République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université des Frères Mentouri Constantine
Faculté des Sciences de la Nature et de la Vie
Département : Biologie Animale

Intitulé du Master : Toxicologie

Intitulé de la matière : Bio-informatique

Semestre : 01

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 sedemander 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

* 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 : ***Devries*** 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 expériences 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 de1024 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".

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).

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).

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 àarchitecture 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).

- 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 1981sous 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.
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) (FrederickSanger)
- Annonce du VAX 11/780 : premier super-mini-ordinateur.
1979 : Début de USENET, échanges de émail et Newsgroups.
1980 : 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 logiciels 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 : 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.

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).

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.

1-Définition de la bio-informatique

Un domaine de recherche qui analyse et interprète des données biologiques, au moyen de méthodes informatiques, afin de créer de nouvelles connaissances en biologie

Qu'est-ce que la bioinformatique ?

* L'approche *in silico* de la biologie
* Trois activités principales :
- Acquisition et organisation des données biologiques
- Conception de logiciels pour l'analyse, la comparaison et la modélisation des données
- Analyse des résultats produits par les logiciels

***Quelques liens utiles en bioinformatique**

- La Société Française de Bio-Informatique (SFBI)
  http://sfbi.impg.prd.fr/
- Logiciels pour la biologie de l'Institut Pasteur
  http://bioweb.pasteur.fr/
- Le Pôle Bioinformatique Lyonnais (PBIL)
  http://pbil.univ-lyon1.fr/pbil.html
  http://npsa-pbil.ibcp.fr/
- L'Institut Européen de Bioinformatique (EBI)
  http://www.ebi.ac.uk/
- Les outils de protéomique d'ExPASy
  http://www.expasy.org/tools/
- Le centre national de bioinformatique (NCBI, USA)
  http://www.ncbi.nlm.nih.gov/

2. Banques et basses 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 organisées 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.

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 généralistes et spécialisées

2-1 : Banques de séquences généralistes :

Ce type correspondent à une collecte des données la plus exhaustive possible et qui offrent finalement un ensemble plutôt hétérogène d'informations.

Il existe des banques de données génomiques et des banques de données protéiques.

*Les trois principales banques de données nucléiques sont :

1. GenBank de NCBI (**N**ational **C**enter for **B**iotechnology **I**nformation) : http://www.ncbi.nlm.nih.gov/. Créée par IntelliGenetics en 1982. Jusqu'en octobre 2004 elle contenait 38 941 263 entrées (ou séquences par auteur)

2. EMBL de EMBO (**E**uropen **M**oleculary **B**iology **O**rganization): http://www.ebi.ac.uk/embl/ . La banque EMBL contient 44 538 943 entrées jusqu'en octobre 2004.

3. DDBJ : **D**na **D**ata **B**ase of **J**apan : http://www.ddbj.nig.ac.jp/searches-e.html Créeé en 1986 et diffusée par NIG (National Institute of Genetics, Japan). En octobre 2004, elle contenait 37 926 117 entrées.

L'inconvénient majeur de ces banques de données reste le manque de vérification des données et retardent un peu dans l'insertion de nouvelles séquences.
Elles ont, cependant quelques qualités :

• Un très grand nombre de séquences : par exemple en 2000, EMBL contenait déjà 109 bases nucléiques, SwissProt contenait 31 millions d'acides aminés.
• Une grande variétés d'organismes (homme, animaux, végétaux, microorganismes).

*Les deux principales banques de données protéiques sont :

1. PIR-NBRF : D'abord, elle fut créée par la NBRF (National Biomedical Researche Foundation) en 1984. Actuellement, elle constitut un ensemble du à la fusion de MIPS (Martinsried Institute for Protein sequences, Munich Allemagne) et de JIPID Japan International Protein Information Database). Elle contient 283 416 entrées.

2. SwissProt : Créée par le biochimiste Amos BAIROCH en 1986 à l'Université de Genève. actuellement développée en collaboration entre l'Institut Suisse de BioInformatique (ISB-SIB et l'EBI [http://www.ebi.ac.uk]). Elle contient la séquence de quasiment toutes les protéines découvertes jusque là (http://www.unige.ch/presse/campus/pdf/c48/decouvertes.pdf). Elle contient plus de 320 000 séquences de protéines provenant de quelques 10 000 espèces différentes.

2- 2 :Banques ou bases de séquences spécialisées

Ce type 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.

*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. En citer quelque type :

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 …

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. 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

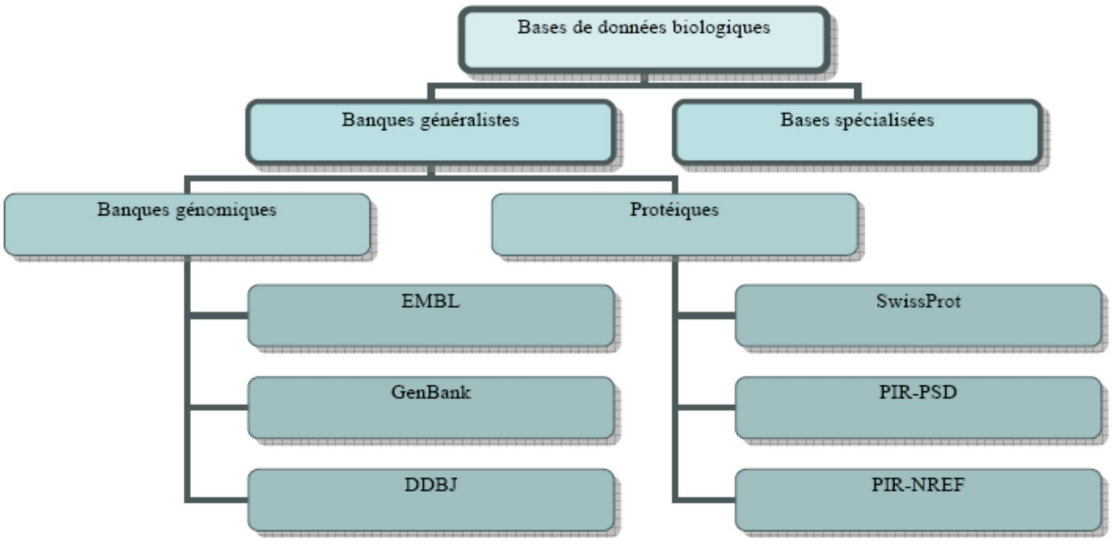
4. Banques immunologiques

Elles sont spécialisées dans les informations suivantes :
- Séquences
- Récepteur (cellule T, par exemple)
- Complex MHC (Major Histocompatibility Complex)
- Système HLA

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

Voici un schéma qui résume et représente les déférentes basses de donnes biologiques



3-Interrogation des bases de données

On peut interroger une BD pour plusieurs raisons :
• Pour connaître la séquence d'un gène ou d'une portion de ce gène
• Pour connaître la structure primaire d'une protéine
• Pour comparer deux séquences, …
Le résultat de l'interrogation des BD est une fiche descriptive de la molécule. On parlera alors d'une **entrée** (ou fiche descriptive de la séquence recherchée). La ùstructure d'une entrée est presque la même quelque soit la BD interrogée.

Pourquoi séquencer les génomes ?

* Intérêt économique
- Médecine
- Biotechnologies
- Environnement
* Intérêt scientifique
- Evolution des espèces
- Fonctionnement des cellules
- Etude des êtres vivants
* Utilité publique
- Nutrition
- Propagation des maladies
- Environnement

- **Une série de TD**

4-ressemblance ou similitude entre séquences (la comparaison)

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, nécessitent 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 ..

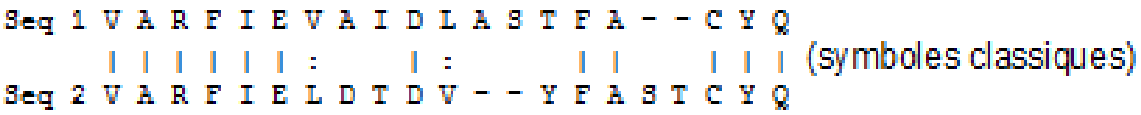
Une similitude entre séquences est souvent un argument en faveur d'une homologie : deux séquences sont homologues si elles ont un ancêtre commun. Remarquons quand-même qu'il n'y a pas de d'équivalence entre similitude et homologie : deux séquences peuvent avoir un degré de similitude conséquent sans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



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) .

En bioinformatique, la comparaison des séquences (ADN, ARN et/ou protéines : ARNm, régions 5'UTR, les EST, des clones, …) repose essentiellement sur la notion de l'**alignement**, et permet de déterminer le degré de ressemblance entre celles-ci (similitude ou identité en révélant des régions proches dans leurs séquences primaires). Cela peut alors indiquer que :
- La structure (primaire, secondaire ou tertiaire) des deux séquences est semblable,
- La fonction biologique est proche ou différente (dans le cas de la dissémilarité),
- L'origine des séquences alignées est commune ou éloignée (notion d'homologie), …

Cependant, la comparaison pour l'obtention d'un alignement optimal entre deux séquences biologiques, nécessite néanmoins la mise en oeuvre de procédures de calcu (algorithmes) et de modèles biologiques permettant de quantifier la notion de ressemblance entre ces séquences.

4-1 : TRAITEMENT DES SEQUENCES NUCLEIQUES (ADN ou ARN)

Il existe différentes méthodes pour la détermination de segments identiques entre deux séquences biologiques (on parle alors de fenêtres, de motifs ou de mots) sur lesquelles une similitude significative peut exister.

Notion de score : Le score élémentaire (noté "s") est une entité numérique que l'on attribue à chaque couple de nucléotides des deux séquences à comparer. Il prend la valeur de 1 lorsque les deux nucléotides des deux séquences sont identiques, et la valeur de zéro sinon.

Exemple :

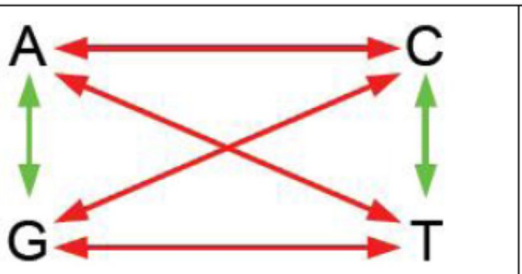
Séquence1	A	G	C	T	A	C	C	T	G	T	Score global : Total des scores
Séquence2	A	A	G	T	A	G	C	T	T	T	
Point de comparaison	1	2	3	4	5	6	7	8	9	10	
Score élémentaire (s)	1	0	0	1	1	0	1	1	0	1	**1+0+0+1+1+0+1+1+0+1=6**

Dans cet exemple, constatez qu'au niveau du premier point de comparaison (ou site de comparaison), les deux séquences contiennent le même nucléotide A, donc le score élémentaire (s) à ce point prend la valeur de 1 (s = 1).
Au deuxième point de comparaison, la séquence 1 contient un G et la séquence 2 contient un A. Elles sont donc différentes en ce point d'où un score élémentaire de zéro (s = 0)…
Au 10ème point de comparaison, les deux séquences contiennent le même nucléotide T donc un score élémentaire de 1.
Constatons que la somme des scores élémentaires est égale à six (s = 6). Donc il y a six points identiques entre les deux séquences ; soit 60% d'identité entre les deux séquences ([6/10] x100). On dit alors que le score global entre les deux séquences est égal à six. Le score a donc permis de quantifier la ressemblance entre les deux séquences.
La relation entre le score global (S) et les scores élémentaires (s) pour deux séquences
Question : Pourquoi avons-nous affecté la valeur de 1 dans le cas de l'identité et zéro dans le cas contraire ?
Il faut savoir qu'il existe une matrice (**matrice d'identité)** qui donne les valeurs de scores d'identité entre les séquences à comparer. Dans cette matrice, on attribue la valeur de 1 lorsque les deux nucléotides sont identiques et zéro s'ils ne le sont pas.

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

Matrice d'identité nucléique

Matrice d'identité nucléique
Il existe une autre matrice de score, qui tient compte de l'analogie structurale entre purines (A et G) et pyrimidines (C, T et U) et affecte des scores en fonction de cette ressemblance :
C'est la matrice de transition/transversion :
La substitution entre purines d'une part, et entre pyrimidines d'autre part est pondérée et n'a pas de score élémentaire nul au moment de la comparaison des séquences :



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

Matrice de transition/transversion

Matrice de transition/transversion
Remarque *: Quelle matrice utiliser ?*
En bioinformatique, on utilise beaucoup plus la matrice d'identité.

Recherche de segments identiques : La matrice de points

Elle permet une vue (méthode visuelle) englobant les similarités entre les régions des séquences à comparer.

Exemple de réalisation: On donne deux séquences x et y :

x=ACTCGGATT et y=AGCTCGGT

Cette méthode consiste à créer une matrice qui va contenir les deux séquences (la séquence x en horizontal et la séquence y en vertical) et de cocher les cases de cette matrice pour le seul cas où les nucléotides sont identiques (Match). Quand il n'ya pas identité on parle de Mismatch:

		Séquence s								
		A	C	T	C	G	G	A	T	T
Séquence t	A	X						X		
	G					X	X			
	C		X		X					
	T			X					X	X
	C		X		X					
	G					X	X			
	G					X	X			
	T			X					X	X

Sur cette matrice, constatons qu'il y a une diagonale formée de cinq cases. Donc le segment identique le plus long entre les deux séquences x et y contient cinq nucléotides identiques et consécutifs qui sont: **CTCGG**

		Séquence s								
		A	C	T	C	G	G	A	T	T
Séquence t	A									
	G									
	C		X							
	T			X						
	C				X					
	G					X				
	G						X			
	T									

Remarque : Dans le cas où les deux séquences sont complètement identiques, le résultat est une diagonal

		Séquence s								
		A	C	T	C	G	G	A	T	T
Séquence t	A	X						X		
	C		X		X					
	T			X					X	X
	C		X		X					
	G					X	X			
	G					X	X	X		
	A	X						X		
	T			X					X	X
	T			X					X	X

La méthode du Dot-Plot

Le dot-plot est utile pour déterminer de combien d'exons est composé un gène en le comparant à son ARNm et pour avoir une idée de la taille des introns et des exons.

Il existe un logiciel de dotplot interactif, Dotlet qui nécessite JAVA. Si JAVA n'est pas installé sur vos machines, vous pouvez utiliser Dottup5.

Le principe du dot-plot est basé sur la comparaison de fenêtres de longueur fixe que l'on déplace le long des séquences.

Soit deux séquences A et B à comparer et l la longueur de la fenêtre. On détermine sur la séquence A une première fenêtre de longueur l que l'on va comparer avec toutes les fenêtres possibles de même longueur, obtenues à partir de la séquence B. Un incrément est alors appliqué pour déterminer une deuxième fenêtre sur la séquence A, puis l'on recommence le balayage des comparaisons sur la séquence B. Si l'on choisit un incrément de 1 et que les séquences ont respectivement une longueur de m et n éléments, on effectuera de l'ordre de n x m comparaisons de fenêtres différentes.

Pour chaque comparaison entre deux fenêtres, un score est obtenu et l'on mémorisera uniquement les comparaisons dont les scores sont jugés significatifs, c'est-à-dire supérieurs ou égaux à un seuil que l'on s'est fixé. Par exemple lorsque le score correspond au minimum à 80% d'identité avec l'utilisation d'une matrice unitaire nucléique comme matrice de scores élémentaires6.

Considérons, par exemple, les deux séquences A et B suivantes :

Séq A = ATGTAATGCATG et Séq B = TATGTGAATG. La taille du motif (fenêtre) étant choisie égale à 5.

5 http://www.fil.univ-lille1.fr/~pupin/MRBS/comp_seq.html

6 http://genet.univ-tours.fr/gen001400_fichiers/chap5/genach5ec9.htm

La fenêtre formée des cinq premiers nucléotides de la séquence A est : ATGTA. Il faut la comparer avec toutes les fenêtres possibles de taille égale à cinq retrouvées sur la séquence B. Ces séquences sont :

1. TATGT
2. ATGTG
3. TGTGA

4. GTGAA
5. TGAAT
6. GAATG

Remarque : Au-delà du nucléotide G en 6ème position dans la séquence B, on ne peut plus avoir une fenêtre de taille égale à cinq nucléotides.

La première comparaison concerne les deux motifs suivants :

Fenêtre de la séquence A = ATGTA

Fenêtre de la séquence B = TATGT

La comparaison de ces deux segments donne un score égal à zéro car il n'ya aucun nucléotide de la séquence A qui soit identique à celui de la séquence B quelque soit le site de comparaison :

Séquence A	A	T	G	T	A	Score global
Séquence B	T	A	T	G	T	
Scores élémentaires	0	0	0	0	0	S = 0

Remarque : Sur la matrice qui contient la totalité des deux séquences A et B, allez au à la case d'intersection qui rejoint le milieu du premier segment de A et celui de B pour insérer la valeur du score global :

	A	T	G	T	A	A	T	G	C	A	T	G
T												
A												
T			0									
G												
T												
G												
A												
A												
T												
G												

En fixant le motif de la séquence A (ATGTA), vous passez au deuxième motif de la séquence B qui est ATGTG :

Séquence A	A	T	G	T	A	Score global
Séquence B	A	T	G	T	G	
Scores élémentaires	1	1	1	1	0	S = 4

La comparaison de la fenêtre de A avec les cinq fenêtres possibles de B donne les résult suivants :

	ATGTA	ATGTA	ATGTA	ATGTA	ATGTA	ATGTA
	TATGT	ATGTG	TGTGA	GTGAA	TGAAT	GAATG
Nucléotides identiques (score)	0	4	1	3	0	1

Ce qui donnera sur la matrice globale :

	A	**T**	**G**	**T**	**A**	**A**	T	G	C	A	T	G
T												
A												
T			**0**									
G			**4**									
T			**1**									
G			**3**									
A			**0**									
A			**1**									
T												
G												

Une fois la comparaison effectuée avec toutes les fenêtres de B, nous incrémentons de un la séquence de A pour avoir la nouvelle fenêtre de cinq autres nucléotides qui sont : **TGTAA.**
C'est cette nouvelle fenêtre de A que nous allons devoir comparer avec les fenêtres de B que nous connaissons toutes maintenant.
Le résultat final est :

Il y a cinq segments formés de cinq nucléotides chacun entre les séquences A et B. Ces segments contiennent tous quatre nucléotides identiques:

1. ATGTA de la séqA et ATGGA de la séqB
2. TAATG de la séqA et GAATG de la séqB
3. TGTAA de la séqA et TGGAA de la séqB
4. TGCAT de la séqA et GGAAT de la séqB
5. GCATG de la séqA et GAATG de la séqB

- **Une série de TD**