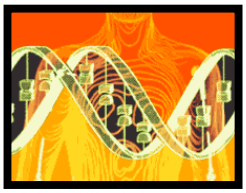
MASTER IBM 2010

INTRODUCTION À LA BIOINFORMATIQUE

COURS MASTER IBM 2010



JEAN-DANIEL ZUCKER

DR À L'IRD UR UMMISCO
(MODÈLISATION MATHÉMATIQUES ET INFORMATIQUES DES SYSTÈMES COMPLEXES)
UMRS 872 EQUIPE NUTRIOMIQUE INSERM/PARIS 6 ET LIM&BIO

UPMC PARISUNIVERSITAS — IRD Institut de recherche pour le développement — Inserm Institut national de la santé et de la recherche médicale — Lim&Bio

MODULE BIOINFO M2 IBM : DIDACTIQUE

- **Objectifs:**
  - COMPRENDRE LE DOMAINE DE LA BIO-INFORMATIQUE
  - LE RÔLE DE L'INFORMATIQUE ET L'IA DANS CES PROBLÉMATIQUES
  - COMPRENDRE LES ALGORITHMES DE BASE DE LA FOUILLE DE DONNÉES POUR LA BIOINFORMATIQUE
  - UTILISER R ET DES APPLETS

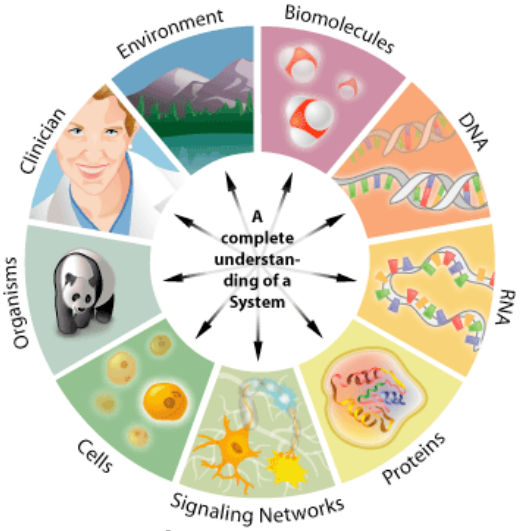
Administratif: Module NT-BIO Master IBM

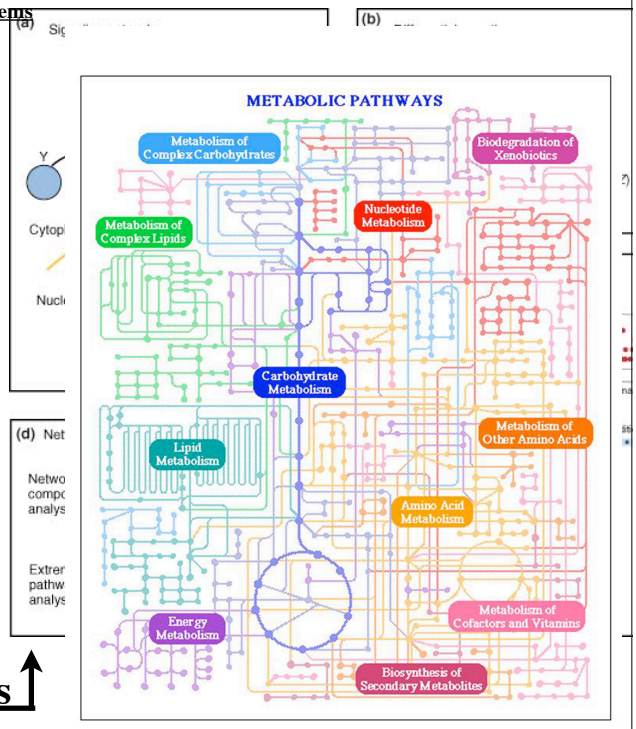
- **Lundi 18 Oct. 2010 –** Intro Générale
  - **La bioinformatique : les 'omics'**
  - **La génomique : BD et algorithmes**
- **Mardi 19 Oct. 2010 –** Algorithme pour la Bioinfo et Omique
  - **Suite et fin sur l'appariement.**
  - **Les réseaux, les BD, les ontologies.**

Aujourd'hui : Biologie des systèmes

Systems Biology

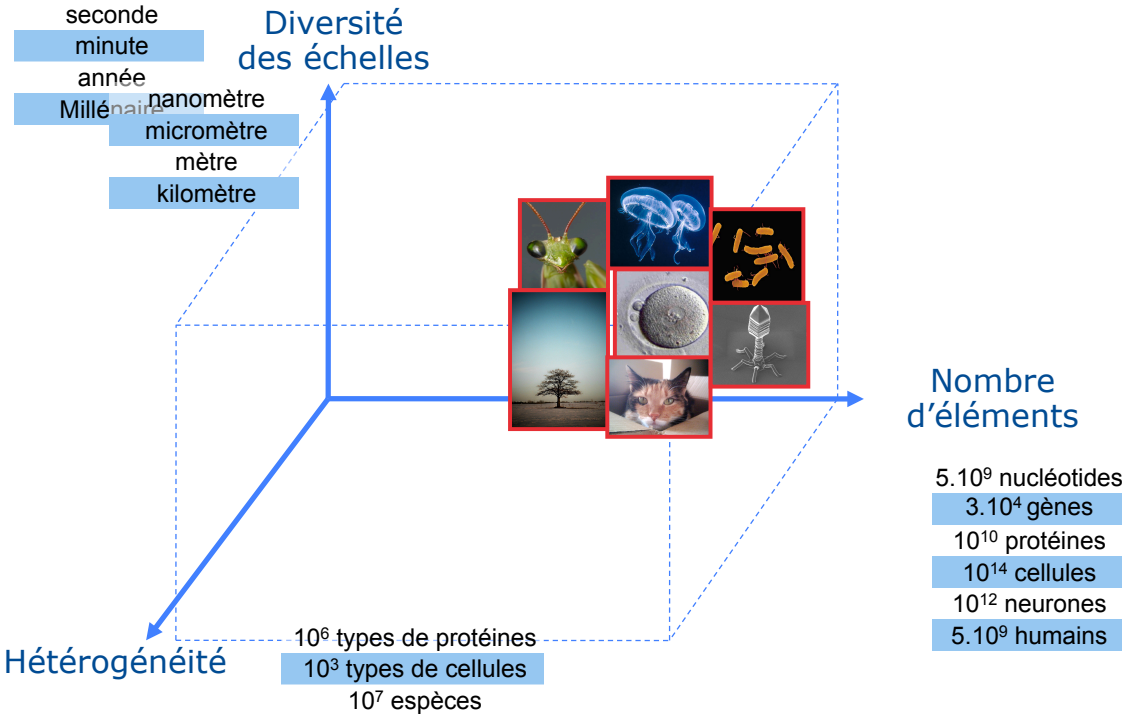
focuses on the systematic study of complex interactions in biological systems





Complex interactions

Systèmes biologiques complexes …



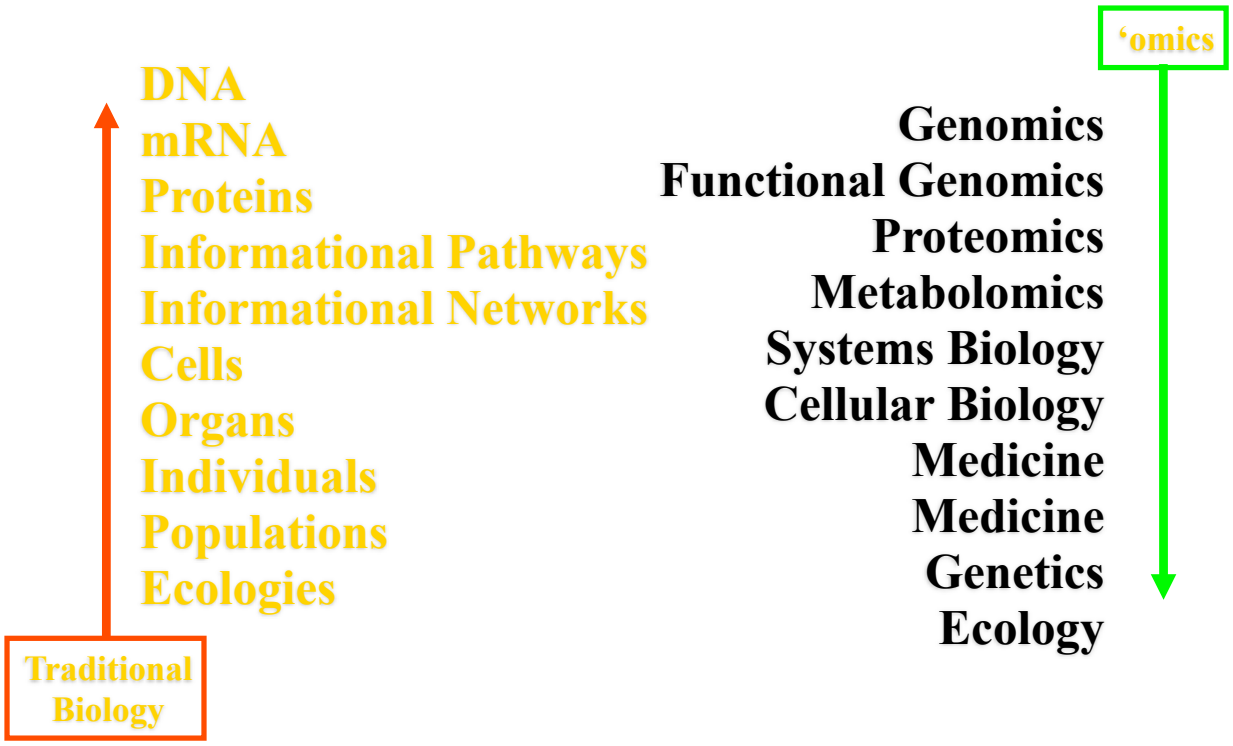
Systèmes biologiques complexes …

- **Un point de départ : la complexité**

« Un système complexe est un système composé d'un grand nombre d'éléments en interaction, la dynamique de ces interactions dirigeant le comportement du système en lui donnant une apparence d'unité aux yeux d'un observateur extérieur. »

- **Plus que les systèmes complexes, il importe de définir la « science des systèmes complexes »**
  - **Toute science est déterminée par le regard qu'elle porte sur son objet**
  - **Caractériser/Comprendre/Reproduire le caractère d'unité d'un système**
  - **Questionner les frontières : une science intrinsèquement pluridisciplinair**

Niveau de l'information Biologique



DEFINITIONS (IN ENGLISH)

- 'Bioinformatics is an integration of mathematical, statistical, and computer methods to analyze biological, biochemical and biophysical data' (Georgia Tech: www.biology.gatech.edu/bioinformatics/whatis.html)
- The Science of developing computer databases and algorithms for the purpose of speeding up and enhancing biological research..' (Whatis.com)
- A combination of Computer Science, Information Technology and Genetics to determine and analyse genetic information (BitsJournal.com)

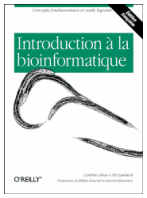
UNE DÉFINITION À RETENIR

La bioinformatique est une science interdisciplinaire
- Analyser l'info. biologique disponible et produire de nouvelles K.
- Propose et développe des modèles, des méthodes et des outils
- Requiert des K. en mathématique, en informatique et en biologie.

(Introduction à la bioinformatique Cynthia Gibas & Per Jambeck , O'Reilly.

↓

Langage: BIOPERL
Algorithmique : ex. Algo Appariement BLAST
Réseaux : ex. Grille de calcul GBIO
BD: ex: NCBI (1988 Nat. Res. Mol. Biol)
Modèlisation: Réseaux de régulations Equa Diff.



OBJECTIFS DE LA BIO-INFORMATIQUE?

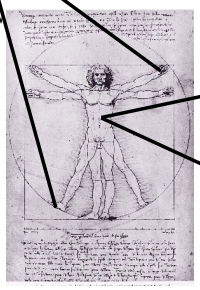
Champs multi-disciplinaire qui utilise des méthodes informatiques (mathématiques, statistiques, combinatoires…) pour:

- Formaliser des problèmes de biologie moléculaire;
- Concevoir des solutions computationelles à la portée des machines,
- Développer et valider des outils;

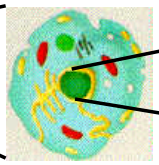
- Analyser, structurer, comparer, traiter les information biologiques
- Stocker, accéder, filtrer ces informations
- Modéliser des processus biologiques
- Prédire des résultats biologiques; etc.

Discipline relativement nouvelle, qui évolue en fonction des nouveaux problèmes posés par la biologie moléculaire.

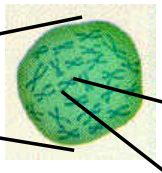
Niveaux DE l'information biologique



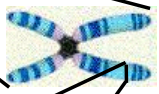
Cellule

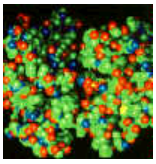


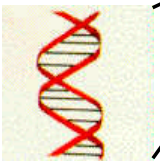
Noyau



Chromosome







Protéine

Gene (ARNm), Simple brin

Gene (ADN)

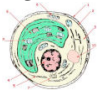
National Human Genome Research Institute

l'étymologie de **chromosome**: *chromo* = couleur et *soma* = corps: des corps colorables

MASTER IBM 2008UCKER©

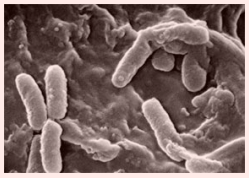
Quelles catégories d'organism

Eucaryotes. Ce groupe d'êtres vivants - dont font partie les hommes, les animaux, les plantes - est caractérisé par une organisation cellulaire complexe. Chaque cellule est composée d'un noyau et de composants membranaires.

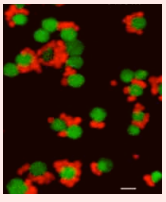


Prokaryote

Bactéries, ou eubactéries. De forme sphérique ou ovoïde, ces organismes, composés le plus souvent **d'une seule cellule**, sont présents dans l'air, l'eau, le sol, etc. Leur taille est généralement de l'ordre du micron.



Archæa, ou archébactéries. (Woese, 77) **micro-organismes unicellulaires** sont aussi éloignés des bactéries que ces dernières le sont des eucaryotes. La plupart sont adaptées à des conditions physico-chimiques extrêmes (température, pH, etc.).



http://tolweb.org/tree/

virus ?)

Quelles Types d'informations: «OMES» ?

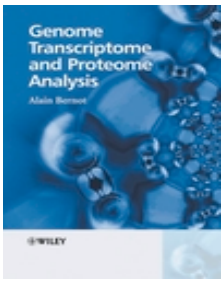
Génome (l'ensemble du matériel génétique d'un individu ou d'une espèce.)

Transcriptome (l'ensemble des ARN messagers transcrits à partir du génome)

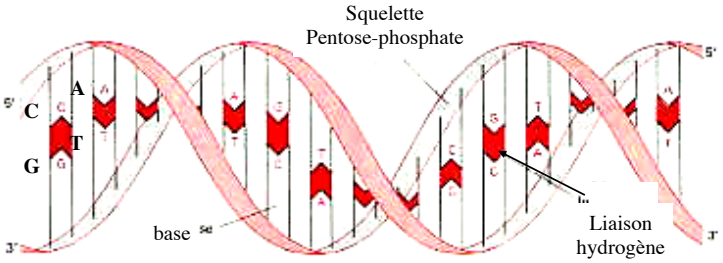
Protéome (l'ensemble des protéines exprimés à partir du génome)

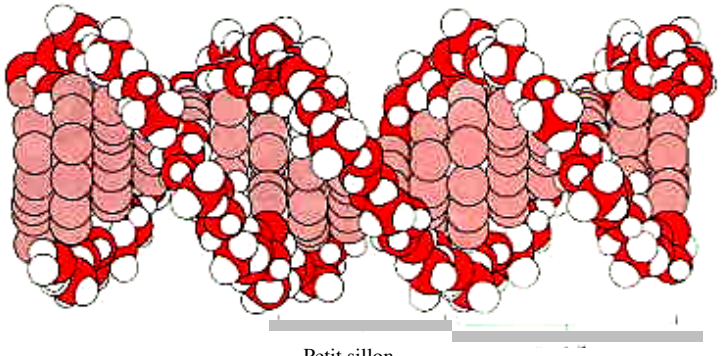
Métabolome (l'ensemble des composés organiques (sucres, lipides, amino-acides, …)

Intéractome (l'ensemble des interactions protéine-protéine)…



Bioinformatics in the Post-Genomic Era: Genome, Transcriptome, Proteome, and Information-Based Medicine





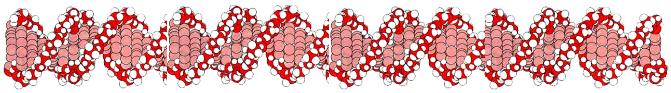
Chaîne polydésoxyribonucléotidique (liaison 5 '=>3 ' phosphodiester)

Double Hélice droite

Bases : Adénine (**A**), Guanine (**G**),
Cytosine (**C**) et Thymine (**T**)

MASTER IBM 2008UCKER©

Du gène à la protéine



TRANSCRIPTION
- Recopie d 'un brin
Avec T==>U
Désoxyribose ==> Ribose

ARN primaire

NOYAU

MATURATION
- épissage
- capping
- polyadénylation

ARN messager **CAP** AAUAAAAAAAAAAAAAAAAAAAAA.....AAAAA

TRADUCTION

CYTOPLASME

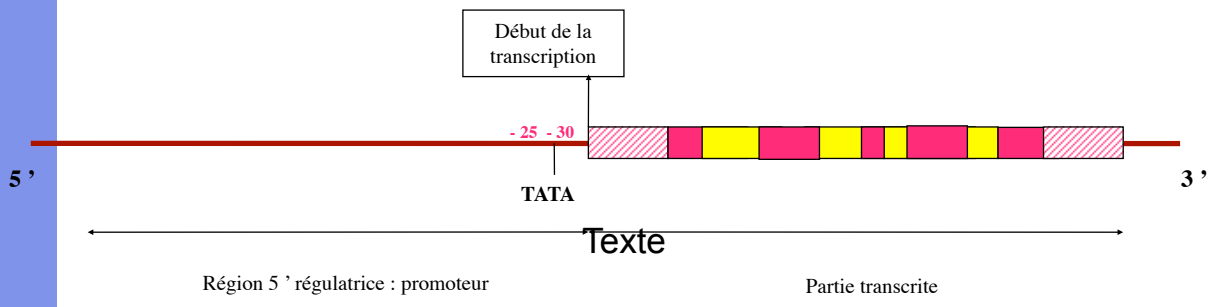
Protéine

CAP AAUAAAAAAAAAAAAAAAAAAAAA.....AAAAA

Ribosomes

MASTER IBM 2008UCKER©

Structure d'un gène codant une protéine



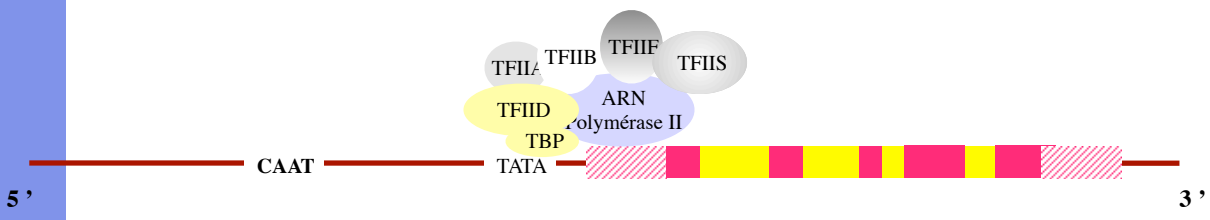
Exons: séquences retouvées dans l'ARNm

Introns: séquences transcrites puis éliminées (épissage)

Séquences exoniques retrouvées dans l'ARNm mais non traduites

MASTER IBM 2008UCKER©

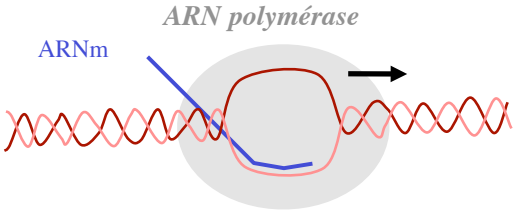
Transcription



- ARN polymérase II : Formée de plusieurs sous-unités (10 à 12). Elle se fixe par l 'intermédiaire d 'un ensemble de co-facteurs.

- Elle possède un double rôle : séparation des brins, copie d'un des brins.

- Elle possède également une activité de « vérification » de la copie.



MASTER IBM 2008UCKER©

TRADUCTION : LE CODE GÉNÉTIQUE

- L'ADN a une structure primaire formée de l 'enchaînement de 4 bases. Pour traduire l 'information ADN ==> acides aminés, il faut donc au minimum un enchaînement de trois bases (codon) pour définir un acide aminé. Il existe donc 64 possibilités.

1ère base	2ème base U	2ème base C	2ème base A	2ème base G	3ème base
U	UUU *Phe* UUC *Phe* UUA *Leu* UUG *Leu*	UCU *Ser* UCC *Ser* UCA *Ser* UCG *Ser*	UAU *Tyr* UAC *Tyr* UAA Stop UAG Stop	UGU *Cys* UGC *Cys* UGA Stop UGG *Trp*	U C A G
C	CUU *Leu* CUC *Leu* CUA *Leu* CUG *Leu*	CCU *Pro* CCC *Pro* CCA *Pro* CCG *Pro*	CAU *His* CAC *His* CAA *Gln* CAG *Gln*	CGU *Arg* CGC *Arg* CGA *Arg* CGG *Arg*	U C A G
A	AUU *Ile* AUC *Ile* AUA *Ile* AUG *Met*	ACU *Thr* ACC *Thr* ACA *Thr* ACG *Thr*	AAU *Asn* AAC *Asn* AAA *Lys* AAG *Lys*	AGU *Ser* AGC *Ser* AGA *Arg* AGG *Arg*	U C A G
G	GUU *Val* GUC *Val* GUA *Val* GUG *Val*	GCU *Ala* GCC *Ala* GCA *Ala* GCG *Ala*	GAU *Asp* GAC *Asp* GAA *Glu* GAG *Glu*	GGU *Gly* GGC *Gly* GGA *Gly* GGG *Gly*	U C A G

- Les protéines sont formées de 20 acides aminés différents qui s 'enchaînent de façon bien précise.

Régulation de l'expression des gènes

- Bien que la majorité des cellules de l 'organisme contienne le même ADN nucléaire, elles n 'expriment pas toutes les mêmes protéines. Par exemple, seules les cellules du foie produisent de l 'albumine ou les cellules ß-pancréatiques de l 'insuline.

- D 'autre part, au sein d 'une même cellule, une protéine peut être nécessaire à certains moments mais non à d 'autres.

→ **Nécessité d'une régulation de l 'expression des gènes**

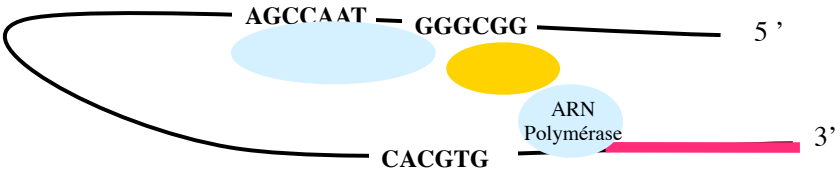
Régulation de l'expression des gènes : Séquences *cis*, facteurs *trans*

- Certaines séquences de l'ADN sont capables de modifier le taux de transcription d 'un gène.

- Situées dans la partie promotrice du gène, ces séquences régulatrices sont appelées séquences *cis*.

- Elles confèrent soit une spécificité d'expression tissulaire, soit une spécificité de réponse à des stimuli extra- ou intracellulaires (hormones, ions...). Dans ce dernier cas on les appelle éléments de réponse.

- Ces séquences cis fixent pour la plupart des protéines. Ces protéines sont appelées facteurs de transcription.



Partie promotrice d 'un gène

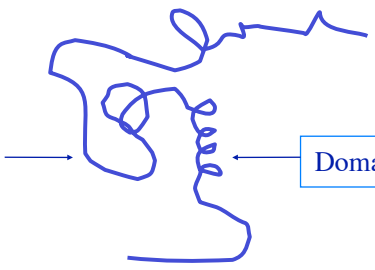
Les facteurs de transcription.

- Plusieurs centaines de facteurs de transcription ont été caractérisés.

- On peut dégager schématiquement un certain nombre de structures communes aux facteurs de transcription :

Domaine d 'interaction avec le complexe basal de transcription

Domaine de liaison d 'un ligand



Domaine de régulation

Domaine de dimérisation

Domaine de liaison à l'ADN

LE TRANSCRIPTOME: ENTRE LE GÈNE ET LA PROTÉINE

ADN

AAG CGC TTG CGC TGT ATG CGT CCG

TRANSCRIPTION

Recopie d'un brin où U à la place de T

ARN primaire

AAG CGC UUG CGC UGU AUG CGU CCG

MATURATION

Epissage alternatif, …

ARN messager

AAAAAAAAAAAAAAAAAA.....AAAAA

TRADUCTION

Protéine

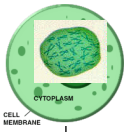
AAUAAAAAAAAAAAAAAAAAAAAAA.....AAAAA

Ribosomes

Expression d'un gene

NOYAU

CYTOPLASME



MASTER IBM 2008UCKER©

ADN

- Localisation: noyau de chaque cellule
- Composition: Nucleotides
- 4 Types de Nucleotide Bases: A, G, C, T
- Humain: 3 millions de kilobases ADN, 1 mètre
- 90% "Junk DNA"; 10% Genes

ARN

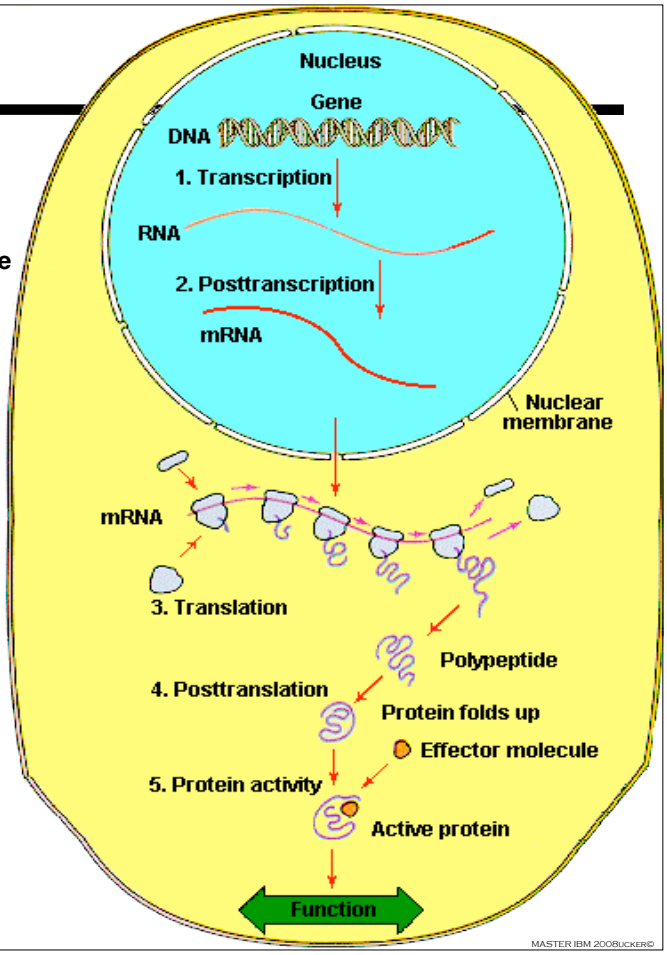
.....ACGTGGTTAAATCGTA.....

....UGCACCAAUUUAGCAU.....

- T->U
- Modifications

PROTÉINES

- 3 Nucleotides →1 acide aminé
- 20 acides aminé
- Modifications Post-traduction



MASTER IBM 2008UCKER©

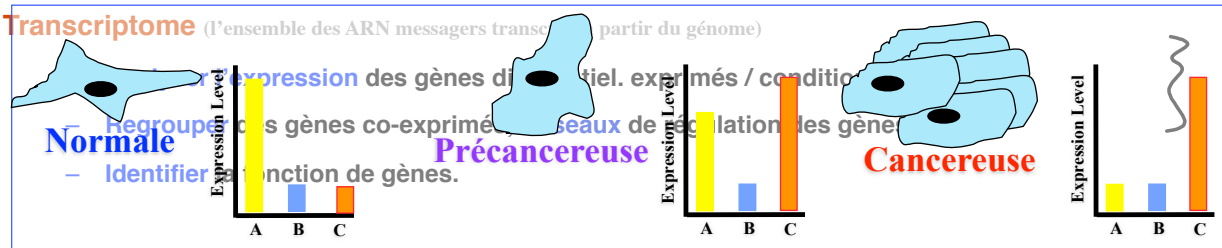
Les "omes" et les problematiques associées

Génome (l'ensemble du matériel génétique d'un individu ou d'une espèce.)

- **Identifier, localiser** les gènes dans une séquence
- **Analyser les** modifications epigénétiques : modifications de la chromatine (méthyla
- **Aligner** et comparer de **séquences** ex: BLAST (Basic Local Alignment Search Tool)

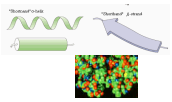
Transcriptome (l'ensemble des ARN messagers transc partir du génome)

- expression des gènes di iel. exprimés / conditio
- Regroupe s gènes co-exprimé eaux de régulation les gène
- Identifier e fonction de gènes.



Protéome (l'ensemble des protéines traduites des ARN messager)

- **Prédire** de la **structure** secondaire, la **fonction** des protéines, …
- **Analyser, mesurer l'expression** en fonction des organes

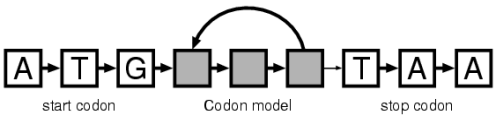


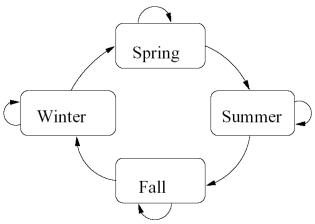
Métabolome (l'ensemble des composés organiques (sucres, lipides, amino-acides, …)

- **Analyser**, **visualiser**, **prédire, mesure** des concentrations

APPRENTISSAGE ET GENOME

- **Les observations sont ordonnées**
- **Un processus aléatoire peut être représenté par une machine à état stochastique.**

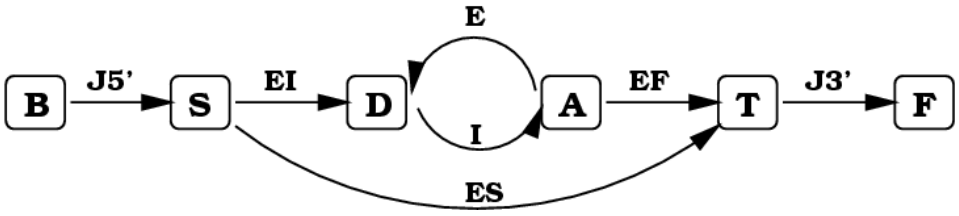




- **Evaluer la probabilité d'une séquence donnée étant donné le model (Forward)**
- **Trouver le chemin le plus probable dans un modèle étant donnée une observation (Viterbi)**

APPRENTISSAGE ET GENOME

- **Problème:**
  - Trouver les régions codantes (exons) et non-codantes (introns) d'un brin d'AND. Les sites des "splice": accepteur et donneurs
- **Application:**
  - Annotation de genes séquencés
  - Compréhension des mécanismes de transcriptions, etc.



==> Apprentissage des parametres à partir d'exemples

ANALYSE DE L'ADN

- **Séquenceur : séquençage du génome entier (cartographie physique), identification de mutations (investissements lourds genome center)**
- **CGH (comparative genome hybridation)**
  - **Mise en évidence de remaniements chromosomiques par comparaison à un génome de référence : délétions, amplifications, mais résolution limitée (microdélétion), certains types de remaniement ne peuvent pas être détectés (translocation équilibrée)**
- **SNP array (single nucleotide polymorphism)**

LE FUTUR: À CHACUN SON GÉNOME POUR 1000$

nature Vol 452|17 April 2008|doi:10.1038/nature06884

LETTERS

The complete genome of an individual by massively parallel DNA sequencing

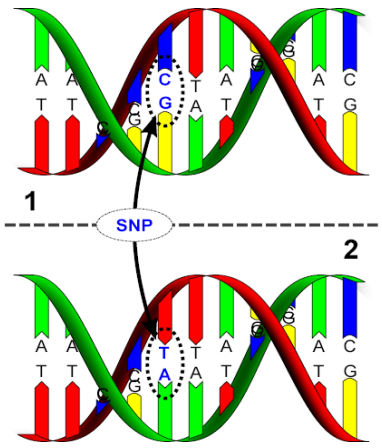
Next-generation technology brings sequencing closer to routine

QUICKER, SMALLER, CHEAPER

Genome sequenced (publication year)	HGP (2003)	Venter (2007)	Watson (2008)
Time taken (start to finish)	13 years	4 years	4.5 months
Number of scientists listed as authors	> 2,800	31	27
Cost of sequencing (start to finish)	$2.7 billion	$100 million	< $1.5 million
Coverage	8-10 ×	7.5 ×	7.4 ×
Number of institutes involved	16	5	2
Number of countries involved	6	3	1

Today	Goal
$ 100 000	$ 1000

SNP ARRAY (SINGLE NUCLEOTIDE POLYMORPHISM)

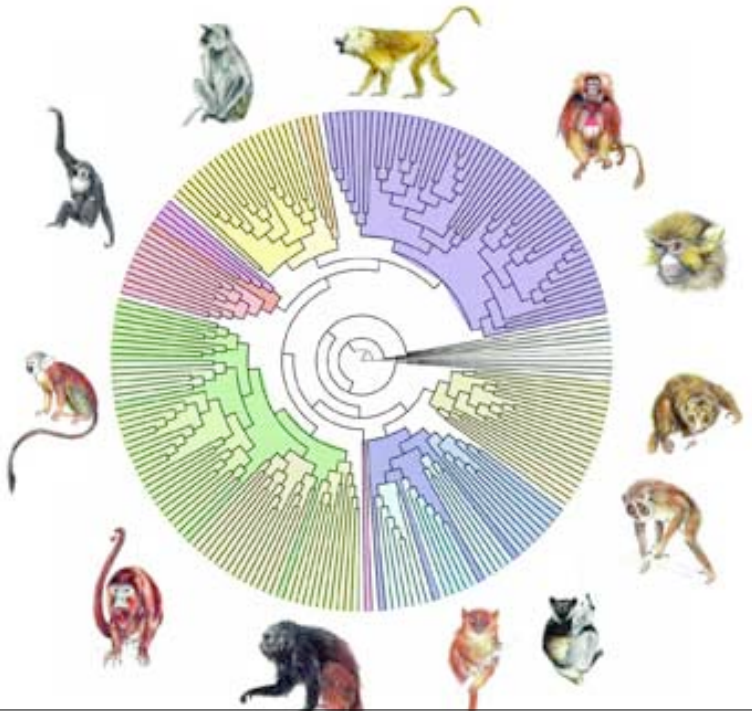


- Variations non pathologiques de la séquence d'ADN au niveau d'une seule base
- Plusieurs millions sont connues pour le génome humain
- Projet HAPMAP (catalogue des variations génétiques)
- Analyse possible par la technique des micro-arrays à ADN
  - ~10000 à 1 million de SNPS

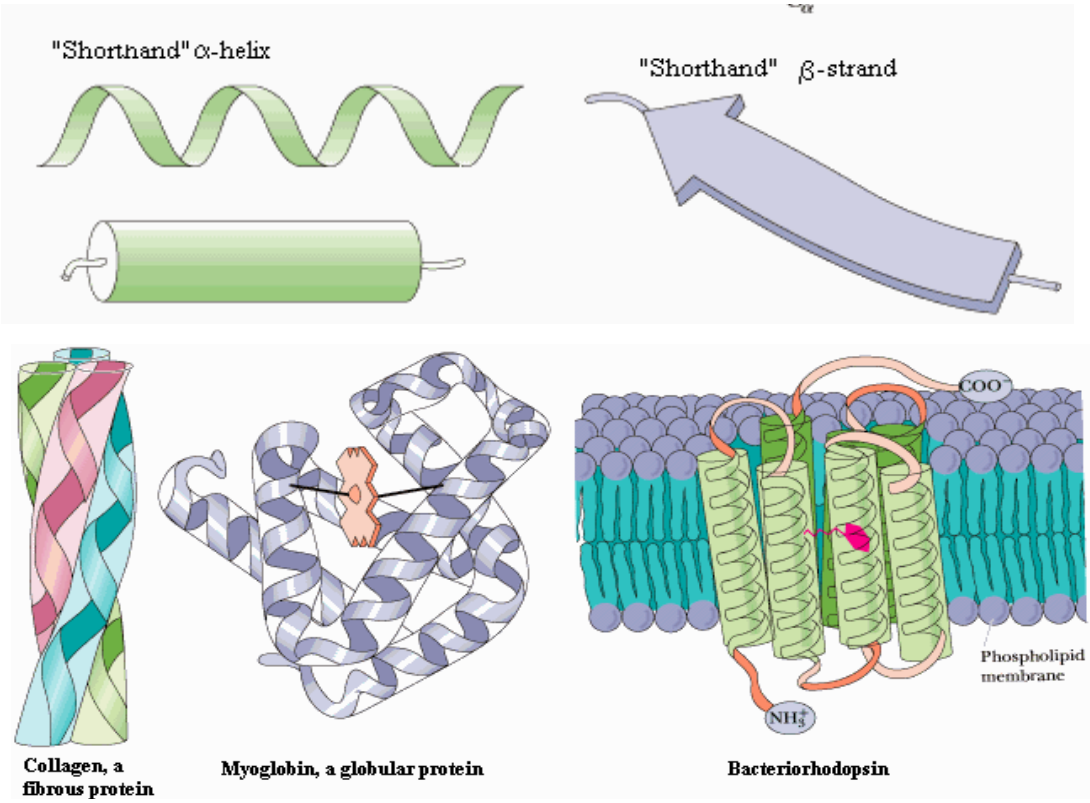
Application des SNP arrays

- **Etudes génétiques**
  - **Analyses de liaison ou d'association génétiques**
    - **»recherche de régions et de gènes candidats**
- **Analyse du nombre de copies (copy number) et de la perte d'hétérozygotie**
  - **Remaniements fréquents dans les cancers ++gènes clés dans la tumorogénèse et la prolifération**

Etude Phylogénétique



Prédire le repliement des protéines



… UN VASTE CHAMPS D'ÉTUDE

La Bio-Informatique s'applique à tous les types de donnée biologique :

- **Les séquences d'ADN et de protéines**
- **Les structures d'ARN et de protéines**
- **Les contenus en gènes des génomes**
- **Les transcrits (ARNm)**
- **Les réseaux d'interactions entre protéines**
- **Les réseaux métaboliques**
- **Les arbres de phylogénie**
- **…**

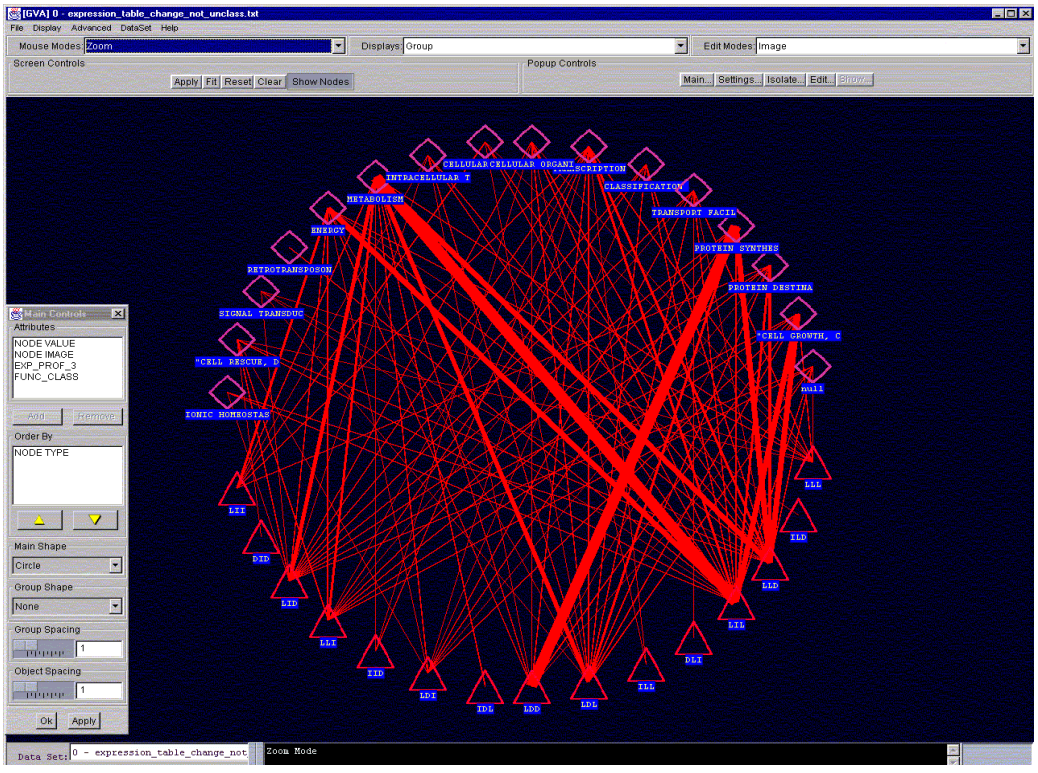
Utilités:

- **Faire avancer les connaissances en biologie, en génétique humaine, en théorie de l'évolution…**
- **Aider à la conception de médicaments**
- **Comprendre les maladies complexes, ...**

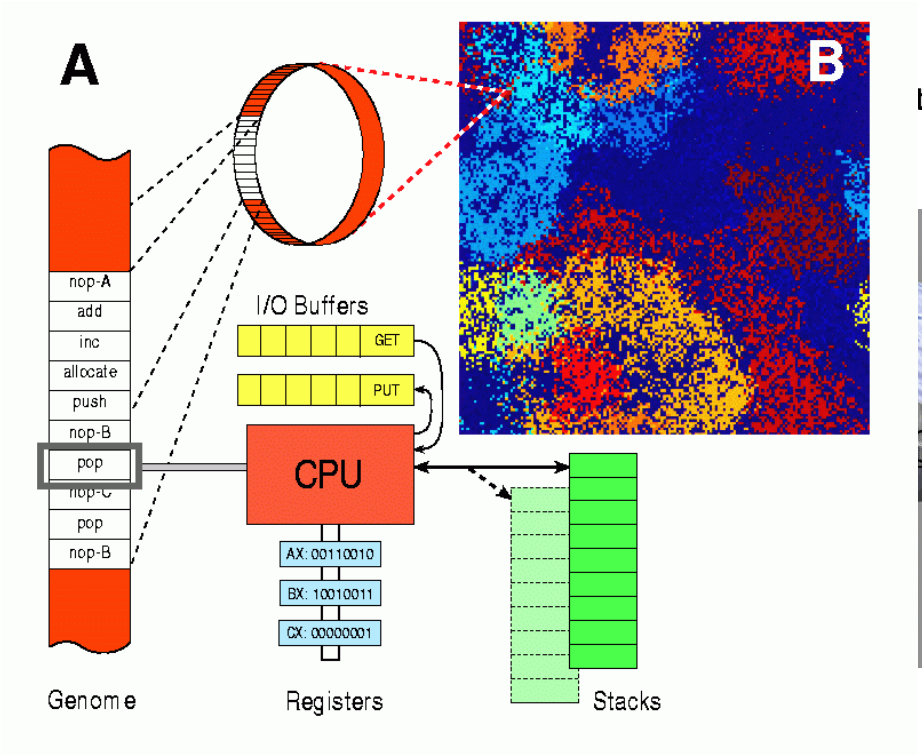
De très nombreux domaines de recherche en informatique, automatique et mathématiques appliquées sont concernés

- **Algorithmique sur les séquences, sur les graphes…**
- **Statistique, analyse de données**
- **Apprentissage symbolique et numérique**
- **Visualisation de données**
- **Modélisation et simulation dynamiques**
- **Calcul parallèle**
- **Bases de données et de connaissances**
- **Représentation des connaissances**
- **Algorithme Evolutionniste**
- **Algorithme Génétique**

VISUALISATION



Biologie Computationelle



Quelles données de l'analyse des «omes» ?

GATCACC
TCACTAC
GGGTCAG
GGGAAGG
AAAGGGG
AACTGAG
AGATTT..

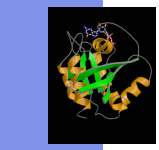
GAUCACC
UCACUAC
GGGUCAG
GGGAAGG
AAAGGGG
AACUGAG
AGAUUU..

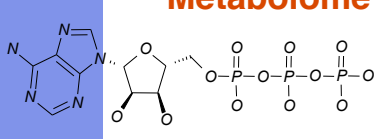
Génome

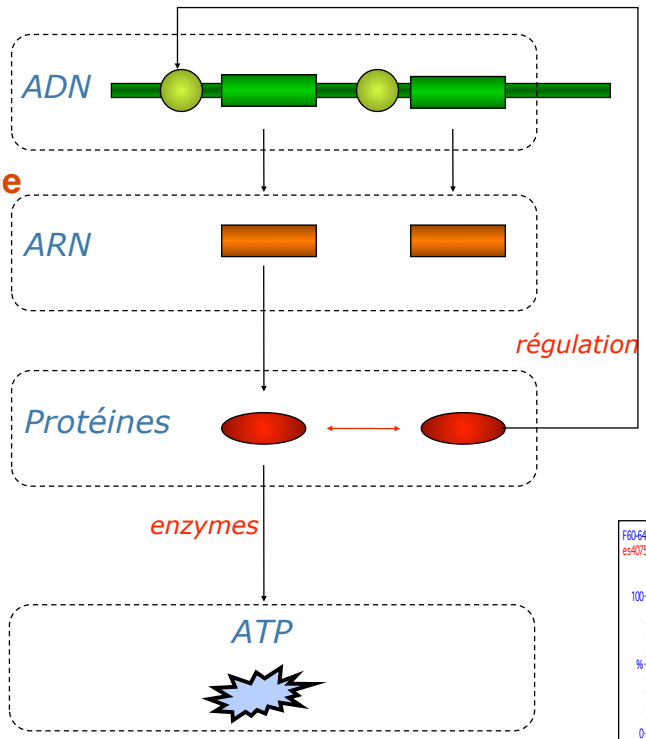
Transcriptome

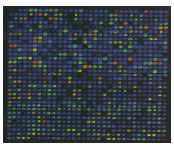
Protéome

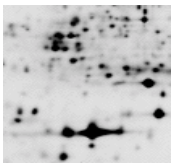
Métabolome

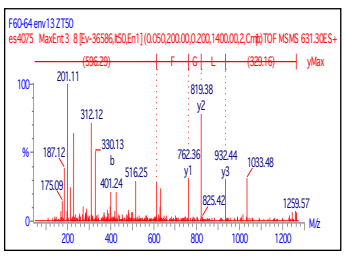








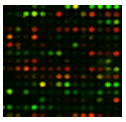




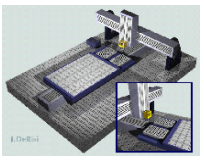
Acquisition de données Biologique

Manuel ⇒ Automatique

- Sequences ADN (génome)
  - High Throughput DNA Sequencers
- Données ARN (transcriptome)
  - Biopuces
- Données sur les protéines (protéome)
  - Crystallographie au rayon X
  - Puces à Protéines

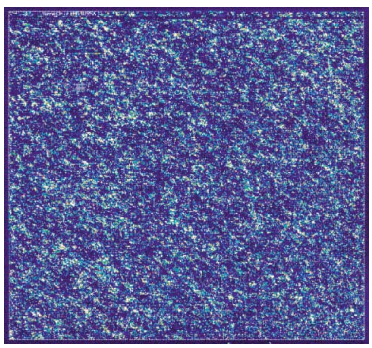


⇒ EXPLOSION de données



Analyse de l'expression des gènes : expression arrays

- Mesure simultanée de l'abondance de quelques centaines (puce ciblée) à plusieurs dizaines milliers de transcripts (puce pangénomique) au niveau d'une ou plusieurs cellules (lignées) ou d'un tissu
- => Niveau d'activité du gène
- Remarque pas de corrélation stricte avec l'expression protéique



Données en bioinformatique

- Explosion de la quantité de données (ADN 73 Gb, arrivée des données biopuces, metabolic pathways, …)
- Croissance exponentielle des données (11-15% tous les 3 mois), plus traitable localement
- Données hétérogènes dans leur structure et leur sémantique
- Systèmes d'information hétérogènes
- Beucoup de connaissances cachées, privées ou inconnues.
- …

Des BD et encore des BD…

- Base de Données ADN
  - GenBank, DDBJ, EMBL,…
- Base de Données Protéines
  - PIR, Swiss-Prot, PRF, GenPept, TrEMBL, PDB,…
- Base de Données EST
  - dbEST, DOTS, UniGene, GIs, STACK,…
- Base de Données Structure
  - MMDB, PDB, Swiss-3DIMAGE,…
- Base de Données voies métabol.
  - KEGG, BRITE, TRANSPATH,…
- Base de Données intégrées
  - SRS

Base de Données de Motifs
- Prosite, Pfam, BLOCKS, TransFac, PRINTS, URLs,…

Base de Données sur les maladies
- GeneCards, OMIM, OMIA,…

Base de Données taxonomique

Base de données littérature scient.
- PubMed, Medline,…

Base de données de brevets
- Apipa, CA-STN, IPN, USPTO, EPO, Beilstein,…

Autres…
- RNA databases, QTL…

ID

AmiGO Get this GO term as RDF XML.

chaperonin ATPase

Accession: GO:0003763
Synonyms:
chaperonin
GO:0003764
GO:0003765
GO:0003766

synonymes

Definition: The hydrolysis of ATP involved in maintaining an unfolded polypeptide structure before folding or to entry into mitochondria and chloroplasts.

définition

Term Lineage Graph view.

GO:0003673 : Gene_Ontology (31411)
P GO:0003674 : molecular_function (21780)
I GO:0003754 : chaperone (229)
I **GO:0003763 : chaperonin ATPase (15)**
I GO:0003824 : enzyme (7782)
I GO:0016787 : hydrolase (3280)
I GO:0016817 : hydrolase, acting on acid anhydrides (1090)
I GO:0016818 : hydrolase, acting on acid anhydrides, in phosphorus-containing anhydrides (1090)
I GO:0016887 : ATPase (783)
I GO:0016821 : hydrolase, acting on acid anhydrides, involved in cellular and subcellular movement (120)
I **GO:0003763 : chaperonin ATPase (15)**
I GO:0005488 : ligand binding or carrier (6876)
I GO:0000166 : nucleotide binding (1357)
I GO:0017076 : purine nucleotide binding (1336)
I GO:0030554 : adenyl nucleotide binding (891)
I GO:0005524 : ATP binding (891)
I GO:0016887 : ATPase (783)

cross-reference

External References
EC
3.6.4.9 EC AmiGO

UNE RÉVOLUTION POUR LA BIOLOGIE

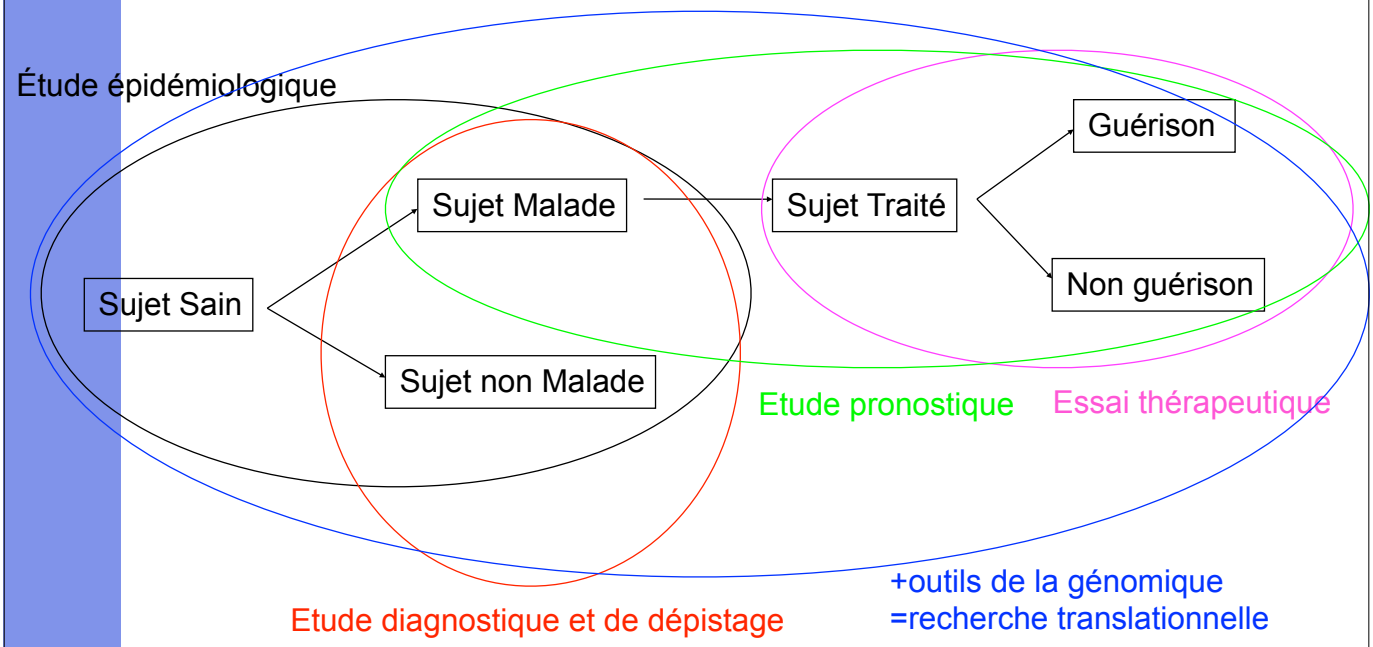
Biologie ***in vivo, in vitro, …….. in silico***

La bioinfo ce n'est pas simplement stocker des données

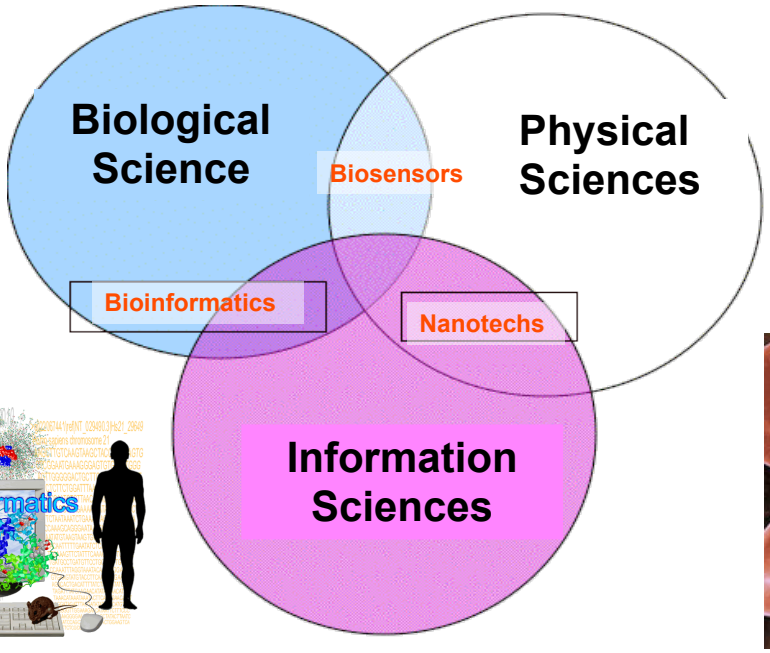
Ce que la bioinformatique a permis de faire:

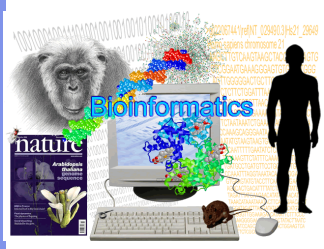
- Les organisations simples sont des modèles pour des systèmes plus complexes
- Les similitudes entre ADN ou protéine illustrent la variation évolutionnaire de gène d'histoire des populations sont importantes pour la découverte de drogue
- La variation de gène des populations sont importante pour la découverte de médicaments
- Les données génétiques ont des implications personnelles et sociales importantes
- L'ADN non codante est importante pour la commande de gène
- Que le nombre de gènes n'est pas corrélé à la complexité des organismes

UNE RÉVOLUTION POUR LA RECHERCHE CLINIQUE

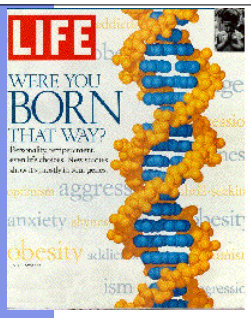


Pluridisciplinarité





LES DÉFIS



End of the beginning

Lincoln D. Stein

Just over three years ago, it was announced that a first draft of the human genome sequence had been completed. Gaps and errors remained, but the job of fixing those problems is now largely done.

- **Décoder** l'information contenue dans les séquences d'ADN et de protéines
  - Trouver les gènes
  - Différencier entre introns et exons
  - Analyser les répétitions dans l'ADN
  - Identifier les sites des facteurs de transcription
  - Étudier l'évolution des génomes
- **Génomique Comparative**
  - Construire les relations de parenté entre organismes
- **Génomique fonctionnelle**
  - Étudier l'expression des gènes
  - Étudier la régulation des gènes
  - Déterminer les réseaux d'interaction entre les protéines
- **Génomique structurale**:
  - Modéliser les structures 3D des protéines et des ARN structurels
  - Déterminer la relation entre structure et fonction
- **Pharmacogénomique**

INTÉRÊT DES SÉQUENCES

- La séquence nucléotidique d'un gène détermine la séquence d'AA de la protéine
- La séquence d'une protéine détermine sa structure et sa fonction
- Généralement, une similarité de séquence implique une similarité de structure et de fonction (l'inverse n'est pas toujours vrai)

Évolution basée, en grande partie, sur la duplication suivie de modification. D'où, beaucoup de redondance dans les bases de données

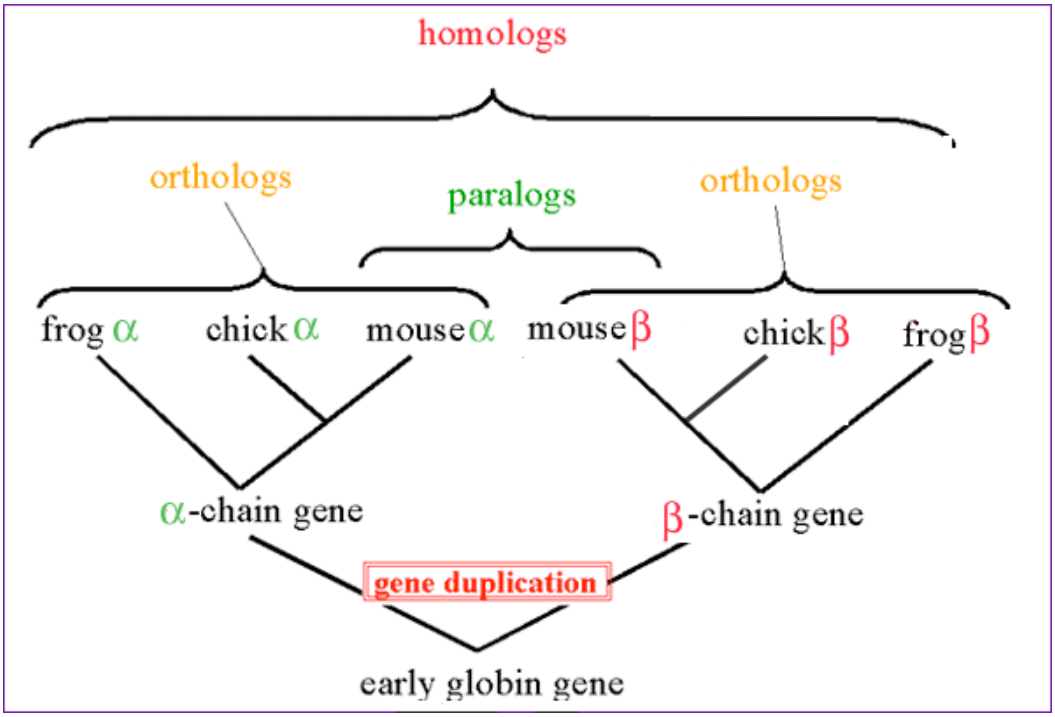
Taches courantes d'un biologiste

Est-ce qu'une nouvelle séquence a déjà été complètement ou partiellement déposée dans les bases de données?

- Est-ce que cette séquence contient un gène?
- Est-ce que ce gène appartient à une famille connue? Quelle est la protéine encodée?
- Existe-t-il d'autres gènes homologues?
- Existe-t-il des séquences non-codantes similaires. Répétitions ou séquences régulatrices

Logiciels les plus connus:
Smith-Waterman, FASTA et BLAST

Biological Definitions for Related Sequences



Biological Definitions for Related Sequences

- ***Homologs*** are similar sequences in two different organisms that have been derived from a common ancestor sequence. Homologs can be described as either orthologous or paralogous.
- ***Orthologs*** are similar sequences in two different organisms that have arisen due to a speciation event. Orthologs typically retain their functionality throughout evolution.
- ***Paralogs*** are similar sequences within a single organism that have arisen due to a gene duplication event.
- ***Xenologs*** are similar sequences that do not share the same evolutionary origin, but rather have arisen out of horizontal transfer events through symbiosis, viruses, etc.

Les BD de séquences : structure d'une entrée

Identification de la séquence

Numéro unique d'accession (Accession Number - AC)

Données taxinomiques

Références bibliographiques

Annotations

Références croisées avec d'autres banques de données

Mots-clefs

```
ID   IL6_HUMAN      STANDARD;      PRT;   212 AA.
AC   P05231;
DT   13-AUG-1987 (Rel. 05, Created)
DT   13-AUG-1987 (Rel. 05, Last sequence update)
DT   01-MAR-2002 (Rel. 41, Last annotation update)
DE   Interleukin-6 precursor (IL-6) (B-cell stimulatory factor 2) (BSF-2)
DE   (Interferon beta-2) (Hybridoma growth factor).
GN   IL6 OR IFNB2.
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., AND PARTIAL SEQUENCE.
RX   MEDLINE=87065033; PubMed=3491322; [NCBI, ExPASy, EBI, Israel, Japan]
RA   Hirano T., Yasukawa K., Harada H., Taga T., Watanabe Y., Matsuda T.,
RA   Kashiwamura S.-I., Nakajima K., Koyama K., Iwamatsu A., Tsunasawa S.,
RA   Sakiyama F., Matsui H., Takahara Y., Taniguchi T., Kishimoto T.;
RT   "Complementary DNA for a novel human interleukin (BSF-2) that induces
RT   B lymphocytes to produce immunoglobulin.";
RL   Nature 324:73-76(1986).
CC   -!- FUNCTION: IL6 IS A CYTOKINE WITH A WIDE VARIETY OF BIOLOGICAL
CC       FUNCTIONS: IT PLAYS AN ESSENTIAL ROLE IN THE FINAL DIFFERENTIATION
CC       OF B-CELLS INTO IG-SECRETING CELLS, IT INDUCES MYELOMA AND
CC       PLASMACYTOMA GROWTH, IT INDUCES NERVE CELLS DIFFERENTIATION, IN
CC       HEPATOCYTES IT INDUCES ACUTE PHASE REACTANTS.
CC   -!- SUBCELLULAR LOCATION: Secreted.
CC   -!- SIMILARITY: BELONGS TO THE IL-6 SUPERFAMILY.
DR   EMBL; X04430; CAA28026.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence]
DR   EMBL; M14584; AAA52728.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence]
..
DR   PDB; 1IL6; 04-FEB-98. [ExPASy / RCSB]
DR   PDB; 2IL6; 04-FEB-98. [ExPASy / RCSB]
...
DR   Pfam; PF00489; IL6; 1.
KW   Cytokine; Glycoprotein; Growth factor; Signal; Polymorphism;
KW   3D-structure.
```

BD de séquences biologiques : structure d'une entrée de la banque

Annotation de la séquence

```
FT   SIGNAL       1    29
FT   CHAIN       30   212       INTERLEUKIN-6.
FT   DISULFID    72    78
FT   DISULFID   101   111
FT   CARBOHYD    73    73       N-LINKED (GLCNAC...).
FT   VARIANT     32    32       P -> S.
FT                              /FTId=VAR_013075.
FT   VARIANT    162   162       D -> V.
FT                              /FTId=VAR_013076.
FT   MUTAGEN    173   173       A->V: ALMOST NO LOSS OF ACTIVITY.
FT   MUTAGEN    185   185       W->R: NO LOSS OF ACTIVITY.
FT   MUTAGEN    204   204       S->P: 13% ACTIVITY.
FT   MUTAGEN    210   210       R->K,E,Q,T,A,P: LOSS OF ACTIVITY.
FT   MUTAGEN    212   212       M->T,N,S,R: LOSS OF ACTIVITY.
SQ   SEQUENCE   212 AA;  23718 MW;  1F1ED1FE1B734079 CRC64;
     MNSFSTSAFG PVAFSLGLLL VLPAAFPAPV PPGEDSKDVA APHRQPLTSS ERIDKQIRYI
     LDGISALRKE TCNKSNMCES SKEALAENNL NLPKMAEKDG CFQSGFNEET CLVKIITGLL
     EFEVYLEYLQ NRFESSEEQA RAVQMSTKVL IQFLQKKAKN LDAITTPDPT TNASLLTKLQ
     AQNQWLQDMT THLILRSFKE FLQSSLRALR QM
//
```

Séquence

Fin de l'entrée

- La séquence peut être formatée : le format FASTA

Entrée de SwissProt — Numéro unique d'accession — **Informations diverses (nom, espèce, …)**

```
>sp|P05231|IL6_HUMAN Interleukin-6 precursor (IL-6) - Homo sapiens (Human).
MNSFSTSAFGPVAFSLGLLLVLPAAFPAPVPPGEDSKDVAAPHRQPLTSSERIDKQIRYI
LDGISALRKETCNKSNMCESSKEALAENNLNLPKMAEKDGCFQSGFNEETCLVKIITGLL
EFEVYLEYLQNRFESSEEQARAVQMSTKVLIQFLQKKAKNLDAITTPDPTTNASLLTKLQ
AQNQWLQDMTTHLILRSFKEFLQSSLRALRQM
```


ALIGNEMENT LOCAL ET GLOBAL (1/2)

Alignement de deux séquences:

Méthodes naturelle pour comparer deux séquences.
On compte le nombre de ``différences'' (insertion, suppression, substitutions)

Alignement global:

```
C A G C A – C G T G G A T T C T C G G
|   |   |   |   |   |   |       | |
T A T C A G C G T G G – C A C T A G C
```

Alignement local:

```
CAGCAC T T – G G A T TCTCGG
       | |   | |   |
TAGT T T A G G - T GGCAT
```

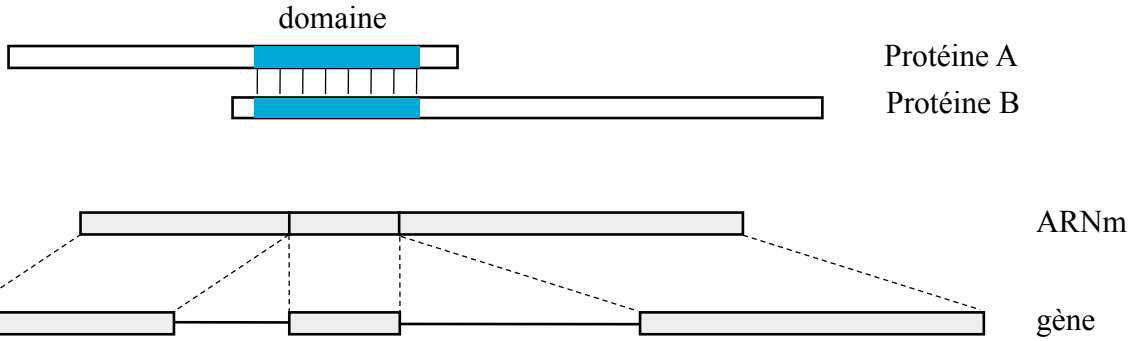
Recherche de motif/gène:

```
C A G C A – C T T G G A T T C T C G G
    | |   |     | | |
    C A G C G T G G
```

Alignement local et global (2/2)

• Alignement global (Needlman & Wunsch, 1970)

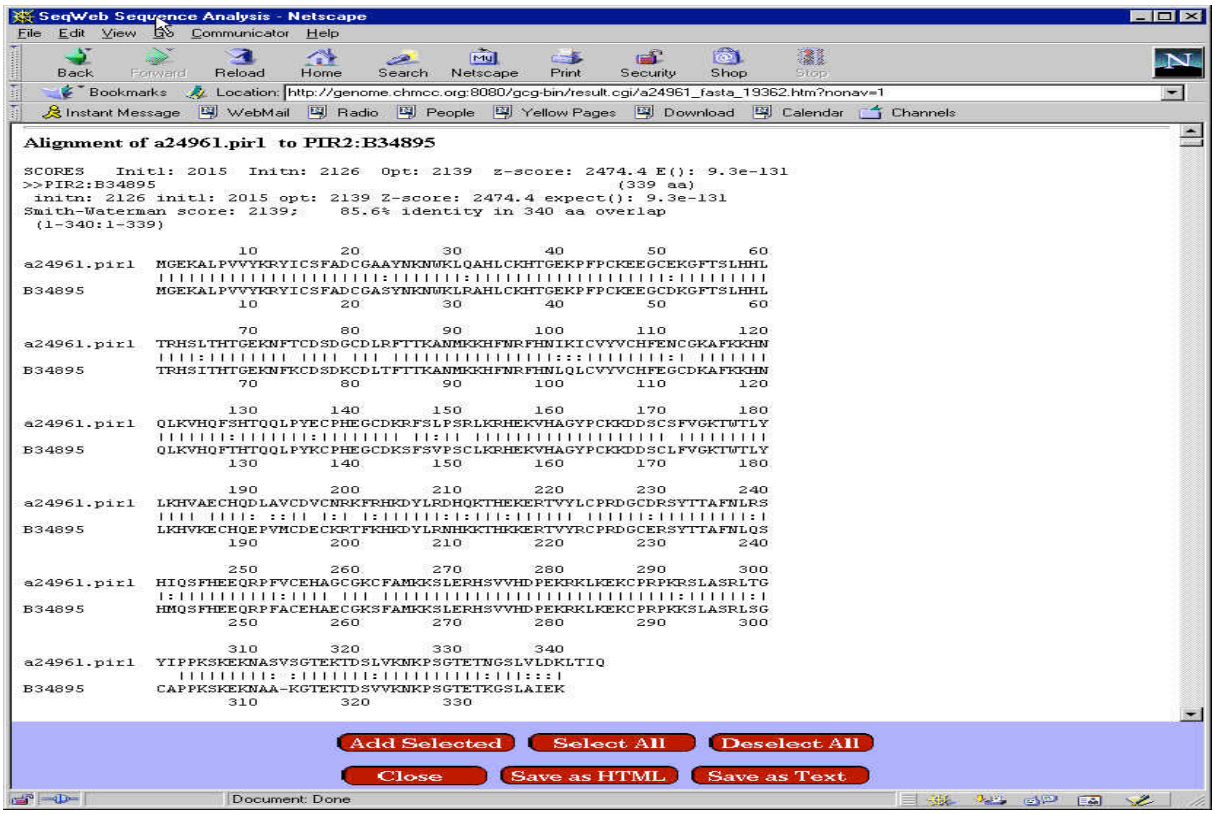
• Alignement local (Smith & Waterman, 1981 ; FASTA, 1988 ; BLAST, 1990)



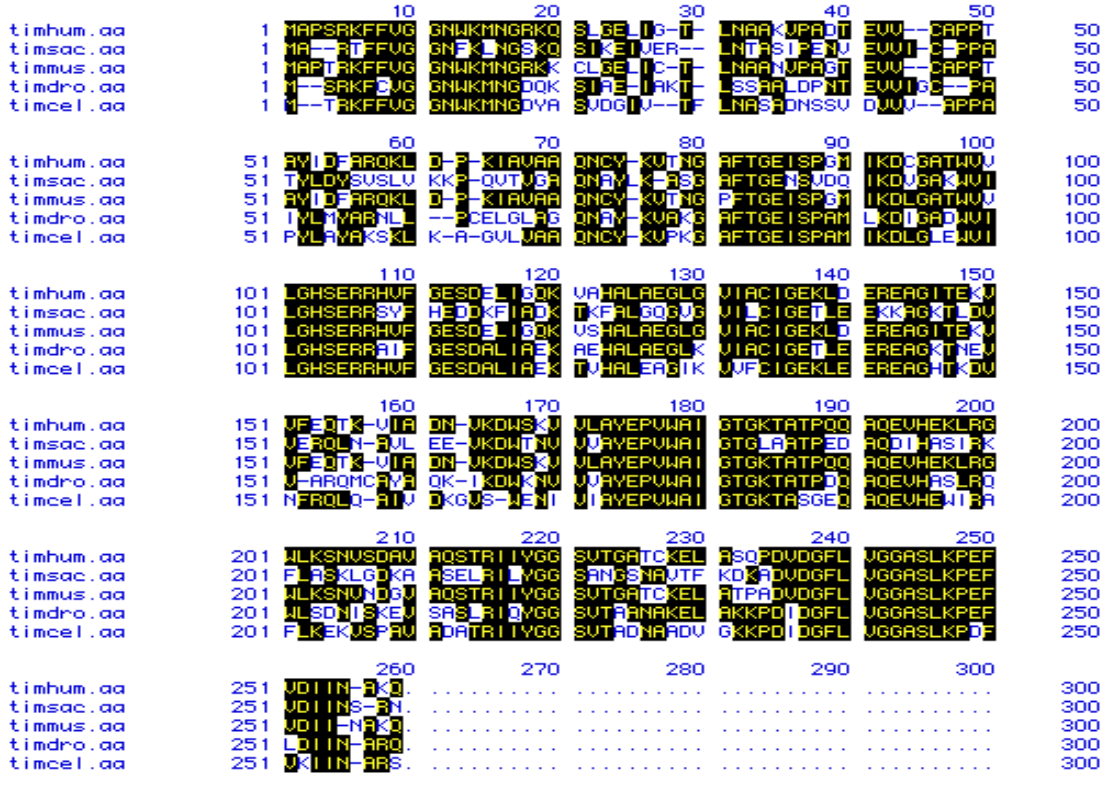
Alignements multiples

- **Trouver des caractéristiques communes à une famille de protéines**
- **Relier la séquence à la structure et à la fonction**
- **Caractériser les gènes homologues**
- **Caractériser les régions conservées et le régions variables**
- **Déduire des contraintes de structures pour les ARN**
- **Construire des arbres de phylogénie**

Exemple de comparaison de 2 séquences

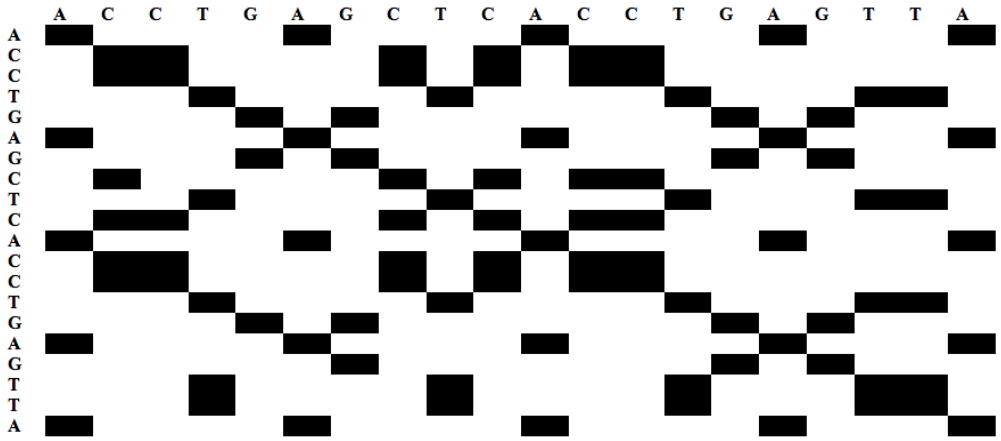


Exemple de comparaison de 5 séquences



Alignements Visuels: Dot Plots

Dot plots of sequence similarity are created using a matrix where the rows in the matrix correspond to the characters in the first sequence and the columns in the matrix correspond to the characters in the second sequence.



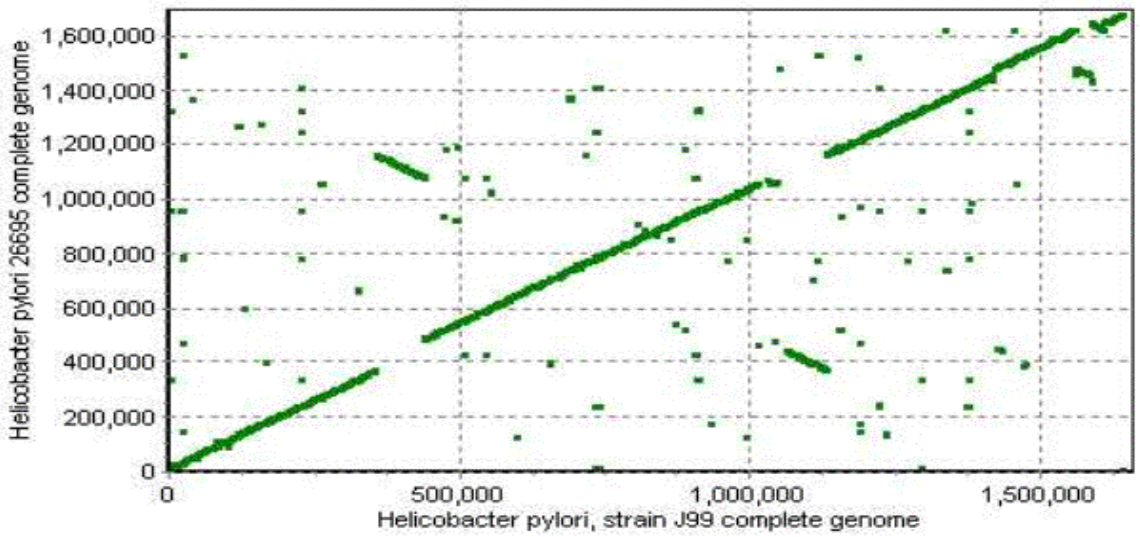
Visual Alignment Algorithme

The dot plot is created as follows:

1. loop through each row.

2. For the current row, take the character in that row and compare it to the character in each column.
   - If they are equal, place a dot in the matrix.
3. Continue until all nodes in the matrix have been considered.

Exemple de comparaison de deux génomes



PLAN

- Algorithme d'alignement

Alignement : représentation

• Opérations élémentaires d'édition :

opérations permettant de « passer » d'une séquence à une autre ;

• **insertions** (i) :

```
A A - B C A A
* *   * * * *
A A C B C A A
```

• **délétions** (d) :

```
A A B C A A
* *   * * *
A A - C A A
```

INsertion / DELétion
INDEL distance

• **substitutions** (s) :

```
A A B C A A
* *   * * *
A A C C A A
```


Quel est le meilleur alignement ?

```
GTTACGA
***  **
GTT-GGA
```

5 matches
1 délétion
1 substitution (C>G)

```
GTTACGA
***  **
GTTG-GA
```

5 matches
1 délétion
1 substitution (A>G)

```
GTTAC-GA
***   **
GTT--GGA
```

5 matches
2 délétions
1 insertion

Méthode d' évaluation :

$Score = \sum se - \sum sp$

se = score élémentaire
sp = pénalité d'*indel*

Le meilleur

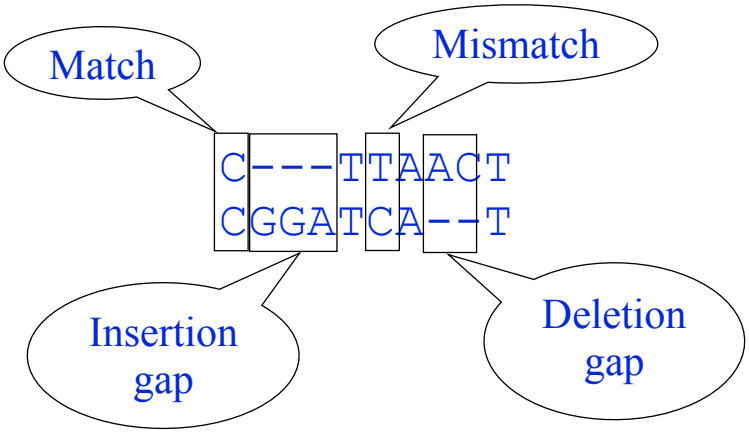
Celui qui optimise le score entre SeqA et SeqB

Coût d'alignement

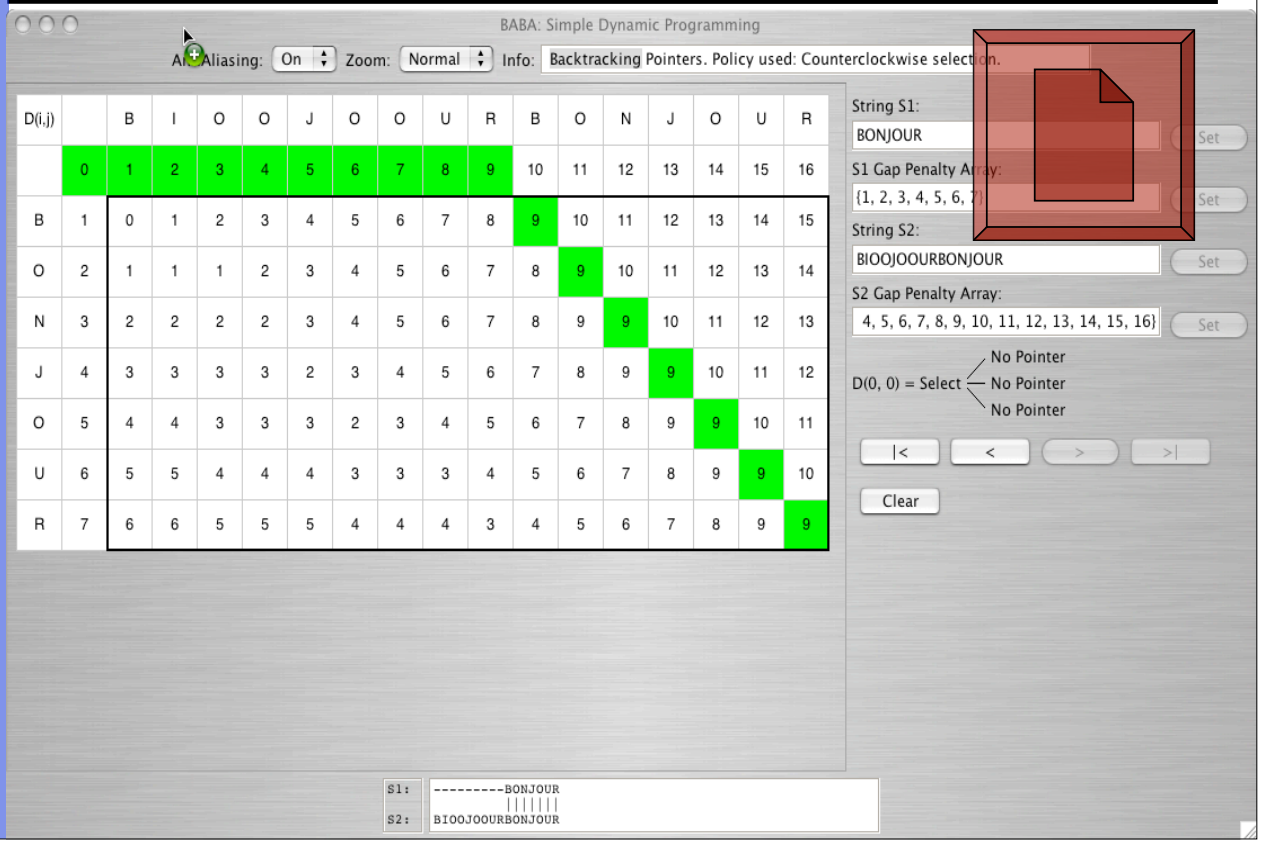
- **Match: +8 ($w(x, y) = 8$, if $x = y$)**
- **Mismatch: -5 ($w(x, y) = -5$, if $x != y$)**
- **Each gap symbol: -3 ($w(-,x)=w(x,-)=-3$)**

Séquence SeqA: CTTAACT
Séquence SeqB: CGGATCAT

Un alignement de SeqA et SeqB:



Distance entre séquences: BABA

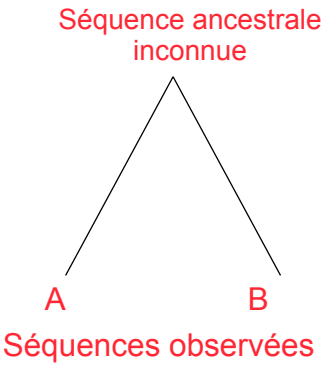


Complexité des algorithmes

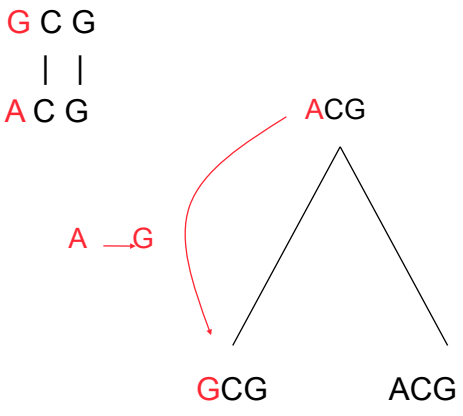
- **Needleman-Wunsch**
  - optimal global alignment
  - dynamic programming, $O(n^3)$
  - Gotoh's improvement, $O(n^2)$, too slow
- **Smith-Waterman**
  - optimal local alignment
  - dynamic programming, $O(n^2)$, too slow
- **FASTA and BLAST**
  - heuristic algorithms
  - BLAST is faster than FASTA
  - FASTA is better than BLAST for nucleotide sequence search

Signification de l'alignement de séquences

Modèle sous-jacent: Mutations ponctuelles



Exemple: Substitution de caractère



Alignements pondérés

Au lieu de compter le nombre de « mutations »
-> matrice de score pour les substitutions.

Matrices empiriques basées sur des alignements de séquences ``homologues''.

Sensées refléter le taux de mutation d'un AA en un autre

PAM250, BLOSUM62 les plus utilisées

Exemple:

Case (i,j) de PAM250: Fréquence avec laquelle A_i remplacée par A_j dans les séquences qui divergent de 250 PAM (taux d'évolution entre les séquences comparées) Point Accepted Mutation=PAM

Case (i,j) de BLOSUM: Fréquence avec laquelle A_i est aligné avec A_j dans des alignements de la base de données BLOCKS

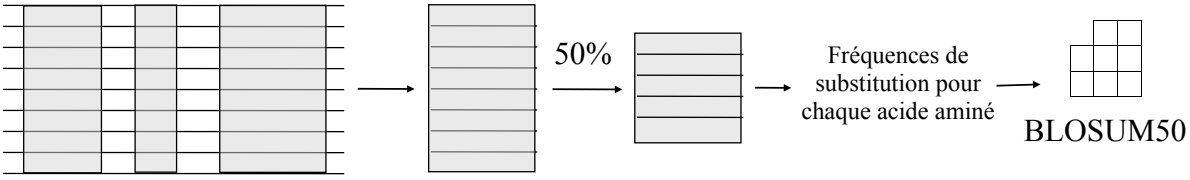
Matrice BLOSUM 62

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A Ala	4																			
R Arg	-1	5																		
N Asn	-2	0	6																	
D Asp	-2	-2	1	6																
C Cys	0	-3	-3	-3	9															
Q Gln	-1	1	0	0	-3	5														
E Glu	-1	0	0	2	-4	2	5													
G Gly	0	-2	0	-1	-3	-2	-2	6												
H His	-2	0	1	-1	-3	0	0	-2	8											
I Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
L Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
K Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
M Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
F Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
P Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
S Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
T Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
W Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Y Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
V Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

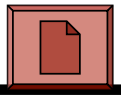
Score positif pour les identités, et négatif pour les mismatchs

Les matrices protéiques de substitution

- **Les matrices de type BLOSUM** (*BLOcks Substitution Matrix*)
  - observation de blocs d'acides aminés issus de protéines relativement éloignées ;
  - chaque bloc provient d'alignements multiples sans insertions / délétions de courtes régions conservées ;
  - les blocs sont utilisés pour regrouper tous les segments de séquences ayant un pourcentage d'identité minimum au sein de leur bloc ;
  - on en déduit des fréquences de substitution pour chaque paire d'acides aminés ;
  - on calcule une matrice logarithmique de probabilité ;
  - à chaque pourcentage d'identité correspond une matrice :
    - BLOSUM50 avec un seuil d'identité de 50 % ;
    - BLOSUM62 avec un seuil d'identité de 62 %.



Quelle matrice de substitution choisir ?

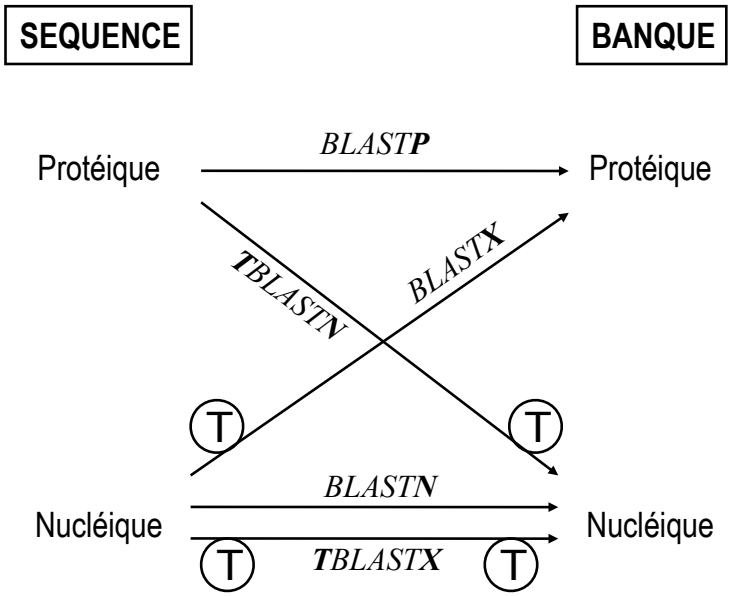


- **Pas** de matrice **idéale** ;
- Les matrices dérivées des **mutations observées** donnent, pour les protéines, de **meilleurs résultats** que les matrices basées sur l'identité, le code génétique ou les propriétés physico-chimiques.
- **Matrices PAM établies par M. Dayhoff** (1978) :
  - donnent un trop grand poids aux identités ;
  - négligent trop les ressemblances structurales ;
  - PAM250 : séquences éloignées, faible identité ;
  - PAM125 : séquences proches, identité élevée.
- **Matrices BLOSUM** (1992) :
  - construites à partir de plus de données ;
  - BLOSUM62 : séquences proches, identité élevée ;
  - BLOSUM30 : séquences éloignées, identité faible.

BLOSUM	Identité	PAM
	100	
90	90	
	80	50
	70	
62	60	
50	50	100
	40	120
30	30	
	20	250
	10	
	0	

Barton, G.J. (1996), « Protein Sequence Alignment and Database Scanning », In: *Protein Structure Prediction : A Practical Approach* (Ed. M.J.E. Sternberg), IRL Press at Oxford University Press.
http://barton.ebi.ac.uk/papers/rev93_1/rev93_1.html

BLAST : Choix du programme

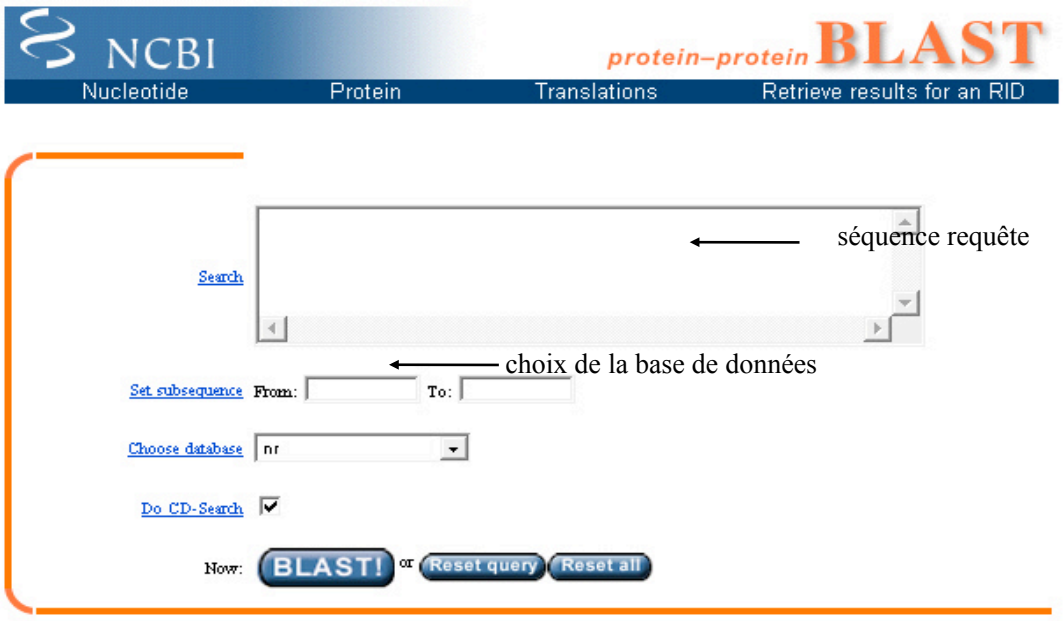


PLAN

- BLAST SUR LE WEB

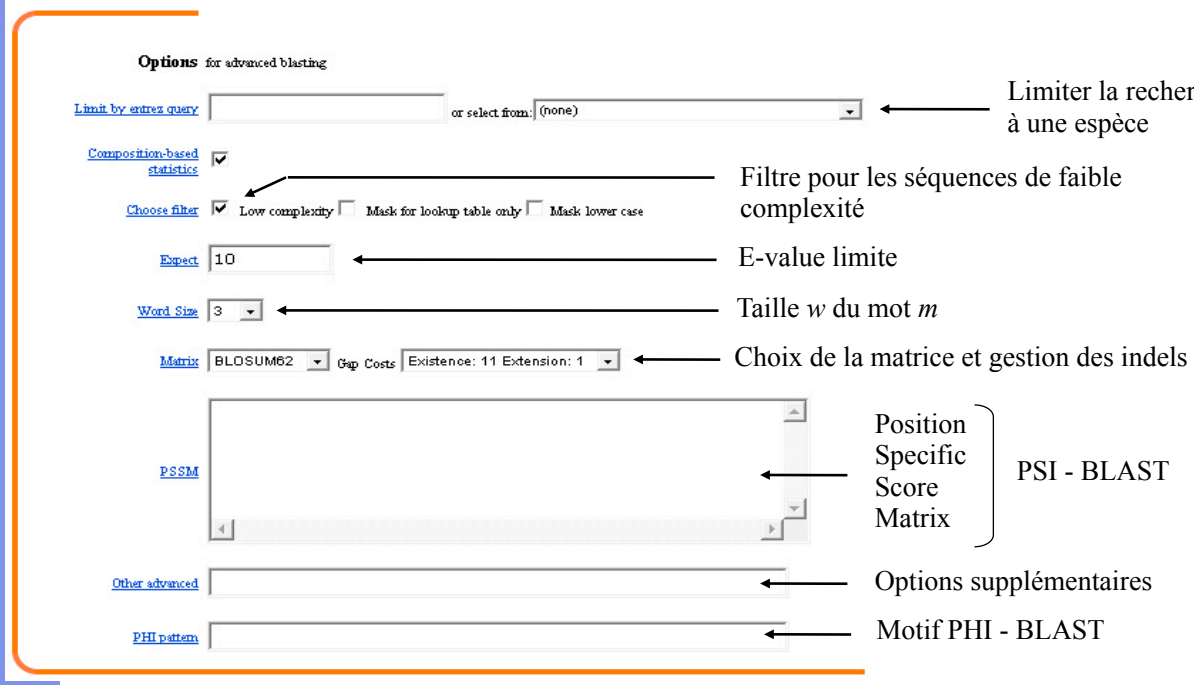
BLAST sur Internet : Requête(1/8)

http://www.ncbi.nlm.nih.gov/blast/



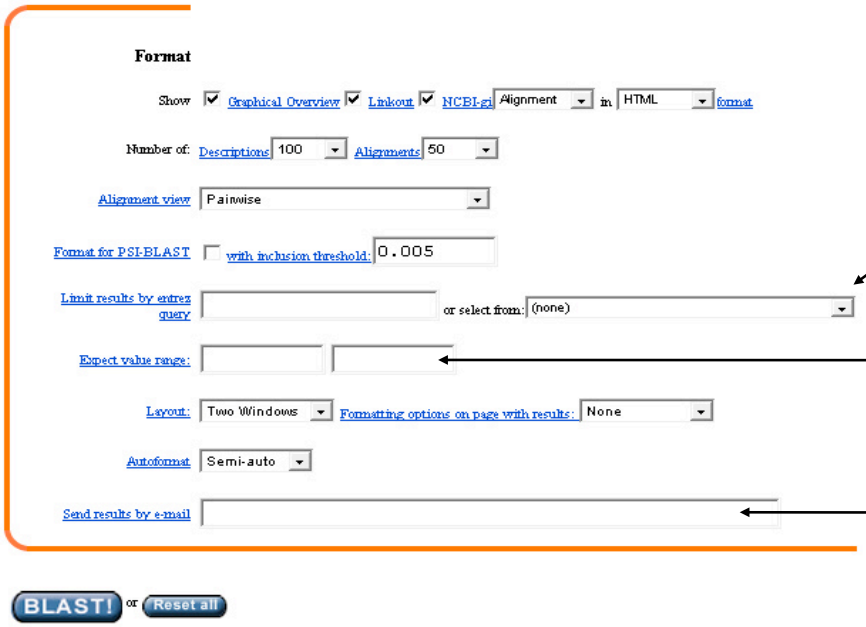
BLAST sur Internet : Paramètres (2/8)

http://www.ncbi.nlm.nih.gov/blast/



BLAST sur Internet : Format de la sortie (3/8)

http://www.ncbi.nlm.nih.gov/blast/



Limiter l 'affichage des résultats à une espèce

Limiter l'affichage des résultats à une plage de valeur d'E-value

Pour recevoir les résultats par e-mail

BLAST sur Internet : Fichier de sortie (5/8)

http://www.ncbi.nlm.nih.gov/blast/

NCBI results of BLAST

BLASTP 2.2.3 [Apr-24-2002]

Reference:
Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs", Nucleic Acids Res. 25:3389-3402.

RID: 1019992403-09001-19888

Query= gp131 proteine
(662 letters)

Database: Non-redundant SwissProt sequences
108,418 sequences; 39,574,701 total letters

If you have any problems or questions with the results of this search
please refer to the **BLAST FAQs**

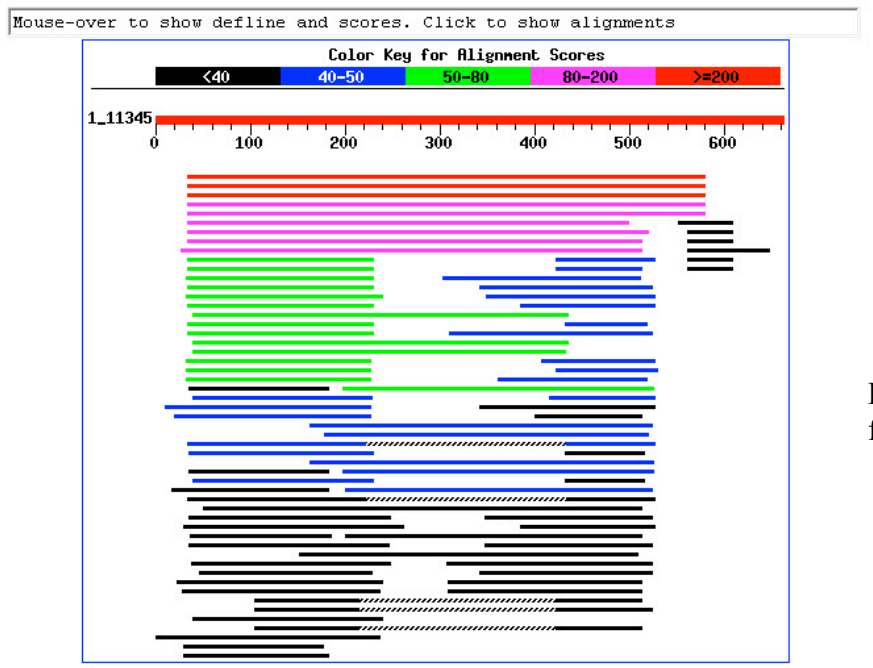
Taxonomy reports

Séquence requête

Banques de données choisies

BLAST sur Internet : Fichier de sortie (6/8)

Distribution of 134 Blast Hits on the Query Sequence ← Nombres de hits



Répartition des hits en fonction du score

http://www.ncbi.nlm.nih.gov/blast/

BLAST sur Internet : Fichier de sortie (7/8)

Sequences producing significant alignments:	Score (bits)	E Value
gi|729833|sp|P40189|IL6B_HUMAN Interleukin-6 receptor beta ...	229	1e-59
gi|729834|sp|Q00560|IL6B_MOUSE Interleukin-6 receptor beta ...	223	8e-58
gi|729835|sp|P40190|IL6B_RAT Interleukin-6 receptor beta ch...	218	3e-56
gi|729564|sp|Q99062|GCSR_HUMAN Granulocyte colony stimulati...	147	9e-35
gi|729565|sp|P40223|GCSR_MOUSE GRANULOCYTE COLONY STIMULATI...	134	4e-31
gi|1170784|sp|P42702|LIFR_HUMAN Leukemia inhibitory factor ...	108	2e-23
gi|1170785|sp|P42703|LIFR_MOUSE Leukemia inhibitory factor ...	106	1e-22
gi|12229836|sp|Q99665|I12S_HUMAN Interleukin-12 receptor be...	103	1e-21
gi|12229832|sp|P97378|I12S_MOUSE Interleukin-12 receptor be...	103	1e-21
gi|2494727|sp|Q90374|PRLR_COLLI PROLACTIN RECEPTOR PRECURSO...	67	7e-11
gi|730343|sp|Q08501|PRLR_MOUSE Prolactin receptor precursor...	65	4e-10
gi|548527|sp|Q04594|PRLR_CHICK PROLACTIN RECEPTOR PRECURSOR...	64	1e-09
gi|2494725|sp|Q28235|PRLR_CEREL PROLACTIN RECEPTOR PRECURSO...	63	1e-09
gi|2494728|sp|Q91094|PRLR_MELGA PROLACTIN RECEPTOR PRECURSO...	62	2e-09
gi|2506457|sp|P05710|PRLR_RAT Prolactin receptor precursor ...	62	3e-09
gi|6166563|sp|Q62959|LEPR_RAT Leptin receptor precursor (LE...	62	4e-09
gi|2494724|sp|Q28172|PRLR_BOVIN PROLACTIN RECEPTOR PRECURSO...	60	1e-08
gi|130323|sp|P14787|PRLR_RABIT PROLACTIN RECEPTOR PRECURSOR...	59	4e-08
gi|1352612|sp|P48356|LEPR_MOUSE Leptin receptor precursor (...	57	7e-08
gi|1352611|sp|P48357|LEPR_HUMAN Leptin receptor precursor (...	53	2e-06
gi|1168985|sp|Q08406|CNTR_RAT Ciliary neurotrophic factor r...	52	4e-06
gi|2494726|sp|Q91513|PRLR_ORENI PROLACTIN RECEPTOR PRECURSO...	52	5e-06
gi|1352099|sp|P26992|CNTR_HUMAN Ciliary neurotrophic factor...	51	8e-06
gi|125978|sp|P10586|PTPF_HUMAN LAR protein precursor (Leuko...	51	8e-06
gi|2494719|sp|O09030|I131_MOUSE Interleukin-13 receptor alp...	49	4e-05
gi|12229808|sp|Q91735|EPB3_XENLA Ephrin type-B receptor 3 p...	48	6e-05
gi|130321|sp|P16471|PRLR_HUMAN Prolactin receptor precursor...	47	2e-04
gi|1705964|sp|P51641|CNTR_CHICK CILIARY NEUROTROPHIC FACTOR...	45	5e-04
gi|12229805|sp|Q07498|EPB3_CHICK Ephrin type-B receptor 3 (...	44	9e-04
gi|3182940|sp|Q99715|CA1C_HUMAN Collagen alpha 1(XII) chain...	44	0.001
gi|12229834|sp|Q60837|I12R_MOUSE Interleukin-12 receptor be...	43	0.002

http://www.ncbi.nlm.nih.gov/blast/

BLAST sur Internet : Fichier de sortie (8/8)

```
>gi|729833|sp|P40189|IL6B_HUMAN  Interleukin-6 receptor beta chain precursor (IL-6R-beta)
          (Interleukin 6 signal transducer) (Membrane glycoprotein
          130) (GP130) (Oncostatin M receptor) (CDw130) (CD130
          antigen)
         Length = 918
```

http://www.ncbi.nlm.nih.gov/blast/

```
 Score =  229 bits (583), Expect = 1e-59
 Identities = 154/550 (28%), Positives = 250/550 (45%), Gaps = 12/550 (2%)

Query: 35  PAKPENISCVYYYRKNLTCTWSPGKETSY-TQYTVKRTYAFGEKHDNCTTNSSTSENRAS 93
           P KP+N+SC+    K + C W  G+ET   T +T+K  +A   K  +C     T     S
Sbjct: 126 PEKPKNLSCIVNEGKKMRCEWDGGRETHLETNFTLKSEWA-THKFADCKAKRDTP---TS 181

Query: 94  CSFFLPRITIPDNYTIEVEAENGDGVIKSHMTYWRLENIAKTEPPKIFRVKPVLGIKRMI 153
           C+     +    N  + VEAEN  G + S    +      K  PP    V     +  ++
Sbjct: 182 CTVDYSTVYFV-NIEVWVEAENALGKVTSDHINFDPVYKVKPNPPHNLSVINSEELSSIL 240

Query: 154 QIEWIKPELAPVSSDLKYTLRFRTVNSTSWMEVNFAKNRKDKNQTYNLTGLQPFTEYVIA 213
           ++ W  P +  V   LKY +++RT ++++W ++   ++      ++ +  L+PFTEYV
Sbjct: 241 KLTWTNPSIKSVII-LKYNIQYRTKDASTWSQIP-PEDTASTRSSFTVQDLKPFTEYVFR 298

Query: 214 LRCAVKESK-FWSDWSQEKMGMTEEEAPC-GLELWRVLKPAEADGRRPVRLLWKKARGAP 271
           +RC  ++ K +WSDWS+E  G+T E+ P      W  + P+   G R V+L+WK
Sbjct: 299 IRCMKEDGKGYWSDWSEEASGITYEDRPSKAPSFWYKIDPSHTQGYRTVQLVWKTLPPFE 358

Query: 272 VLEKTLGYNIWYYPESXXXXXXXXXXXXXXXXLHLGGESFWVSMISYNSLGKSPVATLRI 331
              K L Y +                      ++L  + +  ++   N +GKS  A L I
Sbjct: 359 ANGKILDYEVTL--TRWKSHLQNYTVNATKLTVNLTNDRYLATLTVRNLVGKSDAAVLTI 416

Query: 332 PAIQEKSFQCIEVMQACVAEDQLVVKWQSSALDVNTWMIEWFPDVDSEPTTLSWESVSQA 391
           PA  ++  +  ++A  ++ L V+W +   V  +++EW   D  P   W+
Sbjct: 417 PACDFQATHPVMDLKAFPKDNMLWVEWTTPRESVKKYILEWCVLSDKAPCITDWQQEDGT 476
```

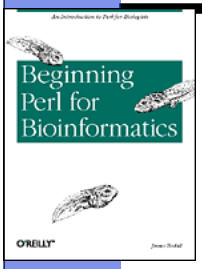

Webliographie

SÉQUENCES

- http://www.iro.umontreal.ca/~casagran/baba.html
  (Pour comprendre l'alignement)
- http://www.ncbi.nlm.nih.gov/blast/
  (Blast sur NCBI)
- http://www.isrec.isb-sib.ch/java/dotlet/dotlet_examples.html
  (Dotlet)

WEBLIOGRAPHIE

PERL



http://changjiang.whlib.ac.cn/pylorus/download/book/Beginning%20Perl%20for%20Bioinformatics/index.html

Randal L. Schwartz and Tom Christiansen, "Learning Perl" 2nd ed., O'Reilly

- Randal L. Schwartz and Tom Phoenix, "Learning Perl" 3rd ed., O'Reilly
- http://archive.ncsa.uiuc.edu/General/Training/PerlIntro excellente introduction au Perl
- http://www.comp.leeds.ac.uk/Perl/start.html bonne introduction au Perl </A>;
- http://www.ebb.org/PickingUpPerl/ Livre d'introduction au Perl redistribuable gratuitement
- http://www.cclabs.missouri.edu/things/instruction/perl/perlcourse.html Intro au Perl courte mais pas mal

LINKS

Cours PERL en Français: http://www.med.univ-rennes1.fr/~poulique/cours/perl/perl_html/introperl02.html

Perl

http://bioperl.org/

How Perl Saved the Human Genome Project

http://www.ddj.com/documents/s=928/ddj9718e/9718e.htm

Bioinformatics and open source

http://www.opensourcebioinformatics.org/

http://open-bio.org/

The Human Genome Project

http://www.ornl.gov/hgmis/

Articles

http://www.linux-mag.com/2002-06/bioinformatics_01.html

http://www.oreillynet.com/pub/a/network/2002/01/28/bioday1.html

Compléments

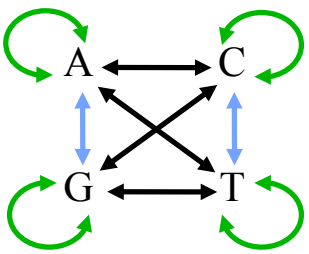
Les matrices nucléiques de substitution

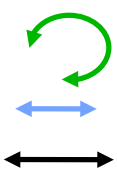
- Unitaire et unitaire inverse :

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

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

- Matrices évolutives :





Matches : 3

Transitions : 1

Transversions : 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

Les matrices protéiques de substitution

- Matrice de substitution protéique liée au code génétique
- Nombre de mutations nécessaires pour passer du codon d'un acide aminé au codon d'un autre acide aminé

Mutation GLU ⟶ LYS

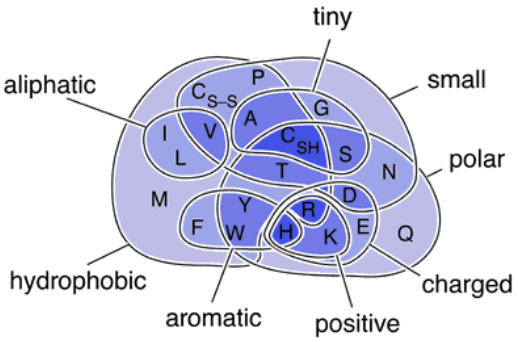
d 'ou

GAA ⟶ AAA
GAG ⟶ AAG

1 mutation sur la première base du codon

Les matrices protéiques de substitution

Les matrices liées aux propriétés physico-chimiques :



- hydrophobie / hydrophilie (Levitt, 1976) ;
- propension d'un acide aminé à intégrer une structure secondaire précise (Levin, 1986).

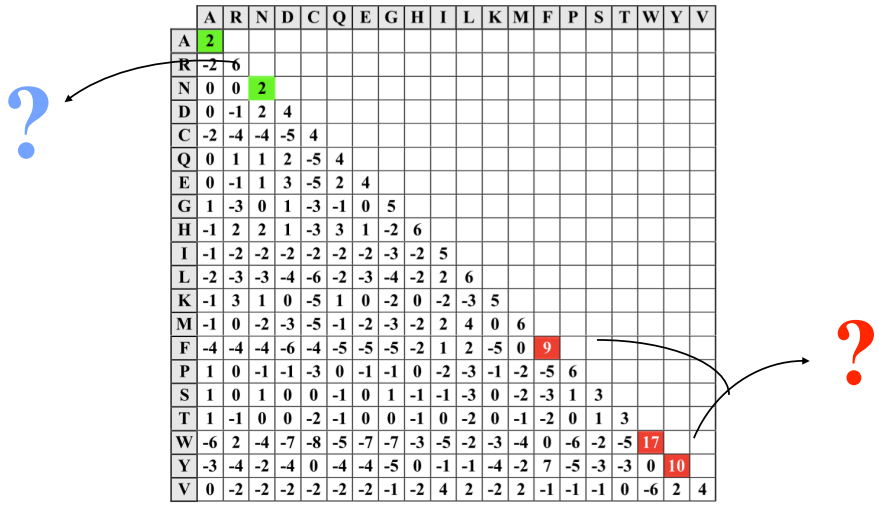
Les matrices protéiques de substitution

- Les matrices liées à l'évolution : matrices PAM
  - représentent les échanges possibles et acceptables d'un acide aminé par un autre lors de l'évolution des protéines (Dayhoff, 1978).

- Principe de base : si deux séquences appartiennent au même processus évolutif, et qu'un acide aminé de l'une a été muté pour donner l'autre, alors on peut supposer que les deux acides aminés sont similaires :
  - les mutations sont dites acceptées (**P**oint **A**ccepted **M**utation)
  - elles ont été conservées au cours de l'évolution de part leur caractère à ne pas altérer la fonction de la protéine.

*Les protéines évoluent via des successions de **mutations ponctuelles indépendantes** les unes des autres et **acceptées** dans la population.*

La matrice de substitution PAM250 (Dayhoff, 1978)



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2																			
R	-2	6																		
N	0	0	2																	
D	0	-1	2	4																
C	-2	-4	-4	-5	4															
Q	0	1	1	2	-5	4														
E	0	-1	1	3	-5	2	4													
G	1	-3	0	1	-3	-1	0	5												
H	-1	2	2	1	-3	3	1	-2	6											
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5										
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6									
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5								
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6							
F	-4	-4	-4	-6	-4	-5	-5	-5	-2	1	2	-5	0	9						
P	1	0	-1	-1	-3	0	-1	-1	0	-2	-3	-1	-2	-5	6					
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	3				
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-2	0	1	3			
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17		
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	2	4

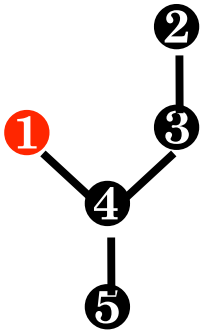
Les matrices de substitution PAM (Dayhoff, 1978) : construction (1/4)

❶ Fréquences d'échanges de paires d'acides aminés

Mutation fixée durablement au cours de l'évolution

↓

unité de distance évolutive
1 unité PAM
=
une mutation ponctuelle acceptée
entre deux séquences pour 100 résidus



Utilisation de 71 familles (1300 protéines)
de protéines à 85% identiques

↓

Construction d'arbres phylogénétiques

↓

Construction des matrices
pour chaques couples
d'acides aminés échangés

A_{ij} : nombre de mutations $i \longrightarrow j$ acceptées observées

MASTER IBM 2008UCKER©

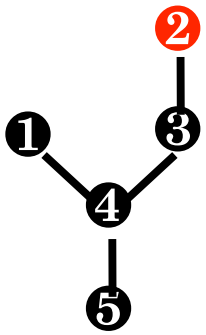
Les matrices de substitution PAM (Dayhoff, 1978) : construction (2/4)

❷ Mutabilité des acides aminés (m_j)

• Propension d'un acide aminé a être remplacé par un autre acide aminé
• Soit l'alignement suivant :

```
A D E F R E
A D D W R E
```

• Les acides aminés en jeu sont les suivants : **A**, **D**, **E**, **F**, **W** et **R**.



	A	D	E	F	W	R
Nombre de changements	0	1	1	1	1	0
Nombre d'ocurence	2	3	3	1	1	2
Mutabilité	**0**	**1/3**	**1/3**	**1**	**1**	**0**

MASTER IBM 2008UCKER©

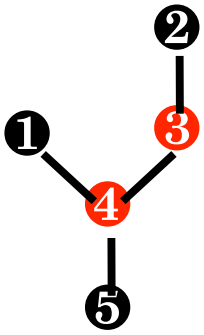
Les matrices de substitution PAM (Dayhoff, 1978) : construction (3/4)

❸ Matrice de probabilité de mutation

• Calcul de la probabilité qu'un acide aminé *i* de la matrice remplace un acide aminé *j* :

$$M_{ij}^{1} = m_j \frac{A_{ij}}{\sum_{i=1}^{20} A_{ij}}$$

• La matrice ainsi construite est appelée 1-PAM : elle définit l'unité de changement évolutif, l'unité PAM. Deux séquences sont séparées par une distance évolutive de 1 PAM si il y a eu 1 changement observé et accepté au cours de l'évolution pour 100 acides aminés.



❹ Matrice « odds »

• Chaque élément de la matrice est divisé par la fréquence d'occurrence de chaque acide aminé

$$R_{ij}^{1} = \frac{M_{ij}^{1}}{p_i}$$

MASTER IBM 2008UCKER©

Les matrices de substitution PAM (Dayhoff, 1978) : construction (4/4)

❺ Matrice « log-odds »

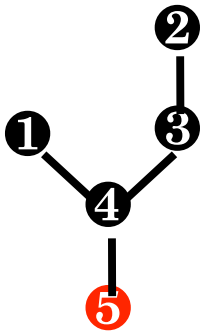
• Elle est calculée en prenant le logarithme de la matrice précedente :

$$S_{ij}^{1} = \log R_{ij}^{1}$$

• Cette matrice peut être utilisée pour simuler l'évolution en utilisant un générateur de nombres aléatoires afin de prédire le devenir de chaque résidu et ceci en accord avec les probabilités données dans la table. Les opérations suivantes sont équivalentes :
- applications successives de 1-PAM sur une séquence ;
- multiplication de la matrice 1-PAM par elle-même puis application à une séquence :

$$R_{ij}^{k} = \left(R_{ij}\right)^{k} \quad ou \quad S_{ij}^{k} = \sum_{k=1}^{n} S_{ij}^{k}$$

- rééchelonnage des éléments de 1-PAM par une constante de proportionnalité suivi de l'application à une séquence :



$$M_{ij}^{1} = \ddot{e} m_j \frac{A_{ij}}{\sum_{i=1}^{20} A_{ij}}$$

MASTER IBM 2008UCKER©

Niveau de l'information Biologique

DNA
mRNA
Proteins
Informational Pathways
Informational Networks
Cells
Organs
Individuals
Populations
Ecologies

Traditional Biology

'omics

Genomics
Functional Genomics
Proteomics
Metabolomics
S..stems Biology
..ular Biology
Medicine
Medicine
Genetics
Ecology

The Future!

MASTER IBM 2008UCKER©