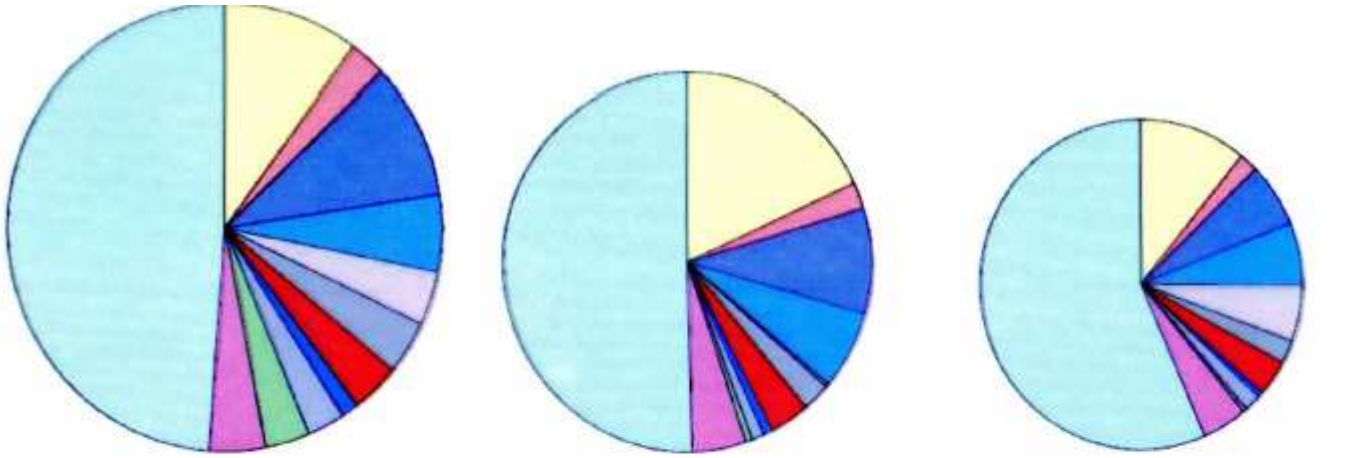
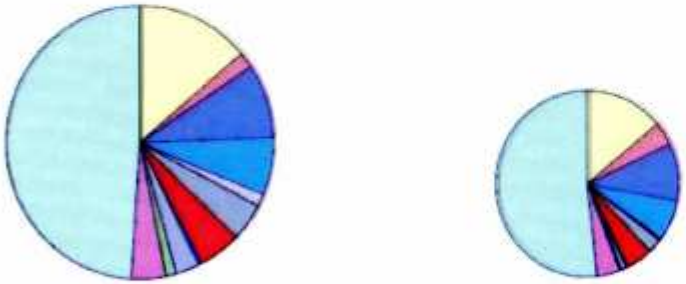
Bioinformatique M2: Lecture 1

P. Derreumaux

Comment déterminer la fonction d'une protéine



Organism	Human	*Arabidopsis* (plant)	*C. elegans* (roundworm)
Genes	~32,000	25,706	18,266

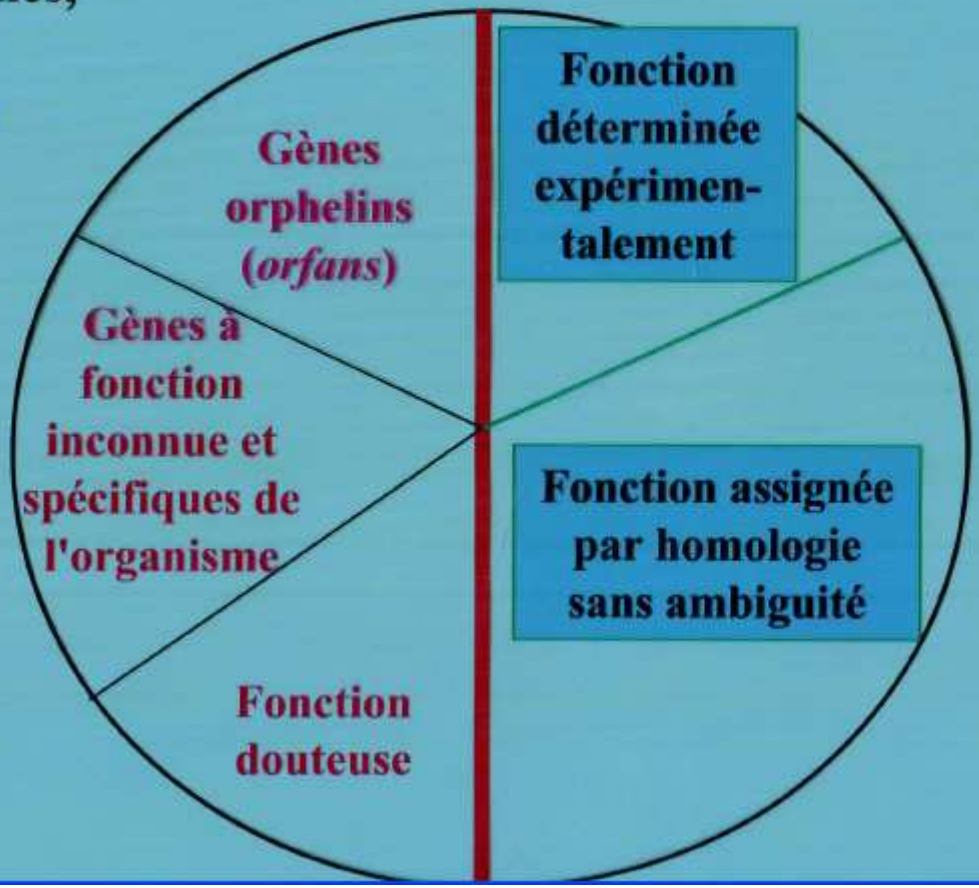


Organism	*Drosophila* (fly)	*Saccharomyces* (yeast)
Genes	13,338	~6000

- Metabolism
- DNA replication/modification
- Transcription/translation
- Intracellular signaling
- Cell-cell communication
- Protein folding and degradation
- Transport
- Multifunctional proteins
- Cytoskeleton/structure
- Defense and immunity
- Miscellaneous function
- Unknown

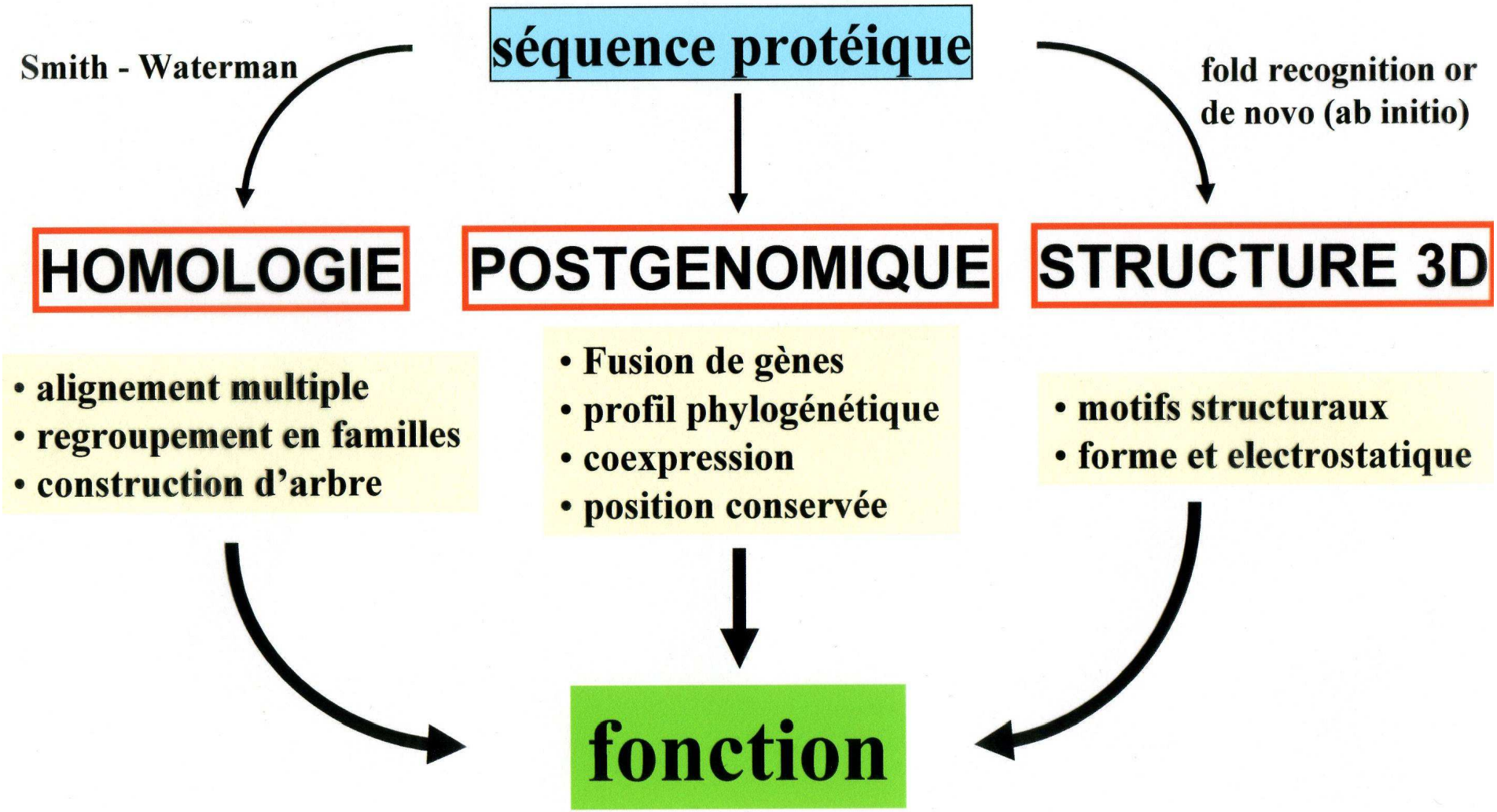
La surprise des gènes orphelins

Pour la majorité des génomes, on a une répartition 50/50 entre le connu et l'inconnu



Déterminer la fonction d'une protéine

Les différents méthodes



I. Homologie et alignement de séquences

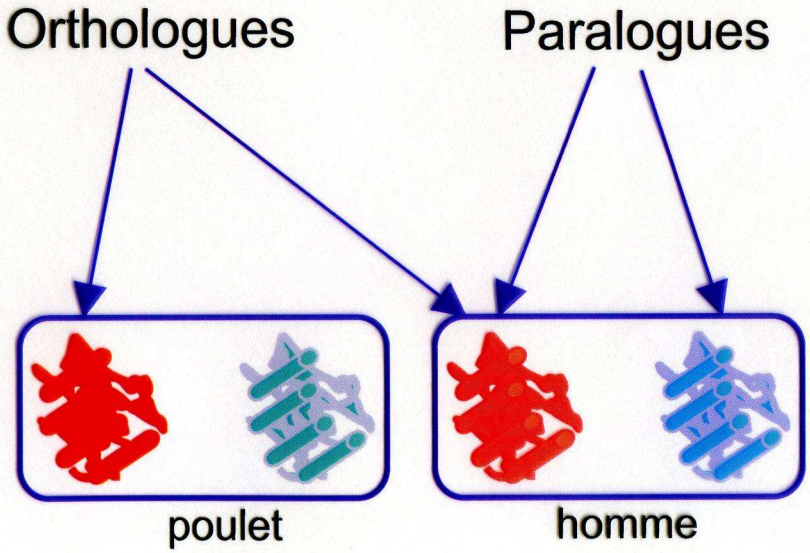
La déduction par homologie, ou le « dogme central » de la bioinformatique

★ 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 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)

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 d'un phénomène de spéciation
- ★ Transfert horizontal: par endosymbiontes, etc. Fitsch a aussi introduit "xénologue" pour évoquer ce cas.



Comment observer une homologie?

Plusieurs approches:

- La comparaison de séquences 2 à 2. C'est la méthode principale. La comparaison de séquence permet d'observer les régions **conservées**. On déduit l'homologie de la conservation.
- La recherche de domaines communs (utilisation des banques de domaines: Prodom, Blocks)
- La recherche de motifs communs (utilisation de la banque de motifs: Prosite)

Comparaison de deux séquences

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

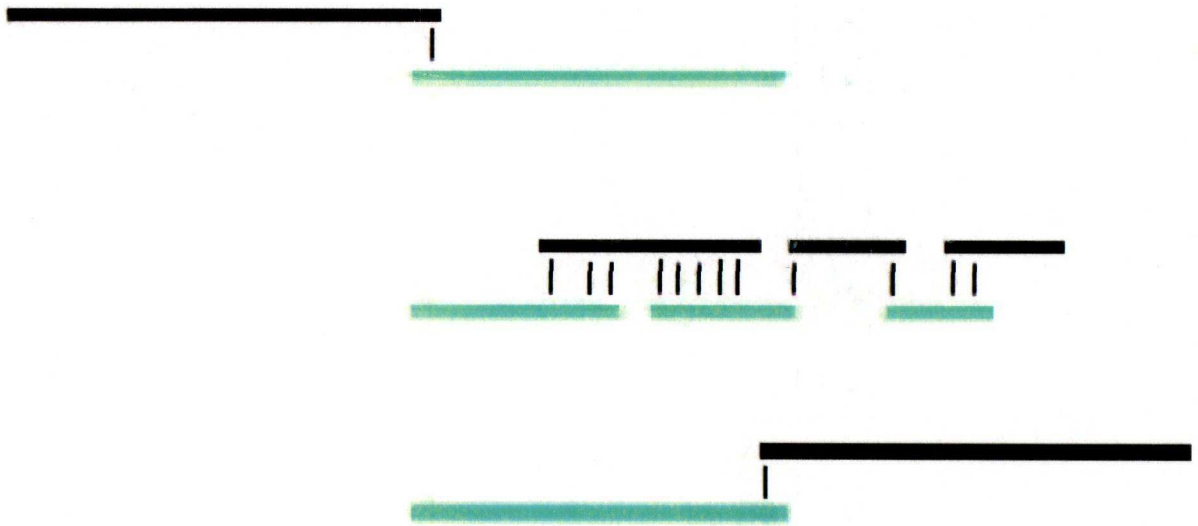
• substitutions (s) :

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

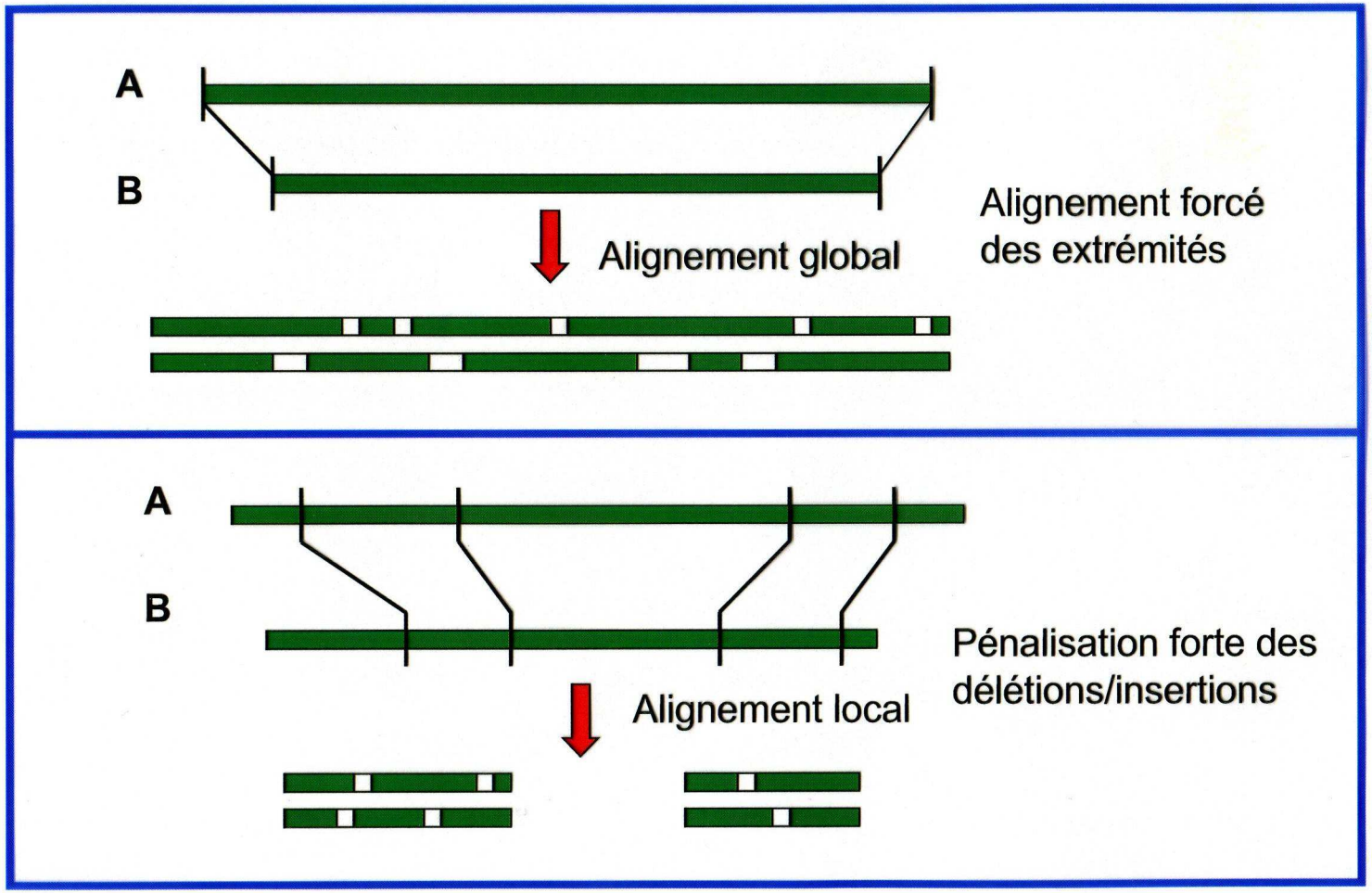
Définition du Score Total.

Why is alignment hard?

Number of ways to align 2 sequences



Alignements globaux *versus* locaux



here global alignment

The sequence alignment problem

```
1. THESESENTENSESALIGN--NICELY        2 (gap number)
   ||||| || |     ||||| ||||||      19
2. THESEQENCE----ALIGNEDNICELY         4

1. THESESENTENSESALIGN--NICELY        2
   |||||      || | ||||| ||||||      19
2. THESE-Q--ENCE-ALIGNEDNICLEY         4

1. THESESENTENSESALIGN--NICELY        2
   |||  ||  || | ||||| ||||||      19
2. THE--SEQ-ENCE-ALIGNEDNICLEY         4
```


→ matrices de scores + scores gaps.

Fig.3 PAM 250 Amino Acid Similarity Matrix

Cys	12																			
Gly	-3	5																		
Pro	-3	-1	6																	
Ser	0	1	1	1																
Ala	-2	1	1	1	2															
Thr	-2	0	0	1	1	3														
Asp	-5	1	-1	0	0	0	4													
Glu	-5	0	-1	0	0	0	3	4												
Asn	-4	0	-1	1	0	0	2	1	2											
Gln	-5	-1	0	-1	0	-1	2	2	1	4										
His	-3	-2	0	-1	-1	-1	1	1	2	3	6									
Lys	-5	-2	-1	0	-1	0	0	0	1	1	0	5								
Arg	-4	-3	0	0	-2	-1	-1	-1	0	1	2	3	6							
Val	-2	-1	-1	-1	0	0	-2	-2	-2	-2	-2	-2	-2	4						
Met	-5	-3	-2	-2	-1	-1	-3	-2	0	-1	-2	0	0	2	6					
Ile	-2	-3	-2	-1	-1	0	-2	-2	-2	-2	-2	-2	-2	4	2	5				
Leu	-6	-4	-3	-3	-2	-2	-4	-3	-3	-2	-2	-3	-3	2	4	2	6			
Phe	-4	-5	-5	-3	-4	-3	-6	-5	-4	-5	-2	-5	-4	-1	0	1	2	9		
Tyr	0	-5	-5	-3	-3	-3	-4	-4	-2	-4	0	-4	-5	-2	-2	-1	-1	7	10	
Trp	-8	-7	-6	-2	-6	-5	-7	-7	-4	-5	-3	-3	2	-6	-4	-5	-2	0	0	17
	Cys	Gly	Pro	Ser	Ala	Thr	Asp	Glu	Asn	Gln	His	Lys	Arg	Val	Met	Ile	Leu	Phe	Tyr	Trp

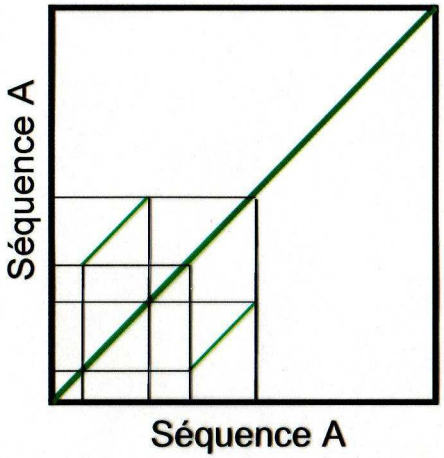
How similar are amino-acids?

- Use chemical properties
- Use mutation data.
- Use structures.

"Dotplot"

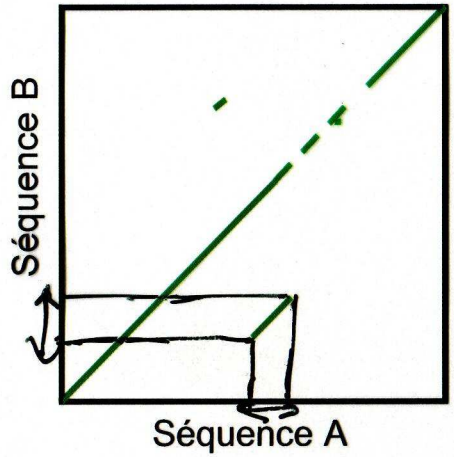
Auto-comparaison

A



Séquences très similaires

B

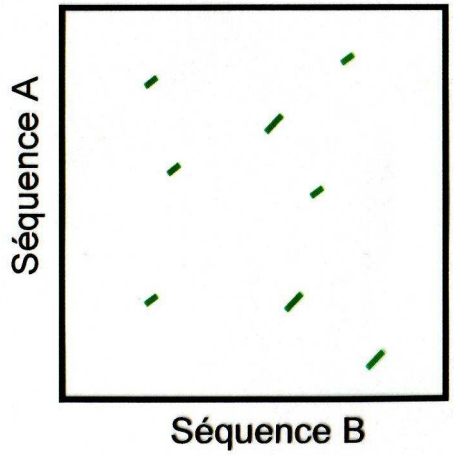


Séquences apparentées

C

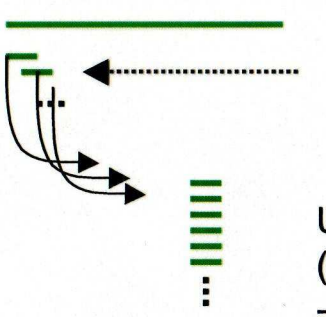
Pas de similitude

D



Algorithme BLAST (Altschul *et al.*, 1990)

(1) Recherche de la liste des mots de longueurs w à hauts scores

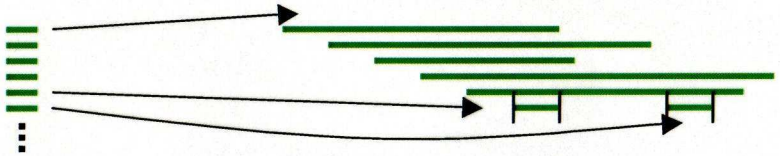


séquence source de longueur L

Au maximum L-W+1 mots (w~3 pour protéines)

Utilisation d'une matrice
(PAM ou BLOSSUM, etc.)
-> liste des mots qui ont un score ≥ T

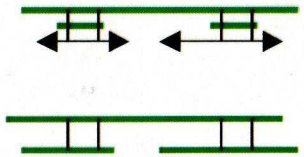
(2) Comparaison liste de mots / banque de données -> appariements exacts



Séquences
Banque de données

Appariements exacts

(3) Extension des appariements pour trouver les alignements de score ≥ seuil S



Paires de segments maximaux

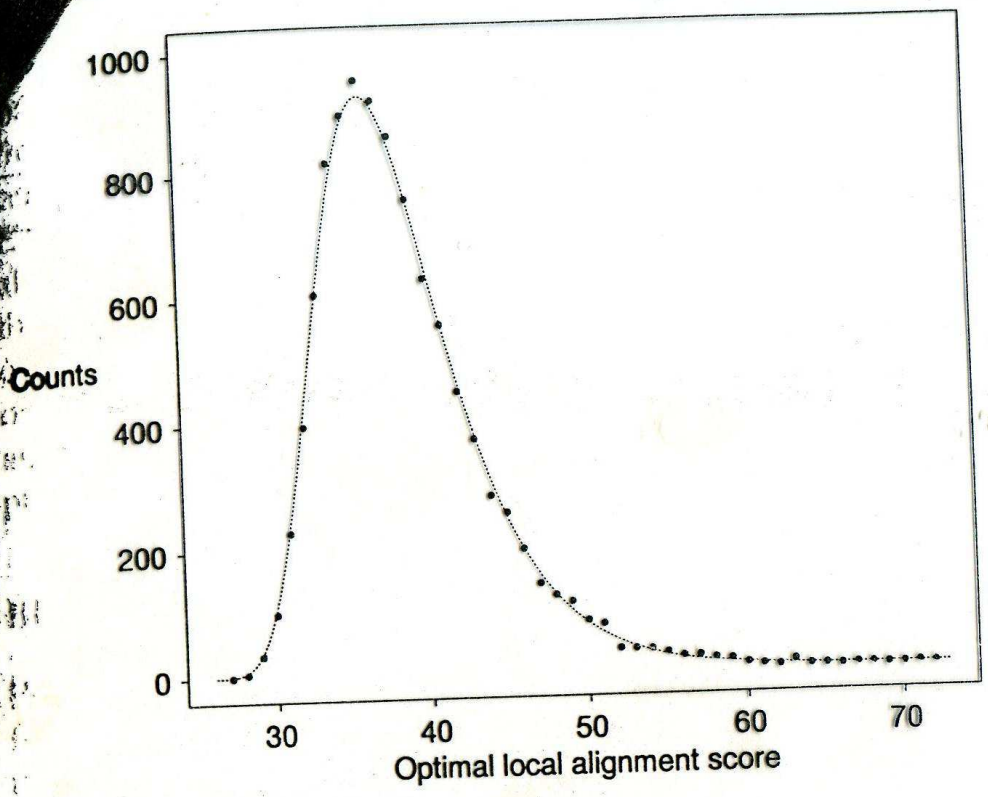


Figure 6. The distribution of optimal local alignment scores from the comparison of a position-specific score matrix with 10 000 random protein sequences. The score matrix was constructed by PSI-BLAST from the 128 local alignments with E-value ≤0.01 found in a search of SWISS-PROT using as query the length-567 influenza A virus hemagglutinin precursor (27) (SWISS-PROT accession no. P03435). The random sequences, each of length 567, were generated using the amino acid frequencies of Robinson and Robinson (20). Optimal local alignment scores were calculated using the position-specific matrix in conjunction with $10 + k$ gap costs. The extreme value distribution that best fits the data (3,15) is plotted. A χ^2 goodness-of-fit test with 34 degrees of freedom has value 41.8, corresponding to a P-value of 0.20.

Estimation statistique du score d'un alignement local.

- un score S est associé à E-value

$$\text{E-value}(S) = K m n e^{-\lambda S}$$

m, n longueur des séquences comparée.
K, λ = f(m, n, composition AA, matrice, gap)
calculés automatiquement.

- probabilité de trouver au hasard au moins un match avec score ≥ S :

$$P = 1 - e^{-\text{Evalue}}.$$

E-value :	P
10	0.999954.
1	0.63
0.0001	0.0001

Position Specific Scoring Matrices

Sequence Information

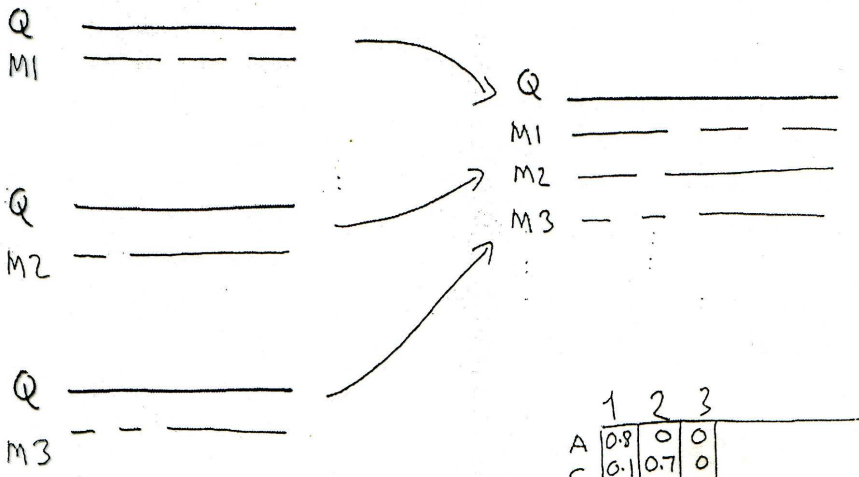
- Ligands to Calcium
- Hydrophobic patches that stabilize structure

```
MCEG     FKEAFSLFDKDGDGTITTKELGTVMRSL
MCDO     FKEAFSLFDKDGDGSITTKELGTVMRSL
MCSP     FKEAFSLFDKDGDGCITTKELGTVMRSL
MCCHM    IREAFRVFDKDGNGYISAAELRHVMTNL
MCUR1C   IKAIIQKADANKDGKIDREEFMKLIKS.
MCUR2C   IDAIIKKADGNNDGKIRVQEFVKMIESS
KLBOB    FNKAFELYDQDGDGYIDENELDALLKDL
KLCHI    FNKAFEMYDQDGNGYIDENELDALLKDL
KLBOI    LDELFEELDKNGDGEVSFEEFQVLVKKI
KLPGI    LDDLFQELDKNGNGEVSFEEFQVLVKKI
```

PSI-BLAST Sequence Matching

(1) Search query sequence against large database using scoring matrix (1st step) or profile (2nd, ...)
Keep significant hits

(2) Align them all to query



(3) Make a sequence profile

	1	2	3	...	103
A	0.8	0	0		0.1
C	0.1	0.7	0		0.1
⋮					
W	0	0.1	1.0		0.1
Y	0	0.2	0		0.1

Search with the profile (return to (1)) until the threshold value for inclusion in the position specific matrix is satisfied → (2nd E-value parameter)

S. Altschul et al.

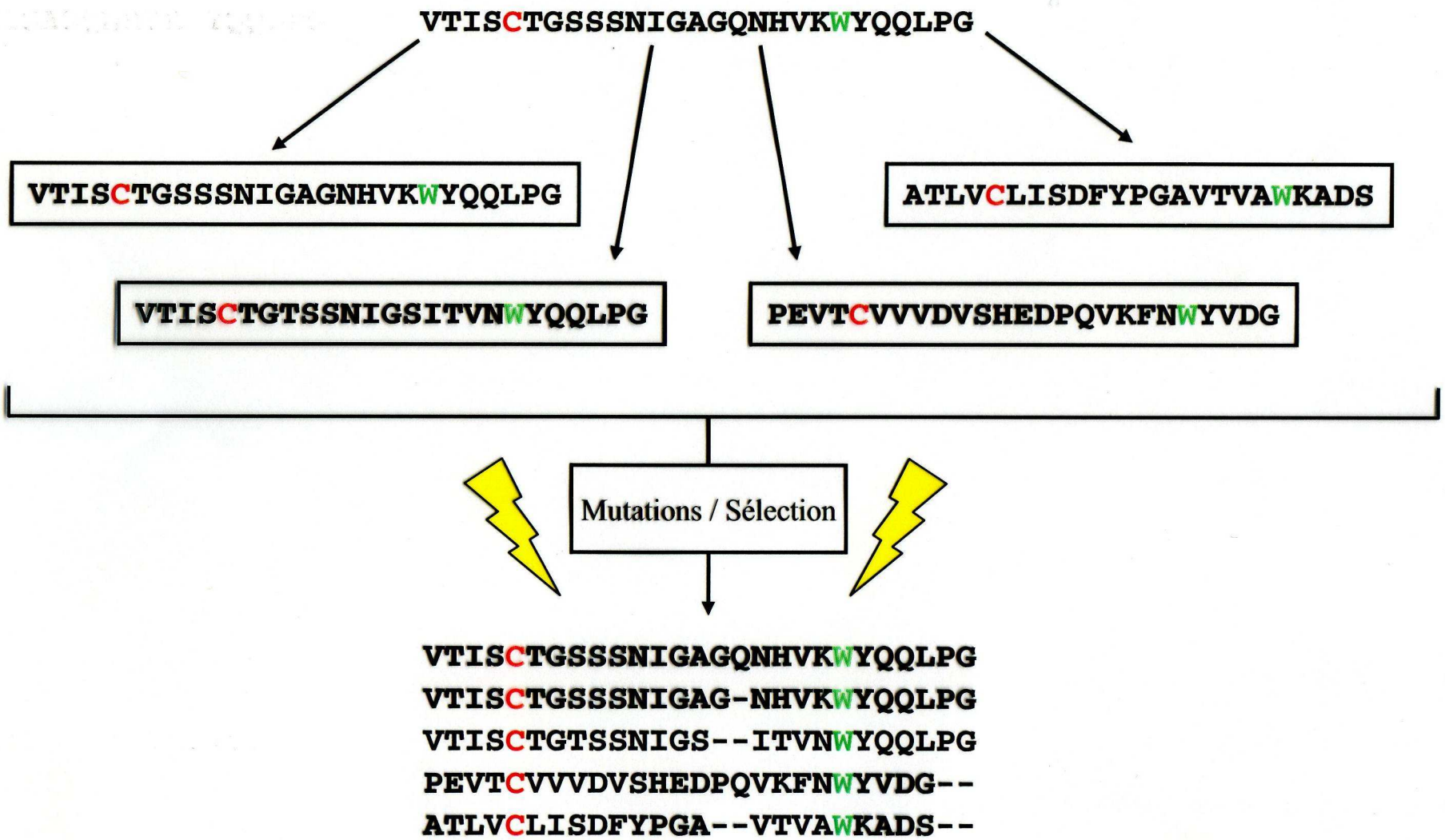
Nucleic Acids Research, 1997, Vol. 25, No. 17

3. The number of SWISS-PROT sequences yielding alignments with *E*-value ≤0.01, and relative running times, for Smith–Waterman and various version of BLAST

Protein family	Query	Smith–Waterman	Original BLAST	Gapped BLAST	PSI-BLAST
Serine protease	P00762	275	273	275	286
Serine protease inhibitor	P01008	108	105	108	111
Ras	P01111	255	249	252	375
Globin	P02232	28	26	28	623
Hemagglutinin	P03435	128	114	128	130
Interferon α	P05013	53	53	53	53
Alcohol dehydrogenase	P07327	138	128	137	160
Histocompatibility antigen	P10318	262	241	261	338
Cytochrome P450	P10635	211	197	211	224
Glutathione transferase	P14942	83	79	81	142
H^+-transporting ATP synthase	P20705	198	191	197	207
Normalized running time		36	1.0	0.34	0.87

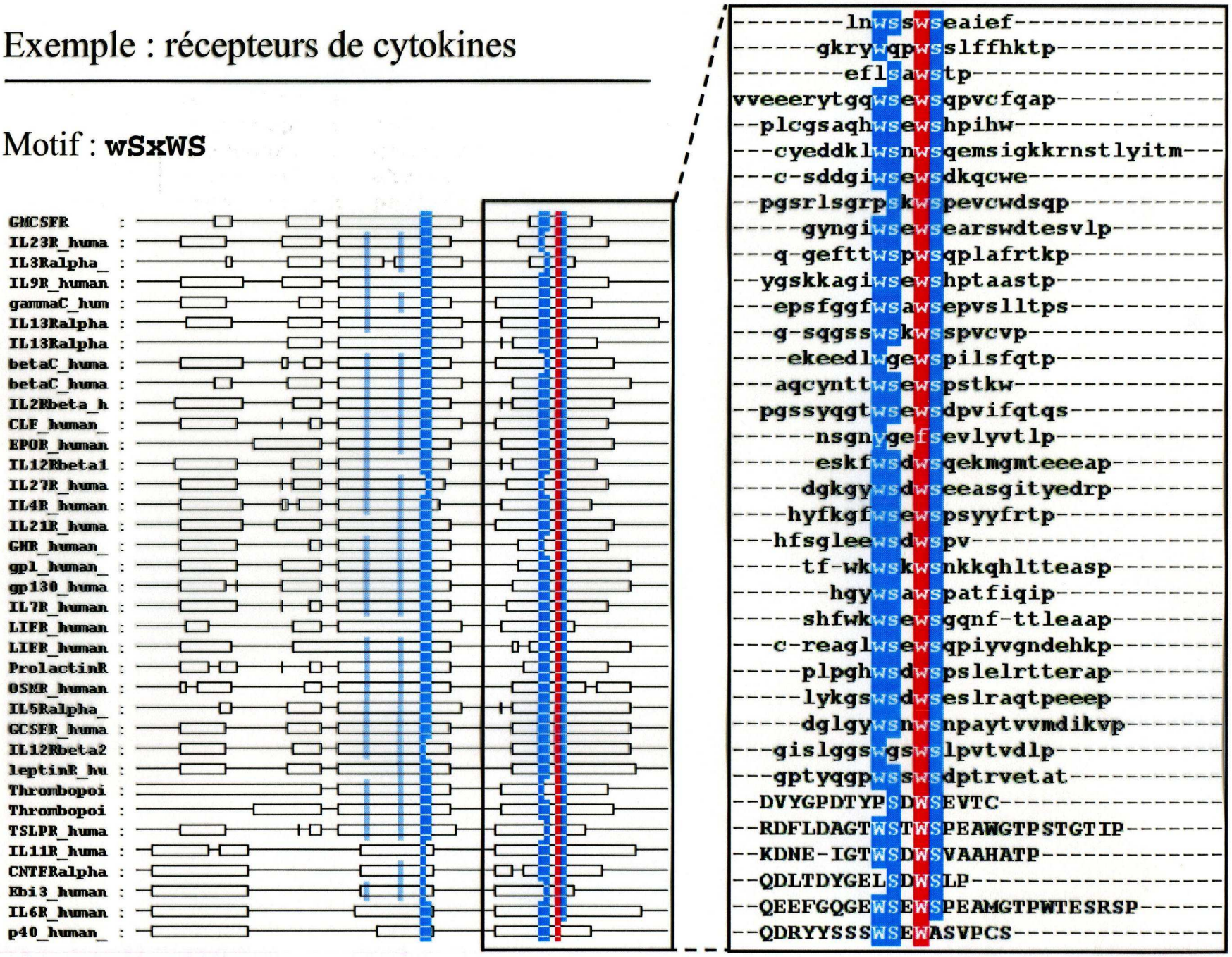
To score and evaluate the significance of the alignments found, the original BLAST program uses BLOSUM-62 substitution scores (18) and sum-statistics (21,22). The Smith–Waterman and gapped BLAST programs use BLOSUM-62 substitution scores, $10 + k$ gap costs, and the statistics of equations 1 and 2, in conjunction with the experimentally determined parameters $\lambda_g = 0.255$ and $K_g = 0.035$ (3). PSI-BLAST uses the same gap costs and λ_g, but applied to the position-specific score matrix constructed from the output of the gapped BLAST run. Only one PSI-BLAST iteration is executed. All three BLAST programs use the same parameter settings as in Table 2, except that T is set to 11. Normalized running times are the mean ratio of program running time to that for the original BLAST. The time for PSI-BLAST includes the time for the initial BLAST search.

Alignement multiple : une histoire



Exemple : récepteurs de cytokines

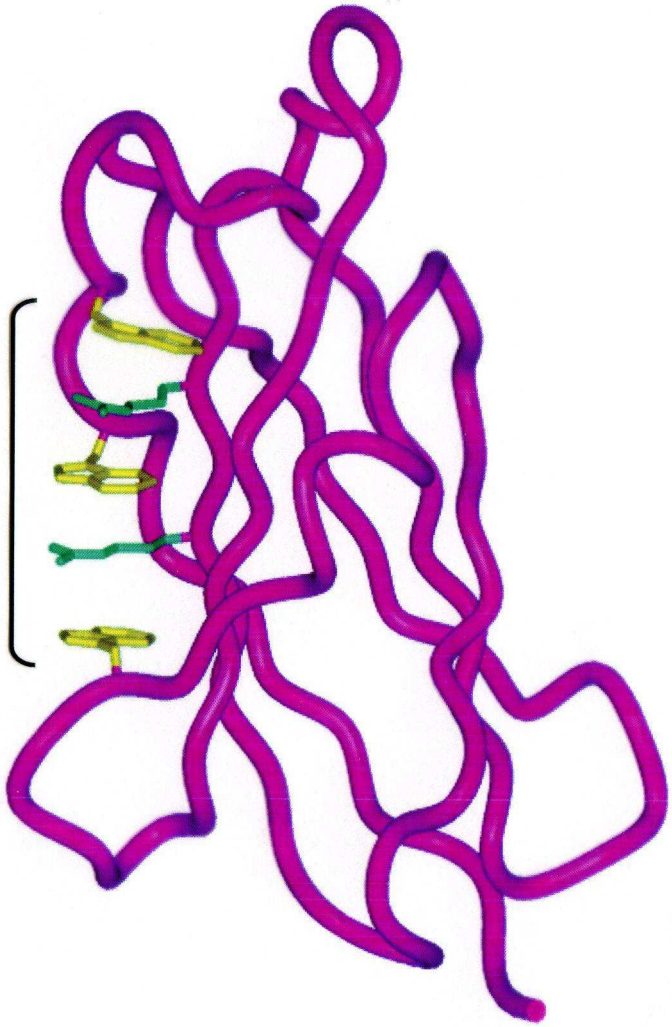
Motif : **wSxWS**



Exemple : récepteurs de cytokines

Motif : **wSxWS**

<u>Motif structural</u> : interactions entre résidus W et R



GLOBAL ALIGNMENT

CLUSTALW Program [Thompson, Higgins and Gibson, 1994]

CLUSTALW is one widely used implementation of profile-based progressive multiple alignment.

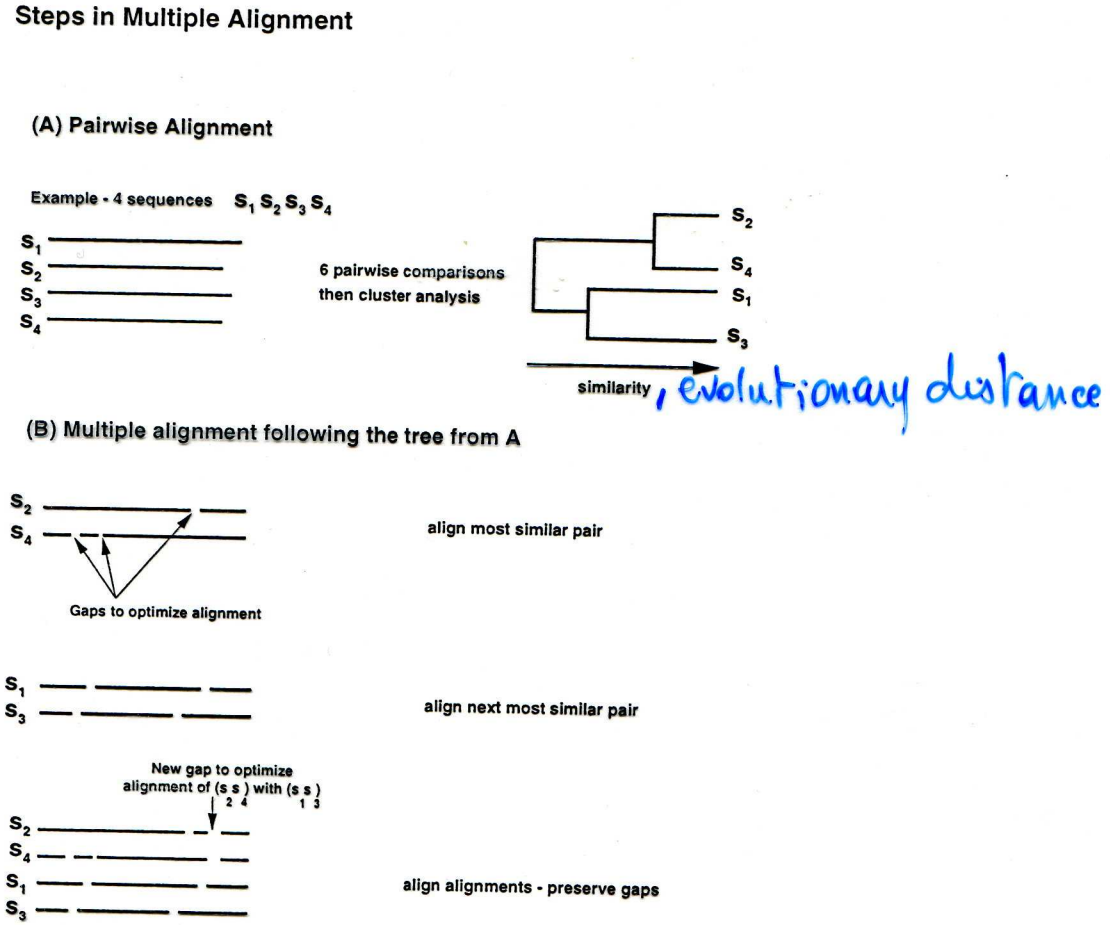
It is very similar to the Feng - Doolittle algorithm and it works as follows:

***1.** Construct a distance matrix of all N(N-1)/2 pairs of sequences by pairwise sequence alignment. Then convert the similarity scores to evolutionary distances using a specific model of evolution proposed by Kimura in 1983.*

***2.** Construct a guide-tree from this matrix using a clustering method called neighbor-joining proposed by Saitou and Nei in 1987.*

***3.** Progressively align nodes of the tree in order of decreasing similarity using sequences vs sequences, sequences vs profile and profile vs profile alignments.*

Figure 14: Progressive Alignment



Scoring along a tree is the main alternative to the simple "sum-of-pairs" cost model; only pairs of sequences that are adjacent (neighboring) in the tree are taken into consideration (or, at least they're weighted higher). Indeed, by weighting the pairs differently, we can score along a tree, yet employ Carrillo-Lipman and try out all possibly optimal alignment paths in the hyperlattice, see [AlL89] ! "Tree Alignment" subsumes methods that involve reconstructing ancestral sequences, too.

Problems with Progressive Alignments

(also used for genome alignment, e.g. programs MGA, MAUVE ...)

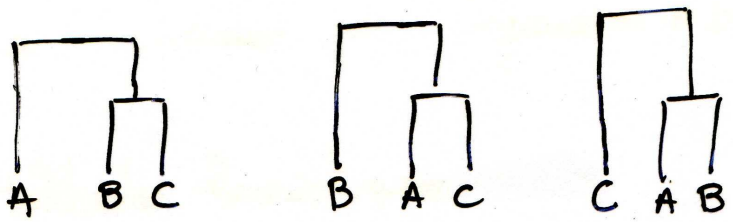
1. Local Minimum Problem

- It stems from greedy nature of alignment (mistakes made early in alignment cannot be corrected later) "Once a gap, always a gap"
- A better tree gives a better alignment (UPGMA<neighbour-joining tree method)

↳ Statistical methods for evaluating calculated trees. (BOOTSTRAP, JACK-KNIFE)

Possible ways of drawing a tree

- Only 3 trees possible for 3 species



- For 4 species → 15 different topologies

- number of possible trees for n OTU's

→ rooted trees = $\frac{(2n-3)!}{2^{n-2}(n-2)!}$

→ unrooted trees = $\frac{(2n-5)!}{2^{n-3}(n-3)!}$

ex 20 OTU → $8\ 10^{21}$ rooted trees
$2\ 10^{20}$ unrooted trees

Problem is a NP (non-polynomial) problem
∝ exponential in character

→ Exhaustive enumeration impossible

T-COFFEE: global vs. local

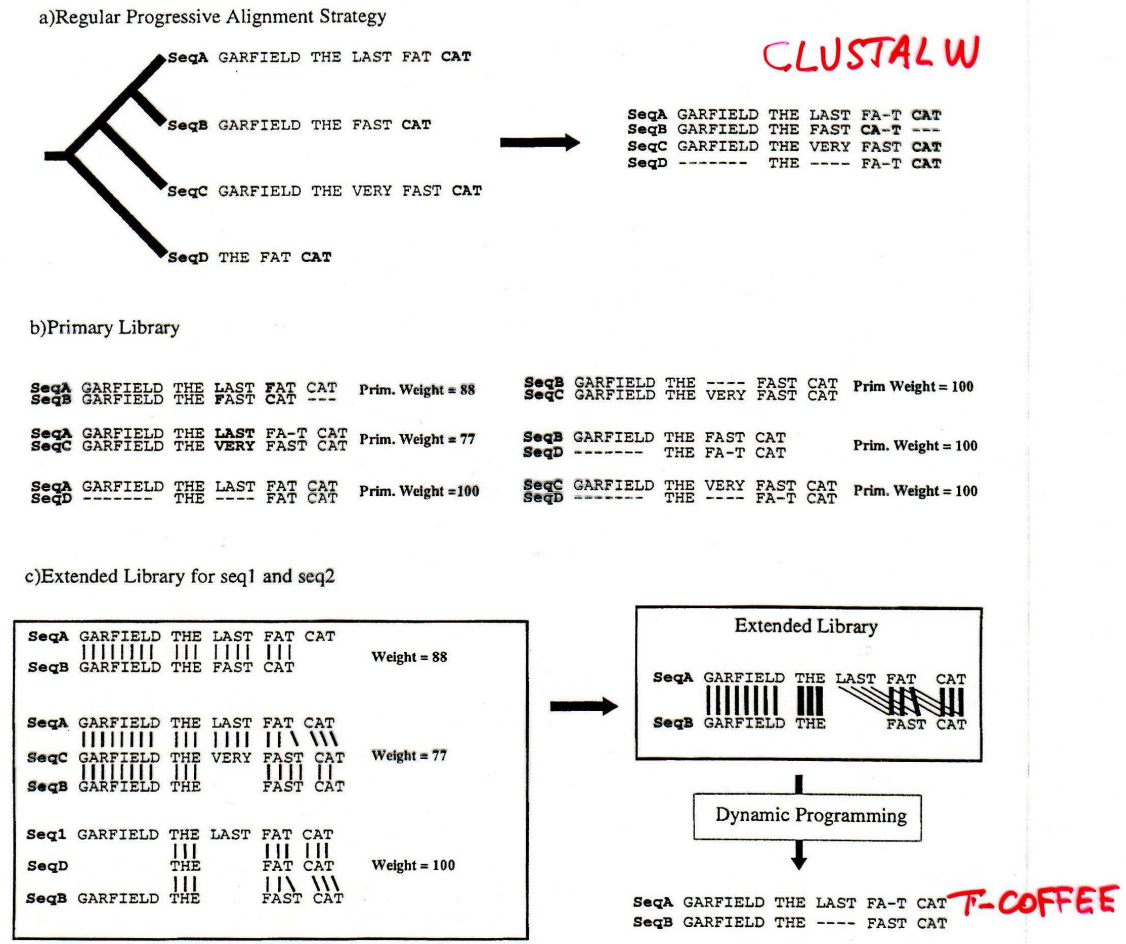


Figure 2. The library extension. (a) Progressive alignment. Four sequences have been designed. The tree indicates the order in which the sequences are aligned when using a progressive method such as ClustalW. The resulting alignment is shown, with the word CAT misaligned. (b) Primary library. Each pair of sequences is aligned using ClustalW. In these alignments, each pair of aligned residues is associated with a weight equal to the average identity among matched residues within the complete alignment (mismatches are indicated in bold type). (c) Library extension for a pair of sequences. The three possible alignments of sequence A and B are shown (A and B, A and B through C, A and B through D). These alignments are combined, as explained in the text, to produce the position-specific library. This library is resolved by dynamic programming to give the correct alignment. The thickness of the lines indicates the strength of the weight.