La Revue de Modulad

UTILISATION DES BANQUES DE SEQUENCES POUR LA RECHERCHE TAXINOMIQUE EN PHYTOVIROLOGIE ET APPLICATION DE L'ANALYSE FACTORIELLE DES CORRESPONDANCES A LA CLASSIFICATION DES GEMINIVIRUS

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

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. Des références bibliographiques ainsi qu'une liste d'adresses (postales, téléphoniques, télex, fac-similés et électroniques) permettront au lecteur de compléter aisément sa documentation sur une banque de séquences, un logiciel d'analyse ou un service télématique particulier.

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, Ressources informatiques dans un pays en développement francophone.

---

* Cellule Analyse des Données, Centre Universitaire de Traitement de l'Information, Université Nationale de Côte d'Ivoire, BP V34, ABIDJAN 01, RCI.
Adresse électronique : *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.
Adresse électronique : *FAUQUET @ BIOLGY.WUSTL*

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

I) INTRODUCTION

Les développements de la biologie moléculaire intervenus depuis les années 50 corrélativement aux progrès enregistrés au début des années 60 concernant le traitement automatisé de l'information, tant sur le plan du stockage que de la manipulation des données, ont incité à réunir l'ensemble des données et des faits de la biologie moléculaire pour les mettre à la disposition des chercheurs concernés. Certains projets, initiés pour la plupart dans le courant des années 70, arrivent désormais à maturité et proposent des services que ne saurait ignorer la pratique du chercheur au cours de la prochaine décennie. Notre propos concernera principalement les banques de séquences qui constituent le noyau central de l'offre de services informationnels en direction des chercheurs en biologie, en particulier pour des travaux à caractère taxinomique.

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

La maîtrise de ce phénomène conduit à une révolution non seulement dans les méthodes d'acquisition de séquences mais également dans les techniques de traitement de l'information, tant sur le plan du matériel (vidéodisque, architecture parallèle, .) que du logiciel (techniques de stockage et d'organisation de l'information, méthodes d'accès et de distribution) de par le caractère massif de l'information déjà saisie ou potentiellement stockable. Les réseaux de la recherche ont un rôle essentiel à jouer dans la diffusion de cette information scientifique et technique, en particulier en direction des pays en développement défavorisés au plan de la balance des échanges informationnels. Du fait de l'étendue des domaines prospectés et de la spécificité du champ couvert (biologie moléculaire) ces banques de séquences ouvrent des perspectives intéressantes dans le domaine de la taxinomie numérique en offrant l'opportunité d'analyser des données d'un type nouveau et donc de développer de nouvelles approches.

II) LES BANQUES DE SEQUENCES

Rappelons tout d'abord les principes et les objectifs qui ont présidé au développement de ces banques de séquences :

i) élimination de la redondance des données;
ii) flexibilité du langage de requête et de manipulation des données;
iii) expansion automatique de la banque.

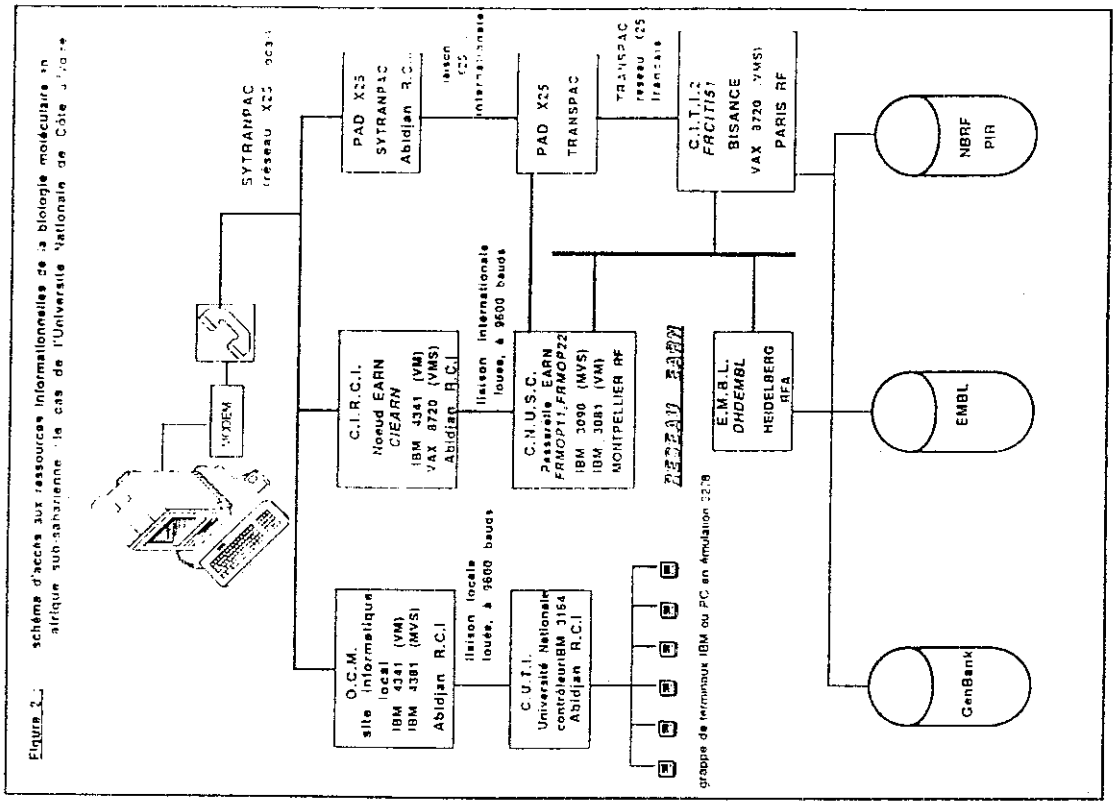

Figure 2 : schéma d'accès aux ressources informationnelles de la biologie moléculaire en afrique sub-saharienne : le cas de l'Université Nationale de Côte d'Ivoire

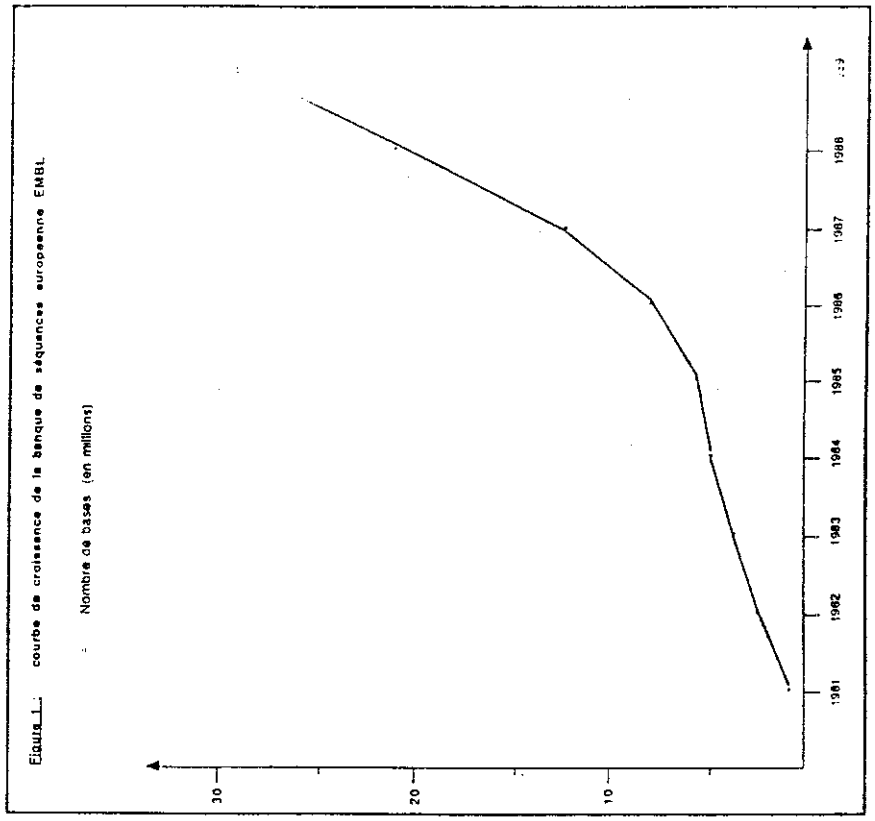

Figure 1 : courbe de croissance de la banque de séquences européenne EMBL

Le mouvement actuel tend à intégrer les connaissances heuristiques aux informations factuelles au sein de bases de connaissances grâce aux techniques de l'intelligence artificielle. 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).

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 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**).

Encadre 1 : une reference de la base de donnees EMBL sur les geminivirus.

Encadre 2 : une reference de la base de donnees GenBank sur les geminivirus.

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 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) Acquisition des données

La méthode traditionnelle de collecte des données consiste à identifier les différentes sources bibliographiques susceptibles de contenir des séquences. D'une part, on dépouille systématiquement d'un certain nombre de revues scientifiques ayant publié un nombre jugé significatif de séquences originales sur la base du matériel antérieurement accumulé; d'autre part, les articles issus des publications n'appartenant pas à cette catégorie sont repérés d'après une indexation spécifique fournie par les grandes banques bibliographiques (*Current Contents, MedLine, Chemical Abstracts, ...*); des collaborations sont également établies avec d'autres projets catalographiques d'intérêt connexe tels que la *Human Gene Mapping Library* (**HGML**) afin de développer un système de références croisées.

---

[1] Le *Committee on Data for Science and Technology* est un des douze comités spécialisés créés par l'*International Council of Scientific Unions* (**ICSU**). Depuis sa création en 1966, le CODATA a encouragé la production et la diffusion de banques de données scientifiques et techniques d'abord dans les domaines de la physique et de la chimie puis de la géologie et plus récemment de la biologie 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, les différents travaux menés par ce groupe CODATA ont abouti à une proposition de norme (George, Mewes & Kihara 1987).

[2] La *Backus-Naur Form* est un métalangage utilisé en informatique comme formalisme permettant l'écriture des règles de syntaxe des langages *context-free*

Encadré 3: sequence d'une proteine capsidaire de geminivirus au format CODATA.

```
ENTRY            UCCUWU #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        A25005\ A24356
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 AAU for residue 5
                   as Arg
COMMENT          This virus is a member of the group Geminivirus
SUPERFAMILY      #Name maize streak virus coat protein
KEYWORDS         coat protein
SUMMARY          #Molecular-weight 29408   #Length 260   #Checksum 2085
SEQUENCE
               5         10        15        20        25        30
     1 A V T H K D S R D K D K R K M E E D E S S N R K G A V K
    31 R R K D A V K V V P V K P P R L D F R V H W L M S Q A T H
    61 I V Q D M T P R V D L I T C F A D D K R Q N R R H T M Q T V
    91 L V K F N I Q Q T C V N S D R S A P F I D P V R L V H W L V
   121 V D R E P K Q A M P Q A T D I F T M P N M L L P S T W T V Q
   151 A R V S H R F V V K R K W T V M L D T D Q R K V D S K T V Q
   181 D R V R D V V Q K H I V D R M K F F K G L S V T T E W R N T
   211 D D Q K I G D I K K G A L V L I S S T R Q D V T D D S R S T
   241 H F D V V C R V T H R C F K R D V I Q
..
```

Encadré 4 : corpus des sequences de geminivirus.

Acronyme	Phytovirus	Type	Vecteur	Numero d'acces	Source
ACMV-KE	African cassava mosaic virus	2c	Mouche-blanche	J02057	(15)
ACMV-NG	African cassava mosaic virus	2c	Mouche-blanche	X17095	(10)
BCTV-US	Beet curly top virus	1c	Cicadelle	M24597	(14)
BGMV-PR	Bean golden mosaic virus	2c	Mouche-blanche	M10070	(4)
BGMV-GA	Bean golden mosaic virus	2c	Mouche-blanche	n.e.	(9)
BGMV-DR	Bean golden mosaic virus	2c	Mouche-blanche	n.e.	(8)
CSMV-AU	Chloris striate mosaic virus	1c	Cicadelle	M20021	(1)
DSV-VU	Digitaria streak virus	1c	Cicadelle	n.e.	(2)
MSV-NG	Maize streak virus	1c	Cicadelle	X01633	(11)
MSV-ZA	Maize streak virus	1c	Cicadelle	J02514	(5)
MYMV-TH	Mung bean yellow mosaic virus	2c	Mouche-blanche	n.e.	(9)
SqLCV-US	Squash leaf curl virus	2c	Mouche-blanche	n.e.	(5)
TGMV-BR	Tomato golden mosaic virus	2c	Mouche-blanche	K02029	(3)
TYLCV-TH	Tomato yellow leaf curl virus	2c	Mouche-blanche	n.e.	(13)
TYLCV-IL	Tomato yellow leaf curl virus	1c	Mouche-blanche	n.e.	(12)
WDV-CS	Wheat dwarf virus	1c	Cicadelle	n.e.	(16)
WDV-SE	Wheat dwarf virus	1c	Cicadelle	X02869	(7)

Legende :

1c	1 composant genomique	2c	2 composants genomiques
c.p.	communication personnelle	n.e.	non-enregistre

Les numeros d'acces sont ceux de la banque EMBL

La provenance geographique du virus est indiquee par le code ISO du pays à deux lettres :

AU	Australie	BR	Brésil
CS	Tchecoslovaquie	DR	Republique dominicaine
GA	Guatemala	KE	Kenya
IL	Israel	NG	Nigeria
PR	Porto-Rico	SE	Suede
TH	Thailande	US	Etats-Unis d'Amerique
VU	Vanuatu	ZA	Afrique du Sud

Sources bibliographiques des sequences 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 [illegible]
(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

Cependant cette approche rencontre déjà ses limites pour de multiples raisons :

(i) l'extraction des séquences de la littérature scientifique est un processus intensif en travail humain spécialisé donc coûteux;

(ii) l'accroissement du nombre de séquences publiées et du nombre de revues publiant des séquences devrait rendre ce procédé impraticable à brève échéance;

(iii) l'annotation interprétative des séquences devrait relever de la responsabilité des auteurs et non de celle du gestionnaire des données;

(iv) beaucoup de séquences ne sont pas publiées faute de signification biologique manifeste

Le développement du processus de soumission directe des séquences par leur auteurs sous forme standardisée et sur support informatique constitue donc un objectif prioritaire pour les gestionnaires de ces banques afin, tout en diminuant le taux d'erreurs, de réduire les délais d'édition (désormais une semaine pour une soumission directe, 48 heures pour une soumission par courrier électronique) et faire face à la production induite par les grands projets de séquençage du génome envisagés (homme, levure, bactérie, riz). A cet effet, des accords ont été conclus entre ces banques et les principales revues scientifiques pour inclure la soumission directe dans le processus de révision par les pairs d'un article scientifique (e.g. accords de *Virology* avec GenBank, de *Protein Sequences & Data Analysis* avec la PIR, de *Nucleic Acids Research* et *Plant Molecular Biology* avec l'EMBL). Signalons que désormais 60% (EMBL) à 70% (GenBank) des soumissions sont effectuées directement.

4) Supports de diffusion

La communauté de chercheurs susceptibles de consulter ces banques de données se révèle numériquement importante : à titre d'exemple, sur la base d'une enquête effectuée en 1988 les responsables d'EMBL estimaient à près de 10 000 le nombre d'utilisateurs scientifiques.

Leur distribution s'effectue classiquement par bandes magnétiques (densité 1600 bpi). Signalons également des versions allégées ou des sous-ensembles (e.g. les virus) de certaines banques disponibles sur disquettes (aux différents formats 5 25" ou 3.5" pour compatibles IBM-PC ou MacIntosh) : les séquences, les annotations et les fichiers d'index y sont stockés en mode compressé, des programmes de décodage permettent la reconstitution des données en mode caractère; néanmoins la version 55 de GenBank (août 1987) comportait déjà 58 disquettes et il est probable qu'à terme, ce type de support sera abandonné en raison des limites imposées par la technologie d'enregistrement. Par contre, les procédés de lecture optique permettent une évolution rapide vers des formats haute-densité : la diffusion sur CD-ROM a récemment été introduite pour ce type de banques (c'est le cas de GenBank et d'EMBL).

Le téléchargement de ces banques, réalisable sur certains réseaux télématiques, n'est actuellement pas utilisé comme mode de distribution en raison du volume des données à transmettre qui aboutirait pour les réseaux longue-distance (dont les débits sont actuellement de l'ordre du kilobit par seconde) à des temps de transmission prohibitifs et à un risque de saturation du réseau. Cependant, l'accès en ligne aux mises à jour quotidiennes de ces bases de données, est possible dans le cadre d'un réseau européen spécialisé, EMBNet, par l'intermédiaire de centres serveurs nationaux ; sont actuellement opérationnels les noeuds EMBNet suivants : le *SERC Laboratory* (Daresbury au Royaume-Uni), le CITI2 (Paris en France), *COAS/CAMM Center* (Nijmegen aux Pays-Bas), Hoffman-La Roche (Basel en Suisse) et bien sûr EMBL (Heidelberg en Allemagne).

On trouvait encore récemment les principales séquences d'acides nucléiques éditées sous forme de livres (Gautier & al 1981, Anon 1986) mais le support papier est abandonné depuis 1987, date de la dernière édition imprimée de GenBank (Atencio & al 1987).

5) 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. Pour nos traitements locaux, l'UNCI ne disposant pas de site informatique propre, nous utilisons celui du Ministère de l'Economie et des Finances (**OCM** - Office Central de la Mécanographie, offrant les meilleures garanties au plan de la fiabilité), grâce à une liaison synchrone point à point (protocole BSC, 9600 bauds) avec un grappe de terminaux IBM en télétraitement et des ordinateurs personnels connectables en émulation 3278. 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).

Les adresses postales, téléphoniques et électroniques des banques de séquences et des différents serveurs, listées en annexe, permettront aux chercheurs intéressés de compléter leur documentation[3].

---

[3] Un guide documentaire intitulé *Banques de données moléculaires : inventaire des services existants* sera prochainement édité par l'Université Nationale de Côte d'Ivoire (Abidjan) à l'intention des chercheurs des pays francophones

III) CLASSIFICATION DES GEMINIVIRUS

1) Le groupe des géminivirus

1.1) Définition

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).

1.2) Intérêt économique

Depuis 1974, on a identifié une grande variété de géminivirus infectant des cultures importantes dans les régions tropicales, chaudes et tempérées (Harrison 1985), en particulier certaines plantes vivrières d'Afrique ou d'Asie :

- le virus de la mosaïque africaine du manioc (*african cassava mosaic virus* - **ACMV**) provoque des pertes allant jusqu'à 70% du poids du tubercule et son incidence est particulièrement élevée en Afrique (80% des plants sont infectés);
- avant l'introduction de cultivars résistants, le virus de l'enroulement apical de la betterave (*beet curly top virus* - **BCTV**) fut la cause principale de l'arrêt de toute industrie sucrière dans l'ouest des Etats-Unis d'Amérique;
- en Afrique, le **MSV** est à l'origine de pertes sévères, surtout en période de sécheresse, dans les champs de maïs;
- le virus de l'enroulement jaune de la tomate (*tomato yellow leaf curl virus* - **TYLCV**) est prévalent dans tout le Moyen-Orient et en Afrique. Les plants de tomate ensemencés à l'automne subissent des pertes variant de 50 à 75%;
- le virus de la mosaïque jaune du pois d'Angola (*mungbean yellow mosaic virus* - **MYMV**) génère des dommages importants dans les cultures de pois en Thaïlande ou en Inde;
- les cultures de coton et de courges sont sérieusement endommagées dans le sud-ouest des USA par des géminivirus transmis par mouches-blanches, respectivement le virus du rabougrissement du coton (*cotton leaf crumple virus* - **CLCV**) et le virus de l'enroulement de la courgette (*squash leaf curl virus* - **SqLCV**).

1.3) Intérêt scientifique

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 l'**ICTV**[4] (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, avec k(i,j') défini par 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, avec k(i,j") défini par 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[5], avec k(i,j'") défini par le nombre d'occurrences du codon j'" au sein de la séquence i.

Ces trois tableaux de contingence nous permettent d'analyser les profils respectifs des différentes séquences pour leur composition en acides aminés, en dinucléotides et en codons.

4) Les méthodes d'analyse

4.1) L'analyse factorielle des correspondances

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

---

[4] 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 tous les trois ans une nomenclature permettant à chaque virologiste de disposer d'une base cohérente et actualisée de classification.

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

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 en utilisant le programme (ANAFAC CORR) de la librairie ADDAD[6].

4.2) La classification ascendante hiérarchique

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 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 en utilisant 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

---

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

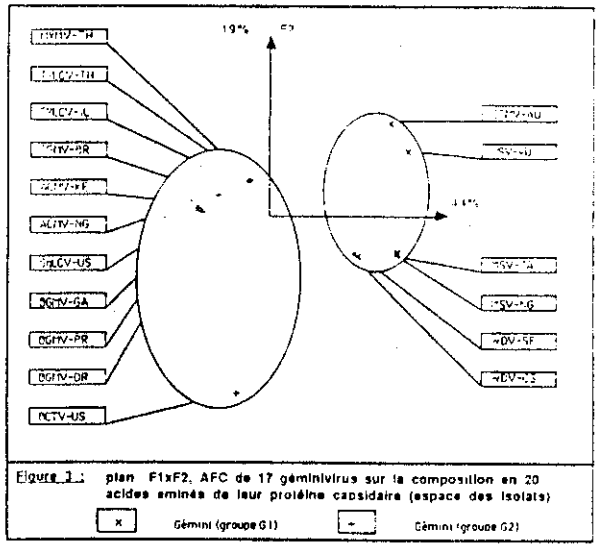

Figure 3 : plan F1xF2, AFC de 17 géminivirus sur la composition en 20 acides aminés de leur protéine capsidaire (espace des isolats)

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

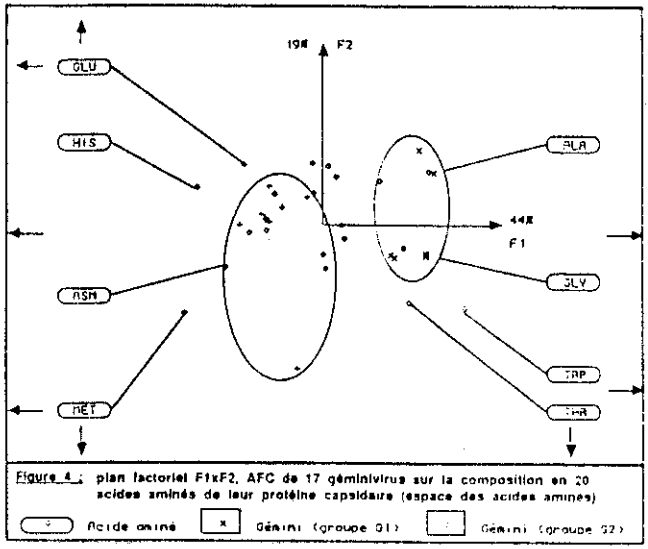

Figure 4 : plan factoriel F1xF2, AFC de 17 géminivirus sur la composition en 20 acides aminés de leur protéine capsidaire (espace des acides aminés)

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

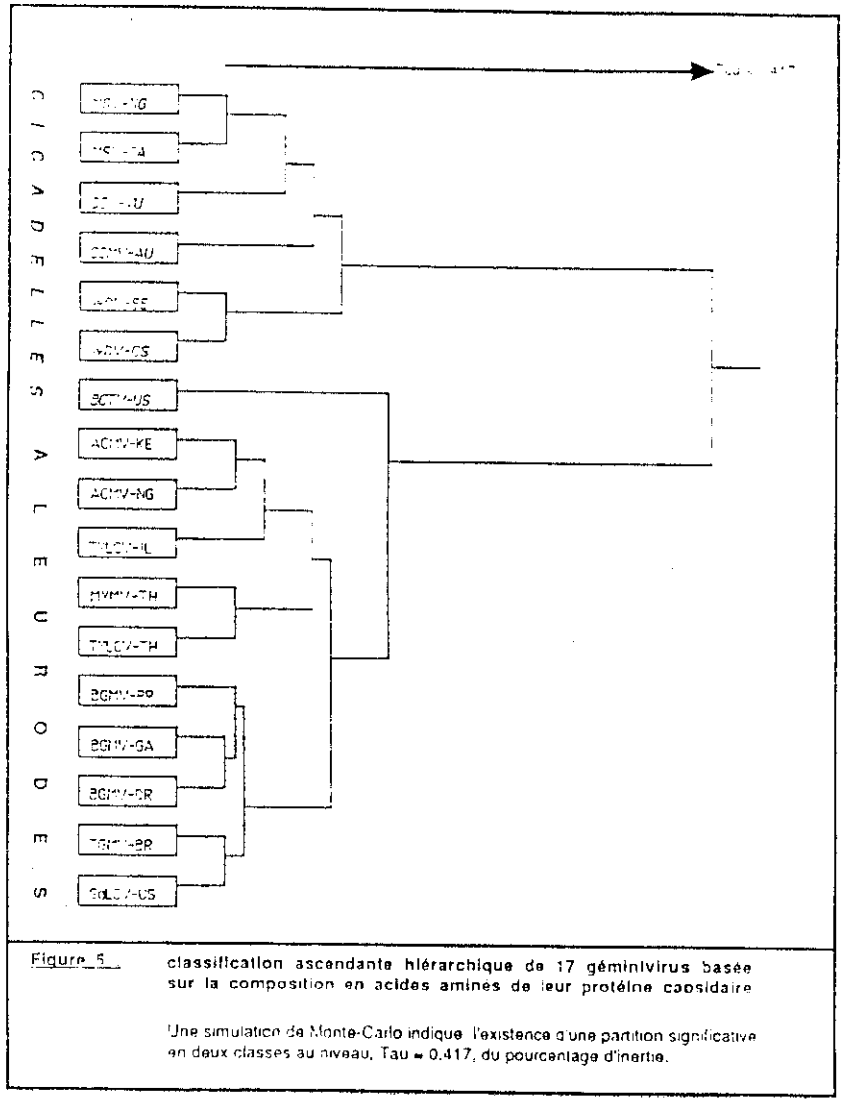

Figure 5 classification ascendante hiérarchique de 17 géminivirus basée sur la composition en acides aminés de leur protéine capsidaire

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

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**[7] 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%).

L'examen des axes de rang supérieur, dont le pourcentage d'inertie décroît rapidement, ne permet pas d'apporter d'autres conclusions, le niveau de l'indice de corrélation des projections n'autorisant pas 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.

---

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

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 troisième 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

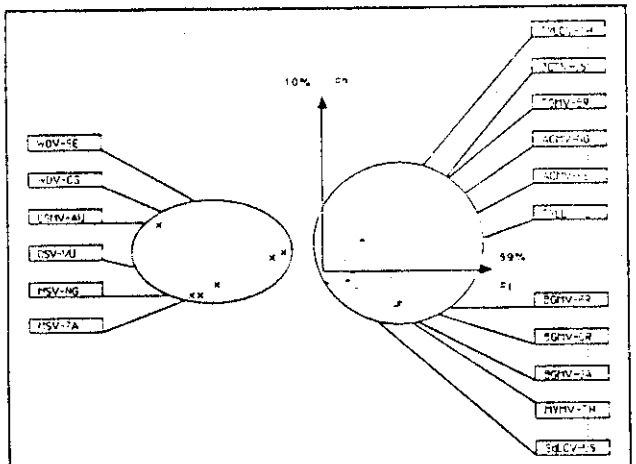

Figure 6 : plan factoriel F1xF2, AFC de 17 géminivirus sur la composition en 16 dinucléotides de leur protéine capsidaire (espace des isolats)

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

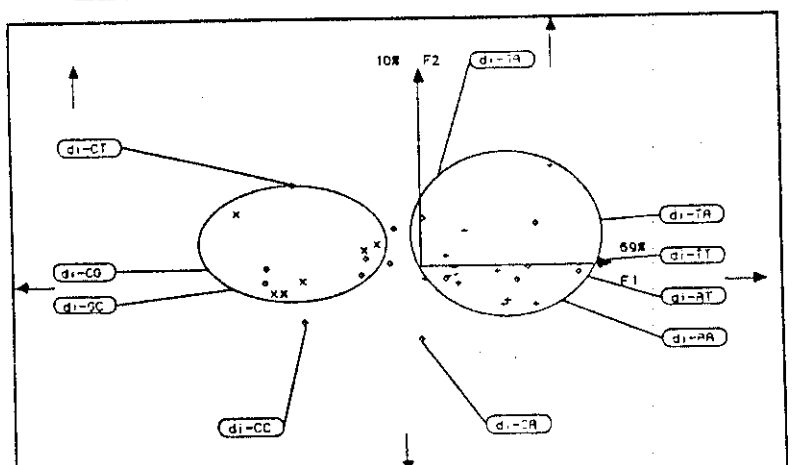

Figure 7 : plan factoriel F1xF2, AFC de 17 géminivirus sur la composition en 16 dinucléotides de leur protéine capsidaire (espace des dinucléotides)

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

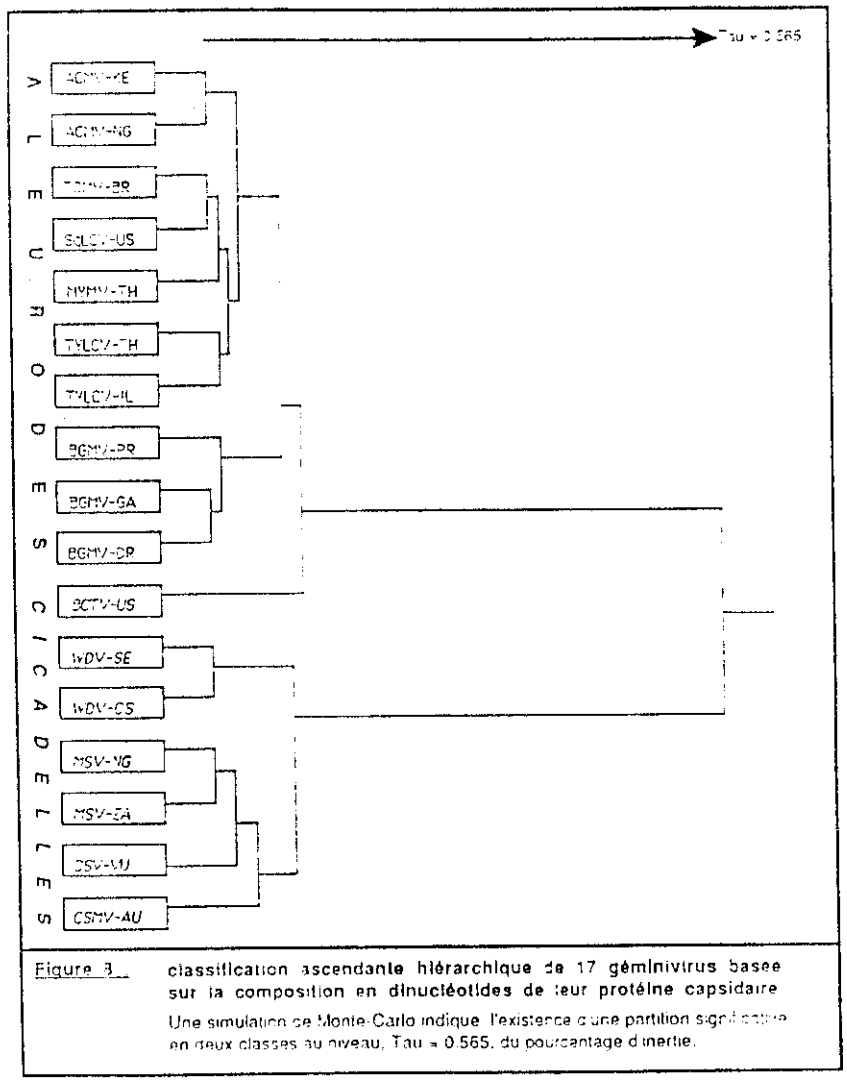

Figure 8 : classification ascendante hiérarchique de 17 géminivirus basée sur la composition en dinucléotides de leur protéine capsidaire

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

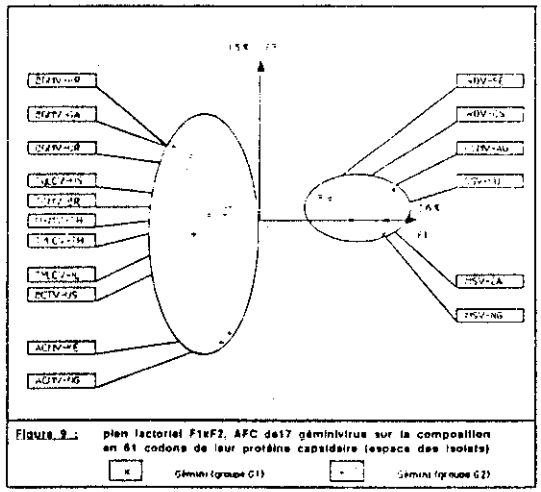

Figure 9 : plan factoriel F1xF2, AFC de17 géminivirus sur la composition en 61 codons de leur protéine capsidaire (espace des isolats)

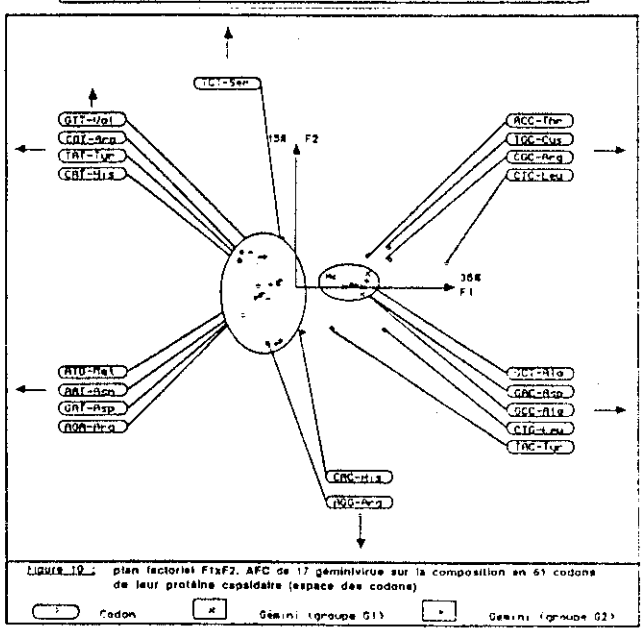

Figure 10 : plan factoriel F1xF2, AFC de 17 géminivirus sur la composition en 61 codons de leur protéine capsidaire (espace des codons)

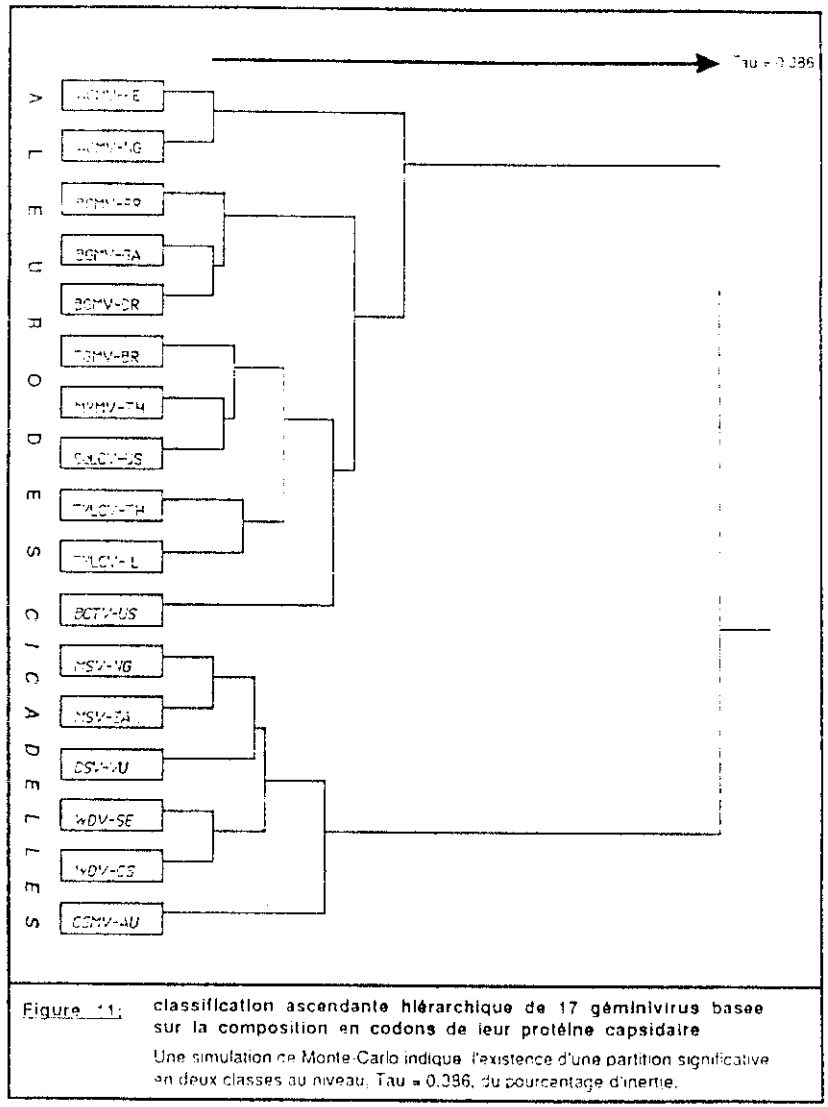

Figure 11 : classification ascendante hiérarchique de 17 géminivirus basée sur la composition en codons de leur protéine capsidaire

Une simulation de Monte-Carlo indique l'existence d'une partition significative en deux classes au niveau Tau = 0.396, du pourcentage d'inertie.

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. Afin de vérifier cette hypothèse, il conviendrait néanmoins de confirmer ce résultat sur la base d'un corpus de souches et de géminivirus plus important, en essayant de l'étendre à 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'analyse factorielle des correspondances, 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 :

Anon (1986). Nucleotide sequences 1986. Oxford, IRL Press.

Atencio, E. J. & al. (1987). Nucleotide sequences 1986/1987 : a compilation from the GenBank and EMBL data libraries. Orlando, Floride, Academic Press.

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

Gautier, C., Gouy, M., Jacobzone, M. & Grantham, R. (1981). Nucleic acid sequences handbook. Londres, Praeger Publishers

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

Harrison, B. D. (1985). Advances in geminivirus research. Ann. Rev. Phytopathol. 23 : 55-82.

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.

Annexe : banques de séquences, logiciels et services

(I)

A) BANQUES DE SEQUENCES NUCLEIQUES

* DNA/JDB - National Institute of Genetics
-information : <P> = < c/o Hisao Uchida, Teikyo University 2-11-1, Kaga, Itabashi-ku TOKYO 173 JAPAN>
* EMBL - European Molecular Biology Laboratory
-sousmission : <P> = < Meyerhofstrasse 1, Postfach 10 2209, 6200 HEIDELBERG, RFA>
<E> = <DATASUB@DHDEMBL EARN>
-souscription : <P> = < Meyerhofstrasse 1, Postfach 10 2209, 6200 HEIDELBERG, RFA>
<E> = <DATALIB@DHDEMBL EARN>
* GenBank - Genetic Sequence Data Bank
-sousmission : <P> = < Genbank/T-10,MS K170, Los Alamos National Laboratory, LOS ALAMOS, NM 87545, USA>
<E> = <WBG@LAN ARPA BITNET,gb-sub%life@lanl gov, genbank%life@lanl gov>
-souscription : <P> = < GenBank, IntelliGenetics, 700 El Camino Real East MOUNTAIN VIEW CA 94040 USA>
<E> = <genbank@bionet-20 arpa, genbank @ genbank ig com>
<T> = <415-962-7364>
* MIPS/MPIB - Martinsrieder Institüt fur Proteinsequenzen
-information : <P> = < am Max - Plank - Institute fur Biochemie, Am Klopferspitz, D 8033 MARTINSRIED bei München, RFA>
<E> = <MEWES@DMONPB51 EARN> <T> = <089-8578-2656>

B) BANQUES DE SEQUENCES PROTEIQUES

* NBRF-PIR National Biomedical Research Foundation - Protein Identification Resource
-information : <P> = < Georgetown University Medical Center for Molecular Genetics, University of California at San Diego, LA JOLLA, CA 92 093, USA>
<E> = <NBRF@GUVM BITNET,PIRSUB@GUNBRF BITNET, DIALCOM: 42: CDT0105>
<T> = <202-687-2121> <X> = <4909975103>

C) LOGICIELS

1)sur sites centraux ou départementaux
* UWGCG - University of Wisconsin Genetics Computer Group
-information : <P> = <-c/o John Devereux, Laboratory of Genetics, University of Wisconsin, WI 53706, USA;
-University of Wisconsin Biotechnology Center, 1710 University Avenue, University of Wisconsin, MADISON, WI 53705 USA>
<E> = <NBRF@GUVM BITNET,PIRSUB@GUNBRF BITNET, DIALCOM: 42: CDT0105>
<T> = <608-263-8970>
* LMB/MRC - Laboratory of Molecular Biology / Medical Research Council
-information : <P> = < Hills Road, CAMBRIDGE CB2 2HQ, ENGLAND UK>
<E> = <RS@UK.AC.CAM.MRC-LMB>
<F> = <0223-213556> <T> = <0223-248011>
<X> = <81532>
* GenBank Software Clearing House
-information : <P> = < Los Alamos National Laboratory, LOS ALAMOS, NM 87545, USA>
<E> = <WBG@LAN ARPA BITNET>

2)sur micro-ordinateurs ou stations de travail
* DNASIS - DNA Sequence Input and analysis System
-information : <P> = <-Pharmacia Ltd, Pharmacia House, Midsumer Boulevard, MILTON KEYNES MK9 3HP, UK;
-Hitachi Software Engineering Co Ltd., 6-81, Onoemachi, Nakaku, YOKOHAMA 231 JAPAN>
-Hitachi Software Engineering Co Ltd., 950 Elm Avenue SAN BRUNO, CA 94006, USA>
<F> = <UK 0908-690091> <T> = <UK 0908-661161>
<X> = <UK 826778 PH GBG>

" DNASTAR - Comprehensive Microcomputer System for Molecular Biology (II)
-information : <P> = <-Dnastar Ltd , 8 Walpole Gardens LONDON W4 4HG UK;
-Dnastar Inc , 1801 University Avenue MADISON
WI 53705 USA>
<F> = <UK 01-747-4748> <T> = <UK 01-994-0619; USA 608-233-5525>
<X> = <UK 826778 PH GBG>

• The IBI/PUSTELL DNA and Protein Sequence Analysis System
-information : <P> = <-IRL Press Ltd, P.O Box 1, Eynsham, OXFORD OX8 1JJ, UK;
-IRL Press Inc., P.O. Box Q, McLEAN, VA 22101, USA>
<F> = <UK 0865-882890; USA 703-689-0660>
<T> = <UK 0865-882283; USA 608-233-5525>
<X> = <UK 83147 IRL USA 888579>

• The MICROGENIE Sequence Analysis Program Analysis System
-information : <P> = < Beckman - RIICLTD Progress Road, Sands Industrial Estate,
High Wycombe Buckinghamshire HP12 4SL, UK>
<T> = <494-41181>

• PC/Gene
-information : <P> = < GENOFIT S.A., Case Postale 239, 1212 Grand-Lancy, GENEVA,
SWITZERLAND>
<F> = <022-71-31-28> <T> = <022-71-44-44>
<X> = <423 082 gfit ch>

• LGBC - Laboratoire de Génétique et de Biologie Cellulaire
-information : <P> = < c/o B. Bellon CNRS Case 907, Faculté de Luminy,
13288 MARSEILLE cedex 09, FRANCE>

• DNA Strider - CEA Commissariat à l'Energie Atomique
-information : <P> = < c/o C. Marck, Service de Biochimie - Bat 412, CEN-Saclay
91191 GIF-SUR-YVETTE cedex, FRANCE>

D) SERVICES EN LIGNE

• BIONET
-information : <P> = < c/o D. Kristofferson, BIONET Resource Manager, 700 El Camino Real East
MOUNTAIN VIEW, CA 94040, USA>
<E> = <kristofferson@bionet-20 arpa>

• PROHET
-information : <P> = < c/o R. DuBois, Building 31, Room 5B-43, National Institute of Health,
BETHESDA, MD 20892 USA>
<E> = < 301-496-5411>

• MBCCR - Molecular Biology Computer Resource
-information : <P> = < c/o S Tolman, User Coordinator, Dana-Farber Cancer Institute BOSTON MA
USA>

• UCSS - University of Cambridge Computer Resource
-information : <P> = < c/o External Receptionist, Computer Laboratory, University of Cambridge,
Corn Exchange Street, CAMBRIDGE, CB2 3QG, UK>

" EUS- Edinburgh University Service
-information : <P> = < c/o A Coulson, Department of Molecular Biology, King s Building, Mayfield Road,
EDINBURGH, EH9 3JR, UK>

• CITI2 - Centre Interuniversitaire de Traitement de l'Information 2
-information : <P> = < c/o C. Fondrat, CITI2, 45 rue des Saints-Pères, 75006 PARIS FRANCE>

<E> = <FONDRAT@FRCITI51.EARN ; TRANSPAC (3600) 175001236 ;
TRANSPAC (3601) 175001236 ;
RTC (1) 42-96-34-97, (1) 42-97-49-06>
<T> = <(1)42-96-24-89>
<X> = <CITIUM 670602F>

ADRESSES : <P> = <postale>; <E> = <électronique>; <F> = <Fax>; <T> = <Téléphone>; <X> = <Telex>

extrait de *Banques de données en biologie moléculaire : inventaire des services existants* , D. DESBOIS & G VIDAL, Université Nationale de Côte d Ivoire, Abidjan, 1990 (sous presse, ISBN 2-7166-0351-0)