Initiation à la Bioinformatique

Daniel Gautheret
ESIL, Université de la Méditerranée

V.2004.4 http://www.esil.univ-mrs.fr/~dgaut/Cours

Bioinformatique

Deux définitions possibles

- ★ Applications de l'informatique à la biologie (en anglais: *computational biology*)
- ★ Analyse de l'information biologique (en anglais: *bioinformatics*)

C'est cette bioinformatique que nous abordons ici.

L'information est:

- ★ La séquence
- ★ La structure
- ★ La fonction, les interactions etc.

Pour quoi faire?

La bioinformatique est d'abord utilisée pour identifier les gènes, étudier leur fonction et leur évolution.

- ★ "Fonction" peut être entendu dans un sens général (ATPase, RNA-Polymérase, etc.) ou dans un sens beaucoup plus précis, avec identification des résidus essentiels, éléments structuraux, sites de fixation aux ligands, site catalytiques, etc.

Par exemple…

- ★ Pour rechercher les gènes/mutations impliqués dans une pathologie/susceptibilité particulière
- ★ Pour chercher chez un organisme modèle un gène homologue à un gène humain d'intérêt
- ★ Pour rechercher chez un pathogène des gènes de résistance
- ★ Pour concevoir une expérience de mutagénèse dirigée sur une protéine
- ★ Pour trouver tous les gènes présents sur un chromosome/ génome/contig nouvellement séquencé

En fait, la liste des questions est illimitée

La déduction par homologie, ou le « dogme central » de la bioinformatique

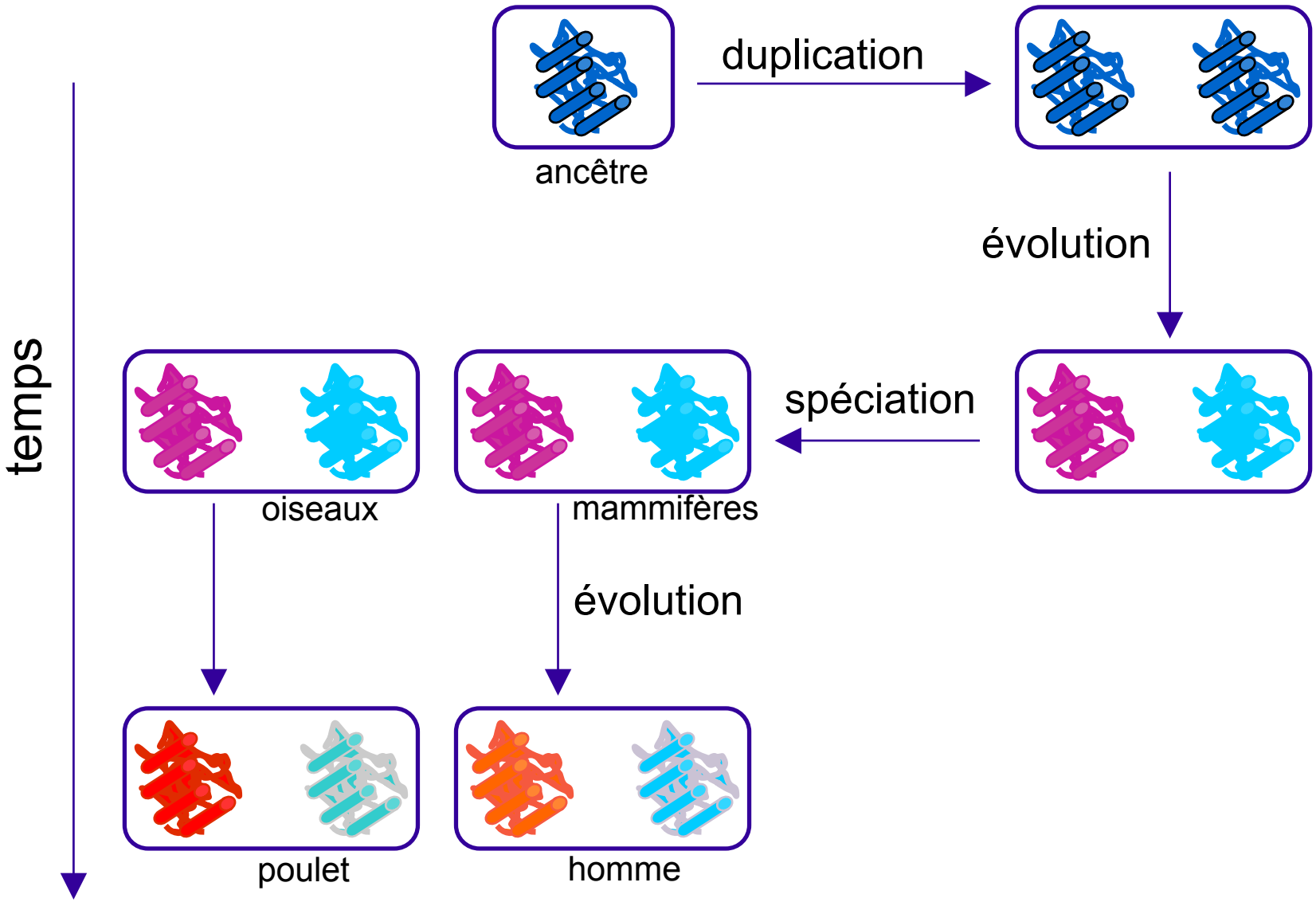
- ★ Si la bioinformatique « marche », c'est parce que l'évolution des gènes laisse une trace parfaitement visible lorsque l'on compare leur séquence
  - Les régions fonctionnelles des gènes (sites catalytique, de fixation, etc.) sont soumises à sélection. Elles sont relativement préservées par l'évolution car des mutations trop radicales sont désavantageuses.
  - Les régions non fonctionnelles ne subissent aucune pression de sélection et divergent rapidement à mesure que s'accumulent les mutations.
  - Les nouveaux gènes apparaissent surtout par remaniement de gènes ancestraux: on peut donc déduire la fonction de la plupart des gènes par comparaison avec les gènes « homologues » d'autres espèces.
  - (Evolution des gènes=mutations, insertions, délétions, recombinaisons)

L'homologie de séquence

En bioinformatique: Homologie = parenté = ancêtre commun

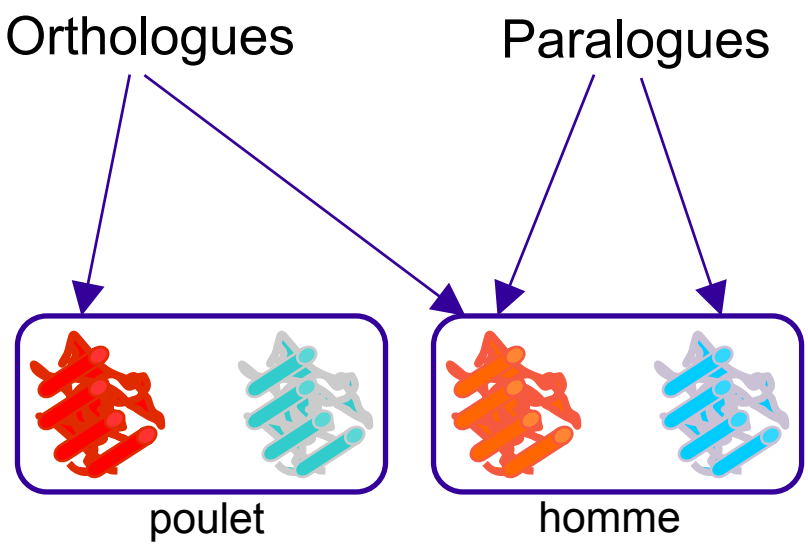
- ★ Le bras humain est homologue à l'aile de l'oiseau
- ★ Le bras humain n'est pas homologue à l'aile de la mouche
- ★ On est homologue ou on ne l'est pas.
- ★ Donc on ne dit pas: "très homologue", "faible homologie", « 28% d'homologie », etc.
- ★ Pour une notion quantitative, on parle de similitude ("très similaire", etc.) ou d'identité (28% d'indentité)

Apparition de nouveaux gènes au cours de l'évolution



Paralogues et orthologues (Fitsh, 1970)

- ★ Homologues: gènes provenant d'un ancêtre commun
- ★ Paralogues: gènes homologues issus d'un phénomène de duplication
- ★ Orthologues: gènes homologues issus de la spéciation
- ★ Transfert horizontal: par endosymbiontes, etc. Fitsch a aussi introduit "xénologue" pour évoquer ce cas.



Fonction et homologie

- Homologie n'implique pas même fonction: par exemple l'aile de l'oiseau et le bras humain n'ont pas la même fonction
- Des orthologues rapprochés (p. ex. homme/souris) ont le plus souvent la même fonction dans l'organisme.
- Des orthologues distants (p. ex. homme/mouche) ont plus rarement le même rôle *phénotypique*, mais peuvent exercer le même rôle dans une *voie* donnée.
- Les paralogues acquièrent rapidement des fonctions différentes

Comment détecter une homologie?

Principe: comparaison de séquences

- La comparaison de deux séquences est la méthode principale. Elle permet d'observer les régions conservées. On déduit l'homologie de la conservation.
- D'autres méthodes existent:
  - Analyse statistique des « mots » contenus dans la séquence
  - Méthodes dérivées de la comparaison de séquence: recherche de domaines ou motifs communs

A quel point des séquences homologues se ressemblent-elles?

- ★ De 100% à quelques nucléotides/aminoacides en commun.
- ★ Il n'y a pas vraiment de limite, mais en dessous de 25% d'identité (*twilight zone*), il devient difficile de distinguer une homologie d'une ressemblance fortuite. 2 séquences d'ADN prises au hasard ont 25% de nt communs.
- ★ Des séquence sans ressemblance apparente peuvent parfaitement être homologues (on le retrouve par ex. au niveau 3D)
- ★ Par contre, étant donné la dimension de l'espace des séquences possibles, une ressemblance importante est généralement interprétée comme une homologie, et non pas comme une évolution convergente.

La comparaison de séquences

- Comparer des séquences serait relativement simple si elles avaient toutes la même longueur. Comme ce n'est pas le cas, il faut les aligner, c'est à dire trouver où se trouvent les insertions et délétions, représentées par des « indels » (« gaps »)

★ Distance d'édition

- Selon ce concept, le bon alignement est celui qui minimise les opérations à réaliser pour passer d'un séquence à l'autre.
- Opérations: conservation, remplacement/mutation, délétion, insertion. Une pénalité peut être affectée à chaque opération, par exemple c=0, m=1, d=2, i=2. La distance finale entre les deux séquences (distance d'édition) est la somme de ces pénalités.

Seq 1	**CAGTGGT-GC**
Seq 2	**CA-TCGTAGC**
distance	**ccicmccdcc = 0+0+2+0+1+0+0+2+0+0 = 5**

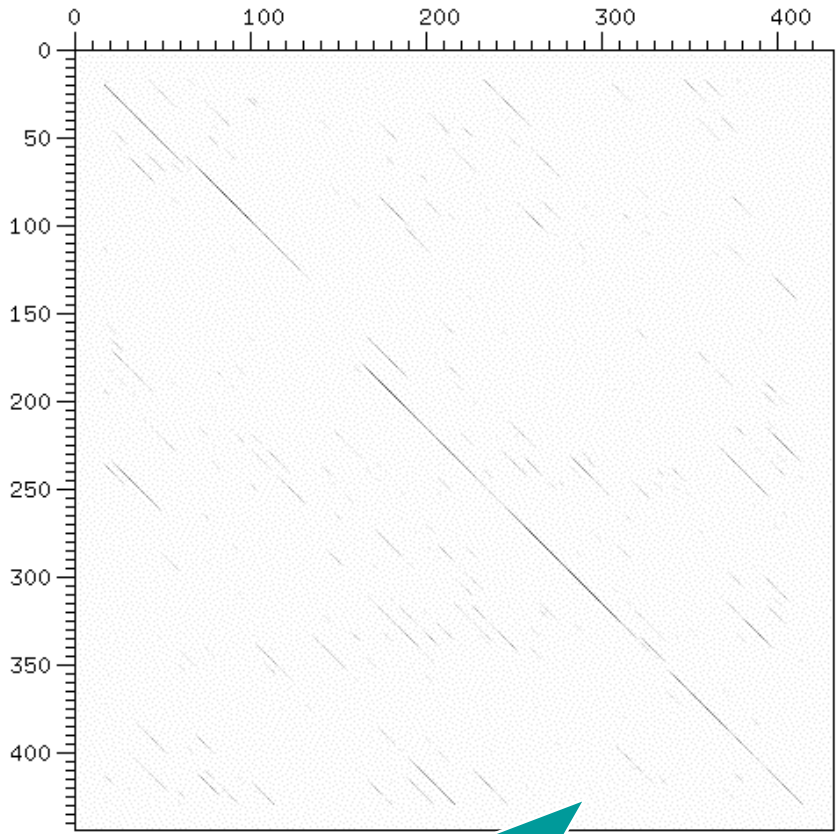
★ Comment trouver le meilleur alignement?

- Le nombre d'alignements possibles est trop élevé: on ne peut pas les essayer tous pour trouver celui qui minimise la distance.

Comparaison de séquences

★ Les "dot plots"

- Deux séquences à comparer sont représentées (ici 2 gènes de globine), une horizontalement, l'autre verticalement. On dessine ensuite un point dans la matrice lorsque les deux positions correspondantes sont identiques. Lorsque des régions se ressemblent, on voit apparaître une diagonale. Les décalages entre les diagonales correspondent à des insertions ou délétions. Plusieurs diagonales parallèles indiquent une répétition.
- Pour "nettoyer" le dot plot, on utilise souvent non pas un point par base, mais un point lorsque n bases sont identiques, ou n bases identiques dans une fenêtre de N. Cela réduit considérablement le nombre de points.
- Les dot plots sur des génomes complets permettent de visualiser les évènements à grande échelle, la synthénie, etc.

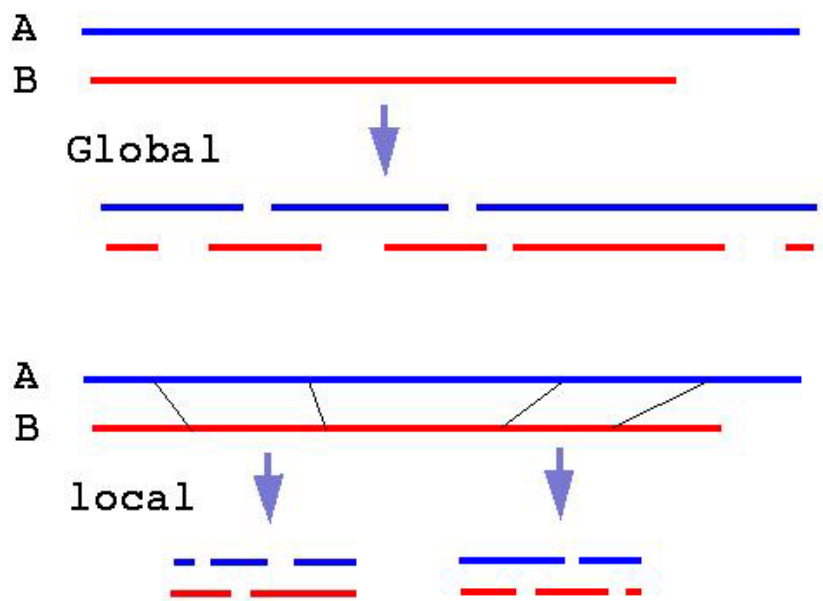


Alignement: trouver le meilleur chemin dans ce graphe

Alignement local ou global

Les finalités sont très différentes.

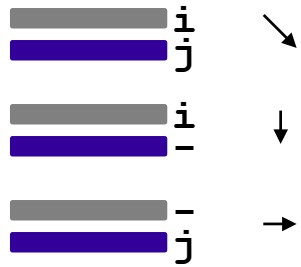
- ★ l'alignement global est conçu pour comparer des séquences homologues sur toute leur longueur.
- ★ L'alignement local est conçu pour rechercher des régions semblables entre A et B.



Algorithme de programmation dynamique

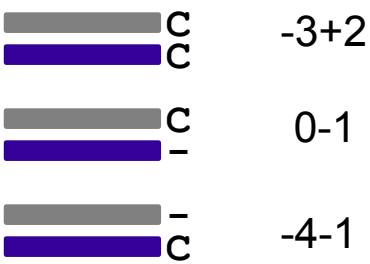
Etape 1: Remplissage de la matrice

- ★ On veut aligner les séquences CAGTG et ACTCGT.
- ★ Une matrice de scores est remplie par:




```
v ([0->i], [0->j]) = max {
  v ([0->i-1], [0->j-1]) + v ([i],[j])
  v ([0->i-1], [0->j ]) + v ([i],[-])
  v ([0->i ], [0->j-1]) + v ([-],[j]) }
Avec par exemple:
  v ([],[])=0
  v ([-],[])=-1
  v ([x],[y])=-1 (x different de y)
  v ([x],[x])=+2
```

Par ex. l'entrée {4,1} (colorée) est obtenue par max (-3+2, 0-1, -4-1), c'est à dire -1.



-3+2

0-1

-4-1

i \ j		C	A	G	T	G
	0	-1	-2	-3	-4	-5
A	-1	-1	1	0	-1	-2
C	-2	1	0	0	-1	-2
T	-3	0	0	-1	2	1
C	-4	-1	-1	-1	1	1
G	-5	-2	-2	1	0	3
T	-6	-3	-3	0	3	2

Exemple tiré du cours Bioinformatics & Computational Genomics du Weizmann Institute

Programmation dynamique

Etape 2: Reconstitution de l'alignement

- ★ Le chemin (alignement) optimal est déterminé par l'algorithme:
- ➢ en partant de la cellule (m,n), retrouver quelle cellule était responsable de la cellule courante (chemin suivi représenté par les traits rouges).
- ➢ Ici 3 Solutions:

```
ACTCGT-
-C-AGTG

ACTCGT-
-CA-GTG

-ACTCGT
CAGT-G-
```

i \ j		C	A	G	T	G
	0	-1	-2	-3	-4	-5
A	-1	-1	1	0	-1	-2
C	-2	1	0	0	-1	-2
T	-3	0	0	-1	2	1
C	-4	-1	-1	-1	1	1
G	-5	-2	-2	1	0	3
T	-6	-3	-3	0	3	2

Matrices de Substitution

- Matrice 4X4 (nt) ou 20x20 (aa) décrivant la distance ou la similitude entre résidus.
- Estiment le coût ou le taux de remplacement d'1 résidu par un autre (distance).
- Le choix d'une matrice affecte fortement le résultat de l'analyse. Chaque matrice de score représente implicitement une théorie évolutive donnée

Matrices DNA

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

Matrice identité

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

Matrice transition/transversion

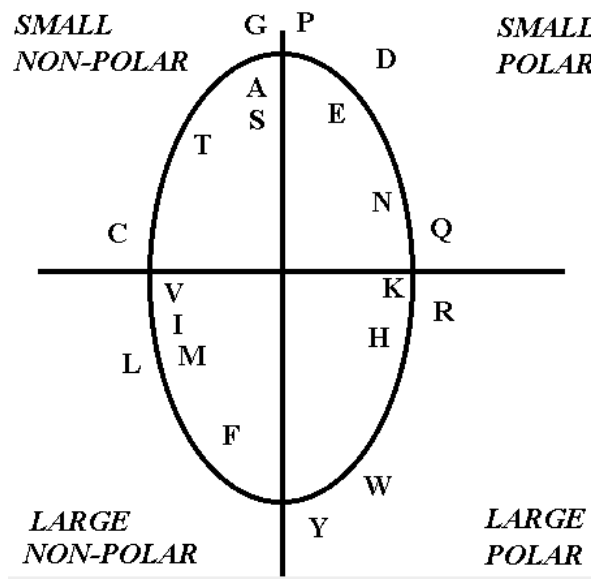
Matrices de Substitution

Matrices fondées sur le code génétique

- ★ Les scores sont déterminés en fonction du nombre commun de nucléotides présents dans les codons des acides aminés, ce qui revient à considérer le minimum de changements nécessaires en bases pour convertir un acide aminé en un autre.

Matrices fondées sur les propriétés physicochimiques

- ★ Les plus courantes sont celles basées sur le caractère hydrophile ou hydrophobe des protéines. Ces matrices sont peu utilisées.



Une représentation bidimensionnelle des propriétés des aa calculée d'après la matrice de Dayhoff par G. Vriend, Centre for Molecular and Biomolecular Informatics, University of Nijmegen

Matrices de Dayoff ou PAM

- ★ **PAM = Percentage of Accepted point Mutation**
- ★ Probabilité d'observer la mutation X->Y après un temps évolutif donné.
  Basé sur alignement de protéines conservées à + de 85%.

•Proposé par Margaret Dayhoff en 1978.

•Chaque case représente la probabilité de voir ces deux résidus remplacés l'un par l'autre dans un alignement. (matrice lod-score, de "log-odds" ou "log des chances").

•Un exemple de lod-score est:
S = log (Fij / (Fi x Fj))
Où Fij est la fréquence de remplacement du résidu i par j, et Fi et Fj sont les fréquences respectives des résidus i et j.

•Dans cette matrice de similitude, plus la valeur est négative, plus la probabilité est faible, plus le remplacement est rare.

•La table est valable pour une certaine distance évolutive.

•La distance est mesurée en PAM: nbre de mutations ponctuelles par 100 aa.

•2 Séquences séparées par une unité PAM: 1 mutation par 100 aa.

•Les valeurs sont déterminées initialement pour des protéines séparées de 6 à 100 PAM, puis extrapolées pour 150, 250 PAM, etc.

•Pour des protéines éloignées, on ne pourrait pas directement extrapoler à partir de valeurs tirées par ex. de PAM 10, car la *nature* des mutations change avec la distance évolutive. Le code génétique, par exemple, influence les mutations permises sur une courte durée, mais pas sur une longue durée.

Matrice de Dayoff (1979)

A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Z	
0.4	0.0	-0.4	0.0	0.0	-0.8	0.2	-0.2	-0.2	-0.2	-0.4	-0.2	0.0	0.2	0.0	-0.4	0.2	0.2	0.0	-1.2	-0.6	0.0	A
	0.5	-0.9	0.6	0.4	-1.0	0.1	0.3	-0.4	0.1	-0.7	-0.5	0.4	-0.2	0.3	-0.1	0.1	0.0	-0.4	-1.1	-0.6	0.4	B
		2.4	-1.0	-1.0	-0.8	-0.6	-0.6	-0.4	-1.0	-1.2	-1.0	-0.8	-0.6	-1.0	-0.8	0.0	-0.4	-0.4	-1.6	0.0	-1.0	C
			0.8	0.6	-1.2	0.2	0.2	-0.4	0.0	-0.8	-0.6	0.4	-0.2	0.4	-0.2	0.0	0.0	-0.4	-1.4	-0.8	0.5	D
				0.8	-1.0	0.0	0.2	-0.4	0.0	-0.6	-0.4	0.2	-0.2	0.4	-0.2	0.0	0.0	-0.4	-1.4	-0.8	0.6	E
					1.8	-1.0	-0.4	0.2	-1.0	0.4	0.0	-0.8	-1.0	-1.0	-0.8	-0.6	-0.6	-0.2	0.0	1.4	-1.0	F
						1.0	-0.4	-0.6	-0.4	-0.8	-0.6	0.0	-0.2	-0.2	-0.6	0.2	0.0	-0.2	-1.4	-1.0	-0.1	G
							1.2	-0.4	0.0	-0.4	-0.4	0.4	0.0	0.6	0.4	-0.2	-0.2	-0.4	-0.6	0.0	-0.4	H
								1.0	-0.4	0.4	0.4	-0.4	-0.4	-0.4	-0.4	-0.2	0.0	0.8	-1.0	-0.2	-0.4	I
									1.0	-0.6	0.0	0.2	-0.2	0.2	0.6	0.0	0.0	-0.4	-0.6	-0.8	0.1	K
										1.2	0.8	-0.6	-0.6	-0.4	-0.6	-0.6	-0.4	0.4	-0.4	-0.2	-0.5	L
											1.2	-0.4	-0.4	-0.2	0.0	-0.4	-0.2	0.4	-0.8	-0.4	-0.3	M
												0.4	-0.2	0.2	0.0	0.2	0.0	-0.4	-0.8	-0.4	0.2	N
													1.2	0.0	0.0	0.2	0.0	-0.2	-1.2	-1.0	-0.1	P
														0.8	0.2	-0.2	-0.2	-0.4	-1.0	-0.8	0.6	Q
															1.2	0.0	-0.2	-0.4	0.4	-0.8	0.6	R
																0.4	0.2	-0.2	-0.4	-0.6	-0.1	S
																	0.6	0.0	-1.0	-0.6	-0.1	T
																		0.8	-1.2	-0.4	-0.4	V
																			3.4	0.0	-1.2	W
																				2.0	-0.8	Y
																					0.6	Z

Autres matrices de substitution

BLOSUM

- ★ Le but est de détecter des relations entre protéines plus éloignées.
- ★ Avec les matrices PAM, les valeurs pour des protéines éloignées sont extrapolées. Avec BLOSUM, ces valeurs sont obtenues en comparant des blocs facilement alignables (sans gaps) dans des familles de protéines très éloignées.
- ★ Ces matrices sont reconnues pour mettre en valeur les similitudes biologiquement importantes (celles qui sont présentes dans les régions alignées sans gaps).
- ★ BLOSUM62: faite à partir d'un alignement de séquences ayant 62% de similitude, BLOSUM45: 45%, etc.

Matrices d'après alignement 3D

- ★ Basées sur la structure secondaire ou tertiaire. Evaluent la propension d'un acide aminé à adopter une certaine conformation. Fiables car fondées sur le meilleur alignement possible. Encore incomplètes en raison de la taille des banques de données 3D.

Les programmes d'alignement global

- ★ Méthode employée pour aligner des séquences dont on soupçonne l'homologie. L'alignement est optimisé sur toute la longueur des séquences. L'algorithme de référence est celui de Needleman & Wunsh (1970).
- ★ Utilisé principalement aujourd'hui dans le cadre de l'alignement multiple (voir plus loin)

Les programmes d'alignement local

- Aligne seulement les régions dont le score est supérieur à un seuil donné. Utilisé lorsque l'on veut aligner deux séquences de taille très différente. (par ex. dans une recherche de sous-séquence). Beaucoup plus rapide que l'alignement global.

★ **Smith-Waterman**

- Programmation dynamique avec arrêt de la procédure sur un critère de score. Selection du meilleur alignement local.

★ **Fasta (Lipman & Pearson, 1985)**

- Heuristique: recherche d'abord des segments de longueur k exactement semblables (k-mots), raccorde ces segments si présents sur une même diagonale où sur des diagonales proches, puis réaligne la région par programmation dynamique. Une seule solution par couple de séquences comparées.

Blast (Lipman, Karlin, Altschul, 1990)

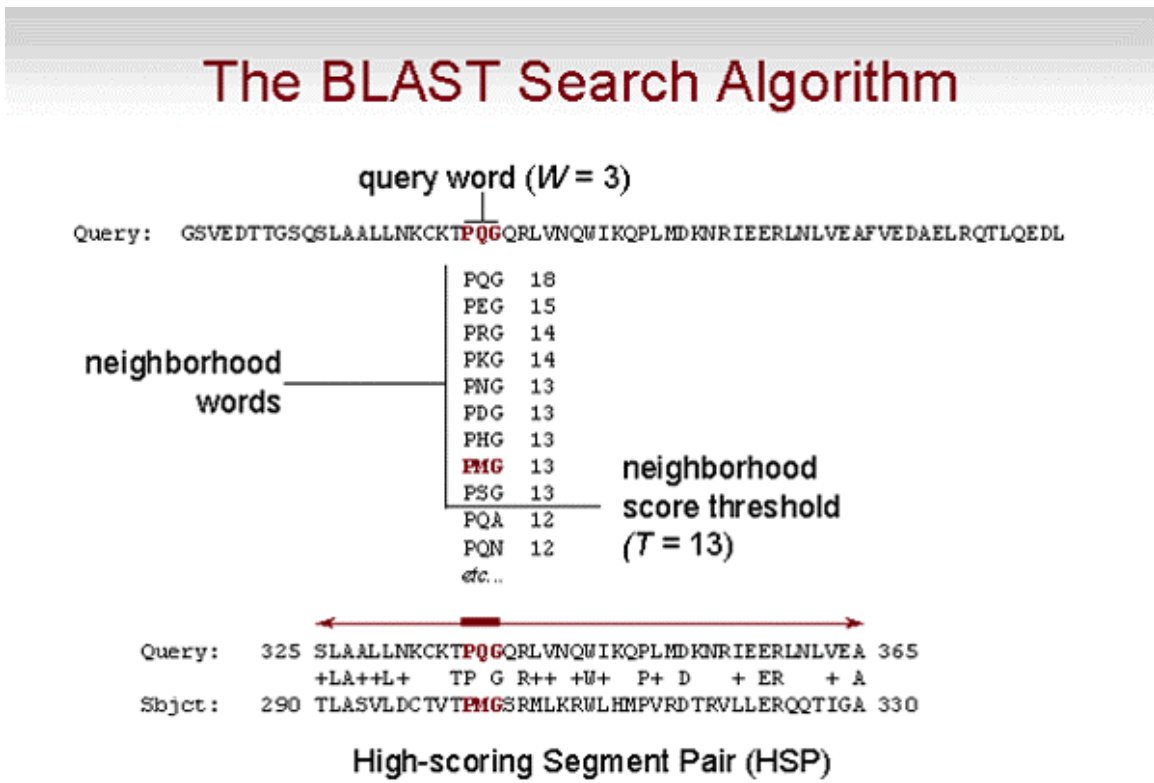
★ Le plus utilisé des programmes d'alignement local

- k-mots également, mots approchés permis au dessus d'un certain score.
- Pré-codage de la base de données et de la requête pour recherche plus rapide des k-mots.
- Version 1: sans Gaps
- Version 2: avec Gaps

★Points forts

•Rapidité

•Calcul de la valeur statistique des scores.

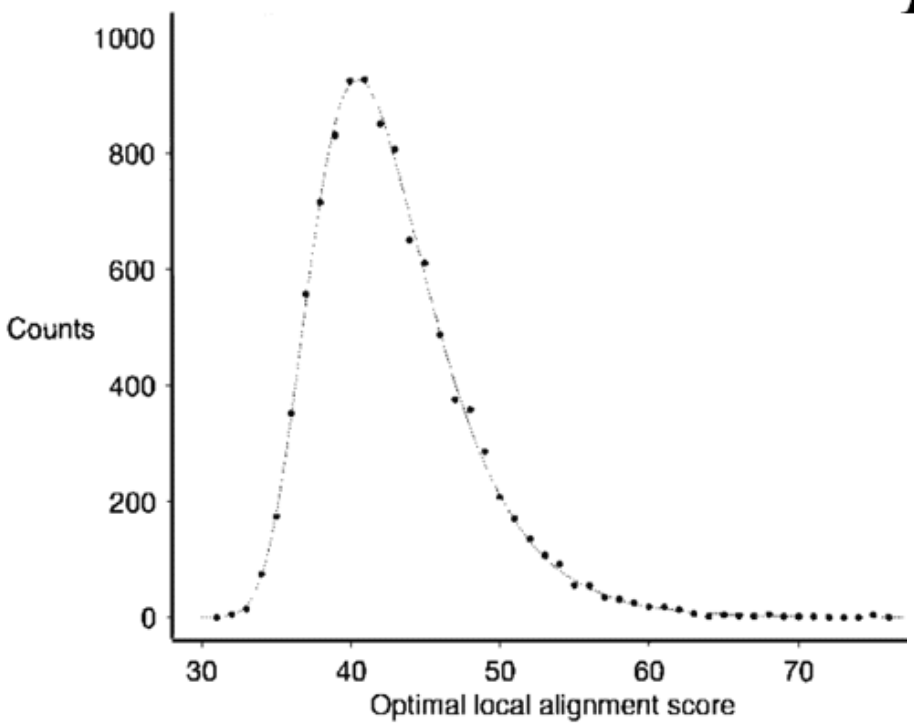


Statistiques de Blast

E value: nombre de solutions attendues par chance avec un score *S* ou plus

- En raison de l'algorithme d'alignement, qui recherche le meilleur score possible pour une position donnée, les scores des HSP ne suivent pas une distribution normale, mais une **distribution des valeurs extrêmes**. En comparant 2 séquences de longueurs n et m, le nombre attendu de HSP ayant un score S ou plus est défini par:

$$E = Kmne^{-\lambda S}$$



où K et lambda sont des paramètres statistiques dépendant du système de score et de la composition de fonds en aminoacides. Blast estime ces paramètres a priori pour les différents systèmes de score (BLOSUM62, etc.). Pour un alignement sans gaps, K et lambda peuvent être calculés. Pour un alignement avec gaps, il a fallu recourir à des simulations sur un grand nombre de séquences aléatoires.

exemple de tracé de scores d'alignement optimaux

Statistiques de Blast

P value

- La probabilité de trouver au moins un HSP de score >= S est:

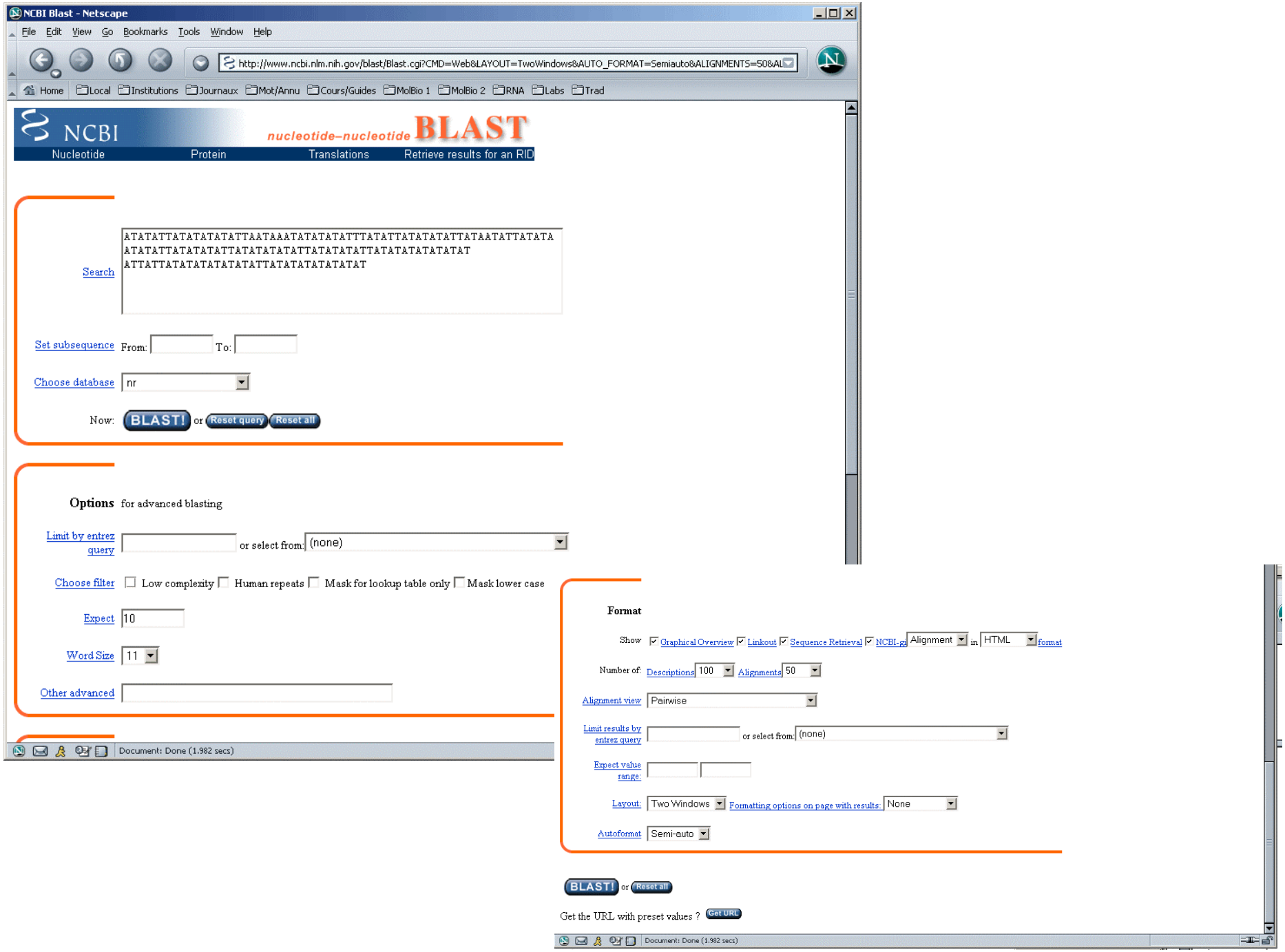
$$P = 1 - e^{-E} = 1 - \exp(-Kmn\, e^{-\lambda x})$$

Recherche dans les banques

- Les équations de E et P s'appliquent à la comparaison de 2 séquences. Si l'on compare une séquence à une banque en contenant un grand nombre, les chances d'obtenir un certain score sont bien sûr plus élevées. Blast fait comme si la recherche s'effectuait dans une longue séquence de longueur N (longueur totale de la banque), en tenant compte en outre des effets de bordure (en raison de leur longueur, les séquences "requêtes" ne peuvent arriver trop près des bords des séquences de la base).

Blast en pratique

- ★ Visiter Le serveur Blast du NCBI:
- ★ Programs: *blastn:* AN contre AN *blastp:* Prot contre Prot *blastx:* AN 6 cadres contre prot *tblastn:* Prot contre AN 6 cadres *tblastx:* AN 6 cadres contre AN 6 cadres
- ★ Database: nr (non redondant) est automatiquement selectionné en version "protéines" ou "acides nucléiques" selon qu'on utilise blastp ou blastn.
- ★ Masquage: les régions risquant de produire des solutions non spécifiques peuvent-être remplacées pasr des X
  - Régions de basse complexité
  - Séquences répétées eucaryotiques


NCBI Blast - Netscape
File Edit View Go Bookmarks Tools Window Help
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?CMD=Web&LAYOUT=TwoWindows&AUTO_FORMAT=Semiauto&ALIGNMENTS=50&AL
Home Local Institutions Journaux Mot/Annu Cours/Guides MolBio 1 MolBio 2 RNA Labs Trad
NCBI
nucleotide–nucleotide BLAST
Nucleotide Protein Translations Retrieve results for an RID
Search
ATATATTATATATATATTAATAAATATATATATTTATATTATATATATTATAATATTATATA
ATATATTATATATATTATATATATATTATATATATTATATATATATATAT
ATTATTATATATATATATATTATATATATATAT
Set subsequence From: To:
Choose database nr
Now: BLAST! or Reset query Reset all
Options for advanced blasting
Limit by entrez query or select from: (none)
Choose filter Low complexity Human repeats Mask for lookup table only Mask lower case
Expect 10
Word Size 11
Other advanced
Document: Done (1.982 secs)
Format
Show Graphical Overview Linkout Sequence Retrieval NCBI-gi Alignment in HTML format
Number of: Descriptions 100 Alignments 50
Alignment view Pairwise
Limit results by entrez query or select from: (none)
Expect value range:
Layout: Two Windows Formatting options on page with results: None
Autoformat Semi-auto
BLAST! or Reset all
Get the URL with preset values ? Get URL
Document: Done (1.982 secs)

Sortie de Blast

Query=
(734 letters)

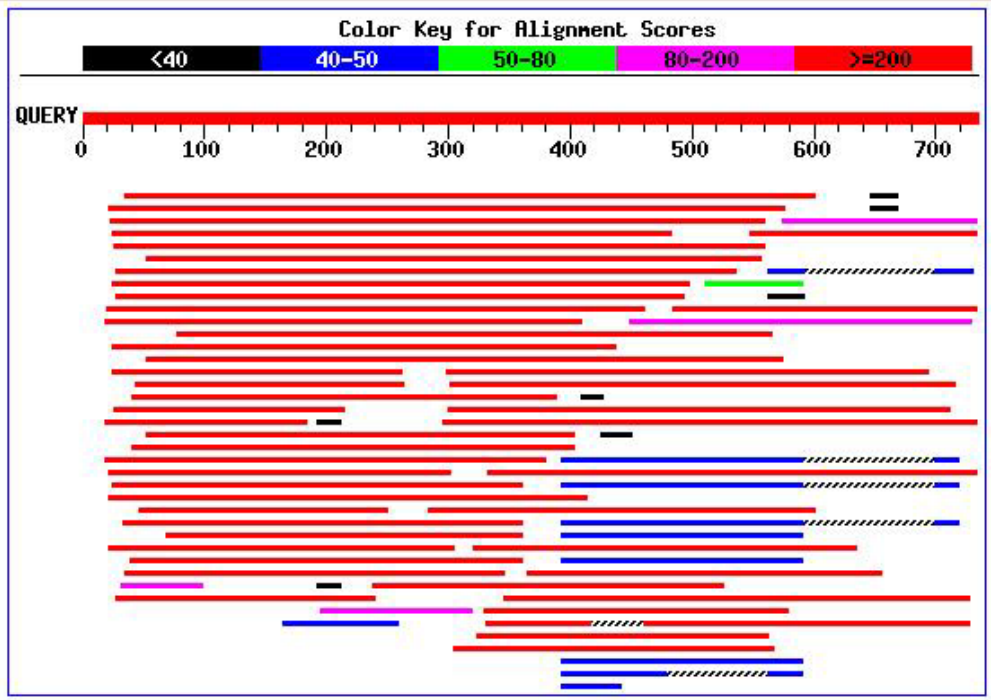
Database: Non-redundant Database of GenBank EST Division
1,938,225 sequences; 736,227,809 total letters

Searching.....done

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

Distribution of 76 Blast Hits on the Query Sequence

Mouse-over to show defline and scores. Click to show alignments



Sortie de Blast

```
gb|AA421570|AA421570  zu25d04.s1 Soares NhHMPu S1 Homo sapiens c...  765  0.0
gb|AA670218|AA670218  ad19g11.s1 Soares NbHFB Homo sapiens cDNA ...  739  0.0
gb|AA628778|AA628778  af42c05.s1 Soares total fetus Nb2HF8 9w Ho...  722  0.0
gb|AI246150|AI246150  qi29a04.x1 Soares_NhHMPu_S1 Homo sapiens c...  704  0.0
gb|AA588195|AA588195  no23h07.s1 NCI_CGAP_Pr22 Homo sapiens cDNA...  690  0.0
gb|AI143969|AI143969  qe01c10.x1 Soares_testis_NHT Homo sapiens ...  690  0.0
gb|AA977500|AA977500  on60d06.s1 Soares_NFL_T_GBC_S1 Homo sapien...  680  0.0
gb|AA779757|AA779757  af44e01.s1 Soares total fetus Nb2HF8 9w Ho...  642  0.0
gb|AA487274|AA487274  aa94e08.s1 Stratagene fetal retina 937202 ...  624  e-177
gb|AA953515|AA953515  on80a09.s1 Soares_NFL_T_GBC_S1 Homo sapien...  624  e-177
gb|AA129683|AA129683  zn91b03.s1 Stratagene lung carcinoma 93721...  617  e-175
gb|AA938880|AA938880  op74c01.s1 Soares_NFL_T_GBC_S1 Homo sapien...  585  e-165
ares_NFL_T_GBC_S1 Homo sapien...  585  e-165
ares parathyroid tumor NbHPA ...  575  e-162
ares_NFL_T_GBC_S1 Homo sapien...  533  e-150
I_CGAP_Thy1 Homo sapiens cDNA...  398  e-109
apiens cDNA clone 286077 3'.      274  1e-71
ratagene fetal retina 937202 ...  202  3e-50
ares NhHMPu S1 Homo sapiens c...  200  1e-49
ares NhHMPu S1 Homo sapiens c...  200  1e-49
ratagene lung carcinoma 93721...  151  1e-34

gb|N92166|N92166 yz89b07.r1 Homo sapiens cDNA clone 290197 5'.
          Length = 479

 Score = 52.0 bits (26), Expect = 7e-05
 Identities = 89/105 (84%), Positives = 89/105 (84%), Gaps = 9/105 (8%)


Query: 7   ccctaatttgtagccagtccaacccttcattccttggaggatttagttttgggataaaa 66
           ||| |||||||||||||| |  |||||||| | |||||||||||| |||||||  |||
Sbjct: 384 cccnaatttgtagccagtcaa--cctttcatnc-ttggaggatttag-tttggga-naaa 438


Query: 67  attttggtcccttgggcacagagacattccactattaatgaagta 111
           || ||||   ||||||||||||||||   |||||||||||||
Sbjct: 439 atnttgg--ncttgggcacagagacat--nactattaatgaagta 479


 gb|AI044766|AI044766 UI-R-C1-kb-g-03-0-UI.s1 UI-R-C1 Rattus norvegicus cDNA
          UI-R-C1-kb-g-03-0-UI 3'
          Length = 432

 Score = 44.1 bits (22), Expect = 0.017
 Identities = 31/34 (91%), Positives = 31/34 (91%)


Query: 243 gaaaaaatttttggtaaacagtattgtgtaaaat 276
           |||||| ||| ||||||||||||| ||||||||||
Sbjct: 295 gaaaaagtttatggtaaacagtaatgtgtaaaat 262


 Score = 40.1 bits (20), Expect = 0.27
 Identities = 20/20 (100%), Positives = 20/20 (100%)


Query: 391 tgtgtaaatttaatataaca 410
           ||||||||||||||||||||
Sbjct: 135 tgtgtaaatttaatataaca 116
```

A connaître:

- Score
- Identities
- Expect
- Query
- Subject

Exemples usage Blast (suivre les liens)

- Blastx cDNA / Prot
- Blastn cDNA / Gb
- Blastn cDNA / dbEST
- Pieges: vecteurs
- Pièges: basse complex.
- Pièges: Alu
- Pièges: transmembrane

- Caractériser le gène correspondant à un cDNA
- Trouver d'autres cDNA ou mRNA semblables
- Trouver des EST correspondant à un cDNA (p.ex. pour profil d'expression)
- Attention: il reste souvent des séquences de vecteurs attachées aux ARNm
- Séquences riches en AT, riches en résidus hydrophobes, etc.
- Séquences répétées=40% du génome humain
- Séquences transmembranaires (hydrophobes): même comportement que basse complexité

Vaut-il mieux comparer les protéines ou l'ADN pour rechercher des homologues d'une séquence?

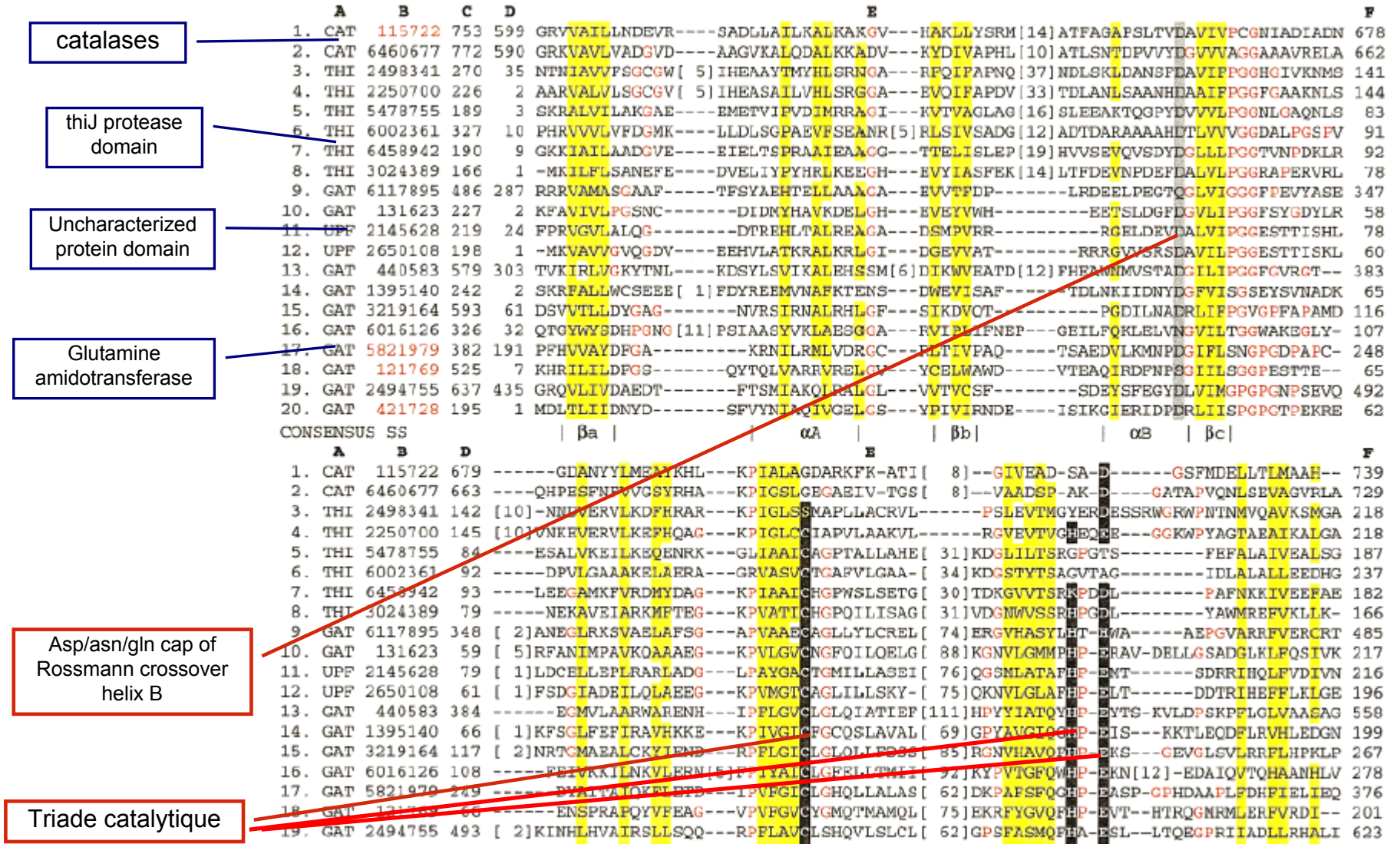
- La meilleure façon de détecter des similitudes entre séquences est généralement la *comparaison au niveau protéique.*
  1. Il existe 20 aa contre 4 bases. La probabilité de trouver une "lettre" donnée par hasard est donc plus importante pour les bases.
  2. Plusieurs codons produisent le même aa. 134 / 549 substitutions de bases sont synonymes. Les séquences protéiques sont plus informatives.
  3. La raison principale est en fait l'existence d'outils de comparaison plus puissants pour les aa: utilisation des propriétés physicochimiques ou des substitutions observées dans l'évolution. Même lorsque les aa sont différents, on est capable de retrouver des similitudes. On en est tout à fait incapable au niveau des bases.
- Il existe en fait des cas où la séquence d'ADN est plus conservée que la séquence protéique, ce qui enlève du poids à l'argument 1
- Les comparaisons avec les séquences protéiques ne permettent de détecter que les régions codantes. Evidemment, on utilisera toujours la séquence ADN/ARN pour analyser ce qui n'est pas traduit!

Un protocole pour l'analyse fonctionnelle complète d'une protéine

- Déterminer avec fiabilité la fonction d'une protéine d'après sa séquence demande l'application d'un protocole bien précis. On peut lire à ce sujet le papier de Bork et Koonin (Nature genetics, 1998, 18, 313). Les principaux points sont les suivants:
- **1. Eléments structuraux**
- L'identification des éléments suivants permet de s'assurer que les recherches d'homologies se déroulent sans artefact.
  - Régions de basse complexité (pour masquage)
  - Régions transmembranaires
  - Répétitions internes (ex: D1-D2-D1)
- **2. Homologies**
- C'est par homologie que l'on identifie les domaines/motifs fonctionnels de la protéine...
  - Identification de domaines connus (Prosite-Prodom)
  - Recherches itératives pour l'extension de la famille d'homologues
    - Manuellement (avec BLAST) ou automatiquement (avec PSI Blast)
    - PAM250 au lieu de BLOSUM62
    - Si protéine modulaire: masquer les domaines identifiés et lancer la recherche sur la partie restante.
    - Précautions essentielles: ne pas considérer que les meilleurs scores Blast, ne pas croire systématiquement les annotations, Tenir compte du contexte (par ex: pas de séquence signal dans un doigt de zinc)
- **3. Synthèse**
  - Alignement multiple pour visualisation finale des domaines conservés/absents.
  - Vérifier la présence de paralogues (par ex: substrats différents). La meilleure approche est de réaliser l'arbre phylogénétique.

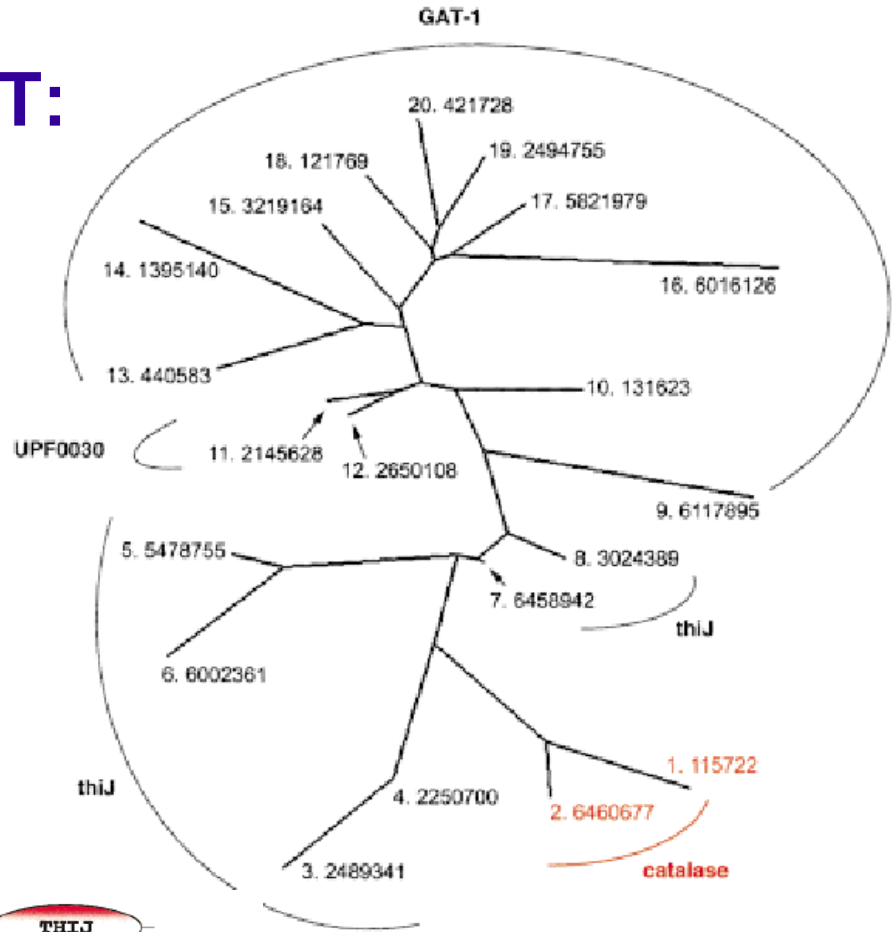
Exemple d'analyse bioinformatique: les Glutamine Aminotransferases (GAT)

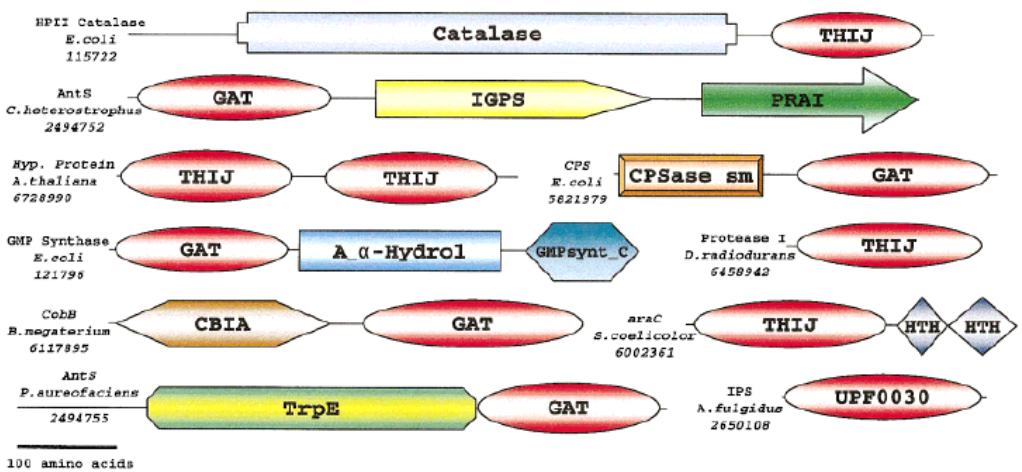
Horvath & Grishin, Proteins, 2001



1, Escherichia coli HPII catalase; 2, D. radiodurans HPII catalase; 3, Danio rerioES1; 4, Homo sapiensKNP-I a; 5, Rattus norvegicusSP22; 6, S. coelicolorAraC; 7, D. radioduransprotease I; 8, P. furiosusPfpI protease I; 9, S. coelicolor CobB; 10, Bacilis subtilisFGAM synthase I; 11, Mycobacterium leprae amidotransferase hisH; 12, A. fulgidus imidazole glycerol-phosphate synthase subunit H; 13, Saccharomyces cerevisiaeCTP synthetase; 14, Acinetobacter sp. aniline dioxygenase; 15, Arabidopsis thaliana glutamine amidotransferase; 16, A. thaliana g-glutamyl hydrolase precursor; 17, E. coli carbamoyl phosphate synthetase; 18, E. coliGMP synthase, P. aureofaciens AntS, S. sulfataricus AntS.

Arbre des GAT:





Organisation des domaines (de l'importance de séparer les domaines pour l'analyse)