Alignment of biological sequences

PhD Program on Computational Biology 2005

Laurent Duret

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

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

Bioinformatics and Evolutionary Genomics

- Molecular evolution: understand genome organization, function and evolution

- Bioinformatics: develop software and databases for comparative genomics and phylogenetics (Pôle Bioinformatique Lyonnais)

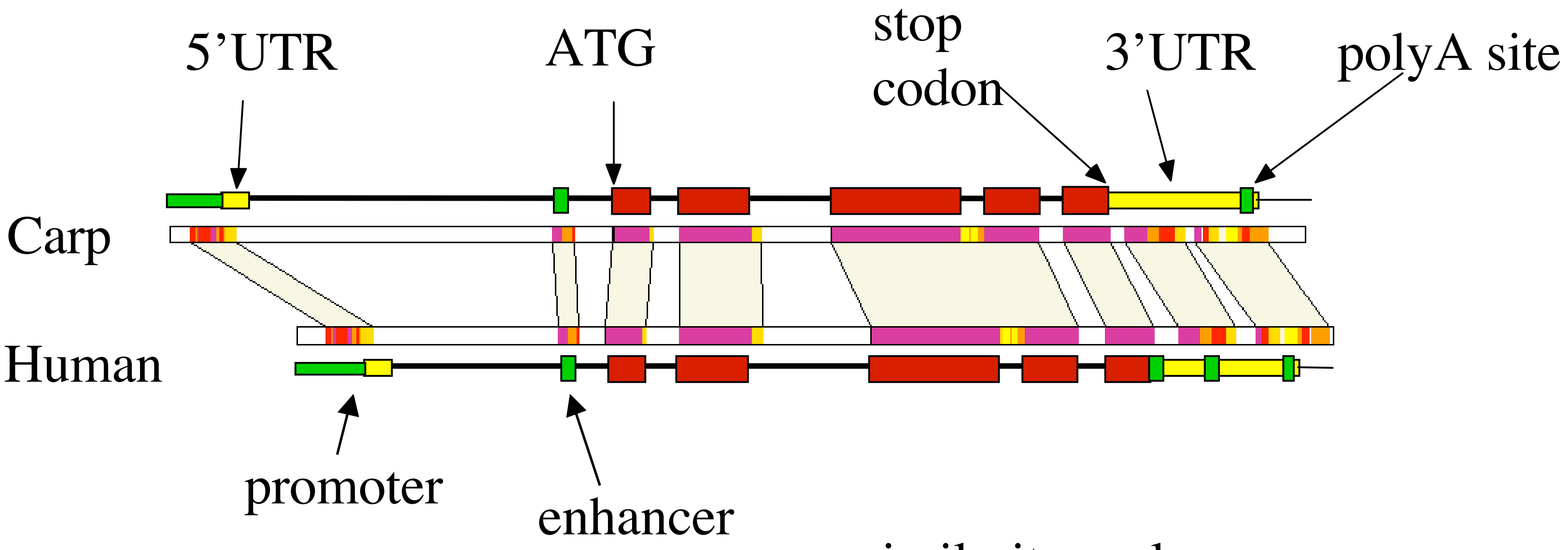
Sequence alignment

- 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



5'UTR

ATG

stop codon

3'UTR

polyA site

Carp

Human

promoter

enhancer

introns: —

coding regions : 🟥

regulatory elements 🟩

similarity scale :

□ no significant similarity

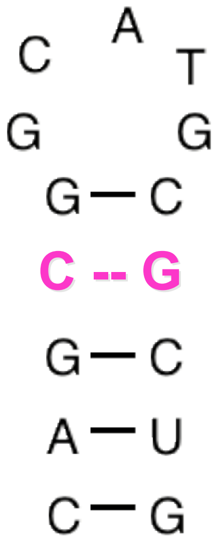
🟧 80 - 90% identity

🟪 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

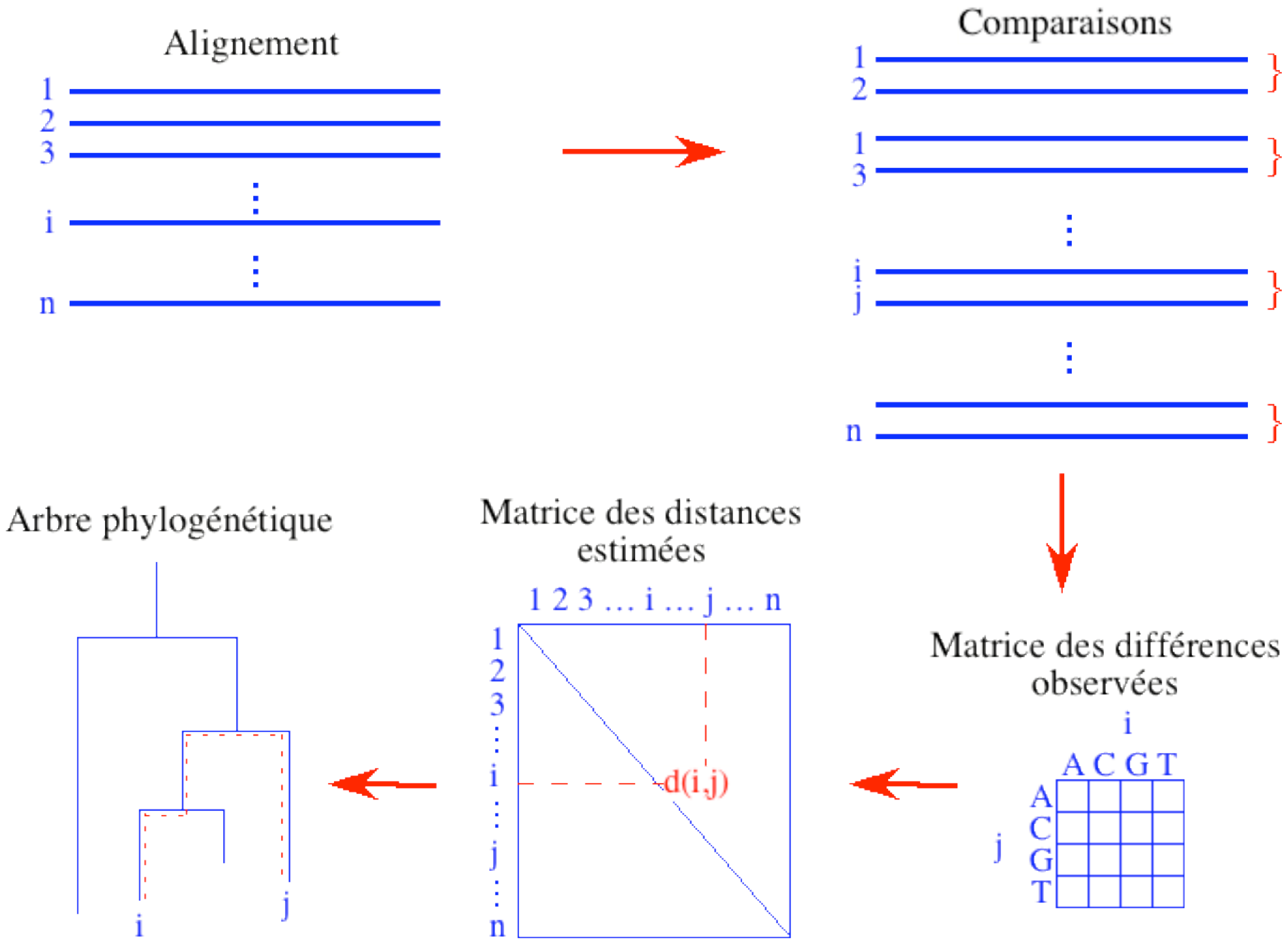


```
C A G U G G C A U G C A C U G
C A G C G G C G U G C G C U G
C A G C G G T A U G C G C U G
C A A U G G T A U G C A U U G
C A G U G G C A U G C A C U G
* *     * *     * * *     * *
```

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



Alignement

Comparaisons

Matrice des différences observées

Matrice des distances estimées

Arbre phylogénétique

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

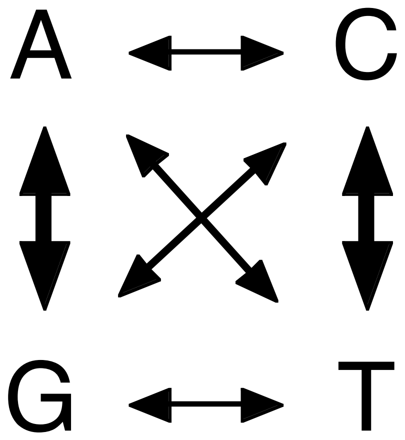
How do we measure sequence similarity ?

```
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)

$$\begin{array}{c|cccc} & A & C & G & T \\ \hline A & 1 & 0 & 0.5 & 0 \\ C & 0 & 1 & 0 & 0.5 \\ G & 0.5 & 0 & 1 & 0 \\ T & 0 & 0.5 & 0 & 1 \end{array}$$

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

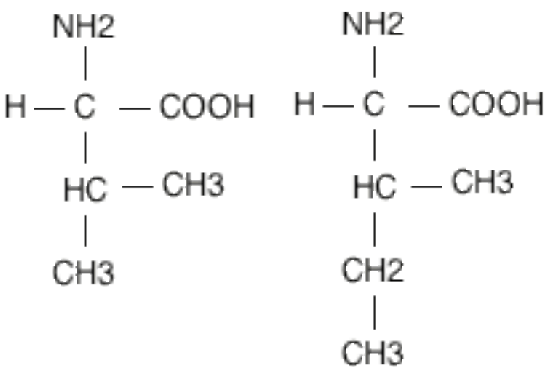
Models of evolution (proteins)

- 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.)

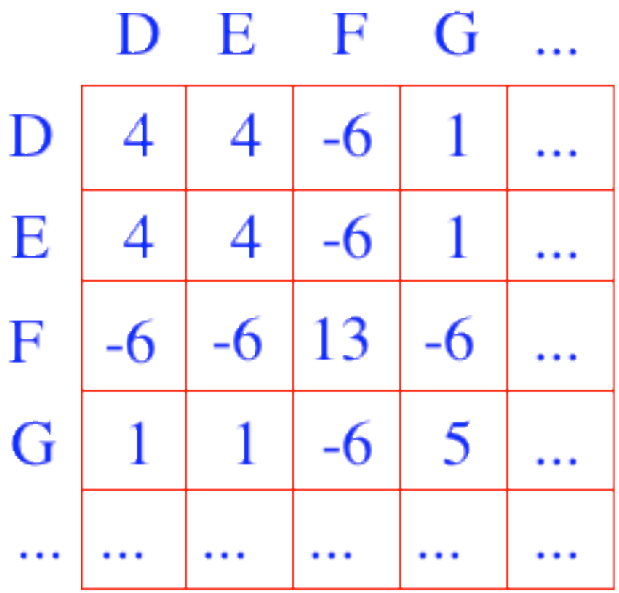
conservative substitutions

Val Ile



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	...
...	...	...	...	...	...

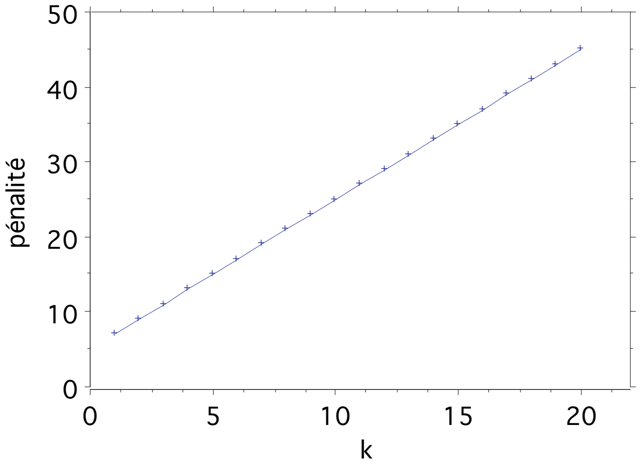
Weighting of gaps

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

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

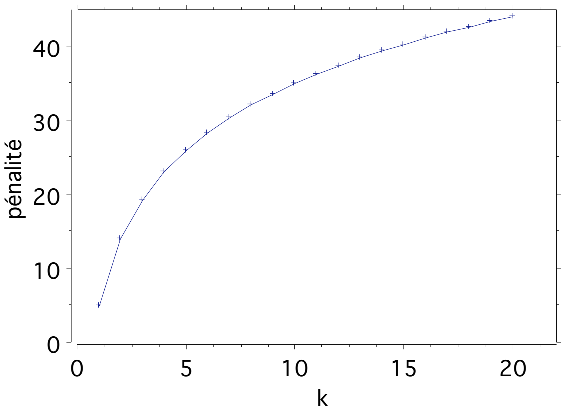
  δ_o : penalty for gap opening

  δ_e : penalty for gap extension



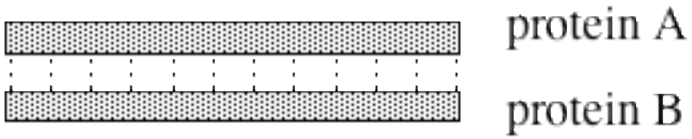
Weighting of gaps (more realistic)

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

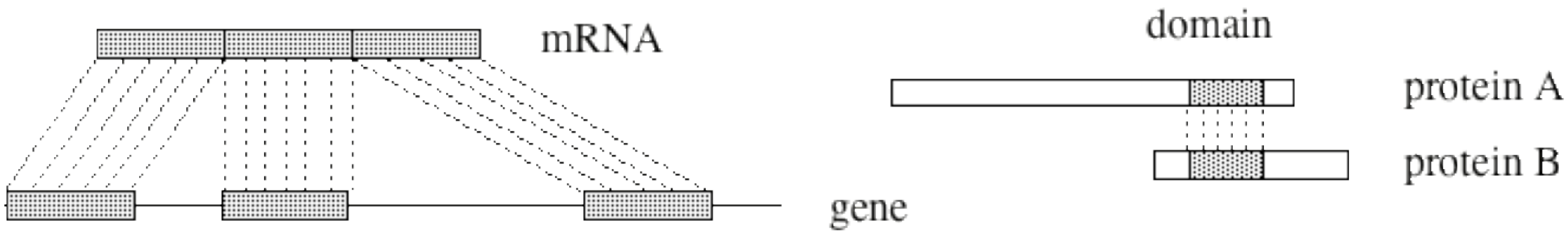


  - $w = \mathrm{f}(\mathit{log(k), log(PAM), residue, structure})$
    - *PAM: the probability of a gap increases with the evolutionary distance*
    - *Resides, structure: the probability of a gap is higher in a loop (hydrophilic) than in the hydrophobic core of proteins*

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--  ...
```

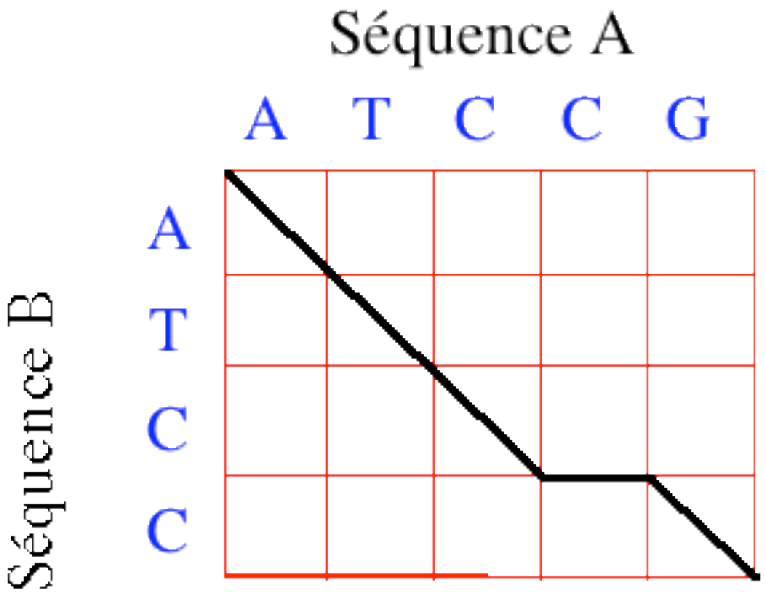
- Objective: for a given scoring scheme, find the best alignment(s), i.e. the optimal alignment(s)

- Problem: the number of possible alignments between two sequences increases exponentially with the length of sequences

Algorithms for aligning two sequences

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

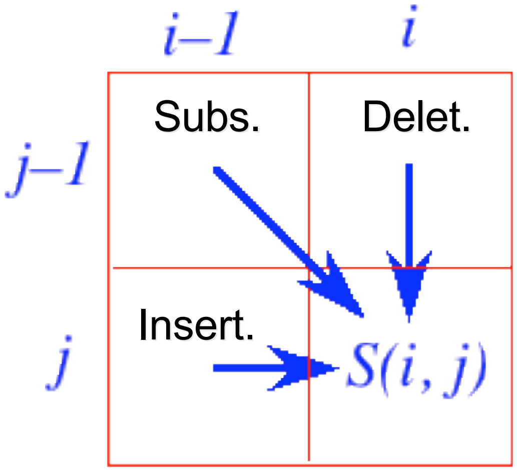
Alignment representation: a path in a matrix



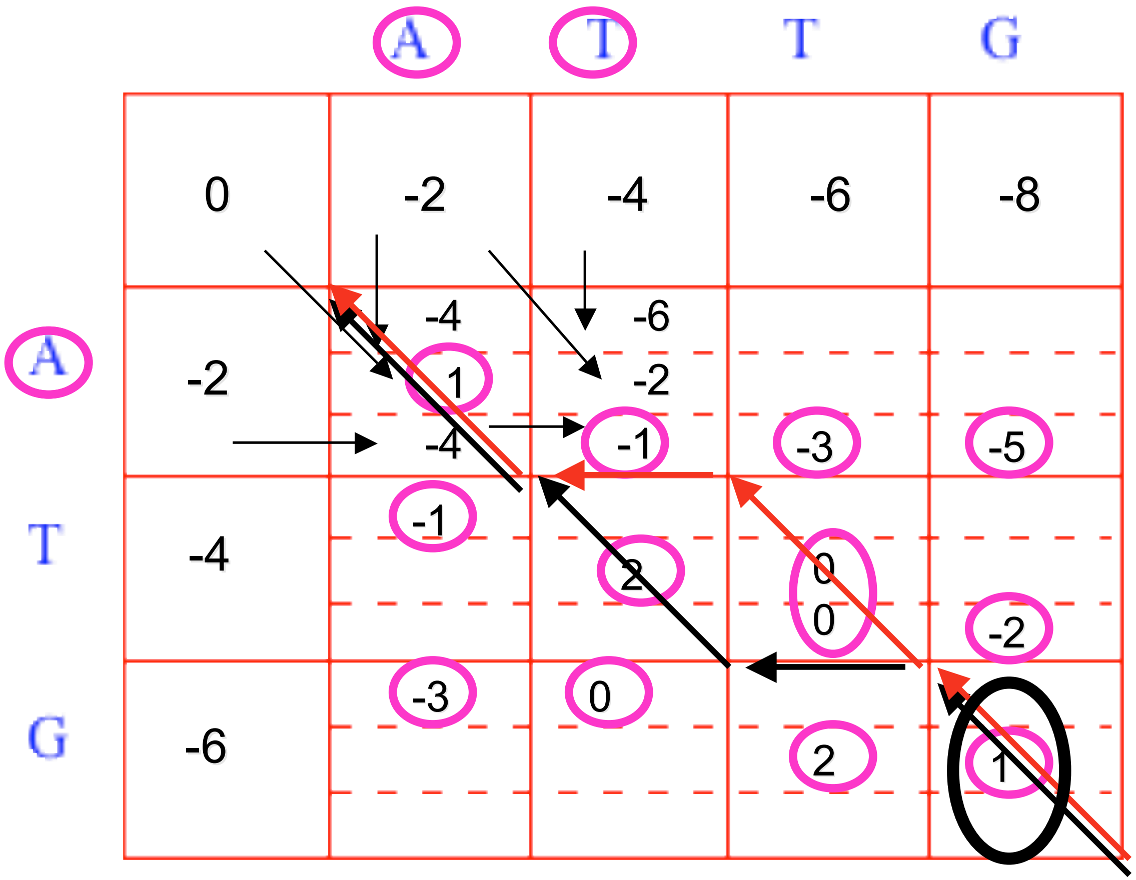
```
A T C C G
A T C - G
```

Recursive computation of the matrix

- Needleman & Wunsh, 1970



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



		A	T	T	G
	0	-2	-4	-6	-8
A	-2	-4 1 -4	-6 -2 -1	-3	-5
T	-4	-1	2	0 0	-2
G	-6	-3	0	2	1

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

Dynamic programming: time and memory requirements

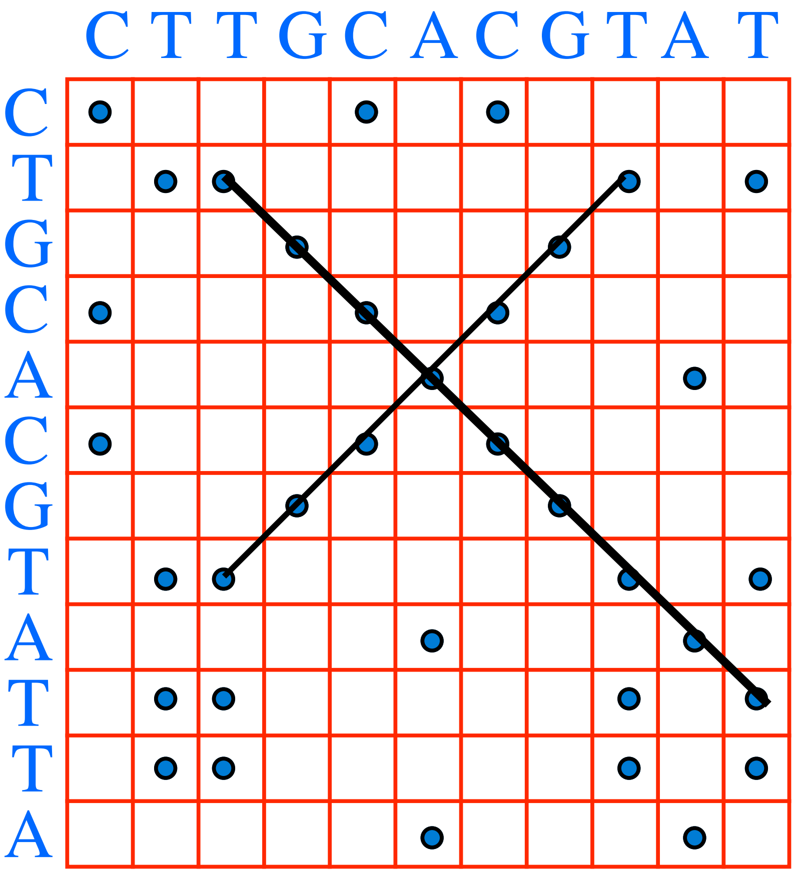
- Alignment of two sequences of length M and N
- Needleman-Wunsh (global alignments), Smith-Waterman (local alignments):
  - Time: O(N . M)
  - Memory: O(N . M)

- Improvement of Smith-Waterman (Huang & Miller 1991):
  - Time: O(N . M)
  - Memory: O($N + M$)

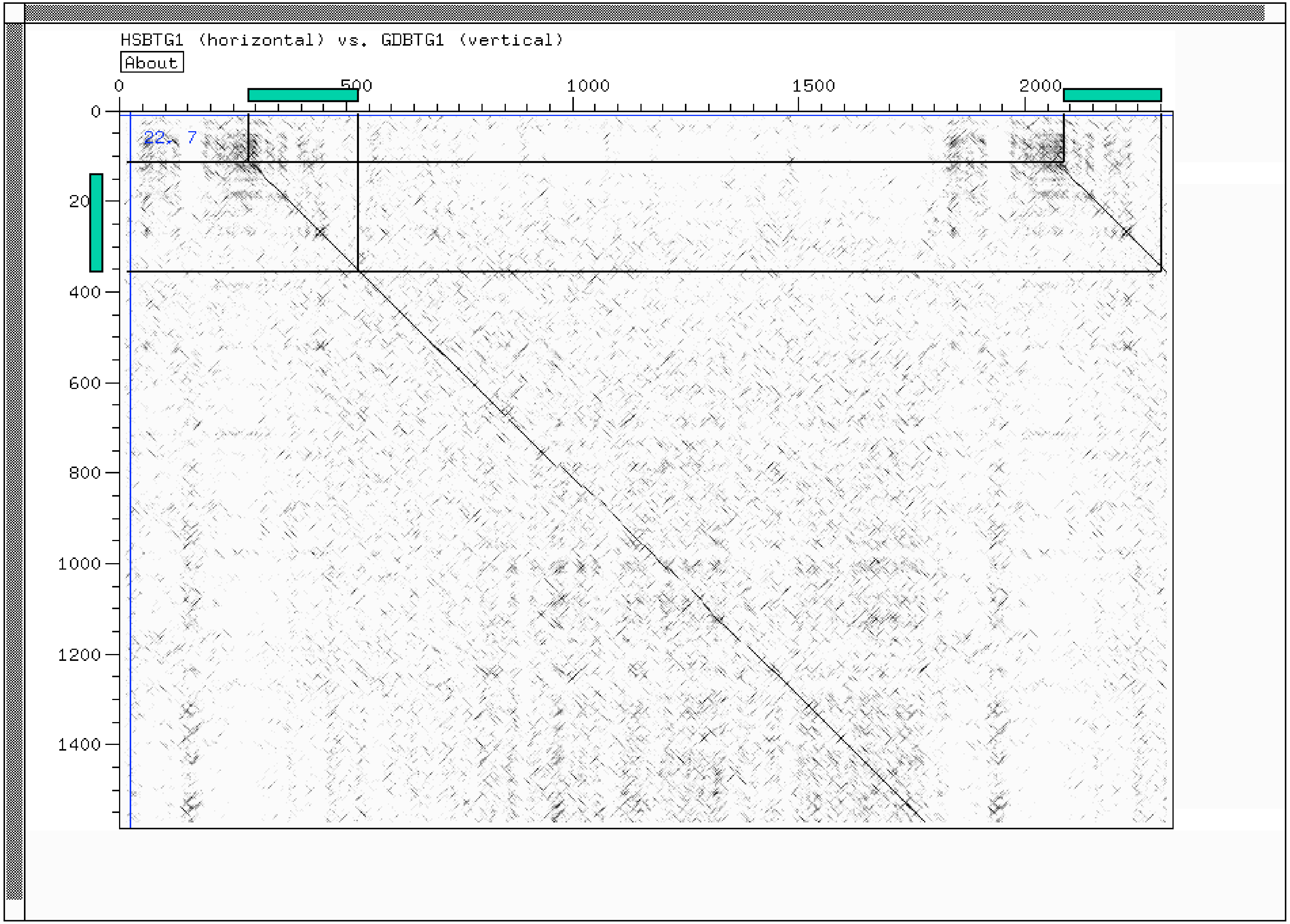
- SIM, LALIGN

Dot Plot

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



	C	T	T	G	C	A	C	G	T	A	T
C	●				●		●				
T		●	●						●		●
G				●				●			
C	●				●		●				
A						●				●	
C	●				●		●				
G				●				●			
T		●	●						●		●
A						●				●	
T		●	●						●		●
T		●	●						●		●
A						●				●	



HSBTG1 (horizontal) vs. GDBTG1 (vertical)

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 ’

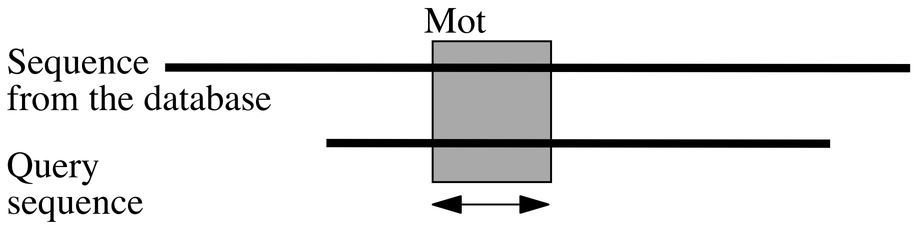
    2 - global alignment, anchored on the region of similarity
  - BLAST

    1 - search for similar ‘words’

    2 - extend blocks of similarity

BLAST

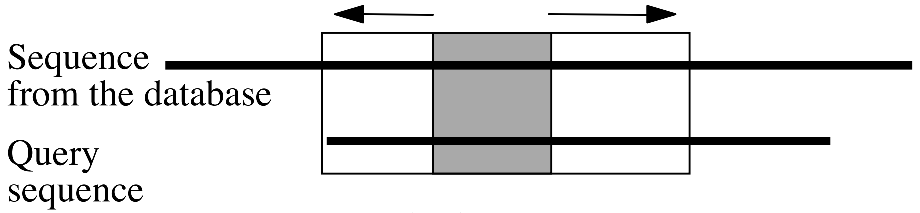
Step 1: detect similar ‘words’



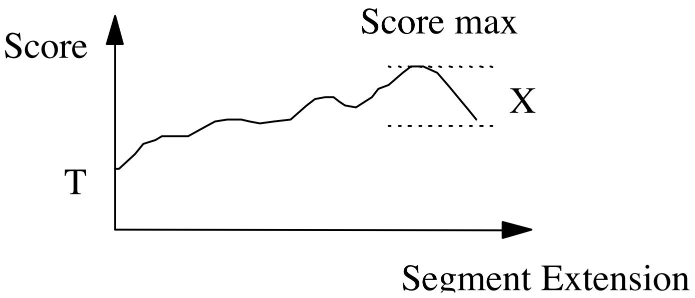
Word length = W

Score ≥ T (Threshold score)

Step 2: extend blocks of similarity



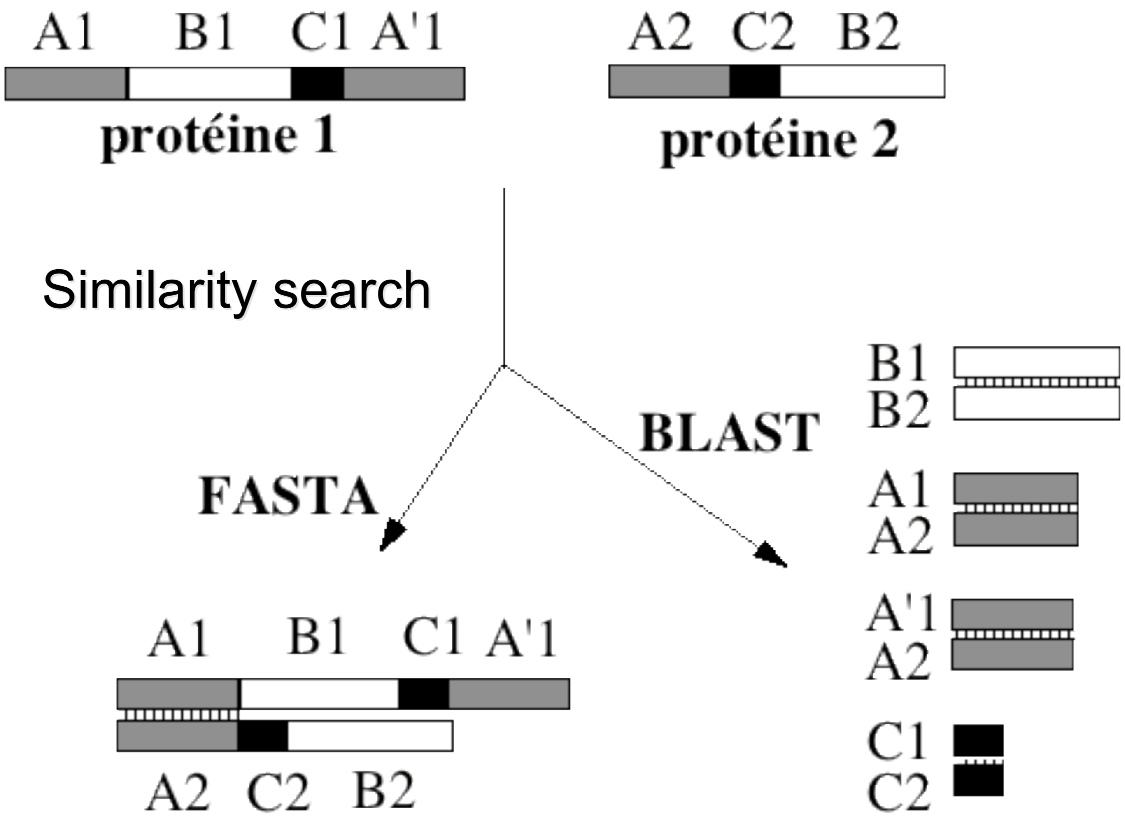
HSP: high score segment pair)



Stop extension if:

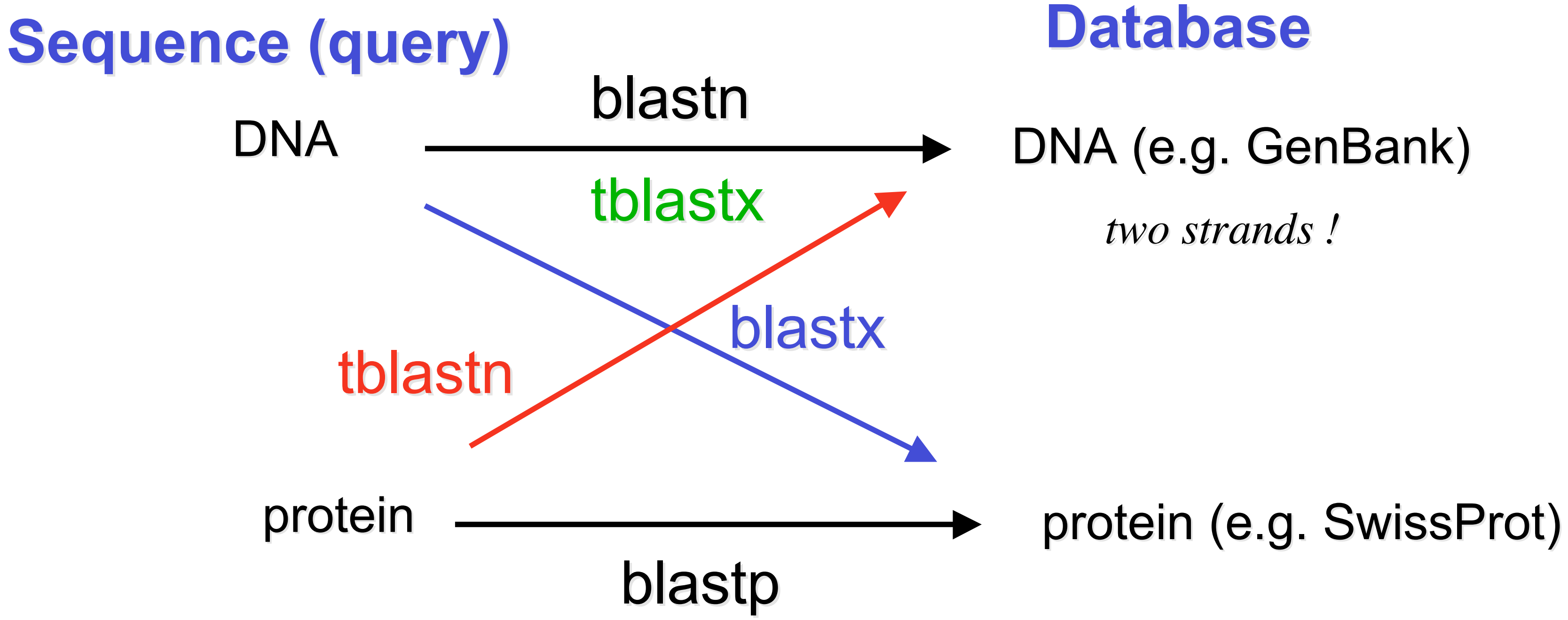
- reach the end of a sequence
- score ≤ 0
- score ≤ score_max - X

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, ...



Sequence (query)

Database

DNA

blastn

DNA (e.g. GenBank)

tblastx

two strands !

blastx

tblastn

protein

protein (e.g. SwissProt)

blastp

Different versions of BLAST for different problems

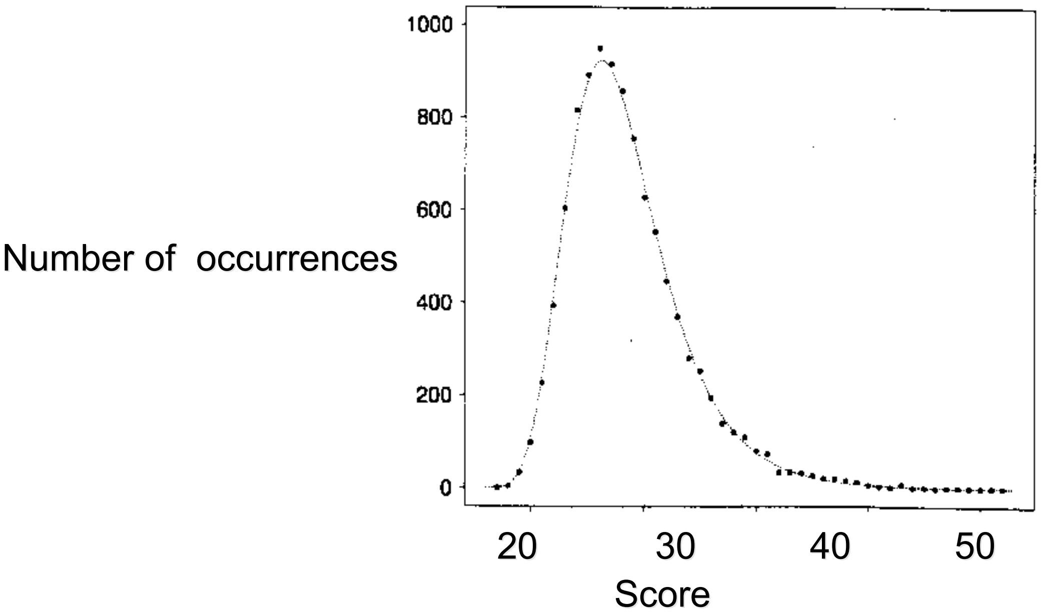
- blastp: protein/protein
- blastn: DNA/DNA (useful for non-coding sequences)
- blastx: DNA-translated/protein (useful for query sequences with unidentifed coding regions; more sensitive than blastn)
- tblastn: protein/DNA-translated (useful for database sequences with unidentifed coding regions; e.g. 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

                                                        Score  E
  Sequences producing significant alignments:           (bits) 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

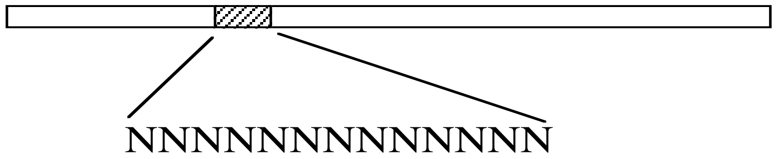
Filtering programs: SEG, XNU, DUST

```
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

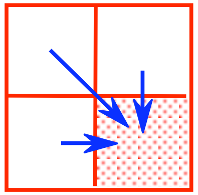
- Search for similarities with very short DNA sequences (e.g. PCR primers):
  - decrease W (11 → 7)

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

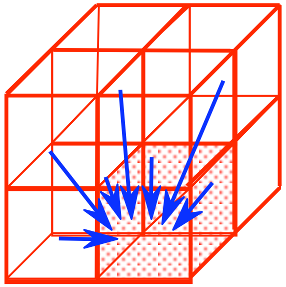
Multiple sequence alignment

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

⇒ Use **heuristic methods**.

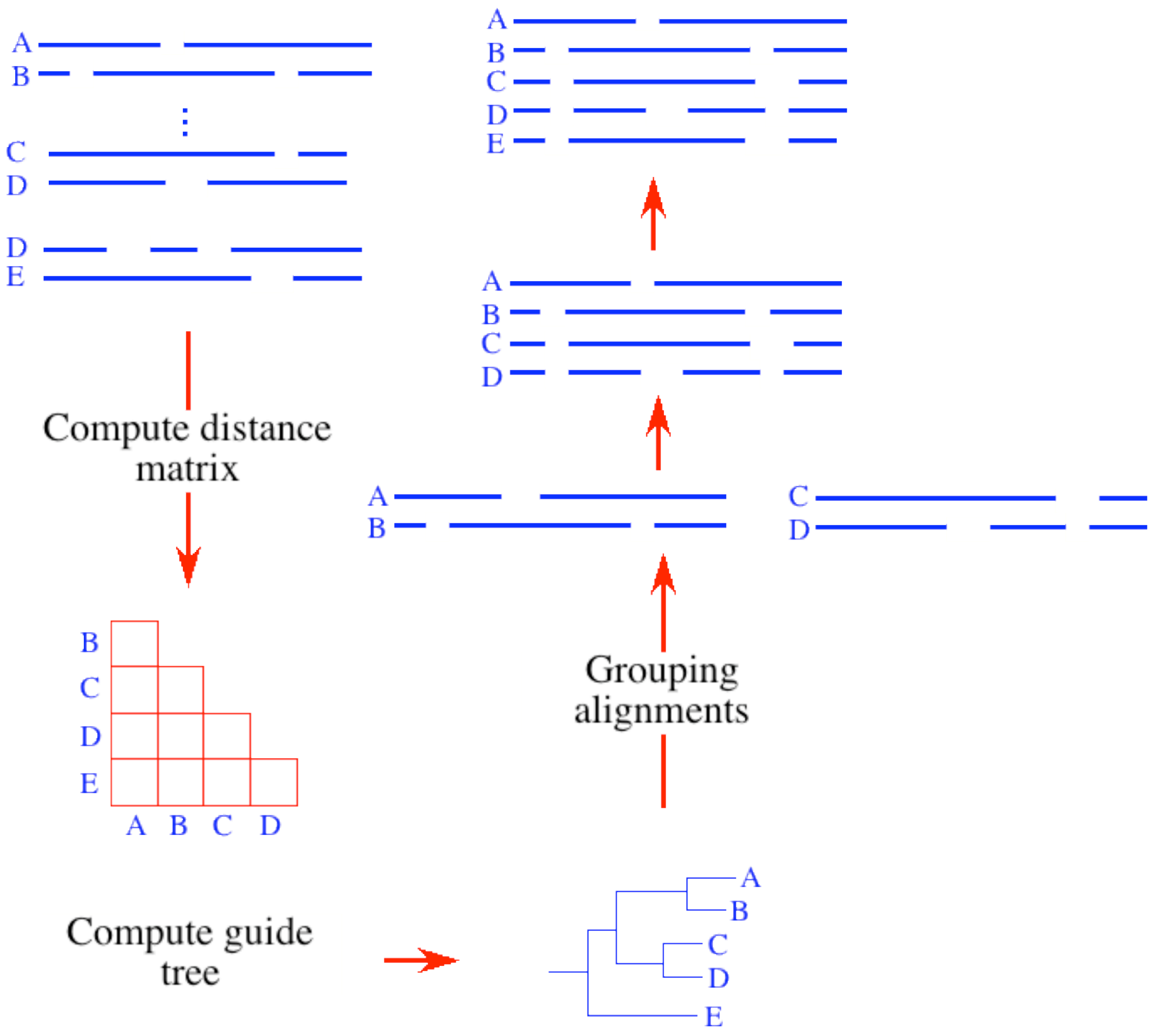
Progressive Alignment

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

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

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

- MULTALIN, PILEUP, T-Coffee, Muscle



Compute distance matrix

Compute guide tree

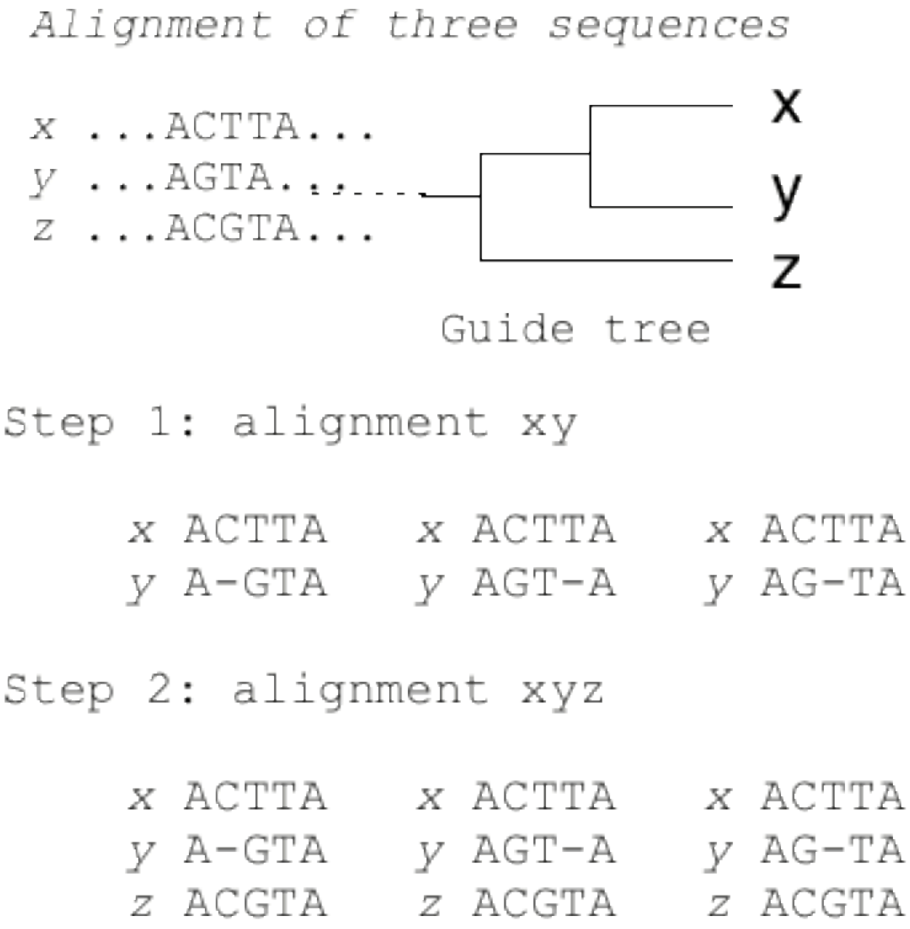
Grouping alignments

Position specific gap penalty

λ Decrease gap penalty in **hydrophilic** regions (≥ 5 residues).

λ 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

T-Coffee

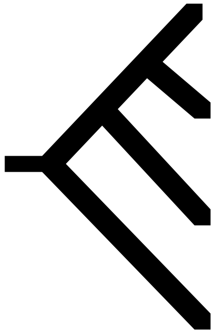
Notredame, Higgins, Heringa (2000) JMB 302:205

Pairwise Alignments

```
SeqA GARFIELD THE LAST FAT CAT
SeqB GARFIELD THE FAST CAT ---

SeqA GARFIELD THE LAST FA-T CAT
SeqC GARFIELD THE VERY FAST CAT

SeqA GARFIELD THE LAST FAT CAT
SeqD ---------THE ---- FAT CAT
```

```
SeqB GARFIELD THE ---- FAST CAT
SeqC GARFIELD THE VERY FAST CAT

SeqB GARFIELD THE FAST CAT
SeqD ---------THE FA-T CAT

SeqC GARFIELD THE VERY FAST CAT
SeqD ---------THE ---- FA-T CAT
```

Progressive Alignment



```
SeqA GARFIELD THE LAST FAT CAT

SeqB GARFIELD THE FAST CAT

SeqC GARFIELD THE VERY FAST CAT

SeqD THE FAT CAT
```

→

```
SeqA GARFIELD THE LAST FA-T CAT
SeqB GARFIELD THE FAST CA-T ---
SeqC GARFIELD THE VERY FAST CAT
SeqD ---------THE ---- FA-T CAT
```

T-Coffee

Notredame, Higgins, Heringa (2000) JMB 302:205

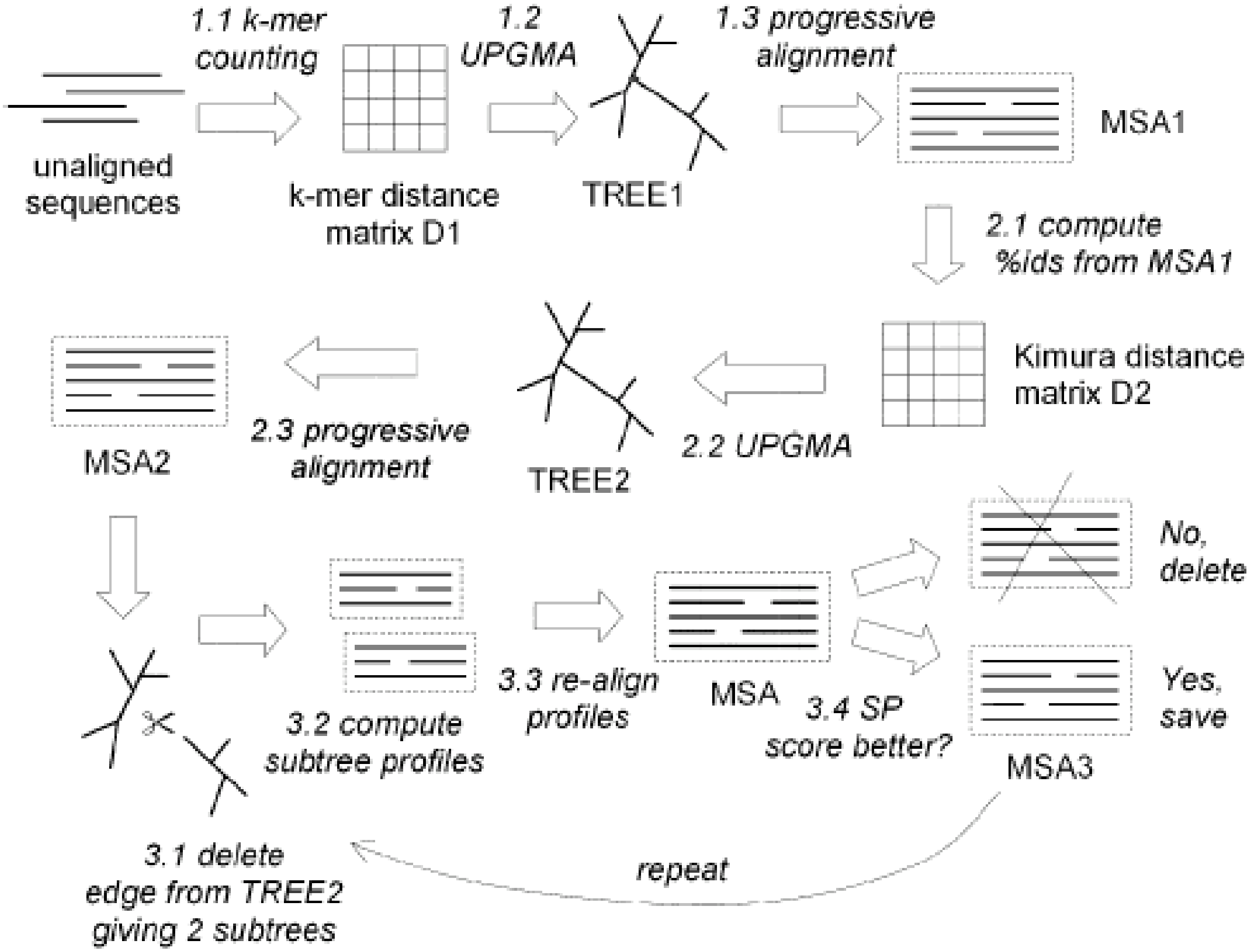
http://igs-server.cnrs-mrs.fr/~cnotred/

- Progressive Alignment
- during the progressive alignment, takes into account all pairwise alignements
- Possibility to introduce other informations (structure, etc.)

Muscle

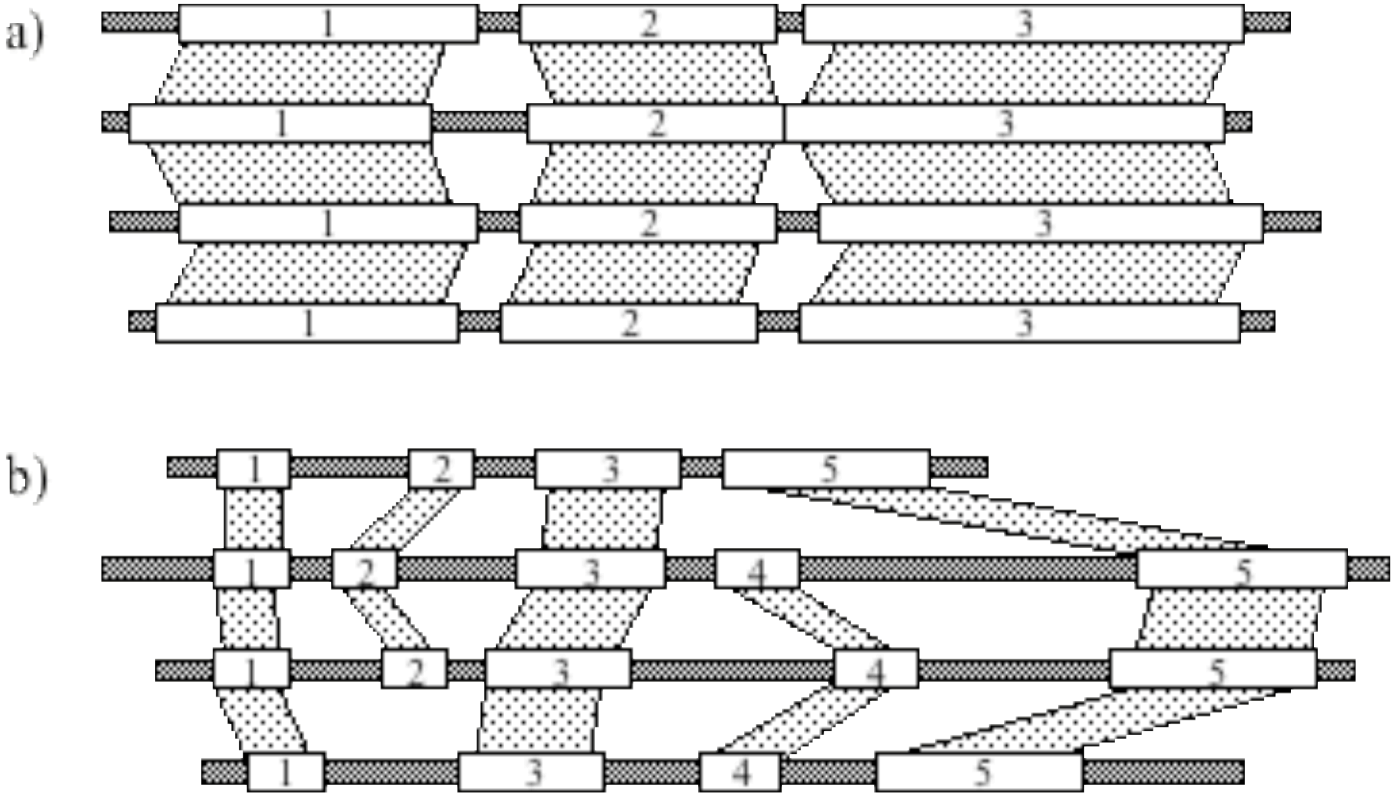
Edgar (2004) Nucleic Acids Res. 32:1792

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



1.1 k-mer counting
1.2 UPGMA
1.3 progressive alignment
unaligned sequences
k-mer distance matrix D1
TREE1
MSA1
2.1 compute %ids from MSA1
Kimura distance matrix D2
2.2 UPGMA
TREE2
2.3 progressive alignment
MSA2
No, delete
MSA
3.3 re-align profiles
3.2 compute subtree profiles
3.4 SP score better?
Yes, save
MSA3
3.1 delete edge from TREE2 giving 2 subtrees
repeat

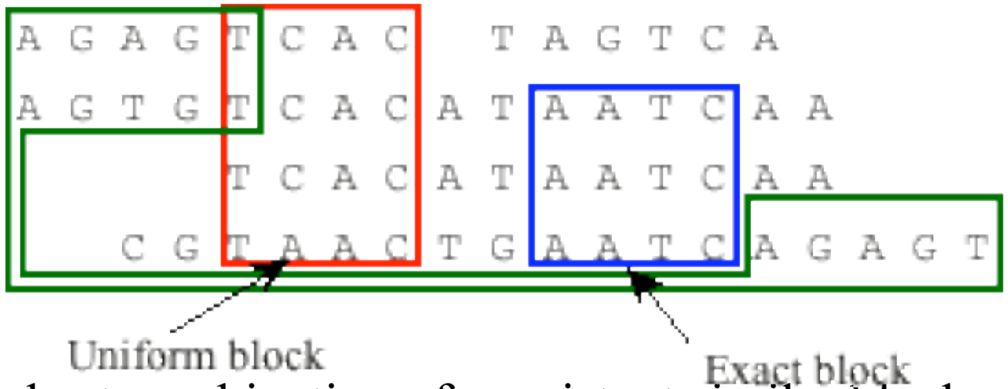
Global Alignments, Block alignments



Dialign

Morgenstern et al. 1996 PNAS 93:12098

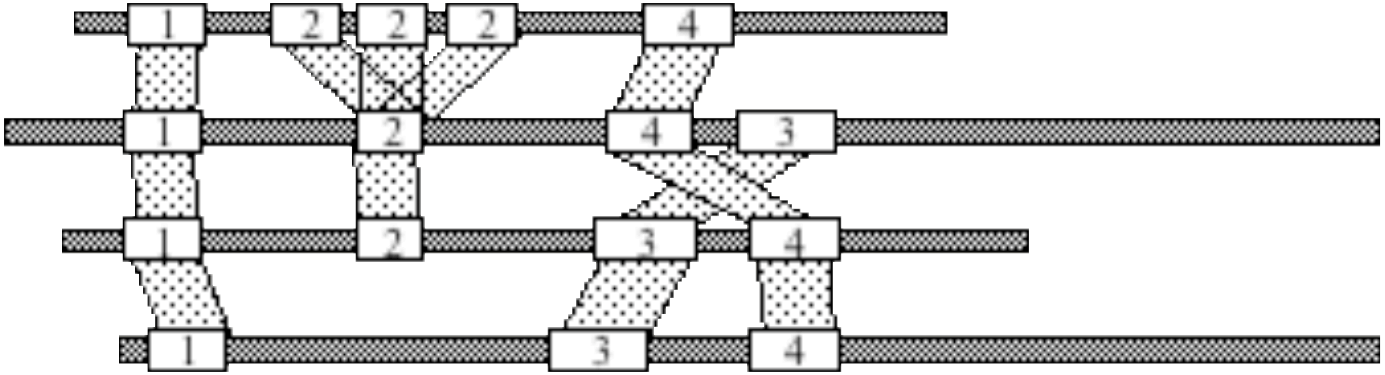
- Search for similar blocks without gap



```
A G A G T C A C   T A G T C A
A G T G T C A C A T A A T C A A
        T C A C A T A A T C A A
    C G T A A C T G A A T C A G A G T
```

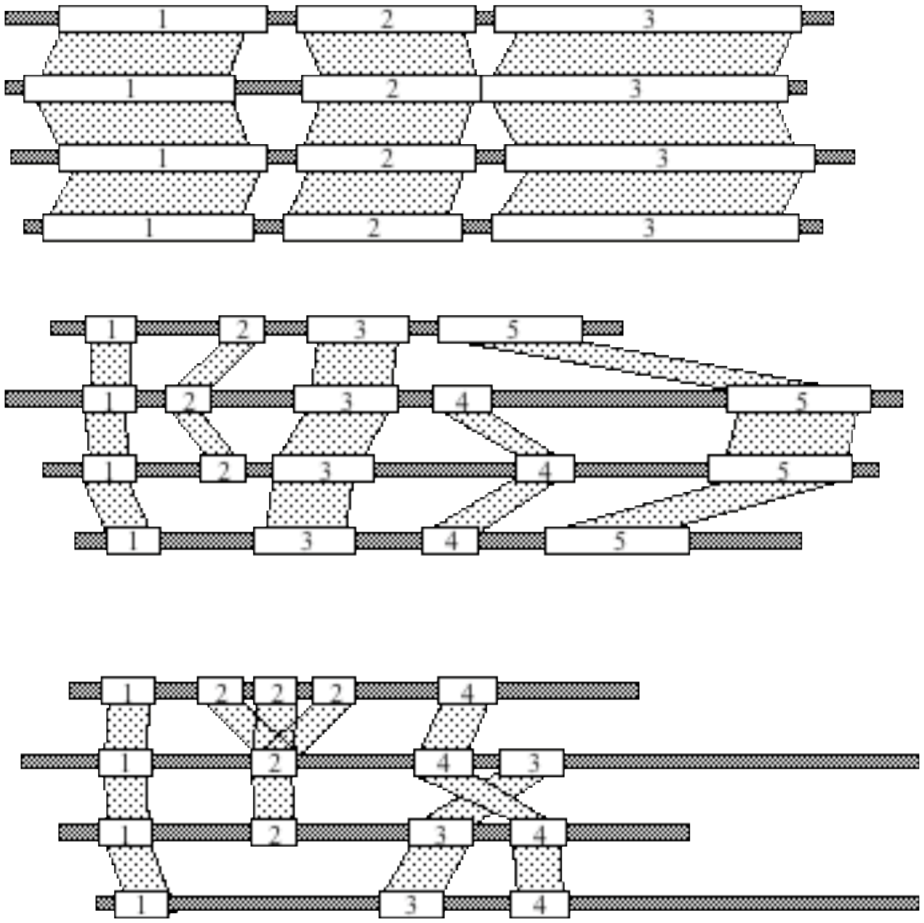
- Select the best combination of consistent similar blocks (uniforms or not) : heuristic (Abdeddaim 1997)
- Alignment anchored on blocks
- Slower than progressive alignment, but better when sequences contain large indels
- Do not try to align non-conserved regions

Local Multiple Alignments



- MEME
- MATCH-BOX
- PIMA

Overview

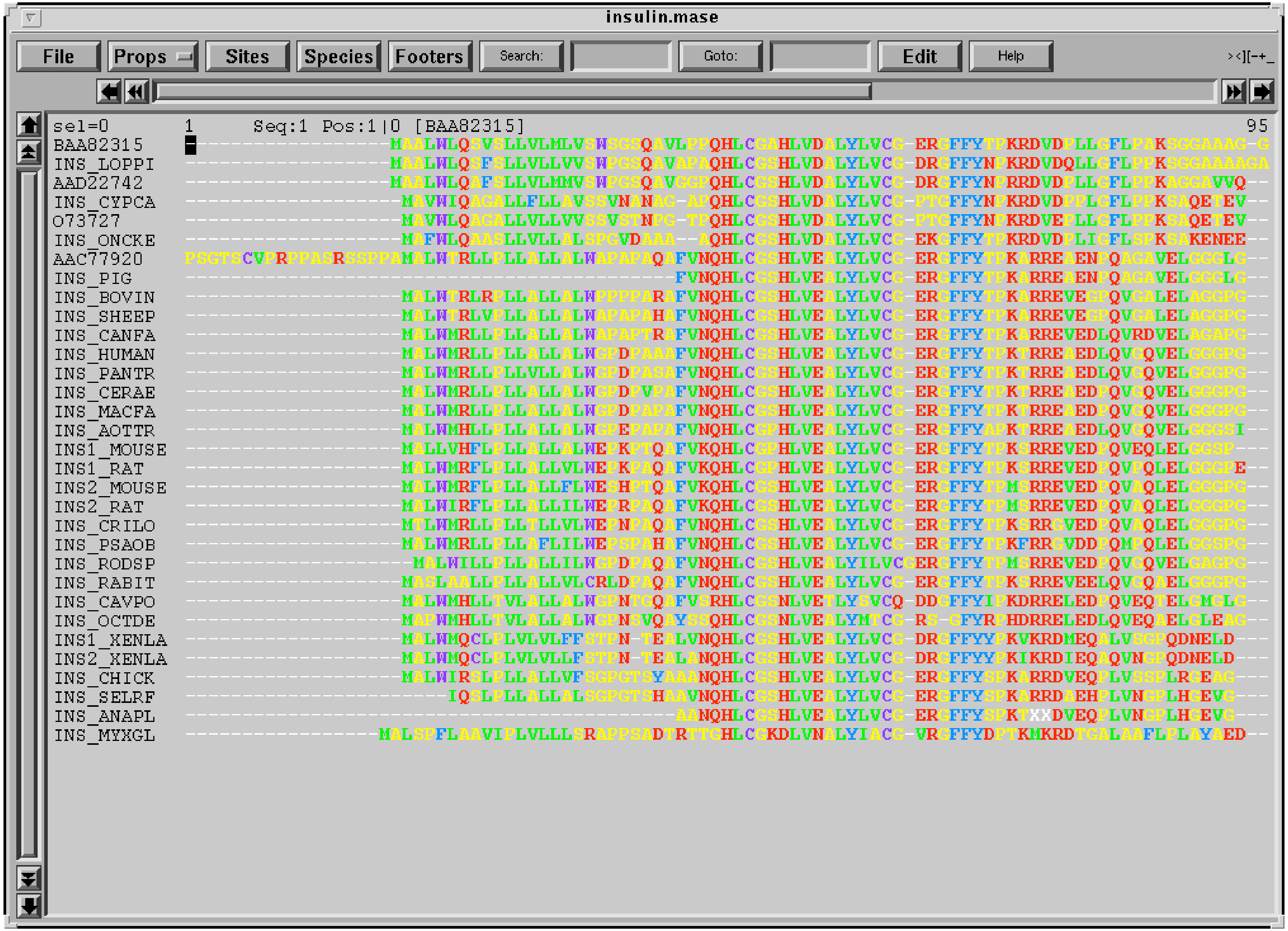


- ClustalW

- Muscle
- Dialign
- T-coffee

- MEME

Multiple alignment editor



```
sel=0         1     Seq:1 Pos:1|0 [BAA82315]                                                   95
BAA82315      ------------------MAALWLQSVSLLVLMLVSWSGSQAVLPPQHLCGAHLVDALYLVCG-ERGFFYTPKRDVDPLLGFLPAKSGGAAAG-G
INS_LOPPI     ------------------MAALWLQSFSLLVLLVVSWPGSQAVAPAQHLCGSHLVDALYLVCG-DRGFFYNPKRDVDQLLGFLPPKSGGAAAAGA
AAD22742      ------------------MAALWLQAFSLLVLMMVSWPGSQAVGGPQHLCGSHLVDALYLVCG-DRGFFYNPRRDVDPLLGFLPPKAGGAVVQ--
INS_CYPCA     -------------------MAVWIQAGALLFLLAVSSVNANAG-APQHLCGSHLVDALYLVCG-PTGFFYNPKRDVDPPLGFLPPKSAQETEV--
O73727        -------------------MAVWLQAGALLVLLVVSSVSTNPG-TPQHLCGSHLVDALYLVCG-PTGFFYNPKRDVEPLLGFLPPKSAQETEV--
INS_ONCKE     -------------------MAFWLQAASLLVLLALSPGVDAAA--AQHLCGSHLVDALYLVCG-EKGFFYTPKRDVDPLIGFLSPKSAKENEE--
AAC77920      PSGTSCVPRPPASRSSPPAMALWTRLLPLLALLALWAPAPAQAFVNQHLCGSHLVEALYLVCG-ERGFFYTPKARREAENPQAGAVELGGGLG--
INS_PIG       -------------------------------------------FVNQHLCGSHLVEALYLVCG-ERGFFYTPKARREAENPQAGAVELGGGLG--
INS_BOVIN     -------------------MALWTRLRPLLALLALWPPPPARAFVNQHLCGSHLVEALYLVCG-ERGFFYTPKARREVEGPQVGALELAGGPG--
INS_SHEEP     -------------------MALWTRLVPLLALLALWAPAPAHAFVNQHLCGSHLVEALYLVCG-ERGFFYTPKARREVEGPQVGALELAGGPG--
INS_CANFA     -------------------MALWMRLLPLLALLALWAPAPTRAFVNQHLCGSHLVEALYLVCG-ERGFFYTPKARREVEDLQVRDVELAGAPG--
INS_HUMAN     -------------------MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCG-ERGFFYTPKTRREAEDLQVGQVELGGGPG--
INS_PANTR     -------------------MALWMRLLPLLVLLALWGPDPASAFVNQHLCGSHLVEALYLVCG-ERGFFYTPKTRREAEDLQVGQVELGGGPG--
INS_CERAE     -------------------MALWMRLLPLLALLALWGPDPVPAFVNQHLCGSHLVEALYLVCG-ERGFFYTPKTRREAEDPQVGQVELGGGPG--
INS_MACFA     -------------------MALWMRLLPLLALLALWGPDPAPAFVNQHLCGSHLVEALYLVCG-ERGFFYTPKTRREAEDPQVGQVELGGGPG--
INS_AOTTR     -------------------MALWMHLLPLLALLALWGPEPAPAFVNQHLCGPHLVEALYLVCG-ERGFFYAPKTRREAEDLQVGQVELGGGSI--
INS1_MOUSE    -------------------MALLVHFLPLLALLALWEPKPTQAFVKQHLCGPHLVEALYLVCG-ERGFFYTPKSRREVEDPQVEQLELGGSP---
INS1_RAT      -------------------MALWMRFLPLLALLVLWEPKPAQAFVKQHLCGPHLVEALYLVCG-ERGFFYTPKSRREVEDPQVPQLELGGGPE--
INS2_MOUSE    -------------------MALWMRFLPLLALLFLWESHPTQAFVKQHLCGSHLVEALYLVCG-ERGFFYTPMSRREVEDPQVAQLELGGGPG--
INS2_RAT      -------------------MALWIRFLPLLALLILWEPRPAQAFVKQHLCGSHLVEALYLVCG-ERGFFYTPMSRREVEDPQVAQLELGGGPG--
INS_CRILO     -------------------MTLWMRLLPLLTLLVLWEPNPAQAFVNQHLCGSHLVEALYLVCG-ERGFFYTPKSRRGVEDPQVAQLELGGGPG--
INS_PSAOB     -------------------MALWMRLLPLLAFLILWEPSPAHAFVNQHLCGSHLVEALYLVCG-ERGFFYTPKFRRGVDDPQMPQLELGGSPG--
INS_RODSP     --------------------MALWILLPLLALLILWGPDPAQAFVNQHLCGSHLVEALYILVCGERGFFYTPMSRREVEDPQVGQVELGAGPG--
INS_RABIT     -------------------MASLAALLPLLALLVLCRLDPAQAFVNQHLCGSHLVEALYLVCG-ERGFFYTPKSRREVEELQVGQAELGGGPG--
INS_CAVPO     -------------------MALWMHLLTVLALLALWGPNTGQAFVSRHLCGSNLVETLYSVCQ-DDGFFYIPKDRRELEDPQVEQTELGMGLG--
INS_OCTDE     -------------------MAPWMHLLTVLALLALWGPNSVQAYSSQHLCGSNLVEALYMTCG-RS-GFYRPHDRRELEDLQVEQAELGLEAG--
INS1_XENLA    -------------------MALWMQCLPLVLVLFFSTPN-TEALVNQHLCGSHLVEALYLVCG-DRGFFYYPKVKRDMEQALVSGPQDNELD---
INS2_XENLA    -------------------MALWMQCLPLVLVLLFSTPN-TEALANQHLCGSHLVEALYLVCG-DRGFFYYPKIKRDIEQAQVNGPQDNELD---
INS_CHICK     -------------------MALWIRSLPLLALLVFSGPGTSYAAANQHLCGSHLVEALYLVCG-ERGFFYSPKARRDVEQPLVSSPLRGEAG---
INS_SELRF     -----------------------IQSLPLLALLALSGPGTSHAAVNQHLCGSHLVEALYLVCG-ERGFFYSPKARRDAEHPLVNGPLHGEVG---
INS_ANAPL     -------------------------------------------AANQHLCGSHLVEALYLVCG-ERGFFYSPKTXXDVEQPLVNGPLHGEVG---
INS_MYXGL     -----------------MALSPFLAAVIPLVLLLSRAPPSADTRTTGHLCGKDLVNALYIACG-VRGFFYDPTKMKRDTGALAAFLPLAYAED--
```

Some special cases of sequence alignments

Alignment of protein-coding DNA sequences

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

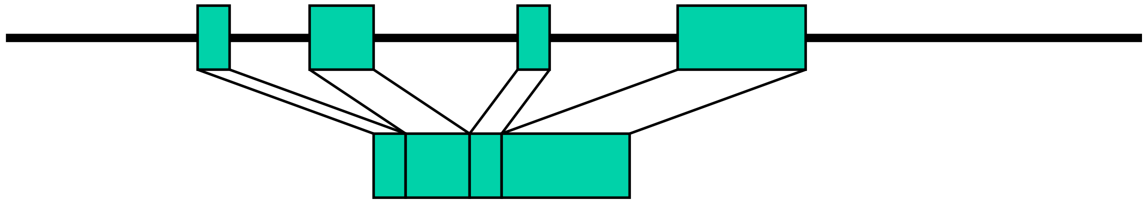
(1) alignment of protein sequences

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

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

Spliced alignment (1)

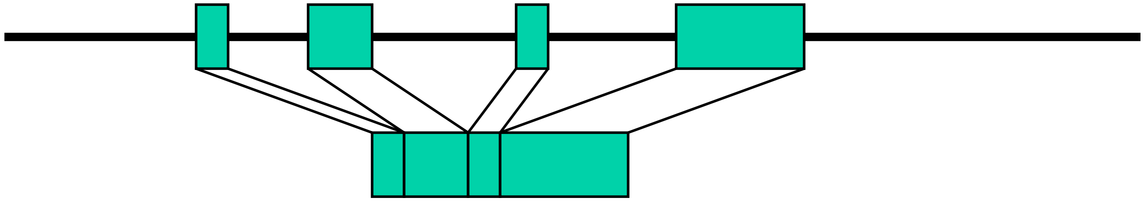
- Align an mRNA with its cognate genomic DNA => gene finding



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

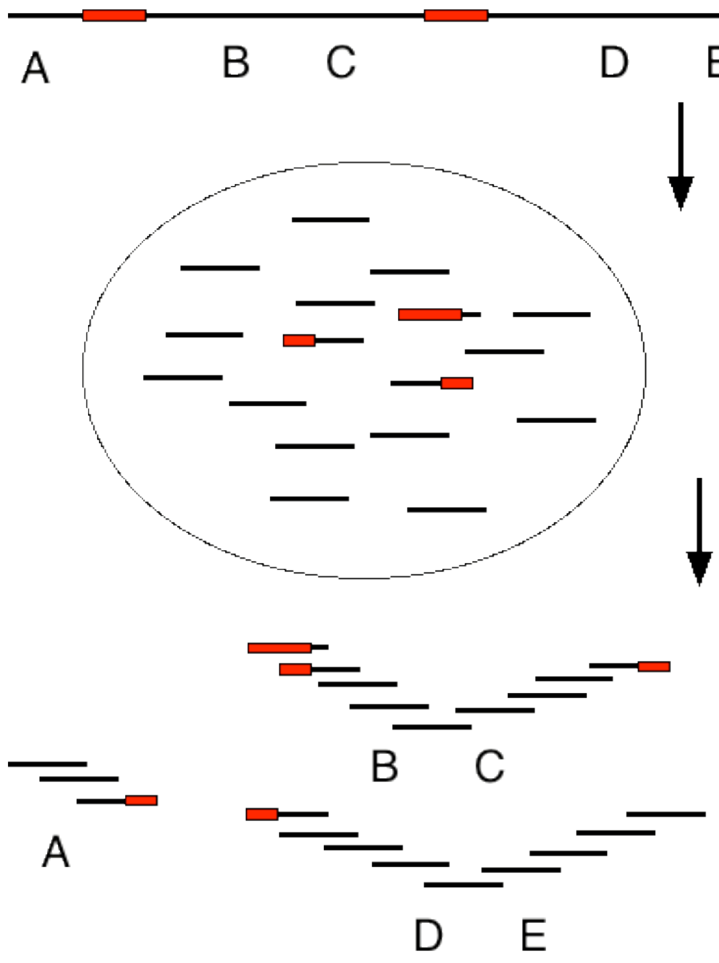
Spliced alignment (2)

- Align a protein with genomic DNA => gene prediction



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

Shotgun sequencing



Genome

Fragmentation (±2kb), cloning

Genomic DNA library

Sequencing (500 to 1000 bp)
Assembly

Contigs

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	0	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: 83

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	0	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. CLUSTAL, 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