BANQUES DE SEQUENCES POUR LA TAXINOMIE DES PHYTOVIRUS : L'ANALYSE FACTORIELLE DES CORRESPONDANCES APPLIQUEE A LA CLASSIFICATION DES GEMINIVIRUS

D. DESBOIS*, C. FAUQUET**, D. FARGETTE***, G. VIDAL*

Mots-clés : Banques de données moléculaires, Analyse statistique multivariée (Analyse factorielle des correspondances, Classification ascendante hiérarchique), Application à la phytovirologie (Taxinomie du groupe des géminivirus), Services télématiques, Pays en développement.

Résumé : Cet article présente un exemple d'utilisation des banques de données moléculaires, comme source d'information, et de l'analyse factorielle des correspondances, comme technique d'analyse, pour la recherche taxinomique, ceci dans un pays en développement. L'application de cette technique statistique multivariée à la classification des virus de plantes permet de dégager des critères taxinomiques cohérents basés sur la composition en acides aminés, dinucléotides et codons de la protéine capsidaire. Ces critères aboutissent à une typologie pertinente du groupe des géminivirus, issue d'un algorithme ascendant de classification hiérarchique.

I) INTRODUCTION

Base d'expression du génome, les séquences d'acides nucléiques et de protéines font l'objet d'un enregistrement systématique dans le but de faciliter l'identification de nouvelles séquences ou de pouvoir élaborer et tester certaines hypothèses sur l'organisation, la fonction et l'évolution de ces molécules. Cet effort catalographique pour répertorier l'ensemble des séquences moléculaires a débuté avec l'enregistrement des séquences d'acides aminés au début des années 60, le *Dayhoff Atlas* (Dayhoff 1972) constituant l'exemple le plus connu de répertoire moléculaire. La mise au point (Sanger & Coulson 1975) de méthodes rapides de séquençage de l'ADN (approches *shotgun, M13,* ...) a provoqué une explosion du nombre des publications. Pour les banques de données moléculaires, la publication de nouvelles séquences d'acides nucléiques a constitué un facteur direct de croissance auquel est venu s'ajouter la prolifération induite des séquences d'acides aminés obtenues par dérivation des séquences nucléotidiques codant pour des protéines. La figure [1], concernant l'évolution au cours de la décennie 80 du nombre de bases contenues dans la banque du Laboratoire européen de biologie moléculaire (*European Molecular Biology Laboratory* - EMBL), indique que cette croissance possède manifestement un caractère exponentiel. Corpus de données spécifiques à la biologie moléculaire, ces banques de séquences ouvrent des perspectives intéressantes dans le domaine de la taxinomie numérique par l'originalité des approches méthodologiques qu'elles suscitent et l'étendue des domaines à prospecter. Les principales banques de séquences sont situées en Europe (EMBL), au Japon (DNA-JPDB) et aux USA (GenBank, NBRF-PIR). Schématiquement, on classe les banques de séquences selon leur contenu (nucléique ou protéique).

II) LES BANQUES DE SEQUENCES

1) Banques nucléiques

1.1) EMBL

La Bibliothèque de séquences nucléotidiques (*Nucleotide Sequence Library* - NSL) de l'EMBL est située à Heidelberg en Allemagne; fédérant l'ensemble des efforts européens, cette banque fut constituée en octobre 1980 pour rassembler une collection fiable et exhaustive des séquences d'acide nucléique; sa création visait à favoriser l'émergence d'une norme susceptible de promouvoir l'échange des séquences au sein de la communauté des chercheurs européens en biologie moléculaire; la première version disponible de cette banque fut

* Cellule Analyse des Données, Centre Universitaire de Traitement de l'Information, Université Nationale de Côte d'Ivoire, BP V34, ABIDJAN 01, RCI, e.mail : *DESBOIS @ CIEARN.BITNET, CADCUTI @ FRMOP11.BITNET*

** International Cassava-Trans Project, Washington University of St Louis, Department of Biology, CB 1137, One Brookings Drive, St Louis, MISSOURI 63130, USA, e.mail : *FAUQUET @ BIOLGY.WUSTL*

*** Virology Division, Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA, G-B
e.mail : *FARGETTE%EDINBURGH @ RL.AC.UK*

Figure 1 : courbe de croissance de la banque de séquences européenne EMBL

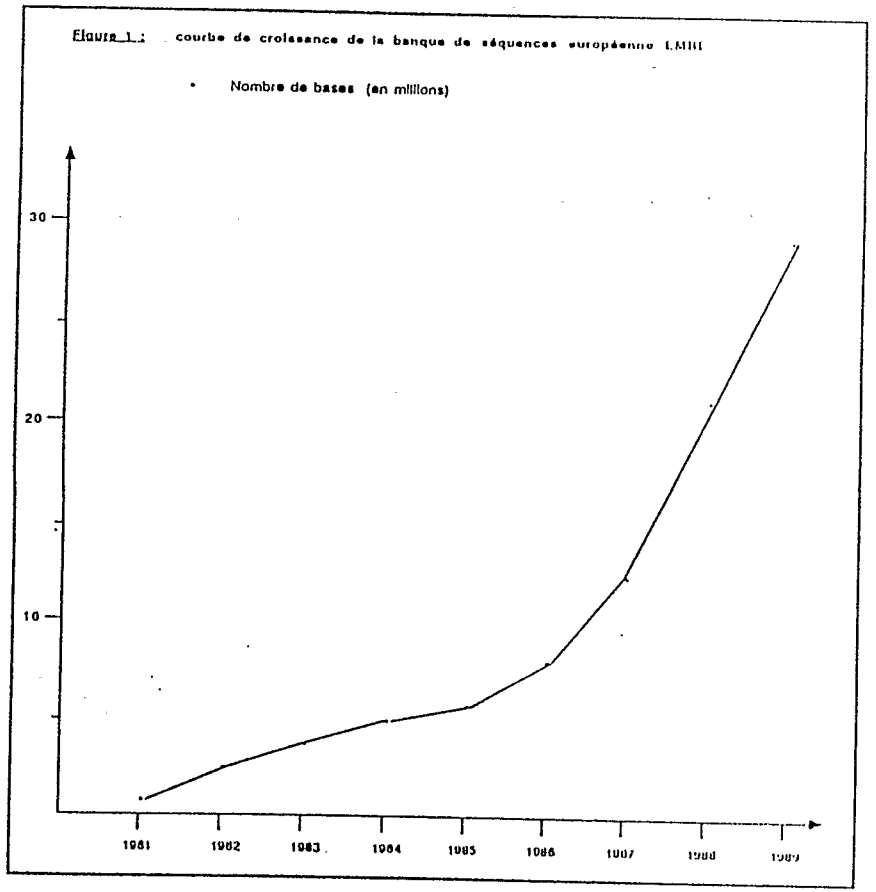
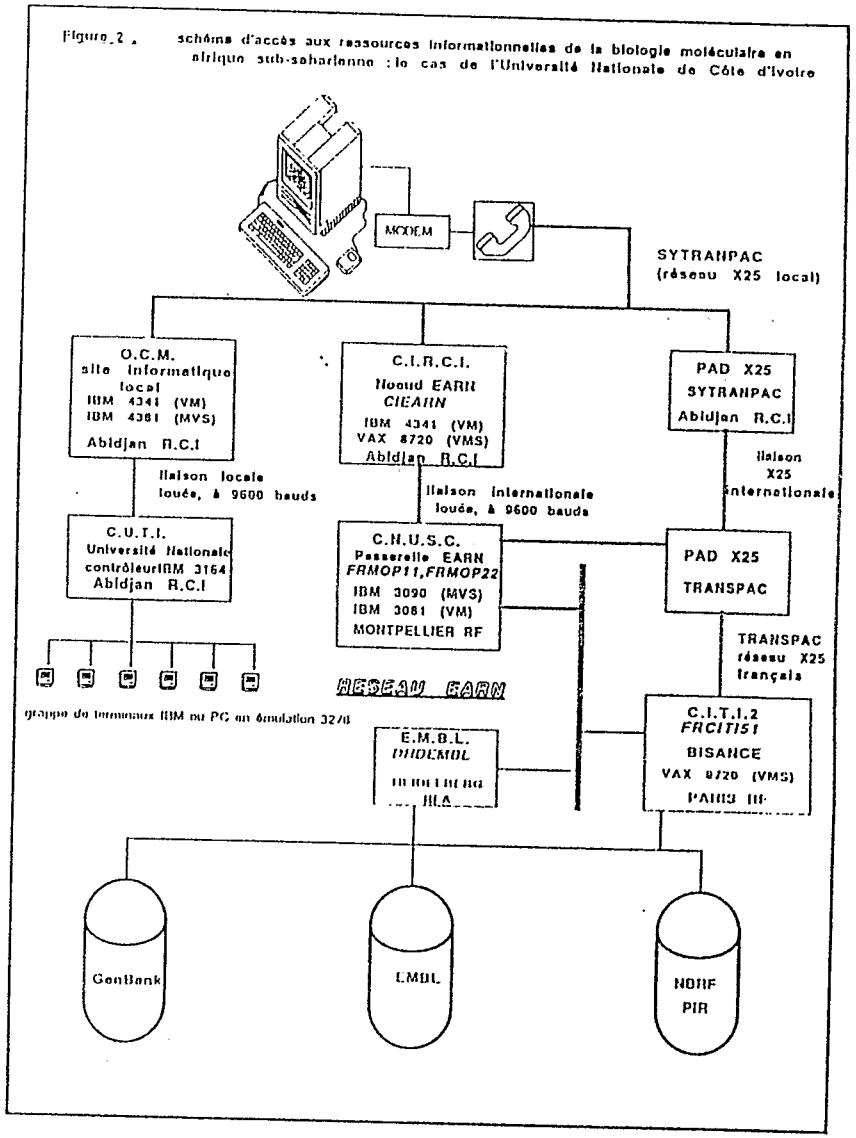


Figure 2 : schéma d'accès aux ressources informationnelles de la biologie moléculaire en afrique sub-saharienne : le cas de l'Université Nationale de Côte d'Ivoire



Encadré 1 : une référence de la base de données EMBL sur les géminivirus.

```
ID   BEMSV5IE       preliminary;  DNA;   2690 BP
XX
AC   V00314;
XX
DT   02-JAN-1985   (Annotation)
XX
DE   Maize streak virus (MSV-B) infectious genome
XX
KW   circular, capsid protein; unidentified reading frame;
KW   genome
OS
OS   Maize streak virus (MSV)
OC   Viridae, ss-DNA nonenveloped viruses; Geminiviridae.
XX
CC   *source, strain=South African isolate; library=pEMBL
CC   ORFs for 10.9 kd and 28.8 kd proteins correspond to those reported
CC   for MSV-M and MSV-K both by sequence and position homology;
CC   overlapping arrangement of ORFs for 31.3 kd and 17.2 kd proteins
CC   is identical to that reported for MSV-M. ORF for put. 8.7 kd
CC   protein remains unsure.
XX
XX
RN   [1]    (bases 1-2690)
RA   Lazarowitz S.G.;
RT   "Infectivity and complete nucleotide sequence of the genome of a
RT   South African isolate of maize streak virus.
RL   Nucl. Acids Res. 16: 229-246(1987).
XX
FH   Key             From   To          Description
FH
FT   TRANSCRIPT      (1113          longer "c"-strand transcript (circular
FT                                     genome I)
FT   CDS             <1     301         10.9 kd protein (circular genome I)
FT                                     (1 is 3rd base in codon)
FT   CDS             315    1046        28.8 kd virion capsid protein
FT   TRANSCRIPT      >2690          longer "c"-strand transcript (circular
FT                                     genome I)
FT   CDS             2889   >2690       10.9 kd protein (circular genome I)
FT                                     (2890 is 2nd base in codon)
FT   CDS             2377   1362 (C)    31.3 kd protein
FT   CDS             1684   1331 (C)    17.2 kd protein
FT   CDS             503     331 (C)    17.2 kd protein
XX
SQ   Sequence  2690 BP;      693 A; 836 C; 688 G; 6731;

GGATCCACAGAACGCCCTGTATTATCAACCGCGGCTACCCACTGCAGCTCCGACATCCGG
AGGAGTACCGTGGAGTCGCGTAGCGAAGGTAGCTATTTGGTGAGCTTGTTGTTGCATTGATTG
CTTTACCTGCTTTACCTTTGGGTGCTGAGAGCCTTATCTTATTCTGAAGCTCGACA
AGGCAGATCCACACGGAGGAGCTGATATTTGGTGAACAAGCTGTGATAGAGCACAACCTAT
CCCTAATCTACCAGCACCACCAAATCAAGGCAATCCGGGCCATTTGTCCAGGCACAGA
ATAAGCAATCAACCATGTCCACGTCCAAGGAAGAACGGGAAATGTGCAATTGAGATTGAGATA
AACGGGTGCCTAAGAAGAAGCCTTCTTCAGCTGGCTAAAGGGCTGAAGCAAGCCG
ATAGGCCATCCCTCCAAATCCAGAACACTCCAGCATGCTGGAACCACCATGATAACTGTCC
CATCCGAAGAATATGTGACCTCATCAACACCTATGCCGAAGATCTGAGAGAGAGCAAGG
GCCAGCACCAGCAAGAGCTCTAAGATGCAGAAGTCCGCTCAACTACGACTCTGTTGCCAAGA
CGGCTGCCTGCCACCACCAACCGAAGCCGATGTAATGTGGCTGGTGTATGACAGCCA
CTCCCGCGACAAGCTCCGACCCCGCAAACTATATTGCCIACCCTGACACGCTAAAAG
CGTGGCCGGCCACATGGAAAGTGAGCCGGAGCTGTATCATCGCTTCGTGGTGAAACGGC
AATGGTGTCAACATGAAGACCAGCGGTCAAATTGGTICGAATACCCTCCTCAAATA
CAAGTTGAAGCCTTGCAAGCAGCAACATCTACTTCCAGCAAGTACTCCAGTGGTAAGCGT
TGAAACGCAGAGTGCAAAAATGTAACGAGCGGAGAGTTGGTGCCATCCAAGAAGAACTT
TGTACATGGTCATTGCCCCCGCAATGGCCTTACATTTACTGCCCATGGGCAACCCCGCG
TGTACTTTAAAGAGTGTTGGCAACCAGTAATGAATAAAAACGCCCGTTTTATTATATCTGA
TGAATGCTGAAAGCTTACATTATATGTGCTGCGTGAAGAAGCACAGCAGATCA
ATACAGGGGGGGTTGCTGGCGGCGGCTAGGTGCTCGGCGGGCAAACATCAGAAA
ATCAAGTATCTATATGAATTCAGCTCCTCCGTAGGAGAAAGCACAGGGGGAGAATACCAC
TICTCCCCCGGCGACATGATGTAAATGATGCAGTTTGCCTCGAAATACTCCAGCTGCCCT
GGAGTCATTTCCTTCATCCAATCTTCATCCGAGTTGGCAGGATTATTGTAGCTTAGAC
TTCTTCTGCAGCTTTTTCTTCTACGATACTTGGGGTTACAATGAAATCCCTCTGAGAG
CCAACTAACTGTTTCCAACGAGGGCAGAATTTAAACGGAATATCATCTACAGATGTTGAG
ATTGCATCTTCGTTGTATGAAGCCAATCAACATTATTTTGCCAGTAATTATGAACCCGT
AGGCTTCTGGCCCAAGTAGTTTCCGGTTCTTGTTGGCCGACGTGTAAGAGCTCTGC
TTTCTTGATCTTTCATCTGATGACTGGATACAGAATCCATCCATTGGAGGTCAGAGATTG
CATCCTCGAAGGTATAACAGGTAGGTTGAAGGAGCATGTAAGCTTCGGGACTAACTGGA
AGATGTAGGCTGGAGCCAATCATTGATTGACTCATTACAAAGTAAATCAGGTGAGGAG
ATGGATGAGAATTGGTGAACTCTTCCTGAATCTCAGGAAAAAGCTATTTGCAGAGTATT
CAAAATACTGCAATTTGTGGACCAACAAGGAAGCTCTTTCTGATCATGAGAGAGT
ACTCTGCTTTGAAATGTAGCGTGTGAAATAATGTCTCGCATTATTTCATCTTTGAAGGCT
TTTTTCGTTTAGCTCTGAATCAGATTTCCTAGGAAGGGGACTCCTAGGAAGAAG
TACCTCTCTCAAACACAGCCAGGTCCTGAAAGTGTAATCCCTCACTCTGTTAACTG
ACTTGACACTCTGAATATTTGGATGAAACCATTATATCAAGGACCTTGAGTCAGATA
TCCTTACCGCCTCTGTTGTGAGCAATGCATGTAAATGCAAACTTCCATCTTTATGTA
CCTCTCGGGCACATAGAATATACTTGGAATCCAACGAACGACGAGCTCCCAGATCATCT
GACAAGCGATTTCAGGATTTTCTGAACACTTTGAATAGTTAGGAACGTGTTAACGTTCC
TGTGTGAGAACTGACGGTTGAATGAGGAGGCCCATAGCAGCAGCGGGGCTGAGGGT
GAGGATGGCAAGCTGGGAGCTCCAAACTCATAATACCGTGCGCCTGAAATCCA
CCACTCCCTGTCTATAGTGGTTGAAATGGCCGAACCGTCCGCCCAGCAGCAAGAAA
GAAGCGCGCACTAATATTACCGCAGCCTTCTTTCCTGCGAGGGCCCGGTAGGGACCGAG
CGCTTGATTAAAGCCIGGTCTACTTTGTATGATTATCTAAAGCAGCCCAATGTAAA
GAAGCGGTCCCGGGCACTATAAATTGCCTAACAAGTGCGATTCATTCAT
```

Encadré 2 : une référence de la base de données GenBank sur les géminivirus.

```
LOCUS       NEIUNA          2588 bp     DNA                 entered    11/20/86
DEFINITION  Tenuis golden mosaic virus  component A of complete genome
ACCESSION   X02029
KEYWORDS    coat protein, complete genome, unidentified reading frame
SOURCE
  ORGANISM  Tenuis golden mosaic virus
            Viridae  ss-DNA nonenveloped viruses, Geminiviridae
REFERENCE   1 (bases 1 to 2588)
  AUTHORS   Hamilton,W.D.O., Stein,V.E., Coutts,R.H.A. and Buck,K.W.
  TITLE     Complete nucleotide sequence of the infectious cloned DNA
            components of tenuis golden mosaic virus   Potential coding
            regions and regulatory sequences
  JOURNAL   EMBO J. 3, 2197-2205 (1984)
COMMENT     History   (Converted from EMBL format to GenBank format)
            20-NOV-1986 (Incorporated)
            Geminiviruses are characterised by twin icosahedral virions,major
            capsid polypeptides of about 29 kd and ss-DNA genomes.The genomes
            of cassava latent virus (CLV) and tenuis golden mosaic virus
            (TGMV) consist of two circular components,while that of maize
            streak virus (MSV) consists of a single circle doesn't appear to
            be similar to either of these [1].
            [1] identifies the following additional open reading frames on
            the complementary strand that would code for proteins with >10kd
                AC1 - 13-1543 (crossing through origin) AC2 - 1801-1212 AC3 -
            1443-1067. The sequence at 1-233 is highly homologous to an
            equivalent region on component B, it doesn't appear to code for
            protein and has the potential to form a stable hairpin [1]. An
            analogous region is found in CSV. The virion-sense (+) is shown
            below.
FEATURES             from    to     description
                     327     1070   coat protein (AR1)
BASE COUNT          672 A       513 C       603 G       700 T
SEQUENCE
     1 GATACGATGG CATTTTTGTA ATTAAGAGGC TTACTACCAA TTGAGGAGGA GCTCCAAAAG
    61 TTATATGAAT TGGTAGTAAG GTAGCTCTTA TATATTAGAA GTTCCTAAAG GGCACATGGC
   121 GCCCATCCGT TTAATATTAC CGGATGGCCG CGCGATCGTC AGCCGACCCG CTTCCGCAAA
   181 ITACGCCGCA TTGTCGTCTA AGTGGTCCCG CATATGTGAA GGCCCAATCA TATTTGGCCC
   241 TGAAATCTAA GATATTTTTA AAGACTTGTG GTTAAGTTGT TAAABTTATA TAAAACGACA
   301 TGCBATTCGT GDATCTTTAA TTCAAAATGC CTAAGCGGGA TGCCCCATGG CGTTTAATGG
   361 CAGGGACCTC AAAGGATTCC CGCTCTGCTA ATTATTCTCC TCGAAGGAAT TTACCTAAGC
   421 GTGATGCTTG GGTTAACGGG CCCATGTACA GGAAACCCAG GATATATCAA TCACTAAAGG
   481 GCCCCGATGT TCCTAAAGGA TGTGAAGGGC CTTGTAAAGT CCAGTCATAC GAGCAGCGTC
   541 ATGATATTTC CCTAGTTTGG AAGGTCATGT GTATATGTBA TGTGACACGT GGTAACGGTA
   601 TTACCCACCG TGTTGGTAAB CGTTTCTGCB TTAAGTCTGT ATATATCTTB GGCAAGATAT
   661 GAATGGATGA GAACACATCAAB TTGAAGAATC ACACGAACAB TGTCATGTTC TGGTTGGTTA
   721 GGGATCGGAG ACCTTATGAC ACTTCTGAGA AGTGTCAAAC ATGTTCGATA
   781 ATGAGCCAAG IACTGCAACB GTAAAGAACB ACGTACGGGA TGATTCCAAA ATGATCCACA
   841 AGTTTCACGC CAAGGTTACT GGTGGTCAAT ATGCCAGCAA CGAGCAGGCT CTGGTTAGGA
   901 GATTCTGGAA GGTCAATAAC AATGTCGTCT ACAACCACCA GGAGGCAGGA AAATATGAGA
   961 ATCATACTGA GAACGCCCTB TTATTGTATA TGGCATGTAC TCATGCCTCT AACCCTGTAT
  1021 ATGCGCAGTT GAAAATTCGA ATCTATTTTT ATGATTCGAT AACAAATTAA TAAAATTTAT
  1081 ATTTTATTGA ATGATTTTCB AGTACATGCB TTATATATBA TCTGTCTGTT GCGAAACGAA
  1141 CAGCTCTAAT AACATTGTTA ATACATATAA CGCCTAACTB TTCAAGGTAC AACATCACTA
  1201 AGTATTAAAA TCTATTTAAA TAAGTTCTCC CAGAAGCTGT CGTCGATGTC GTCCATAGTC
  1261 GGAAGTGGAB AAATGCCTTB TGGAGAICCA ATGCTCTCCT CAAGTTGTGG TTGAACCTAA
  1321 CTTGTAAATG GTATATCCTG GTGTTGGTGT AGAGGGAGTC CTCTACGCTA AGTATCTTGA
  1381 AATAAAGGGB ATTTGTTATC TCCCAGATAT AGACBCCATT CTCTGCTTBA GGCACAGTGA
  1441 TAGGTTCCCC IGTGCGTGAA TCCATTGTTT CTBCAGTCGA TGTGAATGTA TATGGAACAG
  1501 CCACAGTTCA GGTCAATTCG TCGCCTTCTA ATAGCTCTTC GTTTAGCTBC TCTGTGTTGA
  1561 ACTTTAATAG AGGGGGAAGT GAATTTCGCA TTATGGAAAG TCCAGGTCTT
  1621 TAGTGAAATG TTTICLTCTT TGTCAAGAAA AACTTTATAG CTAACACCCT CTCCAAAATT
  1681 GLACAGAGCA ATTGACTGAA TACCTCCTTT AATTTGAACT GGCTTTCCAT AGTTACAATT
  1741 AGTCTGLCAA TCTCTTTGAB CCCCAATGAG TTCTTTCCAA TGTTTCAACT TTAGATATTG
  1801 CGGTGTGACA TCATCGATAA CGTTATACTC AACCTTGTTT GAGTAAACCC TAGAATTGAG
  1861 ATCCAAATGC CCGCTCAAAT AATTATGTGB GCCTAGTGAA GAAGCCCACA TAATCTTTCC
  1921 CGTCCGACTA TCGCCCTCAA TGATAATACT AATAGGTCTC TCCGGCCACG CAACGGAACT
  1981 CTTTCCAAAA TAATTTTCAG CGCATTGTCT CATCTCGTCT GGCACGTTAG TAATGATGA
  2041 GACGTGGAAC GGAGGAAGCC ATGDTTCAGG AGTCTTATCA AATAICCTAT CTAAATTGCT
  2101 ATTTAGGATG TGGAACTGAA ATAAATATTT TTCTGGGATT TTCTCTCTAA TTATCTGCAG
  2161 GGCTICTTCT TTGGAAGAAG CATTTAACGC CTCTGCTGCA GCGTCGTTAG ATGTTTGGCA
  2221 ACCICCTCTA GCACTTCGAC CGTCAACGTG GAATTCTCCC CATACAAGAA TATCTCCGTC
  2281 TTTGTCGATG TACGTCTTGA CGTCGGAAGA CGATTTAGCT CTCTGAATAT TTGGATGGAA
  2341 ATGTGCTGAC CTTGTTGGGG ATACCAGGTC GAAGAATCGT TGATTTTGGC AGCAGTATTT
  2401 TCCCTCGAAC TGAATAAGCA CGTGGAGGTG AGGTTGCCCA TCTTCATGAA GCTCTCTGCA
  2461 GATTTTTATG AATTTTTTGT TAATCGGAGT GTTTAGGGCT TGTAATTGAG AAAGTGATTC
  2521 TTCTTTGGAC AAGGAGCACT GAGGATATGT AAGAAAATAA TTTTTGGCAT TTATTTAAAA
  2581 CTGTTTTG
///
```

Encadré 1 : une référence de la base de données EMBL sur les géminivirus.

```
ID   GEMSVSTR    preliminary;  DNA;  2690 BP
XX
AC   V00314;
XX
DT   02-JUN-1988  (annotation)
XX
DE   Maize streak virus (MSV-S) infectious genome
XX
KW   circular, capsid protein; untranslated reading frame;
KW   genome
OS
OS   Maize streak virus (MSV)
OC   Viridae; ss-DNA nonenveloped viruses; Geminiviridae.
XX
CC   /source=  /strain=South African isolate; /library=pEMBL
CC   ORFs for 10.9 kd and 26.8 kd proteins correspond to those reported
CC   for MSV-M and MSV-K both by sequence and position homology;
CC   overlapping arrangement of ORFs for 31.3 kd and 17.2 kd proteins
CC   is identical to that reported for MSV-M; ORF for put. 8.7 kd
CC   protein remains unsure.
XX
XX
RN   [1]    (bases 1-2690)
RA   Lazarowitz S.G.;
RT   "Infectivity and complete nucleotide sequence of the genome of a
RT   South African isolate of maize streak virus.";
RL   Nucl. Acids Res. 16: 229-249(1987).
XX
FH   Key             From   To          Description
FH
FT   1AAMSCR1        1113               longer (+)-strand transcript (circular
FT                                      genome)
FT   CDS             <1     301         10.9 kd protein (circular genome)
FT                                      (1 is 3rd base in codon)
FT   CDS             315    1046        26.8 kd virion capsid protein
FT   1AAMSCR2685     >2690              longer (+)-strand transcript (circular
FT                                      genome)
FT   CDS             2689   >2690       10.9 kd protein (circular genome)
FT                                      (2690 is 2nd base in codon)
FT   CDS             2377   1362 (C)    31.3 kd protein
FT   CDS             1664   1221 (C)    17.2 kd protein
FT   CDS             303     231 (C)    17.2 kd protein
XX
SQ   Sequence  2690 BP;  693 A; 826 C; 588 G; 673 T;
```

esql. 1

Encadré 2 ; une référence de la base de données GenBank sur les géminivirus.

```
LOCUS       ALTDNAVA        2588 bp     DNA              entered   11/20/88
DEFINITION  Tomato golden mosaic virus  component A of complete genome.
ACCESSION   X02029
KEYWORDS    coat protein, complete genome, unidentified reading frame
SOURCE
  ORGANISM  Tomato golden mosaic virus
            Viridae, ss-DNA nonenveloped viruses, Geminiviridae
REFERENCE   1  (bases 1 to 2588)
  AUTHORS   Hamilton,W.D.O., Stein,V.E., Coutts,R.H.A. and Buck,K.W.
  TITLE     Complete nucleotide sequence of the infectious cloned DNA
            components of tomato golden mosaic virus: Potential coding
            regions and regulatory sequences
  JOURNAL   EMBO J. 3, 2197-2205 (1984)
COMMENT     History: (Converted from EMBL format to GenBank format)
            20-NOV-1988 (incorporated)
            Geminiviruses are characterized by twin icosahedral virions,major
            capsid polypeptides of about 28 kd and ss-DNA genomes.The genomes
            of cassava latent virus (CLV) and tomato golden mosaic virus
            (TGMV) consist of two circular components,while that of maize
            streak virus (MSV) consists of a single circle.Wheat doesn't appear to
            be similar to either of these [1].
            [1] identifies the following additional open reading frames on
            the complementary strand that would code for proteins with >10kd
               AL1 - 13-1343 (passing through origin) AL2 - 1401-1712 AL3 -
            1445-1067. The sequence at 1-233 is highly homologous to an
            equivalent region on component B, it doesn't appear to code for
            protein and has the potential to form a stable hairpin [1]. An
            analogous region is found in CSV. The virion-sense (+) is shown
            below.
FEATURES             from    to      description
                     327     1070    coat protein (AR1)
BASE COUNT           672 A   513 C   603 G   798 T
SEQUENCE
     1 GATCCGATGG CATTTTTGTA ATTAAGAGGC TTACTACCAA TTGAGCAGGG GCTCCAAAAG
    61 TTATATGAAT TGGTAGTAAG GTAGCTCTTA TATATTAGAA GTTCCTAAGG GGCACGTGGC
   121 GGCCATCCGT TTANTATTAC CGGATGGCCG CGCGATCGTC ACCGGACCCG CTTCCGCAAA
   181 TTACGCCGCA TTGTCGTCTA AGTGGTCCCG CATATGTGAA GGGCCAATCA TATTTGCCCC
   241 TGAAATCTAA GATATTTTTA AAGACTTGTG GTTAAGTTGT TAAATTTATA TAAAACGACA
   301 TCCGTTTCGT GGATCTTTAA TTCAAAATGC CTAAGCGGGA TGCCCCATGG CGTTTAATGG
   361 CAGGGACCTC AAAGGTTTCC CGCTCTGCTA ATTATTCTCC TCGAGGAAGT TTGCCTAAGC
   421 GTGATGCTTG GGTTAACAGG CCCATGTACA GGAAGCCCAG GATATATCGA TCACTAAGAG
   481 GCCCCGATGT TCCTAAAGGA TGTGAAGGGC CTTGTAAAGT CCAGTCATAC GAGCAGCGTC
   541 ATGATATTTC CCTAGTTGGG AAGGTCATGT GTATATCTGA TGTGACACGT GGTAACGGTA
   601 TTACCCACCG TGTTGGTAAG CGTTTCTGCG TTAAGTCTGT ATATATCTTG AGCAAGGATA
   661 GGATGGATGA GAACATCAAG TTGAAGAAGC ACACGAACAG TGTCATGTTC TGGTTGGTTA
   721 GGGATCGGAG ACCTTATGGC ACTCGTATGG ATTTCGGACA AGTGTTCAAC ATGTTCGATA
   781 ATGAGCCAAG TACTGCAACG GTAAAGAACG ACCTACGGGA TCGTTTCGAA GTGATCCACA
   841 GGTTTCACGC CAAGGTTACT GGTGGTCAAT ATGCCAGCAA CGAGCAGGCT CTGGTTAGGA
   901 GATTCTGGAA GGTCAATAAC AATGTCGTCT ACAACGACCA GGAGGCAGGG AAATATGAGA
   961 ATCATACTCA GAACGCCCTG TTATTGTATA TGGCATGTAC TCATGCCTCT AACCCTGTGT
  1021 ATGCGACGTT GAAAATTCGA ATCTATTTTT ATGATTCGAT AACAAATTAA TAAAATTTAT
  1081 ATTTTATTGA ATGATTTTCG AGTACATGCA TTATATATGA TCTATCTGTT GCGAAACGAA
  1141 CAGCTCTAAT AACATTGTTA ATACATATAA CGCCTAACTA TTCAAGGTAC AACATCACTA
  1201 AGTATTTAAA TCTATTTAAA TAAGTTCTCC CAGAAGCTAT CGTCGATGTC GTCCATACTT
  1261 GGAAGTTGAG AAATGCCTTG TGGAGATCCA ATGCTCTCCT CAAGTTGTGG TTGAACCTGA
  1321 TTTGTAAGTG GTATATCCTG GTGTTGGTGT AGAGGGGATC CTCTACGCTG ATTATCTTGA
  1381 AATAGAGGGG ATTGTTATC TCCCAGATAT AGACGCCATT CTCTGCTTGA GGCACAGTGA
  1441 TAGGTTCCCC TGTGCGTGAA TCCATTGTTT CTGCAGTCGA TGTGAATGTA TATGGAACAG
  1501 CGACAATTCA GGTCAATTCG TCGCCTTGTA ATAGCTCTTC GTTTAGCTGC TCTGTGTTGA
  1561 GCTTTAATAG AGAGAGGAGT TAGGGAAGAC GAATTTCGCA TTATGGAAAG TCCAGTTCTT
  1621 TAGTAGAGTG TTTTCTTCTT TGTCGGGAAC GAATTTAAGT GGCTTTCCGT ATTTACAGTT
  1681 GTACAGCACA ATTGACGAGA TACTCCTTT AATTTGAACT GGCTTTCCGT ATTTACAGTT
  1741 AGTCTGTCAA TCTCTTTGAG CCCCAATGAA TTCTTTCCAA TGTTTCAACT TTAGATATTG
  1801 CGGTGTGACA TCATCGATGA CGTTATACTC AACCTTGTTT GAGTAAACCC TAGAATTGAG
  1861 ATCCAAATGC CCGCTCAAAT AATTATGTGG GCCTAGTGAA CGAGCCCACA TAGTCTTTCC
  1921 CGTCCGACTA TCGCCCTCGA TGATAATACT AATAGGTCTC TCCGCCGCGA CAGCGGAACT
  1981 CTTTCCAAAA TAATTTTCAG CCCATTGTCT CATCTCGTCT GGCACGTTAG TAAATGATGA
  2041 GACGTGGAAC GGAGGAAGCC ATGGTTCAGG AGTCTTATCA AATATCCTAT CTAAATTGCT
  2101 ATTTAGATTG TGGAACTGAA ATAAATATTT TTCTGGGATT TTCTCTCTAA TTATCTGCAG
  2161 GGCTTCTTCT TTGGAAGAAG CATTTAACGC CTCTGCTGCA GCGTCGTTAG ATGTTTGGCA
  2221 ACCTCCTCTA GCACTTCGAC CGTCGACCTG GAATTCTCCC CATACAAGAG TATCTCCGTC
  2281 TTTGTCGATG TACGTCTTGA CGTCGGAAGA CGATTTAGCT CTCTGAATAT TTGGATGGAA
  2341 ATGTGCTGAC CTTGTTGGGG ATACCAGGTC GAAGAATCGT TGATTTTGGC AGCAGTATTT
  2401 TCCCTCGAAC TGAATAAGCA CGTGGAAGTG AGGTTGCCCA TCTTCATGAA GCTCTCTGCA
  2461 GATTTTTATG AATTTTTTGT TAATCGGAGT GTTTAGGGCT TGTAATTGAG AAAGTGATTC
  2521 TTGTTTGGGC AAGUAGCAGT GAGGATATGT AAGAAAATAA TTTTTGGCAT TTATTGAAAA
  2581 CCGTTTTG
///
```

Encadré 3 : séquence d'une protéine capsidaire de géminivirus au format CODATA.

```
ENTRY               VCCOUV *Type Protein
TITLE               Coat protein - Wheat dwarf virus
DATE                30-Sep-1987 *Sequence 30-Sep-1987 *Text 31-Dec-1989
PLACEMENT           1679.0 2.0    1.0    1.0    1.0
SOURCE              Wheat dwarf virus
ACCESSION           A23005\ A24336
REFERENCE           (Sequence translated from the DNA sequence)
      *Authors      MacDowell S.W., MacDonald H., Hamilton W.D.O.,
                    Coutts R.H.A., Buck K.W.
      *Journal      EMBO J. (1985) 4:2173-2180
      *Comment      The authors translated the codon AAG for residue 3
                      as Arg
COMMENT             This virus is a member of the group Geminivirus.
SUPERFAMILY         *Maize maize streak virus coat protein
KEYWORDS            coat protein
SUMMARY             *Molecular-weight 29408     *Length 260    *Checksum  2063
SEQUENCE
                5         10        15        20        25        30
     1 M V T M K D S A Q K Q K A K N E E Q E S E G R W K D A V K
    31 R R K Q A V K V V P V K P P A L C V F A Y H W L N S D A T H
    61 I V V G H T P A V D L I T C F A Q Q K A Q H H A H T M Q T V
    91 L Y K F H I Q Q T C Y M S D A S A P F I Q P V A L Y H W L V
   121 Y D A E P K Q A M P D A I D I F T M P H M L L P S T H T V Q
   151 A A W S H R F V V K A K V T V N L V T D G R K V Q S K T V D
   181 Q A Y H W V V Q K H I V D A N K F F K G L A V T T E H A N T
   211 D D Q K I G D I K K G A L V L I S S T R G G V T D D S A S T
   241 R F D V V C A Y T H A C V F K A I D I Q
\\\
```

Encadré 4 : corpus des séquences de géminivirus.

Acronyme	Phytovirus	Type	Vecteur	Numéro d'accès	Source
ACMV-KE	African cassava mosaic virus	2 c	Mouche blanche	J02057	(15)
ACMV-NG	African cassava mosaic virus	2 c	Mouche-blanche	X17095	(10)
BCTV-US	Beet curly top virus	1 c	Cicadelle	M24597	(14)
BGMV-PR	Bean golden mosaic virus	2 c	Mouche-blanche	M10070	(4)
BGMV-GA	Bean golden mosaic virus	2 c	Mouche-blanche	c.p	(8)
BGMV DR	Bean golden mosaic virus	2 c	Mouche-blanche	c.p.	(8)
CSMV-AU	Chloris striate mosaic virus	1 c	Cicadelle	M20021	(1)
DSV-VU	Digitaria streak virus	1 c	Cicadelle	n.e.	(2)
MSV-NG	Maize streak virus	1 c	Cicadelle	X01633	(11)
MSV-ZA	Maize streak virus	1 c	Cicadelle	Y00514	(5)
MYMV TH	Mung bean yellow mosaic virus	2 c	Mouche blanche	c.p	(9)
SqLCV US	Squash leaf curl virus	2 c	Mouche blanche	c.p	(3)
TGMV DR	Tomato golden mosaic virus	2 c	Mouche blanche	K02029	(3)
TYLCV TH	Tomato yellow leaf curl virus	2 c	Mouche blanche	c.p	(13)
TYLCV II	Tomato yellow leaf curl virus	1 c	Mouche-blanche	c.p	(12)
WDV CS	Wheat dwarf virus	1 c	Cicadelle	n.e.	(16)
WDV SE	Wheat dwarf virus	1 c	Cicadelle	X02869	(7)

Légende :

1c 1 composant génomique — 2c 2 composants génomiques
c.p communication personnelle — n.e. non enregistré

Les numéros d'accès sont ceux de la banque EMBL

La provenance géographique du virus est indiquée par le code ISO du pays à deux lettres :

AU	Australie	BR	Brésil
CS	Tchécoslovaquie	DR	République dominicaine
GA	Guatemala	KE	Kenya
IL	Israël	NG	Nigéria
PR	Porto-Rico	SE	Suède
TH	Thaïlande	US	Etats-Unis d'Amérique
VU	Vanuatu	ZA	Afrique du Sud

Sources bibliographiques des séquences de géminivirus

(1) Andersen M. T. & al. (1988), Virology 164 : 443-449.
(2) Donson J. & al. (1987), Virology 161 : 160-169.
(3) Hamilton W. D. O. & al. (1984) EMBO J. 3 : 2197-2205.
(4) Howarth A. J. & al. (1985) Proc. Nat. Acad. Sci. USA 82 : 3572-3576.
(5) Lazarowitz S. G. c.p.
(6) Lazarowitz S. G. (1987) Nucl. Acids Res. 16 : 229-249.
(7) MacDowell S. W. & al. (1985), EMBO J. 4 : 2173-2180.
(8) Maxwell c.p.
(9) Morinaga T. & al. (1987), VII Intern. Congress of Virology, Abstracts : 258.
(10) Morris B. & al. (1990), Nucl. Acids Research: 18 : 197-198
(11) Mullineaux P. M. & al. (1984) EMBO J. 3 : 3063-3068
(12) Navot N. c.p.
(13) Rochester c.p. (1989)
(14) Stanley J. & al. (1986), EMBO J. 5 : 1761-1767
(15) Stanley J. & Gay M. R. (1983) Nature 301 : 260-262
(16) Woolston C. J. & al. (1988) Plant Molecular Biology 11 : 35-43

distribuée en avril 1982; cette banque utilise désormais le système de gestion de bases de données (SGBD) relationnel ORACLE. La référence extraite de la banque EMBL [Encadré 1] concerne la séquence complète du génome du virus de la mosaïque en tirets du maïs (*maize streak virus* - MSV), un géminivirus à 1 composant génomique; chaque séquence est identifiée par un mnémonique (ID="GEMSVSXX") et un numéro d'accès (AC="Y00514"), puis comporte, une date (DT) une définition (DE), des mots-clés (KW) et une classification phylogénique (OC) servant à l'indexage, des commentaires (CC) annotant la séquence, des références bibliographiques avec les rubriques auteur (RA), titre (RT) et publication (RL), des caractéristiques (FT) signalant les différents cadres de lecture (*Open Reading Frames* - ORF) et enfin la séquence primaire (SQ).

1.2) GenBank

La Banque de données des séquences génétiques (*Genetic Sequence Data Bank* - GenBank) est produite par le *Los Alamos National Laboratory* (LANL) sous les auspices du *Department of Energy* (DoE) des USA qui en a confié l'exploitation commerciale à la firme *IntelliGenetics* (IG); subdivisée en fichiers spécifiques selon les différents types d'organismes, le schéma conceptuel des données de cette banque a été restructuré dans le nouveau contexte informatique caractérisé par l'introduction d'un SGBD relationnel spécifique. La référence extraite de GenBank [Encadré 2] concerne la séquence du composant A du virus de la mosaïque dorée de la tomate (*tomato golden mosaic virus* - TGMV), un géminivirus à deux composants génomiques; chaque séquence est identifiée par un mnémonique (LOCUS="GETGMVA") et un numéro d'accès (ACCESSION="K02029"), puis comporte une définition (DEFINITION), des mots-clés (KEYWORDS) et une classification phylogénique (SOURCE, ORGANISM) servant à l'indexage, des références bibliographiques (REFERENCE) avec les rubriques auteur (AUTHORS), titre (TITLE) et publication (JOURNAL), des caractéristiques (FEATURES) codifiant les différents signaux biologiques (e.g. la localisation de la protéine capsidaire), des annotations (COMMENT) sur les différents ORF de la séquence, et pour finir la séquence primaire (SEQUENCE).

2) Banques protéiques

Les principales ressources informationnelles dédiées à la collecte des séquences de protéines, celles de la *National Biological Research Foundation* (NBRF) aux USA, du *Martinsrieder Institute für Proteinsequenzen* (MIPS) en Europe et de l'*International Protein Information Database* (JIPID) au Japon, collaborent désormais au sein d'une structure coopérative internationale, la Ressource pour l'identification des protéines (*Protein Identification Resource* - PIR) afin de produire une banque de séquences protéiques unique. L'ensemble des séquences de protéines connues est organisé au sein d'une hiérarchie fondée sur la similarité des séquences et comprenant les niveaux suivants : super-familles, familles et sous-familles, références et sous-références. La PIR est désormais diffusée dans le format CODATA[1] d'échange généralisé pour les séquences. Chaque référence correspondant à une protéine [Encadré 3] comprend un identificateur univoque (ENTRY="VCCVWV"), un titre (TITLE), la séquence primaire des acides aminés (SEQUENCE) et au moins la référence bibliographique (REFERENCE) originale de la séquence, spécifiant les auteurs (#Authors) et la publication (#Journal). D'autres descripteurs permettent de situer l'origine (SOURCE) de la séquence et de la classer (SUPERFAMILY, PLACEMENT) dans la hiérarchie des protéines; les mots-clés (KEYWORDS) servent à l'indexage de la séquence tandis qu'un résumé (SUMMARY) en donne les caractéristiques essentielles (e.g. #Molecular-weigth="29408") Les principaux avantages du format CODATA sont sa

[1] Le *Committee on Data for Science and Technology* est l'un des douze comités spécialisés créés par l'*International Council of Scientific Unions* (ICSU) en 1966 pour encourager la production et la diffusion de banques de données scientifiques et techniques. Réuni à Paris en 1984, le groupe de travail intitulé *Coordination of Protein Sequence Data Banks* a, dans un premier temps, recommandé l'utilisation d'un format commun pour les séquences afin d'en promouvoir l'échange et la distribution. Depuis, ces travaux ont abouti à une proposition de norme (George, Mewes & Kihara 1987).

meilleure lisibilité pour l'opérateur humain (chaque information étant étiquetée par un descripteur ou un sous-descripteur explicite (e.g. #Length="260"), et l'existence d'une spécification formelle (George, Mewes & Kihara 1987) en BNF[2] autorisant une conception fiable et une élaboration aisée des logiciels d'interface avec les formats internes des différentes banques de séquences.

3) Accès aux banques de séquences

Notre équipe de recherche est actuellement dispersée sur trois continents : l'analyse des données est réalisée en Afrique à l'Université Nationale de Côte d'Ivoire (**UNCI**, Abidjan) tandis que l'interprétation des résultats est effectuée par deux phytovirologues de l'ORSTOM, l'un travaillant à l'Université Washington à Saint-Louis dans le Missouri (USA), l'autre au *Scottish Crop Research Institute* à Dundee en Ecosse (Royaume-Uni). La figure [2] donne un aperçu du schéma de communication retenu dans le contexte d'un pays à revenu intermédiaire de l'Afrique sub-saharienne disposant d'un réseau de transmission par paquets. L'accès aux réseaux de la recherche s'effectue par l'intermédiaire du **CIRCI** (Centre Inter-régional de Côte d'Ivoire), premier noeud africain du réseau **EARN** - *European and Academic Research Network* (Desbois & Vidal 1988) en utilisant une liaison synchrone transcontinentale (protocole BSC, 9600 bauds) avec le **CNUSC** (Centre National inter-Universitaire Sud de Calcul) de Montpellier (France), les passerelles créées sur le réseau EARN nous permettant l'accès aux différents domaines (BITNET, JANET, EMBNet, ...) de l'inter-réseau qui nous intéressent. Afin de parer à d'éventuelles défaillances, développer l'autonomie informationnelle des chercheurs et aboutir à une diffusion plus large des usages, nous avons expérimenté une solution alternative, plus légère, ne requérant qu'un micro-ordinateur et un modem en liaison asynchrone (accès sur le réseau téléphonique commuté en 300 et 1200 bauds) grâce à l'implantation récente du réseau ivoirien de transmission par paquets, **SYTRANPAC**. La consultation des différentes banques de séquences (GenBank, EMBL, NBRF-PIR) s'effectue alors en temps réel par l'intermédiaire du système **BISANCE** (Base Informatique sur les Séquences de Biomolécules pour les Chercheurs Européens) implanté sur le site du **CITI2** (Centre Inter-universitaire de Traitement de l'Information 2)

III) CLASSIFICATION DES GEMINIVIRUS

1) Le groupe des géminivirus

Les géminivirus sont des virus de plantes qui possèdent des virions isométriques de forme géminée et dont le génome est constitué d'un ou deux brins d'ADN monocaténaire circulaire. Le caractère géminé des particules fut décrit pour la première fois par Bock, Guthrie et Woods (1974) à propos du virus de la mosaïque en tirets du maïs (*maize streak virus* - **MSV**). La présence de molécules circulaires d'ADN monocaténaire fut mise en évidence par Goodman pour le virus de la mosaïque dorée du haricot (*bean golden mosaic virus* - **BGMV**) en 1977. Les géminivirus sont transmis aux plantes-hôtes soit par cicadelles, soit par aleurodes (mouches-blanches). L'intérêt majeur des géminivirus réside dans la nature (l'ADN est une molécule plus stable que l'ARN) et la taille (relativement courte - 2700 à 2800 nucléotides) de leur génome. Pour la biologie moléculaire, ils constituent donc des modèles faciles à étudier et des vecteurs potentiels de gènes incorporables au génome des plantes qu'ils peuvent infecter.

2) La composition en acides aminés de la protéine capsidaire

La protéine capsidaire est une macro-molécule, élément de base dont l'assemblage formera la capside protégeant l'acide nucléique constituant le génome des virus de plante. La composition en acides aminés (**CAA**) de la protéine capsidaire (**PC**) a été utilisée en 1969 par Gibbs ainsi que par Tremaine et Argyle (1969) comme critère de classification pour les virus de plante mais les résultats sont jugés peu probants, en raison sans doute d'un corpus trop restreint (66 virus pour Gibbs). Cependant, Fauquet, Déjardin et Thouvenel (1986a) ont montré sur un corpus plus étendu (122 isolats issus de 23 groupes différents), que la CAA de la PC des phytovirus permet de retrouver la spécificité des groupes de virus de plante définis par

---

[2] *Backus-Naur Form*, métalangage utilisé comme formalisme pour l'écriture des règles de syntaxe des langages *context-free*.

l'ICTV[3] (Matthews 1982). Une étude ultérieure (Fauquet Déjardin & Thouvenel 1986b) portant sur 126 isolats, a mis en évidence une relation entre la CAA de la PC et certains critères de classification tels que la structure des particules et le mode de transmission. Nous avons testé récemment la stabilité de ce schéma de classification sur un corpus plus étendu comportant 174 isolats distincts (Fauquet & al. 1987). Enfin, poursuivant cet axe de recherche, nous avons établi une typologie des phytovirus en bâtonnet (Desbois & al. 1989) permettant de distinguer un nouveau groupe (furovirus) des groupes déjà répertoriés (hordeivirus, tobamovirus, tobravirus). Soulignons que ces différents corpus ne comprenaient que des virus à ARN.

3) Le corpus des données

Afin d'établir une typologie semblable pour le groupe des géminivirus, nous avons réuni un ensemble de 17 séquences nucléiques de la protéine capsidaire concernant 11 géminivirus distincts [Encadré 4]. L'existence de séquences nucléiques permet de ne pas se limiter à la composition en acides aminés et de constituer trois tableaux de contingence distincts :

i) K_{IxJ1}, croisant l'ensemble I des 17 isolats et l'ensemble J1 des 20 acides aminés, où k(i,j') est le nombre d'occurrences de l'acide aminé j' au sein de la séquence i;
ii) K_{IxJ2}, croisant l'ensemble I des 17 isolats et l'ensemble J2 des 16 dinucléotides, où k(i,j'') est le nombre d'occurrences du dinucléotide j'' au sein de la séquence i;
iii) K_{IxJ3}, croisant l'ensemble I des 17 isolats et l'ensemble J3 des 61 codons[4], où k(i,j''') est le nombre d'occurrences du codon j''' au sein de la séquence i.

autorisant ainsi l'analyse comparative des profils respectifs des différentes séquences pour leur composition en acides aminés, en dinucléotides et en codons.

4) Les méthodes d'analyse

L'Analyse Factorielle des Correspondances (AFC) est une technique multidimensionnelle développée par Benzécri et ses collaborateurs (1973) comme méthode exploratoire de description statistique pour l'analyse des données. Elle permet de construire des représentations graphiques des lignes (17 isolats-géminivirus) et des colonnes (respectivement 20 acides aminés, 16 dinucléotides ou 61 codons) d'un tableau de contingence : les profils-lignes et les profils-colonnes correspondants sont projetés dans un espace factoriel de dimension réduite. Des diagrammes ou **graphiques-plans** croisant les premiers facteurs (correspondant aux axes de variabilité les plus importants) permettent alors d'étudier la forme globale du nuage de points ainsi que leurs positions respectives. Chaque point-ligne représente ainsi la composition du profil (acide-aminé, dinucléotide ou codon) de la PC d'un isolat-géminivirus. La **distance du χ^2** entre deux profils-lignes donne ainsi une mesure de l'éloignement ou de la proximité entre les PC des différents isolats. De plus, les diagrammes réalisés, d'une part pour les points-lignes et d'autre part pour les points-colonnes, peuvent être superposés de telle sorte qu'on puisse interpréter les agrégats homogènes de points-lignes au sein du nuage des isolats (en tant que groupes putatifs de géminivirus) en termes de profil de composition spécifique décrit par la position relative des points-colonnes (acide-aminé, dinucléotide ou codon) vis à vis de ces classes. Les calculs ont été effectués par le programme (ANAFAC CORR) de la librairie ADDAD[5]. Une classification ascendante hiérarchique (**CAH**, Jambu & Lebeaux 1983) est ensuite effectuée sur les 7 premières coordonnées factorielles de l'AFC pour regrouper les isolats en classes homogènes. Ainsi les dissimilarités entre unités taxinomiques sont-elles définies par la métrique du χ^2. La contribution de chaque acide aminé, dinucléotide ou codon à la distance totale entre deux profils-isolats distincts est

---

[3] Comité international pour la taxinomie des virus (*International Committee on the Taxonomy of Viruses*). L'ICTV est organisé en "sous-comités spécialisés" chargé pour chaque groupe-hôte de virus (infectant les vertébrés, les invertébrés, les plantes, les bactéries et les champignons) de la mise en place de "groupes de travail" constitués d'experts de différents pays, élaborant des propositions taxinomiques. L'ICTV publie régulièrement une nomenclature permettant à chaque virologiste de disposer d'une base cohérente et actualisée de classification.

[4] Les 3 codons "stop" (TGA, TAG, TAA) sont exclus de l'analyse.

[5] Association pour le Développement de l'Analyse des Données, 22 rue Charcot, Paris 75013.

pondérée par la fréquence marginale respective de l'acide aminé, du dinucléotide ou du codon dans la distribution considérée (propriété de **distance distributionnelle**). Le critère d'agrégation est la **maximisation du moment centré d'ordre 2** (une variante de l'algorithme de Ward introduite par Benzécri en 1968) qui permet d'obtenir des classes homogènes et bien séparées. Le taux d'inertie du noeud, rapport de l'inertie du noeud à l'inertie totale, fournit une mesure de la distance entre classes variant de la valeur 0 (identique) à la valeur 1 (différent); on l'interprète comme un indice de dissimilarité entre les deux précurseurs du noeud (aîné et benjamin). Comme test de validité, on utilise une procédure de Monte-Carlo pour effectuer des simulations (au nombre de 10) en soumettant à l'algorithme ascendant de classification hiérarchique les tableaux de données obtenus par permutation aléatoire des coordonnées factorielles analysées afin de comparer les taux d'inertie observés aux taux d'inertie simulés. Les partitions significatives sont détectées selon la règle suivante : les seuils d'agrégation observés doivent être supérieurs aux seuils d'agrégation simulés (Jambu 1978). Les calculs sont effectués par le programme (SIMCAH1) de la bibliothèque ADDAD.

5) Résultats

5.1) Le critère de la composition en acides aminés

L'examen des coordonnées factorielles de l'AFC de la CAA de la PC des 17 isolats révèle immédiatement ([Figure 3]; graphique-plan F1xF2, espace des profils-lignes) une partition en deux classes constituée :

- d'une part de géminivirus à 1 composant génomique infectant des monocotylédones, transmis par cicadelles (CSMV, DSV, MSV, WDV) et projetés du coté positif du premier axe factoriel F1 (taux d'inertie : 44 %); les contributions à l'inertie de l'axe factoriel (indice **CTR**) les plus importantes proviennent des isolats DSV-VU(14,5%), MSV-NG et MSV-ZA (12,4% chacun) ainsi que du CSMV-AU (11%), la contribution des isolats du WDV étant plus faible (6% chacun), ce groupe contribuant au total à plus de 62% de l'inertie de cet axe; ces profils sont bien représentés par leurs projections sur F1 puisque leur niveau de corrélation à l'axe est élevé (indice **CO2**[6] variant de 0,74 à 0,45);
- d'autre part de géminivirus infectant des plantes dicotylédones, transmis par mouches-blanches et possédant 2 composants génomiques (ACMV, BGMV,TGMV, SqLCV, TYLCV) projetés du coté négatif de l'axe F1; leur niveau de contribution est plus modeste, variant de 8,9% pour l'ACMV à 3,7% pour le SqLCV-US; les projections restent assez fidèles puisque les corrélations de ces profils à l'axe F1 sont assez élevées (l'indice CO2 varie de 0,62 à 0,30 pour ces isolats).

Les acides aminés caractérisant le demi-axe F1>0 ([Figure 4]; graphique-plan F1xF2, espace des profils-colonnes) sont l'ALAnine , la THRéonine, la GLYcine et le TRyPtophane; leur contribution représente environ 43% de l'inertie de l'axe F1 et leur corrélation à l'axe est élevée (l'indice CO2 varie de 0,69 à 0,54). Le demi-axe F1<0 est caractérisé par la METhionine, L'ASparagiNe, l'HIStidine et la TYRosine, l'ensemble représentant environ 41% de l'inertie de l'axe avec des profils correctement représentés (l'indice CO2 varie de 0,73 à 0,49).

L'inertie de l'axe factoriel F2 (19%) est principalement constituée par la contribution du BCTV-US (CTR=56,3%) dont le profil possède un excellent niveau de corrélation à cet axe (indice CO2=0,83) et dont la projection se situe du coté négatif de l'axe F2. Ce demi-axe F2<0 est caractérisé par les acides aminés THréonine, METhionine et LYSine totalisants une contribution de plus de 37% à l'inertie de l'axe avec un niveau de corrélation des profils à l'axe F2 acceptable (CO2 varie de 0,62 à 0,35).

Signalons également que les profils les plus excentrés, de par leur contribution totale (indice **INR**) à l'inertie du nuage des points-lignes, sont ceux des isolats BCTV-US (INR=13%), DSV-VU (INR=12%) et CSMV-AU (INR=11%).

Les projections sur les axes de rang supérieur, dont le pourcentage d'inertie décroît

---

[6] L'indice CO2 s'interprète comme le carré d'un coefficient de corrélation.

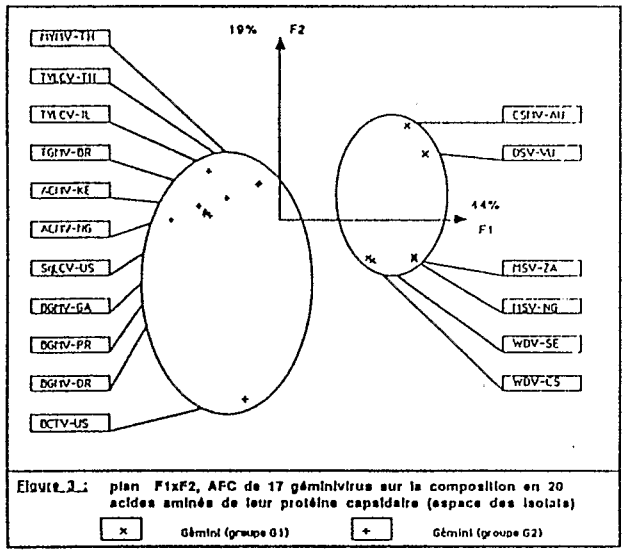


Figure 3 : plan F1xF2, AFC de 17 géminivirus sur la composition en 20 acides aminés de leur protéine capsidaire (espace des isolats)

× Gémini (groupe G1) + Gémini (groupe G2)

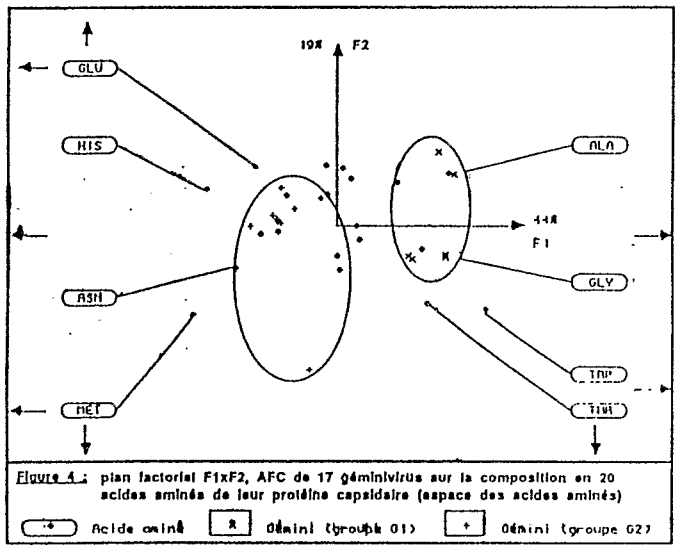


Figure 4 : plan factoriel F1xF2, AFC de 17 géminivirus sur la composition en 20 acides aminés de leur protéine capsidaire (espace des acides aminés)

Acide aminé R Gémini (groupe G1) + Gémini (groupe G2)

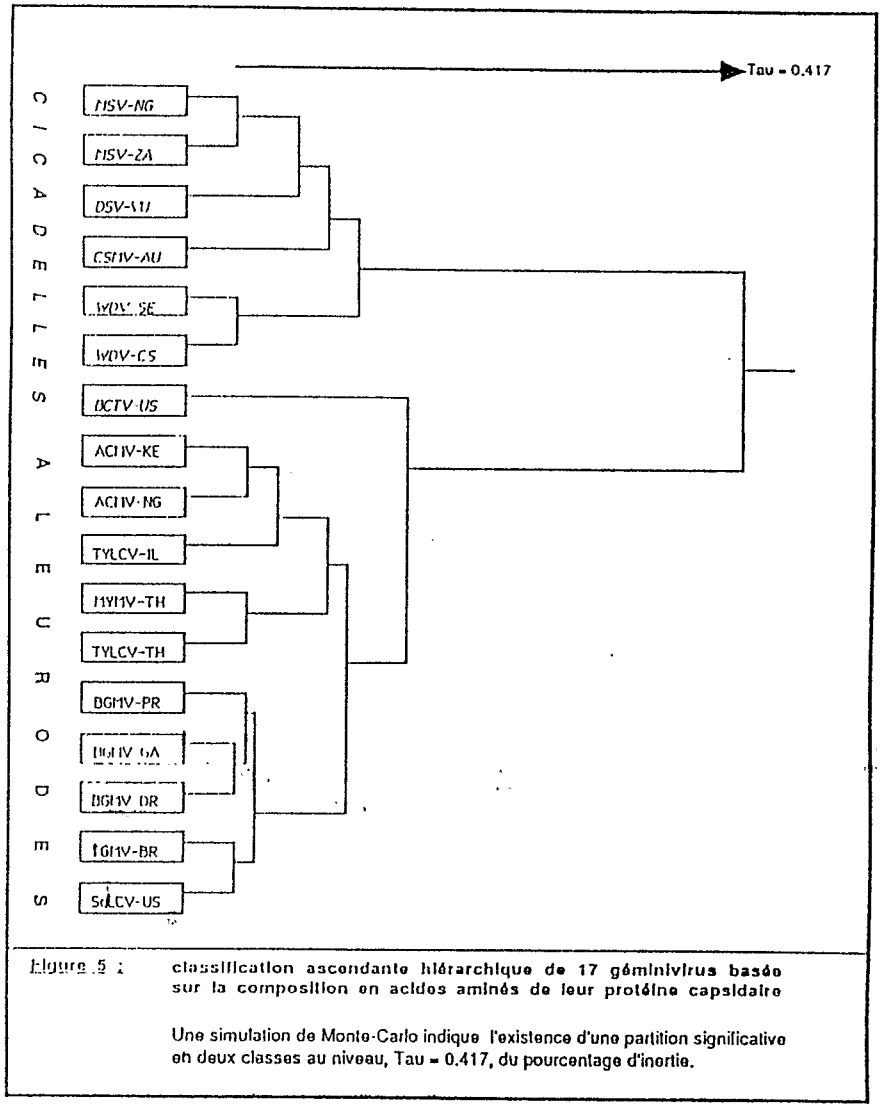


Figure 5 : classification ascendante hiérarchique de 17 géminivirus basée sur la composition en acides aminés de leur protéine capsidaire

Une simulation de Monte-Carlo indique l'existence d'une partition significative en deux classes au niveau, Tau = 0.417, du pourcentage d'inertie.

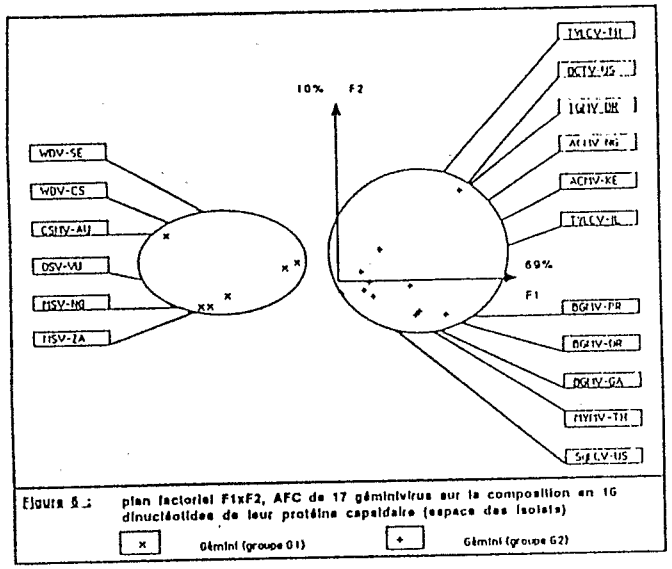


Figure 6 : plan factoriel F1xF2, AFC de 17 géminivirus sur la composition en 16 dinucléotides de leur protéine capsidaire (espace des isolats)

× Gémini (groupe G1) + Gémini (groupe G2)

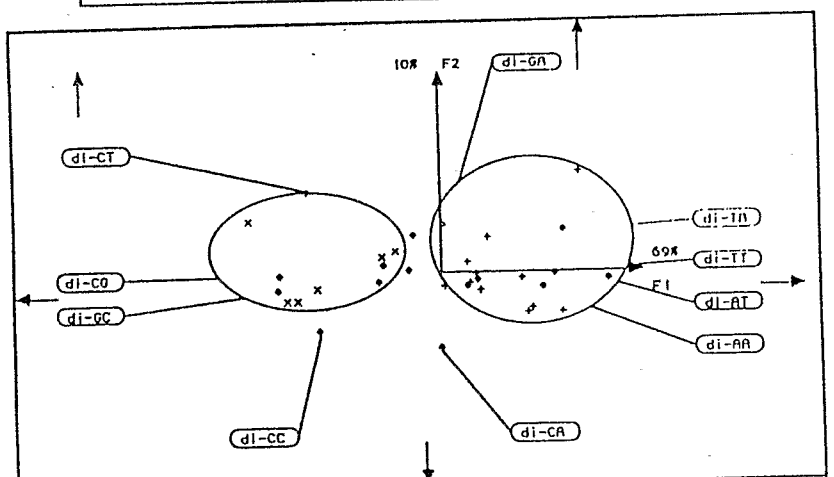


Figure 7 : plan factoriel F1xF2, AFC de 17 géminivirus sur la composition en 16 dinucléotides de leur protéine capsidaire (espace des dinucléotides)

• Dinucléotide × Gémini (groupe G1) + Gémini (groupe G2)

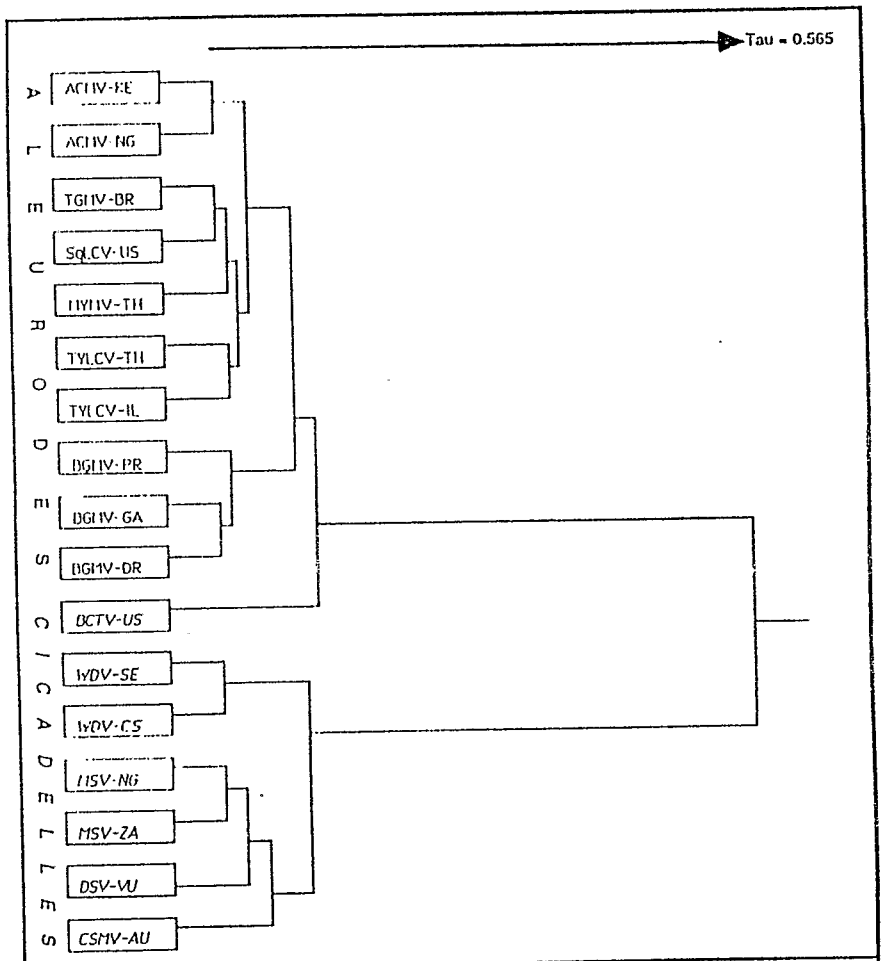


Figure 8 : **classification ascendante hiérarchique de 17 géminivirus basée sur la composition en dinucléotides de leur protéine capsidaire**

Une simulation de Monte-Carlo indique l'existence d'une partition significative en deux classes au niveau, Tau = 0.565, du pourcentage d'inertie.

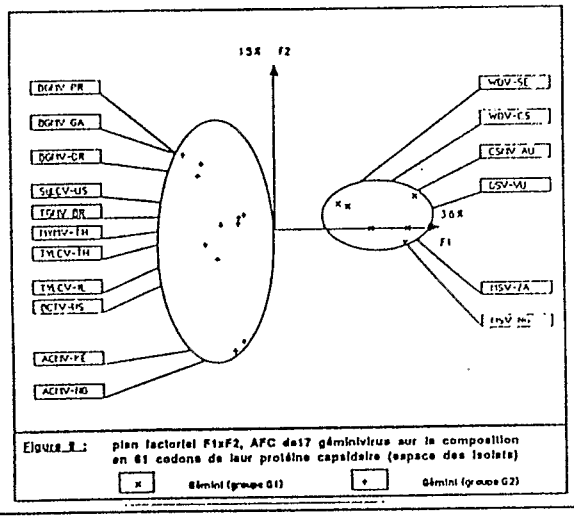


Figure 9 : plan factoriel F1xF2, AFC de17 géminivirus sur la composition en 61 codons de leur protéine capsidaire (espace des isolats)

× Géminl (groupe G1) + Géminl (groupe G2)

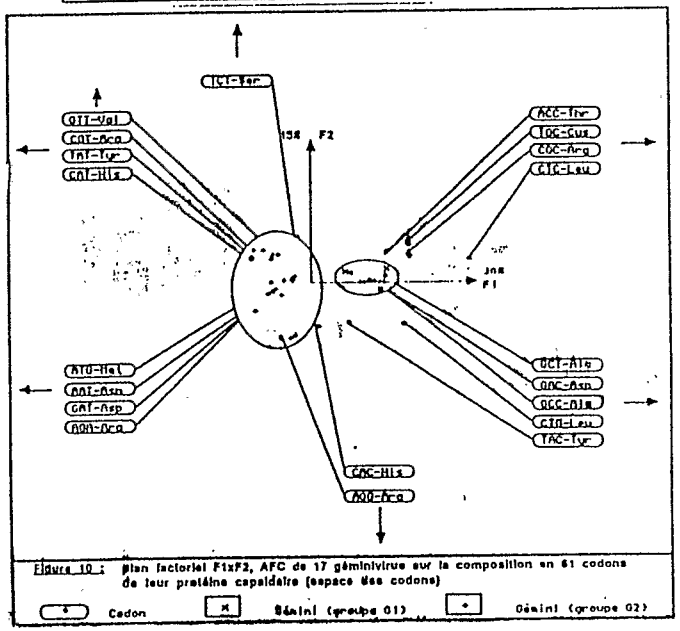


Figure 10 : plan factoriel F1xF2, AFC de 17 géminivirus sur la composition en 61 codons de leur protéine capsidaire (espace des codons)

+ Codon × Géminl (groupe G1) + Géminl (groupe G2)

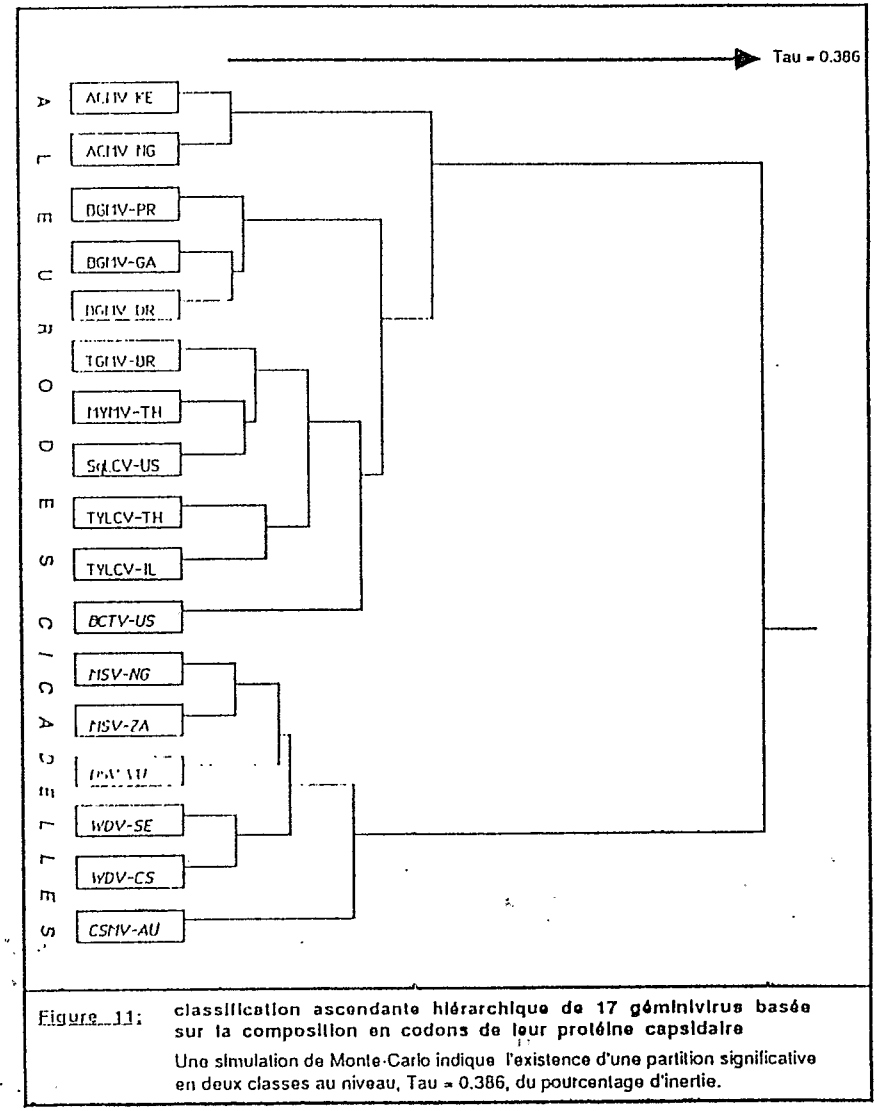


Figure 11: **classification ascendante hiérarchique de 17 géminivirus basée sur la composition en codons de leur protéine capsidaire**

Une simulation de Monte-Carlo indique l'existence d'une partition significative en deux classes au niveau, Tau = 0.386, du pourcentage d'inertie.

rapidement, ne bénéficient pas d'un indice de corrélation suffisant pour autoriser leur interprétation.
La partition en deux classes des profils-lignes que nous avons décelée sur le premier plan factoriel F1xF2, est mis en évidence par la simulation de Monte-Carlo réalisée à partir des résultats de l'algorithme ascendant de classification hiérarchique appliqué aux 7 premières coordonnées factorielles des isolats ([Figure 5]; dendogramme issu de la CAH). Bien que les seuils d'agrégation aux niveaux médians de la hiérarchie ne permettent pas de juger les partitions correspondantes comme significatives, les agrégats observés au sein de la classe des géminivirus transmis par mouches-blanches suggèrent une logique de regroupement par continents (Amérique : SqLCV, TGMV, BGMV; Asie : MYMV, TYLCV-TH; Afrique et Proche-Orient : ACMV, TYLCV-IL). Aux niveaux inférieurs de la hiérarchie, remarquons que les différents isolats se regroupent au sein d'une classe spécifique à chaque virus ([ACMV-KE, ACMV-NG], [BGMV-PR, BGMV-GA, BGMV-DR], [MSV-NG, MSV-ZA], [WDV-SE, WDV-CS]) à l'exception du TYLCV dont l'isolat thaïlandais s'agrège au MYMV, de même provenance.

5.2) Le critère de la composition en dinucléotides

L'analyse des résultats du tableau K_{IxJ2} conduit à des conclusions similaires au plan des profils-lignes [Figure 6] : une partition évidente en deux groupes des 17 isolats selon l'axe F1 (69% d'inertie) avec les géminivirus transmis par cicadelles projetés sur le demi-axe F1<0 et ceux transmis par mouches-blanches sur le demi-axe F1>0, ainsi que la spécificité du BCTV-US qui se manifeste sur l'axe F2 (10% d'inertie). Signalons que les isolats les plus excentrés sont le CSMV-AU, le BCTV-US et les deux souches du MSV.
La figure [7] représente les différents nucléotides contribuant de façon majeure à l'inertie des axes F1 et F2 (les flèches indiquent le sens et la direction des contributions), dans un plan factoriel cumulant près de 80% de l'inertie totale du nuage des profils. La décroissance des valeurs propres s'avère beaucoup plus rapide que dans l'analyse précédente et conduit à délaisser l'interprétation des axes de rang supérieur dont l'inertie apparaît trop faible pour nous révéler des phénomènes significatifs.
Au niveau du dendogramme issu de la CAH [Figure 8], nous retrouvons cette partition en deux classes avec, toutefois, un taux d'inertie plus élevé (t = 0,565 contre t = 0,417 dans l'analyse précédente). Signalons également de légères discordances avec la hiérarchie précédente aux niveaux inférieurs : ainsi les deux isolats du TYLCV se trouvent réunis, tandis que le couple [TGMV, SqLCV] s'agrège à la classe des ACMV plutôt qu'à celle des BGMV.

5.3) Le critère de la composition en codons

La décroissance des valeurs propres de l'AFC du tableau K_{IxJ3}, se révèle beaucoup plus lente : ainsi, le premier plan factoriel F1x F2 ne totalise que 51% de l'inertie totale. Cependant, on observe [Figure 9] toujours la partition des 17 isolats en deux classes selon l'axe F1 (36% de l'inertie) : les géminivirus transmis par mouches-blanches projetés sur le demi-axe F1<0 et ceux transmis par cicadelles sur le demi-axe F1>0. Par contre l'axe F2 (15%de l'inertie) marque la spécificité de l'ACMV; ses deux isolats, projetés du coté négatif de l'axe, contribuent à plus de 64% de son inertie. La particularité du BCTV s'exprime désormais dans le plan factoriel F3xF4 cumulant environ 18,5% de l'inertie totale.
La figure [10] indique les associations de groupes de codons permettant de caractériser les deux premières directions factorielles. Notons que pour certains acides aminés (e.g. ARGinine, acide ASPartique, HIStidine, TYRosine) l'identité de la troïsième base du triplet utilisée pour coder l'acide aminé contribue à différencier les deux classes d'isolats selon l'axe F1 : ainsi le triplet TAT-Tyr s'oppose-t-il au codon TAC-Tyr, le triplet AGA-Arg contribue à l'axe F1 (demi-axe<0) tandis que le codon AGG-Arg contribue à l'axe F2 (demi-axe<0), le triplet GAT-Asp (F1<0) s'oppose au codon GAC-Asp, CGT-Arg (F1<0) et CGC-Arg (F1>0) sont projetés de part et d'autre de l'axe F1, CAT-His contribue à l'axe F1 (demi-axe<0) tandis que CAC-His contribue à l'axe F2 (demi-axe<0).
Le dendogramme [Figure 11] montre un taux d'inertie plus faible (t = 0,386) pour la partition en deux classes mais qui reste significatif selon les règles du test utilisé dans la procédure de simulation. L'examen des niveaux inférieurs de la hiérarchie indique que l'ensemble des

isolats se regroupe au sein de classes spécifiques à chaque géminivirus. Dans cette analyse le TGMV, le MYMV et le SqLCV forment un triplet qui s'agrège au couple des TYLCV.

6) Discussion

Les conclusions de cette étude peuvent être résumées par les points suivants :

i) le consensus sur les typologies obtenues démontre la cohérence des trois types d'analyse de la composition (en acides aminés, dinucléotides et codons) de la protéine capsidaire pour le groupe des géminivirus;

ii) l'étude des profils respectifs de composition de la protéine capsidaire en acides aminés, en dinucléotides et en codons révèle
   - l'existence d'une partition des protéines capsidaires en deux classes
     - d'une part, la classe G1 des PC représentant les géminivirus suivants [CSMV, DSV, MSV, WDV],
     - d'autre part, la classe G2 des PC représentant les géminivirus suivants [ACMV, BGMV, BCTV, MYMV, SqLCV, TGMV, TYLCV];
   - la spécificité de la protéine capsidaire du BCTV;
   - l'homogénéité des isolats d'un même virus relativement à ces trois critères (à l'exception du TYLCV, l'ensemble des isolats du corpus s'agrègent au sein de classes spécifiques des virus dont ils sont les représentants);
   - l'usage différentiel, selon les deux classes G1 ou G2, de la troisième base du triplet pour coder certains acides aminés.

Après avoir signalé que la majorité des géminivirus peuvent être classés en deux sous-groupes :

- d'une part, celui des géminivirus à 1 composant génomique transmis par cicadelles et infectant les plantes monocotylédones,
- d'autre part, celui des géminivirus à 2 composants génomiques, transmis par mouches-blanches et infectant les plantes dicotylédones,

Stanley et alii (1986) soulignent la nature hybride du BCTV qui possède les attributs du premier sous-groupe au niveau de la vection (transmis par cicadelles) et du nombre de composants (les auteurs ayant démontré l'infectivité du seul composant génomique décelé) tandis que son organisation génomique et les plantes-hôtes qu'il infecte l'apparente au second sous-groupe. Ils proposent alors de le considérer comme le prototype d'un troisième sous-groupe. La partition en deux classes que nous avons obtenue s'accorde donc avec la classification communément admise des géminivirus en deux sous-groupes et la spécificité du profil de la protéine capsidaire du BCTV vient confirmer l'hypothèse de Stanley et alii. Afin d'établir l'existence putative d'un troisième sous-groupe, il conviendrait d'obtenir sinon les séquences primaires des protéines capsidaires d'autres géminivirus, du moins leur composition en acides aminés.

De même, l'hypothèse d'un facteur géographique de variabilité pour les géminivirus infectant les dicotylédones, émise par Howarth et Vandemark (1989) à partir de la comparaison de séquences de protéines associées à la réplication provenant de 16 isolats distincts de géminivirus, bien qu'elle soit étayée par les regroupements continentaux signalés dans la hiérarchie établie à partir de la composition en 20 acides aminés (cf. § 5.1), n'est pas confirmée par les hiérarchies basées sur les deux autres critères, composition en dinucléotides et en codons; des analyses complémentaires portant sur un plus grand nombre d'isolats de provenance diverses et incluant l'étude de la composition des protéines associées à la réplication permettraient de tester cette hypothèse. Des séquences du génome d'autres isolats du TYLCV seraient également nécessaires pour décider de la parenté de ses deux souches (israélienne et thaïlandaise) et de leur relation avec la souche thaïlandaise du MYMV.

Enfin, la préférence différentielle de la troisième base du triplet selon le sous-groupe constituerait une illustration de la théorie *de l'usage interspécifique du codon dans la stratégie de codage du génome* (Grantham, Perrin & Mouchiroud 1986) au niveau du sous-groupe dans la classification des géminivirus. Pour vérifier cette hypothèse, il conviendrait néanmoins de confirmer ce résultat sur la base d'un corpus de souches et de géminivirus plus important et d'étudier d'autres gènes que celui de la protéine capsidaire.

IV) CONCLUSION

A l'instar de ce qui a été établi pour l'ensemble des phytovirus à ARN (Fauquet, Déjardin et Thouvenel 1986a et 1986b), le profil de composition de la protéine capsidaire des géminivirus, phytovirus à ADN, est en relation avec le critère du vecteur de transmission, cette démonstration généralisant le résultat aux différents profils de composition (non seulement en acides aminés, mais aussi en dinucléotides et en codons). De plus, l'AFC, de par les propriétés de la métrique du χ^2 (distance distributionnelle), s'avère un outil tout à fait adapté à l'étude des profils de composition de la protéine capsidaire, révèlant un ensemble cohérent de critères taxinomiques pertinents pour la classification des géminivirus. La réalisation de ces travaux a été rendue possible par l'accès aux banques de séquences protéiques et nucléiques. Incidemment, le contexte de cette étude plaide pour l'aménagement de procédures et de supports informationnels permettant à la communauté scientifique des pays en développement une exploitation plus systématique des différentes sources d'information scientifique et technique. Il souligne, en outre, la contribution centrale des réseaux télématiques de la recherche à la collecte et à la diffusion de cette information scientifique et technique ainsi que l'utilité manifeste pour la recherche dans les pays en développement d'infrastructures sub-régionales de télécommunications.

Remerciements : Les auteurs remercient C. Fondrat du CITI2 pour son aimable collaboration lors du test des différents protocoles de connexion au système BISANCE ainsi que le Professeur J.-P. Benzécri pour ses conseils bienveillants.

Références bibliographiques :

Benzécri, J.-P. & al. (1973). L'Analyse des Données. I La Taxinomie. II L'Analyse des Correspondances. Paris. Dunod. 363 p., 374 p.

Bock, K. R., Guthrie, E. J. & Woods R. D. (1974). Purification of maize streak virus and its relationship to virus associated with streak diseases of sugarcane and 'Panicum maximum'. Ann. appl. Biol. 77 : 289-296.

Dayhoff, M. O. (1972). Atlas of protein sequence and structure. Washington DC, National Biomedical Research Foundation. 544 p.

Desbois, D. & Vidal, G. (1988). Abidjan devient le premier noeud africain du réseau EARN. Revue Tiers-Monde. XXIX : 1237-1243.

Desbois, D., Fauquet, C., Fargette, D. & Vidal, G. (1989). Typologie des virus de plante en bâtonnet d'après la composition en acides aminés de leur protéine de capside. Les Cahiers de l'Analyse des Données. 14 : 385-392.

Fauquet, C., Déjardin, J. & Thouvenel, J.-C. (1986a). Evidence that the amino acid composition of the particle proteins of plant viruses is characteristic of the virus group : I multidimensional classification of plant viruses. Intervirology. 25 : 1-13.

Fauquet, C., Déjardin, J. & Thouvenel J.-C. (1986b). Evidence that the amino acid composition of the particle proteins of plant viruses is characteristic of the virus group : II discriminant analysis according to structural, biological and classification properties of plant viruses. Intervirology. 25 : 190-200.

Fauquet, C., Desbois, D., Fargette, D. & Vidal, G. (1987). Classification des virus de plante par la composition en acides aminés de leur protéine capsidaire. Rencontres de Virologie Végétale, 1-5 février 1987. INRA/CNRS. Aussois (France). 9.

George, D. G., Mewes, H. W. & Kihara, H. (1987). A standardized format for sequence data exchange. Protein Seq. Data Anal. 1 : 27-39.

Gibbs, A. J. (1969). Plant virus classification. Adv. Virus Res. 14 : 263-328.

Goodman, R. M. (1977). Single-stranded DNA genome in a white-fly transmitted plant-virus. Virology. 83 : 171-179.

Grantham, R., Perrin, P. & Mouchiroud, D. (1986). Patterns in codon usage of different kinds of species. Oxford Surveys in Evolutionary Biology. 3 : 48-81.

Howarth, A. J. & Vandemark, G. J. (1989). Phylogeny of Geminiviruses. J. gen. Virol. 70 : 2717-2727.

Jambu, M. (1978). Classification automatique pour l'analyse des données. Paris. Dunod. 310 p.

Jambu, M. & Lebeaux, M. O. (1983). Cluster analysis and data analysis.Amsterdam. North-Holland. 898 p.

Matthews, R. E. F. (1982). Classification and nomenclature of viruses. Fourth report of the International Committee on Taxonomy of Viruses. Intervirology. 17 : 1-199.

Sanger, F. & Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J. Mol. Biol. 94 : 441-448.

Stanley, J., Markham, P. G., Callis, R. J. & Pinner M. S. (1986). The nucleotide sequence of an infectious clone of the geminivirus beet curly top virus. EMBO J. 5 : 1761-1767.

Tremaine, J. H. & Argyle, E. (1969). Cluster analysis of viral proteins. Phytopathology. 60 : 654-659.

INSTITUT NATIONAL DE RECHERCHE
EN INFORMATIQUE ET EN AUTOMATIQUE

Actes du 1er Colloque Africain sur la Recherche en Informatique

Proceedings of the 1st African Conference on Research in Computer Science

Yaoundé - Cameroun
14 - 20 octobre 1992

Volume I

Editeur/*Editor*
Maurice TCHUENTE