populations of 1,000,000 (top) and 100,000 (bottom) individuals. Each bar represents the incidence for one trimester.

confidence interval on the estimation. This is one major advantage of the likelihood methods.

We will not present again the likelihood method here (see box 20.2 for more details). Instead, we are interested in this section on the derivation of parameter values, almost from direct reading from the data, after playing a little bit with the model equations. Contrary to the likelihood methods, this method does not produce a confidence interval on the parameter estimation – though such an interval can be produced by Monte Carlo simulations (see section 20.8.1.5). However, this method of parameter estimation is easy, direct, and much faster to implement than the likelihood methods.

20.8.1.1 The basic reproduction ratio R_0 In the previous sections we have seen two expressions of the basic reproduction number which can all help to estimate it from the data. The first one evidenced in section 20.6.1 is relative to the endemic equilibrium value of the proportion of susceptibles in the population. Indeed, by searching the equilibrium point of the system of differential equations we ended with the fact that, at endemic equilibrium, the proportion of susceptibles in the population is equal to the inverse of the basic reproduction number. It is intuitively expected that the higher the basic reproduction number, the lower the proportion of susceptibles at endemic equilibrium in the population. A standard serological survey can easily determine the proportion s^* of susceptibles of an endemic disease. From this proportion one can thus determine the basic reproductive ratio simply as

$$R_0 = \frac{1}{s^*} \tag{20.23}$$

Such estimations of the basic reproduction number of a variety of viral and bacterial infections are listed in table 20.2. In section 20.6.2 we showed an even simpler form of the basic reproduction ratio, provided we have the age of each case notification. From this data one can easily calculate the mean age at infection A. The life expectancy L is a demographic information that is available for many human populations. Dividing it by the mean age at infection produces a good estimate of the basic reproduction ratio of a disease in a given population:

$$R_0 = \frac{L}{A} \tag{20.24}$$

Again this relation seems reasonable as it is intuitively expected that the higher the basic reproduction number R_0, the lower the mean age at infection A.

20.8.1.2 The force of infection λ In section 20.6.2 too we arrived at the intuitively plausible conclusion that the mean age at infection is the reciprocal of the force of infection. Thus, knowing the age of the disease cases, one can easily calculate the

mean age at infection A and deduce the force of infection:

$$\lambda = \frac{1}{A} \tag{20.25}$$

For the cases dealing with an age-structured model as in section 20.7.2, we need to evaluate the force of infection by age cohort. By definition, the force of infection is the probability for a susceptible to get infected. As the events of disease transmission are independent, the number of susceptible follows negative exponential distribution of parameter equal to the force of infection[2]. Said in other words the ratio $S(a+1)/S(a)$ decreases exponentially at a rate equal to the force of infection:

$$\frac{S(a+1)}{S(a)} = \exp(-\lambda(a)) \tag{20.26}$$

From equation 20.26, the force of infection by age can thus be easily calculated as long as we are in the possession of disease prevalence $I(a)$ by age cohorts. Indeed, the number of susceptible at age 0 is simply equal to the number of newborns ($S(0) = \mu N$) and the other values of $S(a)$ are then obtained recursively:

$$S(a+1) = S(a) - I(a) \tag{20.27}$$

[2]Incidentally, we find again equation 20.25. Indeed the mean of a negative exponential distribution is, by definition, the inverse of the distribution parameter (see box 20.3). Thus the mean age at infection is the reciprocal of the force of infection, as in equation 20.25

Table 20.2 Some disease parameter values taken from the literature. all parameters are estimated in Western countries unless otherwise specified.

Diseases	γ^a	A^b	R_0^c	p_c^d	T_{obs}^e	T_{calc}^f
Measles	6-7	4-6^g, 1-3^h	16-17	90-95%	1-2	1-2
Mumps	4-8	6-7	7-8^g, 11-14^h	85-90%	3, 2-4	3, 2-4
Whooping cough	7-10	4-5	16-17	90-95%	3-4	3-4
Rubella	11-12	9-10^g, 2-3^h	6-7^g, 15-16^h	82-87%	3.5	4-5
Chickenpox	10-11	6-8	7-8^g, 10-12^h	85-90%	2-4	3-4
Smallpox	—	—	—	70-80%	5	4-5
Malaria	—	—	—	99%	—	—

[a]Recovery rate. Data from [24, 16, 10].
[b]Mean age at infection. Data form [5].
[c]Basic reproduction ratio. Data from [3, 5, 46].
[d]Critical mass vaccination coverage [7].
[e]Observed inter-epidemic period. Data from [5].
[f]Model-predicted inter-epidemic period. Data from [5].
[g]Western countries.
[h]Developing countries.

20.8.1.3 The coefficient of transmission β In the case of airborne diseases, the force of infection λ is formally related to the coefficient of transmission β. The simplest of such relations is a linear one (see section 20.5.1):

$$\lambda = \beta i \tag{20.28}$$

The coefficient of transmission β can thus be estimated from the value of the force of infection λ, as estimated in the above section 20.8.1.2, and the prevalence i.

For age-structured models, things become a little bit tougher. Indeed, from equation 20.20 we have a system of n equations with n^2 unknown variables $\beta_{i,j}$:

$$\widehat{\lambda}_i = \sum_{j=1}^{n} \widehat{\beta}_{i,j}\widehat{i}_j \qquad i = 1, ..., n \tag{20.29}$$

In order to solve this system, it is necessary to formulate hypotheses allowing us to decrease the number of unknown variables down to n. The first of these hypotheses is an assumption of symmetry [7]:

$$\widehat{\beta}_{i,j} \equiv \widehat{\beta}_{j,i}, \qquad (i,j) \in \{1, ..., n\}^2 \tag{20.30}$$

However, this hypothesis has the effect of decreasing the number of unknown variable only to $n(n+1)/2$. An alternative consists in defining $\widehat{\beta}_{i,j} \equiv \widehat{\beta}_i$ [7]. In any cases these hypotheses necessitate the definition of a WAIFW (who acquire the infection from whom) matrix. Considering $n = 5$ age cohorts, the four most usually used WAIFW matrices are the followings [7]:

$$\text{WAIFW}_1 = \begin{bmatrix} \beta_1 & \beta_1 & \beta_3 & \beta_4 & \beta_5 \\ \beta_1 & \beta_2 & \beta_3 & \beta_4 & \beta_5 \\ \beta_3 & \beta_3 & \beta_3 & \beta_4 & \beta_5 \\ \beta_4 & \beta_4 & \beta_4 & \beta_4 & \beta_5 \\ \beta_5 & \beta_5 & \beta_5 & \beta_5 & \beta_5 \end{bmatrix} \qquad \text{WAIFW}_2 = \begin{bmatrix} \beta_1 & \beta_1 & \beta_1 & \beta_4 & \beta_5 \\ \beta_1 & \beta_2 & \beta_3 & \beta_4 & \beta_5 \\ \beta_1 & \beta_3 & \beta_3 & \beta_4 & \beta_5 \\ \beta_4 & \beta_4 & \beta_4 & \beta_4 & \beta_5 \\ \beta_5 & \beta_5 & \beta_5 & \beta_5 & \beta_5 \end{bmatrix}$$

$$\text{WAIFW}_3 = \begin{bmatrix} \beta_1 & \beta_1 & \beta_1 & \beta_1 & \beta_1 \\ \beta_2 & \beta_2 & \beta_2 & \beta_2 & \beta_2 \\ \beta_3 & \beta_3 & \beta_3 & \beta_3 & \beta_3 \\ \beta_4 & \beta_4 & \beta_4 & \beta_4 & \beta_4 \\ \beta_5 & \beta_5 & \beta_5 & \beta_5 & \beta_5 \end{bmatrix} \qquad \text{WAIFW}_4 = \begin{bmatrix} \beta_1 & \beta_5 & \beta_5 & \beta_5 & \beta_5 \\ \beta_5 & \beta_2 & \beta_5 & \beta_5 & \beta_5 \\ \beta_5 & \beta_5 & \beta_3 & \beta_5 & \beta_5 \\ \beta_5 & \beta_5 & \beta_5 & \beta_4 & \beta_5 \\ \beta_5 & \beta_5 & \beta_5 & \beta_5 & \beta_5 \end{bmatrix}$$

where classes 1, 2, 3, 4, and 5 usually refer to the 0-4 yr, 5-9 yr, 10-14 yr, 15-19 yr, and 20 yr and more respectively. Each of those matrices yields a system of n equations with n unknown. The $\widehat{\beta}_i$ can thus be determined from the observed mean incidences by age cohort (i) and the mean forces of infection by age cohort calculated in the above section 20.8.1.2.

For a temporally varying coefficient of transmission, there is no other means of estimation than maximum likelihood, whatever the exact form of the variation on β.

20.8.1.4 The rate of recovery γ Quite generally, under steady-state conditions, the quantity in a given compartment is equal to the product of the rate of inflow times the expected sojourn time. In the case of the infective compartment of an SIR model this remark translates into *incidence* $\times$ *expected sojourn time* = *prevalence* [18]. As mentioned in section 20.6.1, when the outflow of a compartment occurs at a constant rate, the sojourn time in the compartment follows a negative exponential distribution with parameter equal to the rate of outflow. The mean of a negative exponential distribution is equal to the reciprocal of its parameter, thus the expected sojourn time in the infective compartment is equal to the inverse of the recovery rate. In consequence, we end up with the following:

$$\gamma = \frac{incidence}{prevalence} \tag{20.31}$$

Epidemiological data generally contain either incidence or prevalence. However, one can be easily be calculated form the other as incidence is equal to the variation of prevalence.

20.8.1.5 Monte Carlo simulations We have presented here simple methods to estimate the values of both model parameters (like the rate of recovery) and emerging quantities (such as the basic reproduction number). These estimation are fast and easy to implement with most available epidemiological data. However, and contrary to likelihood methods, they do not provide any confidence interval on the estimation. One classic method to cope with this is to use Monte Carlo simulations.

The idea of Monte Carlo simulations is to generate a distribution of a parameter by resampling the data [42]. A confidence interval can then be found based on this distribution. In practice, the generation of such a distribution is done as follows. (1) an artificial data of the same length as the original data set is generated by sampling with replacement in the original data set. (2) the parameter is estimated on this artificial data set and its value kept in memory. Steps (1) and (2) are repeated a large number of time and the values of the parameters estimated on each artificial data set give a distribution of the parameter. From this distribution one can find a confidence interval. One crucial point in Monte Carlo simulations is related to the choice of the number of time steps (1) and (2) should be repeated. This number should be large enough for the generated parameter distribution to be considered in a steady state. One way to check for the convergence of the distribution towards a steady state is to follows the evolution of the value of one distribution's statistic (such as the mean) at each new repetition and stop when this statistic seems to have converged to a steady value.

20.8.2 Tools for time series analysis

Longitudinal epidemiological surveys produce time series. The object of time series analysis is to look for periodic patterns in the data. Because of the time component, data in time series are not independent. The consequence is that the classic statistical tools that assume independence of data cannot be used on time series [15]. In this section we briefly present the basic tools of time series analysis, from the simplest to the most recent and elaborated.

20.8.2.1 Stationary time series The first two methods require the time series to be in a stationary state. A time series is said to be in a stationary state if there is no systematic change in mean (no trend) and in variance [15]. This basically supposes that the signal has constant period and amplitude, which is not the case for numerous real time series. The trend can be removed by considering the residuals from a regression or a nonparametric smoothing such as a B-spline or loess regression [15]. Another mean for removing the trend consists in applying a moving average to the series [15]: each point a time t in the series is replaced by the average of the points between times $t-T/2$ and $t+T/2$, where the length T of the moving window is to be defined. When T increases the edges effects increases too and the averaged series will be reduced from T data points compared to the original series. Lastly, the trends can be removed by considering the variations in the time series [15]: each data points is replaced by the difference with the previous data point. Square transformation of the data generally has the effect of stabilizing the variance and logarithm transformation is usually used to linearized data, the three methods presented here requiring linear data sets where the effect is proportional to the cause.

20.8.2.2 Autocorrelograms This method applies to stationary time series. The idea is to calculate the correlation between a time series and a lagged copy of itself [15] (figure 20.10). The autocorrelogram plots the value of an autocorrelation coefficient against r the value of the lag. At lag = 0 the autocorrelation is by definition equal to 1. When the lag increases the value of the autocorrelation coefficient decreases, then becomes negative, and oscillates around the $r = 0$ horizontal line. The period of the oscillations of the periodogram is the same as the period of the original signal. Moreover the oscillations are damped. Indeed, because of the dependency between the points of the time series, the noise accumulate additionally, thus hiding any autocorrelation signal when the lag becomes too large. In the same way one can trace correlograms between two different series (called cross-correlograms) to get an idea of the synchronicity or phase difference between two data sets.

20.8.2.3 Fourier spectra Like the autocorrelogram this method applies to stationary time series. The Fourier theorem states that any periodic signal $s(t)$ of frequency

F_0 can be decomposed into a sum of sinusoids [15, 13] (figure 20.11):

$$s(t) = a_0 + \sum_{n=1}^{+\infty} [a_n \cos(2\pi F_0 nt) + b_n \sin(2\pi F_0 nt)] \tag{20.32}$$

Where a_0 is the mean of the signal and a_n and b_n ($n \in \mathbb{N}$) are the Fourier coefficients, basically referring to the weight that each harmonic of frequency nF_0 has in the whole signal. This can also be thought of in terms of the magnitude of the correlation between the signal $s(t)$ and the sinusoid of frequency nF_0. The use of complex

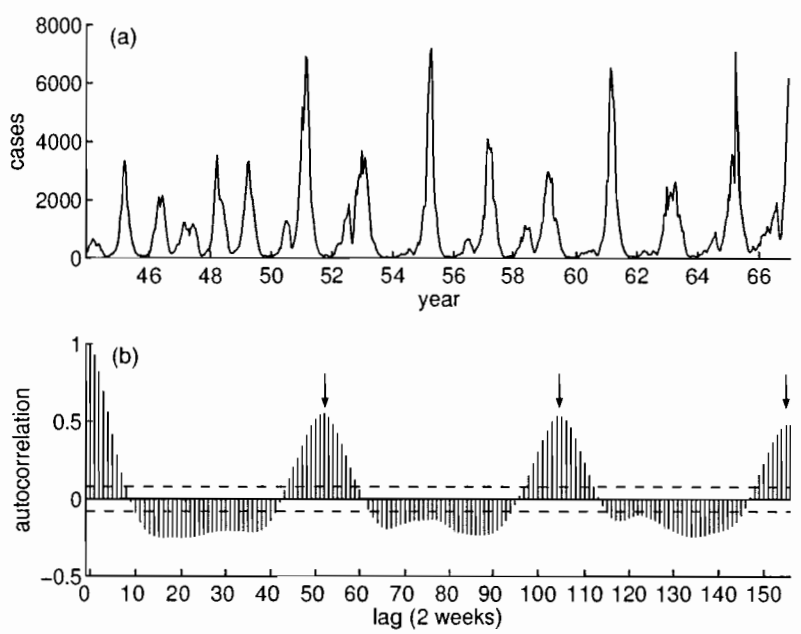


Fig. 20.10 Autocorrelation plot of the biweekly measles notification cases for London from 1944 to 1966. (a) times series of the cases. The correlation between the time series and a lagged copy of itself is calculated. As the notifications are biweekly, the lag is equal to 2 weeks. The autocorrelogram (b) shows the value of the autocorrelation against the lag. The dashed line represent the 95% confidence limits about zero. When lag=0 the series is correlated with itself and thus the autocorrelation is equal to 1. The autocorrelation coefficient then reaches maxima at lag = 53, 105, and 156 (106, 210, 312 weeks respectively, see vertical arrows), thus evidencing a period of about 2 years. Data downloaded from `http://www.zoo.cam.ac.uk/zoostaff/grenfell/measles.htm` [54].

numbers renders this formula much simpler[3]:

$$s(t) = \sum_{n=-\infty}^{+\infty} c_n e^{jn(2\pi F_0)t} \tag{20.33}$$

where j is the imaginary number and the Fourier coefficients c_n are now complex. The advantage of this form is that we have only one coefficient c_n for each frequency nF_0. Decomposing a time series into a Fourier sum consists in estimating the coefficients of the Fourier sum. For equation 20.33 the coefficients are equals to

$$c_n = \frac{1}{T_0} \int_0^{T_0} s(t) e^{-j2\pi F_0 nt} dt \tag{20.34}$$

It is important to realized that the time series $s(t)$ and the series of Fourier coefficients c_n describe exactly the same reality, the time series in the time domain and the Fourier coefficients in the frequency domain. The Fourier transform of equation 20.34 allows passing from the time to the frequency domains whereas the reverse Fourier transform of equation 20.33 does the opposite transformation. The time or the frequency domains are thus two different ways of looking at the same reality. As in time series analysis we are interested into the regular patterns of a series, it often more convenient to work into the frequency domain (at least when the series is stationary). This decomposition of a periodic signal into a sum of sinusoids can be generalized to the decomposition of an aperiodic signal where an aperiodic signal is simply a periodic signal with a period equal to ∞ [15, 13]. Equation 20.34 then takes the more general form

$$S(f) = \int_{-\infty}^{+\infty} s(t) e^{-j2\pi ft} dt \tag{20.35}$$

where the Fourier transform $S(t)$ is now a continuous function of frequency f. A Fourier spectrum plots the values of $S(f)$ against f. Inspection of such a spectrum gives a clear idea of which frequencies contribute the most to the signal.

20.8.2.4 Wavelet analysis Direct and inverse Fourier transforms force us to visualize a stationary time series either in the time or the frequency domain. Analyzing non-stationary time series one may however be interested in visualizing it in the time and frequency domain at the same time in order to able to say that the period of the signal is equal to T_1 between times t_1 and t_2, T_2 between times t_2 and t_3, and so on. One first attempt into this direction has been the use of the Fourier transform on a moving window. The major disadvantage of this *ad hoc* method is that the fixed size of the windows gives different weights to the different frequencies. This inconvenient has been coped by the invention of the wavelets.

[3]After the application of the key mathematical relationship: $e^{j\theta} = \cos\theta + j\sin\theta$.

The last decade has witnessed the emergence of an impressive number of wavelets. Certainly the most used in ecology is the Morlet one which is essentially a complex exponential with a Gaussian envelope. The key advantage of wavelets relative to sinusoids used in Fourier analysis is that they can not only be moved along the signal (as in windowed Fourier analysis) but also stretched to account equally for the different frequencies [13, 55]. A wavelet spectrum is thus a three dimensional graph which plots the correlation of the wavelet with the signal as a function of both the location of the wavelet along the signal (time domain) and the stretching of the wavelet (frequency domain). Figure 20.12 shows an example for the measles cases of London from 1944 to 1966. With such a graph one is able to say the frequency of a signal at any time.

20.9 APPLICATIONS TO VACCINATION POLICIES

After an overview of the basic epidemiological models and results as well as statistical tools for the epidemiologist, the last section is devoted to practical applications for

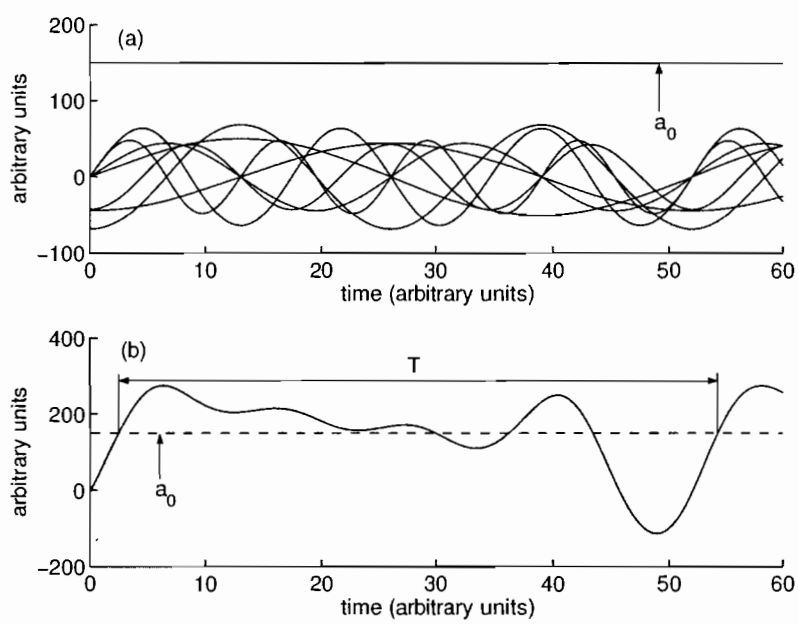


Fig. 20.11 Decomposition of a periodic signal into Fourier sum of sinusoids. a: plot of the sinusoidal component of the whole signal (b). With reference to equation 20.32, $a_0 = 150$, $a_1 = -44.16$, $b_1 = 20.22$, $a_2 = -68.66$, $b_2 = 44.22$, $a_3 = -42.34$, $b_3 = 63.82$, $a_4 = 0$, $b_4 = 47.83$, $T = 1/F_0 = 52$ arbitrary units.

the development of vaccination policies. In public health, vaccination policies are decisions made by governments and applied on large spatial and temporal scales. The ultimate aim of a vaccination policy is the eradication of a disease from a population. This goal is extremely difficult to achieve in practice and most vaccination strategies are imperfect in the sense that they only decrease (sometime dramatically) the number of cases, without however eradicating the disease [7]. In this context, vaccination can yield side-effects on the disease statics and dynamics that are important to evaluate. There currently exist two major vaccination strategies, the mass vaccination, which is the most ancient and still the most applied, and the recently developed pulse vaccination which is used in a increasing number of countries.

20.9.1 Mass vaccination strategy

Mass vaccination strategy is the most ancient and still the most widely used vaccination scheme. It consists in vaccinating a large proportion of infants before the mean age at infection [7], *e.g.* the 0-2 age cohort for the measles-mumps-rubella (MMR) vaccine in the USA. Its first applications started in the sixties against measles in the

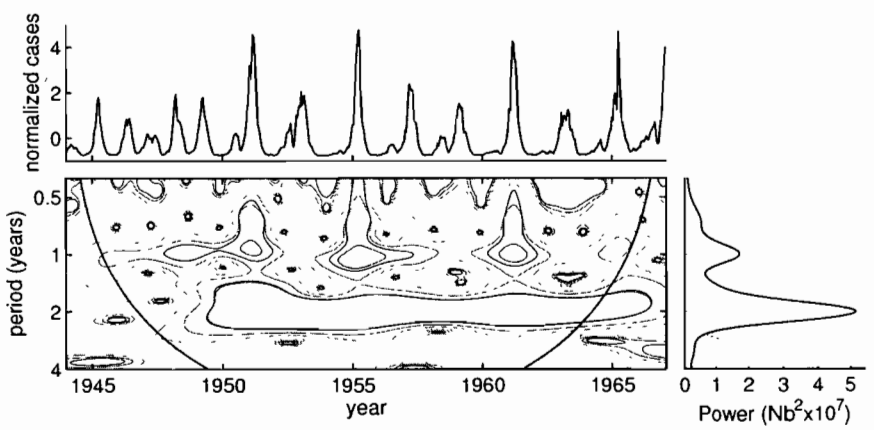


Fig. 20.12 Wavelet power spectrum of the measles notification cases for London between 1944 and 1966. The magnitude of the correlation between the series and the wavelet increases from blue to black. The parabolic curves represent the cone of influence. Because of edge effects, everything below the cone of influence cannot be interpreted. The above graph shows the time series of normalized measles cases and the right graph is the Fourier spectrum. It is clear that the wavelet power spectrum combines information of both time and frequency domains. Data downloaded from `http://www.zoo.cam.ac.uk/zoostaff/grenfell/measles.htm` [54].

North American and European countries where they have caused a dramatic decrease of the number of cases.

20.9.1.1 Calculating the vaccination coverage The derivation of the optimal vaccination coverage is based on the properties of the endemic equilibrium. At equilibrium the replacement number R (see section 20.5.3) is equal to the basic reproduction number R_0 times the proportion of susceptibles: $R = R_0 s^*$. Applying a vaccination coverage equal to p has the effect of diminishing the proportion of susceptibles by p: $s^* = 1 - p$. A condition for disease eradication is that the reproduction number be less than 1: $R = R_0 s^* = R_0(1-p) < 1$, or $p > 1 - 1/R_0$. Thus the critical vaccination coverage p_c for disease eradication is $p_c = 1 - 1/R_0$ [7]. Note that this result shows that we do not need to vaccinate each individual to protect the whole population. Note too that this property known as herd immunity is not evident from the data and emerges only from the model. The higher the basic reproduction number, the higher the vaccination coverage should be. Vaccination coverages of major human infectious diseases are given in table 20.2. We can see that many infectious diseases require vaccination coverage which are far too higher to be achieved in practice. This is further complicated by others mechanisms such as the vaccine efficacy. Consider for example measles and rubella for which estimates of the critical vaccination coverage based on R_0 are 0.94 and 0.86 respectively. A vaccine efficacy of 0.95 means that 5% of those vaccinated do not become immune. In consequence, tacking into account vaccine efficacy necessitates coverages of 0.99 and 0.91 for measles and rubella respectively [31]. This explains why the only human infectious disease which has been successfully worldwide eradicated so far is smallpox which has the lowest critical vaccination coverage.

20.9.1.2 Consequences on the statics A first and expected effect of vaccination is that fewer people will experience infection. But the decrease of the force of infection due to vaccination means that the mean age at infection of the smaller number of people who do acquire infection increases [7] (recall equation 20.25). If the probability of disease complications increases with age it is thus possible that some vaccination programmes could actually increase the absolute number of serious cases. The likelihood of such a perverse outcome again can only be evaluated thanks to disease models. A classical example is the one of the congenital rubella syndrome (CRS) treated in detailed by Anderson and May [4]). They have evidenced that the absolute number of CRS can actually increases with the vaccination coverage when the vaccination coverage is low (figure 20.13).

20.9.1.3 Consequences on the spatial and/or temporal dynamics The spatio-temporal dynamics of a disease is of primary interest if we are to design efficient country-wide vaccination policies [25]. For example, the global eradication of a disease would be easier if the local dynamics are synchronous. In the case where local dynamics are completely asynchronous, local extinctions would be quickly followed by migration of infectious individuals from neighbor communities experiencing an

epidemic outbreak. With a very simple endemo-epidemic model Earn *et al.* [21] have shown that the disease dynamics complexity – as given by the length of the period and the number of different attractors, see box 20.4 – increases with vaccination coverage. Furthermore, a simple island model such as the one presented in section 20.7.3 evidences that an increase in the vaccination coverage results in a decrease of the spatial synchrony of disease dynamics. This is intuitive since the coupling between the different sub-populations is assured by the migration of infective (see equation 20.21). Vaccination has the effect of decreasing the number of infectives and thus the synchrony between the different sub-populations. In conclusion, mass vaccination has the effect of (1) increasing disease dynamics complexity and (2) desynchronizing local dynamics. The first consequence accentuates the second as the probability of dynamics synchronicity naturally decreases with the level of complexity. These model predictions have been successfully confirmed on real data analysis [49]. As the mass vaccination tends to desynchronize the spatial dynamics of the disease, it renders global disease eradication even more difficult to achieve in practice than expected from the above theoretical predictions (section 20.9.1.1).

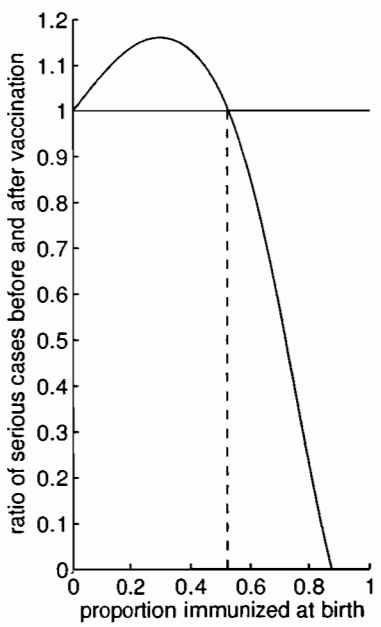


Fig. 20.13 Model predicted effect of a mass vaccination policy at birth on the number of congenital rubella syndromes (CRS). The graph plots the ratio of CRS before and after the start of the mass vaccination policy against the vaccination coverage. It shows that low vaccination coverages (less than 50% here) should be avoided as they can increase the absolute number of serious cases. Model from [4].

20.9.2 Pulse vaccination strategy

As highlighted in the previous section 20.9.1, mass vaccination strategy requires a too high systematic vaccination coverage to be achieved in practice. A recently proposed and potentially less expensive strategy is vaccination in pulses [1, 47]. This approach consists in vaccinating a certain proportion of the population at regular intervals of time. The rationale behind this is to vaccinate sufficiently and frequently enough to maintain the percentage of susceptibles below the threshold necessary for an epidemic to start (figure 20.14). What makes this policy less expensive than the mass vaccination strategy is that it explicitly accounts for the dynamics of the host population through the birth rate. Several theoretical works have been carried out to express the optimal vaccination coverage and frequency as a function of the host demographic characteristics [52, 53, 20]. The simplest one is derived from the Pythagore theorem, see figure 20.14. This theory has been successfully applied in campaigns against poliomyelitis and measles in Central and South America, and measles in the UK in 1994.

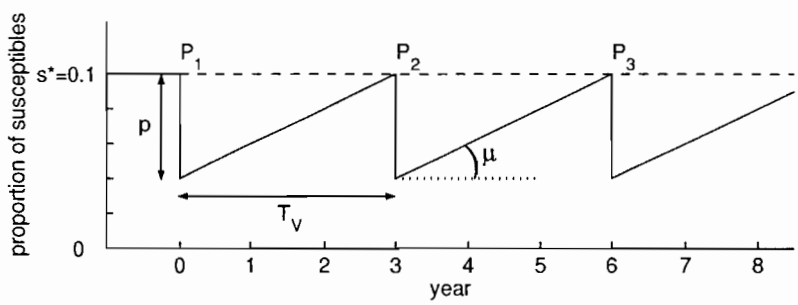


Fig. 20.14 Pulse vaccination scheme. The graph shows the proportion of susceptibles as a function of time. The aim is to find a vaccination coverage p and a frequency $1/T_V$ of the vaccination pulses P_1, P_2, P_3, *etc...* such as the proportion of susceptibles stays below the critical value s^* necessary for an epidemic to start (represented by the dashed line). Births fill the stock of susceptibles at a constant rate μ.

20.9.2.1 Spatial dynamics Using a simple endemo-epidemic model Earn *et al.* [22] have shown that a same pulse vaccination strategy tends to synchronize disease dynamics in independent localities (see figure 20.15). This phenomenon is a case of Moran effect where the same causes produce the same effects [44, 27, 12]. Indeed, by its periodical nature, an imperfect pulse vaccination strategy acts as a forcing driving the disease dynamics. Two independent populations submitted to the same pulse vaccination scheme will thus exhibit similar and synchronous disease dynamics. The effect of pulse vaccination on the spatial dynamics of a disease is thus opposite to the one of mass vaccination and this facilitates the achievement of a global disease eradication [35].

20.9.2.2 ***Resonance*** A second side-effect that can be associated to the periodic nature of pulse vaccination is the phenomenon of resonance. Theory of oscillator dynamics predicts that an oscillating dynamical system (such as an epidemiological one) submitted to a periodic forcing (such as an imperfect pulse vaccination) can produce phenomena of resonance [33]. Resonance is a generic term indicating that the amplitude of observed oscillations depends on the period of the forcing and has a maximum, called peak of resonance. These theoretical predictions and their epidemiological consequences have been investigated on a disease system by numerical simulations and data analyzes (Choisy *et al.*, submitted). Figure 20.16 shows that the mean annual number of infectives globally decreases as the frequency of vaccination pulses increases. However, resonance is responsible for the peaks observed on this general trend. The major one occurs at a vaccination frequency close to 2 years, the others simply being harmonics of it. The practical consequence of these peaks is that, locally on the vaccination frequency dimension, the mean annual number of infectives counter-intuitively increases with the frequency of vaccination.

20.10 CONCLUSION

20.10.1 What we have seen

Statistical analyzes of epidemiological data help to characterize, quantify, and summarize the way diseases spread in host populations (section 20.8). The aim of epidemiological modeling is to understand the behavior of diseases in nature (section 20.2). Because of ethical and practical impossibility to perform experiments in public

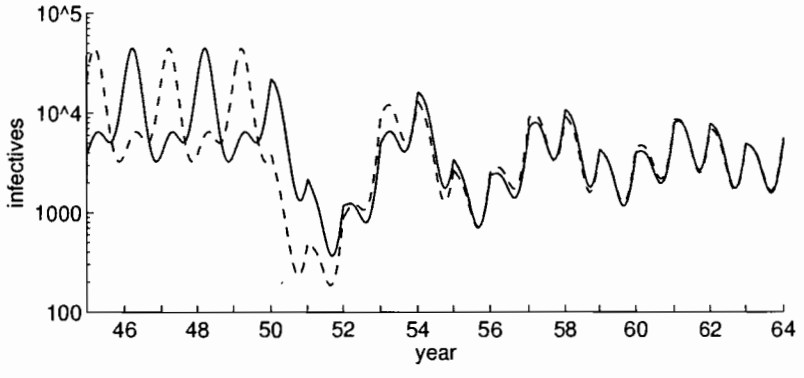


Fig. 20.15 Effect of a same pulse vaccination strategy on independent disease dynamics. The simulations start so that the disease dynamics are in opposition of phase. At year 50 a pulse vaccination strategy is started with p=20% and T_V=1 year, progressively synchronizing the two dynamics.

health, mathematical models appear as a cheap and efficient way to explore and test hypotheses (section 20.2.2). In addition to force the investigator to think rigourously (sections 20.5.1 and 20.5.2, box 20.3), models provide powerful conceptual results such as the basic reproduction number and threshold effects (20.5.3), or the herd immunity (20.9.1.1). Even if very interesting pure theoretical works have been realized, the key element of epidemiological modeling is to link model with data. Likelihood methods are modern and efficient ways to do so (box 20.2). Models thus allow to estimate epidemiological parameters and also to identify crucial data that needs to be collected.

This chapter was centered on the SIR model. Although one of the simplest epidemiological models, it is still one of the most used, particularly to study childhood viral and bacterial infections. There is a multitude of ways to complexify this simple model in order to account for more and more phenomena. However, the more complex is not necessary the best and it is the purpose of a model that dictates its degree of complexity (section 20.2, box 20.1). We have explored models to study single epidemics (section 20.5), endemic diseases (section 20.6), spatial disease dynamics (section 20.7.3) and have illustrated how these mathematical tools can be used for the development of public health policies in helping defining optimal vaccination strategies (section 20.9).

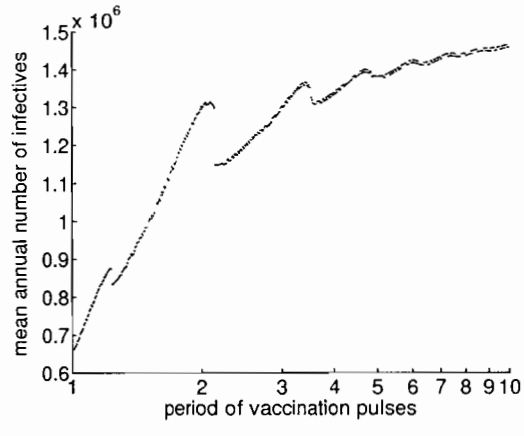


Fig. 20.16 Mean annual number of infectives as a function of the period T_V of vaccination pulses. The general trend is a decrease of the mean annual number of infectives as the frequency of vaccination ($1/T_V$) increases. However, resonance is responsible for these peaks on this general trend. The consequence of such peaks is that, locally on the vaccination frequency dimension, the mean annual number of infectives counter-intuitively increases with the frequency of vaccination.

20.10.2 What we have not seen

Of course, the list of what we have seen about epidemiological models is much longer than the list of what we have glanced at.

Some classes can be added to the simple SIR model. For example a commonly used model for childhood diseases is the SEIR one which adds a class of exposed (E) individuals, *i.e.* individuals which are infected but not infectious yet. For most childhood infectious diseases the latence phase is often as long as the infectious and should be accounted for as it can substantially change the epidemiological conclusions of the models. MSEIR models further add a class accounting for the post-birth period during which new-born are protected from infections by maternal antibodies.

All the models covered in this chapter assume that the host population is of constant size. This thus excludes both diseases in exponentially growing populations like in most developing countries and disease-induced mortality as the case for many infections including childhood diseases in developing countries, malaria, *etc* ... Accounting for a non-constant host population size requires the explicit modeling of the host population dynamics, in addition to the disease dynamics.

For sexually transmitted diseases (STD), contagious contacts are not established randomly as usually assumed for airborne infections. Besides a strong heterogeneity in the sexual activity, sexual contact occurs preferentially between people of the same sexual activity, creating this core effect in the epidemiology. Models for STD should thus account for all these degrees of heterogeneity in the contacts. In addition, many STD result in little or no acquired immunity following recovery and SIRS models would be more appropriate.

Other particular epidemiological systems requiring adapted models include, among others, mother-to-child diseases for which not all children are born into the susceptible compartment and diseases propagated by syringe-sharing intravenous drug users such as HIV/AIDS.

The epidemiology of multiple-host diseases is far more complicated than the one of directly transmitted infections. Compartmental models of such diseases need to account for the dynamics of the disease in the different hosts or reservoirs, and possibly also for the dynamics of these different hosts or reservoirs. For these diseases, the modeling of passage from one host to the other is not always an easy task.

Lastly, so far we have dealt primarily we micro-parasitic infections. Contrary to micro-parasites (section 20.4), macro-parasites refer to large-size parasites (helminths, arthropods) with direct reproduction in the definitive host. Macro-parasites generally have longer generation time (often an appreciable proportion of the host life span) than micro-parasites. The immunity following the recovery from a macro-parasitic infection is generally of short duration and the number of parasites per host is a strong determinant of the epidemiology. From a modeling point of view, this means that the simple compartmental models presented in this chapter for micro-parasites should be replaced by more complicated models accounting for

the distribution of parasites among the hosts. This is totally an other subject and we refer the reader to the classical book by Anderson and May [7] for more details.

20.11 SUMMARY

By clarifying rigorously the assumptions, the variables and the parameters, mathematical modeling allows understanding the observed spread of diseases in space and time. Epidemiological model further provide important conceptual results including the basic reproduction number, the threshold effects and the herd immunity. For evident ethical and practical reasons, experiments in public health are often impossible to perform and mathematical models thus appear as a cheap and efficient way to explore and test hypotheses. This is for example of particular practical utility in the design of vaccination policies. One key aspect of epidemiological models is their link to real data. Such data often stand under the form of time series which necessitate specific statistical tools for their analysis. Models can always be complicated to improve their fit to real data. However, more complex models are not always the best and it is the question under investigation that should dictate the optimal level of complexity.

Acknowledgments

Benjamin Roche drawn figure 20.9. MC is funded by a BDI CNRS/Région Languedoc-Roussillon, and JFG is sponsored by IRD and CNRS.

References

1. Agur, Z., Cojocaru, L., Anderson, R. M., and Danon, Y. L. (1993). Pulse mass measles vaccination across age cohorts. *Proceedings of the National Academy of Sciences of the USA*, 90:11698–11702.

2. Akaike, H. (1973). *Information theory as an extension of the maximum likelihood principle*, pages 267–281. Akademiai Kado.

3. Anderson, R. M. (1982). *Transmission dynamics and control of infectious disease agents*, pages 149–176. Springer-Verlag, Berlin.

4. Anderson, R. M. and May, R. M. (1983). Vaccination against rubella and measles: quantitative investigations of different policies. *Journal of Hygiene, Cambridge*, 90:259–325.

5. Anderson, R. M. and May, R. M. (1985). Vaccination and herd immunity to infectious diseases. *Nature*, 318:323–329.

6. Anderson, R. M. and May, R. M. (1986). The invasion, persistence and spread of infectious diseases within animal and plant communitites. *Philosophical Transactions of the Royal Society of London*, B314:533–570.

7. Anderson, R. M. and May, R. M. (1991). *Infectious Diseases of Humans. Dynamics and Control.* Oxford University Press, Oxford.

8. Bailey, N. J. T. (1957). *The Mathematical Theory of Infectious Diseases and its Application.* Griffin, London.

9. Bartlett, M. S. (1960). The critical community size for measles in the United States. *Journal of the Royal Statistical Society*, 123:37–44.

10. Benenson, A. S. (1975). *Control of Communicable Diseases in Man.* American Public Health Association, Washington D.C.

11. Bernoulli, D. (1760). Essai d'une nouvelle analyse de la mortalité causée par la petite vérole et des avantages de l'inoculum pour la prévenir. *Mém. Math. Phys. Acad. Roy. Sci., Paris*, pages 1–45.

12. Blasius, B., Huppert, A., and Stone, L. (1999). Complex dynamics and phase synchronization in spatially extended ecological systems. *Nature*, 399:354–359.

13. Burke Hubbard, B. (1998). *The World According to Wavelets: The Story of a Mathematical Technique in the Making*. AK Peters.

14. Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multi-model Inference: a Practical Information-theoretic Approach*. Springer-Verlag, Berlin.

15. Chatfield, C. (1996). *The Analysis of Time Series: An Introduction*. Chapman & Hall, London.

16. Christie, A. B. (1974). *Infectious Diseases: Epidemiology and Practice*. Churchill Livingstone, London.

17. Daley, D. J. and Gani, J. (1999). *Epidemic Modelling: An Introduction*. Cambridge University Press, Cambridge.

18. Diekmann, O. and Heesterbeek, J. A. P. (2000). *Mathematical Epidemiology of Infectious Diseases. Model Building, Analysis and Interpretation*. Wiley and Sons, Chichester.

19. Dietz, K. (1976). The incidence of infectious diseases under the influence of seasonal fluctuations. *Lecture Notes in Biomathematics*, 11:1–5.

20. d'Onofrio, A. (2002). Stability properties of pulses vaccination strategy in SEIR epidemic model. *Mathematical Biosciences*, 179:57–72.

21. Earn, D. J. D., Rohani, P., Bolker, B. M., and Grenfell, B. T. (2000). A simple model for complex dynamical transitions in epidemics. *Science*, 287:667–670.

22. Earn, D. J. D., Rohani, P., and Grenfell, B. T. (1998). Persistence, chaos and synchrony in ecology and epidemiology. *Proceedings of the Royal Society of London*, B265:7–10.

23. Ellner, S., Bailey, B. A., Bobashev, G. V., Gallant, A. R., Grenfell, B. T., and Nychka, D. W. (1998). Noise and nonlinearity in measles epidemics: combining mechanistic and statistical approaches to population modeling. *American Naturalist*, 151:425–440.

24. Fenner, F. and White, D. O. (1970). *Medical Virology*. Academic Press, New-York.

25. Grenfell, B. T., Bjørnstad, O. N., and Kappey, J. (2001). Travelling waves and spatial hierarchies in measles epidemics. *Nature*, 414:716–723.

26. Grenfell, B. T. and Harwood, J. (1997). (Meta)population dynamics of infectious diseases. *Trends in Ecology and Evolution*, 148:317–335.

27. Grenfell, B. T., Wilson, K., Finkenstädt, B. F., Coulson, T. N., Murray, S., Albon, S. D., Pemberton, J. M., Clutton-Broack, T. M., and Crawley, M. J. (1998). Noise and determinism in synchronized sheep dynamics. *Nature*, 394:674–677.

28. Grossman, Z. (1980). Oscillatory phenomena in a model of infectious diseases. *Theoretical Population Biology*, 18:204–243.

29. Hamer, W. H. (1906). Epidemic disease in england. *The Lancet*, i:733–739.

30. Harris, T. E. (1963). *The Theory of Branching Processes*. Springer-Verlag, Berlin.

31. Hethcote, H. W. (2000). The mathematics of infectious diseases. *SIAM Review*, 42:599–653.

32. Hilborn, R. and Mangel, M. (1997). *The Ecological Detective. Confronting Models with Data*. Princeton University Press, Princeton.

33. Jackson, E. A. (1992). *Perspectives of Nonlinear Dynamics: Volume 1*. Cambridge University Press, Cambridge.

34. Johnson, J. B. and Omland, K. S. (2004). Model selection in ecology and evolution. *Trends in Ecology and Evolution*, 19:101–108.

35. Keeling, M. and XXXXXX (2004). *Metapopulation Dynamics of Infectious Diseases*, pages XXX–XXX. Elsevier, XXXXXX.

36. Keeling, M. J. and Grenfell, B. T. (2002). Understanding the persistence of measles: reconciling theory, simulation and observation. *Proceedings of the Royal Society of London*, B269:335–343.

37. Keeling, M. J., Rohani, P., and Grenfell, B. T. (2001). Seasonally forced diseases dynamics explored as switching between attractors. *Physica D*, 148:317–335.

38. Kermack, W. O. and McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London*, A115:700–721.

39. Kot, M. (2001). *Elements of Mathematical Ecology*. Cambridge University Press, Cambridge.

40. Lloyd, A. L. (2001). Destabilization of epidemic models with the inclusion of realistic distributions of infectious periods. *Proceedings of the Royal Society of London*, B268:985–993.

41. London, W. P. and Yorke, J. A. (1973). Recurrent outbreaks of measles, chickenpox and mumps. I seasonal variation in contact rates. *American Journal of Epidemiology*, 98:453–468.

42. Manly, B. (1997). *Randomization, Bootstrap and Monte Carlo Methods in Biology*. Chapman & Hall, London.

43. McCallum, H., Barlow, N., and Hone, J. (2001). How should pathogen transmission be modelled? *Trends in Ecology and Evolution*, 16:295–300.

44. Moran, P. A. P. (1953). The statistical analysis of the canadian lynx cycle. II. synchronization and meteorology. *Australian Journal of Zoology*, 1:291–298.

45. Morens, D. M., Folkers, G. K., and Fauci, A. S. (2004). The challenge of emerging and re-emerging infectious diseases. *Nature*, 430:242–249.

46. Nokes, D. J. and Anderson, R. M. (1988). The use of mathematical models in the epidemiology study of infectious diseases and in the design of mass vaccination programmes. *Epidemiology and Infection*, 101:1–20.

47. Nokes, D. J. and Swinton, J. (1997). Vaccination in pulses: a strategy for global eradication of measles and polio? *Trends in Microbiology*, 5(1):14–19.

48. Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1997). *Numerical Recipes in C. The Art of Scientific Computing*. Cambridge University Press, Cambridge.

49. Rohani, P., Earn, D. J. D., and Grenfell, B. T. (1999). Opposite patterns of synchrony in sympatric diseases metapopulations. *Science*, 286:968–971.

50. Rohani, P., Keeling, M. J., and T., G. B. (2002). The interplay between determinism and stochasticity in childhood diseases. *American Naturalist*, 159:569–481.

51. Ross, R. (1911). *The Prevention of Malaria*. Murray, London.

52. Shulgin, B., Stone, L., and Agur, Z. (1998). Pulse vaccination strategy in the sir epidemic model. *Bulletin of Mathematical Biology*, 60:1123–1148.

53. Stone, L., Shulgin, B., and Agur, Z. (2000). Theoretical examination of the pulse vaccination policy in the SIR epidemic model. *Mathematical and Computer Modelling*, 31:207–215.

54. Grenfell, T. B., Bjørnstad, O. N., and Finkenstädt, B. (2002). Dynamics of measles epidemics: scaling noise, determinism and predictability with the TSIR model. *Ecological Monographs*, 72:185–202.

55. Torrence, C. and Compo, G. P. (1988). A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society*, 79:61–78.

56. Turchin, P. and Hanski, I. (2001). Contrasting alternative hypotheses about rodent cycles by translating them into parametrized models. *Ecology Letters*, 4:267–276.

Annexe E

CHOISY M., ROHANI, P. & GUÉGAN J.-F. Dynamics of infectious diseases and pulse vaccination : teasing apart the embedded resonance effects. Soumis à *Theoretical Population Biology*

Dynamics of infectious diseases and pulse vaccination: teasing apart the embedded resonance effects

Marc Choisy

GEMI, UMR CNRS-IRD 2724, Centre IRD, 911 avenue Agropolis BP 64501
34394 Montpellier Cedex 5, France
choisy@mpl.ird.fr

Pejman Rohani

Institute of Ecology, University of Georgia, Athens, Georgia 30602-2202, USA
rohani@arches.uga.edu

Jean-François Guégan

GEMI, UMR CNRS-IRD 2724, Centre IRD, 911 avenue Agropolis BP 64501
34394 Montpellier Cedex 5, France
guegan@mpl.ird.fr

Key words: infectious diseases, population dynamics, parametric resonance, environmental forcing, pulse vaccination.

Running head: Resonance and disease dynamics

Type of manuscript: Article.

Abstract

Dynamical systems theory predicts that inherently oscillatory systems undergoing periodic forcings will exhibit resonance phenomena, which are characterized by qualitative dynamical consequences resulting from the amplification of small external perturbations. In this paper we use extensive numerical simulations to demonstrate that the periodic nature of pulse vaccination strategies can make disease dynamics resonate. We proceed step by step in order to tease apart the dynamical consequences of (i) the intrinsic nonlinearity of the host-pathogen system, (ii) the seasonal variation in transmission and (iii) the additional forcing caused by vaccinating in pulses. We document that the resonance phenomenon associated with pulse vaccination can have quantitative epidemiological implications and produce perverse effects such as an unexpected increase in the number of infectives as the vaccination frequency increases. Our findings emphasize the importance of carefully taking into account the dynamical properties of the disease when designing a pulse vaccination strategy.

Introduction

The dynamics of most directly transmitted childhood diseases are characterized by pronounced oscillations, with alternating "boom" (epidemic, outbreak) and "bust" (inter-epidemic) periods (Anderson et al. 1984; Hethcote 1998; Rohani et al. 1999; Earn et al. 2000; Grenfell et al. 2002). This observation has motivated an impressive number of studies, ranging from the theoretical to the applied perspective. Ecologically, the most commonly addressed questions concern infection persistence (Bartlett 1957; Bolker and Grenfell 1995; Keeling and Grenfell 1997; Rohani et al. 2000), spatial synchrony of epidemics (Bailey 1975; Bolker and Grenfell 1996; Lloyd and May 1996; Earn et al. 1998; Rohani et al. 1999; Grenfell et al. 2001), and the impact of vaccination (Anderson et al. 1984; Tidd et al. 1993; Hethcote 1998; Deguen and Flahault 2000; Earn et al. 2000; Grenfell et al. 2001; Rohani et al. 2002). From a more theoretical point of view, there have been a number of investigations into the complex and nonlinear nature of epidemiological systems (Schaffer and Kot 1985; Bolker and Grenfell 1993; Rhodes et al. 1997; Ellner et al. 1998; Tuckwell et al. 2000; Keeling et al. 2001; Billings and Schwartz 2002). A particularly interesting area has been to develop an understanding of the interaction between the weakly damped oscillations of epidemic systems and external forcing (e.g., the school year cycle), which can give rise to a plethora of complex patterns (Dietz 1976; Grossman 1980; Schwartz 1985; Glendinning and Perry 1997), such as a cascade of period-doubling bifurcations and chaos. Surprisingly, the phenomenon of resonance – i.e. the excitation of oscillations by external forcing – has received little attention by epidemiologists and ecologists. This is all the more surprising given the numerous potential sources of resonance in such systems and their possible dynamical consequences both in qualitative and quantitative terms (see for example Tuckwell et al. 2000). A recent exception is the work of Greenman et al. (2004), who showed that resonance has great potential for shedding light on the dynamics of ecological and epidemiological systems. In this paper, we explore resonance phenomena in models of diseases transmission. We begin by studying the resonance associated with seasonality in transmission (as due to the alternation of holidays and school terms). In a second stage we focus on the resonance related to the periodic nature of pulse vaccination, first without and then together with seasonality in transmission.

The rationale underpinning classical vaccination policy is to ensure the proportion of susceptible individuals in the population remains below the threshold necessary for an epidemic (Anderson and May 1991). The most commonly used scheme for the control of childhood microparasitic infections is called paediatric mass vaccination. It is based on the static properties of the host-pathogen system and involves vaccinating a critical fraction of infants before they reach a specific age cohort, usually 0-2 years (Anderson and May 1991). An alternative and potentially less expensive strategy, called pulse vaccination, has been recently proposed (Agur et al. 1993; Nokes and Swinton 1997). This scheme explicitly accounts for the host population dynamics and involves the periodic immunization of a specified proportion of the susceptible population to prevent invasion of the infection. A number of elegant studies have determined the optimal vaccination coverage and frequency to eradicate common infections, such as measles (Agur et al. 1993; Shulgin et al. 1998; d'Onofrio 2002). Interestingly, d'Onofrio (2002) briefly mentioned the potential for "parametric resonance" (de-

fined below) resulting from the periodic nature of vaccination pulses. However the dynamical consequences of such a resonance phenomenon have never been studied in detail. Moreover, the models exploring periodic vaccination have ignored so far the well-documented seasonality in disease transmission (primarily to facilitate analytical tractability).

The present work addresses this problem through extensive numerical simulations to study in detail the resonance-induced quantitative consequences of pulse vaccination in a context of seasonally varying disease transmission. Given the potential for complex dynamics, we proceed step by step in order to tease apart the consequences of interactions between the inherent nonlinearity of these systems and different external forcings (seasonality and vaccination pulses). We first review the classical SEIR model, highlighting the source of nonlinearity. We then introduce sinusoidal variation in disease transmission (to mimic the alternation of school terms and holidays). Within this simple set-up, we provide a description of the different characteristics of resonance, both linear and nonlinear, and contrast the quantitative predictions of resonance due to seasonality in disease transmission with the patterns observed in epidemiological data. We finally incorporate pulse vaccination into the model. Despite the increased degree of complexity in the dynamics, parametric resonance associated with the periodic vaccination pulses is clearly identified. We focus on its quantitative epidemiological consequences, in terms of incidence, and reveal potential counter-intuitive effects such as an increased number of infectives as the frequency of vaccination rises. The results of this study have strong implications for the design of pulse vaccination schemes and these are discussed at the end of the paper.

The Model

Consider an immunizing, non-fatal, acute disease in a constant population of N individuals. In the classical SEIR framework (Anderson and May 1991), the dynamics of the susceptible (S), infected but not infectious (E), infectious (I), and recovered (R) individuals are described by the following differential equations:

$$\frac{dS}{dt} = \mu N - (\lambda + \mu) S \tag{1}$$

$$\frac{dE}{dt} = \lambda S - (\sigma + \mu) E \tag{2}$$

$$\frac{dI}{dt} = \sigma E - (\gamma + \mu) I \tag{3}$$

$$\frac{dR}{dt} = \gamma I - \mu R \tag{4}$$

where μ is the population turn-over rate, λ is the force of infection (the per capita rate of acquisition of infection) and $1/\sigma$ and $1/\gamma$ are the mean incubation and infectious periods, respectively. For directly transmitted infections (such as measles, whooping cough, influenza, etc.) the force of infection is modeled as proportional to the infection prevalence: $\lambda = \beta I$ (Dietz 1976). The constant of proportionality (coefficient of transmission β) combines epidemiological, environmental, and social factors that affect transmission rates. The

density dependence in the transmission process (bilinearity βIS) is the source of nonlinearity (and thus complex dynamics) of the SEIR system. The basic reproduction ratio R_0, defined as the average number of secondary infections produced by one infected individual introduced into a fully susceptible population (Anderson and May 1991), is expressed as

$$R_0 = \frac{\beta N \sigma}{(\gamma + \mu)(\sigma + \mu)} \tag{5}$$

Note that for most childhood diseases like measles or whooping cough, $\gamma \gg \mu$ and $\sigma \gg \mu$ making R_0 approximately independent of μ.

The system of equations (1–4) has two equilibrium: (i) the disease free scenario, $(S_0^*, E_0^*, I_0^*, R_0^*) = (N, 0, 0, 0)$ and (ii) the endemic case, $(S_1^*, E_1^*, I_1^*, R_1^*) = (\mu N/(\mu+\lambda), \lambda/(\mu+\sigma) \times \mu N/(\mu+\lambda), \sigma/(\mu+\gamma) \times \lambda/(\mu+\sigma) \times \mu N/(\mu+\lambda), N - S_1^* - E_1^* - I_1^*)$, the stability of which depend solely on R_0. If R_0 is less than unity, then the disease-free equilibrium is stable, while $R_0 > 1$ means the endemic equilibrium is stable. Perturbations to the endemic equilibrium result in damped oscillations before the equilibrium is recovered. Linear stability analysis reveals the natural period T of these damped oscillations to be approximated by

$$\widehat{T} = 2\pi\sqrt{AG} \tag{6}$$

where A represents the mean age at infection, $A \simeq 1/\mu(R_0 - 1)$, and G gives the ecological generation length of the infection, i.e. the sum of latent and infectious periods, $G = [1/(\mu + \gamma)] + [1/(\mu + \sigma)]$ (Anderson and May 1991; Rohani et al. 2002). Note that the life expectancy $1/\mu$ is about one order of magnitude higher than the average age at infection A.

In contrast to the model-predicted endemic equilibrium, the observed dynamics of most childhood diseases exhibit sustained (rather than damped) oscillations (London and Yorke 1973; Fine and Clarkson 1982; Anderson et al. 1984; Fine and Clarkson 1986; Hethcote 1998; Rohani et al. 1999; Earn et al. 2000; Grenfell et al. 2002). This may be explained by considering the damping (or return) time of the endemic equilibrium, which stability analysis reveals to be approximately $2A$ (Anderson and May 1991). For most epidemiologically reasonable parameter values, the damping time is typically much longer than the natural period: $2A/T \gg 1$. This renders the endemic equilibrium weakly stable, with relatively small perturbations (intrinsic or extrinsic) "exciting" the inherent oscillatory behavior (Grossman 1980). Alternative mechanisms for this phenomenon have been explored, including demographic stochasticity (Bartlett 1957) and temporal heterogeneity in the transmission rate (London and Yorke 1973; Fine and Clarkson 1982; Schenzle 1984; Fine and Clarkson 1986; Anderson and May 1991). Seasonal forcing – usually assumed to represent the aggregation of children in schools – can be modeled by making the coefficient of transmission a periodic function of time: $\beta = \beta(t)$. Among the numerous forms for $\beta(t)$ proposed in the literature (Dietz 1976; Ellner et al. 1998; Earn et al. 2000; Keeling et al. 2001) the simplest and most widely used one remains a sinusoidal wave:

$$\beta(t) = \beta_0 \left(1 + \beta_1 \cos\left(\frac{2\pi}{T_S} t\right)\right), \qquad 0 \leqslant \beta_1 < 1. \tag{7}$$

The strength of seasonality β_1 measures the amplitude of the oscillations around the baseline coefficient of transmission β_0, and T_S is the period of forcing, in the

same units as time, t. Variations of the coefficient of transmission are annual. However, since we are interested in the value of T_S relative to the natural period T, instead of studying different combinations of the epidemiological parameters β_0, σ, and γ, our approach is to keep T constant and vary T_S. Thus, in the simulations presented in the next section, we will take classically estimated measles values ($\beta_0 = 0.0002$ yr$^{-1}\cdot$individual^{-1}, $\sigma^{-1} = 7.5$ day, and $\gamma^{-1} = 6.5$ day; Anderson & May 1991) and let T_S vary from 0.1 to 10 yr. The population size will be fixed at $N = 5 \times 10^6$ individuals, yielding an R_0 around 17.

Resonance

Definition and exploration process

Given a system with temporal forcing and a weakly stable equilibrium, oscillator theory predicts interesting dynamics, whose complexity increases with nonlinearity (Jackson 1992). Within the context of the SEIR system, the linear approximation becomes reasonable when the number of susceptibles S is large enough for the density dependence in the transmission process (βIS) to be negligible. The key parameter controlling the recruitment of susceptibles is μ. Since host life expectancy $1/\mu$ is far higher than the average age at infection A (see previous section), S increases linearly with μ. Consequently, the higher the susceptible recruitment rate, the better the linear approximation and we would expect "simpler" dynamics. Conversely, the lower the recruitment rate, the larger the nonlinear influence (density dependence) and more "complicated" dynamics are expected. This influence of the recruitment rate on the complexity of the dynamics has been verified in measles dynamics using both simulations and time-series analysis (Earn et al. 2000; Grenfell et al. 2002). These authors also demonstrated the correspondence between the recruitment rate μ and the mean coefficient of transmission β_0, predicting similar dynamical consequences of increasing μ and increasing β_0. For simplicity we will vary only μ. Nevertheless, bear in mind hereafter that increasing (decreasing) μ is dynamically equivalent to increasing (decreasing) β_0. In this section we illustrate the dynamical properties of the seasonally forced SEIR model by first considering the simple case of linear dynamics (harmonic resonance) and then exploring the consequences of nonlinearity (parametric resonance and fold-over effects) by progressively decreasing the recruitment rate μ.

We now proceed to define some of the basic terminology of resonance phenomena. "Resonance" is a generic term that indicates the amplitude of observed oscillations depends on the period of the forcing and has a maximum. This maximum is called the "resonance peak", with a corresponding "resonance period" (the forcing period at which the resonance peak occurs). As stated above, model dynamics are approximately linear when μ is large. When periodically excited, linear systems exhibit harmonic oscillations, meaning they oscillate with the same period as the forcing, which is not necessarily equal to the natural period (T in the SEIR model).

Harmonic resonance

A fundamental behavior of harmonic (i.e. linear) oscillators, when driven by a periodic force, is harmonic resonance where the unique resonance period is equal to the natural period (Jackson 1992). In the specific case of the forced SEIR system, this would mean that whenever the period of seasonality T_S equals the natural period T, maxima in the coefficient of transmission $\beta(t)$ and the number of infectives $I(t)$ coincide, thus amplifying the magnitude of the oscillations. These possibilities are explored in fig. 1A, which shows a resonance diagram (Jordan and Smith 1999) where the peak and the trough values in the number of infectives are plotted against the logarithm of the forcing period T_S. The proximity of the resonance period (resonance peak at $T_S = T \simeq 1.05$ yr on fig. 1A) to the estimation of the natural period from equation (6) (vertical line at $T_S = \widehat{T} \simeq 0.95$ yr on fig. 1A) is confirmation of the linear approximation: $(T - \widehat{T})/T \simeq 0.10$. Indeed, as μ further increases, $T - \widehat{T}$ tends towards 0 (results not shown).

Parametric resonance

As the recruitment rate μ further decreases, nonlinearity increases, thus rendering the dynamics even more complex. In contrast to harmonic oscillators, a key property of nonlinear oscillators is that the amplitude and the frequency of oscillations are not independent (Jackson 1992). This property is visible on fig. 1. As μ decreases (from fig. 1A to fig. 1D), the amplitude of fluctuations decreases whereas the natural period T increases. Note also that the approximation $\hat{T}$ of the natural period T by equation (6) (vertical lines on fig. 1) is expected to worsen as μ decreases, with $T - \widehat{T}$ increasing exponentially. Let focus only on the peak values of the number of infectives. Fig. 2A illustrates then the nonlinear effect of a continuous decrease in μ (from 0.10 to 0.01/person/yr) on the dynamics behavior. It is noteworthy that this range of variation is not unnatural as the effective population birth rate varies between 0.01/person/yr in the western countries and 0.05/person/yr in some African countries.

Nonlinear oscillators may exhibit "parametric resonance" when one of the parameters (here the transmission coefficient, β) depends on time. Parametric resonance differs from harmonic resonance in that it is an instability phenomenon stemming from the aforementioned weak stability of the endemic equilibrium (Jackson 1992). Oscillations in one parameter may destabilize this equilibrium through a bifurcation. For the SEIR model, recall that in the case of linear oscillations (see above), the period of oscillations in the number of infectives I is equal to the forcing period T_S and is not necessarily equal to the natural period T. Consider the case where $T_S = T/2$, i.e. where the transmission coefficient undergoes two complete cycles for each natural epidemiological cycle in numbers of infectives and susceptibles. Here, every second peak in the coefficient of transmission coincides with a peak in the number of infectives, thus amplifying every second peak in I. The consequence is to destabilize the attractor with period $T_S = T/2$ through a bifurcation towards an attractor of period T. Such bifurcations are visible for example at $T_S = 2T \simeq 2$ yr in fig. 1A, $T_S = T/2 \simeq 0.7$ yr and $T_S = 2T \simeq 2.8$ yr in fig. 1B, and $T_S = T/2 \simeq 1$ yr and $T_S = 2T \simeq 4$ yr in fig. 1C. The amplitude of the unstable solution grows exponentially with T_S and the above mentioned period–amplitude rela-

tionship causes a rapid shift of the natural period out of the resonance domain. This produces the characteristic asymmetric peak of parametric resonance: for increasing T_S, the dynamics bifurcate and the amplitude gradually increases before a sudden drop to approximately the same behavior at which the biennial behavior emerged (see for example fig. $1B$ at $T_S \simeq 0.7$). When nonlinearity further increases, the period–amplitude relation accentuates the asymmetry of the peak, bending it further and even folding it over itself (see for example fig. $1C$ at $T_S \simeq 1$ or fig. $1D$ where it is even more pronounced). This "fold-over" effect leads to bistability and hysteresis, i.e. the system oscillates either with a large or a small amplitude with an unstable periodic solution in-between. At the end of the interval of bistability this unstable limit cycle annihilates with one of its stable counterparts in a saddle-node bifurcation.

Dependence among parameters in parametric resonance

Parametric resonance is expected at integer fractions of the natural period, once a control parameter has exceeded a certain threshold, with each parametric resonance peak having its own threshold value. Grossman et al. (1977) and Grossman (1980) have reported a threshold effect for the seasonal forcing amplitude β_1 (see also Schwartz and Smith 1983 and Smith 1983 for more rigorous mathematical proofs for the SIR and SEIR models respectively) and this has been verified in our simulations (results not shown). Another threshold effect is associated with the value of the baseline coefficient of transmission β_0. This is visible on fig. 1, recalling the aforementioned correspondence between μ and β_0: parametric resonance occurs around $T_S = 2$ yr in fig. $1A$ and when β_0 decreases (or, equivalently, μ decreases) from fig. $1A$ to fig. $1D$, the number of parametric peaks increases. This is even more apparent on fig. $2A$ where the dependence of the resonance period on β_0 is responsible for the curvature observed on the figure. Note that this effect of β_0 is particularly pronounced for the biologically realistic values of μ ($0.01 < \mu < 0.05$) where the nonlinearity is high. The relation between the thresholds on β_1 and on β_0 is treated in Aron and Schwartz (1984). Grossman (1980) derived analytical formulations of the thresholds associated to each parametric resonance peak and showed a direct correlation between the aforementioned local stability of the system (defined by the $2A/T$ ratio) and its excitability, here defined by the threshold value on β_1 (see Schwartz and Smith 1983 and Smith 1983 for rigorous mathematical treatments). Subharmonic parametric resonance has been suggested to explain the biennial cycles of measles epidemics (London and Yorke 1973; Stirzaker 1975). Moreover, Schwartz and Smith (1983) and Smith (1983) showed that several subharmonics of different periods can be simultaneously stable and further pointed out that random effects in the environment could perturb the state of the system from the domain of attraction of one subharmonic to that of another, producing aperiodic looking levels of incidence. This is greatly enhanced by the fact that the basins of attraction of the different subharmonics are largely intertwined (Schwartz 1985; Earn et al. 2000).

Detection of resonance in measles data

Fig. 1 and fig. $2A$ illustrate, in a general theoretical context, resonance phenomena for a given recruitment rate μ when T_S is varied. Because of the curvature

observed in fig. $2A$, all the phenomena described above are also observed for a given period T_S of the seasonal forcing when μ is varied. This observation is practically relevant for the study of a particular disease since it implies that small variations in birth rate and/or vaccination coverage may dramatically change the severity of the epidemics and not necessarily in an intuitive manner. This generally echoes the findings of Dietz (1976), who studied the effect of R_0 (see equation (5) for the relation between R_0 and μ) in an SIR framework with annual oscillations in the coefficient of transmission. We explored these predictions using weekly notification data for measles, and associated demographic data, for 60 towns and cities of England and Wales in the pre-vaccine era (1944–1966). See Bjørnstad et al. 2002 for more details on the data set.

For measles parameter values, the model predicts parametric resonance in the dynamics: setting $T_S = 1$ yr in fig. $2A$ and increasing μ in a biologically meaning domain (i.e. from 0.01 to 0.05/person/yr), we observe a peak followed by an increase in the number of infectives. We looked for this pattern in the data by calculating the maximum of the attractor and plotting it against the per capita birth rate, with the trend smoothed by a lowess regression (with a tensor parameter of 0.45). In practice, for each city, we truncated the measles notification and birth time series in adjacent intervals of a fixed duration. On each of these intervals we considered the largest reported number of cases and the mean number of births. Both of these quantities were divided by the median city size between 1944 and 1966, producing our estimations of the maximum of the attractor and its corresponding per capita birth rate. Fig. $2B$ shows the results for a time interval of 1 year, though analyses for time intervals of 2, 3, and 4 years gave similar results. Each curve in fig. $2B$ corresponds to an analysis on a specified range of city sizes (see legend of fig. $2B$). The observed patterns fit model predictions remarkably well irrespective of the city sizes: compare fig. $2B$ with fig. $2A$ for $T_S = 1$ yr and μ varying from 0.01 to 0.05/person/yr. Therefore, these results highlight the fact that changes in the recruitment in susceptibles (either due to systematic trends in the crude birth rate or vaccination) not only have qualitative consequences through dynamical transitions but, due to resonance effects, may also have major quantitative consequences such as dramatic changes in the amplitude of the oscillations.

Pulse Vaccination

Definition and theoretical background

The most commonly used strategy for vaccination against infections such as measles is to immunize infants once they have reached a certain age (e.g. 12–25 months for the MMR vaccine in the USA). This process however requires a too high vaccination coverage (around 95% for measles) for disease eradication to be achieved in practice. An alternative (and potentially less expensive) strategy is vaccination in pulses (Agur et al. 1993; Nokes and Swinton 1997). This approach, based on theoretical results on population dynamics in varying environments (Agur 1985), consists in vaccinating a proportion p of the susceptible population every T_V years. The essential aim is to antagonize (or entrain) natural dynamics by a different temporal process. This theory has been successfully applied in campaigns against poliomyelitis and measles in Central and South

America and measles in the UK in 1994 (see references in d'Onofrio 2002).

The theoretical challenge of pulse vaccination is the analytical determination (for specified values of p and μ) of the optimal value $T_V^{\max}$ which ensures the eradication of the disease. The rationale behind this, derived from the SIR model, is simply to ensure the proportion of susceptibles $S(t)/N$ remains below the threshold $s_c = 1/R_0$ required for an increase in the number of infectives. This led to the approximation $T_V^{\max} = A$, where A is the mean age at infection (Agur 1993). Further detailed analyses revealed that to prevent an epidemic, it is sufficient that the mean value of $S(t)/N$, averaged over the pulsing period, remains below s_c. For an SIR model with constant coefficient of transmission, this led to the following relationship between $T_V^{\max}$ and p (Shulgin et al. 1998):

$$T_V^{\max} = \frac{p\,\gamma}{\beta\mu(1 - p/2 - \gamma/\beta)} \tag{8}$$

This value of $T_V^{\max}$ can be substantially larger than the mean age at infection A and remains a good approximation for the SEIR model, as long as β is constant (d'Onofrio 2002). When the transmission rate is periodic, β in equation (8) should be replaced by its mean value β_0 (Shulgin et al. 1998; d'Onofrio 2002). Once $T_V < T_V^{\max}$, the disease is thus expected to disappear. Should $T_V > T_V^{\max}$, however, the pulsed nature of this vaccination strategy may give rise to a rich variety of dynamics (Shulgin et al. 1998). One important potential consequence is an increased likelihood in epidemic synchrony across sub-populations (Earn et al. 1998). Another possible consequence of vaccination in pulses may be resonance effects associated with the frequency of vaccination events (as briefly mentioned by d'Onofrio 2002). We show here, via extensive numerical simulations, the conditions under which pulse vaccination can result in very large epidemics.

Investigating the resonance effect of pulse vaccination

A proportion p of the susceptible population is now vaccinated every T_V years and the following equation should be added to equations (1–4):

$$S(kT_V^+) = (1-p)\cdot S(kT_V^-), \qquad k \in \mathbb{N} \tag{9}$$

where T_V^- and T_V^+ respectively refer to the instants that immediately precedes and follows the vaccination pulse. Given a vaccination proportion p and $T_V > T_V^{\max}$, two parameters may further influence the onset of resonance: the susceptible recruitment rate μ (or, equivalently, the mean coefficient of transmission β_0), and the amplitude of seasonality in transmission (β_1; equation (7)) – see previous section. In the previous section we studied the resonance effects associated with the seasonal forcing of the coefficient of transmission. We are now interested in the resonance phenomenon related to the periodic nature of the pulse vaccination. We proceed step by step in order to distinguish the resonance effects due to pulse vaccination from those due to seasonal variations in transmission. First consider the model given by equations (1–4) with constant transmission (i.e. $\beta_1 = 0$ in equation (7)). In this setup, fig. 3 shows resonance diagrams plotting the peak and trough values of the number of infectives against the logarithm of the period of vaccination T_V. To identify the influence of the system nonlinearities we proceed as before (fig. 1) through variation in population turn-over rate μ. Note nevertheless that the variations on μ are smaller

in fig. 3 than those in fig. 1 and that they encompass biologically realistic values only: 0.05/person/yr (fig. 3*A*), 0.04/person/yr (fig. 3*B*), 0.03/person/yr (fig. 3*C*), and 0.02/person/yr (fig. 3*D*). More precisely, fig. 3*D* corresponds to the birth rate observed in western countries while fig. 3*A* to that observed in some African countries.

Shulgin et al. (1998) pointed out that when $T_V > T_V^{\max}$ the solution $I(t) = 0$ becomes unstable and the number of infectives begins to oscillate with large amplitude; when T_V is further increased, a sequence of period doubling bifurcations interspersed with chaos is observed. Here, we further observe that a slight increase in the recruitment rate μ from 0.02 (fig. 3*D*) to 0.05 (fig. 3*A*) reduces the chaotic region. A second effect of nonlinearity associated with the decrease of the population turn-over rate μ from fig. 3*A* to fig. 3*D* is the shift of the attractor structure towards the high values of the period of the forcing. This phenomenon has been already observed on fig. 1 and is due to the above-mentioned increased dependency between the amplitude and the frequency of the oscillations when the level of nonlinearity increases (Jackson 1992). For simplicity, we will focus our analysis on the structure of the attractor of a non-chaotic dynamics (fig. 3*A*). Please note that the following observations are robust relative to the complexity level of the dynamics. This robustness appears when taking into account the temporal dimension which is not visible in the resonance diagrams of fig. 3. Indeed, when we estimate the duration spent by the dynamics in each part of its chaotic attractor of fig. 3*D*, we get a structure pretty close to that of fig. 3*A* (results not shown).

The amplitude of the oscillations in the number of infectives changes with T_V. Each peak which occurs at a forcing period T_V higher than 2 are harmonics of the main resonance peak observed at $T_V \simeq 2$ yr. The bifurcations associated with each increase in the amplitude indicate instability and are characteristic of parametric resonance. As visible in fig. 3*A*, the resonance phenomenon due to the periodic nature of this vaccination strategy may locally give rise to unexpected increases in the epidemic peaks as the frequency $1/T_V$ of vaccination increases. Moreover, increases in the peaks tend to be associated with deeper troughs (not shown) which thus may increase the probability of disease extinction in the case of small populations. More importantly, the duration of the epidemic is also influenced (see below).

Introducing seasonality to pulse vaccination's resonance

Consider now the case with β oscillating according to equation (7), with a fixed period T_S equal to 1 yr, intended to mimic the seasonal alternation of holidays and school terms in the case of a measle-like disease. In practice, two β_1 values ($\beta_1 = 0.01$ and $\beta_1 = 0.10$) have been added to the dynamics described in fig. 3*A* (where $\beta_1 = 0$). In both seasonality cases, our simulations tracked how four critical quantities behave when T_V varies; namely the number of infectives in a resonance diagram (fig. 4*A*–4*B*), the number of contracting people (fig. 4*C*–4*D*), the annual mean number of infectives (fig. 4*E*–4*F*), and the number of vaccinated people (fig. 4*G*–4*H*). Clearly, increasing β_1 from 0 (fig. 3*A*) to 0.01 (fig. 4*A*) or the more realistic value (Earn et al. 2000) of 0.10 (fig. 4*B*) increases variation in the dynamics of infectives, leading to chaos. Note however that the general structure of the resonance peaks is not affected by the level of complexity that may be present in the dynamics. Indeed, the confusing patterns obtained

when $\beta_1 = 0.10$ (fig. 4B) resemble those obtained for $\beta_1 = 0.01$ (fig. 4A) and for $\beta_1 = 0$ (fig. 3A) when retaining only the spots where the dynamics spend the longest time (details unshown).

The two levels of seasonality ($\beta_1 = 0.01$ and $\beta_1 = 0.10$) also differentially affect the instantaneous number of new infectives (fig. 4C–4D). Here the number tracked corresponds to the individuals entering the infectious compartment between two vaccination pulses, scaled by the between-vaccination duration: $(1/T_V)\int_{kT_V}^{(k+1)T_V}\sigma E(t)dt$. Whatever the value of β_1, the peaks in incidence (fig. 4C–4D) are associated with those observed in the resonance diagrams (fig. 4A–4B). When $\beta_1 = 0.10$ it remains visible: the incidence exhibit a sudden change at $T_V \simeq 2$ yr (fig. 4D) that is precisely where the main peak on the resonance diagram occurs (fig. 4B).

More relevant from an epidemiological point of view may be the mean annual number of infectives. Its variation with T_V is shown in fig. 4E–4F. Here the mean considered was computed over a period encompassing 200 vaccination events, which has the advantage of smoothing the irregularities due to the complexity of the dynamics. Despite the general increase in the mean annual prevalence with T_V, these two plots clearly exhibit peaks wherever parametric resonance is observed in the resonance diagram (compare fig. 4A–4B with fig. 4E–4F). As expected, the observed pattern is even more clear-cut when the number of people actually vaccinated remains fixed as T_V varies (not shown). Practically, these resonance phenomena mean, somewhat counter-intuitively, that the mean annual number of infectives may locally (i.e. on the resonance domain) increase when T_V decreases. For example, in fig. 4F, vaccination every 24 months yields a mean annual number of infectives (1,275,000) that is 10% higher than vaccination every 28 months (1,155,000). As β_1 decreases, this difference becomes even more pronounced: in fig. 4E, vaccination every 25 months yields a mean annual number of infectives equal to 1,315,000 which is 15% higher than a vaccination every 26 months which yields a mean annual number of infectives equal to 1,145,000.

Lastly, let focus on the number of people effectively vaccinated and its variation with β_1 and T_V (in black in fig. 4G–4H). This number is simply calculated from $p \cdot S(kT_V^-)$, $k \in \mathbb{N}$. Fig. 4G–4H also show in grey the minimum number of people to vaccinate to theoretically reach eradication. This is calculated from $p_{\min} \cdot S(kT_V^-)$, $k \in \mathbb{N}$, where $p_{\min}$ is derived from equation (8) which does not explicitly accounts for the seasonality on the coefficient of transmission (Shulgin et al. 1998; d'Onofrio 2002). Interestingly, this resonance phenomenon occurs both for low values of p and relatively high values of p, close to the level needed to achieve eradication. Note that while $T_V < 2$ yr means the number of people effectively vaccinated is higher than the number required to theoretically reach eradication (from equation (8)), we still observe disease persistence. However, the amplitude of outbreaks decreases dramatically, resulting in rapid extinction shortly after $T_V < 1$ yr (results not shown). Either approximate equation (8) is not appropriate for this particular system or, and more probably, parametric resonance delays eradication when T_V decreases (or, equivalently, when p increases). In the latter case, the inexactitude of the criterion behind equation (8) would certainly be due the fact that its derivation is based on models simple enough for analytical calculus but which do not make full account of the dynamics of the disease itself. It is noteworthy again that

the general structure of the graph is not affected by the level of complexity that may be present in the dynamics.

Discussion

Epidemiological systems describing strongly immunizing infections are inherently oscillatory and thus are expected to produce resonance phenomena when undergoing periodical forcing (Jackson 1992). For childhood diseases, one obvious mechanism for forcing is the seasonal variation in transmission due to the alternation of holidays and school terms (London and Yorke 1973; Fine and Clarkson 1982; Schenzle 1984; Fine and Clarkson 1986; Anderson and May 1991). A second cause of external forcing is associated with the periodic nature of pulse vaccination strategy (when the vaccination coverage and/or the pulse frequency are not high enough to eradicate the disease; d'Onofrio 2002). In this study, we have numerically explored the dynamical effects of the phenomenon of resonance that may be associated with this vaccination policy.

We first considered the simple SEIR model with a sinusoidal coefficient of transmission. Here, the source of nonlinearity is the density-dependent transmission process and the severity of nonlinearity increases as the population turn-over rate μ decreases or, equivalently, when the mean coefficient of transmission β_0 decreases (Earn et al. 2000). Resonance diagrams with the period T_S of the seasonal forcing as the control parameter show the existence of two kinds of resonance: harmonic and parametric, depending on the strength of nonlinearity. Harmonic resonance is characterized by a single high peak in the amplitude of epidemics and occurs when $T_S \simeq T$, with T the natural period of the system in the absence of forcing. Parametric resonance is an instability phenomenon and is characterized by a series of small peaks in the amplitude of epidemics. These peaks occur at integer fractions of the natural period T. The appearance of each parametric resonance peak occurs at a threshold value on the level of nonlinearity (as determined by, for example, the population turn-over rate μ) and the values of these thresholds depend on the amplitude β_1 of the seasonal forcing. The curvature observed on fig. $2A$ is responsible for the fact that the effect of resonance may be also detected when the level of nonlinearity varies, for a fixed period of the seasonal forcing. This prediction has been successfully verified on real data of measles cases from England and Wales in the pre-vaccine era.

In a similar way, we investigated the potential for resonance associated with the periodic nature of pulse vaccination. For simplicity, we began by considering pulse vaccination in absence of seasonality in transmission. We clearly detected parametric resonance, which results in higher peaks and deeper troughs in infective numbers. Obviously, deeper troughs may increase the probability of disease extinction, especially in small populations. However, the effects of resonance are substantially more dramatic on the peaks than the troughs (fig. 3). Moreover, we found an unexpected local (in terms of vaccination frequency) increase in the number of infectives as vaccination pulses become more frequent. This finding holds even when the threshold vaccination level required for eradication in the absence of seasonality is exceeded. From then, we added seasonal variation in transmission. Our simulations show that the inclusion of seasonality has little impact on the system behavior (including the structure of the

resonance peaks) other than to force the oscillations into a chaotic mode. For simplicity, the effects of resonance were investigated on situations where the dynamics behave relatively simply (i.e. high population turn-over rate μ, high mean coefficient of transmission β_0, and low amplitude β_1 on the coefficient of transmission). Nevertheless, our conclusions were robust relative to the level of complexity in the dynamics and remained unchanged when considering realistic μ, β_0, and β_1 values.

The results outlined in the paper have public health implications pertaining to pulse vaccination strategies. The classical mass vaccination scheme is based on the static properties of the host-disease system. What makes the recently developed pulse vaccination theory potentially more efficient is that it explicitly accounts for the dynamics of the host population through the influx of susceptibles (via births). Substantial progress has been accomplished in determining the optimal vaccine coverage p and pulsing frequency T_V in different epidemiological contexts (Agur et al. 1993, Nokes and Swinton 1997, Shulgin et al. 1998, Stone et al. 2000, d'Onofrio 2002). However, largely to facilitate mathematical tractability, these models do not take into account the dynamics of the disease itself. Indeed, the threshold susceptible fraction necessary for an epidemic is still determined by the equilibrium properties of the system and corresponds to $1/R_0$ as in the classical mass vaccination strategy. Our simulations pinpointed this caveat since the presumed optimal T_V and p determined by such models (Shulgin et al. 1998; d'Onofrio 2002; see equation (8)) did not obligatorily lead to eradication, contrary to previous expectations. We believe that this discrepancy between theoretical predictions and our simulation results is due to the neglect of the inherent disease dynamics in the analytical derivation of the optimal T_V and p.

We have also showed that the interference between the intrinsic diseases dynamics and the periodic nature of an imperfect pulse vaccination scheme may produce perverse effects, such as an increase in the number of infectives with the frequency of vaccination. In consequence, if we are to optimize further the pulse vaccination strategy, a full account of the disease dynamics should be made. From a theoretical perspective, there is thus a need for the development of analytical epidemiological models that incorporate information such as the natural period of the disease in question. From a practical point of view, we may adopt a more ad hoc approach. In other words, the design of a pulse vaccination policy on a particular epidemiological system should be preceded by the determination of the disease dynamics characteristics and its resonance domain. Simulation models will be of great help for such tasks.

Acknowledgments

We would like to thank Alberto d'Onofrio for his comments and advices which greatly improved the first versions of the manuscript, and Sylvain Gandon for useful discussions, and suggestions. Christine Chevillon brought invaluable comments on the last versions of the manuscript. MC is funded by a BDI CNRS/Région Languedoc-Roussillon, PR is supported by the Ellison Medical Foundation and JFG is sponsored by IRD and CNRS.

Literature Cited

Agur, Z. 1985. Randomness synchrony population persistence. Journal of Theoretical Biology 112:677-693.

Agur, Z., L. Cojocaru, R. M. Anderson, and Y. L. Danon. 1993. Pulse mass measles vaccination across age cohorts. Proceedings of the National Academy of Sciences of the USA 90:11698–11702.

Anderson, R. M., B. T. Grenfell, and R. M. May. 1984. Oscillatory fluctuations in the incidence of infectious disease and the impact of vaccination: time series analysis. Journal of Hygiene, Cambridge 93:587–608.

Anderson, R. M., and R. M. May. 1991. Infectious diseases of humans. Dynamics and control. Oxford University Press, Oxford.

Aron, J., and I. Schwartz. 1984. Seasonality and period-doubling bifurcations in an epidemic model. Journal of Theoretical Biology 110:665–679.

Bailey, N. T. J. 1975. The mathematical theory of infectious diseases. Charles Griffin and Company Ltd, London.

Bartlett, M. S. 1957. Measles periodicity and community size. Journal of the Royal Statistical Society A120:48–70.

Bjørnstad, O. N., B. Finkenstädt, and B. T. Grenfell. 2002. Dynamics of measles epidemics: estimating scaling of transmission rates using a time series SIR model. Ecological Monographs 72:169–184.

Billings, L., and I. B. Schwartz. 2002. Exciting chaos with noise: unexpected dynamics in epidemic outbreaks. Journal of Mathematical Biology 44:31–48.

Bolker, B. M., and B. T. Grenfell. 1993. Chaos and biological complexity in measles dynamics. Proceedings of the Royal Society of London B 251:75-81.

Bolker, B. M., and B. T. Grenfell. 1995. Space, persistence and dynamics of measles epidemics. Philosophical Transaction of the Royal Society of London B 348:309–320.

Bolker, B. M., and B. T. Grenfell. 1996. Impact of vaccination on the spatial correlation and persistence of measles dynamics. Proceedings of the National Academy of Sciences of the USA 93:12648–12653.

Deguen, S., and A. Flahault. 2000. Impact of immunization of seasonal cycle of chickenpox. European Journal of Epidemiology 16:1177–1181.

Dietz, K. 1976. The incidence of infectious diseases under the influence of seasonal fluctuations. Lecture Notes in Biomathematics 11:1–5.

d'Onofrio, A. 2002. Stability properties of pulses vaccination strategy in SEIR epidemic model. Mathematical Biosciences 179:57–72.

Earn, D. J. D., P. Rohani, B. M. Bolker, and B. T. Grenfell. 2000. A simple model for complex dynamical transitions in epidemics. Science 287:667–670.

Earn, D. J. D., P. Rohani, and B. T. Grenfell. 1998. Persistence, chaos and synchrony in ecology and epidemiology. Proceedings of the Royal Society of London B 265:7–10.

Ellner, S., B. A. Bailey, G. V. Bobashev, A. R. Gallant, B. T. Grenfell, and D. W. Nychka. 1998. Noise and nonlinearity in measles epidemics: combining mechanistic and statistical approaches to population modeling. American Naturalist 151:425-440.

Fine, P. E. M., and J. A. Clarkson. 1982. Measles in England and Wales. 1. An analysis of factors underlying seasonal patterns. International Journal of Epidemiology 11:5–14.

Fine, P. E. M., and J. A. Clarkson. 1986. Seasonal influences on pertussis. International Journal of Epidemiology 15:237–247.

Glendinning, P., and L. P. Perry. 1997. Melnikov analysis of chaos in a simple epidemiological model. Journal of Mathematical Biology 35:359–373.

Grenfell, B. T., O. N. Bjørnstad, and B. F. Finkenstädt. 2002. Dynamics of measles epidemics: scaling noise, determinism, and predictability with the TSIR model. Ecological Monograph 72:185–202.

Grenfell, B. T., O. N. Bjørnstad, and J. Kappey. 2001. Travelling waves and spatial hierarchies in measles epidemics. Nature 414:716–723.

Greenman, J., M. Kamo, M. Boots. 2004. External forcing of ecological and epidemiological systems: a resonance approach. Physica D 190:136–151.

Grossman, Z. 1980. Oscillatory phenomena in a model of infectious diseases. Theoretical Population Biology 18:204–243.

Grossman, Z., I. Gumowski, and K. Dietz. 1977. The incidence of infectious diseases under the influence of seasonal fluctuations – analytical approach. Pages 525–546 *in* V. Lakshmikantham, ed. Nonlinear systems and applications. Academic Press, New-York.

Hethcote, H. W. 1998. Oscillations in an endemic model for pertussis. Canadian Applied Mathematics Quarterly 6:61–88.

Jackson, E. A. 1992. Perspectives of nonlinear dynamics: volume 1. Cambridge University Press, Cambridge.

Jordan, D. W., and P. Smith. 1999. Nonlinear Ordinary Differential Equations. Oxford University Press, Oxford.

Keeling, M. J., and B. T. Grenfell. 1997. Disease extinction and community size: modeling the persistence of measles. Science 275:65–67.

Keeling, M. J., P. Rohani, and B. T. Grenfell. 2001. Seasonally forced disease dynamics explored as switching between attractors. Physica D 148:317–335.

Lloyd, A. L., and R. M. May. 1996. Spatial heterogeneity in epidemic models. Journal of Theoretical Biology 179:1–11.

London, W. P., and J. A. Yorke. 1973. Recurrent outbreaks of measles, chickenpox and mumps. I. Seasonal variation in contact rates. American Journal of Epidemiology 98:453–468.

Nokes, D. J., and J. Swinton. 1997. Vaccination in pulses: a strategy for global eradication of measles and polio? Trends in Microbiology 5:14–19.

Rhodes, C. J., H. J. Jensen, and R. M. Anderson. 1997. On the critical behaviour of simple epidemics. Proceedings of the Royal Society of London B 264:1639–1646.

Rohani, P., D. J. D. Earn, and B. T. Grenfell. 1999. Opposite patterns of synchrony in sympatric disease metapopulations. Science 286:968–971.

Rohani, P., and D. J. D. Earn, and B. T. Grenfell. 2000. Impact of immunisation on pertussis transmission in England and Wales. The Lancet 355:285–286.

Rohani, P., M. J. Keeling, and B. T. Grenfell. 2002. The interplay between determinism and stochasticity in childhood diseases. American Naturalist 159:469–481.

Schaffer, W. M., and M. Kot. 1985. Nearly one dimensional dynamics in an epidemic. Journal of Theoretical Biology 112:403–427.

Schenzle, D. 1984. An age-structured model of pre- and post-vaccination measles transmission. IMA Journal of Mathematics Applied in Medicine and Biology 1:169–191.

Schwartz, I. 1985. Multiple stable recurrent outbreaks and predictability in seasonally forced nonlinear epidemic models. Journal of Mathematical Biology 21:347–361.

Schwartz, I., and H. Smith. 1983. Infinite subharmonic bifurcation in an SEIR epidemic model. Journal of Mathematical Biology 18:233–253.

Shulgin, B., L. Stone, and Z. Agur. 1998. Pulse vaccination strategy in the SIR epidemic model. Bulletin of Mathematical biology 60:1123–1148.

Smith, H. 1983. Multiple stable subharmonics for a periodic epidemic model. Journal of Mathematical Biology 17:179–190.

Stirzaker, D. 1975. A perturbation method for the stochastic recurrent epidemic. Journal of the Institute of Mathematics and its Applications. 15:135–160.

Stone, L., B. Shulgin, and Z. Agur. 2000. Theoretical examination of the pulse vaccination policy in the SIR epidemic model. Mathematical and Computer Modelling 31:207–215.

Tidd, C. W., L. F. Olsen, and W. M. Schaffer. 1993. The case for chaos in childhood epidemics. II. Predicting historical epidemics from mathematical models. Proceedings of the Royal Society of London B 254:257–273.

Tuckwell, H. C., L. Toubiana, and J. F. Vibert. 2000. Enhancement of epidemic spread by noise and stochastic resonance in spatial network models with viral dynamics. Physical Review E 61:5611–5619.

Figure Legends

Figure 1 Resonance diagrams showing the effect of seasonality in transmission. The peak and trough values of the number of infectives (as determined by equations (1–4)) are plotted against the period T_S of the seasonal forcing modeled by equation (7). Parameter values are $N = 5 \times 10^6$ individuals, $\sigma^{-1} = 7.5$ days, $\gamma^{-1} = 6.5$ days, $\beta_0 = 0.0002\,\mathrm{yr}^{-1} \cdot \mathrm{individual}^{-1}$, and $\beta_1 = 0.1$, yielding a R_0 approximately equal to 17. The recruitment rate takes the following values: $\mu^{-1} = 10$, 20, 40, 70 yr for A, B, C, and D respectively. The attractors were determined from a 20-year period, after 180 years of transients were discarded. For each diagram, 1,001 dynamics were simulated, regularly spaced on the decimal logarithm scale of the period. Initial conditions were $S = 0.05N$, $E = I = 0.0001N$. The vertical lines correspond to the estimation $\hat{T}$ of the inherent oscillatory periods T from the linear approximation of equation (6). Note the logarithm scale on the x-axes.

Figure 2 A, resonance diagram showing the effects of seasonality in transmission and nonlinearity. The peak values of the number of infectives (as determined by equations (1–4)) are plotted against both the period T_S of the seasonal forcing modeled by equation (7) and the recruitment rate μ. The model is given by equations (1–4) and equation (7). Initial conditions and parameter values are as in fig. 1 except that μ varies from 0.01 to 0.1 yr^{-1} by step of 0.001. B, relationship between the maximum number of infectives and the birth rate in England and Wales in the pre-vaccine era (1944–1966). For each year from 1944 to 1966 the maximum number of infectives and the number of births were recorded. Both of these numbers were divided by the median city size between 1944 and 1966. The plotted points are the smoothed values after a lowess regression with a tensor parameter equal to 0.45. The analysis was performed on different data sets: the complete 60 cities (□), the 13 cities below 100,000 inhabitants (○), the 25 cities between 100,000 and 160,000 inhabitants (+), the 22 cities above 160,000 inhabitants (∗), and the 14 cities above 250,000 inhabitants (△) as this last threshold corresponds to the critical community size yielding regular endemic oscillations.

Figure 3 Resonance diagrams showing the effect of pulse vaccination, ignoring seasonality in transmission. The peak and trough values of the number of infectives (as determined by equations (1–4)) are plotted against the period T_V of the pulse vaccination modeled by equation (9). The attractors were determined from a period encompassing 200 vaccination events after the first 300 years and the initial 100 vaccination events were discarded. For each diagram, 501 dynamics were simulated, regularly spaced on the decimal logarithm scale of the period. Initial conditions were $S = 0.05N$, $E = I = 0.0001N$. Parameter values are the same as in figure 1 except that here $\beta_1 = 0$ and $T_V = 1$ yr in equation (7). The vaccination coverage $p = 0.4$ in equation (9) and the recruitment rate takes the following values: $\mu = 0.05$, 0.04, 0.03, 0.02 yr^{-1} for A, B, C, and D respectively. Note the logarithm scale on the x-axes.

Figure 4 Mixing seasonality and pulse vaccination. Dynamics of the model

given by equations (1–4) with an oscillating coefficient of transmission modeled by equation (7) and pulse vaccination modeled by equation (9). As in fig. 3*A*, $\mu = 0.05$ yr^{-1} and $p = 0.4$ in equation (9). In addition, $T_V = 1$ yr in equation (7). Seasonality is then introduced with $\beta_1 = 0.01$ (left column) and $\beta_1 = 0.10$ (right column). Each row shows the behavior of four critical quantities when T_V varies. *A–B*, resonance diagrams of the number of infectives. *C–D*, instantenous number of people who get diseased between two vaccination pulses. This number is calculated by $(1/T_V)\int_{kT_V}^{(k+1)T_V}\sigma E(t)dt$ (see text). *E–F*, mean annual number of infectives. This is calculated on the period encompassing the last 200 events of vaccination. *G–H*, actual number of vaccinated people in the simulations (in black) and, in grey, the optimal number, as estimated by d'Onofrio (2002). The effective number is simply $p \cdot S(kT_V^-)$. The optimal estimation is $p_{\min} \cdot S(kT_V^-)$ where $p_{\min}$ is calculated from equation (8) and $k \in \mathbb{N}$. Initial conditions were $S = 0.05N$, $E = I = 0.0001N$. Note the logarithm scale on the x-axes.

Figure 1

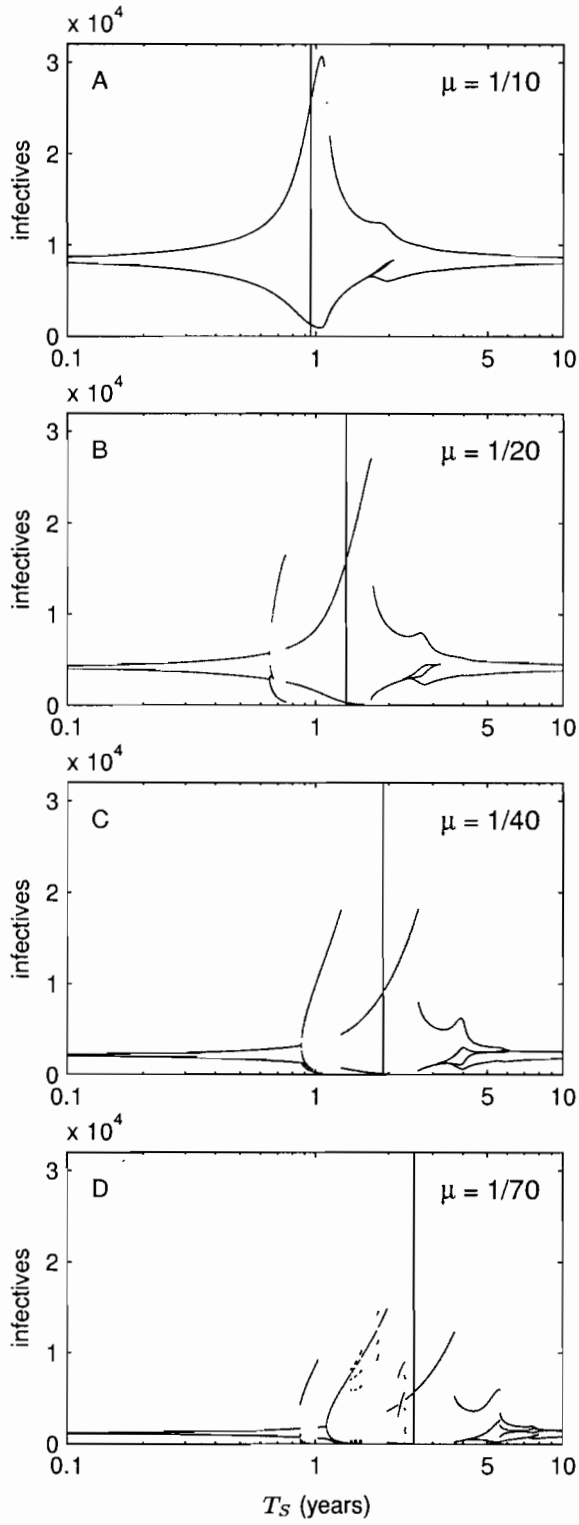

x 10^4
A
μ = 1/10
B
μ = 1/20
C
μ = 1/40
D
μ = 1/70
infectives
0
1
2
3
0.1
1
5
10
T_S (years)

Figure 2

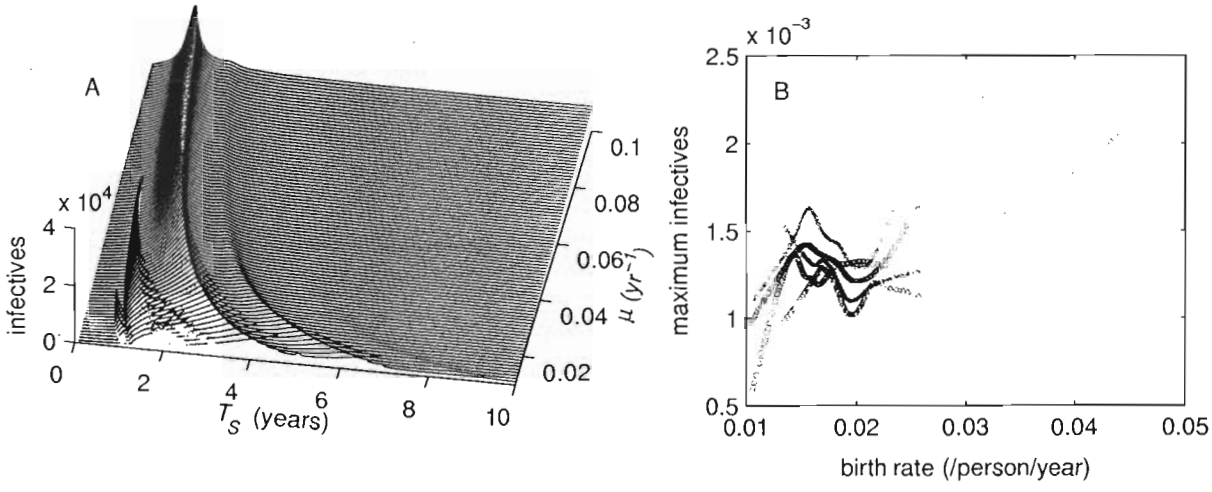

A
x 10^4
infectives
4
2
0
0
2
4
6
8
10
T_S (years)
0.1
0.08
0.06
0.04
0.02
μ (yr^{-1})
B
x 10^-3
maximum infectives
2.5
2
1.5
1
0.5
0.01
0.02
0.03
0.04
0.05
birth rate (/person/year)

Figure 3

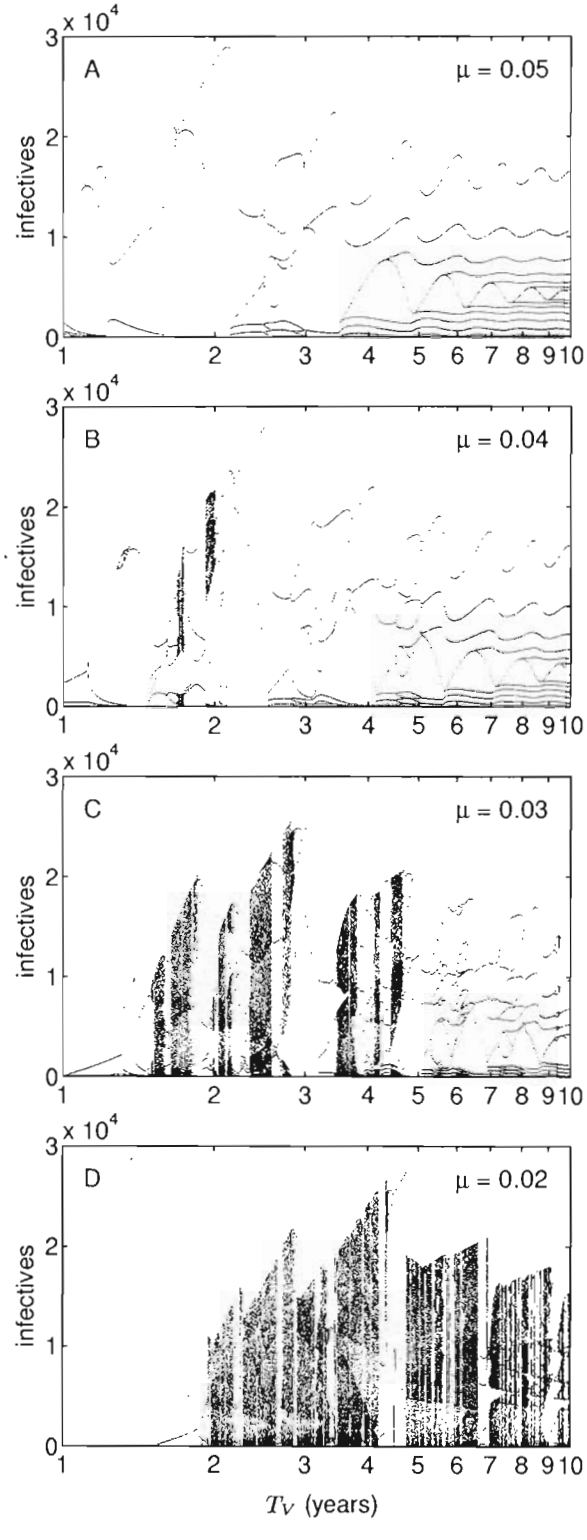
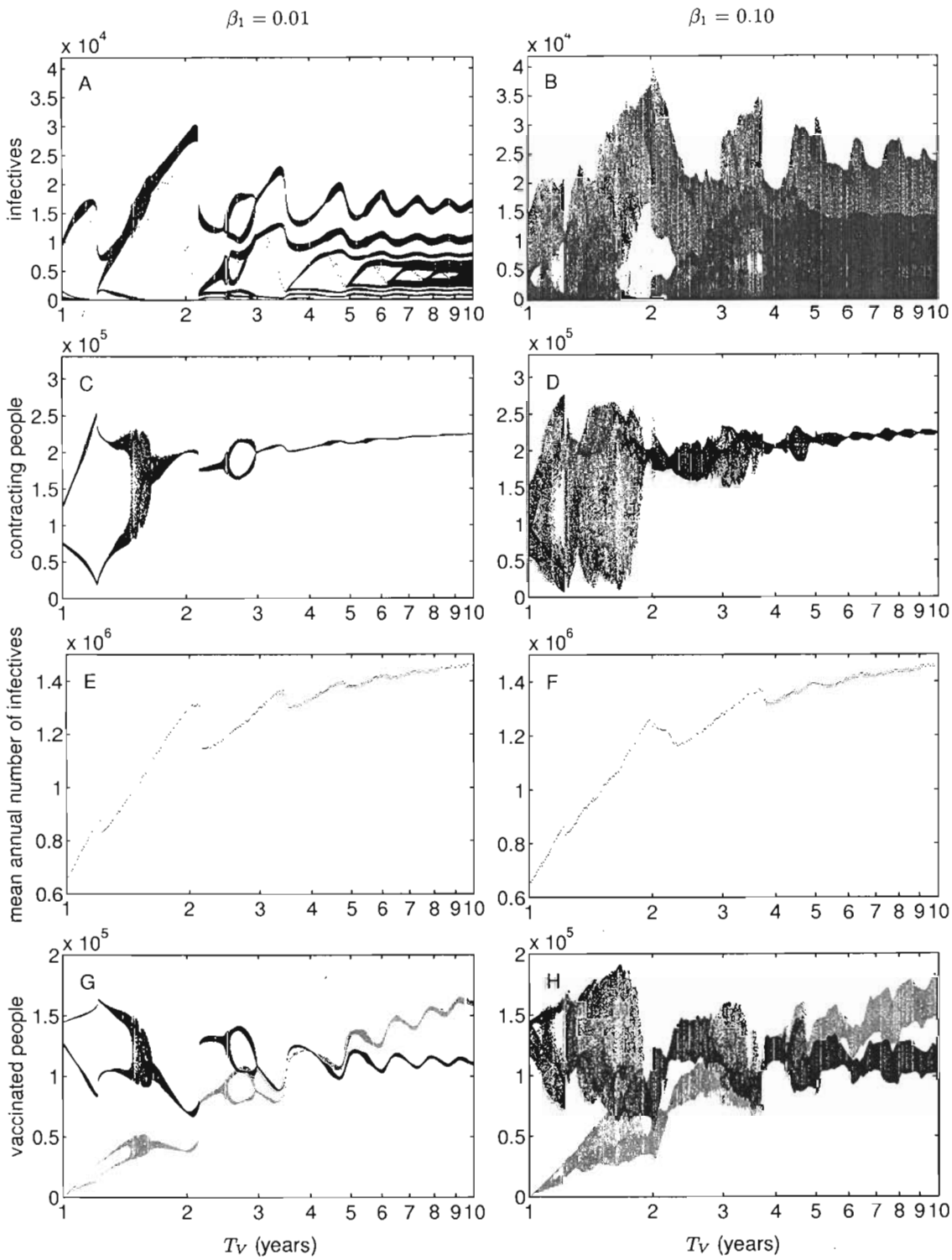

x 10^4
A
μ = 0.05
B
μ = 0.04
C
μ = 0.03
D
μ = 0.02
infectives
0
1
2
3
1 2 3 4 5 6 7 8 9 10
Tv (years)

Figure 4


$\beta_1 = 0.01$
$\beta_1 = 0.10$
A
B
C
D
E
F
G
H
infectives
contracting people
mean annual number of infectives
vaccinated people
T_V (years)
T_V (years)

Approches écologique et évolutive de la vaccination

Résumé

Une approche raisonnée de la vaccination peut s'organiser autour de deux axes majeurs de recherche. Le premier concerne la mise au point du vaccin lui-même et cherche à déterminer les cibles vaccinales optimales, c'est-à-dire les parties du pathogène contre lesquelles stimuler le système immunitaire. Ce choix est crucial puisqu'il détermine les succès vaccinaux. Le deuxième axe de recherche s'intéresse au mode d'administration du vaccin aux individus et aux populations, c'est-à-dire déterminer qui vacciner, quand vacciner, comment vacciner. A ce niveau, on s'intéresse en particulier aux conséquences des politiques vaccinales imparfaites, soit presque toutes les politiques en pratique. Dans cette thèse nous nous sommes intéressés à ces deux axes de recherche. Dans une première partie nous étudions l'évolution moléculaire chez les parasites à partir de la recherche de gènes sous sélection positive qui pourraient constituer des cibles vaccinales potentielles. Dans ce cadre de recherche, deux études sont menées, une s'intéressant à l'analyse et la comparaison de l'adaptation moléculaire sur les trois principaux gènes de lentivirus de primates. La deuxième étude analyse l'adaptation moléculaire d'une famille de gènes codant pour des enzymes fortement impliquées dans les relations hôte-parasite chez les leishmanies. Dans une deuxième partie nous étudions le comportement dynamique des maladies dans les populations. Dans cette partie, deux études sont menées. La première s'intéresse au phénomène de persistance globale et de synchronie entre dynamiques de varicelle en France. La deuxième étude explore les conséquences dynamiques d'une des deux grandes politiques vaccinales actuellement en vigueur, la vaccination par pulsations, en s'intéressant spécifiquement au phénomène de résonance.

Mot-clés : vaccination, évolution moléculaire, résonance, politique vaccinale, VIH, maladies infantiles

Ecological and evolutionary approaches of vaccination

Summary

A modern approach to vaccination implies two major axes. The first one is the vaccine design and consists in the determination of optimal vaccine targets, i.e. the parts of the pathogen against which to stimulate the immune system. This choice is crucial as it determines vaccine successes. Le second axe concerns the way populations are vaccinated. It basically consists in determining who to vaccinate, when to vaccinate, and how to vaccinate. At this scale we are particularly interested in the consequences of imperfect vaccination policies, i.e. almost all policies in practice. This thesis considers these two axes of research. In a first part we study parasite molecular evolution in seeking genes under positive selection which could constitute potential vaccine targets. In this framework, two studies are carried out. The first one is a comparison of molecular adaptation on the three major genes of primate lentiviruses. The second study analyse the molecular adaptation of a gene family coding for enzymes highly implicated in host-parasites relationships in leishmanies. In a second part we study diseases population dynamics. Here, two studies are carried out again. The first one concerns global persistence and synchrony between varicella dynamics in France. The second study is an exploration of the dynamical consequences of one of the two major vaccine policies currently applied. Specifically, we study the phenomenon of resonance associated with the pulse vaccination.

Key-words : vaccination, molecular evolution, resonance, vaccination policies, HIV, childhood diseases.