Alignment of biological sequences

INSA 2010-2011

Laurent Duret

Laboratoire de Biométrie et Biologie Evolutive (CNRS, INRIA), Université de Lyon

http://pbil.univ-lyon1.fr/alignment.html

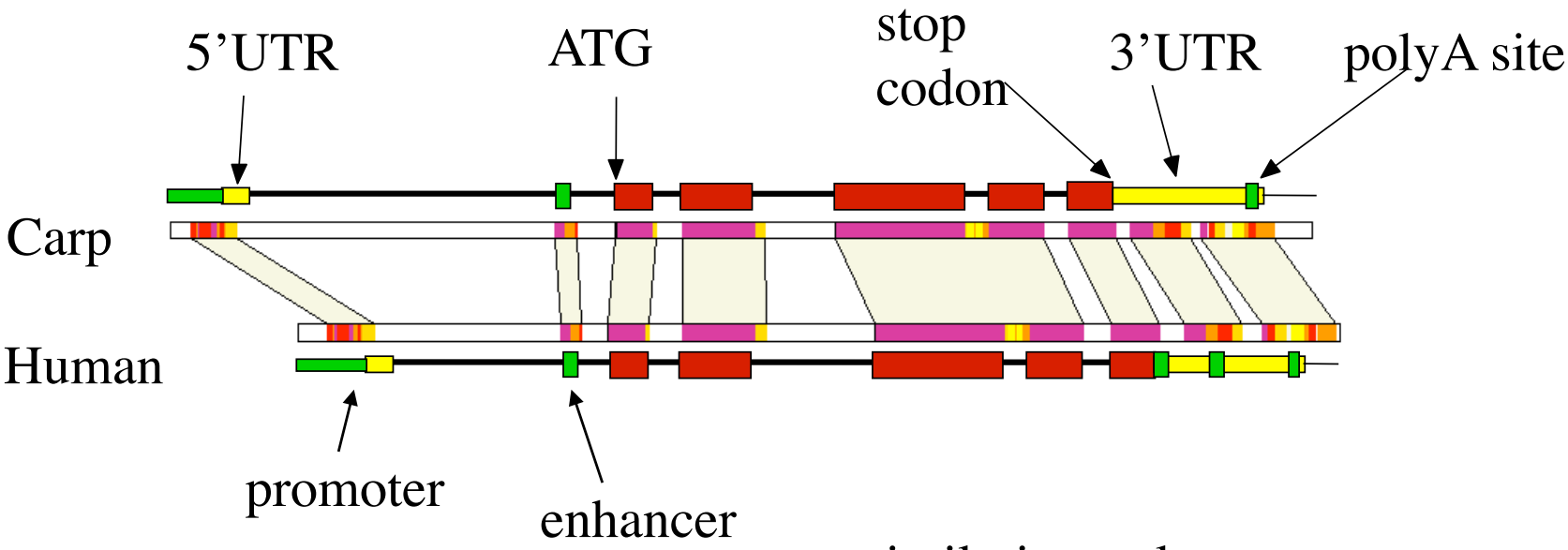
Sequence alignment

- Objectives
- General concepts
- Pairwise sequence alignment
- Database similarity search
  - standard (BLAST)
  - advanced (profile, PSI-BLAST)
- Multiple sequence alignment

Objectives

- Alignments allow the **comparison** of biological sequences. such comparisons are necessary for different studies :
  - Identification of homologous genes
  - Search for **functional constraints** in a set of genes or proteins.

Comparative analysis of human and carp β-actin genes



introns:

coding regions :

regulatory elements

similarity scale :

no significant similarity

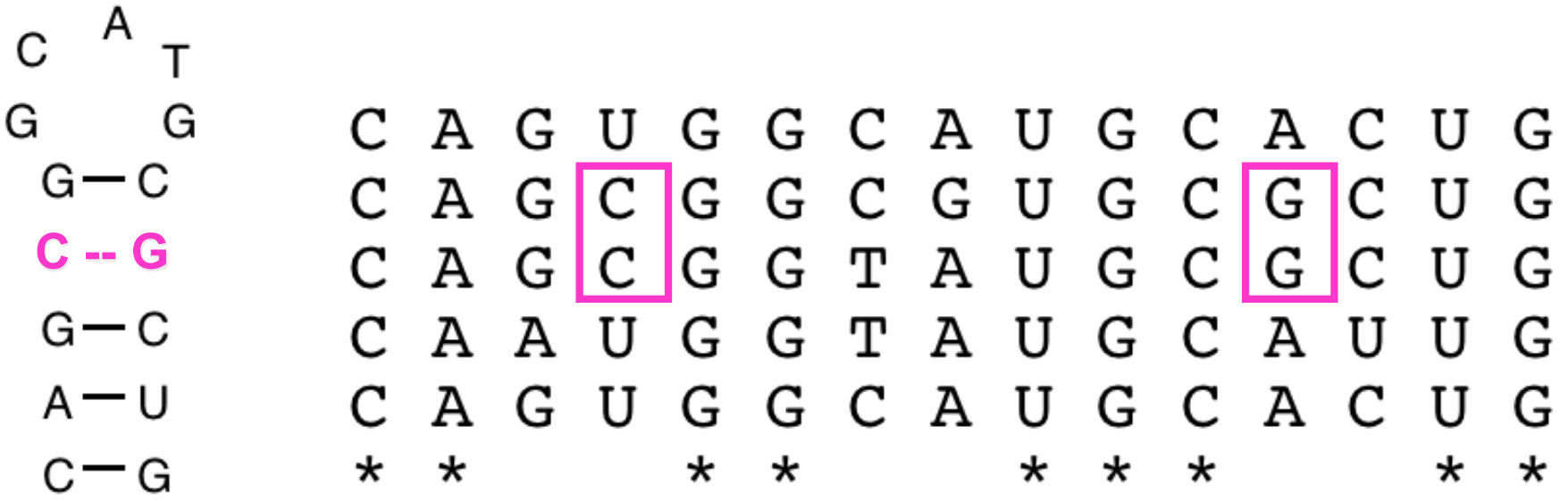
80 - 90% identity

70 - 80% identity

Objectives

- Alignments allow the **comparison** of biological sequences. such comparisons are necessary for different studies :
  - Identification of homologous genes
  - Search for **functional constraints** in a set of genes or proteins.
  - Function prediction
  - Structure prediction

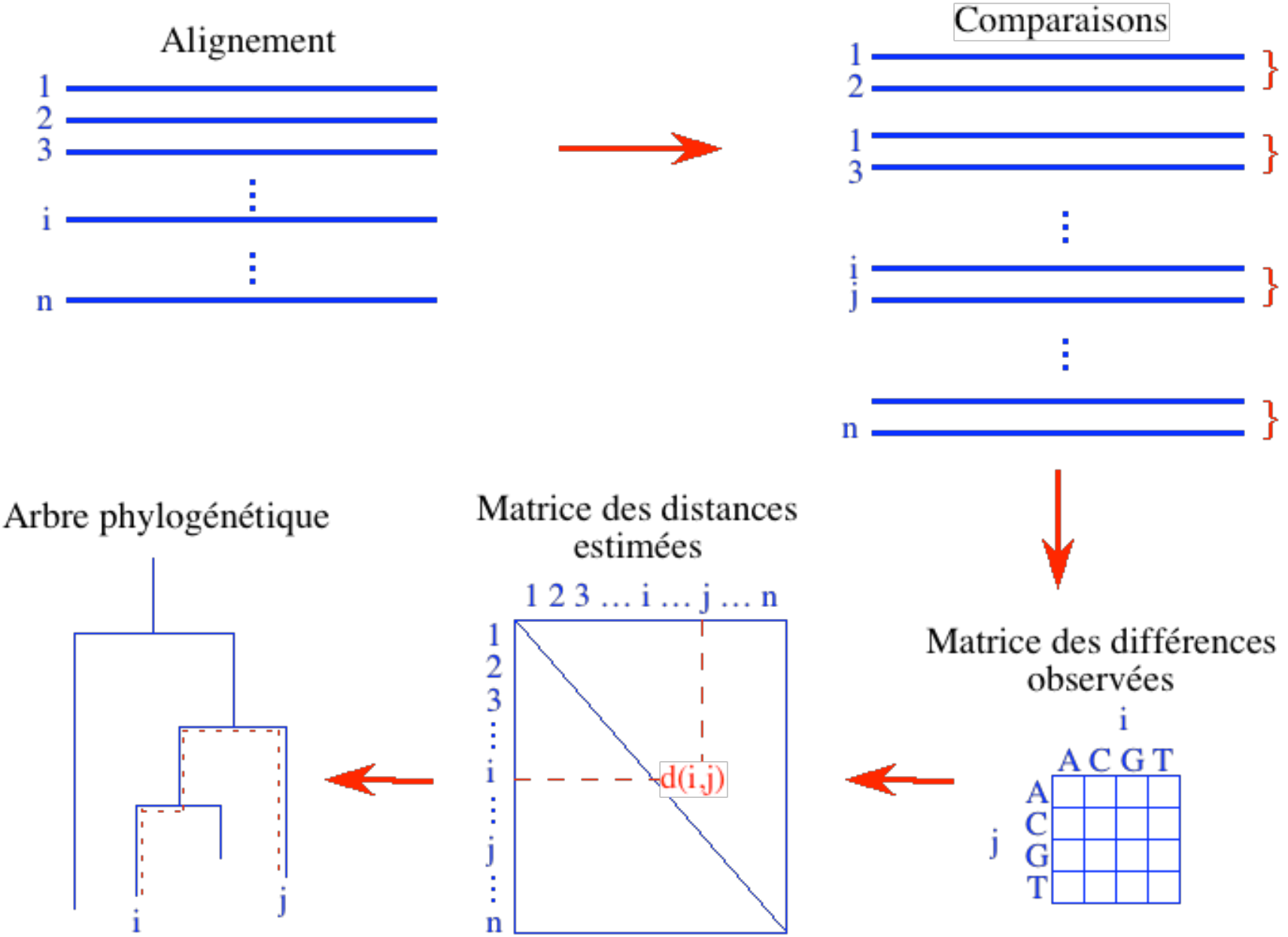
Prediction of RNA structure



Objectives

- Alignments allow the **comparison** of biological sequences. such comparisons are necessary for different studies :
  - Identification of homologous genes
  - Search for **functional constraints** in a set of genes or proteins.
  - Function prediction
  - Structure prediction
  - Reconstruct **evolutionary relationships** between sequences (phylogeny)

Molecular Phylogeny



Objectives

- Alignments allow the **comparison** of biological sequences. such comparisons are necessary for different studies :
  - Identification of homologous genes
  - Search for **functional constraints** in a set of genes or proteins.
  - Function prediction
  - Structure prediction
  - Reconstruct **evolutionary relationships** between sequences (phylogeny)
  - Design of PCR primers
  - Sequence assembly
  - ...

Alignment: representation

- Residues (nucleotides, amino-acids) are superposed so that to maximise the similarity between sequences.

```
G T T A A G G C G - G G A A A
G T T - - - G C G A G G A C A
* * *       * * *   * * *   *
```

- Mutations :
  - Substitution (*mismatch*)
  - Insertion
  - Délétion
- Insertions or deletions : indels (***gap***).

Which one is the good alignment ?

```
G T T A C G A        G T T A C G A
G T T - G G A        G T T G - G A
* * *     * *        * * *     * *
```

OR

```
G T T A C - G A
G T T - - G G A
* * *       * *
```

- For the biologist, the good alignment is the one that corresponds to the most likely evolutionary process

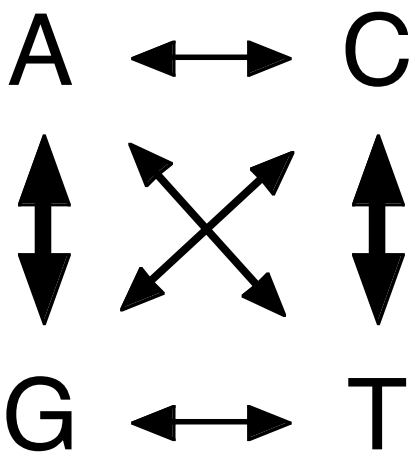
How do we measure sequence similarity ?

```
G T T A A G G C G - G G A A A
G T T - - - G C G A G G A C A
* * *       * * *   * * *   *
```

- $\text{Score} = \sum_{begin}^{end} SubstitutionWeight - \sum_{begin}^{end} GapPenalty$

- Example:
  - identity = 1
  - mismatch = 0
  - gap = -1
  - Score = 10 - 4 = 6

Models of evolution (DNA)



- Transition: A <-> G T <-> C
- Transversions : other substitutions
- p(transition) > p(transversion)

```
G T T A C G A        G T T A C G A
G T T - G G A        G T T G - G A
* * *     * *        * * * .   * *
```

Substitution Matrix (DNA)

	A	C	G	T
A	1	0	0.5	0
C	0	1	0	0.5
G	0.5	0	1	0
T	0	0.5	0	1

Examples :

$\delta(A,A) = 1$

$\delta(A,C) = 0$

$\delta(C,T) = 0.5$

- Gap = -1

```
G  T  T  A  C  G  A
G  T  T  -  G  G  A
1  1  1 -1  0  1  1
```

score = 4

```
G  T  T  A  C  G  A
G  T  T  G  -  G  A
1  1  1 .5 -1  1  1
```

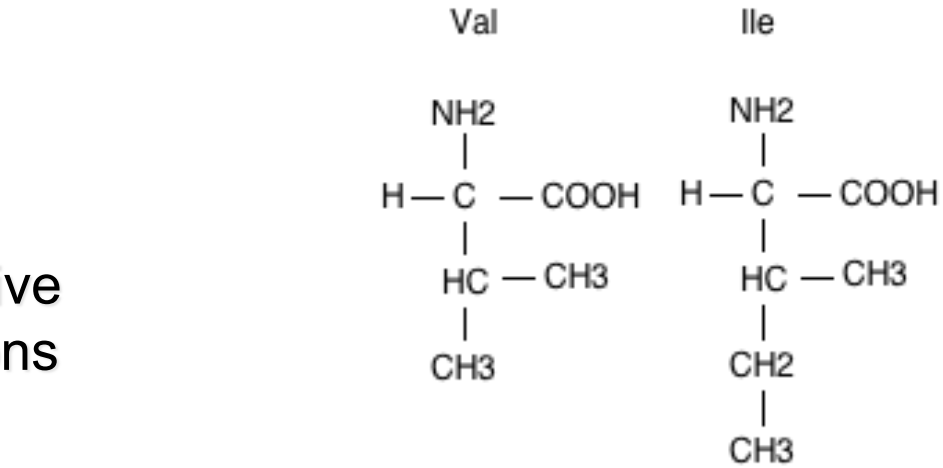
score = 4.5

Models of evolution (proteins)

- Genetic code
  - Asp (GAC, GAU) -> Tyr (UAC, UAU) : 1 mutation
  - Asp (GAC, GAU) -> Cys (UGC, UGU) : 2 mutations
  - Asp (GAC, GAU) -> Trp (UGG) : 3 mutations

- Physico-chemical properties of amino-acids (acidity, hydrophobicity, etc.)

conservative substitutions



Substitution matrix

- Dayhoff (PAM), BLOSUM: measure the frequency of substitutions in alignments of homologous proteins
  - PAM 60, PAM 120, PAM 250 (extrapolations from PAM 15)
  - BLOSUM 80, BLOSUM 62, BLOSUM 40 (based on blocks alignments)

	D	E	F	G	...
D	4	4	-6	1	...
E	4	4	-6	1	...
F	-6	-6	13	-6	...
G	1	1	-6	5	...
...	...	...	...	...	...

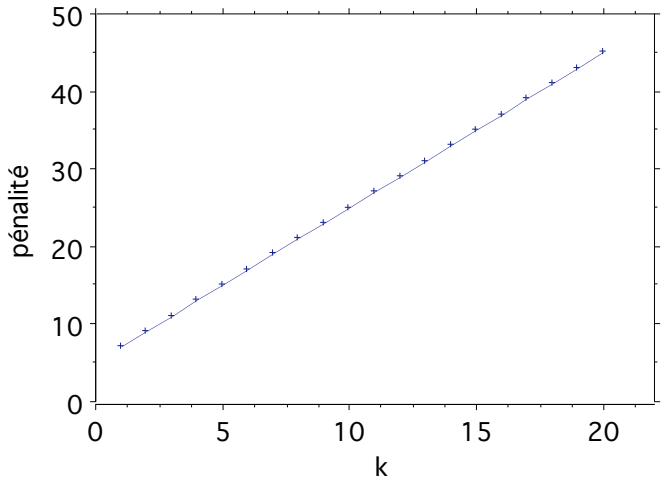
Weighting of gaps

```
TGATATCGCCA          TGATATCGCCA
TGAT---TCCA          TGAT-T--CCA
****    ***          **** *  ***
```

- Gap of length k: Linear penalties: $w = \delta_o + \delta_e k$

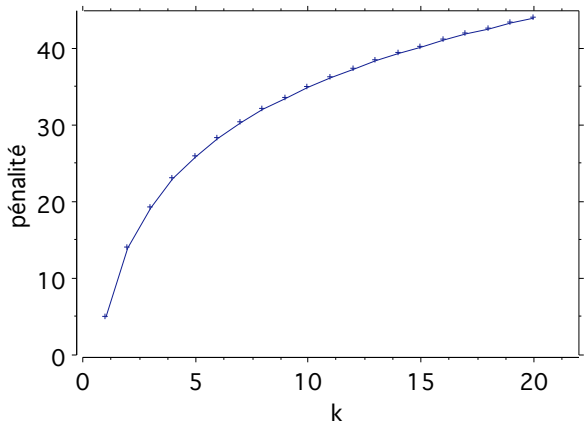
  δ_o : penalty for gap opening

  δ_e : penalty for gap extension



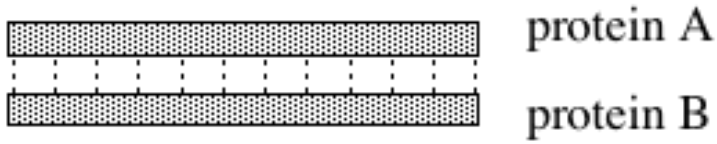
Weighting of gaps (more realistic)

- Estimation of parameters with true alignments (e.g. based on known structures)
- Gap of length *k*:
  - Logarithmic penalty: $w = \delta_o + \delta_e \log(k)$

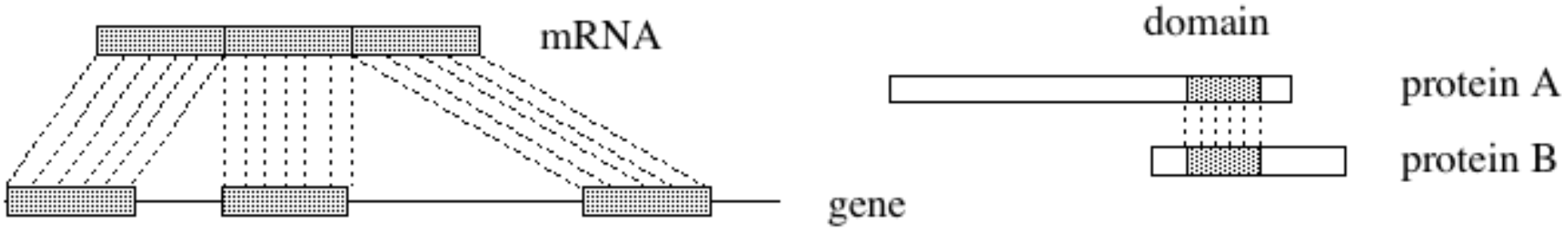


  - w = f(*log(k), log(PAM), residue, structure*)
    - *PAM: the probability of a gap increases with the evolutionary distance*
    - *Resides, structure: the probability of a gap is higher in a loop (hydrophilic) than in the hydrophobic core of proteins*

Similarity: global, local



global similarity



local similarity

Similarity, homology

- *Two sequences are homologous if (and only if) they derive from a common ancestor*

- *30% identity between two proteins => homology, except if:*
  - *Short block of similarity (< 100 aa)*
  - *Compositional biais (low-complexity regions, e.g. Pro-rich, Ala-rich regions)*

The number of alignments

```
AT      A-T     AT-     -AT     -AT     --AT   ...
AC      AC-     A-C     AC-     A-C     AC--   ...
```

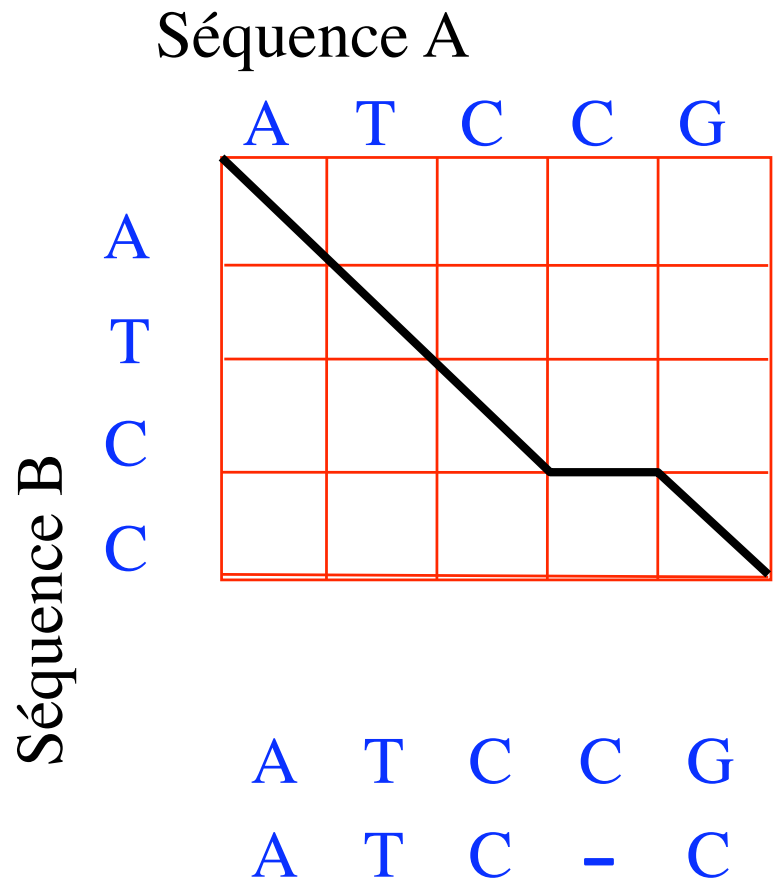
- Objective: for a given scoring scheme, find the best alignment(s), i.e. the optimal alignment(s)

- Problem: the number of possible alignments between two sequences increases exponentially with the length of sequences. For example:
  - Lg1 = 3, Lg2 = 4 : 129 possible alignments
  - Lg1 = 6, Lg2 = 8 : 40,081 possible alignments

Algorithms for aligning two sequences

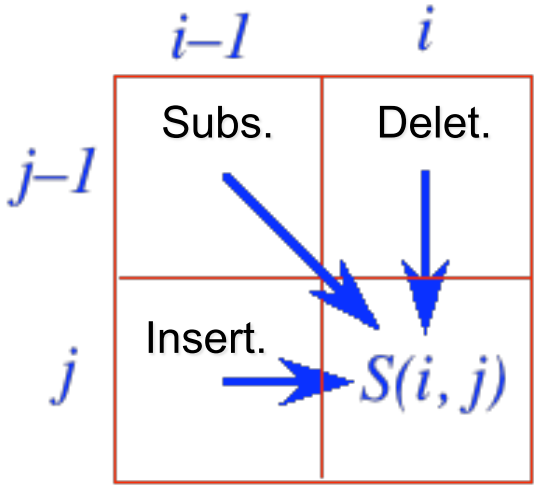
- Dynamic programming
  - Global alignment : Needleman & Wunsh
  - Local alignment : Smith & Waterman

Alignment representation: a path in a matrix



Recursive computation of the matrix

- Needleman & Wunsh, 1970



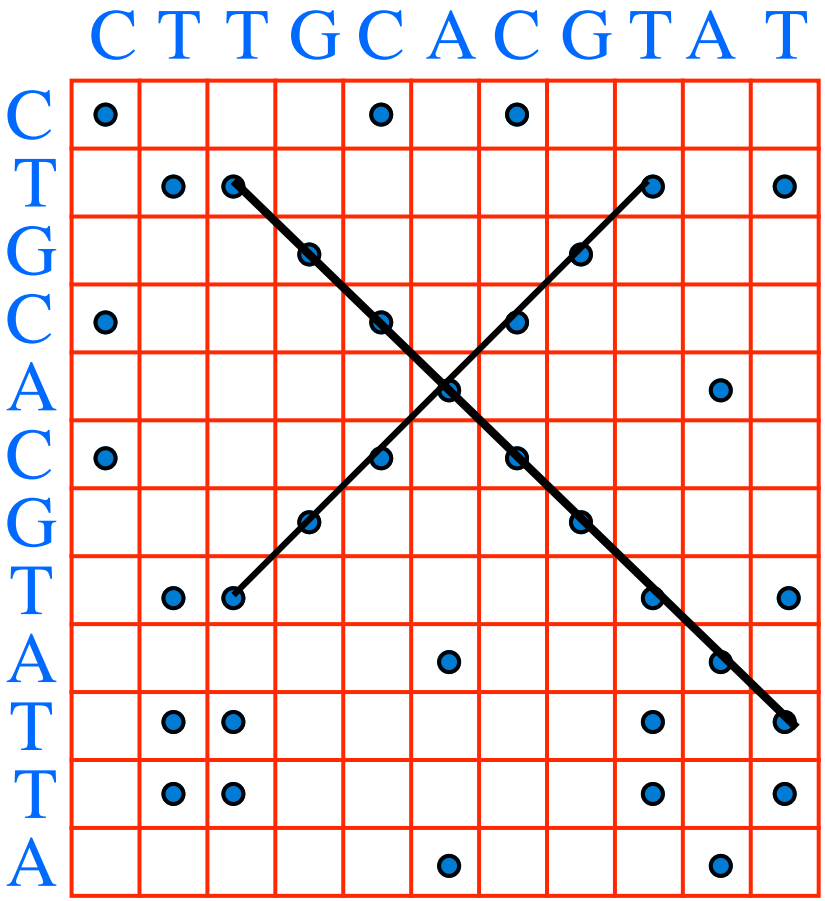
$$S(i,j) = \max \begin{bmatrix} S(i-1,j) + \delta(gap), \\ S(i-1,j-1) + \delta(a_i, b_j), \\ S(i,j-1) + \delta(gap) \end{bmatrix}$$

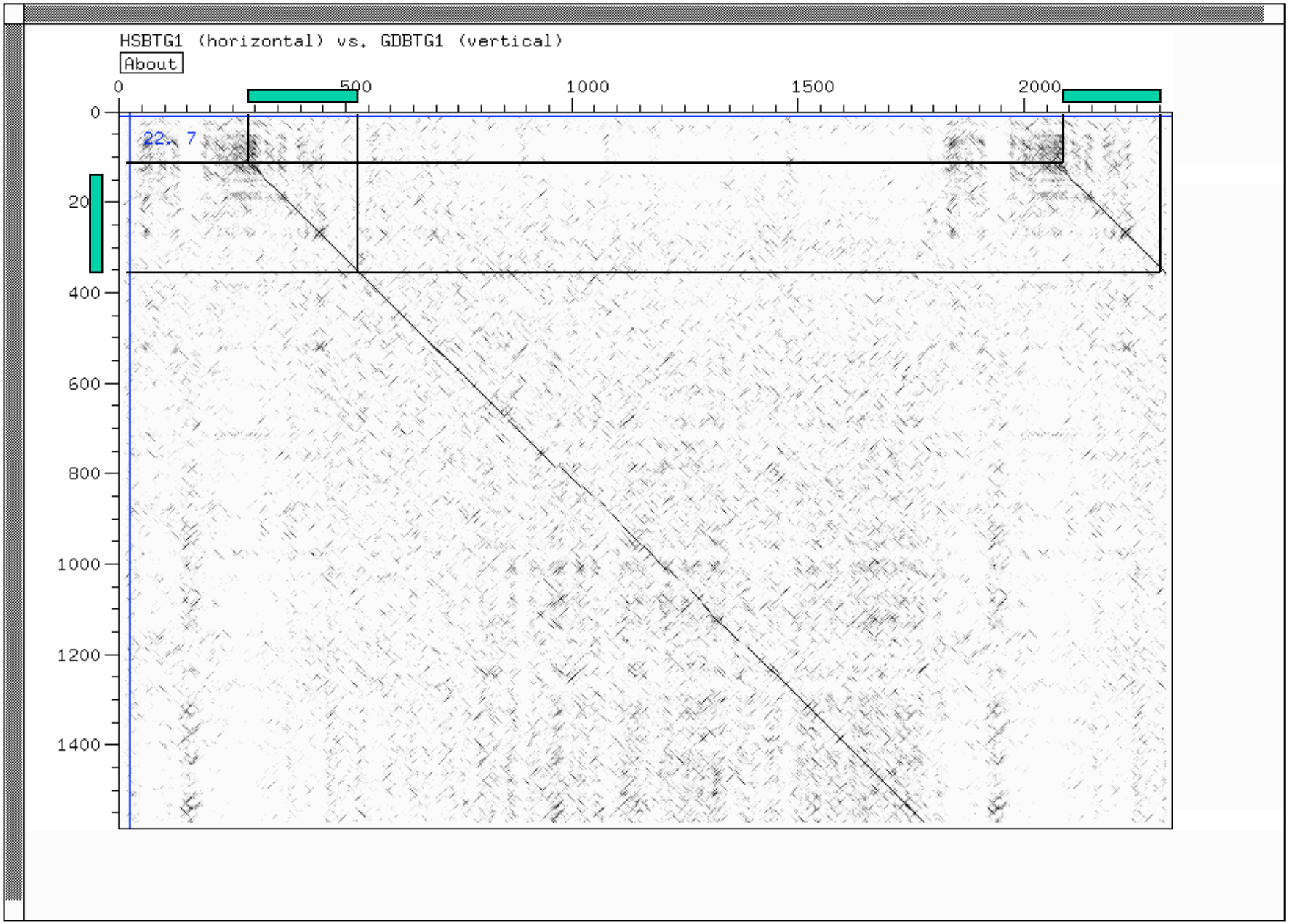
Dynamic programming: time and memory requirements

- Alignment of two sequences of length M and N
- Needleman-Wunsh (global alignments), Smith-Waterman (local alignments):
  - Time: $O(N \,.\, M)$
  - Memory: $O(N \,.\, M)$
- Improvement of Smith-Waterman (Huang & Miller 1991):
  - Time: $O(N \,.\, M)$
  - Memory: $O(N + M)$
- SIM, LALIGN

Dot Plot

- Graphical representation of similarities between two sequences
- Inversion, duplications





DOTTER: http://www.sanger.ac.uk/Software/Dotter/

Searching for similarities in sequence databases

- Objective: compare one sequence to a database of sequences, compare two databases, ...
- e.g. :
  - I have identified a new gene; does this gene have any homologue (known or unknown) in sequence databases ?
  - I want to identify all the genes that belong to a same gene family
  - I want to identify all homologous genes between the genomes of species A and species B

Searching for similarities in sequence databases

- Exact Algorithms (Smith-Waterman)
  - SIM, LALIGN, SSEARCH, …
- Heuristics
  - FASTA

    1 -search for identical ‘ k-tuples ’

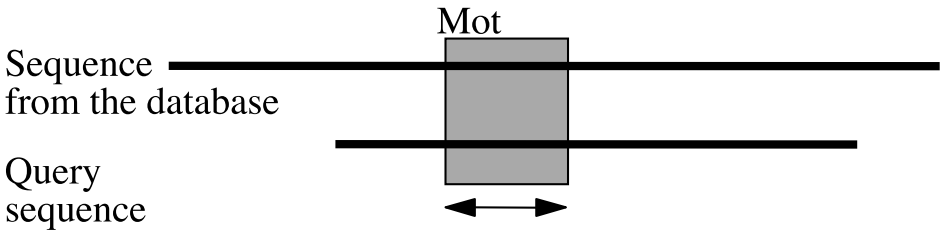
    2 - global alignment, anchored on the region of similarity
  - BLAST

    1 - search for similar ‘words’

    2 - extend blocks of similarity

BLAST

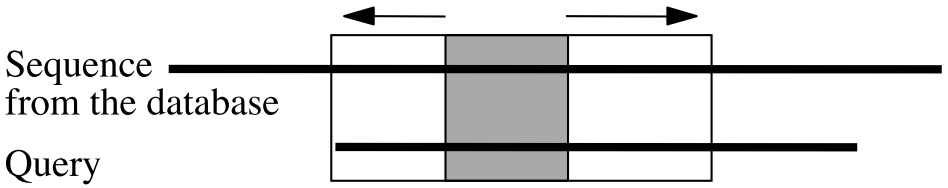
Step 1: detect similar ‘words’



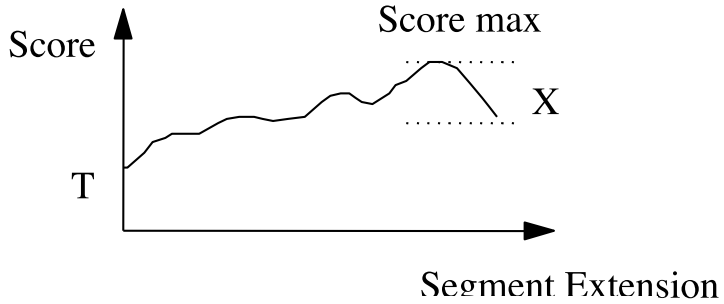
Word length = W

Score ≥ T (Threshold score)

Step 2: extend blocks of similarity



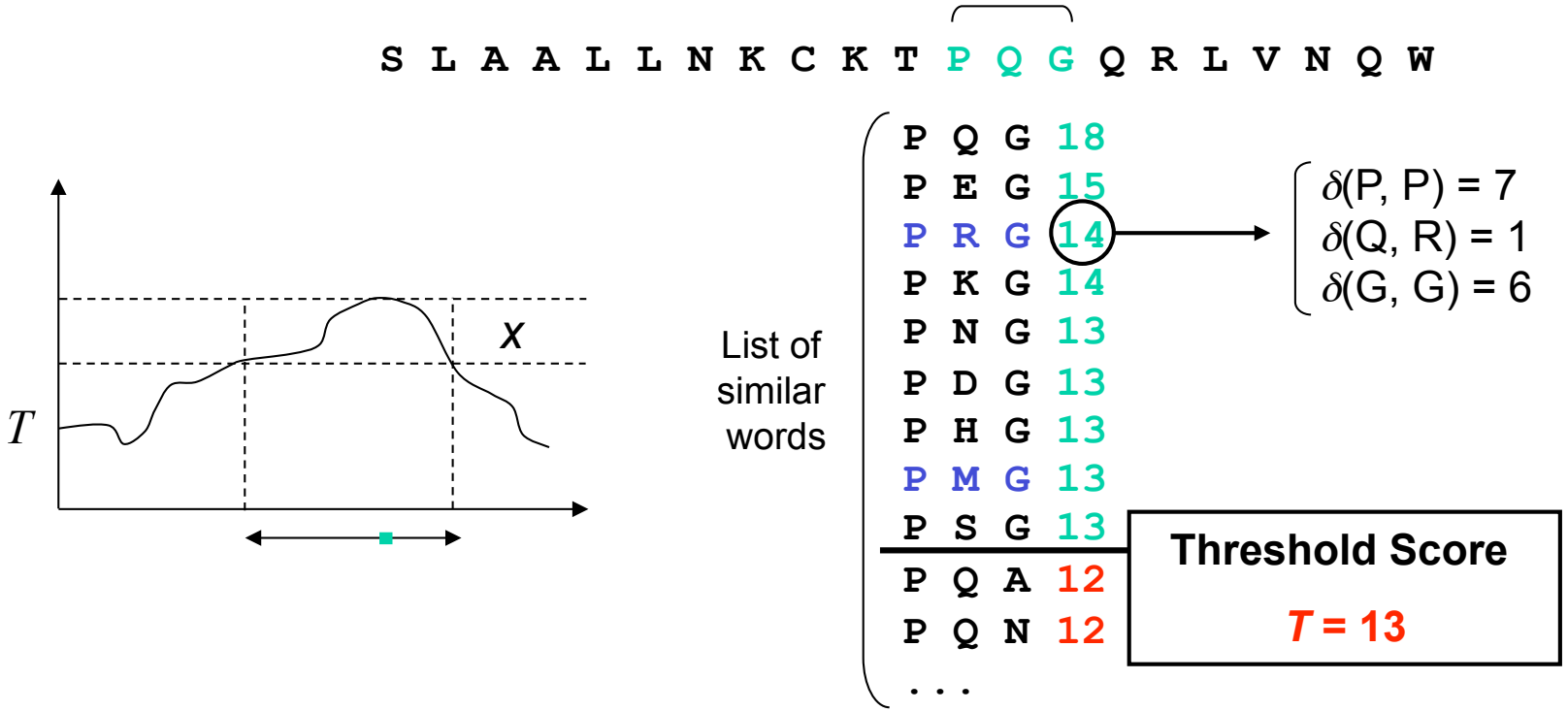
HSP: high score segment pair)



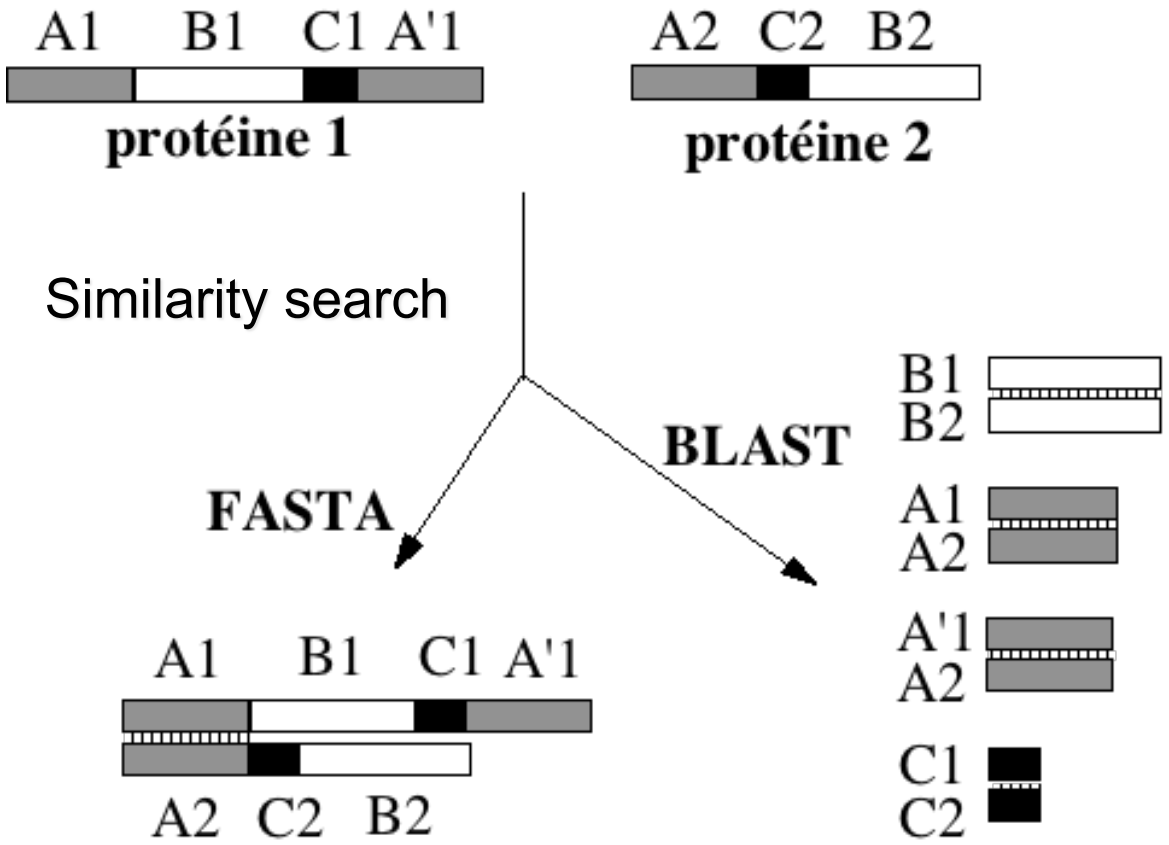
Stop extension if:

- reach the end of a sequence
- score ≤ 0
- score ≤ score_max - X

Example

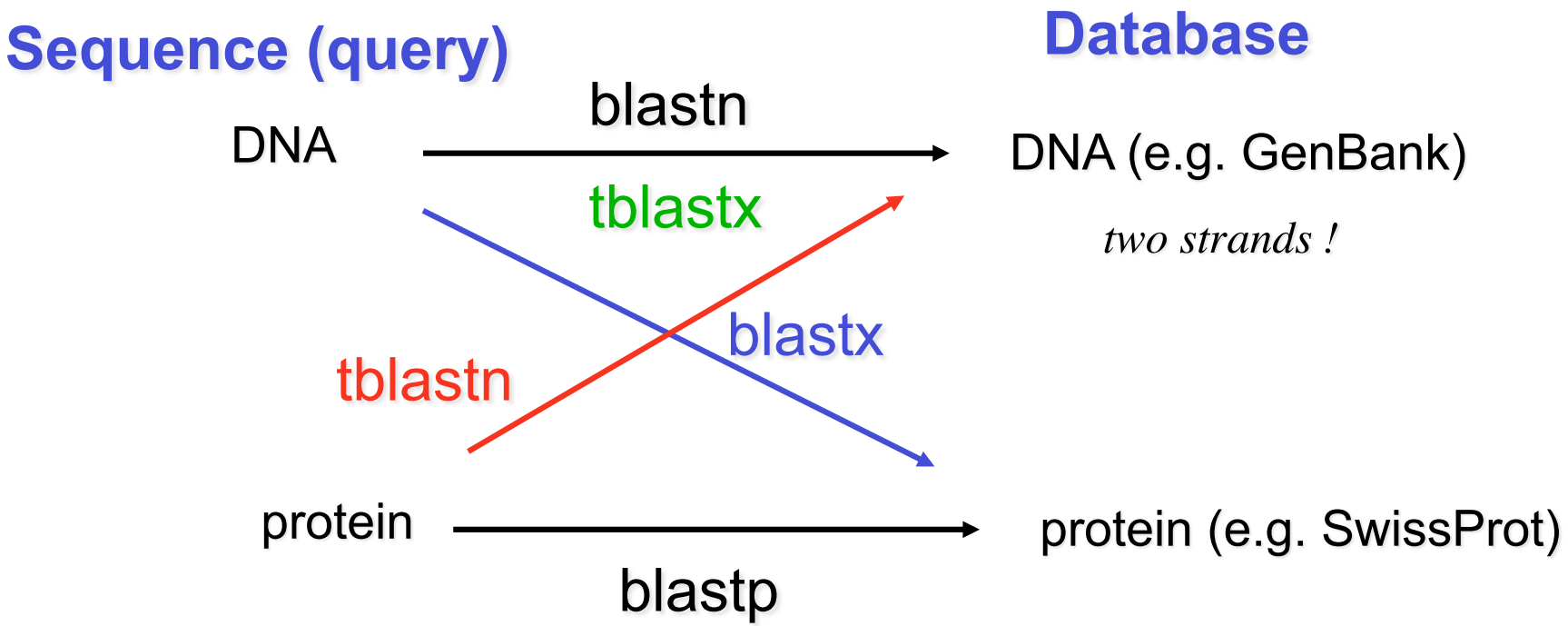

S L A A L L N K C K T P Q G Q R L V N Q W
T
X
List of similar words
P Q G 18
P E G 15
P R G 14
P K G 14
P N G 13
P D G 13
P H G 13
P M G 13
P S G 13
P Q A 12
P Q N 12
. . .
δ(P, P) = 7
δ(Q, R) = 1
δ(G, G) = 6
Threshold Score
T = 13
Query : 325 S L A A L L N K C K T P Q G Q R L V N Q W 345
+ L A + + L + T P G R + + + W
Sbjct : 290 T L A S V L D C T V T P M G S R M L K R W 310

Block alignment or global alignment : comparison
BLAST / FASTA



What to compare: DNA or protein ?

- Limits of DNA similarity search
  - Reduced alphabet (4 letters)
  - Degenerascy of the genetic code
- But … some sequences are non-coding
  - regulatory regions, structural RNAs, ...



Different versions of BLAST for different problems

- blastp: protein/protein
- blastn: DNA/DNA (useful for non-coding sequences)
- blastx: DNA-translated/protein (useful for query sequences with unidentifed coding regions; more sensitive than blastn)
- tblastn: protein/DNA-translated (useful for database sequences with unidentifed coding regions; e.g. search for homologues of a protein gene in an unannotated genome ; more sensitive than blastn)

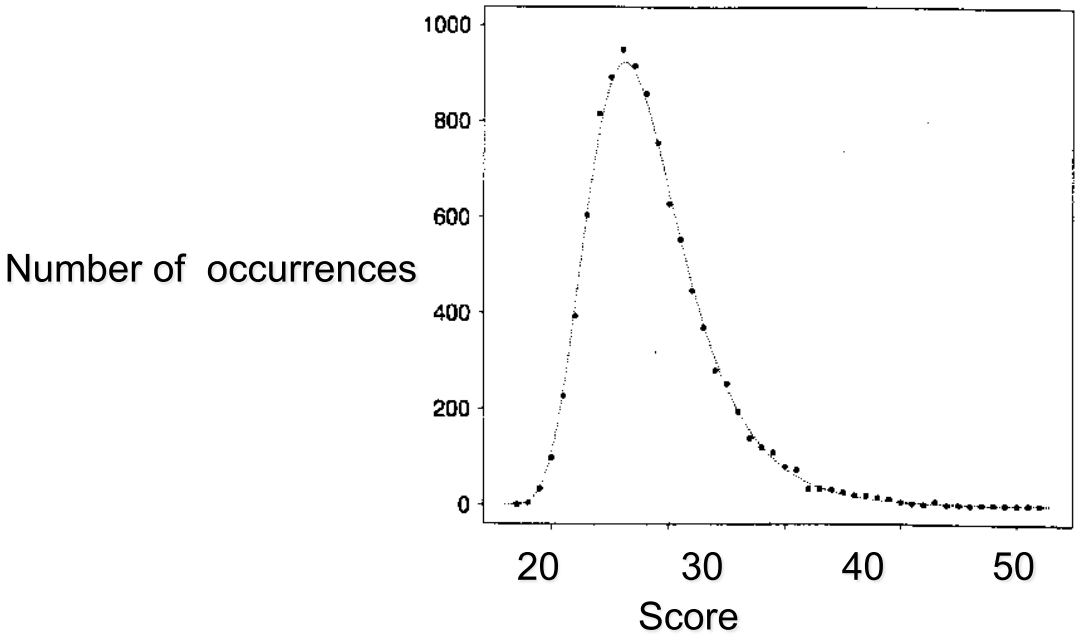
```
BLASTP 2.0.14 [Jun-29-2000]

Query= MyProtein
         (213 letters)

Database: sptrembl: SWISSPROT + TREMBL database ( Sep 23, 2002)
           897,714 sequences; 282,774,038 total letters

Searching.....................................done
```

Sequences producing significant alignments:	Score (bits)	E Value
G6PD_ECOLI 491 Glucose-6-phosphate 1-dehydrogenase (...	432	e-120
Q8XPS9 489 Probable glucose-6-phosphate 1-dehydrogen...	257	1e-37
Q9SUJ9 515 Glucose-6-phosphate 1-dehydrogenase (EC 1...	121	1e-26
AAM51346 625 Putative glucose-6-phosphate dehydr...	93	4e-18
P95611 97 Orf9 protein (Fragment).	72	9e-12
Q9VNW4 581 CG7140 protein.	69	4e-11
Q8T8Z3 526 AT14419p.	50	4e-05
O53176 435 Hypothetical protein Rv2449c.	33	3.6

```
>G6PD_BUCAI 491 Glucose-6-phosphate 1-dehydrogenase (EC 1.1.1.49) (G6PD).
          Length = 491

 Score =  239 bits (603), Expect = 3e-62
 Identities = 110/211 (52%), Positives = 156/211 (73%)

Query: 3   VTQTAQACDLVIFGAKGDLARRKLLPSLYQLEKAGQLNPDTRIIGVGRADWDKAAYTKVV 62
           + +T  ACDLVIFGAKGDL +RKLLP+LY+LEK+ +++  TRII  GRADW    Y + +
Sbjct: 2   IIETNHACDLVIFGAKGDLTKRKLLPALYKLEKSKKIHKYTRIIASGRADWSTEDYIEKI 61

Query: 63  REALETFMKETIDEGLWDTLSARLDFCNLDVNDTAAFSRLGAMLDQKNRITINYFAMPPS 122
           +  ++ F+ E I++ +W  LS+R+ FCN+DV++   F RL  +L QK  I + Y A+P +
Sbjct: 62  KTEVKNFLNEEINDLIWKNLSSRIFFCNIDVHEPLHFFRLKTILKQKKNIIVYYCAVPSN 121

Query: 123 TFGAICKGLGEAKLNAKPARVVMEKPLGTSLATSQEINDQVGEYFEECQVYRIDHYLGKE 182
           T  +I  GLG A LN+ P+R+V+EKPLG  L TS++INDQ+ +YF E Q++RIDHYLGKE
Sbjct: 122 TLNSIFIGLGNAHLNSVPSRIVLEKPLGVCLKTSKKINDQISKYFLESQIFRIDHYLGKE 181

Query: 183 TVLNLLALRFANSLFVNNWDNRTIDHVEITV 213
           ++LNL ALRF+N+    NW+N+TIDH++ITV
Sbjct: 182 SILNLFALRFSNTCLFYNWNNKTIDHIQITV 212
```

Statistical significance of similarities

- Among the similarities that have been detected, which are the ones that reflect biologically meaningful relationships ? which are the ones that are observed simply by chance ?
- Frequency distribution of similarity scores of local alignments between unrelated sequences



- Probability that a similarity of score S be observed by chance

Filtering low complexity sequences and repeated elements

- *Low complexity sequences (proteins, DNA):*

  40% of proteins DNA: microsatellites

  15%of residues example: CACACACACACACACACA

  Ala, Gly, Pro, Ser, Glu, Gln

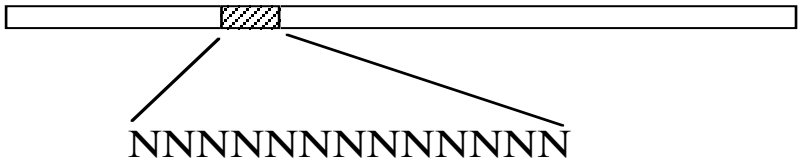
  Filtering programs: SEG, XNU, DUST

```
RSPPR--KPQGPPQQEGNNPQGPPPPAGGNPQQPQAPPAGQPQGPP
. :::    : :: : :  ::::: :  :: :.:  :: : :::::
QGPPRPGNQQCPPPQGG--PQGPPRP--GNQQRP--PPQGGPQGPP
```

- *Repeated sequences: e.g. transposable elements*

  10^6 Alu, 10^5 L1 in the human genome

  Filtering program: RepeatMasker



Searching for homologues: summary

- algorithm
- substitution matrix, weighting of gaps
- search strategy (DNA, protein)
- filtering of low complexity or repeated sequences
- completeness of sequence databases

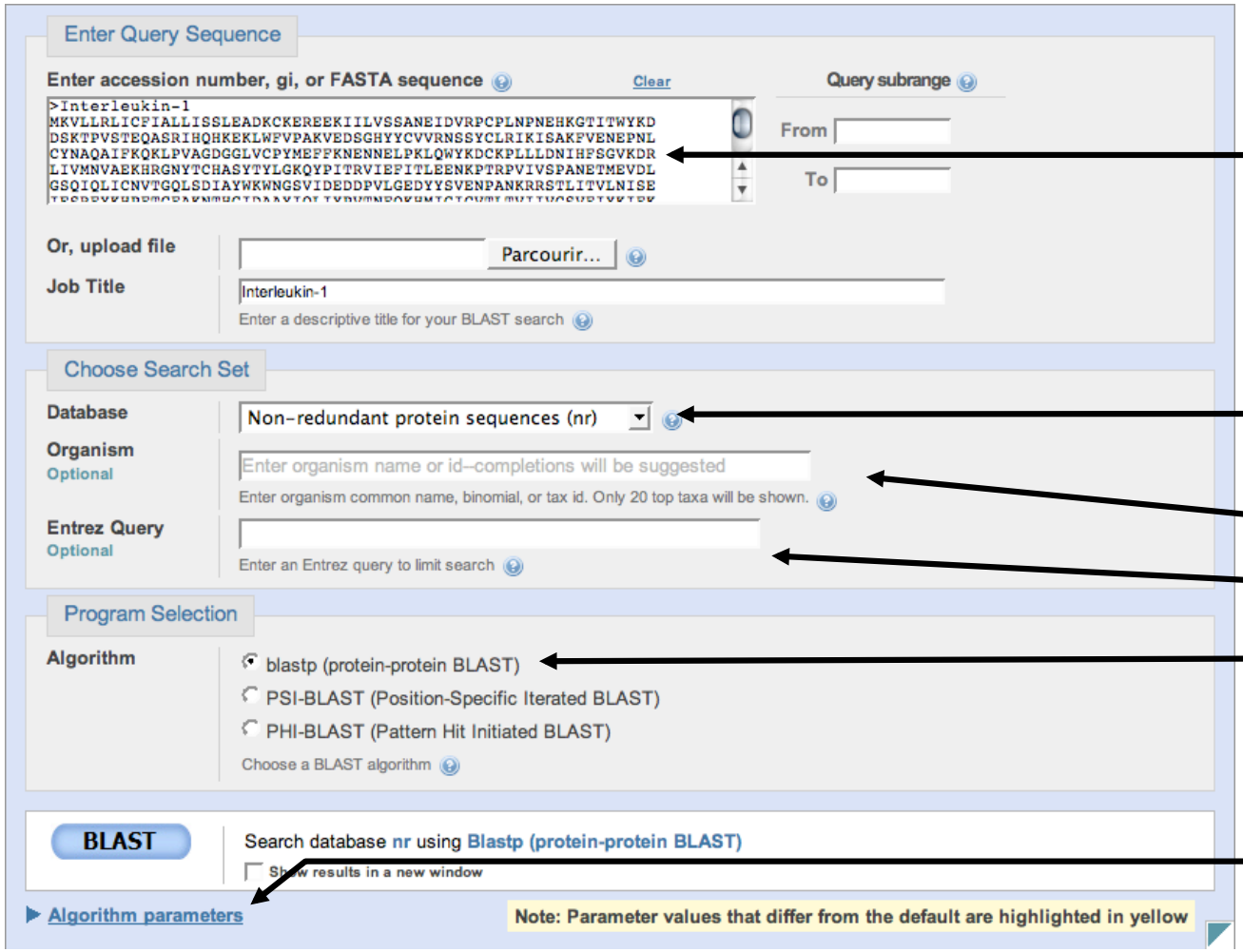
- 1 -rapid software, default parameters
- 2 - filtering (if necessary)
- 3 - change parameters (matrix, W, k, etc.)
- 4 - change algorithm
- 5 - repeat the search regularly

Special cases

- Search for similarities with very short DNA sequences (e.g. PCR primers):
  - decrease W (11 → 7)

- Very rapid search for strong similarities (e.g. cDNA to genome, human vs. chimp, ...) :
  - megablast

BLAST au NCBI (1)



Séquence requête

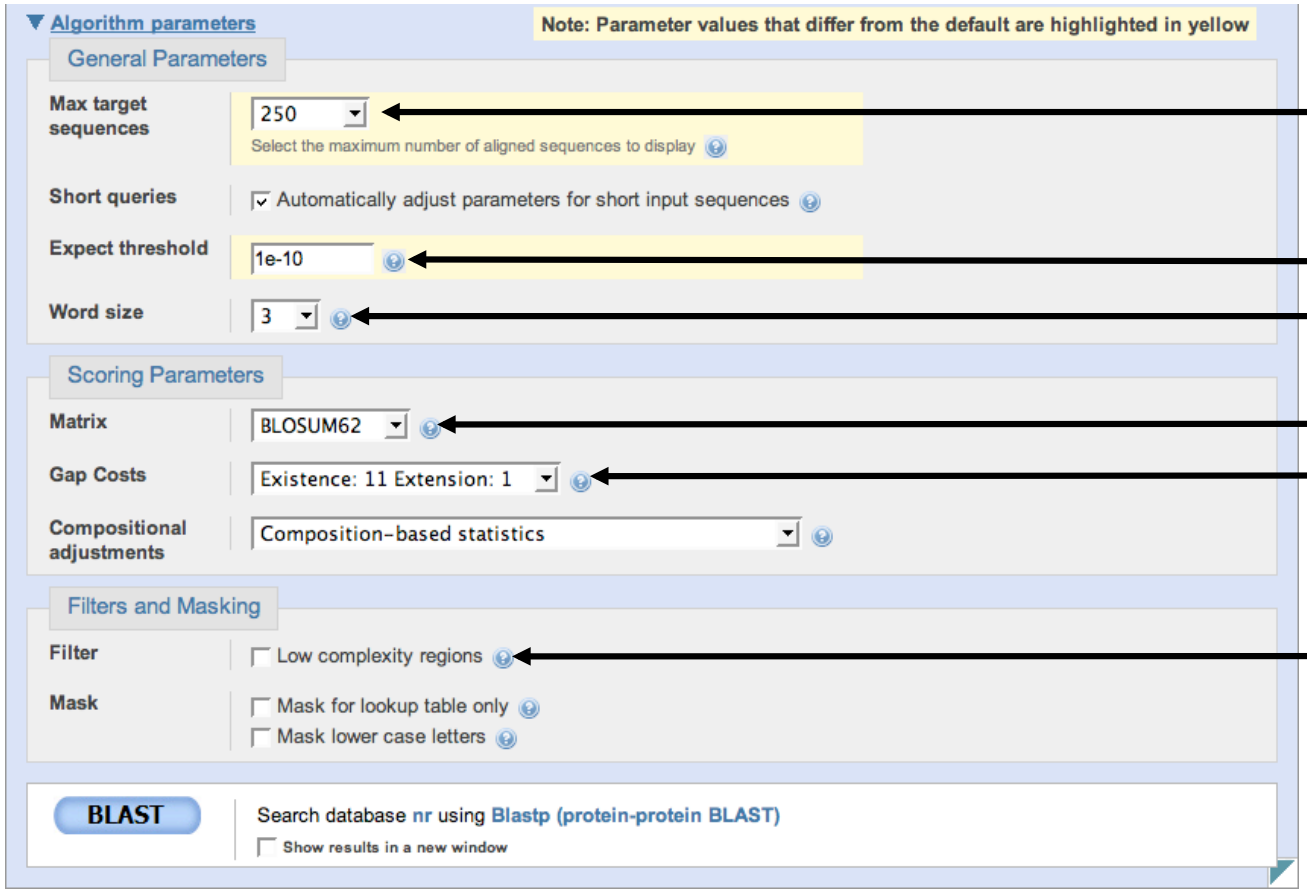
Choix de la base de données

Choix du taxon, requête Entrez

Choix de l'algorithme

Paramètres

BLAST au NCBI (2)



Nb. maximum de séquences à visualiser

E-value limite

Taille du mot

Choix de la matrice

Pénalités des *gaps*

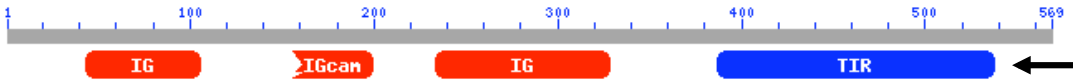
Filtrage (séquences de faible complexité)

BLAST au NCBI (3)

Job Title: Interleukin-1

▼ Show Conserved Domains

Putative conserved domains have been detected, click on the image below for detailed results.



Domaines conservés

BLASTP 2.2.16 (Mar-25-2007)

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.

Reference:
Schäffer, Alejandro A., L. Aravind, Thomas L. Madden, Sergei Shavirin, John L. Spouge, Yuri I. Wolf, Eugene V. Koonin, and Stephen F. Altschul (2001), "Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements", Nucleic Acids Res. 29:2994-3005.

RID: 6P3FH4UM012

Database: All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
5,035,465 sequences; 1,742,204,576 total letters

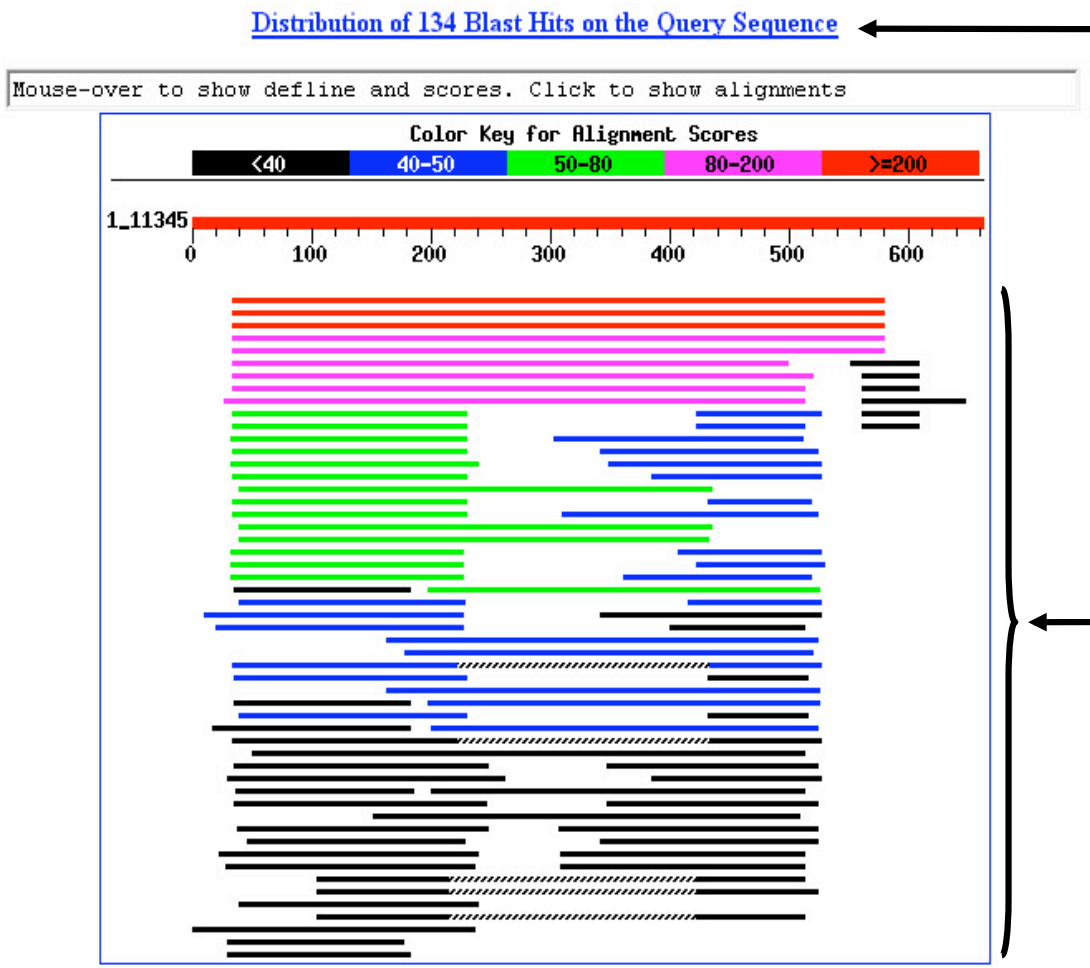
Banque de données

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

Résultats par taxon

Query= Interleukin-1
Length=569

BLAST au NCBI (4)



Nombres de *hits*

Répartition des *hits* en fonction du score

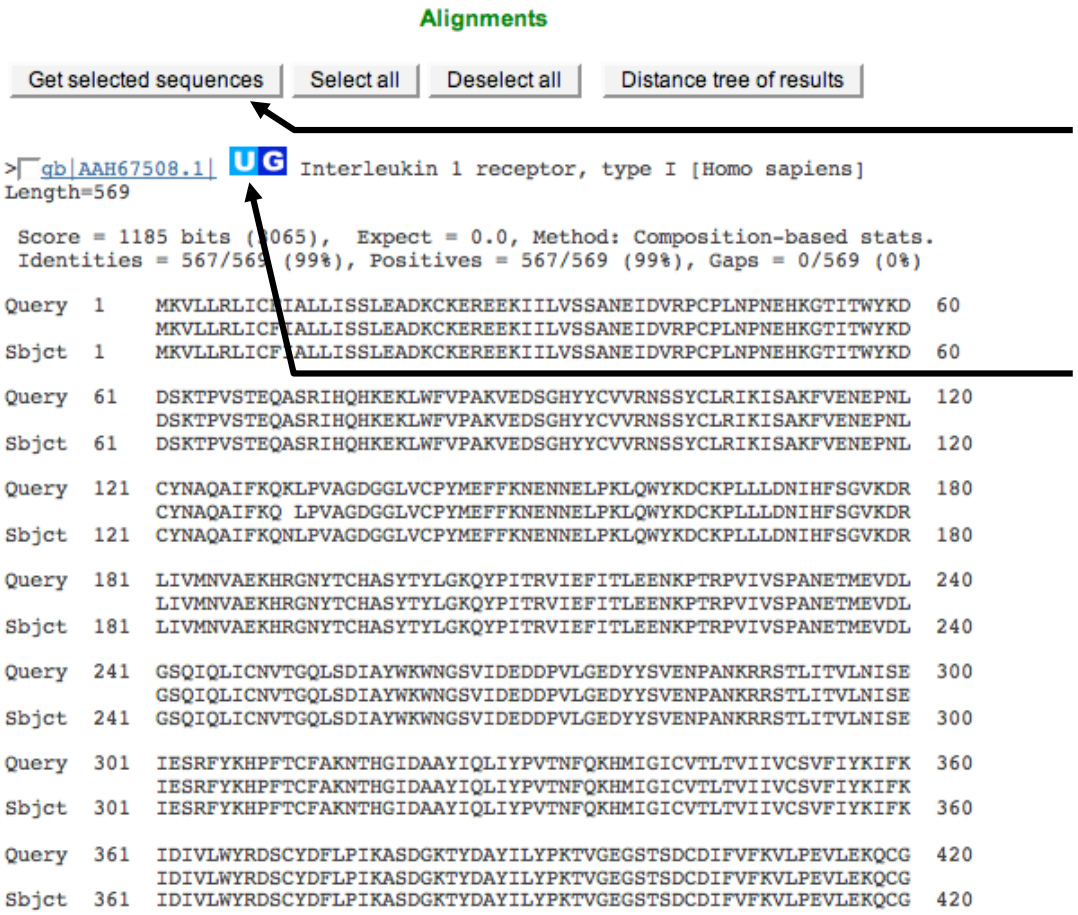
BLAST au NCBI (5)

Distance tree of results NEW Related Structures

Sequences producing significant alignments:	Score (Bits)	E Value	
ref|NP_000868.1| interleukin 1 receptor, type I precursor [Ho...	1189	0.0	U G
gb|AAH67508.1| Interleukin 1 receptor, type I [Homo sapiens]	1185	0.0	U G
gb|AAH67507.1| Interleukin 1 receptor, type I [Homo sapiens]	1181	0.0	U G
ref|XP_001107510.1| PREDICTED: interleukin 1 receptor, type I [M	1147	0.0	U G
ref|XP_001162875.1| PREDICTED: hypothetical protein isoform 1 [P	1144	0.0	G
gb|AAR88996.1| interleukin-1 receptor type I precursor [Macaca f	1130	0.0	U
ref|XP_538449.2| PREDICTED: similar to Interleukin-1 receptor...	932	0.0	U G
ref|NP_001075263.1| interleukin 1 receptor, type I [Equus cab...	923	0.0	G
emb|CAH90532.1| hypothetical protein [Pongo pygmaeus]	908	0.0	
ref|XP_593695.3| PREDICTED: hypothetical protein [Bos taurus]	852	0.0	U G
ref|NP_032388.1| interleukin 1 receptor, type I [Mus musculus...	837	0.0	U G
dbj|BAC33372.1| unnamed protein product [Mus musculus]	837	0.0	U G
dbj|BAC35666.1| unnamed protein product [Mus musculus]	835	0.0	U G
ref|NP_037255.2| interleukin 1 receptor, type I [Rattus norve...	825	0.0	U G
dbj|BAF34664.1| interleukin 1 bete receptor type 1 long trans...	825	0.0	U G
sp|Q02955|IL1R1_RAT Interleukin-1 receptor type I precursor (...	825	0.0	G
dbj|BAF34665.1| interleukin 1 beta receptor type 1 short tran...	825	0.0	U G
ref|XP_001162910.1| PREDICTED: hypothetical protein isoform 2...	799	0.0	G
dbj|BAF34663.1| interleukin 1 beta receptor type 1 soluble fo...	747	0.0	U G
pdb|1IRA|Y Chain Y, Complex Of The Interleukin-1 Receptor Wit...	675	0.0	S
pdb|1ITB|B Chain B, Type-1 Interleukin-1 Receptor Complexed W...	666	0.0	S
pdb|1G0Y|R Chain R, Il-1 Receptor Type 1 Complexed With Antag...	661	0.0	S
emb|CAH93030.1| hypothetical protein [Pongo pygmaeus]	620	6e-176	
ref|XP_001371210.1| PREDICTED: similar to interleukin-1 recep...	562	2e-158	G
ref|NP_990816.1| interleukin 1 receptor, type I [Gallus gallu...	442	4e-122	U G
gb|AAA50243.1| soluble IL-1 receptor type I	410	1e-112	U G
ref|NP_573456.1| interleukin 1 receptor-like 2 [Mus musculus]...	395	3e-108	G
gb|EDL14573.1| interleukin 1 receptor-like 2, isoform CRA_a [Mus	395	3e-108	

Séquences similaires

BLAST au NCBI (6)



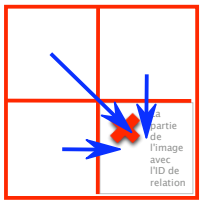
Récupération des séquences

Liens vers des banques NCBI

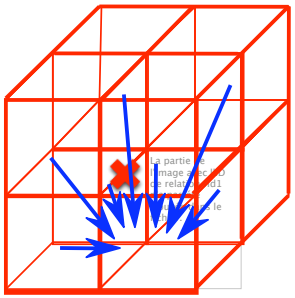
Multiple sequence alignment

Multiple alignments: impossible to use exact algorithms

- The Needleman&Wunsh algorithm can in theory be used for more than two sequences, but it is impossible to use it in practice .



Pairwise Alignment:
three possibilities



Alignment of three
sequences : seven possibilities

- The number of possible paths for aligning n sequences is proportional to $2^n - 1$.
- Computer time and memory increases exponentially with the number of sequences

  $\Rightarrow$ Use **heuristic methods**.

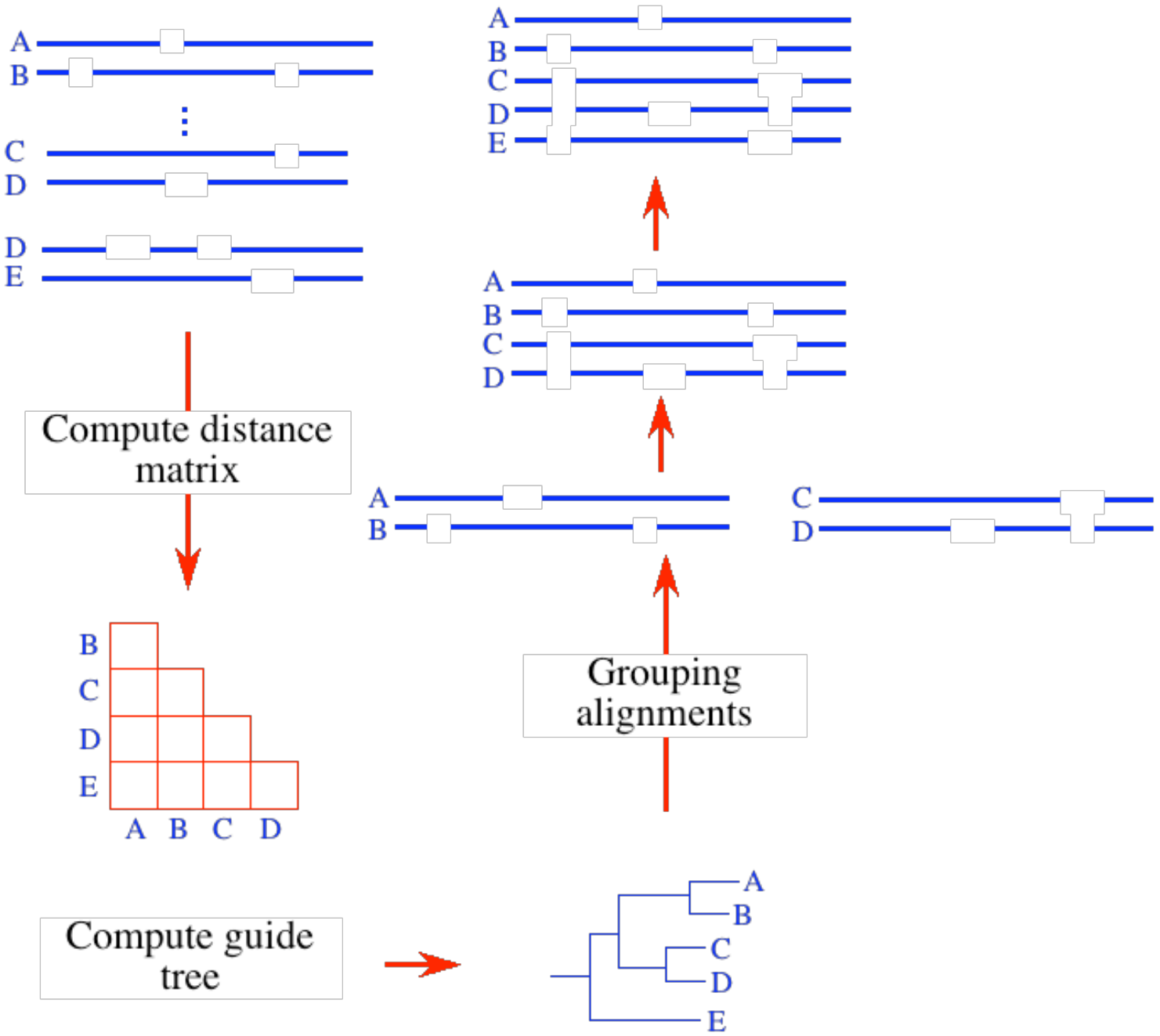
Progressive Alignment

- **Iterative** approach to compute multiple alignments, by grouping pairwise alignments.

- Three steps :
  - Alignment of sequence pairs.
  - Guide tree
  - Grouping of alignments (progressive alignment).

- **CLUSTAL** (Higgins, Sharp 1988, Thompson *et al*., 1994), the most cited multiple alignment program.

- MULTALIN, PILEUP, T-Coffee, Muscle


A
B
C
D
D
E
Compute distance matrix
B
C
D
E
A B C D
Compute guide tree
A
B
C
D
E
Grouping alignments
A
B
C
D
A
B
C
D
E

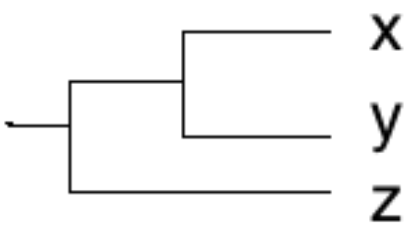
Position specific gap penalty

- Decrease gap penalty in **hydrophilic** regions (≥ 5 residues).
- Amino-acid specific gap penalty (*e.g.* lower gap penalty for Gly, Asn, Pro).

Progressive alignment : not always optimal

Alignment of three sequences

```
x ...ACTTA...
y ...AGTA.....
z ...ACGTA...
```



Guide tree

Step 1: alignment xy

```
x ACTTA    x ACTTA    x ACTTA
y A-GTA    y AGT-A    y AG-TA
```

Step 2: alignment xyz

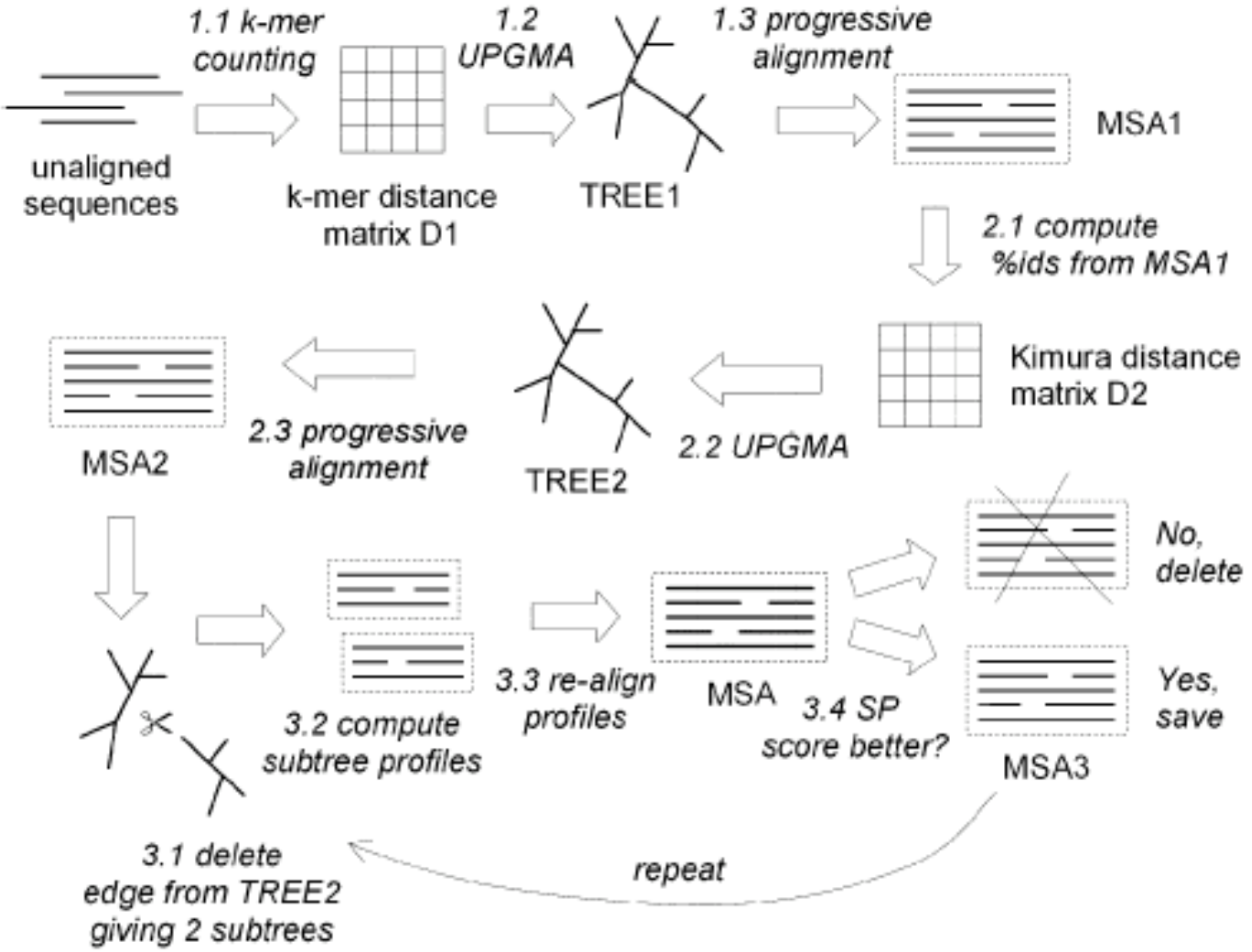
```
x ACTTA    x ACTTA    x ACTTA
y A-GTA    y AGT-A    y AG-TA
z ACGTA    z ACGTA    z ACGTA
```

- Only one of these three alignments is optimal

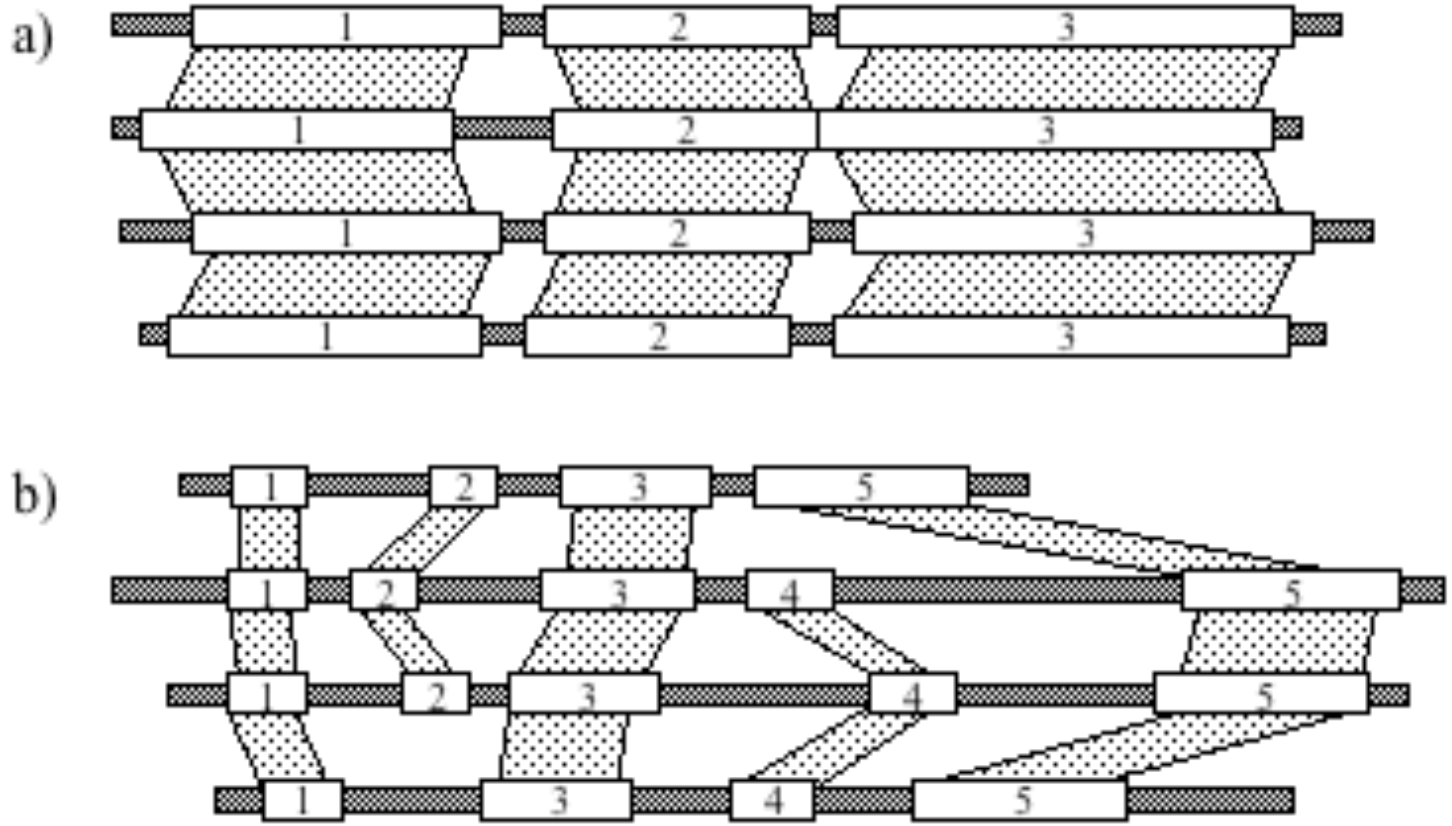
Muscle

Edgar (2004) Nucleic Acids Res. 32:1792

http://www.drive5.com/muscle/



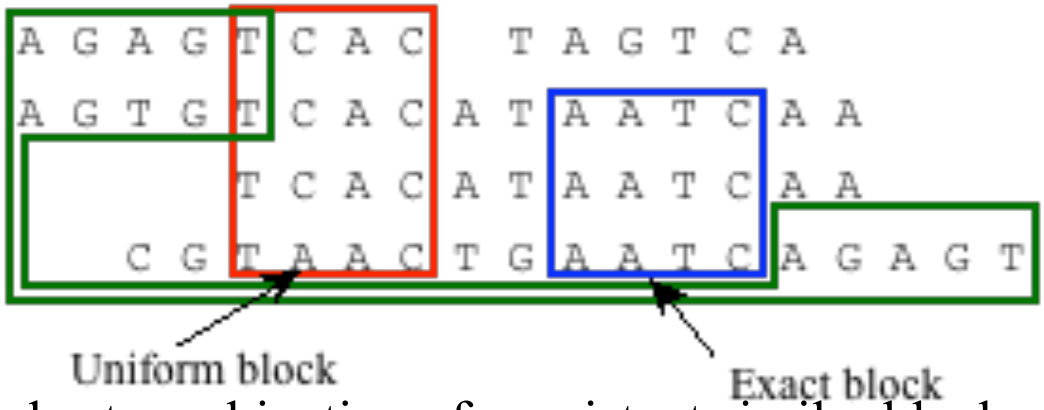
Global Alignments, Block alignments



Dialign

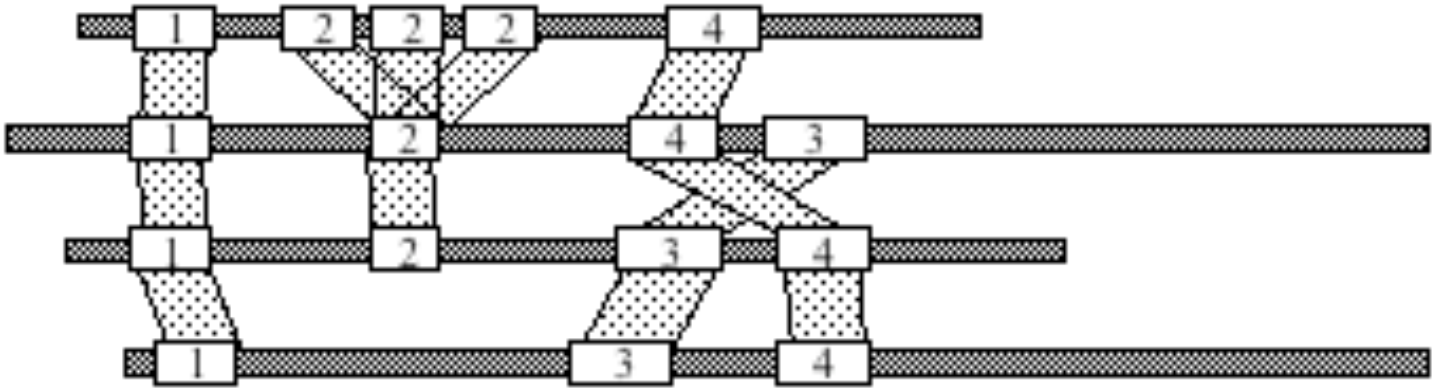
Morgenstern et al. 1996 PNAS 93:12098

- Search for similar blocks without gap



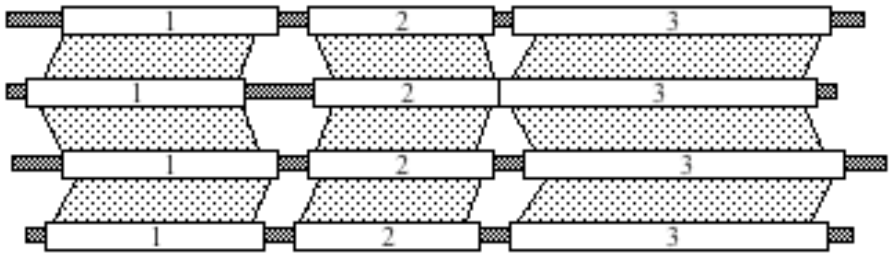
- Select the best combination of consistent similar blocks (uniforms or not) : heuristic (Abdeddaim 1997)
- Alignment anchored on blocks
- Slower than progressive alignment, but better when sequences contain large indels
- Do not try to align non-conserved regions

Local Multiple Alignments

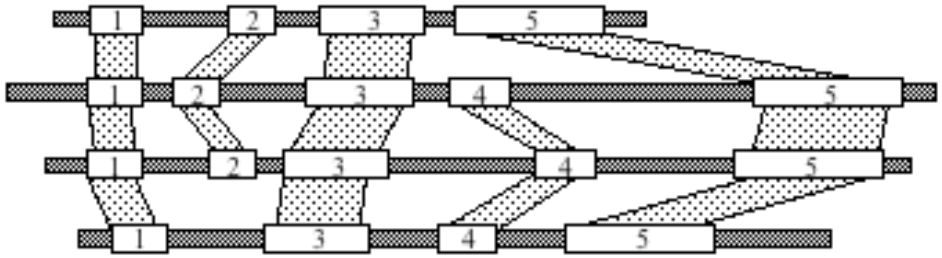


- MEME
- MATCH-BOX
- PIMA

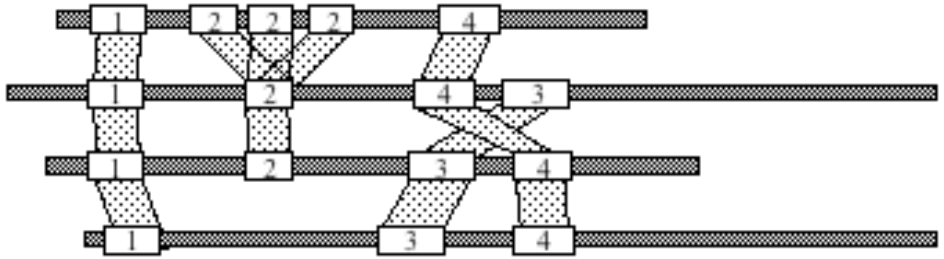
Overview



- Muscle
- ClustalW



- Dialign
- T-coffee



- MEME

Multiple alignment editor

Some special cases of sequence alignments

Alignment of protein-coding DNA sequences

```
 L   F          L   F
CTT TTC        CTT TTC
CTC ---        --- CTC
 L   -          -   L
```

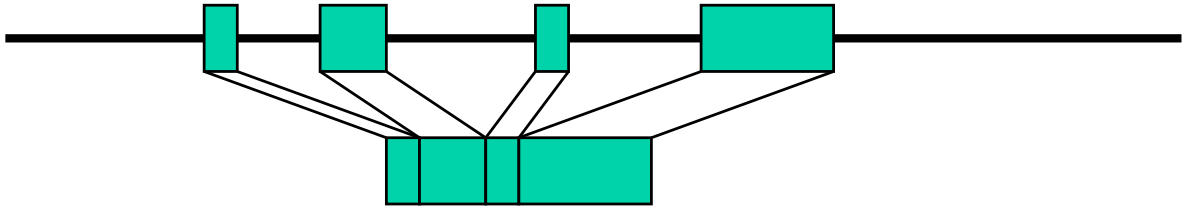
(1) alignment of protein sequences

(2) back-translation of the protein alignment into a DNA alignment

protal2dna: http://bioweb.pasteur.fr/seqanal/interfaces/protal2dna.html

Spliced alignment (1)

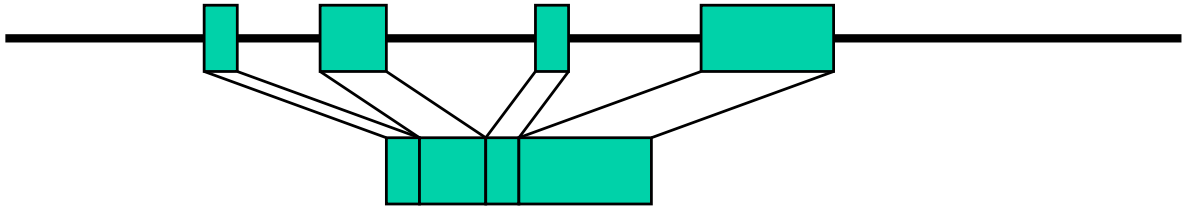
- Align an mRNA with its cognate genomic DNA => gene finding



- No gap penalty at introns => search for splice sites
- sim4, est2genome

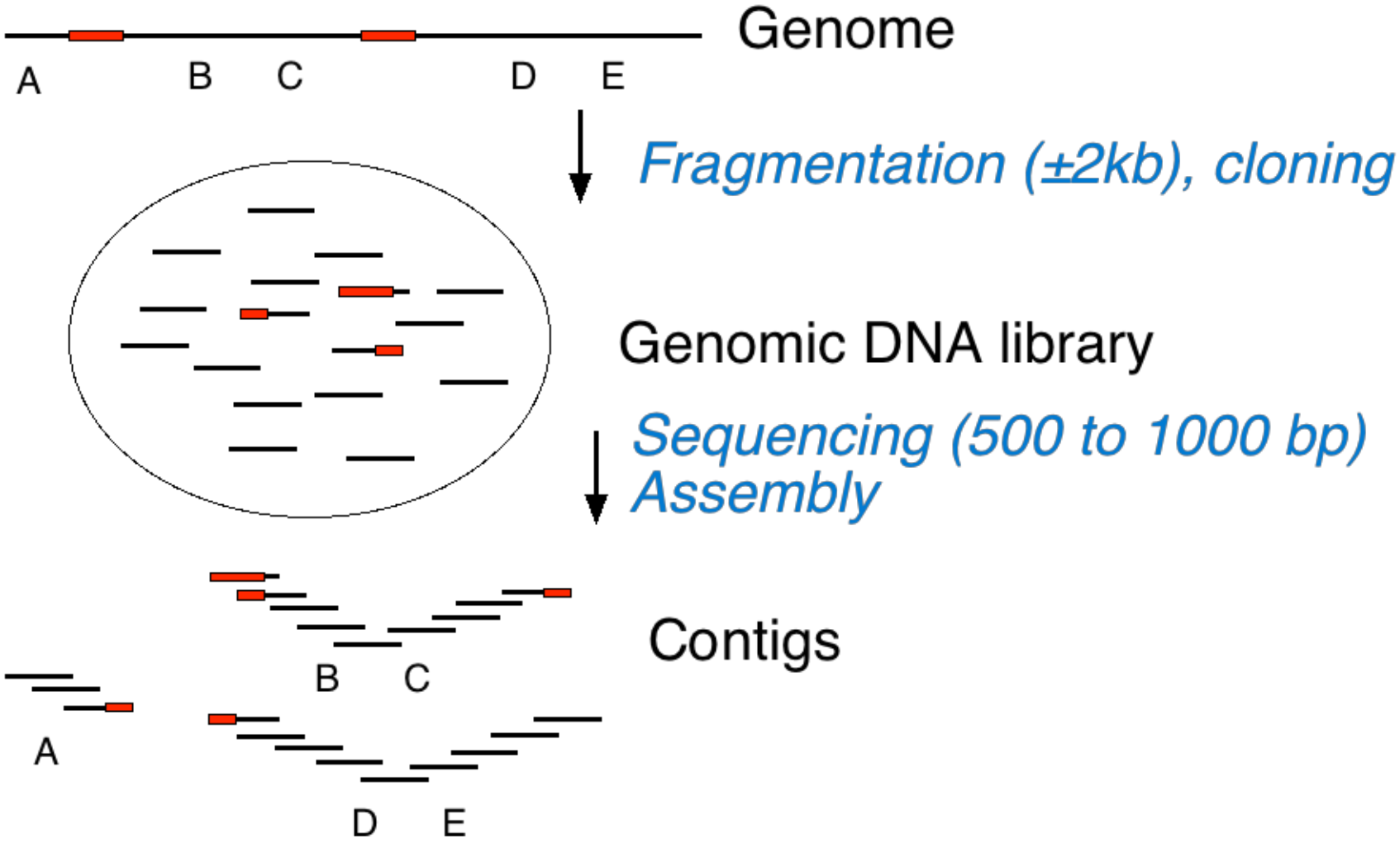
Spliced alignment (2)

- Align a protein with genomic DNA => gene prediction



- No gap penalty at introns => search for splice sites
- genewise

Shotgun sequencing



Sequence assembly

- Search for overlaps between sequence reads
- Allow for sequencing errors (or polymorphism)
- Take into account sequence quality

- cap3, phred/phrap (more complex tools for whole genome assembly)

Sequence similarity search: advanced methods

Searching for weak similarities between distantly related homologs

Limits of pairwise comparison (BLAST, FASTA, ...)

```
Seq A          CGRRLILFMLATCGECDTDSSE … HICCIKQCDVQDIIRVCC
               ::  :       :::            ::   :        :
Insulin        CGSHLVEALYLVCGERGFFYTP … EQCCTSICSLYQLENYCN
                 :::   :     :  :            ::  :  :
Seq B          YQSHLLIVLLAITLECFFSDRK … KRQWISIFDLQTLRPMTA
```

Pairwise comparison:

Insulin / Seq A : 25% identity

Insulin / Seq B : 25% identity

Insulin gene family: sequence alignment

```
                             B-chain                                  A-chain
INSL4     Q14641  ELRGCGPRFGKHLLSYCPMPEKTFTTTPGG...[x]58 ....SGRHRFDPFCCEVICDDGTSVKLCT
INSL3     P51460  REKLCGHHFVRALVRVCGGPRWSTEA.......[x]51 ....AAATNPARYCCLSGCTQQDLLTLCPY
RLN1      P04808  VIKLCGRELVRAQIAICGMSTWS..........[x]109 ....PYVALFEKCCLIGCTKRSLAKYC
BBXA      P26732  VHTYCGRHLARTLADLCWEAGVD..........[x]25 ........GIVDECCLRPCSVDVLLSYC
BBXB      P26733  ARTYCGRHLADTLADLCF--GVE..........[x]23 ........GVVDECCFRPCTLDVLLSYCG
BBXC      P26735  SQFYCGDFLARTMSILCWPDMP...........[x]25 ........GIVDECCYRPCTTDVLKLYCDKQI
BBXD      P26736  GHIYCGRYLAYKMADLCWRAGFE..........[x]25 ........GIADECCLQPCTNDVLLSYC
LIRP      P15131  VARYCGEKLSNALKLVCRGNYNTMF........[x]58 ........GVFDECCRKSCSISELQTYCGRR
MIP I     P07223  RRGVCGSALADLVDFACSSSNQPAMV.......[x]29 ....QGTTNIVCECCMKPCTLSELRQYCP
MIP II    P25289  PRGICGSNLAGFRAFICSNQNSPSMV.......[x]44 ....QRTTNLVCECCFNYCTPDVVRKYCY
MIP III   P80090  PRGLCGSTLANMVQWLCSTYTTSSKV.......[x]30 ....ESRPSIVCECCFNQCTVQELLAYC
MIP V     P31241  PRGICGSDLADLRAFICSRRNQPAMV.......[x]44 ....QRTTNLVCECCYNVCTVDVFYEYCY
MIP VII   P91797  PRGLCGNRLARAHANLCFLLRNTYPDIFPR...[x]86 ..EVMAEPSLVCDCCYNECSVRKLATYC
ILP       P22334  AEYLCGSTLADVLSFVCGNRGYNSQP.......[x]31 ........GLVEECCYNVCDYSQLESYCNPYS
INS       P01308  NQHLCGSHLVEALYLVCGERGFFYTPKT.....[x]35 ........GIVEQCCTSICSLYQLENYCN
IGF1      P01343  PETLCGAELVDALQFVCGDRGFYF.........[x]12 ........GIVDECCFRSCDLRRLEMYCAPLK
IGF2      P01344  SETLCGGELVDTLQFVCGDRGFYF.........[x]12 ........GIVEECCFRSCDLALLETYCATPA
                      *.         .*                                   **   *    .   *
```

Biomolecular Sequence Motif Descriptors

- Consensus: e.g. TATA box: TATAWAWR

- Regular expression: e.g. insulins PROSITE pattern

  C-C-{P}-x(2-4)-C-[STDNEKPI]-x(3)-[LIVMFS]-x(3)-C

- Position-specific weight matrix (profiles, hidden markov models) : position-specific weighting of substitutions and indels

Matrix of position-specific amino-acid frequency (A-chain of insulin)

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	-
1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	16
3	1	0	0	1	0	0	0	0	0	0	1	0	0	3	0	1	0	0	0	0	10
4	2	0	0	0	0	2	0	0	0	0	0	0	1	0	2	1	0	0	0	0	9
5	1	0	0	1	0	0	0	0	0	0	0	0	0	0	2	0	3	0	0	1	9
6	0	0	0	0	0	0	1	0	0	0	0	0	2	0	0	0	4	1	0	0	9
7	1	0	0	0	0	9	0	0	0	0	0	4	0	0	1	2	0	0	0	0	0
8	0	0	0	0	1	0	0	8	0	0	5	0	1	0	0	0	0	2	0	0	0
9	2	0	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	12	0	0	0
10	0	5	6	4	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0
11	0	0	1	12	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	0
12	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	1	5	0	0	0	0	4	1	0	0	0	1	0	1	0	0	4	0
15	0	0	0	0	0	0	0	1	2	0	0	5	0	1	5	2	0	1	0	0	0
16	0	0	0	1	0	2	0	2	0	0	0	0	5	1	0	3	0	2	0	1	0
17	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	4	9	0	0	0	0
19	0	0	1	0	0	0	0	1	1	5	0	1	1	1	0	0	1	4	0	1	0
20	1	0	6	0	0	1	0	0	0	0	0	0	0	2	3	3	0	0	0	1	0
21	0	0	1	3	0	0	0	0	1	1	0	0	0	2	1	1	1	6	0	0	0
22	0	0	0	0	1	0	0	0	0	14	0	0	0	0	0	1	0	1	0	0	0
23	2	0	0	4	0	0	0	0	1	5	0	0	0	1	2	0	0	1	0	1	0
24	1	0	0	1	0	0	0	0	3	1	1	1	0	1	0	4	4	0	0	0	0
25	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	15	0
26	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	2	0	1	0	0	2	0	0	0	0	0	2	2	0	0	0	1	0	0	2	5
28	0	0	0	0	0	0	0	0	1	0	0	0	2	0	1	0	1	0	0	1	11
29	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	0	0	0	0	1	12
30	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	0	13

Alignment of SeqA with the matrix of position-specific amino-acid frequency

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	-	SeqA
1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	-
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	16	-
3	1	0	0	1	0	0	0	0	0	0	1	0	0	3	0	1	0	0	0	0	10	-
4	2	0	0	0	0	2	0	0	0	0	0	0	1	0	2	1	0	0	0	0	9	-
5	1	0	0	1	0	0	0	0	0	0	0	0	0	0	2	0	3	0	0	1	9	-
6	0	0	0	0	0	0	1	0	0	0	0	0	2	0	0	0	4	1	0	0	9	-
7	1	0	0	0	0	9	0	0	0	0	0	4	0	0	1	2	0	0	0	0	0	-
8	0	0	0	0	1	0	0	8	0	0	5	0	1	0	0	0	0	2	0	0	0	-
9	2	0	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	12	0	0	0	-
10	0	5	6	4	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	H
11	0	0	1	12	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	0	I
12	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	C
13	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	C
14	0	0	0	1	5	0	0	0	0	4	1	0	0	0	1	0	1	0	0	4	0	I
15	0	0	0	0	0	0	0	1	2	0	0	5	0	1	5	2	0	1	0	0	0	K
16	0	0	0	1	0	2	0	2	0	0	0	0	5	1	0	3	0	2	0	1	0	Q
17	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	C
18	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	4	9	0	0	0	0	D
19	0	0	1	0	0	0	0	1	1	5	0	1	1	1	0	0	1	4	0	1	0	V
20	1	0	6	0	0	1	0	0	0	0	0	0	0	2	3	3	0	0	0	1	0	Q
21	0	0	1	3	0	0	0	0	1	1	0	0	0	2	1	1	1	6	0	0	0	D
22	0	0	0	0	1	0	0	0	0	14	0	0	0	0	0	1	0	1	0	0	0	I
23	2	0	0	4	0	0	0	0	1	5	0	0	0	1	2	0	0	1	0	1	0	I
24	1	0	0	1	0	0	0	0	3	1	1	1	0	1	0	4	4	0	0	0	0	R
25	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	15	0	V
26	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	C
27	2	17	1	0	0	2	0	0	0	0	0	2	2	0	0	0	1	0	0	2	5	C
28	0	0	0	0	0	0	0	0	1	0	0	0	2	0	1	0	1	0	0	1	11	-
29	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	0	0	0	0	1	12	-
30	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	0	13	-
																				Score:		100

Alignment of SeqB with the matrix of position-specific amino-acid frequency

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	-	SeqB
1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	-
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	16	-
3	1	0	0	1	0	0	0	0	0	0	1	0	0	3	0	1	0	0	0	0	10	-
4	2	0	0	0	0	2	0	0	0	0	0	0	1	0	2	1	0	0	0	0	9	-
5	1	0	0	1	0	0	0	0	0	0	0	0	0	0	2	0	3	0	0	1	9	-
6	0	0	0	0	0	0	1	0	0	0	0	0	2	0	0	0	4	1	0	0	9	-
7	1	0	0	0	0	9	0	0	0	0	0	4	0	0	1	2	0	0	0	0	0	-
8	0	0	0	0	1	0	0	8	0	0	5	0	1	0	0	0	0	2	0	0	0	-
9	2	0	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	12	0	0	0	-
10	0	5	6	4	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	K
11	0	0	1	12	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	0	R
12	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Q
13	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	W
14	0	0	0	1	5	0	0	0	0	4	1	0	0	0	1	0	1	0	0	4	0	I
15	0	0	0	0	0	0	0	1	2	0	0	5	0	1	5	2	0	1	0	0	0	S
16	0	0	0	1	0	2	0	2	0	0	0	0	5	1	0	3	0	2	0	1	0	I
17	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	F
18	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	4	9	0	0	0	0	D
19	0	0	1	0	0	0	0	1	1	5	0	1	1	1	0	0	1	4	0	1	0	L
20	1	0	6	0	0	1	0	0	0	0	0	0	0	2	3	3	0	0	0	1	0	Q
21	0	0	1	3	0	0	0	0	1	1	0	0	0	2	1	1	1	6	0	0	0	T
22	0	0	0	0	1	0	0	0	0	14	0	0	0	0	0	1	0	1	0	0	0	L
23	2	0	0	4	0	0	0	0	1	5	0	0	0	1	2	0	0	1	0	1	0	R
24	1	0	0	1	0	0	0	0	3	1	1	1	0	1	0	4	4	0	0	0	0	P
25	0	17	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	15	0	M
26	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	T
27	2	0	1	0	0	2	0	0	0	0	0	2	2	0	0	0	1	0	0	2	5	A
28	0	0	0	0	0	0	0	0	1	0	0	0	2	0	1	0	1	0	0	1	11	-
29	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	0	0	0	0	1	12	-
30	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	0	13	-
																				Score:		34

Position-specific weight matrix

- Matrix of position-specific amino-acid frequency
- log transformation => position-specific weight matrix = profile
- Similar approach using HMM

DNA weight matrix

- Splice donnor sites of vertebrates: frequency (%) of the four bases at each position

Base	Position								
	-3	-2	-1	+1	+2	+3	+4	+5	+6
A	**33**	**60**	8	0	0	**49**	**71**	6	15
C	**37**	13	4	0	0	3	7	5	19
G	18	14	**81**	**100**	0	**45**	12	**84**	20
T	12	13	7	0	**100**	3	9	5	**46**
Cons.	M	A	G	G	T	R	A	G	T

Searching for distantly related homologues in sequence databases

- 1- search for homologues (e.g. BLAST)
- 2- align homologues (e.g. MUSCLE, MEME)
- 3- compute a profile from the multiple alignment
- 4- compare the profile to a sequence database (e.g. MAST, pfsearch)

- pfsearch: http://www.isrec.isb-sib.ch/profile/profile.html
- MEME/MAST: http://meme.sdsc.edu/meme/website/

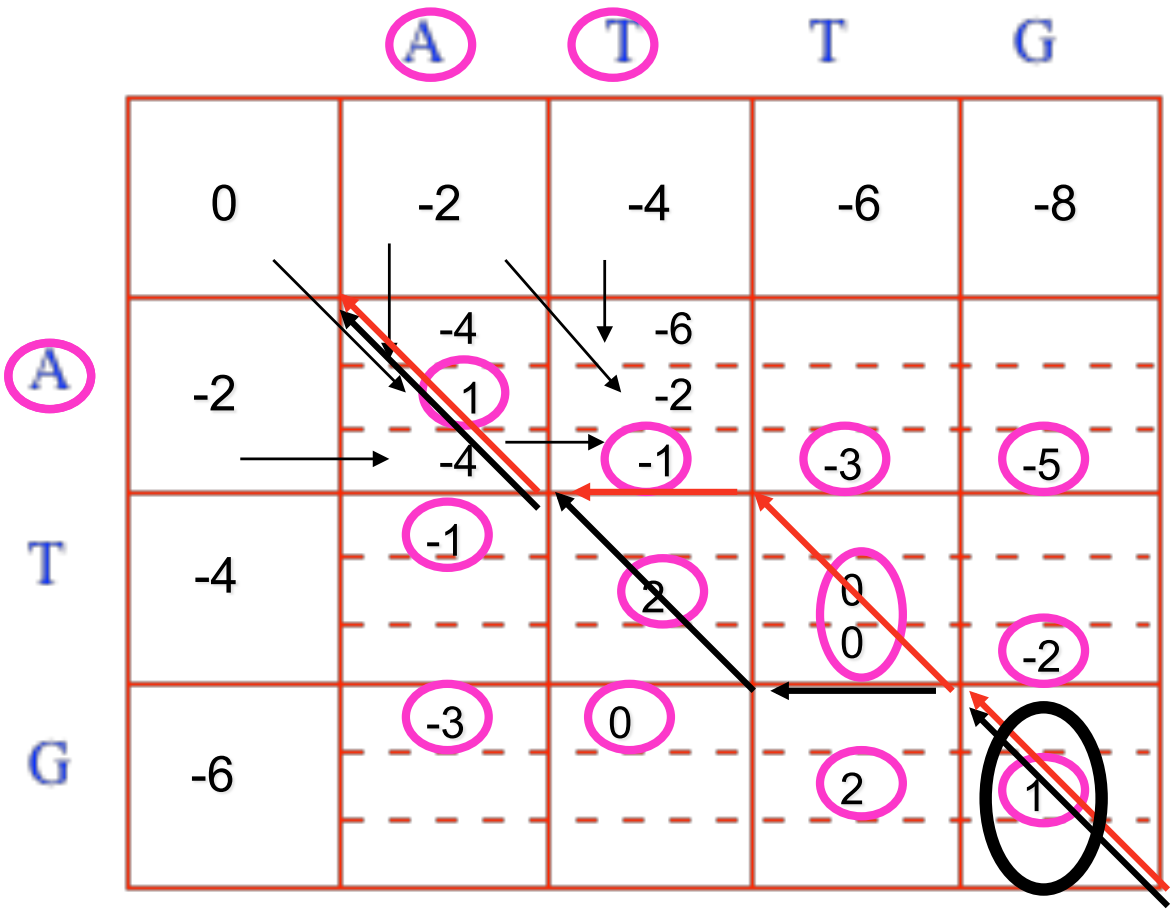
PSI-BLAST

- Position-Specific Iterated BLAST
  - 1- classical BLAST search
  - 2- compute a profile with significant BLAST hits
  - 3- BLAST search based on the profile
  - 4 -repeat steps 2-3 up to convergence

- More sensitive than Smith-Waterman
- 40 times faster

Comparison of a sequence to a database of protein motifs

- Databases: PROSITE, PFAM, PRODOM, …, INTERPRO
- Search tools:
  - ProfileScan : http://hits.isb-sib.ch/cgi-bin/PFSCAN



Identité : +1
Mismatch : +0
Gap : −2

A	T	T	G
A	T	-	G

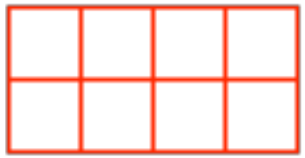
S = 1 + 1 -2 + 1 = 1

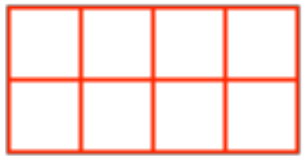
A	T	T	G
A	-	T	G

S = 1 -2 + 1 + 1 = 1

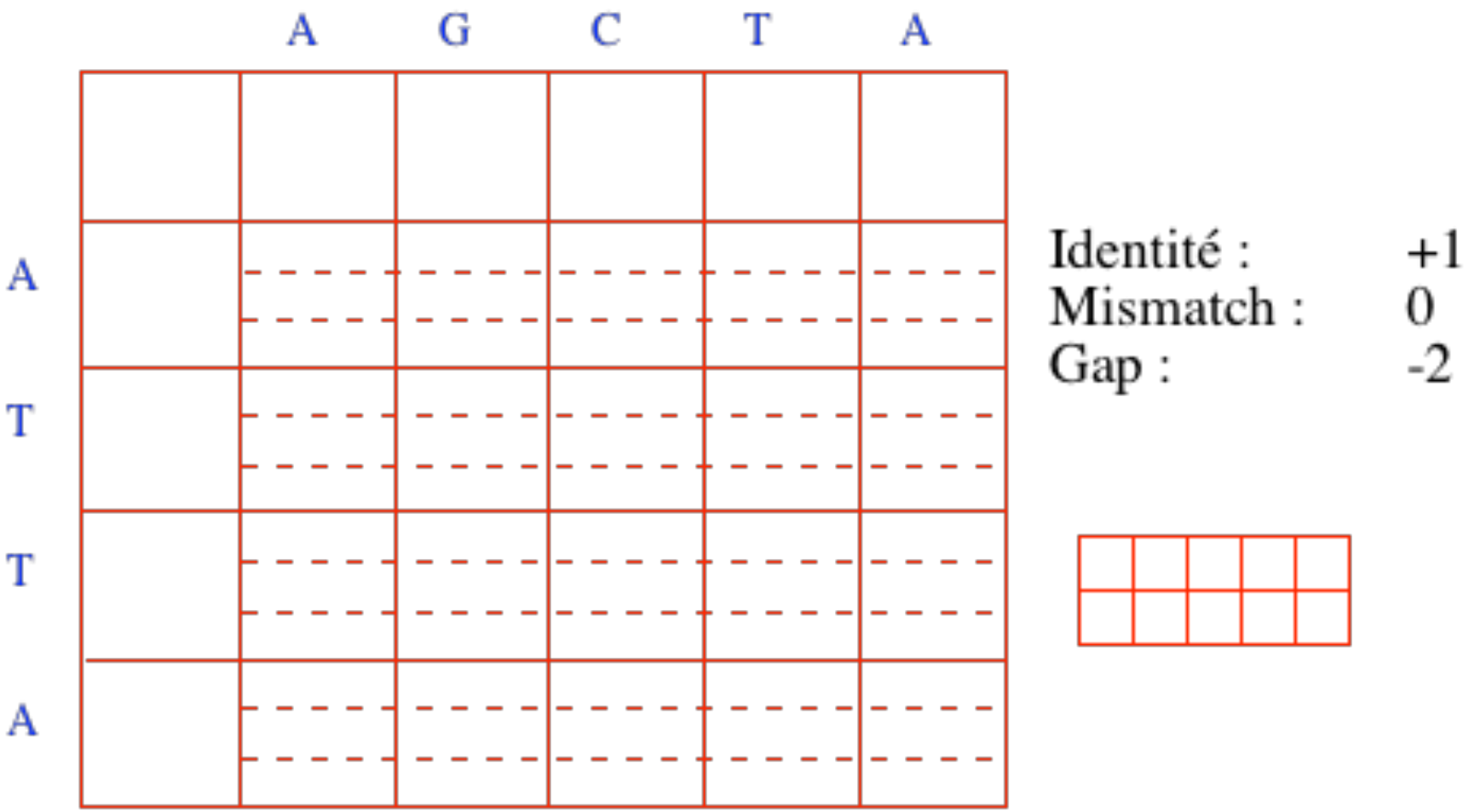
- Needleman & Wunsh, 1970

Identité : +1
Mismatch : +0
Gap : –2

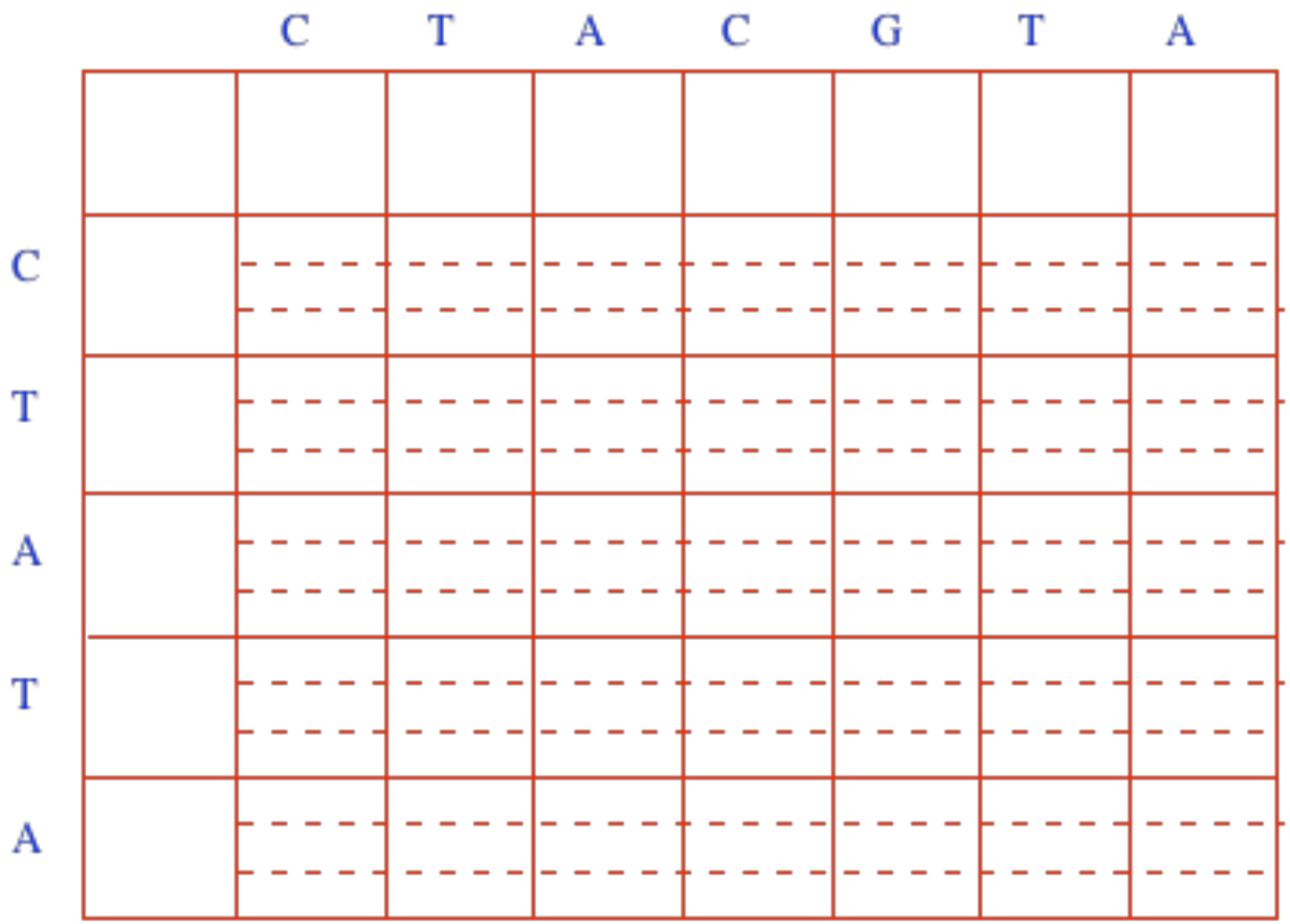




N=129 possible alignments



N=681 possible alignments



Identité : +1
Mismatch : +0
Gap : –2

N=7183 possible alignments