Introduction à la bioinformatique

Introduction à la Bioinformatique

1. Introduction

Où ne trouve-t-on pas maintenant l'utilisation de "l'informatique" en biologie : pilotage d'appareils expérimentaux, archivage de données, traitement de données, analyse de séquences, prédictions sur celles-ci, etc..

Toutefois, les exégètes, que ce soit par une juste perspicacité ou un snobisme effréné, réservent le mot bioinformatique, qui émerge dans les années 1990, à une nouvelle discipline, fusion des disciplines de biologie, informatique et traitement de l'information (on peut se demander pourquoi la "*physinformatique*" et la "*mathinformatique*" n'existent point).

Dans ces quelques pages d'introduction, nous nous intéresserons essentiellement au traitement de l'information des séquences biologiques pour les points particuliers suivants :

- banques, bases de données de séquences
- la question de la ressemblance entre séquences
- analyse et prédiction sur les séquences

2. Quelques notes historiques

Voici une brève promenade historique le long de quelques évènements biologiques ou informatiques :

1646 : *Blaise Pascal invente une machine ("La Pascaline") capable d'effectuer des additions et des soustractions afin d'aider son père, collecteur d'impôts à Rouen.*

1673 : *Gottfried Wilhelm von Leibniz construit une machine effectuant automatiquement les additions, soustractions, multiplications et les divisions.*

1812 : *Charles Babbage, professeur de mathématiques, réalise les plans d'une machine capable d'exécuter n'importe quelle séquence de calculs au moyen d'une cinquantaine de roues dentées qui étaient activées grâce à des instructions lues sur une carte perforée.*

1840 : *Collaboratrice de Charles Babbage et fille du poète Lord Byron, Ada Lovelace, mathématicienne, définit le principe des itérations successives dans l'exécution d'une opération. En l'honneur du mathématicien Arabe Al Khowarizmi (820), elle nomme le processus logique d'exécution d'un programme : algorithme.*

1854 : *George Boole pose les axiomes et règles de l'algèbre booléenne, fondement des ordinateurs à arithmétique binaire*.

1858 : *Premier câble télégraphique transatlantique*.

1866 : Gregor Mendel publie ses lois de l'hérédité à partie d'études menées chez le Pois.

1896 : *Herman Hollerith crée la Tabulating machine et fonde une compagnie, qui deviendra IBM*.

1901 : De Vries redécouvre expérimentalement les lois de Mendel et publie "La théorie de la mutation".

1903 : Walter S. Sutton (1903) et Boveri (1904) proposent pour la première fois d'associer les gènes au chromosome qui deviennent ainsi supports de l'hérédité.

1909 : Wilhem Johannsen dénomme "gènes" les particules de l'hérédité proposées par Mendel puis redécouvertes par de Vries.
- Archibald Garrod propose la relation un gène-une enzyme à partir de l'étude d'une anomalie métabolique humaine: l'alcaptonurie (déficit en acide homogentisique-oxydase sur la voie du catabolisme de la tyrosine).

1913 : Thomas Morgan et Alfred Sturtevant publient la première carte génétique du chromosome X avec la position respective de 3 gènes évaluée par le pourcentage de recombinaison (phénomène de crossing-over).

1915 : Thomas Morgan publie avec Sturtevant, Muller et Bridge: "Le mécanisme de l'hérédité mendélienne" .

1927 : Hermann Muller met au point l'induction artificielle de mutations par les rayons X.

1928 : Fred Griffith fait les premières expérience de la transformation bactérienne.

1930 : *Georges Stibitz construit un additionneur binaire,appelée "Calculateur de Nombres Complexes" , en s'appuyant sur les idées de Georges Boole*.

1931 : *Konrad Zuse construit, le Z1 : premier calculateur digital électromécanique*.

1935 : Max Delbrück étudie le gène par le biais de l'effet induit par des rayonnements sur celui-ci. Il fonde le Groupe du phage, avec Salvador Luria et Alfred Hershey six ans plus tard.

1936 : *Alan Turing définit le concept de la machine de Turing et de là les notions de fonctions calculables*.

1940 : *Alan Turing parvient à décrypter le code Enigma utilisé par l'Amirauté du Reich pour communiquer avec ses sous-marins sillonnant l'Atlantique*.

1941 : George Wells Beadle et Edward Tatum établissent la relation un "gène-une enzyme" chez Neurospora crassa.

1944 : Oswald Avery démontre avec Colin McLeod et McLyn McCarthy que l'ADN transporte l'information génétique responsable de la transformation bactérienne.
- Erwin Schrödinger introduit la notion de programme et de code génétique.
- Howard Aiken termine la construction du Mark I : 1er ordinateur électronique à programme interne (à registre).

1946 : *L'annonce de l'ENIAC (Electronic Numerical Integrator and Computer) par J. Presper Eckert, marque le début de l'histoire moderne des calculateurs*.

1947 : Le DOE (agence fédérale responsable des programmes nucléaires aux Etats-Unis) s'engage dans les recherches génétiques.
- John Mauchly, J.P. Eckert, et John von Neumann travaillent à la conception d'un ordinateur électronique, l'EDVAC (Electronic Discret VAriable Computer) : 1er calculateur à programme enregistré. C'est le descendant direct de l'ENIAC (capacité mémoire est de 1024 mots de 44 bits).

1948 : *Claude Shannon publie "Une théorie mathématique de la communication" et est à l'origine de la théorie de l'information).*

1949 : *John Mauchly présente "Short Order Code", le premier langage de programmation.*
EDSAC (Electronic Delay Storage Automatic Computer) : 1er ordinateur numérique et électronique basé sur l'architecture de John von Neumann.

1950 : *Alan Turing publie le Test de Turing, pour définir l'IA (intelligence artificielle) d'une machine*.

1951 : *William Shotkley met au point le transistor*.

- Le bureau de la statistique US reçoit le premier UNIVAC (UNIversal Automatic Computer) (1000 instructions/s) : 1er ordinateur commercialisé. Il utilise des bandes magnétiques en remplacement des cartes perforées. (UNIVAC Memories)

1952 : Alfred Day Hershey et Chase démontrent que les bactériophages injectent leur ADN dans les cellules hôtes (corrélation entre l'ADN et l'information génétique).

1953 : James Watson, Francis Crick et Maurice Wilkins (prix Nobel) découvrent la structure en double hélice de l'ADN.
- Début de l'IBM 650, le premier ordinateur "commercial".

1954 : *Suicide d'Alan Turing : il croque une pomme remplie de cyanure, suite à une inculpation pour "mœurs controversées".*

1956 : Frédérick Sanger établit la séquence en acides aminés de l'insuline.
- Vernon Ingram montre qu'une mutation liée à une altération héréditaire de l'hémoglobine se traduit par un changement d'un unique acide aminé dans la protéine.
- Création de FORTRAN,premier langage procédural de haut niveau , par John Backus & al. d'IBM.

1959 : *Annonces de l'IBM 1401 (tout transistor).*

1960 : *DEC présente le PDP1, premier ordinateur commercial avec écran/clavier.*

1961 : Marshall Nirenberg et J.Heinrich Matthaei déchiffrent le code génétique.

1962 : *Atlas, Manchester University, premier ordinateur à mémoire virtuelle.*

1964 : *Annonce du IBM/360 : ordinateur de 3e génération.*
CDC 6600 par Seymour Cray, premier supercomputer (9 MFLOPS : 9 millions d'opérations par seconde).

1965 : Jacques Monod, François Jacob et André Wolf (prix Nobel) découvrent les mécanismes de la régulation génétique impliqués dans le dogme central de la biologie moléculaire, énoncé initialement par Crick.
- Théorie de l'horloge moléculaire (Zuckerkandl & Pauling).
- Atlas of Protein Sequences : première compilation de protéines (M. Dayhoff, Georgetown).
- PDP8 (Programmed Data Processor) de DEC : 1er mini-ordinateur diffusé massivement (> 50000 exemplaires).

1967 : "Construction of Phylogenetic Trees" (Fitch & Margoliash).
- Début des circuits intégrés CMOS (voir aussi : Circuits intégrés logiques).

1968 : *Annonce par Seymour Cray du CDC 7600 (40 MFLOPS : 40 millions d'opérations par seconde).*

1969 : *Premières interconnexions ARPANET (réseau).*

1970 : Programme d'alignement global de séquences (algorithme de Needleman & Wunsch).
- Ken Thompson & Dennis Ritchie développent UNIX aux Bell laboratories.

1971 : *Annonce du microprocesseur INTEL 4004 : 1er microprocesseur.*

1972 : Clonage de fragments d'un plasmide bactérien dans le génome du virus SV40 (Paul Berg, David Jackson, Robert Symons)
- Annonce du Cray 1, crée par Seymour CRAY (cf. interview, 1996): 1er super-ordinateur à architecture vectorielle.

1973 : Découverte des enzymes de restriction.
- Obtention d'une méthode fiable de transfection (introduction d'un ADN étranger) des cellules eucaryotes grâce à un virus (vecteur). (Franck Graham et Alex Van der Eb).
- Développement de l'ALTO de Xerox suite aux recherches démarrées en 1970. Ce prototype, pensé pour devenir le bureau du futur, est le premier à introduire l'idée de fenêtres et d'icônes que l'on peut gérer grâce à une souris. Il ne sera introduit sur le marché qu'en 1981 sous le nom de Star 8010 qui connaîtra un échec commercial total.

1974 : Création d'un Comité sur l'ADN recombinant, présidé par Paul Berg (Université de Stanford, Californie), appelant la communauté scientifique à un moratoire sur les expériences de recombinaison génétique.
- Programme de prédiction de structures secondaires des protéines (Chou & Fasman).

1975 : *MITS Altair 8080 : 1er ordinateur personnel (commercialisé en kit).*
- Conférence internationale d'Asilomar (Californie), organisée par Paul Berg et ses collègues sur le risque génétique.
- Mise au point de la technique "Southern blot"

1976 : *Le Cray 1 atteint 138 MFLOPS (138 millions d'opérations par seconde).*

1977 : Frédérick Sanger met au point la méthode de Sanger pour établir le séquençage.
Premier ensemble de programmes sur l'analyse des séquences (Staden).

- *Création d'Apple Computer (Apple II) et de Microsoft.*

1978 : Mutagenèse dirigée. (Michael Smith)
- Séquençage du premier génome à ADN, le bactériophage phiX174 (5386pb) (Frederick Sanger)
- *Annonce du VAX 11/780 : premier super-mini-ordinateur.*

1979 : *Début de USENET, échanges de émail et Newsgroups.*

1980 : David Botstein et Ronald Davis introduisent les marqueurs moléculaires, notamment, les RFLP.
- Découverte de la technique de FISH (hybridation in situ sur chromosome), technique notamment utile dans la construction des banques génomiques (identification d'un fragment d'ADN sur un chromosome)
- Création de la banque EMBL : banque européenne généraliste de séquences nucléiques créée à Heidelberg et financée par l'EMBO (European Moleculary Biology Organisation). Elle est aujourd'hui diffusée par l'EBI (European Bioinformatics Institute, Cambridge, GB)

1981 : *IBM-PC (8088), 16-32kb : 1er IBM-PC (PC-DOS)*
- Programme d'alignement local de séquences (algorithme de Smith et Waterman)
- Extension de l'algorithme de Needleman et Wunsch au problème de recherche de similitude locale.
- Naissance du 1er animal transgénique (une souris) (Franck H.Ruddle et John W. Gordon)
- Découverte des oncogènes humains

1982 : Création de la banque Genbank : banque américaine généraliste de séquences nucléiques créée par la société IntelliGenetics et diffusée aujourd'hui par le NCBI (National Center for Biotechnology Information, Los Alamos, US).
- *Annonce de Internet (TCP/IP).*

1983 : Barbara McClintock découvre les éléments mobiles génétiques (transposons) chez les plantes.
- *IBM-XT Disque dur (10 Mbytes = 10 Moctets).*

1984 : Développement de la réaction de polymérisation en chaîne par Mullis de la PCR : outil devenu indispensable tant en recherche appliquée que fondamentale : séquençage génomique et cartographie, diagnostic génétique, analyse de l'expression des gènes ...
- Création de la banque NBRF : banque américaine généraliste de séquences protéiques créée par la NBRF (National Biomedical Research Foundation).
- *Commercialisation du LISA et du premier Macintosh*

1985 : ACNUC, un des premiers logiciel d'interrogation des banques, a été développé et est maintenu à Lyon.
- Programme Fasta (Pearson- Lipman) : recherche rapide d'alignements locaux dans une banque.
- Publication du 1er article relatant l'utilisation de la PCR.
- L'idée de décrypter les trois milliards de bases du génome humain naît pour la 1ère fois à l'Imperial Cancer Research (ICR) de Londres..
- *Annonce du Cray 2 à un GIPS*.

1986 : Création de la banque DDBJ : banque japonaise généraliste de séquences nucléiques créée par le NIG (National Institute of Genetics, Japon).
- Création de la banque SwissProt : banque généraliste de séquences protéiques créée à l'Université de Genève et maintenue depuis 1987 dans le cadre d'une collaboration, entre cette université (via ExPASy, Expert Protein Analysis System) et l'EBI.
- Le DOE propose de créer des centres du génome pour s'atteler au séquençage du génome humain
- Clonage du gène responsable de la myopathie de Duchenne

1987 : Réalisation et commercialisation du premier séquenceur automatisé par la société Applied Biosystems (Californie).
- Mise au point d'un nouveau vecteur : le YAC (Yeast Artificial Chromosome), premier vecteur permettant de cloner des fragments d'ADN 20 fois plus grands que les plasmides utilisés jusqu'alors.
- Publication de la 1ère carte génétique du génome humain
- Apparition de la technologie des puces à ADN

1988 : Création du projet HUGO (Human Genome Organization) pour coordonner les efforts de cartographie et de séquençage entrepris dans le monde et éviter les doublons.

1989 : *INTERNET succède à ARPANET et BITNET*.
- Découverte des marqueurs microsatellites.
- Découverte du système double hybride permettant d'étudier dans des cellules de levure (ou d'*Escherichia Coli*) l'interaction entre deux protéines hybrides fusionnées à des facteurs de transcription..

1990 : Programme Blast (Altschul et al.) : recherche rapide d'alignements locaux dans une banque.
- Premier essai de thérapie génique.
- Création du 1er Généthon.

- *Tim Bernes-Lee développe le prototype du WEB.*

1991 : Programme Grail (Mural et al.) : localisation de gènes.

1992 : Fondation du Centre de recherche SANGER par le Welcome Trust et le British Medical Research Concil (Cambridge, UK). C'est le centre le plus productif des instituts public de séquençage : il réalise la moitié de la "production" mondiale.
- Publication de la 2e carte génétique du génome humain, établie par le Généthon à partir de 814 fragments génomiques (marqueurs choisis : microsatellites - résolution : 4,4 cM).

1993 : Etzold et Argos créent SRS, logiciel d'interrogation multibanques accessible sur le web

1994 : Publication de la 4e carte génétique du génome humain, établie par le Généthon à partir de 2066 fragments génomiques (marqueurs choisis : microsatellites - résolution : 2,9 cM).
- *Succédant au navigateursLynx et NCSA, Netscape Navigator est disponible.*

1995 : Séquençage de la 1ère bactérie, Haemophilus influenzae (1,83 Mb) (Fleischmann). Séquenceur à capillaire qui a conduit à augmenter les performances des laboratoires d'un facteur dix entre 1995 et la fin de 1997, et d'un nouveau facteur dix à la fin du siècle.

1996 : Séquençage du 1er génome eucaryote, *Saccharomyces cerevisiae* (12 Mb) (Dujon).

1998 : Séquençage du 1er organisme pluricellulaire, *Caenorhabditis elegans* (100 Mb) .

2000 : Séquençage du 1er génome de plante, *Arabidopsis thaliana*
- *ASCI White (RS/6000) : IBM construit le premier superordinateur qui dépasse les 10 TERAFLOPS (dix mille milliards d'opérations par seconde).*

2001 : Annonce du décryptage presque complet du génome humain. (Février) : les travaux de la compagnie américaine privée Celera Genomics et du projet public international Génome Humain (HGP pour Human Research Project) sont sur les sites Internet des deux revues Science et Nature.

2.1. L'implication du Web et du WYSIWYG sur la bioinformatique

Des années 1960 à 1990, l'utilisation de l'informatique par les biologistes se faisait sur des consoles reliées à une machine centrale (un serveur) où étaient implantés les programmes, les banques de données : le serveur devait posséder toutes les ressources nécessaires.

La mise en place d'Internet a permis de développer la mise en commun de ressources, partagées par les biologistes, que ce soit par échange de courrier, de données (ftp) ou encore par l'utilisation à distance d'un serveur. Toutefois, cela obligeait l'utilisateur à apprendre le "jargon" informatique des commandes de bases des systèmes d'exploitation ainsi que les commandes particulières pour chaque programme qu'il veut utiliser.

L'arrivée concomitante des interfaces graphiques des systèmes d'exploitation et des logiciels (WYSIWYG : What You See Is What You Get), du réseau Internet et la création du Web par Tim Bernes-Lee (Ingénieur au CERN) a profondément modifié non seulement l'utilisation de la bioinformatique, mais aussi sa conception et son développement.

2.1.1. Architecture client-serveur

De nombreuses applications fonctionnent selon un environnement client/serveur, cela signifie que des machines clientes (des machines faisant partie du réseau) contactent un serveur qui leur fournit des services. Ces services sont des programmes fournissant des données telles que l'heure, des fichiers, une connexion, .. Les services sont exploités par des programmes, appelés programmes clients, s'exécutant sur les machines clientes. On parle ainsi de client ftp, client de messagerie, client Web, etc …

- Le client émet une requête vers le serveur grâce à son adresse et le **port**, qui désigne un service particulier du serveur. C'est toujours le client qui déclenche une demande de service.
- Le serveur reçoit la demande et répond à l'aide de l'adresse de la machine client et son port. Le serveur attend passivement les requêtes des clients (port d'écoute) et peut traiter plusieurs requêtes en même temps.

Par exemple, le numéro de port du service ftp est le 21, du service smtp (mail) 25 : pour les serveurs Web, le port par défaut est 80.

2.1.2. Le Web

C'est le dernier système d'échanges d'information du modèle client/serveur, protocole http (*HyperText Transport Protocol*) et qui peut se définir comme un système d'information réparti hypermédia. Le serveur gère et héberge l'information sous forme de fichiers. Les documents fournis par le serveur sont dans le format HTML (*HyperText Markup Langage*) qui **contient non seulement du texte, des images mais aussi des liens (*référence de type url*) vers d'autres fichiers ou d'autres serveurs**.

Les adresses des serveurs que vous indiquez à votre logiciel client sont normalisées (URL : *Uniform Resource Locator*) de la forme :
<protocole>://<adresse machine>[port]/<reference locale>

- *protocole* : http, ftp, gopher, news
- *adresse* : machine nom symbolique ou adresse I.P.
- *port* : numéro du service (si aucun port n'est indiqué, c'est le numéro prédéfini qui est utilisé par exemple pour WWW : 80)
- *reference locale* : désigne un fichier ou répertoire qui sont définis par rapport au type de serveur

Pour le cas particulier du protocole "http", il existe une méthode CGI (*Common Gateway Interface*) qui permet d'incorporer des programmes s'exécutant au sein d'un serveur Web, sur requête du navigateur du client. Ceci est largement utilisé pour mettre à disposition de la communauté scientifique des programmes bioinformatiques qui s'exécutent sur le serveur et dont le client reçoit le résultat dans le format html.

2.1.3. Délocalisation des ressources

Les première ressources disponibles pour les biologistes étaient soit sur le même serveur ou sur de serveurs différents mais elles obligeaient l'utilisateur à jongler avec celles-ci et les logiciels.
Le protocole "http" qui permet :

1) d'incorporer des programmes par un appel à partir d'une "forme" (ou formulaire) dans une page html et ce de manière totalement transparente pour l'utilisateur
2) d'incorporer des liens (URL) dans les pages html

a induit une délocalisation complète des ressources. Par exemple, un serveur d'une banque de séquences biologiques peut très bien envoyer à l'utilisateur des informations qui

- sont le résultats de calculs exécutés sur un autre serveur
- qui sont des liens spécifiques vers d'autres banques de données bibliographiques ou de séquences ou autres (voir les références croisées)

Bien évidemment, pour qu'un tel système fonctionne correctement et soit transparent pour l'utilisateur, il faut un minimum d'entente entre les organismes qui mettent des informations disponibles sur leurs serveurs (URL fixe).

3. Banques et bases de données biologiques

Souvent les termes de banque ou base sont utilisées sans distinction particulière. Toutefois il existe une différence non seulement pour l'utilisateur mais aussi pour l'implantation informatique de ces dernières :

Banque de données : ensemble de données relatif à un domaine défini des connaissances et organisé pour être offert aux consultations d'utilisateurs

Base de données : ensemble de données organisé en vue de son utilisation par des programmes correspondant à des applications distinctes et de manière à faciliter l'évolution indépendante des données et des programmes.

Par exemple, on peut considérer la banque GenBank comme un énorme fichier contenant une suite d'enregistrement et pour chacun des champs spécifiques définis, avec une seule clé d'index comme entrée.

Par exemple, MICADO (MICrobial Advanced Database Organization) est une base de données relationnelle (système de gestion PostgreSQL), dédiée aux génomes microbiens. Elle intègre notamment l'ensemble des séquences primaires microbiennes issues de Genbank, les génomes complets microbiens réannotés dans la banque Emglib et les données d'analyse fonctionnelle de la bactérie modèle *B. subtilis*.

Il existe un grand nombre de banques ou bases de données d'intérêt biologique. Cette introduction sera limitée à une présentation des principales banques de données publiques, basées sur la structure primaire des séquences. Nous distinguerons deux types de banques :

- celles qui correspondent à une collecte des données la plus exhaustive possible et qui offrent finalement un ensemble plutôt hétérogène d'informations (banques de séquences généralistes)
- celles qui correspondent à des données plus homogènes établies autour d'une thématique et qui offrent une valeur ajoutée à partir d'une technique particulière ou d'un intérêt suscité par un groupe d'individus (banques ou bases de séquences spécialisées).

La séquence est l'élément central autour duquel les banques de données se sont constituées. Les séquences biologiques, dès qu'elles ont pu être établies, ont très tôt fait l'objet d'une compilation dans les banques de données. La première compilation de protéines est publiée en 1965 par Margaret Dayhoff : c'est l'Atlas of Protein Sequences qui contient alors 50 entrées. D'abord imprimé jusqu'en 1978, il fut ensuite proposé sous forme électronique.

3.1. Les banques de séquences généralistes

C'est au début des années 80 que les premières banques de séquences sont apparues sous l'initiative de quelques équipes dont la première à l'initiative de Grantham et C. Gautier à Lyon. Très rapidement avec les évolutions techniques du séquençage, la collecte et la gestion des données ont nécessité une organisation plus conséquente. Ainsi, plusieurs organismes ont pris en charge la production de telles bases de données. Nous présenterons dans les paragraphes suivants l'information contenue dans les banques telles qu'elle apparaît lors d'une requête et nous ne dirons rien de la structuration informatique de celles-ci.

Trois banques de séquence nucléiques :

- **EMBL** : banque européenne créée en 1980 et financée par l'EMBO (European Moleculary Biology Organization), elle est aujourd'hui diffusée par l'EBI (European Bioinformatics Institute, Cambridge, UK)
- **GenBank** : créée en 1982 par la société IntelliGenetics et diffusée maintenant par le NCBI (National Center for Biotechnology Information, Los Alamos, US)
- **DDBJ** : créée en 1986 et diffusée par le NIG (National Institute of Genetics, Japon)

Ces trois banques s'échangent systématiquement leur contenu depuis 1987 et ont adopté un système de conventions communes : "The DDBJ/EMBL/GenBank Feature Table Definition".

Deux banques protéiques :

- **PIR-NBRF** : créée en 1984 par la NBRF (National Biomedical Research Foundation). Elle est maintenant un ensemble de données issues du MIPS (Martinsried Institute for Protein Sequences, Munich, Allemagne) et de la banque japonaise JIPID (Japan International Protein Information Database)
- **SwissProt** : créée en 1986 à l'Université de Genève et maintenue depuis 1987 dans le cadre d'une collaboration, entre cette université (via ExPASy, Expert Protein Analysis System) et l'EBI. Celle-ci regroupe aussi des séquences annotées de la banque PIR-NBRF ainsi que des séquences codantes, traduites de l'EMBL.

Elles contiennent les protéine obtenues de plusieurs manières différentes :

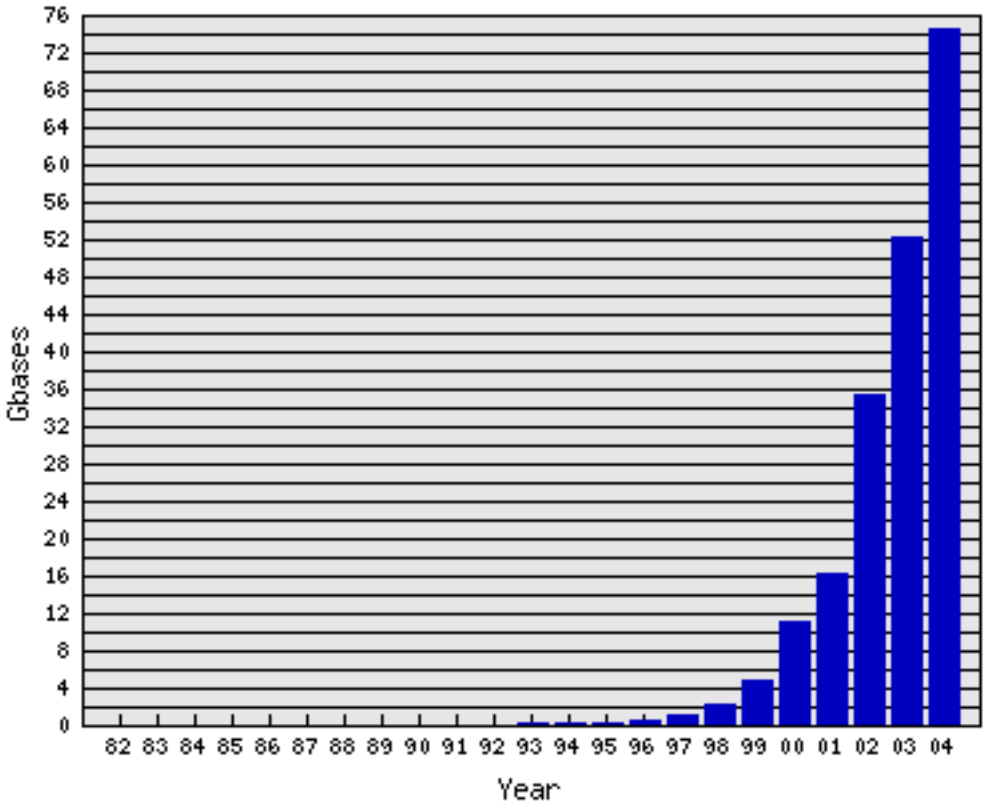
- *in silico* : déduite à partir de la séquence nucléique, par simple traduction du ou des exons la codant
- isolée à partir de la cellule
- ou encore par génie génétique

3.1.1. EMBL (nucléique)

Financée par l'EMBO (European Moleculary Biology Organisation), développée au sein du Laboratoire Européen de Biologie Moléculaire situé à Heidelberg (Allemagne), elle est maintenant diffusée par l'EBI (European Bioinformatics Institute), situé près de Cambridge (Angleterre).

Cette banque contient 74 491 158 213 nucléotides dans 44 538 943 entrées à la date du Vendredi 22 0ctobre 2004.

Voici l'évolution du nombre du nombre de nucléotides depuis sa création :



(extrait des statistiques : `http://www3.ebi.ac.uk/Services/DBStats/`)

L'évolution du nombre d'entrées a un profil similaire. Toute la documentation pour cette banque est disponible sur le serveur de l'EBI :

`http://www.ebi.ac.uk/embl/Documentation/`

Voici un exemple d'entrée :

```
ID   AF148567   standard; mRNA; MAM; 510 BP.
XX
AC   AF148567;
XX
SV   AF148567.1
XX
DT   07-MAY-2000 (Rel. 63, Created)
DT   07-MAY-2000 (Rel. 63, Last updated, Version 1)
XX
```

```
DE   Sus scrofa pancreatic colipase mRNA, complete cds.
XX
KW   .
XX
OS   Sus scrofa (pig)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Cetartiodactyla; Suina; Suidae; Sus.
XX
RN   [1]
RP   1-510
RA   Darnis S., Chaix J.C.;
RT   "Cloning, sequencing and functional expression of porcine pancreatic
RT   colipase";
RL   Unpublished.
XX
RN   [2]
RP   1-510
RA   Darnis S., Chaix J.C.;
RT   ;
RL   Submitted (04-MAY-1999) to the EMBL/GenBank/DDBJ databases.
RL   LBBN, CNRS.ESA6033, Av Escadrille Normandie Niemen, Marseille 13397, France
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..510
FT                   /db_xref="taxon:9823"
FT                   /mol_type="mRNA"
FT                   /organism="Sus scrofa"
FT                   /tissue_type="pancreas"
FT   CDS             44..382
FT                   /codon_start=1
FT                   /db_xref="GOA:P02703"
FT                   /db_xref="UniProt/Swiss-Prot:P02703"
FT                   /note="cofactor; triglyceride lipase"
FT                   /product="colipase"
FT                   /protein_id="AAF67823.1"
FT                   /translation="MEKVLALLLVTLTVAYAVPDPRGIIINLDEGELCLNSAQCKSNCC
FT                   QHDTILSLSRCALKARENSECSAFTLYGVYYKCPCERGLTCEGDKSLVGSITNTNFGIC
FT                   HDVGRSSD"
XX
SQ   Sequence 510 BP; 117 A; 160 C; 121 G; 112 T; 0 other;
     ctgaccttcc agctgctact tacaccagct gtgctcattc atcatggaga aggtccttgc        60
     ccttctgctc gtgaccctca cggtagccta tgcggttcca gacccccgcg gaatcattat       120
     caacctggat gagggcgagc tctgcctgaa cagtgcccag tgcaagagca actgctgcca       180
     gcatgacaca atcctgagcc tgtcccgctg cgcactcaag gccagagaga acagcgagtg       240
     ttctgccttc acgctctatg gggtttacta caagtgtccc tgtgaacggg gcctgacctg       300
     tgagggggac aagagtctcg tgggctccat caccaacacc aactttggta tctgccatga       360
     tgttggacgc tccagtgact gagaacacac accagctgga gcaccgaggg acgcccctcc       420
     ttccaccact actccctggc tggcacctcc gtcttctcat tgggttcttg gcaattaaag       480
     cccctcttgc aaaccttaaa aaaaaaaaaa                                        510
//
```

Le texte en style gras est une information par un lien (URL) vers un serveur Web.
Chaque entrée de la base EMBL est composée de lignes ou champs qui commencent par une **étiquette, code à 2 caractères** indiquant le type d'information contenue dans la ligne et la fin de l'entrée est indiquée par //.

Ces étiquettes sont divisées en cinq parties (chacune délimitée par une bordure):

1) General Information

- Etiquette **ID** : identificateur de l'entrée contenant la séquence. Cette ligne a la structure suivante : nom de l'entrée classe de la donnée ; molécule (DNA, RNA, RNAm, XXX si l'entrée n'a pas été annotée) ; division ; longueur de la séquence en paire de bases (pb)
- Etiquette **XX** : C'est une ligne vide qui sert à limiter les différents champs de l'entrée et à clarifier sa lecture
- Etiquette **AC** : numéro d'accession de l'entrée
- Etiquette **SV** : version de la séquence
- Etiquette **DT** : donne la date d'incorporation dans la base (1ère ligne) et la date de la dernière mise à jour de l'entrée (2ème ligne)

2) Description

- Etiquette **DE** : informations descriptives sur la séquence
- Etiquette **KW** : mot(s)-clé(s)
- Etiquette **OS** : organisme
- Etiquette **OC** : ordre dans la classification

3) Références

- Etiquette **RN** : numéro de la référence (peut être utilisé dans les features)
- Etiquette ***RC*** : commentaires sur la référence
- Etiquette **RP** : région de la séquence
- Etiquette ***RX*** : lien (URL) vers des bases bibliographiques accessibles par le réseau (par exemple Medline, PubMed)
- Etiquette **RA** : auteurs de la publication
- Etiquette **RT** : titre de la publication
- Etiquette **RL** : référence : journal, volume, pages, année (peut aussi porter la mention : unpublished)

Les étiquettes en style italique sont facultatives.

4) Additional Information : Features (facultatif)

La ligne est composée de l'étiquette **FT**, suivie d'un mot-clé (**Key**), lui-même suivi par le champ (**Location/Qualifiers**) qui est un couple de mot-clé/valeur, le mot-clé étant soit de type "location" soit de type "qualifier". Depuis 1987 un système de conventions communes a été adopté par les trois banques généralistes nucléiques : "The DDBJ/EMBL/GenBank Feature Table Definition".

Lorsque le mot-clé (key) est absent, cela signifie que la ligne "FT Location/Qualifiers" est la suite de la précédente.

Un "qualifier" de type *db_xref* est un lien (URL) vers une banque ou base de données, c'est une référence croisée (style gras dans l'exemple donné).

Le nombre de mots-clé est très important et varié, vous en avez une liste exhaustive sur le serveur de l'EBI :

`http://www.ebi.ac.uk/embl/Documentation/FT_definitions/feature_table.html`

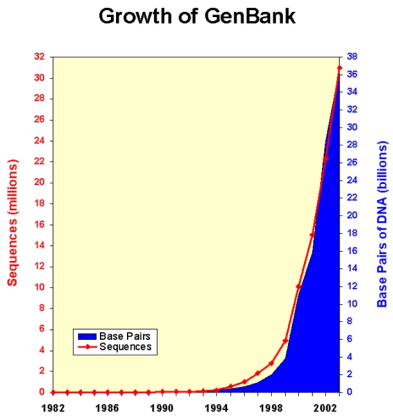
5) Sequence

- Une seule étiquette : **SQ** qui contient le nombre de paires de base et la répartition entre les différents nucléotides.

La séquence est suivie par l'étiquette // qui indique la fin de l'entrée.

3.1.2. GenBank (nucléique)

Créée en 1982 par la société IntelliGenetics et diffusée maintenant par le NCBI, elle contient 43 194 602 655 nucléotides dans 38 941 263 entrées à la date du Vendredi 22 0ctobre 2004.



(extrait des statistiques : `http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html`)
Toute la documentation pour cette banque est disponible sur le serveur du NCBI :

`http://www.ncbi.nlm.nih.gov/Genbank/GenbankOverview.html`

Voici un exemple d'entrée :

```
LOCUS       NM_001832               539 bp    mRNA    linear   PRI 23-AUG-2004
DEFINITION  Homo sapiens colipase, pancreatic (CLPS), mRNA.
ACCESSION   NM_001832
VERSION     NM_001832.2  GI:11496883
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 539)
  AUTHORS   van Tilbeurgh,H., Bezzine,S., Cambillau,C., Verger,R. and
            Carriere,F.
  TITLE     Colipase: structure and interaction with pancreatic lipase
  JOURNAL   Biochim. Biophys. Acta 1441 (2-3), 173-184 (1999)
   PUBMED   10570245
REFERENCE   2  (bases 1 to 539)
  AUTHORS   Sims,H.F. and Lowe,M.E.
```

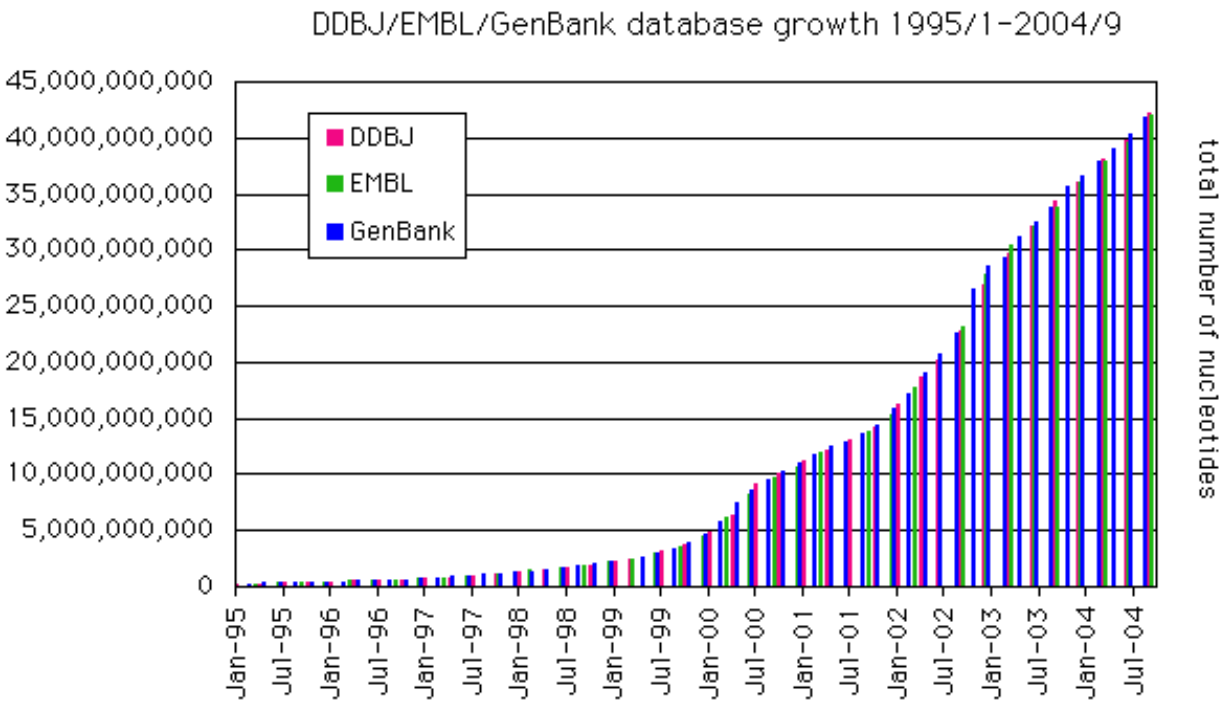
```
  TITLE     The human colipase gene: isolation, chromosomal location, and
            tissue-specific expression
  JOURNAL   Biochemistry 31 (31), 7120-7125 (1992)
   PUBMED   1643046
REFERENCE   3  (bases 1 to 539)
  AUTHORS   Davis,R.C., Xia,Y.R., Mohandas,T., Schotz,M.C. and Lusis,A.J.
  TITLE     Assignment of the human pancreatic colipase gene to chromosome
            6p21.1 to pter
  JOURNAL   Genomics 10 (1), 262-265 (1991)
   PUBMED   2045105
REFERENCE   4  (bases 1 to 539)
  AUTHORS   Lowe,M.E., Rosenblum,J.L., McEwen,P. and Strauss,A.W.
  TITLE     Cloning and characterization of the human colipase cDNA
  JOURNAL   Biochemistry 29 (3), 823-828 (1990)
   PUBMED   2337598
COMMENT     REVIEWED REFSEQ: This record has been curated by NCBI staff. The
            reference sequence was derived from J02883.1 and M95529.1.
            On Dec 1, 2000 this sequence version replaced gi:4502894.

            Summary: The protein encoded by this gene is a cofactor needed by
            pancreatic lipase for efficient dietary lipid hydrolysis. It binds
            to the C-terminal, non-catalytic domain of lipase, thereby
            stabilizing an active conformation and considerably increasing the
            overall hydrophobic binding site. The gene product allows lipase to
            anchor noncovalently to the surface of lipid micelles,
            counteracting the destabilizing influence of intestinal bile salts.
            This cofactor is only expressed in pancreatic acinar cells,
            suggesting regulation of expression by tissue-specific elements.
            COMPLETENESS: full length.
FEATURES             Location/Qualifiers
     source          1..539
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /db_xref="taxon:9606"
                     /chromosome="6"
                     /map="6pter-p21.1"
     gene            1..539
                     /gene="CLPS"
                     /db_xref="GeneID:1208"
                     /db_xref="LocusID:1208"
                     /db_xref="MIM:120105"
     CDS             38..376
                     /gene="CLPS"
                     /note="go_component: extracellular [goid 0005576]
                     [evidence IEA];
                     go_component: soluble fraction [goid 0005625] [evidence
                     NR];
                     go_function: enzyme activator activity [goid 0008047]
                     [evidence IEA];
                     go_process: digestion [goid 0007586] [evidence IEA];
                     go_process: lipid catabolism [goid 0016042] [evidence
                     IEA]"
                     /codon_start=1
                     /product="colipase preproprotein"
                     /protein_id="NP_001823.1"
                     /db_xref="GI:4502895"
                     /db_xref="GeneID:1208"
                     /db_xref="LocusID:1208"
                     /db_xref="MIM:120105"
                     /translation="MEKILILLLVALSVAYAAPGPRGIIINLENGELCMNSAQCKSNC
                     CQHSSALGLARCTSMASENSECSVKTLYGIYYKCPCERGLTCEGDKTIVGSITNTNFG
                     ICHDAGRSKQ"
     sig_peptide     38..88
                     /gene="CLPS"
     proprotein      89..373
                     /gene="CLPS"
     mat_peptide     104..373
                     /gene="CLPS"
                     /product="colipase"
     polyA_signal    517..522
                     /gene="CLPS"
     polyA_site      539
```

```
                     /gene="CLPS"
ORIGIN
        1 ctgtctcccg ccacccacac cagctgtccc actcaccatg gagaagatcc tgatcctcct
       61 gcttgtcgcc ctctctgtgg cctatgcagc tcctggcccc cgggggatca ttatcaacct
      121 ggagaacggt gagctctgca tgaatagtgc ccagtgtaag agcaattgct gccagcattc
      181 aagtgcgctg ggcctggccc gctgcacatc catggccagc gagaacagcg agtgctctgt
      241 caagacgctc tatgggattt actacaagtg tccctgtgag cgtggcctga cctgtgaggg
      301 agacaagacc atcgtgggct ccatcaccaa caccaacttt ggcatctgcc atgacgctgg
      361 acgctccaag cagtgagact gcccacccac tcccacacct agcccagaat gctgtaggcc
      421 actaggcgca ggggcatctc tcccctgctc cagcgcatct cccgggctgg ccacctcctt
      481 gaccagcata tctgttttct gattgcgctc ttcacaatta aaggcctcct gcaaacctt
//
```

Le texte en style gras est une information par un lien (URL) vers un serveur Web.
Les informations et leurs formats sont très similaires à celles de la banque EMBL, à ceci près que les étiquettes ne sont pas des abréviations mais le nom complet, directement explicite. Rappelons que depuis 1987 pour les "Features", un système de conventions communes a été adopté par les trois banques généralistes nucléiques : "The DDBJ/EMBL/GenBank Feature Table Definition".
GenBank contient une sous-banque de protéines, traduction des séquences nucléiques, appelée GenPept.

3.1.3. DDBJ (nucléique)

Créée en 1986 et diffusée par le NIG (National Institute of Genetics, Japon) , elle contient 42 245 956 937 nucléotides dans 37 926 117 entrées à la date du Vendredi 22 0ctobre 2004.



(extrait des statistiques : `http://www.ddbj.nig.ac.jp/ddbjnew/statistics-e.html`)
Toute la documentation pour cette banque est disponible sur le serveur du DDBJ :

http://www.ddbj.nig.ac.jp/

Voici un exemple d'entrée (même gène que l'entrée pour GenBank) :

```
LOCUS       HUMCOLIP                523 bp    mRNA    linear   HUM 01-NOV-1994
DEFINITION  Human colipase mRNA, complete cds.
ACCESSION   J02883
VERSION     J02883.1
KEYWORDS    cofactor; colipase; triglyceride lipase.
SOURCE      Homo sapiens
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 523)
  AUTHORS   Lowe,M.E., Rosenblum,J.L., McEwen,P. and Strauss,A.W.
  TITLE     Cloning and characterization of the human colipase cDNA
  JOURNAL   Biochemistry 29 (3), 823-828 (1990)
  MEDLINE   90248429
   PUBMED   2337598
COMMENT     Original source text: Human adult pancreas, cDNA to mRNA.
            Draft entry and computer-readable sequence for [1] kindly submitted
            by M.E.Lowe, 17-NOV-1989.
FEATURES             Location/Qualifiers
     source          1..523
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /db_xref="taxon:9606"
                     /map="6pter-p21.1"
     gene            1..523
                     /gene="CLPS"
     mRNA            <1..523
                     /gene="CLPS"
                     /product="CLP mRNA"
     CDS             22..360
                     /gene="CLPS"
                     /note="colipase precursor"
                     /codon_start=1
                     /protein_id="AAA52054.1"
                     /db_xref="GI:180886"
                     /db_xref="GDB:G00-127-277"
                     /translation="MEKILILLLVALSVAYAAPGPRGIIINLENGELCMNSAQCKSNC
                     CQHSSALGLARCTSMASENSECSVKTLYGIYYKCPCERGLTCEGDKTIVGSITNTNFG
                     ICHDAGRSKQ"
     sig_peptide     22..75
                     /gene="CLPS"
                     /note="colipase signal peptide"
     mat_peptide     73..357
                     /gene="CLPS"
                     /product="colipase"
BASE COUNT          109 a          173 c          128 g          113 t
ORIGIN
        1 acaccagctg tcccactcac catggagaag atcctgatcc tcctgcttgt cgccctctct
       61 gtggcctatg cagctcctgg ccccgggggg atcattatca acctggagaa cggtgagctc
      121 tgcatgaata gtgcccagtg taagagcaat tgctgccagc attcaagtgc gctgggcctg
      181 gcccgctgca catccatggc cagcgagaac agcgagtgct ctgtcaagac gctctatggg
      241 atttactaca agtgtccctg tgagcgtggc ctgacctgtg agggagacaa gaccatcgtg
      301 ggctccatca ccaacaccaa ctttggcatc tgccatgacg ctggacgctc caagcagtga
      361 gactgcccac ccactcccac acctagccca gaatgctgta ggccactagg cgcaggggca
      421 tctctccccct gctccagcgc atctcccggg ctggccacct ccttgaccag catatctgtt
      481 ttctgattgc gctcttcaca attaaaggcc tcctgcaaac ctt
//
```

Le texte en style gras est une information par un lien (URL) vers un serveur Web.

Les informations et leurs formats sont très similaires à celles de la banque GenBank. Il semble que toutefois les liens soient moins nombreux que dans GenBank. Rappelons que

depuis 1987 pour les "Features", un système de conventions communes a été adopté par les trois banques généralistes nucléiques : "The DDBJ/EMBL/GenBank Feature Table Definition".

3.1.4. PIR-NBRF (protéique)

Créée en 1984 par la NBRF (National Biomedical Research Foundation). Elle est maintenant un ensemble de données issues du MIPS (Martinsried Institute for Protein Sequences, Munich, Allemagne) et de la banque japonaise JIPID (Japan International Protein Information Database).
Elle contient 283 416 entrées. D'autres bases dérivées sont accessibles telles que iProClass basée sur les familles de protéines et décrivant leurs structures et leurs fonctions ou encore PIR-NREF qui contient les séquences PIR, Swiss-Prot, TrEMBL, RefSeq, GenPept, et PDB sans aucune redondance.

Toute la documentation pour cette banque est disponible sur le serveur de PIR :

`http://pir.georgetown.edu/home.shtml`

Voici un exemple d'entrée de PIR-NBRF :

```
ENTRY           XLHU  #type complete    iProClass View of XLHU
TITLE           colipase precursor [validated] - human
ALTERNATE_NAMES procolipase
ORGANISM        #formal_name Homo sapiens #common_name man
   #cross-references taxon:9606
DATE            04-Dec-1986 #sequence_revision 19-May-1995 #text_change
                09-Jul-2004
ACCESSIONS      A42568; A33949; A03163
REFERENCE       A42568
   #authors     Sims, H.F.; Lowe, M.E.
   #journal     Biochemistry (1992) 31:7120-7125
   #title       The human colipase gene: isolation, chromosomal location,
                and tissue-specific expression.
   #cross-references MUID:92353041; PMID:1643046
   #accession   A42568
      ##molecule_type DNA
      ##residues 1-112 ##label SIM
      ##cross-references UNIPROT:P04118; GB:M95529; NID:g180842;
                PIDN:AAB05818.1; PID:g1483624
      ##note sequence extracted from NCBI backbone (NCBIN:110576,
                NCBIN:110578, NCBIP:110580)
REFERENCE       A33949
   #authors     Lowe, M.E.; Rosenblum, J.L.; McEwen, P.; Strauss, A.W.
   #journal     Biochemistry (1990) 29:823-828
   #title       Cloning and characterization of the human colipase cDNA.
   #cross-references MUID:90248429; PMID:2337598
   #accession   A33949
      ##molecule_type mRNA
      ##residues 1-112 ##label LOW
      ##cross-references GB:J02883; NID:g180885; PIDN:AAA52054.1;
                PID:g180886
      ##note evidence of partial N-glycosylation, possibly at Asn-43
REFERENCE       A90652
   #authors     Sternby, B.; Engstrom, A.; Hellman, U.; Vihert, A.M.;
                Sternby, N.H.; Borgstrom, B.
   #journal     Biochim. Biophys. Acta (1984) 784:75-80
   #title       The primary sequence of human pancreatic colipase.
```

```
   #cross-references MUID:84104937; PMID:6691986
   #accession   A03163
      ##molecule_type protein
      ##residues 23-108 ##label STE
COMMENT         Colipase, a cofactor of triacylglycerol lipase (EC
                3.1.1.3), forms a 1:1 stoichiometric complex with it,
                enabling it to hydrolyze its substrate at the lipid-water
                interface. Without colipase the enzyme is washed off by
                bile salts, which are known to have an inhibitory effect
                on the lipase.
GENETICS
   #gene        GDB:CLPS
      ##cross-references GDB:127277; OMIM:120105
   #map_position 6pter-6p21.1
   #introns     28/3; 69/3
CLASSIFICATION  SF002415
   #superfamily colipase
KEYWORDS        lipid digestion; lipid hydrolysis; pancreas
FEATURE
   1-17                 #domain signal sequence #status predicted #label
                        SIG\
   18-22                #domain amino-terminal propeptide #status
                        predicted #label APP\
   23-108               #product colipase #status experimental #label
                        MAT\
   109-112              #domain carboxyl-terminal propeptide #status
                        predicted #label CPP\
   34-104,40-56,44-80,
   45-78,66-86          #disulfide_bonds #status predicted\
   69,72,75,76          #binding_site micellar substrate (Lys, Tyr, Tyr,
                        Tyr) #status predicted
SUMMARY         #length 112 #molecular_weight 11954

SEQUENCE
             5        10        15        20        25        30
    1 M E K I L I L L L V A L S V A Y A A P G P R G I I I N L E N
   31 G E L C M N S A Q C K S N C C Q H S S A L G L A R C T S M A
   61 S E N S E C S V K T L Y G I Y Y K C P C E R G L T C E G D K
   91 T I V G S I T N T N F G I C H D A G R S K Q

PDB structures most related to XLHU:
    1LPAA (19-110) 78.3%
SCOP:   1LPA
CATH:   1LPA
FSSP:   1LPA
MMDB:   1LPA

ALIGNMENTS containing XLHU:
    FA0498 colipase - 488.2 1.0
    M02604 colipase - 2341.0 1.0
```

Le texte en style gras est une information par un lien (URL) vers un serveur Web.

Les informations et leurs formats sont très similaires à celles des banques précédentes. Les étiquettes ne sont pas des abréviations mais le nom complet, directement explicite.

Les "`#cross-references MUID:92353041; PMID:1643046`" sont des liens bibliographiques, les "`##cross-references UNIPROT:P04118; GB:M95529; NID:g180842;PIDN:AAB05818.1; PID:g1483624`" sont des liens vers las banques de données Uniprot, GenBank, GenPept

L'ensemble de lignes qui suivent la séquence sont des liens avec des banques spécialisées, par exemple, les lignes "`PDB structures most related to XLHU - 1LPAA (19-110) 78.3%`" sont un lien vers la banque de coordonnées cristallographiques (PDB).

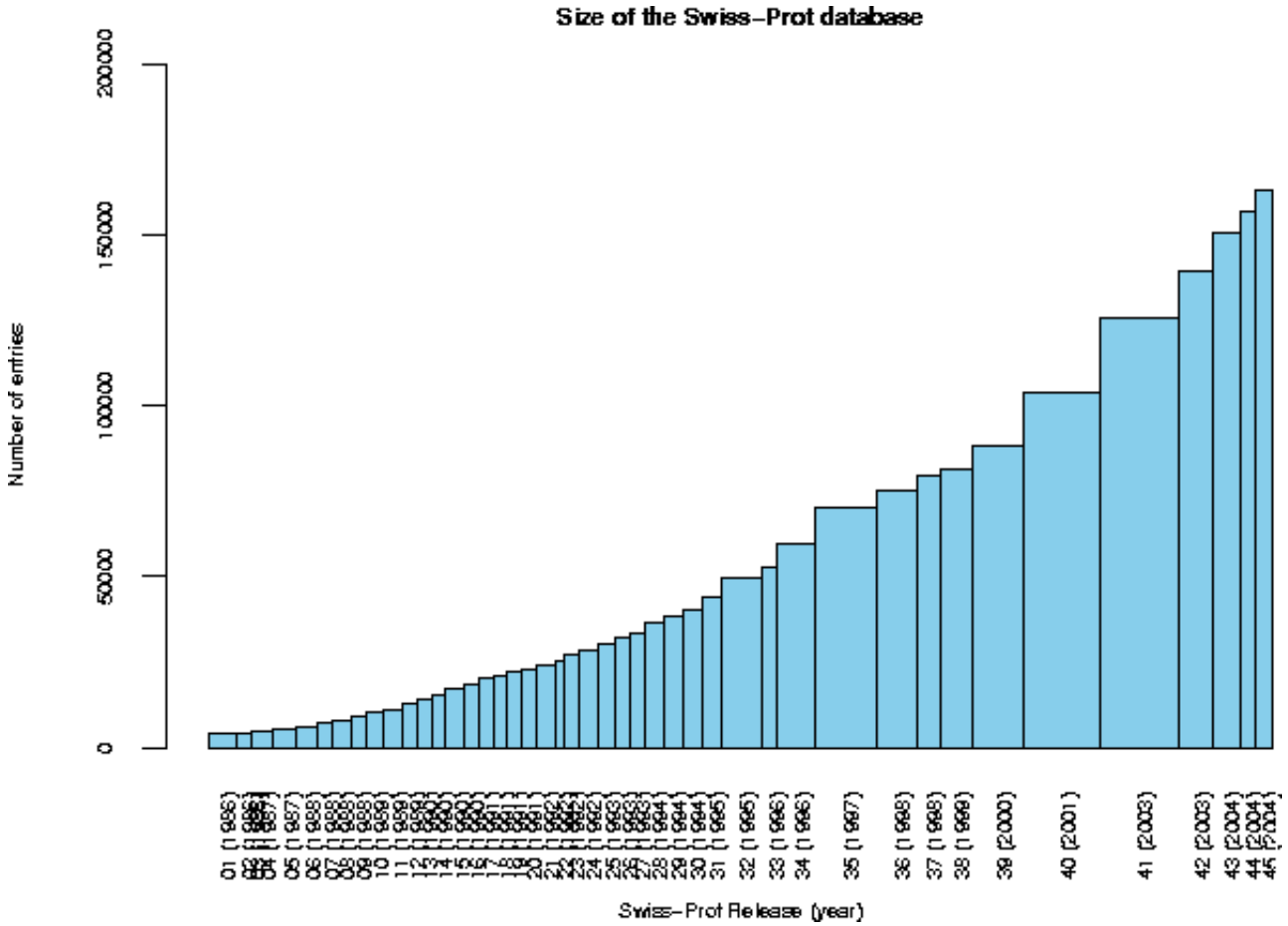
3.1.5. SwissProt (protéique)

Créée en 1986 à l'Université de Genève et maintenue depuis 1987 dans le cadre d'une collaboration, entre cette université (via ExPASy, Expert Protein Analysis System) et l'EBI. Celle-ci regroupe aussi des séquences annotées de la banque PIR-NBRF ainsi que des séquences codantes traduites de l'EMBL.

Elle contient 163 235 entrées, pour un total de 59 631 787 aminoacides avec 120 520 résumés de références bibliographiques à la date du Vendredi 22 0ctobre 2004.

Elle est complétée par la banque TrEMBL qui est un supplément de SwissProt, annotée automatiquement qui contient les traductions des séquences nucléotides de EMBL, pas encore intégrées dans SwissProt.

TrEMBL contient 1 422 984 entrées pour un total de 444 525 054 aminoacides à la date du Vendredi 22 0ctobre 2004.



(extrait des statistiques : http://www.expasy.org/sprot/relnotes/relstat.html)

Toute la documentation pour cette banque est disponible sur le serveur d'Expasy:

`http://www.expasy.org/sprot/sp-docu.html`

Voici un exemple d'entrée de SwissProt :

```
ID   COL_HUMAN      STANDARD;      PRT;   112 AA.
AC   P04118;
DT   01-NOV-1986 (Rel. 03, Created)
DT   01-APR-1990 (Rel. 14, Last sequence update)
DT   05-JUL-2004 (Rel. 44, Last annotation update)
DE   Colipase precursor.
GN   Name=CLPS;
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=90248429; PubMed=2337598 [NCBI, ExPASy, EBI, Israel, Japan];
RA   Lowe M.E., Rosenblum J.L., McEwen P., Strauss A.W.;
RT   "Cloning and characterization of the human colipase cDNA.";
RL   Biochemistry 29:823-828(1990).
RN   [2]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=92353041; PubMed=1643046 [NCBI, ExPASy, EBI, Israel, Japan];
RA   Sims H.F., Lowe M.E.;
RT   "The human colipase gene: isolation, chromosomal location, and tissue-
RT   specific expression.";
RL   Biochemistry 31:7120-7125(1992).
RN   [3]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Pancreas;
RX   MEDLINE=22388257; PubMed=12477932 [NCBI, ExPASy, EBI, Israel, Japan];
     DOI=10.1073/pnas.242603899;
RA   Strausberg R.L., Feingold E.A., Grouse L.H., Derge J.G.,
RA   Klausner R.D., Collins F.S., Wagner L., Shenmen C.M., Schuler G.D.,
RA   Altschul S.F., Zeeberg B., Buetow K.H., Schaefer C.F., Bhat N.K.,
RA   Hopkins R.F., Jordan H., Moore T., Max S.I., Wang J., Hsieh F.,
RA   Diatchenko L., Marusina K., Farmer A.A., Rubin G.M., Hong L.,
RA   Stapleton M., Soares M.B., Bonaldo M.F., Casavant T.L., Scheetz T.E.,
RA   Brownstein M.J., Usdin T.B., Toshiyuki S., Carninci P., Prange C.,
RA   Raha S.S., Loquellano N.A., Peters G.J., Abramson R.D., Mullahy S.J.,
RA   Bosak S.A., McEwan P.J., McKernan K.J., Malek J.A., Gunaratne P.H.,
RA   Richards S., Worley K.C., Hale S., Garcia A.M., Gay L.J., Hulyk S.W.,
RA   Villalon D.K., Muzny D.M., Sodergren E.J., Lu X., Gibbs R.A.,
RA   Fahey J., Helton E., Ketteman M., Madan A., Rodrigues S., Sanchez A.,
RA   Whiting M., Madan A., Young A.C., Shevchenko Y., Bouffard G.G.,
RA   Blakesley R.W., Touchman J.W., Green E.D., Dickson M.C.,
RA   Rodriguez A.C., Grimwood J., Schmutz J., Myers R.M.,
RA   Butterfield Y.S.N., Krzywinski M.I., Skalska U., Smailus D.E.,
RA   Schnerch A., Schein J.E., Jones S.J.M., Marra M.A.;
RT   "Generation and initial analysis of more than 15,000 full-length human
RT   and mouse cDNA sequences.";
RL   Proc. Natl. Acad. Sci. U.S.A. 99:16899-16903(2002).
RN   [4]
RP   SEQUENCE OF 23-108.
RC   TISSUE=Pancreas;
RX   MEDLINE=84104937; PubMed=6691986 [NCBI, ExPASy, EBI, Israel, Japan];
     DOI=10.1016/0167-4838(84)90175-4;
RA   Sternby B., Engstroem A., Hellman U., Vihert A.M., Sternby N.-H.,
RA   Borgstroem B.;
RT   "The primary sequence of human pancreatic colipase.";
RL   Biochim. Biophys. Acta 784:75-80(1984).
RN   [5]
RP   X-RAY CRYSTALLOGRAPHY (3.0 ANGSTROMS).
RX   MEDLINE=93241293; PubMed=8479519 [NCBI, ExPASy, EBI, Israel, Japan];
     DOI=10.1038/362814a0;
RA   van Tilbeurgh H., Egloff M.-P., Martinez C., Rugani N., Verger R.,
RA   Cambillau C.;
RT   "Interfacial activation of the lipase-procolipase complex by mixed
RT   micelles revealed by X-ray crystallography.";
RL   Nature 362:814-820(1993).
CC   -!- FUNCTION: Colipase is a cofactor of pancreatic lipase. It allows
CC       the lipase to anchor itself to the lipid-water interface. Without
CC       colipase the enzyme is washed off by bile salts, which have an
CC       inhibitory effect on the lipase.
CC   -!- FUNCTION: Enterostatin has a biological activity as a satiety
```

```
CC       signal.
CC   -!- SUBUNIT: Forms a 1:1 stoichiometric complex with pancreatic
CC       lipase.
CC   -!- SUBCELLULAR LOCATION: Secreted.
CC   -!- TISSUE SPECIFICITY: Expressed by the pancreas.
CC   -!- SIMILARITY: Belongs to the colipase family.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; J02883; AAA52054.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence]
DR   EMBL; M95529; AAB05818.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence]
DR   EMBL; BC007061; AAH07061.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence]
DR   EMBL; BC017897; AAH17897.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence]
DR   PIR; A42568; XLHU.
DR   HSSP; P02703; 1ETH. [HSSP ENTRY / SWISS-3DIMAGE / PDB]
DR   Genew; HGNC:2085; CLPS.
DR   CleanEx; HGNC:2085; CLPS.
DR   H-InvDB; HIX0005811; -.
DR   MIM; 120105; -. [NCBI / EBI]
DR   GeneCards; CLPS.
DR   GeneLynx; CLPS.
DR   GenAtlas; CLPS.
DR   Ensembl; P04118. [Entry / Contig view]
DR   SOURCE; CLPS.
DR   InterPro; IPR001981; Colipase.
DR   InterPro; Graphical view of domain structure.
DR   Pfam; PF01114; Colipase; 1.
DR   Pfam; PF02740; Colipase_C; 1.
DR   Pfam; Graphical view of domain structure.
DR   PRINTS; PR00128; COLIPASE.
DR   SMART; SM00023; COLIPASE; 1.
DR   PROSITE; PS00121; COLIPASE; 1.
DR   ProDom [Domain structure / List of seq. sharing at least 1 domain]
DR   HOVERGEN [Family / Alignment / Tree]
DR   BLOCKS; P04118.
DR   ProtoNet; P04118.
DR   ProtoMap; P04118.
DR   PRESAGE; P04118.
DR   DIP; P04118.
DR   ModBase; P04118.
DR   SMR; P04118.
DR   SWISS-2DPAGE; GET REGION ON 2D PAGE.
KW   Digestion; Direct protein sequencing; Lipid degradation; Pancreas;
KW   Signal.
FT   SIGNAL        1     17
FT   PROPEP       18     22       Enterostatin, activation peptide
FT                                (Potential).
FT   CHAIN        23    112       Colipase.
FT   DISULFID     34     45       By similarity.
FT   DISULFID     40     56       By similarity.
FT   DISULFID     44     78       By similarity.
FT   DISULFID     66     86       By similarity.
FT   DISULFID     80    104       By similarity.
FT   CONFLICT     68     69       Missing (in Ref. 2).
SQ   SEQUENCE   112 AA;  11954 MW;  772872EBBE7C4DF8 CRC64;
     MEKILILLLV ALSVAYAAPG PRGIIINLEN GELCMNSAQC KSNCCQHSSA LGLARCTSMA
     SENSECSVKT LYGIYYKCPC ERGLTCEGDK TIVGSITNTN FGICHDAGRS KQ
//
```

Le texte en style gras ou le texte sur fond gris clair, est une information par un lien (URL) vers un serveur Web.

Commentons quelques étiquettes :

- **ID** : identificateur de l'entrée contenant la séquence qui se décompose ainsi nom de l'entrée classe de la donnée ; molécule (PRT, XXX si l'entrée n'a pas été annotée) ; division ; longueur de la séquence en nombre d'acides aminés, suivie de AA.
- **AC** : numéro d'accession de l'entrée
- **DT** : date d'incorporation (1ère ligne) dans la base ou date de modification pour les suivantes
- **DE** : informations descriptives sur la séquence
- **KW** : mot(s)-clé(s) qui peuvent être utilisés pour retrouver l'entrée dans la base.
- ***GN*** : noms des gènes codant pour la séquence de protéine.
- **OS** : organisme d'où provient la séquence ; le plus souvent on donne le nom latin suivi du nom anglais entre parenthèses.
- **OC** : ordre dans la classification
- ***OG*** : localisation cellulaire des gènes qui codent pour la séquence
- **OX** : numéro du taxon (lien sur un serveur taxonomique)
- **RN** : numéro unique attribué à chaque référence bibliographique de l'entrée.
- ***RC*** : commentaires sur la référence.
- **RX** : référence bibliographique (lien avec les base bibliographiques, PubMed, Medline)
- **RP** : références associées aux différentes régions de la séquence
- **RA** : auteurs de l'article (chaque auteur a un lien vers un serveur qui renvoie comme résultat toutes les occurrences de celui-ci dans la banque SwissProt)
- **RT** : titre de l'article
- **RL** : références du journal (lien vers l'abstract si le journal a un serveur Web)
- ***CC*** : commentaires
- ***DR*** : liaisons avec d'autres bases de données qui contiennent une information en relation avec cette entrée.
- ***FT*** : "features" : annotation sur la séquence formée par un **mot-clé** suivi de la **région de la séquence** (début .. fin) et de la **description**
- **SQ** : longueur de la séquence (AA) ainsi que la masse molaire (MW) et la valeur du 64-bit CRC de la séquence (Cyclic Redundancy Check), calculé par un algorithme (ISO 3309)
- // : fin de l'entrée.

L'ensemble de lignes avec l'étiquette **DR** sont des liens avec des banques spécialisées. Cette banque généraliste est la plus riche en liens vers d'autres banques, qu'elles soient généralistes ou spécialisées.

3.1.6. Uniprot (protéique)

Le consortium d'UniProt est composé de l'Institut Européen de Bioinformatics (EBI), de l'Institut Suisse de Bioinformatics (SIB), et de la Ressource de l'Information de Protéine (PIR).

En 2002, EBI, SIB, et PIR ont joint leurs forces pour créer le consortium d'UniProt. Jusqu'à récemment, EBI et SIB ont ensembles produit les banques SwissProt et TrEMBL , alors que PIR produisait la base de données de protéine (PIR-PSD) et d'autres telles que iProClass.
Ces bases de données ont coexisté avec des priorités différentes pour l'annotation des protéines. Les membres ont décidé de mettre leurs ressources, efforts, et expertise en commun.

UniProt Release 3.0 est constituée de Swiss-Prot Version 45.0 du 25-Oct-2004 avec 163 235 entrées et de TrEMBL Version 28.0 du 25-Oct-2004 avec 1449 374 entrées.

Toute la documentation pour cette banque est disponible sur le serveur d'Expasy ou de l'EBI :
`http://www.expasy.uniprot.org/index.shtml`
`http://www.ebi.uniprot.org/index.shtml`

3.1.7. Les systèmes d'interrogation des banques

Chaque banque de séquences a son propre système d'interrogation, avec quelquefois des versions différentes proposées par certains serveurs. Pour chaque version, une note explicative donne la syntaxe de la requête (étiquettes, connecteurs logiques, caractères de substitution ..)

Des outils d'interrogation qui permettent des interrogations dans de nombreuses banques de séquences, généralistes ou spécialisées, ont été développés, les plus connus et utilisés sont :

SRS (Sequence Retrieval System)
Logiciel créé par Etzold et Argos en 1993, qui est proposé par de nombreux sites serveurs : il permet une interrogation simple ou croisée sur un éventail large de bases en biologie moléculaire. Chaque serveur SRS met à disposition un ensemble spécifique de bases données. C'est un outil d'accès privilégié aux banques de séquences généralistes et spécialisées.

Le serveur SRS d'Infobiogen (`http://www.infobiogen.fr/srs`) dispose à ce jour de 202 "librairies" dont environ 180 sont des banques de séquences généralistes ou spécialisées.

ENTREZ
Ce serveur permet l'interrogation des banques de séquences Medline et PubMed, GenBank, EMBL, DDBJ, PIR, SwissProt, PRF, PDB, SNP, CDD, UniSTS, OMIM .. (Medline et pubMed sont des bases de données bibliographiques biologiques)

Ce serveur est au NCBI (USA) : `http://www.ncbi.nlm.nih.gov/Entrez/index.html`

ACNUC

Système d'interrogation (au choix) des banques EMBL, Genbank, PIR, Hovergen, NRSub, NRBact , etc .. au total une trentaine de banques.

Ce serveur est accessible au Pôle Bio- Informatique Lyonnais (PBIL)

`http://pbil.univ-lyon1.fr/databases/acnuc/acnuc.html`

DBGET

Système d'interrogation des banques PubMed, EMBL, Genbank, SwissProt, PIR, PRF, LITDB, PDB, PDBSTR, EPD, Prosite, Ligand, PMD, AA- Index, OMIM.

Ce serveur est accessible au GenomeNet (Japon) :

`http://www.genome.jp/dbget/dbget.links.html`

3.1.8. La qualité des données des banques généralistes

Malgré des contrôles lors de la création d'une entrée, ces banques généralistes souffrent de nombreux défauts dont la plupart sont de la responsabilité des auteurs, nous pouvons citer :

- Variabilité de l'état des connaissances sur les séquences : la connaissance des caractéristiques biologiques des séquences et la détermination de leur fonction exige un travail expérimental et une analyse (conduisant à l'annotation de la séquence) qui doivent se surajouter à l'étape automatisée et systématique du séquençage.
- Erreurs dans les séquences : origine du fragment qui peut être contaminé – erreur due à la technologie ou encore à la méthodologie
- Biais d'échantillonnage :
    - biais d'échantillonnage taxonomique (les organismes à partir desquels les séquences ont été extraites sont inégalement représentés)
    - biais d'échantillonnage des séquences (les gènes des génomes étudiés sont inégalement représentés dans chacun d'eux)
    - redondance des données (il est fréquent de trouver plusieurs entrées correspondant à un même gène - certains gènes sont séquencés à la fois sous forme d'ARNm et de fragments génomiques - certaines séquences ont été saisies plusieurs fois dans la banque - certains gènes ont été séquencés à plusieurs reprises.

Malgré cela, il faut souligner l'énorme richesse que représentent ces banques généralistes de données dans le cadre de l'analyse des séquences :

- la majorité des séquences connues y sont réunies en un seul ensemble, c'est un élément fondamental pour la recherche de similitudes avec une nouvelle séquence.

- la grande diversité d'organismes représentée permet d'aborder des analyses de type évolutif.
- un autre intérêt de ces bases réside dans l'information, contenue ou pointée par un lien, qui accompagne les séquences (annotations, expertise, bibliographie, lien vers des banques spécialisées)

3.2. Les banques ou bases de données de séquences spécialisées

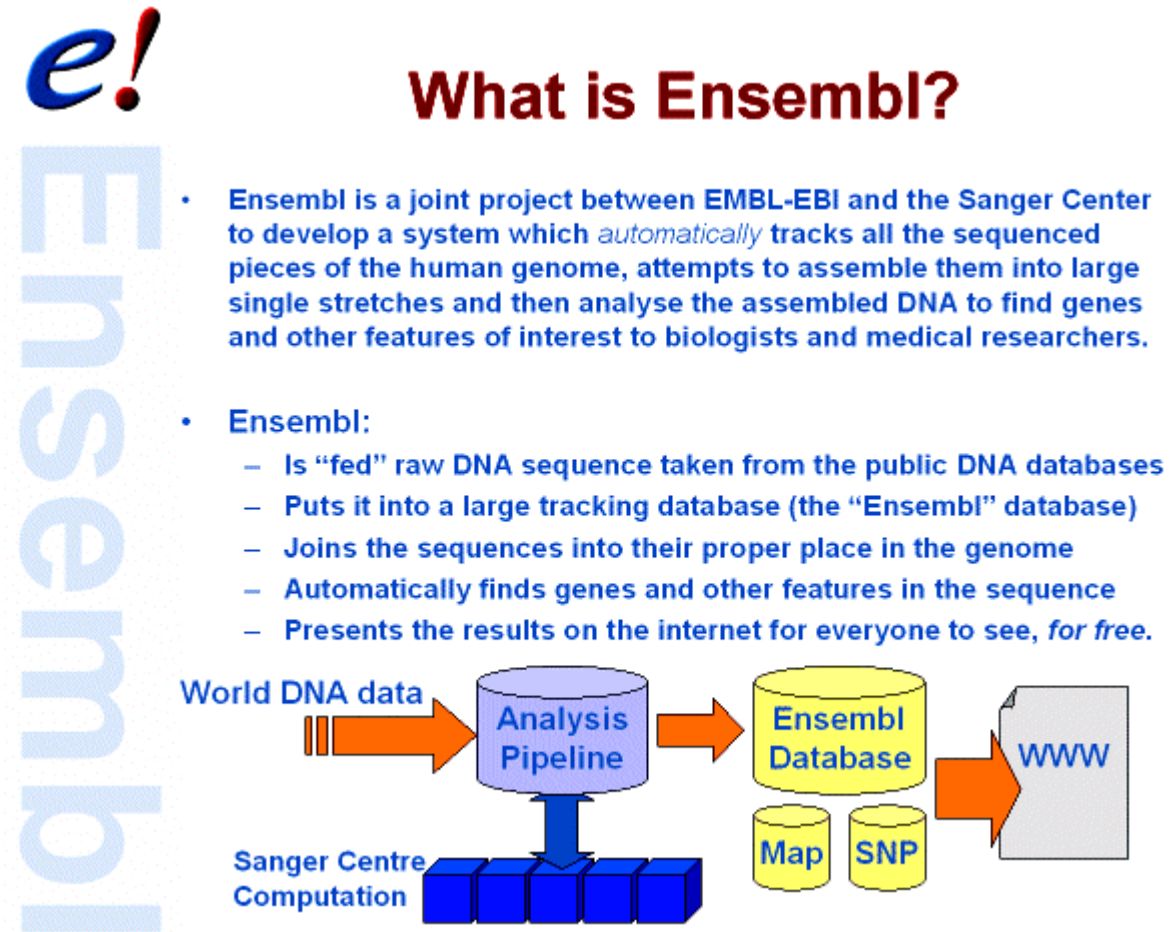
Pour des besoins spécifiques, de nombreuses bases de données spécialisées ont été créées, certaines sont pérennes et continuent d'être développées et mises à jour, d'autres sont laissées à l'abandon et enfin d'autres ont disparu. On en dénombre à cette date un peu plus d'un millier, accessibles directement par le Web. La nature ainsi que la quantité d'informations sont très variable.

3.2.1. Organisme

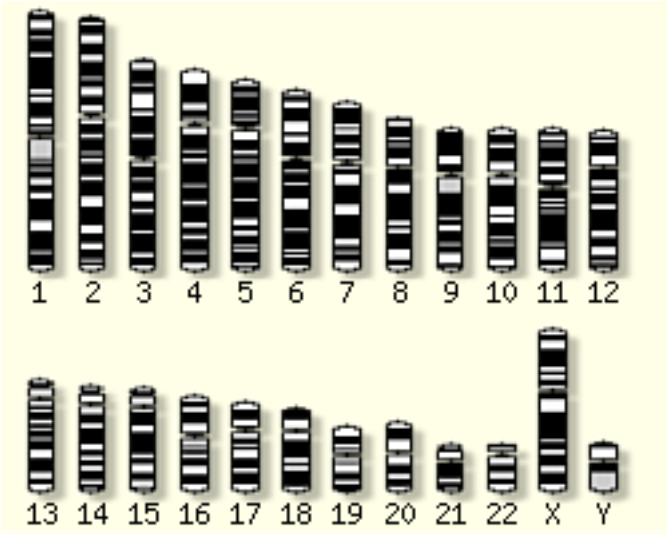
Ces banques regroupent les données pour un organisme particulier, ou un groupe, contenant tout ou partie des informations suivantes :

- carte physique chromosomique
- carte génétique et liaison
- clonage positionnel pour les gènes
- EST (marqueurs de séquences exprimées)
- Banque d'ADNc
- Banque de vecteurs de clonage
- Gène et expression
- Cytogénétique et anomalies chromosomiques
- Gène et maladie - Oncogènes
- etc …

Parmi tous les serveurs accessibles par Internet (de l'ordre de sept cents), citons le projet "Ensembl" mené conjointement par l'EBI (Angleterre) et le "Sanger Institute" (Angleterre).
Ce projet regroupe à peu près toutes les informations disponibles pour un organisme (actuellement 14 organismes disponibles).



Voici une partie de la page d'entrée pour l'homme (Human Genome Browser) :

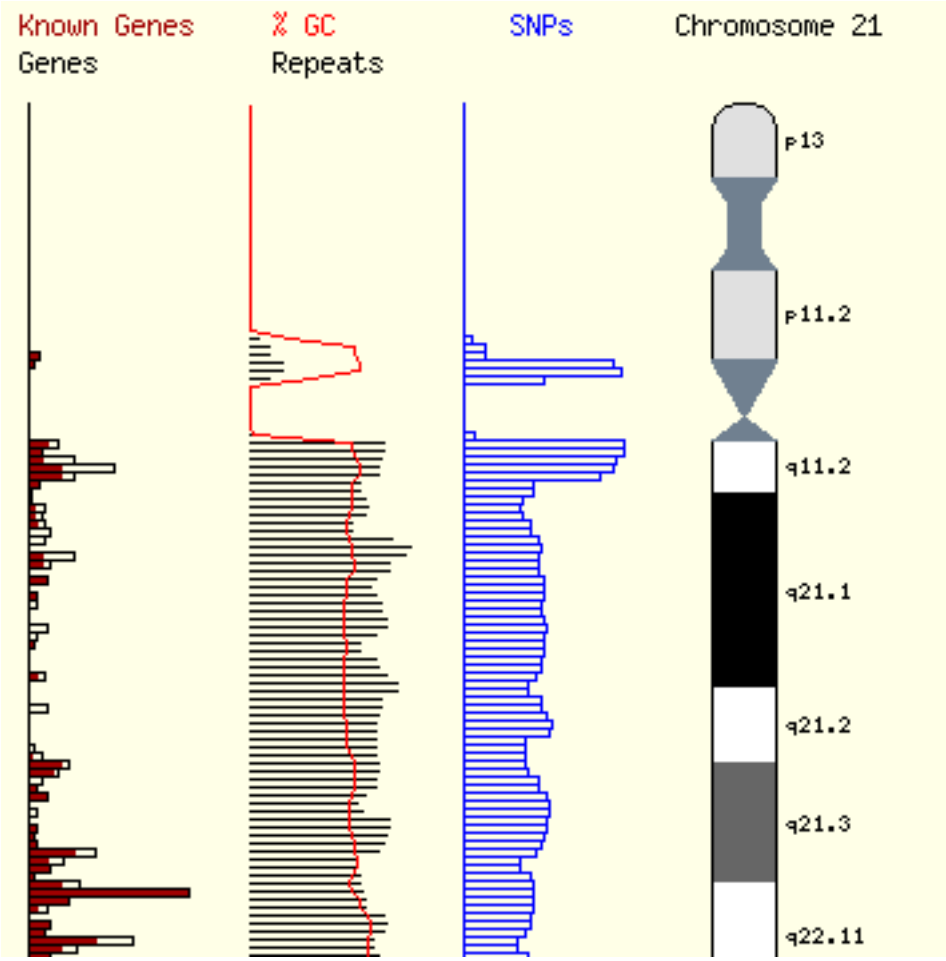


Un simple click permet de visualiser le chromosome entier avec les marqueurs physiques et des informations générales sous forme de diagramme (voir ci-dessous) :

- gènes connus
- pourcentage de GC répétés
- single Nucleotide Polymorphism

ainsi qu'un lien vers la base OMIM*

Voici une partie de la représentation du chromosome 21 :



En cliquant sur une région particulière, celle-ci va être détaillée avec toutes les informations connues : marqueurs, ensemble de gènes (putatifs ou connus), promoteurs, protéines avec lien sur les banques généralistes, carte de restriction, synténie (conservation de groupe de liaison entre espèces), etc ..
Toutes les informations sont disponibles sur le serveur :

`http://www.ensembl.org/`

*OMIM (Online Mendelian Inheritance in Man) : banque de gènes et de désordres génétiques, créée à l'Université de Johns Hopkins (USA) et mise à disposition par le Web par le NCBI (Bethesda – USA) et qui comprend à ce jour 14 000 entrées.
Toutes les informations sont disponibles sur le serveur :

`http://www3.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM`

3.2.2. Banques nucléiques spécialisées

Elles sont spécialisées dans les informations suivantes :
- EST, ADNc
- ARN

- Structure secondaire d'ARN
- Signaux et éléments de régulation
- Sondes, amorces
- Alignements
- Famille de gènes

3.2.3. Banques protéiques spécialisées

Elles sont spécialisées dans les informations suivantes :

- Motifs
- Alignement
- Classification structurale
- Familles de protéines
- Interactions
- Enzymes
- Modifications protéiques post-traductionnelles
- Pathologies
- Gels bidimensionnels
- Bases protéiques sur l'interaction et la thermodynamique des protéines

3.2.4. Banques immunologiques

Elles sont spécialisées dans les informations suivantes :

- Séquences
- Récepteur (cellule T, par exemple)
- Complex MHC (Major Histocompatibility Complex)
- Système HLA

3.2.5. Banques Structure 2D ou 3D

Elles sont spécialisées dans les informations suivantes :

- Coordonnées 3D de protéines *
- Structure secondaire des protéines
- Domaines structuraux
- Centre actif des enzymes
- Complexes récepteurs-ligands
- Atlas de topologie structurale des protéines

* La banque des données 3D des protéines est la "Protein Data Bank" (PDB, créée en 1971 comme archive des données cristallographiques au BNL (Brookaven National Laboratory – USA). Elle comprend à ce jour 27 855 entrées de structures établies soit aux rayons X, soit par résonance magnétique nucléaire ou encore par modélisation théorique.

Toutes les informations sont disponibles sur le serveur :

`http://www.rcsb.org/pdb/`

3.2.6. Les systèmes d'interrogation des banques spécialisées

Chaque banque de séquences a son propre système d'interrogation qui quelquefois peut être réduit à sa plus simple expression tel une liste dans un fichier.
Certaines de ces banques spécialisées ont été intégrées dans les systèmes d'interrogation des banques généralistes comme SRS, ENTREZ, ACNUC, DBGET (paragraphe 3.1.7).

4. Ressemblance ou similitude entre séquences

La caricature du biologiste moléculaire, la plus actuelle, montrerait un biologiste ayant "péché" une séquence et s'exclamerait à quoi tu ressembles ou en quoi diffères-tu!
Nous ne poserons pas la question de la pertinence de tout cartographier et de tout séquencer, sainte quête, en vue d'obtenir le secret de la vie, mais simplement nous allons feuilleter quelques pages du bréviaire.
Deux points de vue pour répondre à la question de la similarité entre deux séquences :

1 - l'analyse du mathématicien qui considère une séquence comme un mot construit à partir d'un alphabet et qui a des méthodes opératoires pour établir des fonctions de mesure

2 - l'analyse ou aussi l'expertise du biologiste qui se référera, au delà des réponses précédentes, à d'autres connaissances que la séquence primaire : toutes les propriétés biologiques

La recherche de similitude entre séquences constitue souvent la première étape des analyses de séquences. La comparaison de séquences biologiques, ainsi que leur alignement, nécessite la mise en oeuvre de procédures de calcul et de modèles biologiques permettant de quantifier la notion de ressemblance ou similitude entre ces séquences.

Une ressemblance entre séquences peut indiquer par exemple :

- une fonction biologique proche
- une structure tridimensionnelle semblable
- une origine commune
- etc ..

Une similitude entre séquences est souvent un argument en faveur d'une homologie : deux séquences sont homologues si elles ont un ancêtre commun. Remarquons quand-même qu'il n'y a pas de d'équivalence entre similitude et homologie : deux séquences peuvent avoir un degré de similitude conséquent sans être homologues et deux séquences peuvent être homologues avec un degré de similitude faible.

Cette notion d'homologie reflète le dogme fondamental de l'évolution biologique :

- les régions fonctionnelles des gènes ou de leurs produits (sites catalytique, de fixation, etc.) sont soumises à la sélection : elles sont relativement préservées par l'évolution car des mutations trop importantes leur feraient perdre leurs fonctions. Cet argument est complété par le principe de parcimonie.
- les régions non fonctionnelles, qui ne subissent aucune sélection, divergent rapidement.
- les nouveaux gènes apparaissent surtout par remaniement de gènes ancestraux : on peut souvent déduire la fonction de la plupart des gènes par comparaison avec les gènes «homologues » d'autres espèces.

Que ce soit par une représentation graphique, un calcul de distance ou de score, la ressemblance entre deux séquences doit pour le biologiste aboutir à la représentation d'un alignement qui est la mise en correspondance des symboles des 2 séquences avec insertion d'espaces pour que les longueurs soient identiques.

```
Seq 1 V A R F I E V A I D L A S T F A - - C Y Q
      | | | | | | :     | :       | |     | | |   (symboles classiques)
Seq 2 V A R F I E L D T D V - - Y F A S T C Y Q
```

Pour une position donnée de l'alignement, nous pouvons avoir :

- identité (|) : les symboles sont identiques dans les deux séquences (*anglais match*)
- insertion/délétion ou ins/del (s/- ou -/s) : le symbole dans l'une des deux séquences est un espace (insertion dans la séquence où le symbole est un espace, délétion dans la séquence où le symbole est autre que l'espace) (*anglais gap*)
- substitution : les symboles ne sont pas identiques (*anglais mismatch*)
- similarité ou substitution conservative (:) : les symboles ne sont pas identiques mais considérés comme similaires dans l'évaluation de la ressemblance (voir les matrices de substitution)

4.1. Méthodes globales

Ce sont des méthodes qui considèrent les séquences dans leur totalité et aboutissent à un alignement de toute la première séquence avec la seconde.

4.1.1. Dot plot

Avec deux séquences de longueur m et n, on construit un tableau (dot-matrix) ainsi :
- une croix (ou un point) si $x_i = y_j$ (où x_i est un élément de la première séquence et y_j un élément de la deuxième séquence, sinon rien. La vision d'une suite de croix consécutives dans une diagonale souligne des identités entre des parties des deux séquences

```
  M Q N W E T T A T T N Y E Q H N A W Y N
N   x               x         x       x
W     x                             x
E       x               x
T         x x   x x
V             ?
T         x x   x x
T               x x
N   x             x
Y                   x             x
D                     ?
Q x                     x
H                         x
```

La ressemblance des séquences est "lue" dans les diagonales du tableau.
La représentation qui est polluée par un bruit de fond non négligeable (point ou petit segment diagonal) peut être améliorée par :

- filtrage : les points ne sont représentés qu'à partir d'un nombre suffisant (seuil) dans une diagonale
- introduction la notion de similarité : le biologiste peut indiquer qu'une substitution de valine (V) par alanine (A) ne change pas les propriétés biologiques (de même acide glutamique (E) par l'acide aspartique (D) : utilisation des matrices de substitution.

Après filtrage et utilisation de la similarité :

```
  M Q N W E T T A T T N Y E Q H N A W Y N
N   x
W     x
E       x
T         x x
V             x
T               x
T                 x
N                   x
Y                     x
D                       x
Q                         x
H                           x
```

Cette méthode peut permettre aussi de visualiser les répétitions internes dans une séquence avec la construction d'un tableau d'une séquence avec elle-même : la diagonale principale sera évidemment une ligne d'identité complète et les petites diagonales indiqueront les répétitions.

4.1.2. Distance d'édition – programmation dynamique

Levenshtein introduit en 1965 la notion de distance entre deux séquences A et B :

- c'est le nombre minimum de substitutions, de délétions ou insertions requises pour transformer A en B

Une distance définie ainsi a les propriétés d'une distance métrique, à savoir:

- $d(A,B) \geq 0$
- $d(A,B) = 0$ si $A=B$
- $d(A,B) = d(B,A)$
- $d(A,B) + d(B,C) \geq d(A,C)$

Cette distance se calcule de manière récursive (on parle de programmation dynamique) :
Soient les séquences A de n lettres $\{a_1 .. a_n\}$ et B de m lettres $\{b_1 .. b_m\}$:

$d(A,B) = d(a_n,b_m)$ avec $d(a_0,b_0) = 0$ et $d(a_i,b_j)$ infini si (i<0 ou j<0)
$d(a_i,b_j)$ est défini par la relation de récurrence :

$$d(a_i,b_j) = \text{minimum} \begin{cases} d(a_{i-1},b_j) + w(a_i,-) & \text{délétion de } a_i \\ d(a_{i-1},b_{j-1}) + w(a_i,b_j) & \text{substitution de } a_i \text{ par } b_j \\ d(a_i,b_{j-1}) + w(-,b_j) & \text{insertion de } a_i \end{cases} \quad \text{(dans cet ordre)}$$

où w est une fonction de poids égale à 0 si $a_i = b_j$ et à 1 dans tous les autres cas
et on définit simultanément un pointeur qui indique la position de la valeur minimum précédente de cette manière :

$$p(i,j) = \text{minimum} \begin{cases} (i-1,j) \\ (i-1,\ j-1) \\ (i,\ j-1) \end{cases} \quad \text{(dans cet ordre)}$$

Exemple pédagogique:

		A	C	G	T	G	C	G	C
	0	1	2	3	4	5	6	7	8
C	1	1	1	2	3	4	4	5	5
G	2	2	2	1	2	2	3	3	4
A	3	2	3	2	2	3	3	4	5
G	4	3	3	2	3	2	3	3	4
C	5	4	3	3	3	3	2	3	3
T	6	5	4	4	3	4	3	3	4

p(8,6)= (7,6) p(7,6)= (6,5)
p(6,5)= (5,4) p(5,4)= (4,3)
p(4,3)= (3,2) p(3,2)= (2,1)
p(2,1)= (1,0) p(1,0)= (0,0)

La distance est de 4 : on peut obtenir B à partir de A par 2 délétions et 2 substitutions.

Un **alignement métrique** entre les deux séquences s'obtient en partant du pointeur p(n,m) et en remontant en arrière (backtrack) jusqu'à la position précédant p(0,0).
L'alignement métrique proposé est donc :

```
A C G T G C G C
  | |   | |
- C G A G C T -
```

Remarque : il n'y a pas un seul alignement entre deux séquences : celui-ci dépend de l'ordre utilisé pour la définition des pointeurs.
Cette notion de distance d'édition n'est pas satisfaisante pour les biologistes et c'est avec quelques modifications que quelques algorithmes dérivés sont proposés.
Les deux problèmes importants sont :

- celui des "gap" : doit-on par exemple utiliser une fonction particulière qui traduise une pénalité non linéaire qui peut être soit sur-pénalisante soit sous-pénalisante. Bien évidemment le score et l'alignement seront dépendants du choix.
- celui de la similarité dans les substitutions : utilisation de matrice de substitution. Bien évidemment le score et l'alignement seront dépendants du choix de la matrice.

4.1.3. Needleman et Wunsch

Ce fût le premier programmes de comparaison de séquences, publié en 1970. Il ne calcule pas la différence entre deux séquences mais la similarité. Considérons deux séquences A(1,n) B(1,m)

Le tableau est rempli ligne après ligne (en partant de la dernière) et pour chaque ligne colonne après colonne (en partant de la dernière) en obéissant à la règle suivante :

- le score S(i,j) est le nombre maximum de correspondance entre les deux parties de séquences A(i,n) et B(j,m) (en prenant tous les chemins possibles à partir de (i,j)) et en appliquant la valuation suivante :

- score (s) pour une identité 1
- score pour une substitution, une insertion ou délétion 0

La formule de récurrence est :

$$S(i,j) = \max \begin{cases} \text{si } a_i = b_{j+1} \;\; S(i, j+1) - 1 + s(a_i,b_j), \text{ si non } S(i, j+1) + s(a_i,b_j) \\ \text{si } a_{i+1} = b_{j+1} \; S(i+1, j+1) - 1 + s(a_i,b_j), \text{ si non } S(i+1, j+1) + s(a_i,b_j) \\ \text{si } a_{i+1} = b_j \;\; S(i+1, j) - 1 + s(a_i,b_j), \text{ si non } S(i+1, j) + s(a_i,b_j) \\ \text{avec évidemment } S(n+1, j) = S(i, m+1) = 0 \end{cases}$$

La similarité entre les deux séquences est égale à la valeur de S(1,1) et l'alignement est un graphe qui a pour origine S(1,1) et parcourt la matrice pour des i et j croissant en recherchant l'élément maximal voisin.

Ce programme a depuis été modifié, en particulier par l'utilisation des matrices de substitution pour la fonction score. Il est toujours utilisé de nos jours pour l'alignement de deux séquences.

Pour "l'alignement d'une séquence contre une banque" qui est tout simplement les alignements d'une séquence avec chacune des séquences d'une banque (en ne retenant que les meilleurs scores), ce programme n'est plus utilisé car il demande des ressources très importantes et de plus, son temps d'exécution s'accroît proportionnellement au nombre de séquences de la banque. Ajoutons aussi que la sensibilité au score défini pour les "gap" peut aboutir à oublier des alignements locaux importants.

4.2. Méthodes locales

Les méthodes globales de recherche de ressemblance de deux séquences ont révélé à l'usage les deux principaux inconvénients suivants et cela qu'elles soient basées sur la distance d'édition ou la similarité :

- une lenteur des programmes augmentant avec l'accroissement du nombre de séquences dans les banques
- une perte d'alignements locaux pour des séquences homologues mais éloignées : pour les biologistes, les ressemblances locales ont une valeur non négligeable.

De nouveaux programmes ont été développés qui rendant compte de similarités locales : ce sont des heuristiques qui supposent que les scores de ressemblance locales indiquent une similarité globale.

Les plus significatifs et utilisés sont les suivants :

4.2.1. Smith et Waterman

Cet algorithme (Smith et Waterman 1981) est directement inspiré de celui Needleman et Wunsch et est utilisé pour des alignements locaux. La principale différence vient du fait que n'importe quelle case de la matrice de comparaison peut être considéré comme point de départ pour le calcul des scores finaux. Si ce score devient inférieur à zéro, la case est réinitialisée à zéro et peut être considérée comme un nouveau point de départ.

L'algorithme identifie les sous-séquences maximales de deux séquences par programmation dynamique. Un matrice de score est construite à l'aide d'une formule de récurrence (en reprenant les mêmes notations):

$$S(i,j) = \max \begin{cases} S(i-1, j-1) + s(ai, bj) \\ S(i-k, j) - Wk \\ S(i, j-k) - Wk \\ 0 \\ \text{avec } S(i, 0) = S(0, j) = 0 \end{cases}$$

S(i,j) est le maximum de similarité entre deux segments se terminant en Ai et Bj. Une séquence maximale est identifiée en trouvant l'élément maximal du tableau et en le parcourant pour des indices décroissants jusqu'à la valeur nulle et en recherchant l'élément voisin maximal.

Dans l'article original, le score d'une identité est nul, celui d'un mismatch (-1/3) et Wk est égal à 1+ (1/3)k où k est la longueur du gap. Les dernières versions de ce programme utilisent pour le score d'une identité ou d'un mismatch une matrice de substitution.

4.2.2. Fasta

Pearson et Lipman (1983) ont fait les remarques suivantes :

- les ressemblances recherchées au "niveau biologique" concernent des fragments de séquences
- de plus, dans ces fragments, la fréquence de substitution est beaucoup plus grande que celle d'insertion ou délétion

Leur programme est basé sur la méthode de la diagonale, que l'on peut approcher intuitivement par la représentation "dot-matrix". La ressemblance se définit par comparaison de paire de fragments de chacune des séquences (fragment : partie de même longueur de chacune des deux séquences, en dot-matrix c'est un morceau d'une diagonale). Ces deux parties contiennent des mots communs séparés par des zones de substitution. Un fonction score est attribué pour un fragment et la ressemblance est mesurée par le fragment de score maximum.

L'algorithme se divise en 4 étapes :

- précodage des séquences en k-uple : mots de longueur k (4 à 6 pour les acides nucléiques, 1 ou 2 pour les protéines). Ceci permet une efficacité beaucoup plus grande pour la deuxième étape
- recherche du fragment de plus haut score pour chaque diagonale qui est le score de la diagonale (fragment = suite de mots séparés par de régions de substitution dont la longueur maximale est prédéfinie)
- les scores des dix meilleures diagonales vont être recalculés en utilisant une matrice de substitution (PAM 250 dans les premières versions). C'est ce score qui est listé dans les résultats sous l'appellation init1dans les programmes antécédents à FASTA (FASTP ..). Pour FASTA, ce score (**initn**) est recalculé en essayant d'enchaîner à partir de la meilleure diagonale les fragments restants des 9 autres diagonales en tenant compte des insertions ou délétions dues au changement de diagonale.
- Les résultats par rapport aux séquences de la banque sont classés à l'aide du score précédent, et pour les meilleurs, un alignement et un nouveau score (**Opt**) entre la séquence requête et la séquence de la banque, sont calculés à partir de l'algorithme de Needleman et Wunsch légèrement modifié.

En 1990, Pearson a ajouté une statistique avec les scores :

- il définit le **z-score** qui correspond au score maximum attendu normalisé (c'est à dire que le z-score est dérivé du score Opt avec une correction en fonction de la longueur de la séquence)
- il définit la **E-value** dont on peut dire que plus elle est faible (plus le nombre de comparaisons présentant un bon score est petit), moins on a de chance de trouver l'alignement par chance dans les banques.

4.2.3. Blast

Karlin et Altschul (1993) ont introduit une statistique pour leur programme BLAST (Basic Local Alignment Search Tool) qui rend compte de la pertinence d'une ressemblance locale.

La stratégie de la recherche consiste à trouver tous les HSPs (fragments similaires) entre la séquence recherchée et les séquences de la base.
Pour déterminer un HSP, des mots de longueur fixe sont identifiés dans une **première étape** entre la séquence recherchée et la séquence de la banque.

- dans le cas des acides nucléiques, cela revient à des recherches d'identité entre les deux séquences sur des segments de longueur fixe (généralement 11).
- dans le cas des protéines, on effectue d'abord une liste de mots similaires pour chaque mot de longueur fixe (généralement 3) de la séquence recherchée et l'on repère ensuite dans la banque les séquences qui possèdent au moins un de ces mots.

Un mot similaire est un mot qui, comparé avec un mot de la séquence recherchée, obtient un score supérieur à un **score seuil**, calculé avec une matrice de substitution.

Dans une **deuxième étape**, on cherche à étendre la similitude dans les deux directions le long de chaque séquence, à partir du mot commun, de manière à ce que le score cumulé puisse être amélioré. L'extension s'arrêtera dans les trois cas suivants:

- si le score cumulé descend d'une quantité x donné par rapport à la valeur maximale qu'il avait atteint.
- si le score cumulé devient inférieur ou égal à zéro.
- si la fin d'une des deux séquences est atteinte.

Dans une **troisième étape** , la signification des segments similaires obtenus est évaluée statistiquement. Le score de la similarité est normalisé et évalué en unité standard d'information (bit). Ensuite la probabilité (**E-value**) d'avoir un tel score au hasard est calculé pour cette longueur de segment (m) dans une banque contenant au total (n) nucléotides ou acides aminés. Seuls seront conservés et classés les HSP significatifs, c'est à dire ceux dont la probabilité est la plus faible.

D'autres versions de BLAST ont été développées depuis :

- gapped blast : introduction de gap pendant la deuxième étape
- PSI-blast (Position Specific Iterated Blast) : il donne la possibilité de relancer itérativement Blast sur les séquences résultats : pour chaque nouvelle itération, celles-ci sont traduites en un "profil ou PSSM" (consensus matérialisé par une matrice) qui est recherché à son tour sur la banque choisie initialement. Les itérations s'arrêtent lorsqu'il y convergence, c'est à dire lorsque les séquences résultats de l'itération n sont identiques à celles de l'itération n-1

- PHI-blast (Pattern Hit Initiated Blast) : à partir d'une séquence protéique donnée et d'un motif spécifique (expression régulière) contenu dans cette séquence, PHI-blast recherche dans une banque protéique les séquences homologues en utilisant le motif comme ancrage pour l'alignement

4.3. Matrices de substitution

Dans tous les programmes de ressemblance, un système de score qui attribue un coût aux opérations élémentaires (identité, substitution, délétion et insertion) est défini.
Ces matrices seront donc fonction :

- de la nature des séquences (nucléique ou protéique)
- de la définition de la ressemblance : soit distance, soit similarité
- des propriétés ou des relations des lettres (nucléotide ou aminoacide) de la séquence que l'on veut mettre en évidence dans la ressemblance : par exemple des propriétés physico-chimiques, des relations de structure, des relations d'homologie, etc..

4.3.1. Matrices pour l'ADN

Les plus utilisées sont :
matrice unitaire identité :

	A	C	G	T
A	**1**	0	0	0
C	0	**1**	0	0
G	0	0	**1**	0
T	0	0	0	**1**

matrice de transition/transversion :

- transition : purine(A,G)↔purine, pyrimidine(C,T)↔pyrimidine : score de 1
- transversion : purine↔pyrimidine : score de 0

	A	C	G	T
A	**3**	0	***1***	0
C	0	**3**	0	***1***
G	***1***	0	**3**	0
T	0	***1***	0	**3**

matrice de Blast (identité) :

	A	C	G	T
A	**5**	-4	-4	-4
C	-4	**5**	-4	-4
G	-4	-4	**5**	-4
T	-4	-4	-4	**5**

4.3.2. Matrices pour les protéines

Plusieurs pondérations ont été proposées pour élaborer des matrices de substitution, basées sur :

- propriétés chimiques des chaînes latérales
- fréquence d'apparition des aminoacides dans les structures secondaires
- distance génétique : en relation avec le nombre de base à modifier dans le codon pour la substitution
- fréquence de substitution observée après superposition de structure 3D
- fréquence de substitution observée dans une série de protéines homologues

Les matrices les plus utilisées et reconnues comme les plus performantes sont :

- la série des PAM (Dayhoff 1978)
- la série des BLOSUM (Henikoff et Henikoff 1992)

Pour les deux séries, les auteurs choisissent un **lot de séquences homologues** et pour chaque paire d'aminoacides, ils vont étudier le nombre de substitutions observées, en déduire une probabilité de cette mutation en la pondérant par la fréquence d'apparition de chacun des deux aminoacides. La différence essentielle entre ces deux séries repose essentiellement sur le choix des lots de séquences et la façon de les aligner.

4.3.3. PAM (Dayhoff)

Dayhoff a utilisé 1572 séquences protéiques groupées en 71 familles avec un total de 1600 mutations et les séquences d'une même famille ont au maximum 15% de différence.

Les hypothèses pour calculer une matrice sont les suivantes :

1) le processus d'évolution est un processus de Markov d'ordre 0 : la probabilité de substitution ne dépend ni de la position ni des évènements antérieurs
2) les évènements de substitution sont indépendants du contexte, c'est à dire des aminoacides adjacents
3) le fait de prendre des protéines homologues avec un maximum de différence de 15% pour chaque famille permet :
   - d'éviter le problème des mutations multiples (X -> Y -> Z)
   - de faciliter les alignements en diminuant l'impact des "gap"

Cette matrice symétrique est construite en plusieurs étapes :

1) Matrice des mutations observées

Un arbre phylogénétique est construit pour chaque famille ainsi qu'une séquence ancestrale. Dans chaque famille on comptabilise les mutations pour chaque aminoacide en prenant les séquences deux à deux et on fait une sommation pour l'ensemble des familles. Chaque mutation observée de type (X -> Y) est comptée dans les deux sens (X -> Y et Y -> X). Soit $A_{X,Y}$ l'élément de la matrice de l'événement substitution de X par Y (X -> Y).

2) Construction de la matrice des probabilités de substitutions (PAM)

Cette construction prendra en compte :

- la normalisation par rapport à la fréquence de l'aminoacide
- la longueur de la séquence
- la distance évolutive entre les séquences

L'élément de la matrice de transition est défini par Dayhoff comme le produit de la probabilité conditionnelle de la substitution et de la mutabilité relative de l'aminoacide considéré :

Le premier terme du produit est égal : $\frac{A_{X,Y}}{\sum_{Y \neq X} A_{X,Y}}$ fréquence de l'événement de la substitution de X en Y parmi tous les évènements de substitution de X.

Le deuxième terme est égal à : $R_X^M = \frac{\sum_{Y \neq Y} A_{X,Y}}{f_X \left(\frac{N_T^M}{L} \right) (100)}$

où $\sum_{Y \neq X} A_{X,Y}$ est le nombre de substitutions de X observées (X ->)

$A_{X,Y}$ est le nombre de substitutions de X en Y observées (X -> Y)

$f_X \left(\frac{N_T^M}{L} \right) (100)$ est le nombre de substitutions de X attendues (X ->) pour 100 aminoacides,

où f_X est la fréquence de l'aminoacide X, N_T^M le nombre total de mutations dans le nombre total de positions L examinées.

Pour obtenir des probabilités, on introduit un facteur d'échelle λ et chacun des termes est multiplié par λ. Un élément de la matrice est :

$$T(X,Y) = \lambda \frac{A_{X,Y}}{\sum_{Y \neq X} A_{X,Y}} \frac{\sum_{Y \neq Y} A_{X,Y}}{f_X \left(\frac{N_T^M}{L} \right) (100)} = \lambda \frac{A_{X,Y}}{f_X \left(\frac{N_T^M}{L} \right) (100)}$$

et $T(X,X) = 1 - \lambda \sum_{Y \neq X} T(X,Y) = 1 - \lambda R_X^M$

Pour tenir compte des distances évolutives (une unité d'évolution est une période où on observe 1% de substitution), Dayhoff a proposé de définir la matrice de base 1PAM (1 *Point Accepted Mutation*) qui est définie pour une conservation de 0,99. Cette valeur de conservation permet de calculer λ (égal à 1 dans ce cas) :

$$\sum_X f_X T(X,X) = \sum_X f_X (1 - \lambda R_X^M) = \sum_X f_X - \sum_X f_X (\lambda R_X^M) = 1 - \sum_X f_X (\lambda R_X^M) = 0.99$$

A partir de cette matrice on peut facilement calculer 2PAM .. NPAM : l'hypothèse d'un processus de Markov d'ordre 0 implique :
2PAM = (1PAM)x(1PAM), … NPAM= $(1PAM)^N$

3) <u>Matrice des "odds"</u>

Le calcul de la matrice a été effectué en prenant des protéines homologues : ce qui nous intéresse c'est de connaître les "chances" d'une substitution pour des protéines non homologues : il faut donc normaliser par rapport au fait d'obtenir les deux aminoacides de manière aléatoire dans chacune des deux séquences (odds ratio) :

$$M_{odd}(X,Y) = \frac{PAM(X,Y)}{f_X f_Y}$$ (f : fréquence de l'aminoacide)

4) <u>Matrice des "log-odds" (MDM)</u>

Pour calculer un score de similarité à l'aide de la matrice précédente (odds-matrix), il faut multiplier chacune des positions : en construisant la matrice des logarithmes des éléments de la matrice des "odds", il suffira de sommer. La matrice MDM (Mutation Data Matrix) est calculée en prenant le logarithme des éléments de la matrice des "odds", en le multipliant par 10 et en l'arrondissant à l'entier supérieur. C'est cette matrice qui est utilisée dans les programmes : par abus de langage on l'appelle PAM.

Voici la matrice de mutation (MDM) :
PAM 250 (MDM 250)

```
C 12
S  0  2
T -2  1  3
P -3  1  0  6
A -2  1  1  1  2
G -3  1  0 -1  1  5
N -4  1  0 -1  0  0  2
D -5  0  0 -1  0  1  2  4
E -5  0  0 -1  0  0  1  3  4
Q -5 -1 -1  0  0 -1  1  2  2  4
H -3 -1 -1  0 -1 -2  2  1  1  3  6
R -4  0 -1  0 -2 -3  0 -1 -1  1  2  6
K -5  0  0 -1 -1 -2  1  0  0  1  0  3  5
M -5 -2 -1 -2 -1 -3 -2 -3 -2 -1 -2  0  0  6
I -2 -1  0 -2 -1 -3 -2 -2 -2 -2 -2 -2 -2  2  5
L -6 -3 -2 -3 -2 -4 -3 -4 -3 -2 -2 -3 -3  4  2  6
V -2 -1  0 -1  0 -1 -2 -2 -2 -2 -2 -2 -2  2  4  2  4
F -4 -3 -3 -5 -4 -5 -4 -6 -5 -5 -2 -4 -5  0  1  2 -1  9
Y  0 -3 -3 -5 -3 -5 -2 -4 -4 -4  0 -4 -4 -2 -1 -1 -2  7  10
W -5 -2 -5 -6 -6 -7 -4 -7 -7 -5 -3  2 -3 -4 -5 -2 -6  0   0 17
   C  S  T  P  A  G  N  D  E  Q  H  R  K  M  I  L  V  F   Y  W
```

Remarque : la valeur s de (N)MDM(X,Y) sera :

- s = 0 : les probabilités observées et attendues sont identiques
- s < 0 : les probabilités observées sont inférieures aux attendues
- s > 0 : les probabilités observées sont supérieures aux attendues

4.3.4. BLOSUM (Henikoff et Henikoff)

Henikoff et Henikoff (1992) ont étudié les substitutions d'aminoacides dans les blocs ("blocks") qui sont un alignement multiple d'une partie des séquences sans "gap" (segments) dans les régions les plus conservées des familles de protéines. Ces auteurs ont utilisé plus d'une centaine de familles de protéines pour un total de blocs supérieurs à 2000. Les hypothèses pour calculer une matrice sont identiques à celles pour le calcul de PAM :

1) le processus d'évolution est un processus de Markov d'ordre 0 : la probabilité de substitution ne dépend ni de la position ni des évènements antérieurs
2) les évènements de substitution sont indépendants du contexte, c'est à dire des aminoacides adjacents

Le calcul de la matrice de substitution se fait en 3 étapes :

1) <u>Matrice des probabilités des substitutions observées</u>

Pour une position donnée dans un bloc, on recense toutes les paires d'aminoacide (i,j) : pour un bloc de n séquences (alignement multiple de n séquences), nous avons pour une position donnée à examiner toutes les combinaisons deux à deux, soit $C_n^2 = \frac{n!}{(n-2)!\,2!} = \frac{n(n-1)}{2}$.

La probabilité d'une paire d'aminoacide (i,j) est : $q_{i,j} = \frac{N_{i,j}}{\sum_{i=1}^{20}\sum_{j=1}^{i} N_{i,j}}$ $(1 \le j \le i \le 20)$ pour les vingt aminoacides classiques, $N_{i,j}$ étant le nombre de paires (i,j).

2) Calcul d'un odds-ratio

A partir de la probabilité précédente, calculons le rapport $r_{i,j} = \frac{q_{i,j}}{e_{i,j}}$ où $q_{i,j}$est la probabilité d'une paire (i,j) et $e_{i,j}$ la probabilité attendue d'une paire (i,j). Calculons $e_{i,j}$ à l'aide de $q_{i,j}$:
La probabilité attendue d'une paire (i,j) sera la probabilité d'avoir i dans une paire multipliée par la probabilité d'avoir j :
probabilité d'avoir i dans une paire (i,j) : $p_i = q_{i,i} + \sum_{j \neq i} \frac{q_{i,j}}{2}$ (dans les paires (i,j) on a une chance sur deux d'avoir i).

d'où : $\begin{cases} e_{i,j} = p_i\, p_i \text{ pour } i = j \\ e_{i,j} = p_{i,j} + p_{j,i} = 2 p_i\, p_j \text{ pour } i \neq j \end{cases}$

3) Matrice des log-odds

Comme dans le cas de PAM, il est plus facile de sommer pour calculer un score : à partir des $r_{i,j}$, on calcule la matrice de substitution S par : $S_{i,j} = \log_2\left(\frac{q_{i,j}}{e_{i,j}}\right)$, logarithme dans la base 2 du rapport $r_{i,j}$.
La valeur s de $S_{i,j}$ sera :

- $s = 0$: les probabilités observées et attendues sont identiques
- $s < 0$: les probabilités observées sont inférieures aux attendues
- $s > 0$: les probabilités observées sont supérieures aux attendues

La matrice utilisée dans les programmes de recherche (BLOSUM : *BLOcks SUbstitution Matrix*) est dérivée de la matrice de substitution (S) en multipliant chacun des éléments de S par 2 et en l'approximant à l'entier le plus proche.
Plusieurs matrices BLOSUM % ont été calculées pour des pourcentages divers d'identité : BLOSUM 62 correspond à un calcul effectué sur des blocs pour lesquels les séquences deux à deux ont au moins 62% d'identité. Toutefois pour minimiser les identités, le poids de séquences ayant plus de % d'identité est minoré dans le calcul (pour BLOSUM62, la minoration est d'environ 25% sur le nombre de séquences ayant une identité supérieure à 62%).

La matrice BLOSUM 62

```
A  4
R -1  5
N -2  0  6
D -2 -2  1  6
C  0 -3 -3 -3  9
Q -1  1  0  0 -3  5
E -1  0  0  2 -4  2  5
G  0 -2  0 -1 -3 -2 -2  6
H -2  0  1 -1 -3  0  0 -2  8
I -1 -3 -3 -3 -1 -3 -3 -4 -3  4
L -1 -2 -3 -4 -1 -2 -3 -4 -3  2  4
K -1  2  0 -1 -3  1  1 -2 -1 -3 -2  5
M -1 -1 -2 -3 -1  0 -2 -3 -2  1  2 -1  5
F -2 -3 -3 -3 -2 -3 -3 -3 -1  0  0 -3  0  6
P -1 -2 -2 -1 -3 -1 -1 -2 -2 -3 -3 -1 -2 -4  7
S  1 -1  1  0 -1  0  0  0 -1 -2 -2  0 -1 -2 -1  4
T  0 -1  0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 -1  1  5
W -3 -3 -4 -4 -2 -2 -3 -2 -2 -3 -2 -3 -1  1 -4 -3 -2  11
Y -2 -2 -2 -3 -2 -1 -2 -3  2 -1 -1 -2 -1  3 -3 -2 -2   2  7
V  0 -3 -3 -3 -1 -2 -2 -3 -3  3  1 -2  1 -1 -2 -2  0  -3 -1  4
   A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T   W  Y  V
```

4.3.5. Utilisation et comparaison des matrices de substitution

La plupart des programmes d'alignement ou ressemblance des séquences qui utilisent les matrices de substitution proposent un choix pour ces dernières. Les matrices retenues sont les PAM et les BLOSUM, considérées comme les plus performantes. Rappelons que pour ces deux dernières, les NPAM sont toutes obtenues à partir de 1PAM alors que pour chaque BLOSUM%, l'ensemble des blocs est recomposé en fonction du pourcentage.

Henikoff et Henikoff, d'une part, et Altschul ont comparé ces matrices de substitution PAM et BLOSUM en calculant leur entropie relative : $H = \sum_{i=1}^{20} \sum_{j=1}^{i} (q_{i,j})\ S_{i,j}$

Voici les correspondances :

PAM120	BLOSUM 80
PAM160	BLOSUM 62
PAM250	BLOSUM 45

en rappelant que pour chaque type de matrice les séquences sont :

BLOSUM 80	<---------->	BLOSUM 45
PAM1	<---------->	PAM250
plus conservées	<---------->	plus divergentes

Il est conseillé par ces différents auteurs d'utiliser pour les programmes de recherche dans les banques :

- des BLOSUM à fort indice et des PAM à faible indice pour chercher des séquences proches
- des BLOSUM à faible indice et des PAM à fort indice pour chercher des séquences peu conservées

Les programmes BLAST , implantés sur le serveur du NCBI, proposent pour des recherches de ressemblance de protéines le choix entre BLOSUM 80, 62, 45, PAM 70, 30, ceux implantés sur le serveur d'Infobiogen donnent un choix plus large : BLOSUM 90, 80, 62, 50, 45, PAM 250, 70, 30 : tous les deux proposent par défaut BLOSUM 62.

5. Analyse et prédiction sur les séquences

Voici un aperçu non exhaustif de quelques domaines d'analyse ou de prédictions sur les séquences nucléiques.

5.1. Analyse de séquences

Les programmes qui sont regroupés dans cette famille ont des algorithmes qui ne sont pas basés sur des règles obtenues à partir de propriétés compilées de séquences déjà connues.

5.1.1. Analyse de séquences nucléiques

- Composition, usage des codons, traduction ..
- Sélection d'amorces (PCR), calcul de la température de fusion (Tm)
- Cartes de restriction (recherche de sous-séquence)
- Régions riches en .. (CpG par exemple)
- Recherche de répétitions
- Recherche de motifs à l'aide d'expression régulières
- Assemblage à partir de fragments séquencés

5.1.2. Analyse de séquences protéiques

- Composition, masse moléculaire ..
- Calcul du coeeficient théorique d'absorption
- Calcul du pH du point isoélectrique (pI) théorique
- Fragments peptidiques obtenus par hydrolyse enzymatique
- Recherche de motifs à l'aide d'expression régulières
- Calcul d'index de d'hydrophilicité, d'hydropathie ..

5.2. Prédiction sur les séquences

Les programmes qui sont regroupés dans cette famille ont des algorithmes qui sont basés sur des règles obtenues à partir de propriétés compilées de séquences déjà connues.

5.2.1. Prédiction sur les séquences nucléiques

- Prédiction de structure d'un gène :
- Prédiction de signaux
- Prédiction de structure secondaire d'un ARN

5.2.2. Prédictions sur les séquences protéiques

- Famille de protéines (Pfam)
- Famille et signature (Prosite)
- Localisation subcellulaire
- Segments nucléaires
- Segments transmembranaires
- Sites spécifiques (O-glycosilation, antigénicité, etc..)
- Peptide signal
- Structure secondaire
- Structure 3D

5.3. Exemple : recherche de signaux pour la prédiction de gènes chez les procaryotes

Vous avez vu la structure classique des gènes chez les procaryotes, qu'on peut décomposer par la suite de signaux suivants:

Le **promoteur**, la région reconnue par la polymérase à ARN, juste en amont du site d'initiation de la transcription comprend trois éléments:

- la boite de Pribnow (TTGACa) vers -35,
- la boite TATA (TAtAAT) vers –10
- le site d'initiation de la transcription

Ce sont ces 3 séquences que les facteurs sigma reconnaissent (lient la polymérase).

L'**opérateur** : séquence reconnue par d'éventuelles protéines régulatrices, il peut se glisser entre la boîte TATA et le site d'initiation de la transcription

La séquence de **Shine-Dalgarno**, ou Ribosome Binding Site (RBS), environ 3 à 9 nucléotides avant le codon ATG.

Le **codon d'initiation** (ATG), quelquefois GUG

Un **cadre de lecture ouvert** (ORF)

Un **codon stop** (TAA, TAG, TGA)

Un **terminateur** (rho-dépendant ou rho-indépendant)

A partir de signaux connus pour différents gènes dans divers organismes, on peut définir un motif consensus pour ces derniers. A partir de ce consensus, par analyse informatique d'une séquence de procaryote, on peut rechercher les signaux potentiels dans cette séquence.

La construction d'un motif consensus consiste à définir à partir d'un lot de motifs connus une matrice de poids décrivant pour chaque position, la fréquence d'apparition de chaque résidu (ici nucléotide).

5.3.1. Exemple : matrice consensus RBS

Un lot de séquences alignées est utilisé pour définir les six positions consécutives du signal.

- matrice de fréquences pour chaque position dans le site

position	*1*	*2*	*3*	*4*	*5*	*6*
G	$f_{1,G}$	$f_{2,G}$	$f_{3,G}$	$f_{4,G}$	$f_{5,G}$	$f_{6,G}$
A	$f_{1,A}$	$f_{2,A}$	$f_{3,A}$	$f_{4,A}$	$f_{5,A}$	$f_{6,A}$
T	$f_{1,T}$	$f_{2,T}$	$f_{3,T}$	$f_{4,T}$	$f_{5,T}$	$f_{6,T}$
C	$f_{1,C}$	$f_{2,C}$	$f_{3,C}$	$f_{4,C}$	$f_{5,C}$	$f_{6,C}$

- matrice du rapport de fréquence observée / fréquence nucléotide dans les séquences (odds)

position	*1*	*2*	*3*	*4*	*5*	*6*
G	$\frac{f_{1,G}}{f_G}$	$\frac{f_{2,G}}{f_G}$	$\frac{f_{3,G}}{f_G}$	$\frac{f_{4,G}}{f_G}$	$\frac{f_{5,G}}{f_G}$	$\frac{f_{6,G}}{f_G}$
A	$\frac{f_{1,A}}{f_A}$	$\frac{f_{2,A}}{f_A}$	$\frac{f_{3,A}}{f_A}$	$\frac{f_{4,A}}{f_A}$	$\frac{f_{5,A}}{f_A}$	$\frac{f_{6,A}}{f_A}$
T	$\frac{f_{1,T}}{f_T}$	$\frac{f_{2,T}}{f_T}$	$\frac{f_{3,T}}{f_T}$	$\frac{f_{4,T}}{f_T}$	$\frac{f_{5,T}}{f_T}$	$\frac{f_{6,T}}{f_T}$
C	$\frac{f_{1,C}}{f_C}$	$\frac{f_{2,C}}{f_C}$	$\frac{f_{3,C}}{f_C}$	$\frac{f_{4,C}}{f_C}$	$\frac{f_{5,C}}{f_C}$	$\frac{f_{6,C}}{f_C}$

où les f_N sont les fréquences de chacun des nucléotides dans l'ensemble des séquences. C'est le rapport de la fréquence observée du nucléotide pour la position du signal sur la fréquence attendue pour ce nucléotide.

- matrice des log-odds (valeurs pour *Bacillus Subtilis*)

position	*1*	*2*	*3*	*4*	*5*	*6*
G	8	5	0	5	5	1
A	1	3	8	1	-2	5
T	1	-1	0	1	-1	0
C	0	-1	2	6	0	4

matrice dérivée de la précédente en prenant le logarithme du rapport des fréquences, de la même manière que lors de la construction des matrices de substitution BLOSUM. De nombreuses matrices ont été construites pour différents organismes pour représenter le signal RBS dont la longueur peut varier de 6 à 9 nucléotides.
Un programme de prédiction pour les séquence génomiques de *Bacillus Subtilis* sélectionnera les signaux de longueur six dont le score est supérieur à un seuil.

5.3.2. Exemple : détermination de la longueur d'un signal

Dans le paragraphe précédent, nous avons vu la construction d'une matrice de consensus pour un signal, mais nous n'avons pas indiqué comment la longueur de celui-ci est déterminé. Cela peut être fait par des expériences du mutagénèse dirigée, mais aussi par analyse informatique.
Connaissant la zone dans laquelle se trouve le signal dans un alignement de séquences, la théorie de l'information pourra nous fournir une mesure du 'contenu informatif' de chaque position :

$I(i) = \sum_{b=A,C,G,T} f_{b,i} \log_2(f_{b,i}) - \sum_{b=A,C,G,T} P_{b,i} \log_2(P_{b,i})$, où $f_{b,i}$ est la fréquence observée de la base (b) à la position (i), et $P_{b,i}$ la probabilité théorique (attendue) de la base (b) à la position (i). Cette valeur sera d'autant plus grande qu'à la position donnée, la composition en base sera "biaisée" et différente de $P_{b,i}$.

En général on considère que $P_{b,i}$ est indépendant de la position (i) et souvent cette valeur est définie comme égale à 0,25.

En simplifiant, on a donc : $I(i) = \sum_{b=A,C,G,T} f_{b,i} \log_2(f_{b,i}) + 2$

L'analyse de la valeur de I(i) pour la zone sélectionnée permettra de :

- délimiter la longueur signal
- trouver les positions les plus significatives dans le signal

5.4. Exemple : prédiction de structure secondaire des protéines

Citons trois méthodes statistiques, qui utilisent la méthode de la fenêtre et ont chacune leurs tableaux de valeurs de référence construits à partir de la connaissance spatiale de structure de protéines.

5.4.1. Chou- Fassman

Les auteurs ont calculé les valeurs des paramètres de conformation d'un aminoacide de se trouver dans une structure α-hélice, β-sheet ou β-turn à partir de la ***structure cristalline de 29 protéines***.

Le paramètre (ou score) pour une position donnée dans la séquence est calculée sur une fenêtre de 4 aminoacides en ne tenant compte que des trois suivants (nucléation d'une structure).

$SC_s(i) = \frac{1}{4} \sum_{j=i}^{j=i+3} sc_s(j)$ où s fait référence aux différentes structure α, β ou β-turn et $sc_s(j)$ est la valeur du score de l'aminoacide dans le tableau de référence.

Est ajoutée à ceci un tableau de fréquence d'apparition des aminoacides participant à une structure de coude. Pour ce paramètre, la valeur est calculée ainsi :

$$F(i) = \prod_{j=i}^{j=i+3} f(j)$$

Une fois ce tableau de quatre valeurs calculées pour chaque position le programme les analyse avec les règles suivantes : il prend tout d'abord en considération la valeur la plus grande des trois structures prédites et puis :

- **α-hélice** : une région de 4 résidus consécutifs, où SC_α est supérieur SC_β et à SC_{turn}, initie une α-hélice et la région est étendue à droite et à gauche jusqu'à la rencontre de SC_α inférieure à 1. Cette région doit remplir les conditions suivantes :
  - longueur au moins de six résidus
  - pas de proline à l'intérieur de l' α-hélice ou du coté C-terminal

- **β-sheet** : une région de trois résidus, où SC_β est supérieur SC_α et à SC_{turn}, initie une structure de β-sheet et la région est étendue à droite et à gauche jusqu'à la rencontre de SC_β inférieure à 1.

- **β-turn** : il faut SC_{turn} supérieur à SC_α et SC_β pour une région avec quatre aminoacides avec pour le premier une valeur de F supérieure à un seuil ($0{,}75\ 10^{-4}$)

5.4.2. Gor method

Cette méthode tient compte du fait que la probabilité d'un aminoacide d'appartenir à un type de structure secondaire dépend de la nature et de la position de ses voisins. Pour cela, les auteurs (Garnier et Robson) utilise la théorie de l'information ou probabilité conditionnelle.

Soient deux événements, x et y, soit P(x|y) la probabilité conditionnelle que x advienne sachant que y est advenu. On appelle l'information associé à l'événement x contraint par y :

$I(x;y) = \log(P(x|y) / P(x))$ { si $P(x|y) = P(x)$ x indépendant de y alors $I(x;y) = 0$ }
{ si $P(x|y) > P(x)$ y favorise x $I(x;y) > 0$ }
{ si $P(x|y) < P(x)$ y défavorise x $I(x;y) < 0$ }

Si y se décompose en deux événements y_1, y_2, nous avons :

$$I(x;y_1, y_2) = \log \{ P(x|y_1, y_2) / P(x) \}$$
$$= \log \{ P(x|y_1, y_2) / P(x|y_1) \} + \log \{ P(x|y_1) / P(x) \}$$
$$= I(x;y_2|y_1) + I(x;y_1)$$

Le premier terme représente l'effet de y_2 sur l'événement x, sachant que y_1 a eu lieu.

ou encore de manière plus générale :

$$I(x;y_1, y_2, y_1, y_3 ..y_n) = I(x;y_1) + I(x;y_2|y_1) + I(x;y_3|y_1, y_2) + .. + I(x;y_n|y_1, .., y_{n-1})$$

Considérons que pour le premier événement, il n'y ait que deux résultats possibles x et son événement contraire $\bar{x}$, nous pouvons définir la préférence de y pour l'événement x comme

$I(S = x : \bar{x} ; y) = I(S = x ; y) - I(S = \bar{x} ; y)$ ou encore :

$$= \log\{ P(S=x|y) / P(S=x) \} - \log\{ P(S=\bar{x} | y) / P(S=\bar{x}) \}$$

Remarquons que $P(S=\bar{x}) = 1 - P(S=x)$ et $P(S=\bar{x} | y) = 1 - P(S=x|y)$

L'application à la prédiction de structure va se faire de la manière suivante :
S sera l'ensemble de l'état α-hélice ou non α-hélice et nous prendrons pour les événements représentés par $y_1..y_n$ les positions dans la séquence (R). Bien sur il y aura les mêmes définitions pour les structures β, β-turn et coil.
La préférence pour une structure α-hélice pour un aminoacide à la position j dans la séquence de longueur n sera évaluée par le paramètre :

$$I(S_j=H: \bar{H}; R_1 .. R_n)$$

Nous ferons les approximations suivantes :

1 - $I(S_j=H: \bar{H}; R_1 .. R_n) \cong I(S_j=H: \bar{H}; R_{j-m} .. R_{j+m})$

Ceci indique que l'influence des amino acides voisins sera limitée à une fenêtre centrée de longueur (2m +1). Les auteurs ont pris m = 8

2 - En se référant au développement précédent, nous avons :

$$I(S_j=H: \bar{H}; R_{j-m} .. R_{j+m}) = I(S_j=H: \bar{H}; R_j) + I(S_j=H: \bar{H}; R_{j-1} | R_j) + I(S_j=H: \bar{H}; R_{j+1} | R_j, R_{j-1}) + ...$$

$$+ \; I(S_j{=}H{:}\ \overline{H};\, R_{j+m} \mid R_{j-m}, .., R_j, .., R_{j+m-1})$$

Cette formule, qui contient (2m+1) facteurs, est simplifiée et nous obtenons les deux méthodes "GOR" suivantes:

<u>information directionnelle</u> (GOR II)

$$I(S_j = H : \overline{H}\ ;\ R_{j-m} .. R_{j+m}) \ \cong \ \sum_{m} I(S_j = H : \overline{H}\ ;\ R_{j+m})$$

L'information du résidu à la position (j+m) est la même quel que soit le résidu à la position j

et **<u>information par paire</u>** (GOR III)

$$I(S_j = H : \overline{H}\ ;\ R_{j-m} .. R_{j+m}) \ \cong \ I(S_j = H : \overline{H}\ ;\ R_j) \ + \ \sum_{m,m\neq 0} I(S_j = H : \overline{H}\ ;\ R_{j+m} \mid R_j)$$

Le deuxième paramètre informationnel ne tient compte que des paires : amino acide à la position j et amino acide à la position (j+m).

Les tableaux de référence sont calculés comme pour la méthode précédente à l'aide de la ***structure cristalline de 75 protéines***.
Pour la méthode "information directionnelle", il comprend pour un type de structure 17 x 20 valeurs ce qui fait pour les quatre structures :
4 x 17 x 20. Elles ont été calculées à partir de la formule :

$$I(S_j{=}x{:}\ \overline{x}\ ;\ R_{j+m}) \ = \ \log\{\ P(S_j{=}x \mid R_{j+m}) \,/\, P(S{=}x)\ \} \ - \ \log\{\ P(S{=}\overline{x}\ \mid R_{j+m}) \,/\, P(S{=}\overline{x})\ \}$$

où les probabilités sont des fréquences observées dans ces 75 protéines, et en tenant compte que x et $\overline{x}$ sont des événements contraires.

5.4.3. Gascuel et Goldmard

Comme la précédente, cette méthode tient compte du fait que la probabilité d'un aminoacide d'appartenir à un type de structure secondaire dépend de la nature et de la position de ses voisins. Un score est calculé pour chacun des états possibles (S : α-hélice, extended et coil) en déclarant que l'état de l'aminoacide considéré est d'autant plus influencé par un autre aminoacide que celui-ci est proche de l'aminoacide considéré :

$$CBLF(i, S) \ = \ N(S) \sum_{j=i-n}^{j=i+m} I(j, S) \ P(j, S)$$

N(S) est un facteur de normalisation associé à chaque état, I(j, S) mesure la préférence pour l'aminoacide en position j pour l'état S et P(j, S) est un poids qui dépend uniquement de la

position relative de j par rapport à i (P(i, S) =1).

Les auteurs ont remarqué que l'influence des aminoacides n'était par forcément symétrique pour chacun des états de structure étudié. Ils ont défini leur fenêtre de calcul ainsi :

pour les structures en	α-hélice la fenêtre est	n = -6	m = 11
	extended	n = -3	m = 3
	coil	n = -6	m = 3

Pour chacune des trois structures, un tableau de valeurs pour les différents paramètres a été calculé à partir de la structure cristallines de 65 protéines.

La probabilité d'un état de structure S à une position considérée est égale à :

$$P(i, S) = \frac{e^{CBLF(i,S)}}{e^{CBLF(i,H)} + e^{CBLF(i,E)} + e^{CBLF(i,C)}}$$

et l'état associé est celui de plus forte probabilité.

5.5. Annotation "in silico" des séquences génomiques

Quelques éléments bibliographiques

Ressemblance

- Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W., Myers, and David J. Lipman (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**:403-10.

- Karlin, Samuel and Stephen F. Altschul (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA* **87**:2264-68.

- Karlin, Samuel and Stephen F. Altschul (1993). Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc. Natl. Acad. Sci. USA* **90**:5873-7.

- Levenshtein V. (1965) Binary codes capable of correcting deletions, insertions, reversals. *Cybernetics and Control Theory* **10** (8): 707-710

- Needleman S. And Wunsch C. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443-453

- Pearson W and Lipman D. (1988) Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* **85**, 2444-2448

- Sellers P. (1974) On the theory and computation of evolutionary distances. *SIAM J. Appl. Math.* **26**, 787-793

- Smith T. and Waterman M. (1981) Identification of common molecular subsequence. *J. Mol. Biol .* **147**, 195-197

- Smith T. Waterman M. and W. Fitch (1981) Comparative biosequence metrics. J. *Mol. Evol.* **18**, 38-46

- Wilbur W. and Lipman D. (1983) Rapid similarity searches of nucleic acid and protein data banks. *Proc. Natl. Acad. Sci. USA* **80**, 726-730

- Wilbur W. and Lipman D. (1984) The context dependant comparison of biological sequences. *SIAM J. APPL. MATH.* **44**, 557-567

Matrices de substitution

- Dayhoff M., Barker W. and Hunt L. (1983) Establishing homologies in protein sequences. *Methods in Enzymol.* **91**, 524-545

- Henikoff S. and Henikoff J. (1992) Amino acid substitution matrices from protein blocks. *Proc. Nat. Acad. Sci. USA* **89**, 10915-10919

Prédiction de structure

- Chou, P.Y. and Fasman, G. D. (1978) *Ann. Rev. Biochem.* **47**, 251-276

- Gascuel O. and Golmard J.L. (1988) *CABIOS* **4**, 357-365

- Gibrat, J.F., Garnier, J. and Robson, B. (1987) *J. Mol. Biol.* **198** , 425-443