Apports futurs de la bioinformatique

E. Rivals[1]

Introduction

Avec le séquençage complet de plusieurs dizaines de génomes, la biologie a entamé l'ère dite « post-génomique ». En outre, les méthodes de mesure d'expression du transcriptome et du protéome permettent de relier l'information statique du génome à la dynamique de l'activité métabolique. Le biologiste peut donc obtenir les séquences d'ensembles complexes de gènes et des descriptions dynamiques de leur activité en grande quantité.

Si ces informations augurent d'un profond changement dans la biologie, de par leur abondance, leur diversité et la complexité des relations qu'elles décrivent, elles requièrent pour leur analyse de nouvelles méthodes informatiques. D'où, l'interaction croissante de la biologie et de l'informatique, nommée *bioinformatique*.

La bioinformatique a jusqu'à présent trouvé deux utilités principales : 1) l'organisation de connaissances biologiques dans des bases de séquences (EMBL, Genbank, etc.) ; 2) l'analyse d'une nouvelle séquence par alignement avec les séquences existantes (Blast, Fasta, etc.) Aujourd'hui, la bioinformatique peut apporter une aide plus importante à la recherche biologique :

[1] Équipe « Méthodes et algorithmes pour l'analyse de séquences », département IFA, Lirmm, 161, rue Ada, 34392 Montpellier cedex 5, France.

– dans l'analyse automatique d'un génome, c.a.d. d'un grand groupe de séquences en tant qu'ensemble (ex. de la génomique comparative) ;

– dans l'étude des relations complexes entre objets biologiques, comme les interactions au sein d'une population, les réseaux métaboliques, les cascades transcriptionnelles. L'étude formelle de systèmes complexes tels que les réseaux, les graphes, qui sont des domaines de recherche anciens en informatique peut aider à comprendre les propriétés intrinsèques de ces organisations biologiques ;

– à la vérification objective et systématique d'hypothèses biologiques exprimées par un critère informatique ou mathématique, et ce sur de grands ensembles de données.

Dans les trois prochaines sections, nous illustrons ces aspects par les exemples de la génomique comparative, de la recherche à grande échelle de sites de régulation et de la détection de répétitions en tandem.

Génomique comparative

Lorsque pour une espèce il est possible de prédire de manière fiable tous ses gènes à partir de la séquence des chromosomes, l'obtention du génome complet donne accès à l'ensemble de tous les gènes de l'espèce. Cela s'avère être le cas pour les génomes de procaryotes. À partir de maintenant, je nommerai *génome* l'ensemble de tous les gènes d'une espèce pour le distinguer de la séquence complète de tous les chromosomes.

On peut alors étudier l'ensemble des gènes en correspondance avec l'ensemble des phénotypes ou fonctions d'une espèce. En considérant la nature d'ensemble du génome, on peut appliquer des opérations de comparaisons d'ensembles : intersection, différence, union, à des paires, triplets, n-uplets de génomes. Dans le cas d'une intersection de génomes, on corrèle l'ensemble des gènes communs aux deux espèces avec les phénotypes qu'elles partagent. Ce genre d'expériences « *in silico* » peut être qualifié de génomique

comparative. Ce même type de raisonnement mathématique simple est pratiqué lorsque d'aucun tente de cerner l'ensemble minimal des gènes nécessaires à la vie. Dans ce cas, la vie est considérée comme l'intersection de ce qui caractérise toutes les espèces. Le travail de Huynen et Bork (1998) montre qu'en raffinant ce raisonnement, la génomique comparative peut aider à la prédiction fonctionnelle des gènes.

Huynen et Bork (1998) l'ont appliqué à neuf génomes d'archéobactéries et de bactéries : *Haemophilus influenzae* (Hi), *Helicobacter pylori* (Hp), *Escherichia coli* (Ec), *Mycoplasma genitalium, Mycoplasma pneumoniae, Synechocystis* sp., *Methanococcus jannaschi, Methanobacterium thermoautotrophicum, Bacillus subtilis.* Pour cela, ils ont comparé deux à deux toutes les séquences de protéines d'une espèce avec celles de chaque autre espèce. Ce type de traitement informatique nommé comparaison « tous contre tous » d'ensembles de séquences devient de plus en plus courant et requiert la conception de nouveaux algorithmes pour une exécution rapide. La génomique comparative a par là stimulé la recherche en informatique « pure », comme en témoigne le travail sur ce sujet de Burkhard *et al.* (1999).

Afin de déterminer les protéines communes entre ces espèces, les auteurs ont considéré la protéine p du génome G comme orthologue à la protéine p' du génome G' si la similarité entre p et p' est significative et s'étend sur au moins 60 % de la séquence de p ou p', et si cette similarité est maximale par rapport à celle de toute autre paire (p, q') ou (q, p') où q, q' sont des protéines de G, G' respectivement. Ensuite, ils ont pu effectuer des comparaisons entre génomes, comme les comparaisons multiples entre *H. influenzae* (Hi), *H. pylori* (Hp) et *E. coli* (Ec).

Dans la figure 1, les trois génomes sont schématisés par des cercles respectivement de couleur noire (Hi), gris foncé (Ec) et gris clair (Hp). Ces cercles se chevauchent et définissent des sous-ensembles : par exemple la portion à l'extérieur du cercle gris foncé (Ec) et à l'intérieur des cercles noir et gris clair (Hi et Hp) est noté A. Mathématiquement, $A = (Hi \cap Hp) \setminus Ec$, c.a.d. l'ensemble des gènes communs à *H. influenzae* et *H. pylori* mais qui n'ont pas d'orthologue chez *E. coli.* Huynen et Bork (1998) ont trouvé que A contenait majoritairement des facteurs d'interaction avec l'hôte,

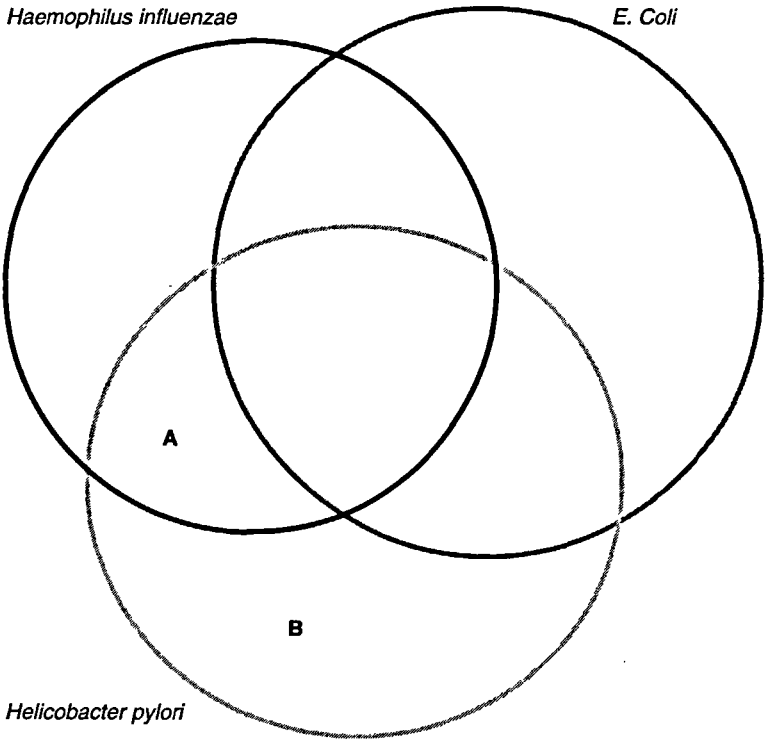

Figure 1
Vue schématique de la comparaison de trois génomes : *E. coli*, *H. pylori* et *H. influenzae*.

plus spécifiquement des facteurs de virulence, des toxines ou des protéines impliquées dans leur production et encore des protéines excrétées ou transmenbranaires.

Ceci reflète bien la pathogénicité qui est un trait commun de *H. influenzae* et *H. pylori*, mais pas de *E. coli*. Les exceptions à cette règle peuvent aussi se révéler informatives du point de vue évolutif. La seule protéine métabolique qui appartient à l'ensemble A, la dehydroquinase type II, est un exemple de déplacement non-orthologue. En effet, la dehydroquinase type I d'*E. coli* assure la même fonction mais n'est pas l'homologue de la dehydroquinase type II.

La portion B de la figure 1 délimite l'ensemble des gènes qui sont propres à *H. pylori*. Outre qu'il est aussi riche en facteurs d'interaction, il contient 23 des 27 protéines de l'ilôt de pathogénicité *cag*. Cet îlot s'est révélé depuis comme un déterminant de la pathogénicité d'*H. pylori* (certaines souches l'ont, mais pas d'autres) et semble contribuer à l'ulcération des tissus et au développement du cancer gastrique (Figura et Valassina, 1999).

Dans ces exemples, on voit comment des expériences à base d'opérations mathématiques et réalisables automatiquement peuvent fournir une prédiction de la fonction de protéines pas encore caractérisées. L'augmentation du nombre de génomes en jeu dans une comparaison ou le croisement avec des critères de synténie permettent de raffiner les prédictions. Enfin, notons que ces expériences ne sont possibles qu'à condition de disposer de l'ensemble exhaustif des gènes ou protéines des espèces considérées.

Analyse des régions amonts de gènes d'expression similaires

Les filtres et puces d'expression, ainsi que d'autres techniques, permettent de mesurer des niveaux d'expression absolus ou relatifs d'un grand ensemble de gènes (Jordan 1998). À partir des résultats bruts, des méthodes de classifications groupent les gènes ayant le même « patron » d'expression (par exemple, même courbe de variation pendant une série de points espacés dans le temps).

Si l'ensemble des gènes inclus dans l'expérience est petit, les auteurs peuvent analyser ou interpréter les « classes » ainsi obtenues à la main. Mais si au contraire, les gènes testés se comptent en milliers, si les expériences de mesure d'expression peuvent être qualifiées d'expériences « à haut débit » ou high-throughput en anglais, alors que fait-on de cette masse de données ?

Qui dit classe de gènes co-régulés suggère de chercher dans les séquences en amont des motifs de régulations. Une possibilité est de chercher les motifs connus de liaison à l'ADN des facteurs de trans-

cription. Ainsi ne retrouve-t-on que des sites d'interactions connus si tant est que les sites soient conservés. Une autre approche est de trouver des nouveaux sites, c.a.d. des portions de séquences conservées dans les séquences d'une même classe sans pour autant exiger que celles-ci soient alignables. C'est l'objet d'une classe d'algorithmes dits *de découverte ou d'extraction de motifs*. Il existe plusieurs définitions pour des motifs, mais la plus courante est celle d'une expression régulière simple (en termes informatiques). Par exemple : GAYRNNC représente toute séquence commençant par GA, suivis d'une pyrimidine (Y), d'une purine (R), de deux résidus quelconques puis d'un C (selon le code IUPAC). C'est l'équivalent des motifs PROSITE chez les protéines. N est ici un joker ou « wildcard ».

Le problème auquel s'attaquent ces algorithmes est de découvrir un motif inconnu conservé dans plusieurs séquences. Il ne faut pas les confondre avec les algorithmes de recherche de motifs (logiciels Patsearch, Patscan, Findpattern, etc.). Ils reçoivent en entrée l'ensemble de séquences non alignées et des paramètres qui limitent leur recherche : taille maximale du motif, nombre maximal de jokers, etc. Le problème est combinatoirement difficile car non seulement le motif est inconnu, mais le nombre de séquences dans lequel il doit apparaître n'est pas donné. Un motif peut être présent seulement dans un sous-ensemble des séquences en entrée. Dans le cas d'un groupe de gènes co-régulés ceci est capital car ils ne sont pas tous nécessairement directement contactés par un même facteur de transcription et sur un même site. Il se peut que certains soient indirectement régulés, voire des faux positifs dus au bruit dans les mesures d'expression.

Brazma *et al.* (1998) ont conçu et appliqué un tel algorithme aux séquences amonts de gènes co-régulés qui furent détectés d'après les expériences de DeRisi *et al.* (1997) sur le transcriptome de levure. DeRisi *et al.* (1997) ont mesuré les variations cinétiques d'expression chez la levure pendant le passage métabolique de la fermentation à la respiration. La série comprend 7 mesures espacées toute les 2 heures pendant la période de manque en glucose (stress). Brazma *et al.* (1998) ont ensuite effectué une classification à partir des courbes cinétiques codées en binaire. Celle-ci fournit 32 classes dont la variation d'expression était significative. Pour chaque classe, ils ont extrait 7 jeux de séquences amonts : de -100 à 0, de -150 à -50, de -200 à -100, etc. En outre, ils sélectionnèrent 2 jeux de séquences de même longueur aléatoirement dans le

génome complet et qui servent à établir une « significativité » des motifs trouvés dans les séquences amonts. Grâce à l'algorithme de Vilo (1998), ils recherchent des motifs tels que ceux évoqués ci-dessus avec au plus un joker. La significativité d'un motif est évaluée par le ratio du nombres d'occurrences dans les séquences de la classe sur le nombre d'occurrences dans les séquences sélectionnées aléatoirement dans le génome (avec un terme correcteur).

Pour valider leurs résultats, tous les 50 motifs les plus significatifs furent comparés aux sites de facteurs de transcription de la base Transfac (Wingender *et al.*, 2000). De nombreux sites furent retrouvés exactement ou approximativement. Par exemple, le motif CCCCT est classé dans les dix premiers pour 4 classes de gènes dont l'expression est croissante. Or on sait que ce motif est lié à la réponse au stress. Pour les 2 classes de gènes dont l'expression décroît, ils trouvent le site de fixation du facteur RAP1 qui contrôle finement la transcription des protéines ribosomales. Certains motifs ne figurent pas dans Transfac, et leur présence parmi les 50 meilleurs motifs constitue un argument en faveur de leur rôle de site de fixation ou de régulation.

Ce genre d'expérimentation est automatique, indépendante de la technologie de mesure d'expression et du type de phénomène biologique étudié. Les résultats des prédictions sont bons malgré les problèmes de bruit dans les mesures. L'étude ne porte que sur les classes dont la variation d'expression est la plus significative, mais prouve que ces mesures sont exploitables. Elle illustre le traitement de données à grande échelle et « en aveugle », c.a.d. sans injonction de connaissances biologiques particulières. Là aussi, l'algorithmique de la découverte de motifs est l'objet de recherches en informatique.

Intérêts des approches mathématiques ou informatiques formelles

Les approches formelles mathématiques ou informatiques donnent une autre vision des objets ou systèmes biologiques. Ce n'est pas

chose nouvelle, mais plutôt importante, car ces approches permettent d'apporter du sens.

Par exemple, les théories de l'information qui forment un champ d'investigation entre les mathématiques et l'informatique s'intéressent à définir l'information et à la mesurer. Que l'information soit celle qui circule sur internet ou celle qui permet la signalisation entre cellules biologiques, elle n'en reste pas moins intrinsèquement de l'information. Ainsi on a vu des concepts de cette discipline s'appliquer avec succès en biologie, par exemple pour construire des consensus de séquences (Mount *et al.* 1992). L'utilisation de critères mathématiques objectivise les procédures en biologie (Rivals *et al.* 1996).

Un autre apport de l'informatique provient de ce qu'elle s'occupe d'étudier les caractéristiques intrinsèques d'objets ou de structures abstraites : les graphes, les réseaux, les automates, etc. Ce type de structures existe en biologie : les réseaux métaboliques ou de régulation ont des structures sous-jacentes de graphe (au sens formel d'un ensemble de noeuds reliés par des arcs). L'application de ces théories éclaire des données biologiques, comme la fréquence de colonisation de l'intestin humain par *E. coli* (Savageau 1998).

Les séquences, les suites finies ou infinies de lettres prises dans un alphabet donné, constituent une autre classe d'objets fort étudiés en mathématiques et informatique. Les algorithmes d'assemblage de séquences, de recherche de motifs ou de comparaison de séquences (Blast, Fasta, Smith et Waterman) tirent pleinement parti des propriétés combinatoires des séquences. Même des problèmes qui peuvent paraître simples en biologie peuvent bénéficier de l'éclairage d'approches formelles, par exemple la localisation de répétitions en tandem.

Une répétition en tandem est l'occurrence d'un motif, par ex. CAG, plusieurs fois de suite côte à côte dans une séquence, les copies n'étant pas systématiquement identiques. Localiser à l'oeil aussi bien qu'automatiquement ce type de régularités est en réalité un problème difficile. La pierre d'achoppement réside dans le taux de conservation admis entre les copies en fonction de la longueur de la répétition et étant donné la petite taille de l'alphabet. Cette dernière pose le problème de la significativité. Plus la séquence est longue, plus il est probable de trouver une occurrence d'une courte répéti-

tion en tandem, par ex. CACACACA. Dès lors que signifie la présence de ce motif dans une séquence d'ADN ? Seule la comparaison avec des séquences aléatoires peut donner une réponse objective à cette question. Le taux de conservation s'il est faible a pour effet de « diminuer » la structure répétitive de la séquence, d'en rendre floue les limites, voire le motif.

CAGCAGCAGCAGCAGCAGCAGCAG
CAGCAGCTGCAGCAGCTGCAGCAGCTG
CAGTCGCTGCAGCTGCGACAGCTGATT

Les trois séquences ci-dessus illustrent ces remarques : la première est exactement 9 fois le motif CAG, mais quel est le motif de la seconde : CAG ou CAGCAGCTG ? Enfin comment déterminer si la troisième est encore une répétition en tandem ou une séquence sans structure répétitive ? Les définitions biologiques de micro- ou mini-satellites ne permettent pas d'élucider ces problèmes.

Les approches formelles utilisant des critères combinatoires (toute suite d'au moins 5 copies d'un motif de longueur comprise entre 3 et 10 pbs avec au plus 3 substitutions, cf. Sagot et Myers, 1998), statistiques (Benson, 1999) ou informationnel (Rivals *et al.*, 1996, Delgrange *et al.*, 1999) ont proposé des définitions objectives de répétition en tandem et permis la conception d'algorithmes capables de localiser exactement toutes les portions de séquences vérifiant ces définitions. L'utilisation de ces algorithmes permet de caractériser précisément les répétitions trouvées et les portions de séquences considérées par l'algorithme comme n'étant pas répétées en tandem. Cette qualité ainsi que l'exactitude sont des avantages cruciaux pour les études biologiques de ce genre de séquences.

Conclusion

Nous espérons que les exemples présentés convaincront le lecteur de l'intérêt du champ interdisciplinaire qu'est la bioinformatique, et montrent que ses apports sont aussi pratiques que fondamentaux. L'accès à de nouveaux niveaux de structures biologiques (génome,

transcriptome, protéome) augure d'une nouvelle ère d'interaction entre les deux disciplines où l'application innovante de concepts formels aidera à la compréhension de ces structures.

À cette présentation prospective, nous voulons apporter quelques précisions. L'application pertinente des méthodes par un biologiste et l'interprétation correcte des résultats passent par une connaissance réelle des caractéristiques de ces méthodes : contexte d'application, paramètres, signification des résultats, complexité. Cela requiert une curiosité, un dialogue et une éducation interdisciplinaires.

Les démarches exposées sont conduites par le raisonnement biologique. Apport bioinformatique ne signifie nullement inutilité de l'expertise biologique. Au contraire, sans elle, pas d'interprétation possible.

En résumé, l'interaction interdisciplinaire apporte à l'informatique de nouvelles problématiques de recherche, et à la biologie la possibilité d'expérimentations « *in silico* » qui permettent d'étudier d'une manière innovatrice des objets complexes et leurs interactions.

Bibliographie

Benson G 1999 — Tandem repeats finder: a program to analyze dna sequences. Nucleic Acids Res 27 (2) : 573-80.

Brazma A, Jonassen I, Vilo J, Ukkonen E 1998 — Predicting gene regulatory elements *in silico* on a genomic scale. Genome Res 8 (11) : 1202-15.

Burkhardt S, Crauser A, Ferragina P, Lenhof HP, Rivals E, Vingron M 1999 — q-gram Based Database Searching Using a Suffix Array (Quasar). In Third Annual International Conference on Computational Molecular Biology (Recomb99), Lyon, France, 11-14 April.

Delgrange O, Dauchet M, Rivals E 1999 — Location of Repetitive Regions in Sequences By Optimizing A Compression Method. In Proc. of the 4th Pacific Symposium on Biocomputing, Hawai, 4-9 Jan.

De Risi J.L., Iyer V.R., Brown P.O., 1997 – Exploring the metabolic and genetic control of gene expression on a genomic scale. Science, 278 (5338) : 680-686.

Figura N, Valassina M 1999 — Helicobacter pylori determinants of pathogenicity. Journal of Chemotherapy 11 (6) : 591-600.

Huynen MA, Bork P 1998 — Measuring genome evolution. Proc. Natl. Acad. Sci. USA 95 : 5849-5856.

Jordan B 1998 — Voyage au pays des puces. Médecines et Sciences 14 : 1097-1102

Mount SM, Burks C, Hertz G, Stormo GD, White O, Fields C 1992 — Splicing signals in drosophila: intron size, information content, and consensus sequences. Nucleic Acids Res 20 (16) : 4255-62.

Rivals E, Dauchet M, Delahaye JP, Delgrange O 1996 — Compression and genetic sequences analysis. Biochimie 78 (4) : 315-322.

Sagot MF, Myers EW 1998 — Identifying satellites and periodic repetitions in biological sequences. J Comput Biol 5 (3) : 539-53.

Savageau MA 1998 — Demand theory of gene regulation. ii. quantitative application to the lactose and maltose operons of *Escherichia coli.* Genetics 149 (4) : 1677-91.

Vilo J 1998 — Discovering frequent patterns from strings. Technical report, Dpt. of Computer Science, Univ. of Helsinki.

Wingender E, Chen X, Hehl R, Karas H, Liebich I, Matys V, Meinhardt M, T Pruss, Reuter I, Schacherer F 2000 — Transfac: an integrated system for gene expression regulation. Nucleic Acids Res 28 (1) : 316-9.