ASAB – Weeks 8 and 9 BLAST and Multiple sequence alignment

Algorithms for Sequence Analysis in Bioinformatics

Arnau Cordomí
arnau.cordomi@esci.upf.edu

BLAST
Basic Local Alignment Search Tool

Global vs. local alignment

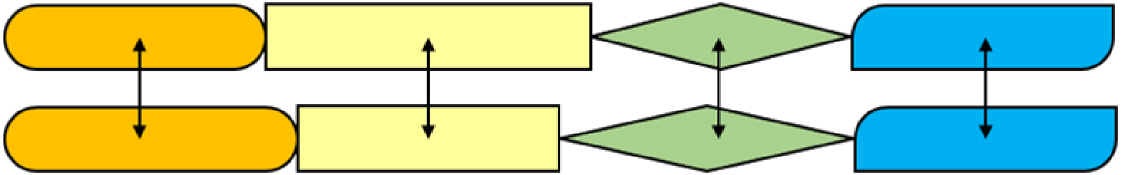
GLOBAL ALIGNMENT

Needleman & Wunsch

Align all letters

Seq1

Seq2



THEFASTFASTCAT
THEFATFASTRAT

THEFASTFASTCAT
THEFA-TFASTRAT

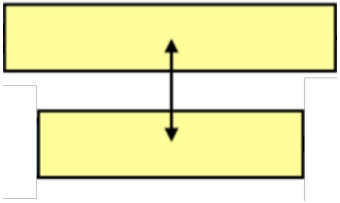
Smith & Waterman

LOCAL ALIGNMENT

Align some letters (one domain)

Seq1

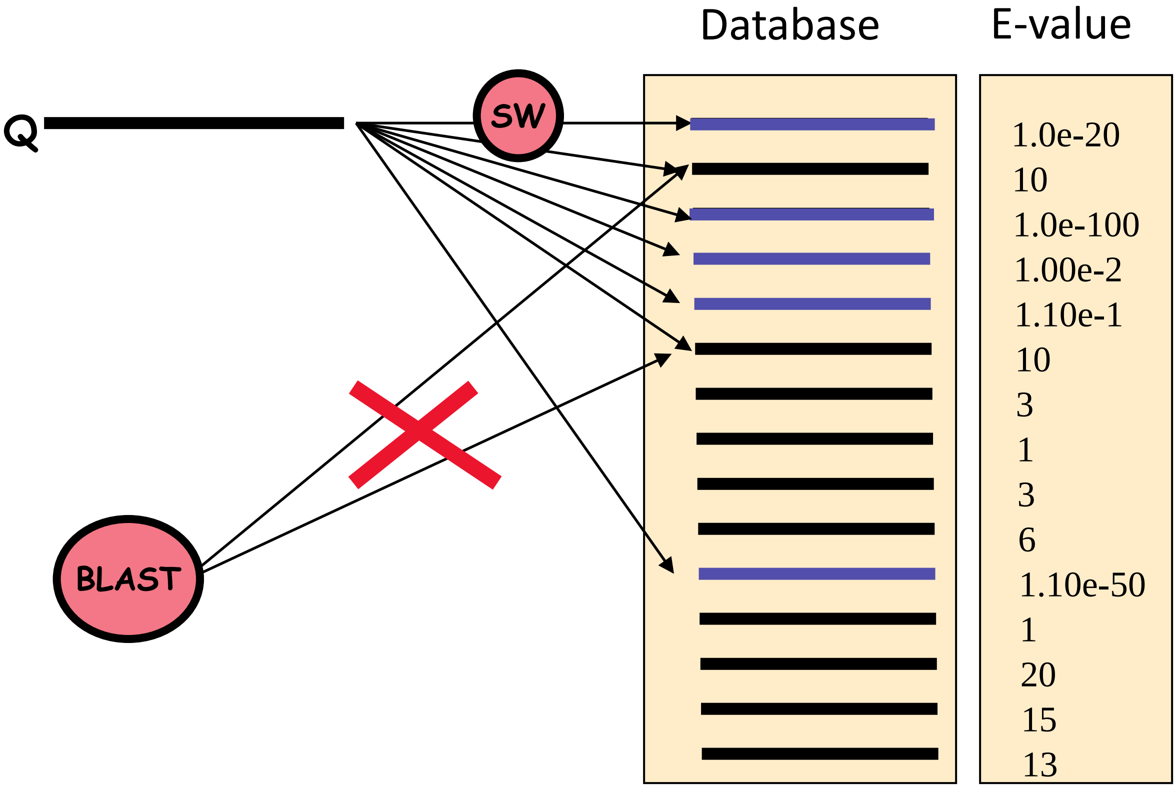
Seq2



AVEYRYFASTCAT
AFATNICERAT

FAST
FA-T

BLAST: main idea



Problem: local alignment (SW) is too slow

A bit of history

Smith and Waterman, 1981

- Exact Local Dynamic Programming.

Heuristic algorithms: Faster than the exact solution (SW), but without a guarantee of finding the best possible alignment.

FASTA: Lipman and Pearson, 1985

- Looks for similar words (k-tup, k: 1 to 6) on a diagonal.
- Comparison of the sequences one by one …

BLAST: Altschul *et al.*, 1990

- Faster and more accurate
- Powerful statistics

BLAST = 3 STEPS

1) Decide who will be compared (seqs with many interesting words)

This is where BLAST SAVES TIME

This is where it LOSES HITS

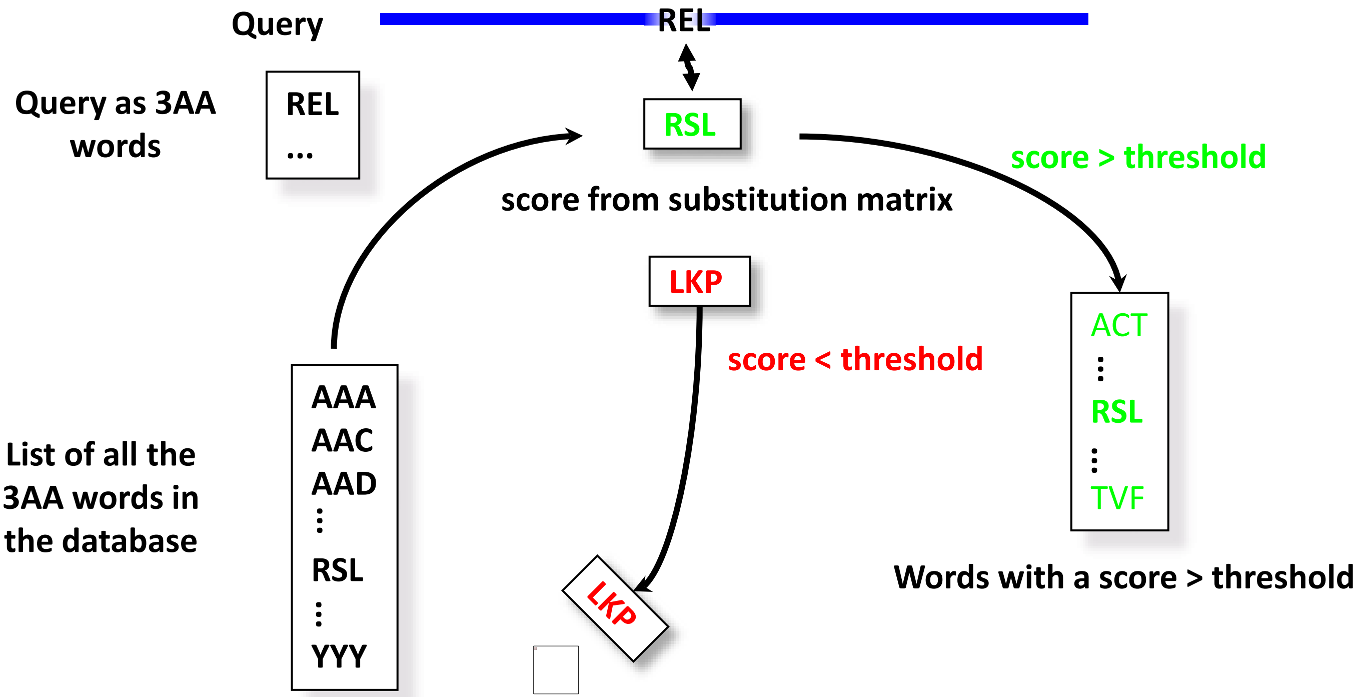
2) Check the most promising hits (SW)

3) Calculate the E-value of the protein hits.

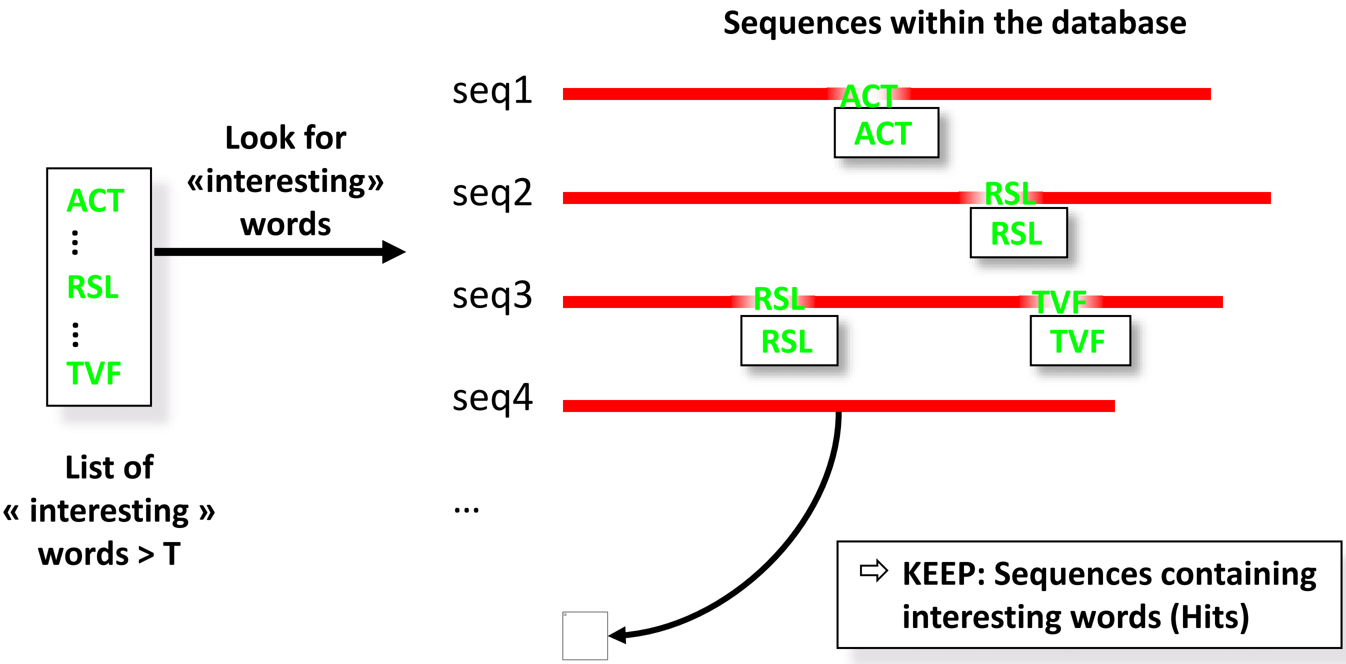
Finding the worthy words

We do not need to align our **query** to all the sequences in the database!

BLAST uses short "word" (*w*) segments to create alignment "seeds."



Discarding sequences



Sequences within the database

seq1

ACT

ACT

seq2

RSL

RSL

seq3

RSL

TVF

RSL

TVF

seq4

...

ACT

⋮

RSL

⋮

TVF

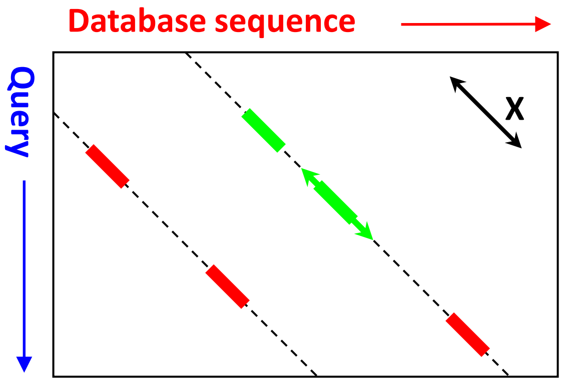
Look for «interesting» words

List of « interesting » words > T

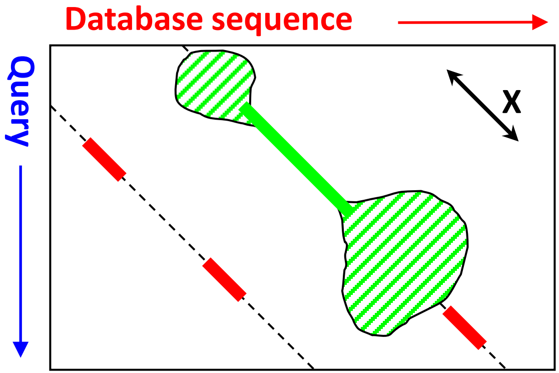
⇨ KEEP: Sequences containing interesting words (Hits)

Extension of the Hits

Merge close "Hits" on the same diagonal



Extension by limited Dynamic Programming



Speed vs. sensitivity

- BLAST increases the speed of alignment **by decreasing the search space** or number of comparisons it makes.
- The **sensitivity** and **speed** of the search are inversely related and controlled by the **word size** and **threshold**
- Larger **word sizes** provide faster search though at a higher risk of losing hits.
- Smaller **thresholds** allows detecting more word pairs and requires a longer processing time

A few definitions (I)

Query and Subject

Query: Your sequence

Subject: The database against which you search

Identity

Proportion of IDENTICAL residues between two sequences (excluding gaps?). Depends on the Alignment. Unit: the % id

Similarity

Proportion of SIMILAR residues

Two residues are similar if their substitution cost is higher than 0. **Depends on the matrix**. Unit: the %similarity

Homology

Sequences SIMILAR enough are sometimes HOMOLOGOUS

HOMOLOGY ⇔ COMMON ANCESTOR

Binary concept: Yes or No!

DIFFERENT sequences can also be Homologous

BLAST statistics (I)

Evaluation of the score

- Raw Score (S)
  - ⇨ Sum of the substitutions and gap penalties.
  - ⇨ Not very informative

- Bit Score (S')
  - ⇨ Evaluates the amount of information in the alignment
  - ⇨ Makes it possible to compare alignments

$$S' = \frac{\lambda S - \ln K}{\ln 2}$$

K: must be calibrated with the database composition

λ: is calibrated with the matrix being used

A few definitions (II)

Hit: A sequence that matches your sequence and is reported by BLAST.

E-Value: Expectation value. The number of times you would expect to find the hit by chance.

Depends on the alignment.

Depends on the matrix

Depends on the database

…

We typically consider HITS with an E-value < 0.0001

A good hit is something you would not expect by chance!

Derived Statistics

Significance via Gumbel extreme value distribution

- p-value
  - ⇨ Probability of finding an alignment with a score (S) at least as good as yours (x) by chance.
  - ⇨ The lower, the better

$$P = 1 - e^{-E}$$

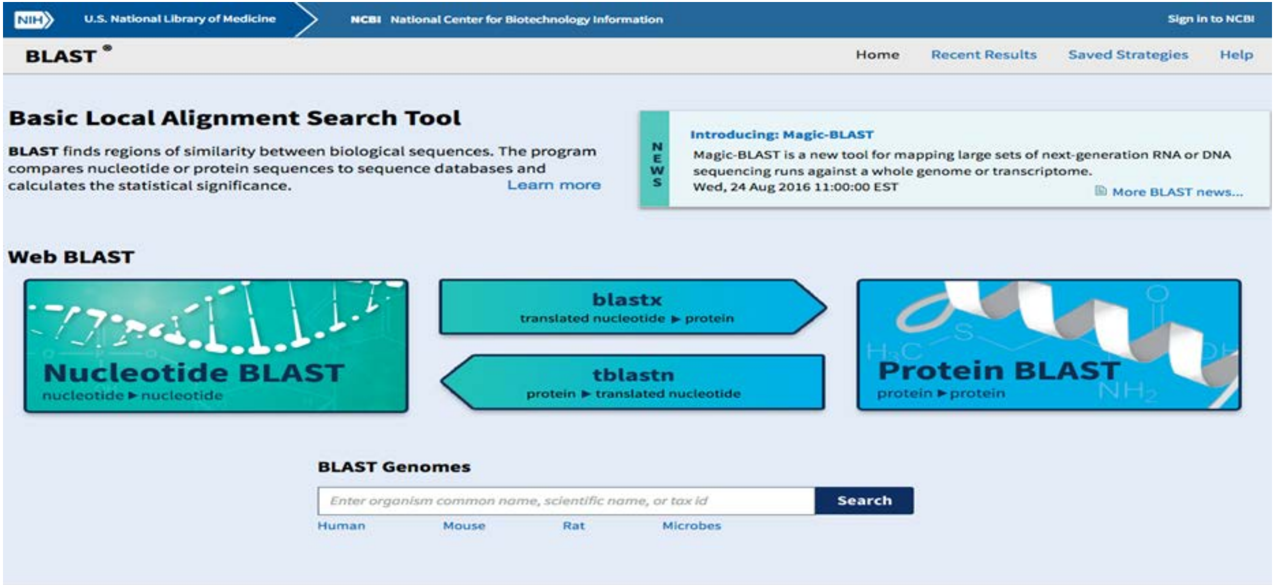
- E-value
  - ⇨The number of times you would expect to find the hit by chance.
  - ⇨The lower, the better: <0.00001

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

x: your obtained score
m: query length
n: database length
K: must be calibrated with the database composition
λ: is calibrated with the substitution matrix used

Basic Local Alignment Search Tool

BLAST is a program designed for rapidly comparing your sequence with every sequence in a database and reporting the most similar sequences



NIH U.S. National Library of Medicine NCBI National Center for Biotechnology Information Sign in to NCBI

BLAST® Home Recent Results Saved Strategies Help

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. Learn more

NEWS

Introducing: Magic-BLAST

Magic-BLAST is a new tool for mapping large sets of next-generation RNA or DNA sequencing runs against a whole genome or transcriptome.
Wed, 24 Aug 2016 11:00:00 EST

More BLAST news...

Web BLAST

Nucleotide BLAST
nucleotide ► nucleotide

blastx
translated nucleotide ► protein

tblastn
protein ► translated nucleotide

Protein BLAST
protein ► protein

BLAST Genomes

Enter organism common name, scientific name, or tax id | Search

Human Mouse Rat Microbes

http://blast.ncbi.nlm.nih.gov/Blast.cgi

BLAST main tools

Program	Query		Database
blastp	protein		proteins
blastn	nucleotide		nucleotide
blastx	nucleotide protein	vs	 protein
tblastn	 protein	vs	nucleotide protein
tblastx	nucleotide protein	vs	nucleotide protein

blastp:

predict protein **function**? (swissprot)

predict the 3d **structure**? (pdb)

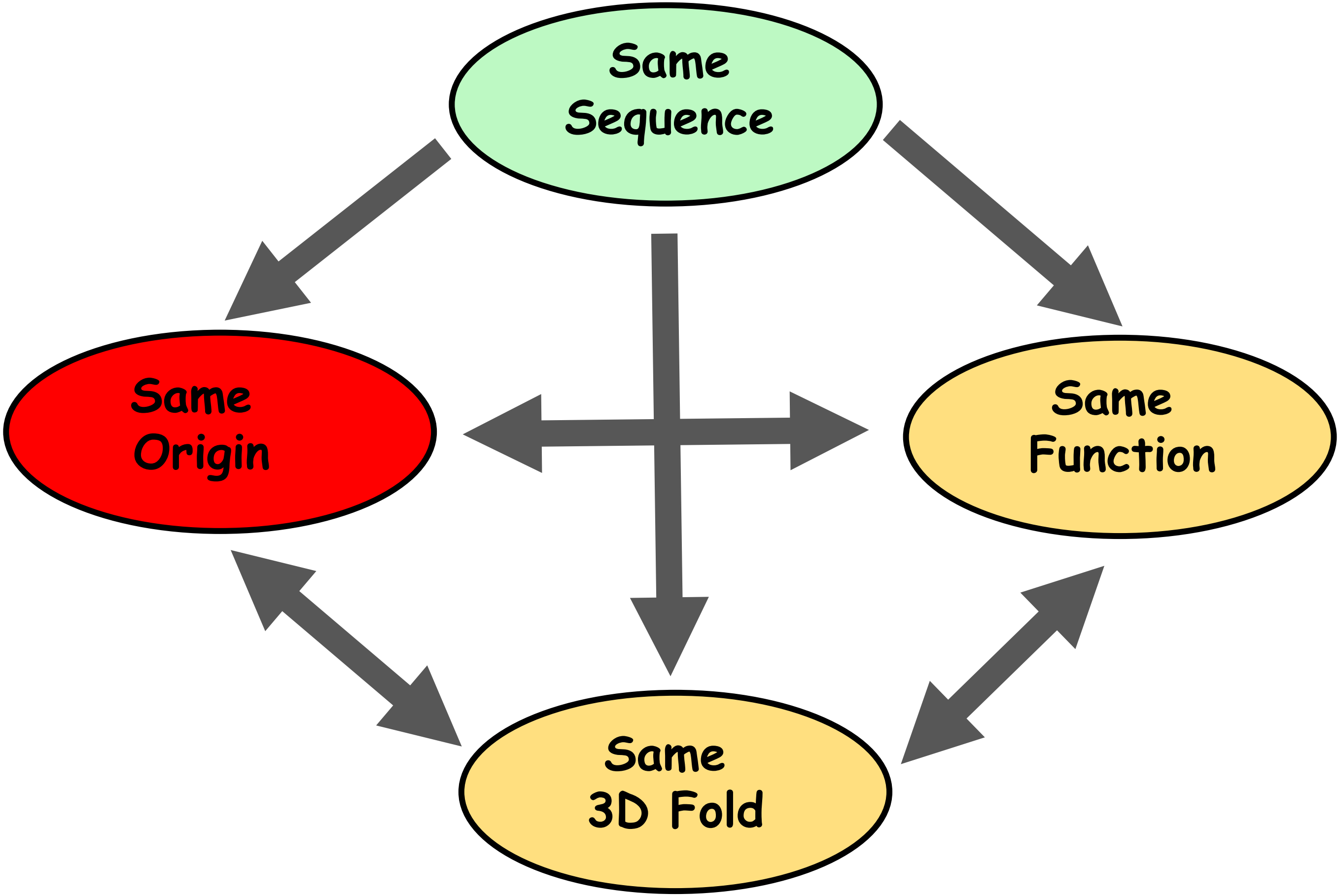
find **all family members**? (non-redundant)

blastn: Interested **in non-coding regions**

blastx: help **annotate coding regions** on a nucleotide sequence

tblastn: Identify **new genes** encoding proteins

tblastx: discover **new proteins**: (detect very distant relationships between nucleotide sequences; the slowest!)



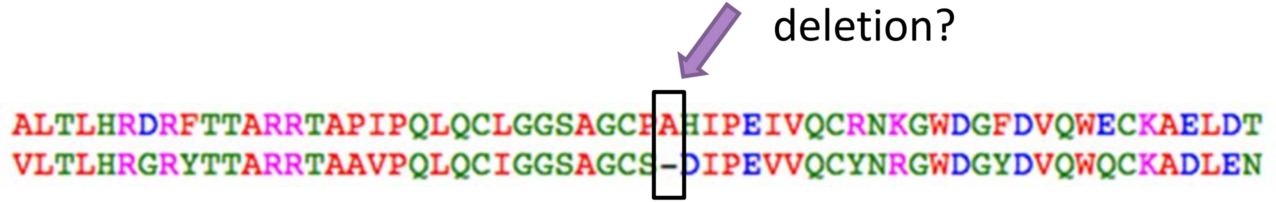
The money graph (by Cedric Notredame)

MSA
Multiple sequence alignment

Limitations of pairwise alignment

How to interpret a disagreement in a column in the pairwise sequence alignment?

Is this an insertion or a deletion?

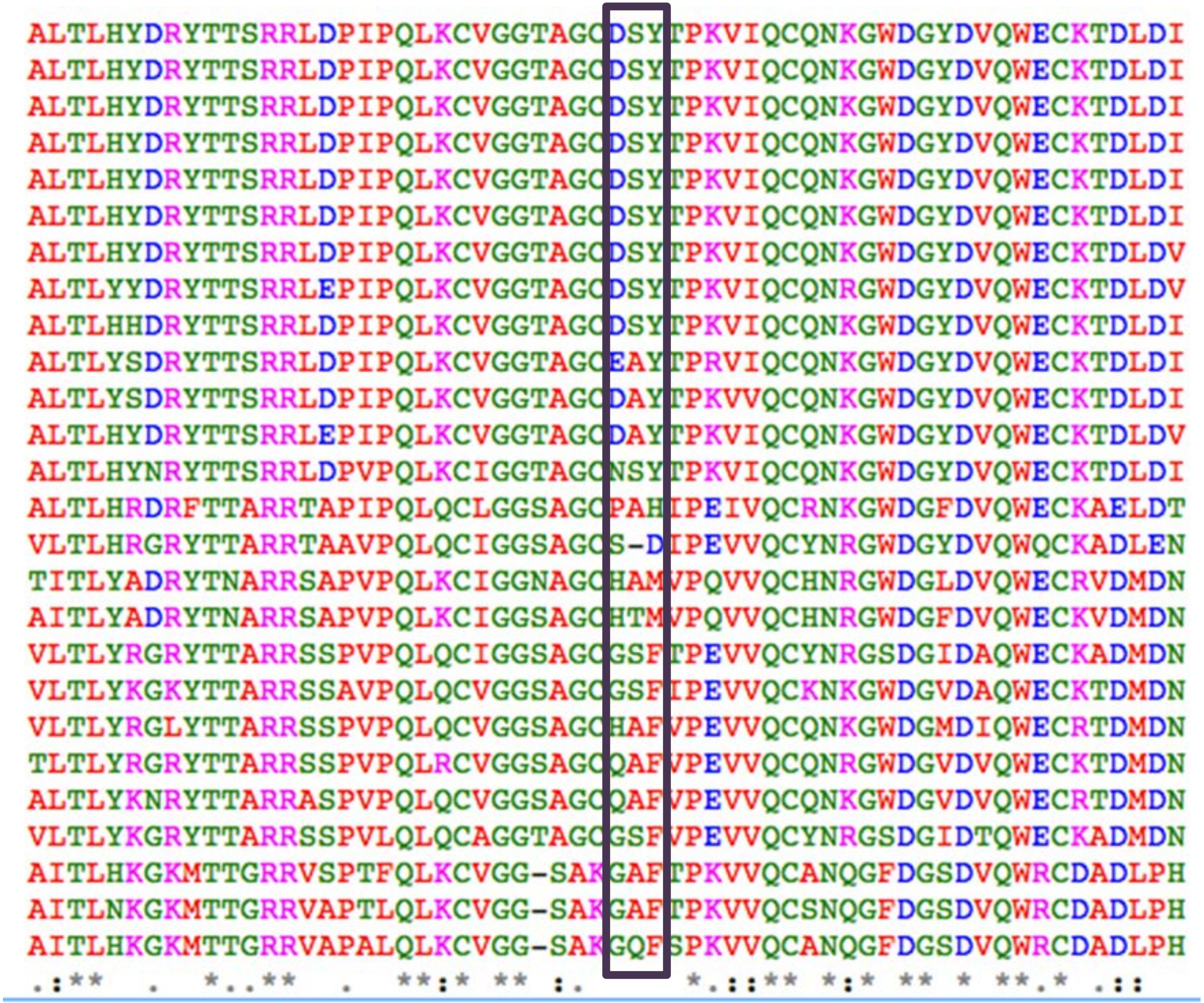


```
ALTLHRDRFTTARRTAPIPQLQCLGGSAGCPAHIPEIVQCRNKGWDGFDVQWECKAELDT
VLTLHRGRYTTARRTAAVPQLQCIGGSAGCS-DIPEVVQCYNRGWDGYDVQWQCKADLEN
```

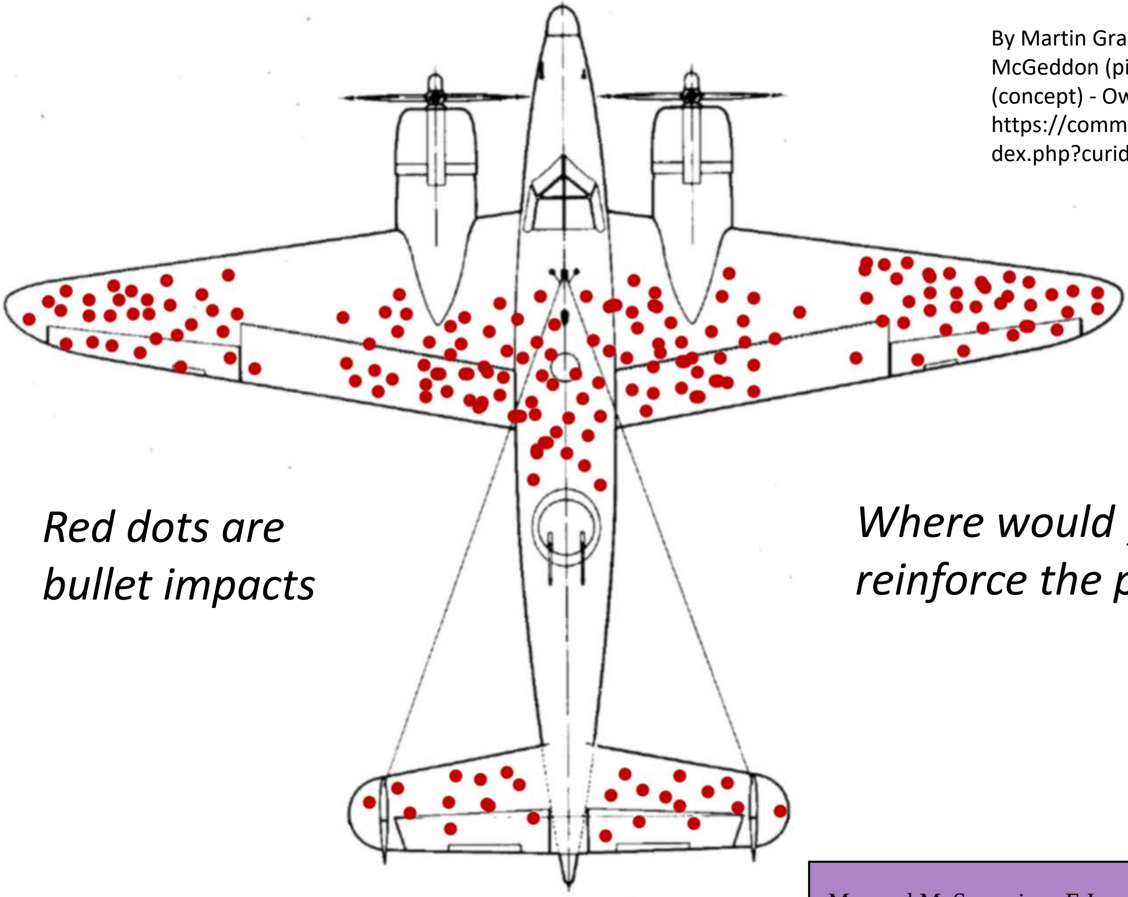
A man with a watch knows what time it is.

A man with two watches is never sure.

Photograph courtesy Getty Images; Collage by Gabe Conte



```
ALTLHYDRYTTSRRLDPIPQLKCVGGTAGCDSYTPKVIQCQNKGWDGYDVQWECKTDLDI
ALTLHYDRYTTSRRLDPIPQLKCVGGTAGCDSYTPKVIQCQNKGWDGYDVQWECKTDLDI
ALTLHYDRYTTSRRLDPIPQLKCVGGTAGCDSYTPKVIQCQNKGWDGYDVQWECKTDLDI
ALTLHYDRYTTSRRLDPIPQLKCVGGTAGCDSYTPKVIQCQNKGWDGYDVQWECKTDLDI
ALTLHYDRYTTSRRLDPIPQLKCVGGTAGCDSYTPKVIQCQNKGWDGYDVQWECKTDLDI
ALTLHYDRYTTSRRLDPIPQLKCVGGTAGCDSYTPKVIQCQNKGWDGYDVQWECKTDLDI
ALTLHYDRYTTSRRLDPIPQLKCVGGTAGCDSYTPKVIQCQNKGWDGYDVQWECKTDLDV
ALTLYYDRYTTSRRLEPIPQLKCVGGTAGCDSYTPKVIQCQNRGWDGYDVQWECKTDLDV
ALTLHHDRYTTSRRLDPIPQLKCVGGTAGCDSYTPKVIQCQNKGWDGYDVQWECKTDLDI
ALTLYSDRYTTSRRLDPIPQLKCVGGTAGCEAYTPRVIQCQNKGWDGYDVQWECKTDLDI
ALTLYSDRYTTSRRLDPIPQLKCVGGTAGCDAYTPKVVQCQNKGWDGYDVQWECKTDLDI
ALTLHYDRYTTSRRLEPIPQLKCVGGTAGCDAYTPKVIQCQNKGWDGYDVQWECKTDLDV
ALTLHYNRYTTSRRLDPVPQLKCIGGTAGCNSYTPKVIQCQNKGWDGYDVQWECKTDLDI
ALTLHRDRFTTARRTAPIPQLQCLGGSAGCPAHIPEIVQCRNKGWDGFDVQWECKAELDT
VLTLHRGRYTTARRTAAVPQLQCIGGSAGCS-DIPEVVQCYNRGWDGYDVQWQCKADLEN
TITLYADRYTNARRSAPVPQLKCIGGNAGCHAMVPQVVQCHNRGWDGLDVQWECRVDMDN
AITLYADRYTNARRSAPVPQLKCIGGSAGCHTMVPQVVQCHNRGWDGFDVQWECKVDMDN
VLTLYRGRYTTARRSSPVPQLQCIGGSAGCGSFTPEVVQCYNRGSDGIDAQWECKADMDN
VLTLYKGKYTTARRSSAVPQLQCVGGSAGCGSFIPEVVQCKNKGWDGVDAQWECKTDMDN
VLTLYRGLYTTARRSSPVPQLQCVGGSAGCHAFVPEVVQCQNKGWDGMDIQWECRTDMDN
TLTLYRGRYTTARRSSPVPQLRCVGGSAGCQAFVPEVVQCQNRGWDGVDVQWECKTDMDN
ALTLYKNRYTTARRASPVPQLQCVGGSAGCQAFVPEVVQCQNKGWDGVDVQWECRTDMDN
VLTLYKGRYTTARRSSPVLQLQCAGGTAGCGSFVPEVVQCYNRGSDGIDTQWECKADMDN
AITLHKGKMTTGRRVSPTFQLKCVGG-SAKGAFTPKVVQCANQGFDGSDVQWRCDADLPH
AITLNKGKMTTGRRVAPTLQLKCVGG-SAKGAFTPKVVQCSNQGFDGSDVQWRCDADLPH
AITLHKGKMTTGRRVAPALQLKCVGG-SAKGQFSPKVVQCANQGFDGSDVQWRCDADLPH
.:**  .  *..**  .  **:* ** :.     *.::** *:* ** * **.* .::
```

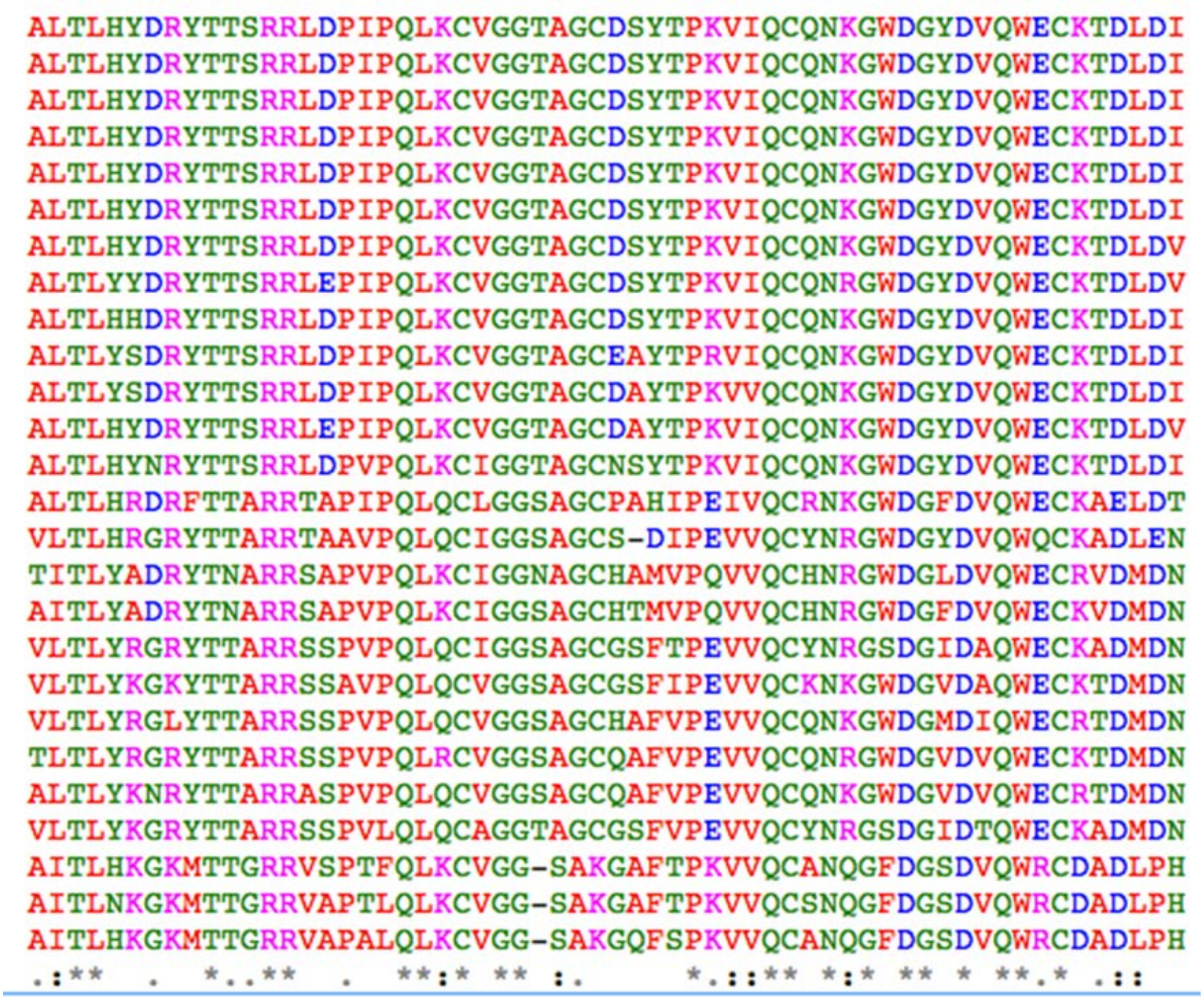


By Martin Grandjean (vector), McGeddon (picture), Cameron Moll (concept) - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=102017718

Red dots are bullet impacts

Where would you reinforce the plane?

Manguel M, Samaniego F.J.,
Abraham Wald's Work on Aircraft Survivability,
J. American Statistical Association. 79, 259-270, (1984)



```
ALTLHYDRYTTSRRLDPIPQLKCVGGTAGCDSYTPKVIQCQNKGWDGYDVQWECKTDLDI
ALTLHYDRYTTSRRLDPIPQLKCVGGTAGCDSYTPKVIQCQNKGWDGYDVQWECKTDLDI
ALTLHYDRYTTSRRLDPIPQLKCVGGTAGCDSYTPKVIQCQNKGWDGYDVQWECKTDLDI
ALTLHYDRYTTSRRLDPIPQLKCVGGTAGCDSYTPKVIQCQNKGWDGYDVQWECKTDLDI
ALTLHYDRYTTSRRLDPIPQLKCVGGTAGCDSYTPKVIQCQNKGWDGYDVQWECKTDLDI
ALTLHYDRYTTSRRLDPIPQLKCVGGTAGCDSYTPKVIQCQNKGWDGYDVQWECKTDLDI
ALTLHYDRYTTSRRLDPIPQLKCVGGTAGCDSYTPKVIQCQNKGWDGYDVQWECKTDLDV
ALTLYYDRYTTSRRLEPIPQLKCVGGTAGCDSYTPKVIQCQNRGWDGYDVQWECKTDLDV
ALTLHHDRYTTSRRLDPIPQLKCVGGTAGCDSYTPKVIQCQNKGWDGYDVQWECKTDLDI
ALTLYSDRYTTSRRLDPIPQLKCVGGTAGCEAYTPRVIQCQNKGWDGYDVQWECKTDLDI
ALTLYSDRYTTSRRLDPIPQLKCVGGTAGCDAYTPKVVQCQNKGWDGYDVQWECKTDLDI
ALTLHYDRYTTSRRLEPIPQLKCVGGTAGCDAYTPKVIQCQNKGWDGYDVQWECKTDLDV
ALTLHYNRYTTSRRLDPVPQLKCIGGTAGCNSYTPKVIQCQNKGWDGYDVQWECKTDLDI
ALTLHRDRFTTARRTAPIPQLQCLGGSAGCPAHIPEIVQCRNKGWDGFDVQWECKAELDT
VLTLHRGRYTTARRTAAVPQLQCIGGSAGCS-DIPEVVQCYNRGWDGYDVQWQCKADLEN
TITLYADRYTNARRSAPVPQLKCIGGNAGCHAMVPQVVQCHNRGWDGLDVQWECRVDMDN
AITLYADRYTNARRSAPVPQLKCIGGSAGCHTMVPQVVQCHNRGWDGFDVQWECKVDMDN
VLTLYRGRYTTARRSSPVPQLQCIGGSAGCGSFTPEVVQCYNRGSDGIDAQWECKADMDN
VLTLYKGKYTTARRSSAVPQLQCVGGSAGCGSFIPEVVQCKNKGWDGVDAQWECKTDMDN
VLTLYRGLYTTARRSSPVPQLQCVGGSAGCHAFVPEVVQCQNKGWDGMDIQWECRTDMDN
TLTLYRGRYTTARRSSPVPQLRCVGGSAGCQAFVPEVVQCQNRGWDGVDVQWECKTDMDN
ALTLYKNRYTTARRASPVPQLQCVGGSAGCQAFVPEVVQCQNKGWDGVDVQWECRTDMDN
VLTLYKGRYTTARRSSPVLQLQCAGGTAGCGSFVPEVVQCYNRGSDGIDTQWECKADMDN
AITLHKGKMTTGRRVSPTFQLKCVGG-SAKGAFTPKVVQCANQGFDGSDVQWRCDADLPH
AITLNKGKMTTGRRVAPTLQLKCVGG-SAKGAFTPKVVQCSNQGFDGSDVQWRCDADLPH
AITLHKGKMTTGRRVAPALQLKCVGG-SAKGQFSPKVVQCANQGFDGSDVQWRCDADLPH
.:**  .  *..**  .  **:* ** :.     *.::** *:* ** * **.* .::
```

Selection pressure

```
chite  ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat  --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
trybr  KKDSNAPKRAMTSFMFFSSDFRS----KHSDLS-IVEMSKAAGAAWKELGP
mouse  -----KPKRPRSAYNIYVSESFQ----EAKDDS-AQGKLKLVNEAWKNLSP
             ***. ::: .: ..  .    :  . .      *  .  *: *

chite  AATAKQNYIRALQEYERNGG-
wheat  ANKLKGEYNKAIAAYNKGESA
trybr  AEKDKERYKREM---------
mouse  AKDDRIRYDNEMKSWEEQMAE
       *   : .* . :
```

Why are some residues "not allowed" to mutate?

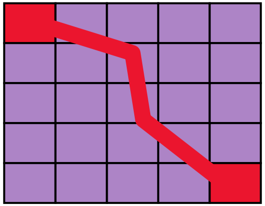
- Residues in catalytic sites, binding sites, molecular gears …
- Stability of the folded state.
- Still, we are diploids and often haplo-sufficient!
- The ultimate **pressure of selection** is:

remaining useful

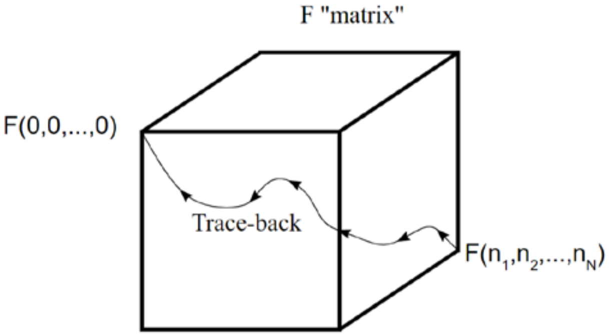
not becoming harmful

Dynamic programming in > 2 dimensions

Align 3 sequences?



from 2D to 3D



Sequences

S: S[0...i]

T: T[0...j]

U: U[0...k]

Scores

F(i, j, k)

↗ ↑ ↖

S T U

$$F_{i,j,k} = \max \left\{ \begin{array}{ll} \left.\begin{array}{l} F_{i-1,j,k} + s\left(S_i, -, -\right) \\ F_{i,j-1,k} + s\left(-, T_j, -\right) \\ F_{i,j,k-1} + s\left(-, -, U_k\right) \end{array}\right\} & \text{1 residue 2 gaps} \\ \left.\begin{array}{l} F_{i-1,j-1,k} + s\left(S_i, T_j, -\right) \\ F_{i-1,j,k-1} + s\left(S_i, -, U_k\right) \\ F_{i,j-1,k-1} + s\left(-, T_j, U_k\right) \end{array}\right\} & \text{2 aligned 1 gap} \\ F_{i-1,j-1,k-1} + s\left(S_i, T_j, U_k\right) & \text{S, T U} \end{array}\right.$$

Computational complexity of N&W

We learned to **globally align a pair of sequences** with N&W

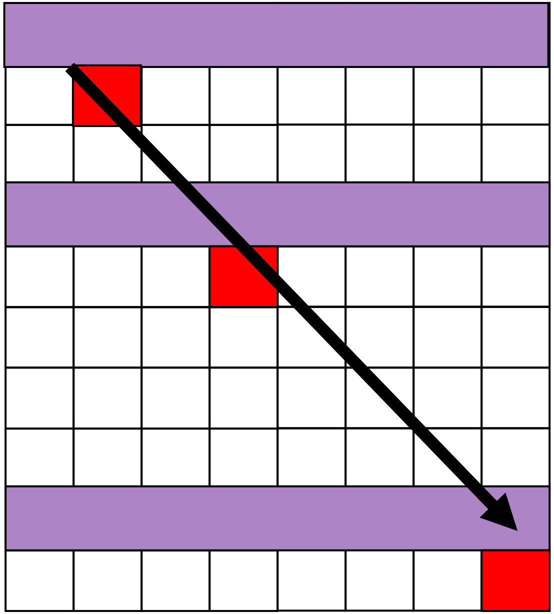
Can we use N&W to align **many sequences**? Not really!!!

2 sequences	O(Lenght^2)	→	~ 1 min
3 sequences	O(Lenght^3)	→	~ 2 h
4 sequences	O(Lenght^4)	→	~ 10 days
5 sequences	O(Lenght^5)	→	~ 3 years
N sequences	O(Lenght^n)	→	forever

Not practical to use dynamic programming for >2 sequences

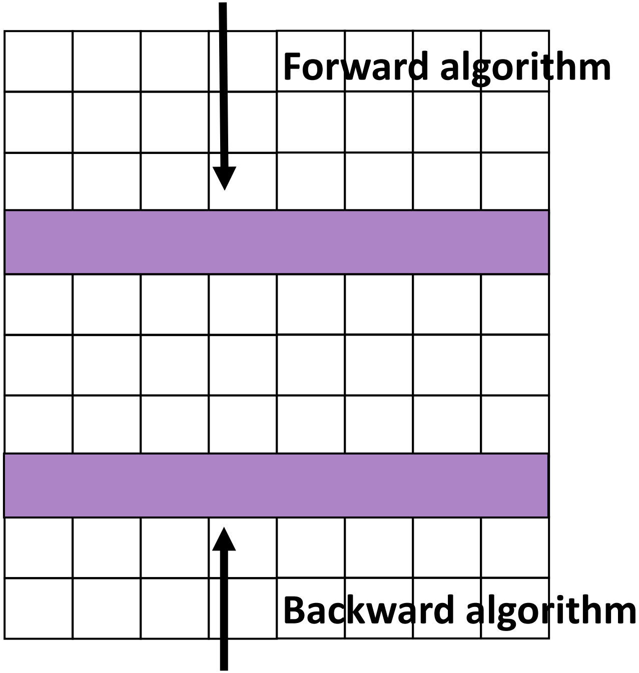
Myers and Miller algorithm

A score in linear space (in memory)



you never need more than the previous row to compute the optimal score

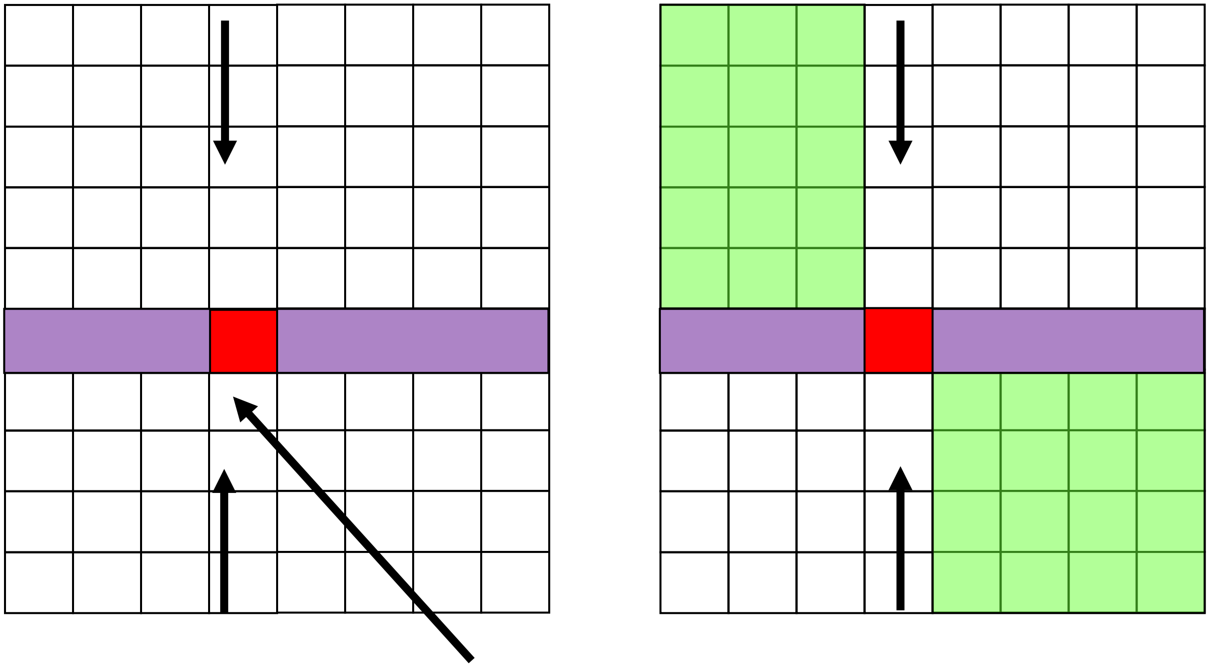
F(i,j)=Optimal score of 0…i Vs 0…j



B(i,j)=Optimal score of M…i Vs N…j

Myers and Miller algorithm

Forward Algorithm

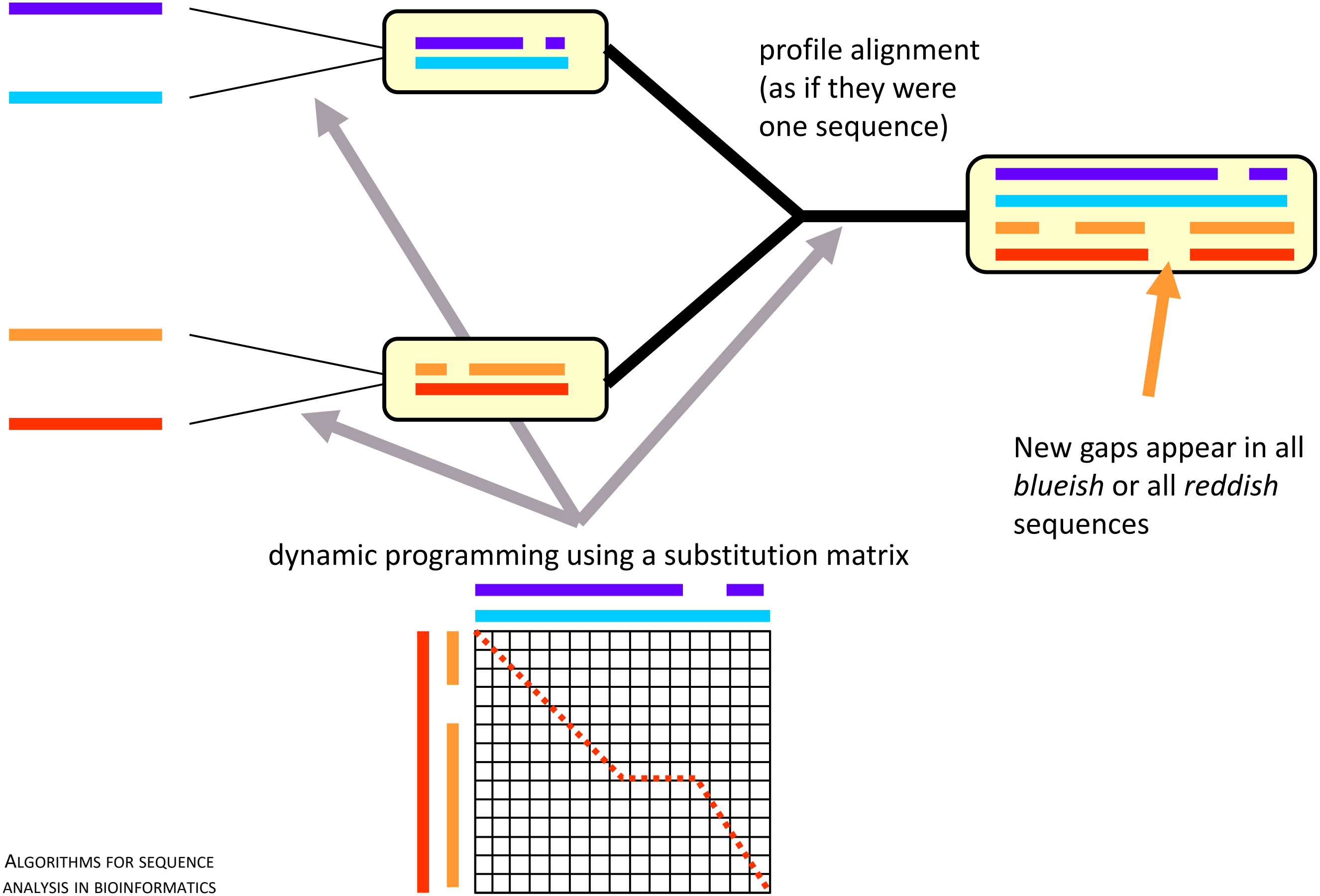


Backward algorithm

Optimal: F(i,j) + B(i,j)

The optimal alignment goes through the red cell

Multiple sequence alignment (Progressive alignment)



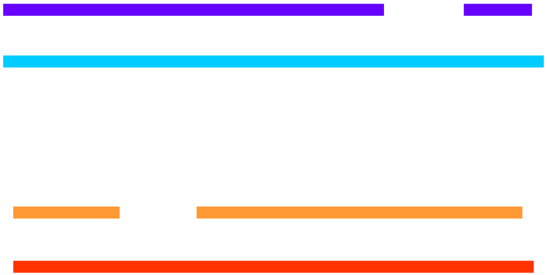
profile alignment (as if they were one sequence)

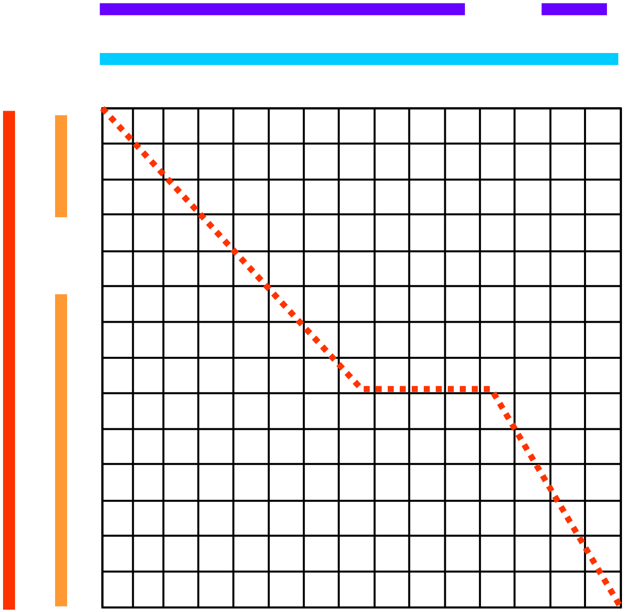
New gaps appear in all *blueish* or all *reddish* sequences

dynamic programming using a substitution matrix

Aligning profiles

The score for matching two columns will be set to the average of the matching scores while ignoring gaps.



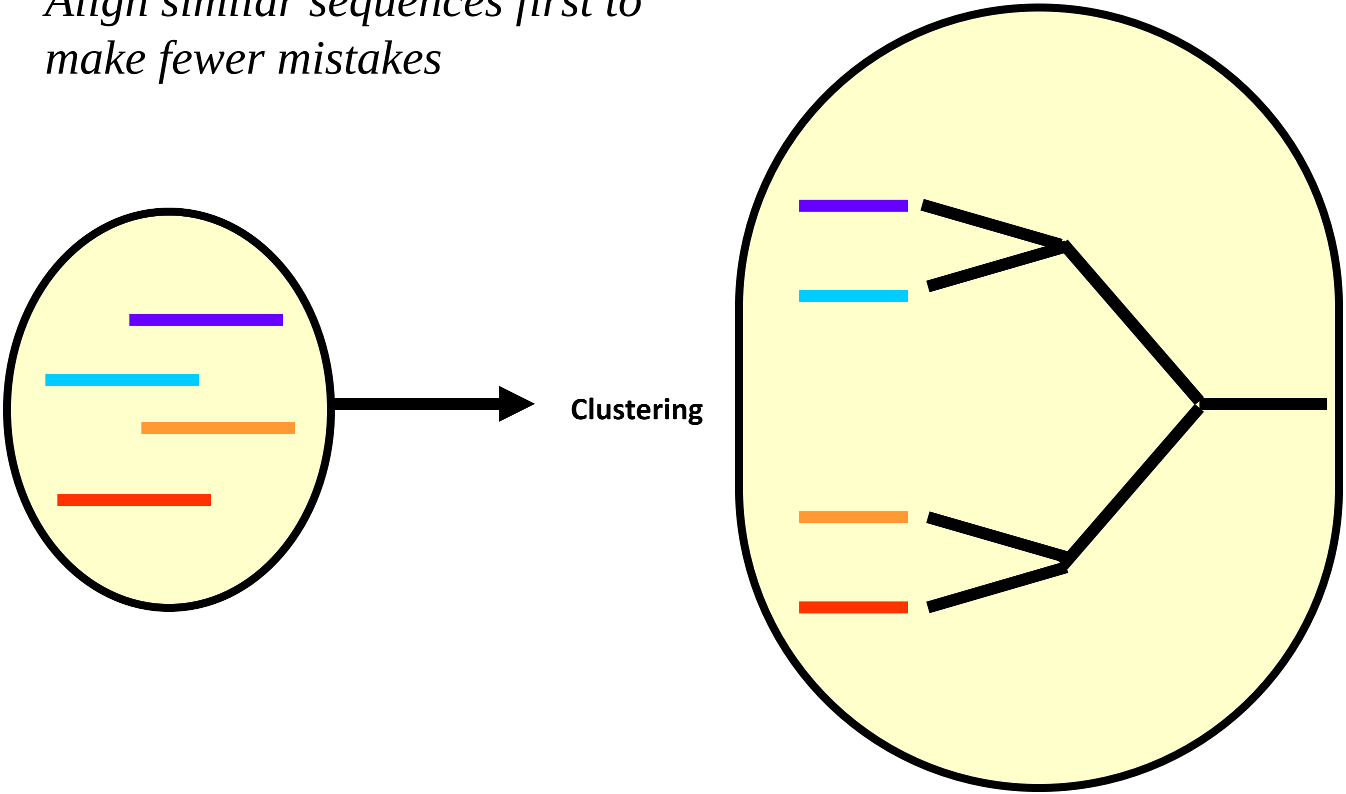


$$S(i, j) = [s(B_1,R_1) + s(B_1,R_2) + s(B_{2,},R_1) + s(B_2,R_2)] / 4$$

If no gaps …

Start with similar sequences

Align similar sequences first to make fewer mistakes



Clustering

Guide tree
(based on sequence similarity)

Progressive alignment algorithm is the most popular

ClustalW

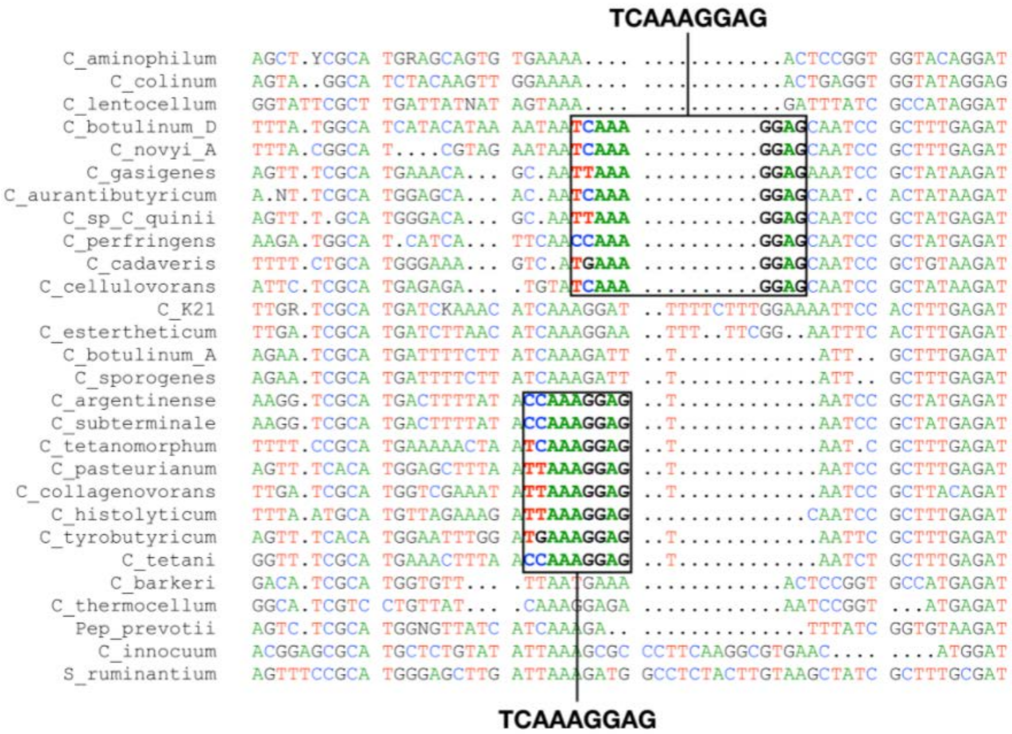
J D Thompson, D G Higgins, and T J Gibson
Nucleic Acids Res. (1994), 22, 4673-4680.
> 50,000 citations

- Feng DF, Doolittle RF (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J Mol Evol 25:351–360
- Taylor WR, Orengo CA (1989) Protein structure alignment. J Mol Biol 208:1–22
- Hogeweg P, Hesper B (1984) The alignment of sets of sequences and the construction of phyletic trees: an integrated method. J Mol Evol 20:175–186

+ Fast Heuristic Algorithm

– Greedy heuristic (No guaranty)

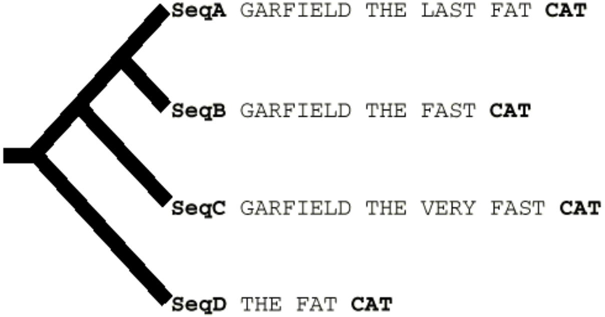
Progressive mis-alignment



```
    C_aminophilum  AGCT.YCGCA TGRAGCAGTG TGAAAA.... ............ACTCCGGT GGTACAGGAT
        C_colinum  AGTA..GGCA TCTACAAGTT GGAAAA.... ............ACTGAGGT GGTATAGGAG
     C_lentocellum  GGTATTCGCT TGATTATNAT AGTAAA.... ............GATTTATC GCCATAGGAT
     C_botulinum_D  TTTA.TGGCA TCATACATAA AATAATCAAA ..........GGAGCAATCC GCTTTGAGAT
        C_novyi_A  TTTA.CGGCA T....CGTAG AATAATCAAA ..........GGAGCAATCC GCTTTGAGAT
       C_gasigenes  AGTT.TCGCA TGAAACA... GC.AATTAAA ..........GGAGAAATCC GCTATAAGAT
C_aurantibutyricum  A.NT.TCGCA TGGAGCA... AC.AATCAAA ..........GGAGCAAT.C ACTATAAGAT
    C_sp_C_quinii  AGTT.T.GCA TGGGACA... GC.AATTAAA ..........GGAGCAATCC GCTATGAGAT
     C_perfringens  AAGA.TGGCA T.CATCA... TTCAACCAAA ..........GGAGCAATCC GCTATGAGAT
       C_cadaveris  TTTT.CTGCA TGGGAAA... GTC.ATGAAA ..........GGAGCAATCC GCTGTAAGAT
   C_cellulovorans  ATTC.TCGCA TGAGAGA... .TGTATCAAA ..........GGAGCAATCC GCTATAAGAT
            C_K21  TTGR.TCGCA TGATCKAAAC ATCAAAGGAT ..TTTTCTTTGGAAAATTCC ACTTTGAGAT
   C_estertheticum  TTGA.TCGCA TGATCTTAAC ATCAAAGGAA ..TTT..TTCGG..AATTTC ACTTTGAGAT
     C_botulinum_A  AGAA.TCGCA TGATTTTCTT ATCAAAGATT ..T............ATT.. GCTTTGAGAT
     C_sporogenes  AGAA.TCGCA TGATTTTCTT ATCAAAGATT ..T............ATT.. GCTTTGAGAT
    C_argentinense  AAGG.TCGCA TGACTTTTAT ACCAAAGGAG ..T............AATCC GCTATGAGAT
    C_subterminale  AAGG.TCGCA TGACTTTTAT ACCAAAGGAG ..T............AATCC GCTATGAGAT
   C_tetanomorphum  TTTT.CCGCA TGAAAAACTA ATCAAAGGAG ..T............AAT.C GCTTTGAGAT
    C_pasteurianum  AGTT.TCACA TGGAGCTTTA ATTAAAGGAG ..T............AATCC GCTTTGAGAT
 C_collagenovorans  TTGA.TCGCA TGGTCGAAAT ATTAAAGGAG ..T............AATCC GCTTACAGAT
    C_histolyticum  TTTA.ATGCA TGTTAGAAAG ATTAAAGGAG ..............CAATCC GCTTTGAGAT
    C_tyrobutyricum  AGTT.TCACA TGGAATTTGG ATGAAAGGAG ..T............AATTC GCTTTGAGAT
         C_tetani  GGTT.TCGCA TGAAACTTTA ACCAAAGGAG ..T............AATCT GCTTTGAGAT
        C_barkeri  GACA.TCGCA TGGTGTT... .TTAATGAAA ............ACTCCGGT GCCATGAGAT
    C_thermocellum  GGCA.TCGTC CTGTTAT... .CAAAGGAGA ............AATCCGGT ...ATGAGAT
      Pep_prevotii  AGTC.TCGCA TGGNGTTATC ATCAAAGA.. ..............TTTATC GGTGTAAGAT
        C_innocuum  ACGGAGCGCA TGCTCTGTAT ATTAAAGCGC CCTTCAAGGCGTGAAC.... ....ATGGAT
     S_ruminantium  AGTTTCCGCA TGGGAGCTTG ATTAAAGATG GCCTCTACTTGTAAGCTATC GCTTTGCGAT
```

Errors made in early pairwise sequence alignment are not corrected later!

Non-optimal solutions



SeqA GARFIELD THE LAST FAT **CAT**

SeqB GARFIELD THE FAST **CAT**

SeqC GARFIELD THE VERY FAST **CAT**

SeqD THE FAT **CAT**

```
CORRECT (Score=24)

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

from T-coffee paper

```
CLUSTALW (Score=20, Gop=-1, Gep=0, M=1)

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

It is very easy to get non-optimal solutions!

Steps

1. **Global Pairwise Alignment** (NW) for all sequence pairs

   → Obtain a distance matrix with the scores

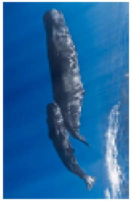
2. Create a **guide tree** from this distance matrix.

   → UPGMA, Neighbor-joining

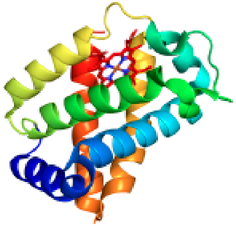
3. **Add sequences progressively** to the alignment according to calculated distances (guide tree).

An example : Globins

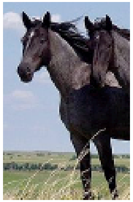
EXAMPLE: Investigate the sequence relation between different globins using the Clustal algorhitm



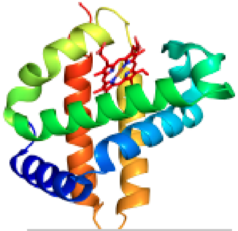
Sperm whale



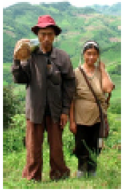
1- Myoglobin



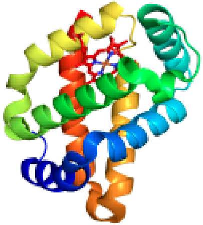
Horse



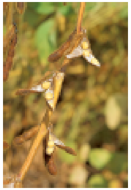
2- Beta-globin



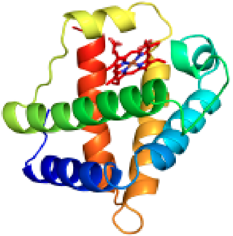
Human



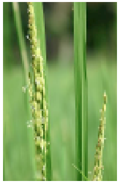
3- Neuroglobin



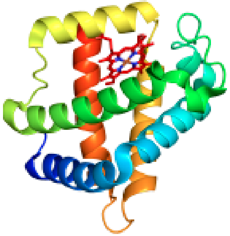
Soybean



4- Leghemoglobin



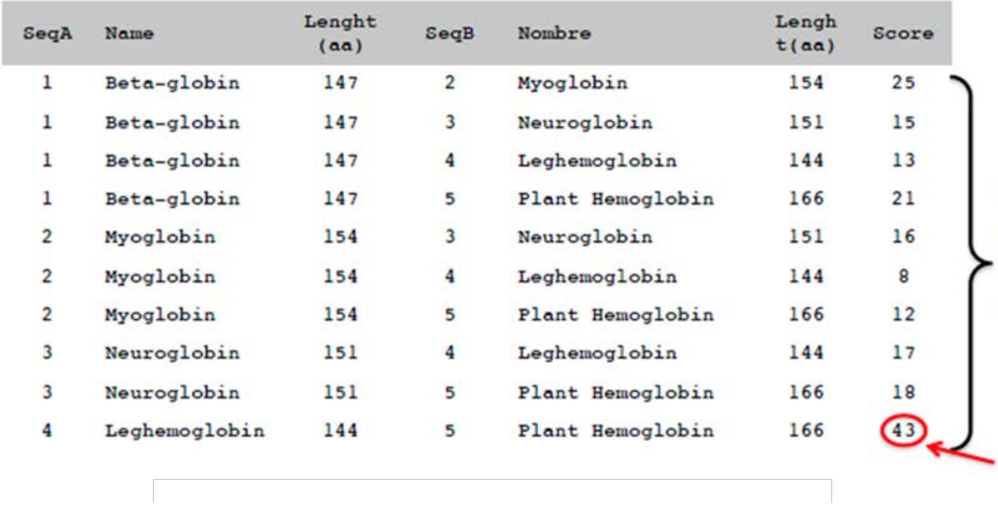
Rice



5- Plant Hemoglobin

Step 1: Global PSA for all sequences

E.g. Five Globins → Beta-globin, Myoglobin, Neuroglobin (vertebrates)
Leghemoglobin, Plant Hemoglobin (plants)



SeqA	Name	Lenght (aa)	SeqB	Nombre	Length(aa)	Score
1	Beta-globin	147	2	Myoglobin	154	25
1	Beta-globin	147	3	Neuroglobin	151	15
1	Beta-globin	147	4	Leghemoglobin	144	13
1	Beta-globin	147	5	Plant Hemoglobin	166	21
2	Myoglobin	154	3	Neuroglobin	151	16
2	Myoglobin	154	4	Leghemoglobin	144	8
2	Myoglobin	154	5	Plant Hemoglobin	166	12
3	Neuroglobin	151	4	Leghemoglobin	144	17
3	Neuroglobin	151	5	Plant Hemoglobin	166	18
4	Leghemoglobin	144	5	Plant Hemoglobin	166	43

Scores are transformed into distances to generate the guide tree

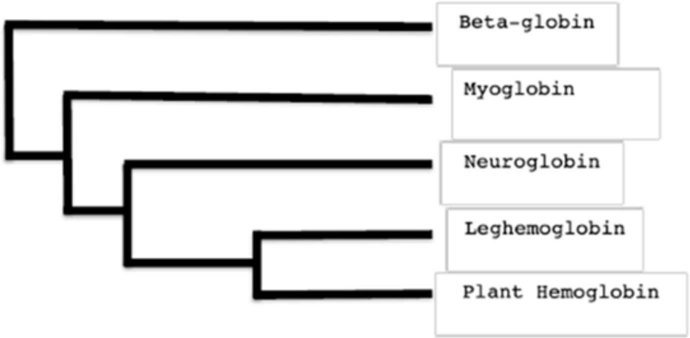
Best Alignment

5 sequences→ 10 alignments
N sequences→ (N-1)! alignments

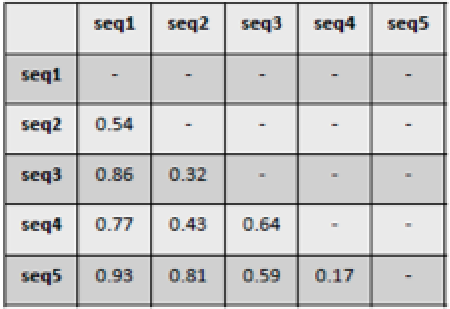
$$D(a,b) = -\log S_{eff(a,b)} = -\log \frac{S_{(a,b)} - S_{rand(a,b)}}{S_{max(a,b)} - S_{rand(a,b)}}$$

Create a guide tree

Step 2: Create a guide tree from the distance matrix



Distance Matrix

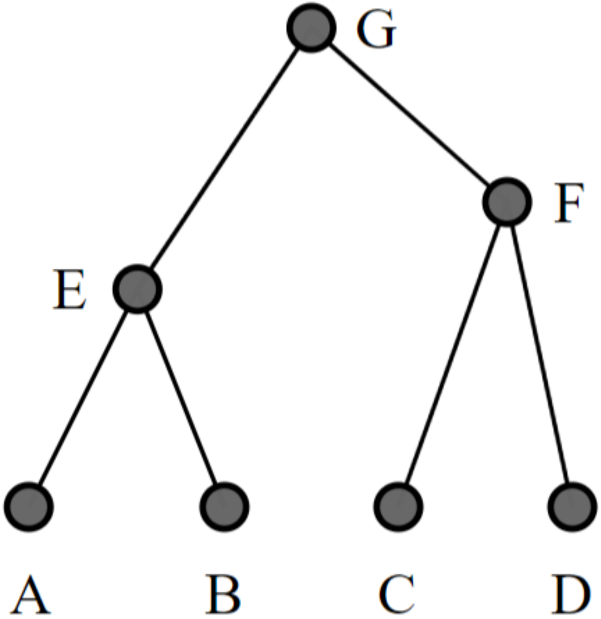


	seq1	seq2	seq3	seq4	seq5
seq1	-	-	-	-	-
seq2	0.54	-	-	-	-
seq3	0.86	0.32	-	-	-
seq4	0.77	0.43	0.64	-	-
seq5	0.93	0.81	0.59	0.17	-

Connect the sequences with smaller distances first (more similar), increment sequence branches following the distance matrix.

The length of the branches are proportional to the distances (greater length => more divergent sequences).

Order of the sequences



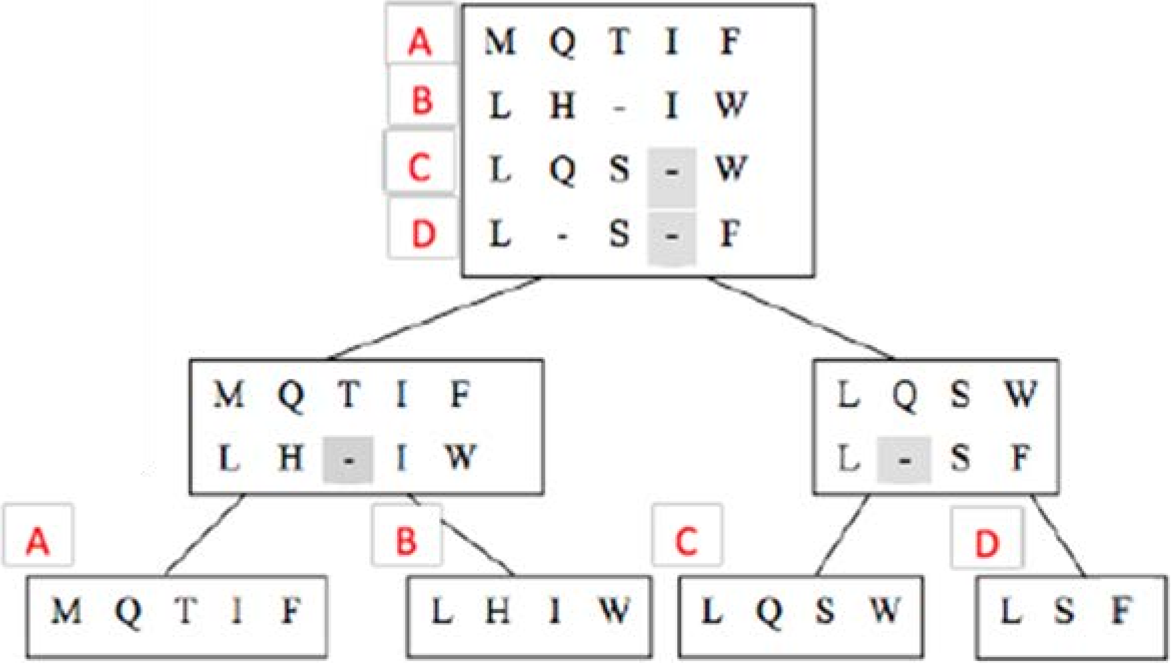
$$d = \begin{array}{c|cc} & E & F \\ E & - & 8 \\ F & & - \end{array}$$

$$\boxed{d = \begin{array}{c|cccc} & A & B & C & D \\ A & - & 4 & 8 & 8 \\ B & & - & 8 & 8 \\ C & & & - & 6 \\ D & & & & - \end{array}}$$

Find nodes with the smallest distance and merge

Progressively align sequences

Step 3: Add sequences progressively to the alignment according to the guide tree



```
A  M Q T I F
B  L H - I W
C  L Q S - W
D  L - S - F
```

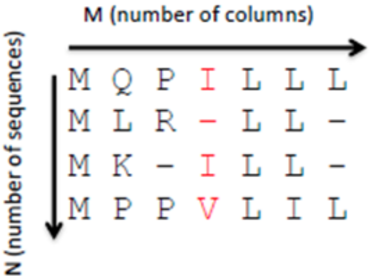
```
M Q T I F
L H - I W
```

```
L Q S W
L - S F
```

```
A  M Q T I F
B  L H I W
C  L Q S W
D  L S F
```

Compute scores

Measured by an objective scoring system such as sum-of-pairs scores (SPS)



```
M Q P I L L L
M L R - L L -
M K - I L L -
M P P V L I L
```

1. Calculate the score of each column
   - Independent of argument order
     `score(I,-,I,V)= score(V,I,I,-)`

For the ith column

$$S_i(\mathtt{I,-,I,V}) = p(\mathtt{I,-}) + p(\mathtt{I,I}) + p(\mathtt{I,V}) + p(\mathtt{-,I}) + p(\mathtt{-,V}) + p(\mathtt{I,V})$$

Sum of scores for all pairs in one column

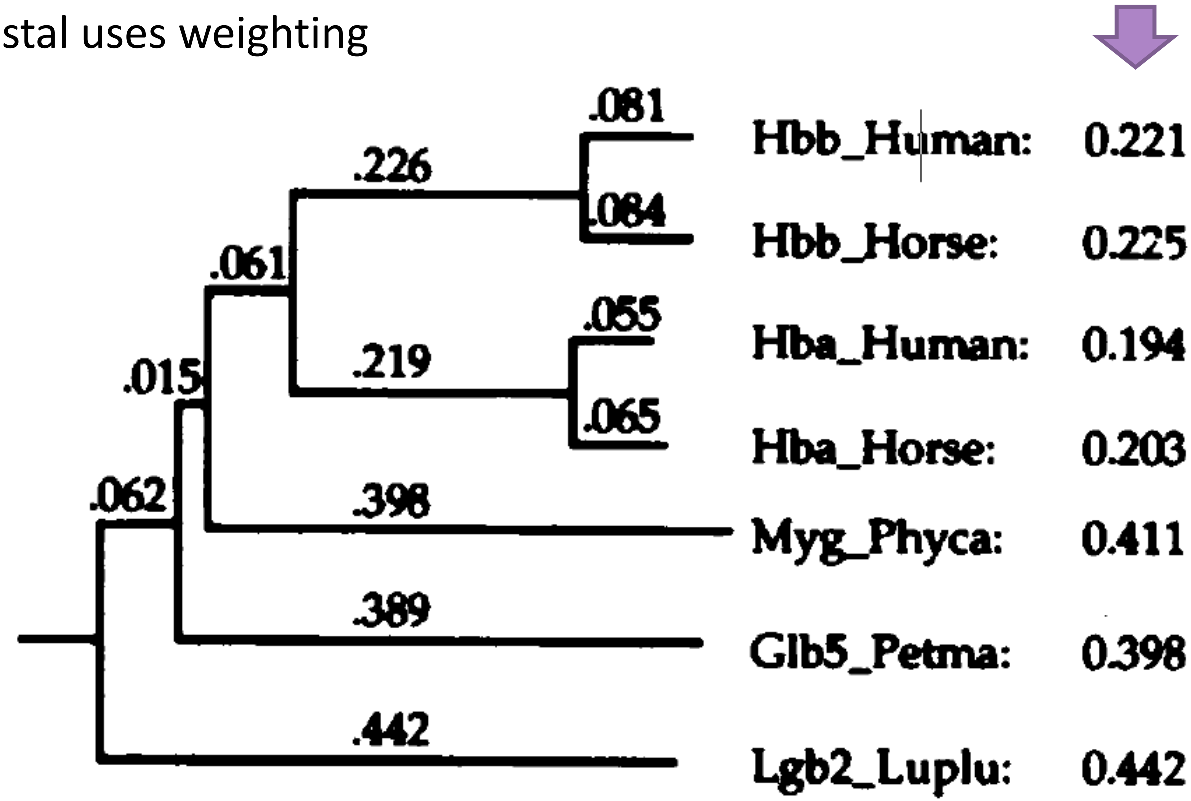
$$S_i = \sum_{j=1, j \neq k}^{N} \sum_{k=1}^{N} p_{ijk}$$

Sum of scores for all aligned columns

$$SPS = \sum_{i=1}^{M} S_i$$

Limit the effect of similar sequences

Clustal uses weighting



from the ClustalW paper

Limitations of MSA

- Same as pairwise alignment problem (but worse)

  We do not know how sequences evolve.

  We do not understand the relation between sequences and structures.

  We would not recognize the correct alignment if we had it in front of our eyes

  …

- Depends on the CHOICE of the sequences

- Depends on the ORDER of the sequences (tree)

- Depends on the PARAMETERS: substitution matrix, gap penalties, scoring system

Applications: Extrapolation

```
chite    ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat    --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
trybr    KKDSNAPKRAMTSFMFFSSDFRS---KHSDLS-IVEMSKAAGAAWKELGP
unknown  -----KPKRPRSAYNIYVSESFQ---EAKDDS-AQGKLKLVNEAWKNLSP
               ***. ::: .: ..  .    :  . .      *  .  *: *

chite    AATAKQNYIRALQEYERNGG-
wheat    ANKLKGEYNKAIAAYNKGESA
trybr    AEKDKERYKREM---------
unknown  AKDDRIRYDNEMKSWEEQMAE
         *   : .* . :
```

< 30 % id (beyond the twilight zone)

BUT

Conserved where it MATTERS

Homology?

Unkown Sequence

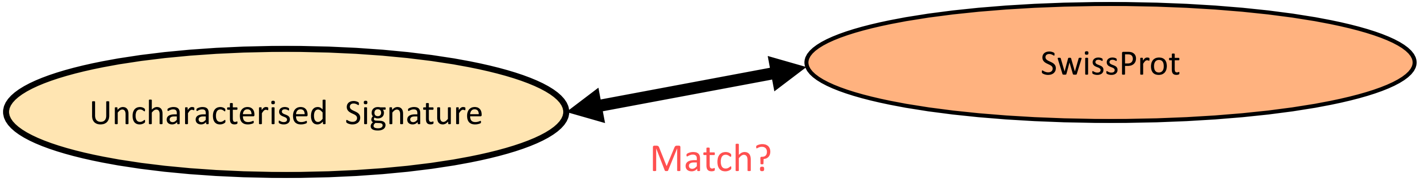
SwissProt

MSA reveals constraints that otherwise would remain invisible

Applications: Patterns

```
chite   ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat   --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
trybr   KKDSNAPKRAMTSFMFFSSDFRS----KHSDLS-IVEMSKAAGAAWKELGP
mouse   -----KPKRPRSAYNIYVSESFQ----EAKDDS-AQGKLKLVNEAWKNLSP
              ***. ::: .: ..  .    :  . .      *  .  *: *

chite   AATAKQNYIRALQEYERNGG-
wheat   ANKLKGEYNKAIAAYNKGESA
trybr   AEKDKERYKREM---------
mouse   AKDDRIRYDNEMKSWEEQMAE
        *   : .* . :
```

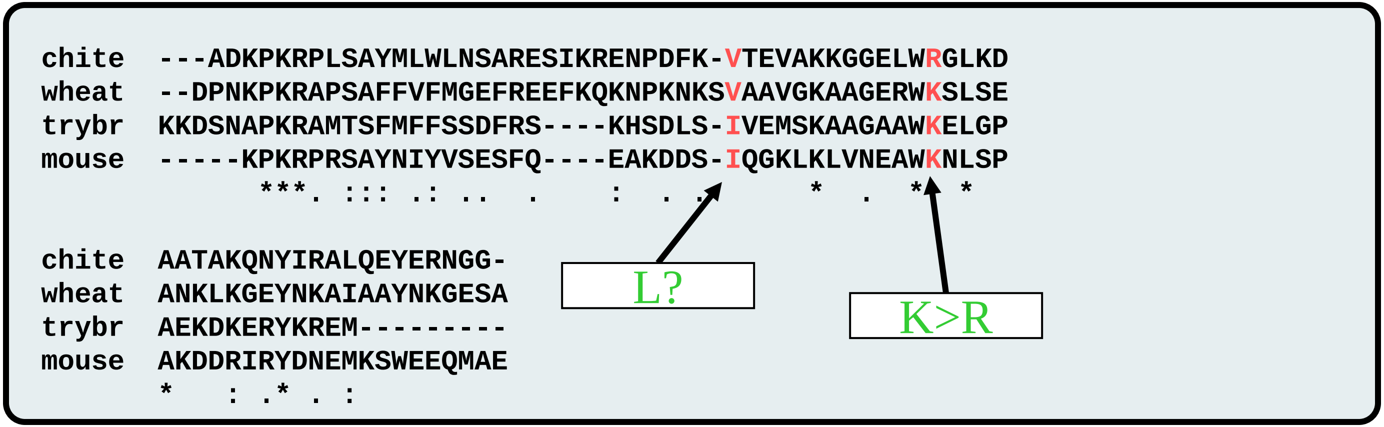
P-K-R-[PA]-x(1)-[ST]… *regular expression*



Applications: Profiles

Position-specific substitution matrices (PSSM)

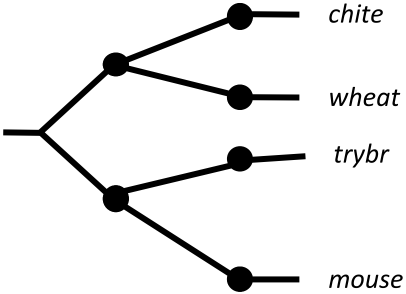
PSI-BLAST



```
chite  ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat  --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
trybr  KKDSNAPKRAMTSFMFFSSDFRS---KHSDLS-IVEMSKAAGAAWKELGP
mouse  -----KPKRPRSAYNIYVSESFQ---EAKDDS-IQGKLKLVNEAWKNLSP
             ***. ::: .: ..  .    :  . .      *  .  *  *

chite  AATAKQNYIRALQEYERNGG-
wheat  ANKLKGEYNKAIAAYNKGESA
trybr  AEKDKERYKREM---------
mouse  AKDDRIRYDNEMKSWEEQMAE
       *   : .* . :
```

Applications: Phylogeny

```
chite  ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat  --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
trybr  KKDSNAPKRAMTSFMFFSSDFRS----KHSDLS-IVEMSKAAGAAWKELGP
mouse  -----KPKRPRSAYNIYVSESFQ----EAKDDS-AQGKLKLVNEAWKNLSP
             ***. ::: .: ..  .    :  . .      *  .  *: *

chite  AATAKQNYIRALQEYERNGG-
wheat  ANLKGEYNKAIAAYNKGESA
trybr  AEKDKERYKREM---------
mouse  AKDDRIRYDNEMKSWEEQMAE
       *   : .* . :
```



- Evolution
- Homology relations

Applications: Structural prediction

```
chite  ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat  --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
trybr  KKDSNAPKRAMTSFMFFSSDKRS----KHSDLS-IVEMSKAAGAAWKELGP
mouse  -----KPKRPRSAYNIYVSESFQ----EAKDDS-AQGKLKLVNEAWKNLSP
             ***. ::: .: ..       :  . .      *  .  *: *

chite  AATAKQNYIRALQEYERNGG-
wheat  ANKLKGEYNKAIAAYNKGESA
trybr  AEKDKERYKREM---------
mouse  AKDDRIRYDNEMKSWEEQMAE
       *   .  *   .
```

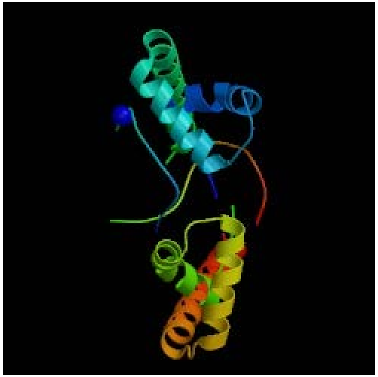
Column Constraint

⇔

Evolution Constraint

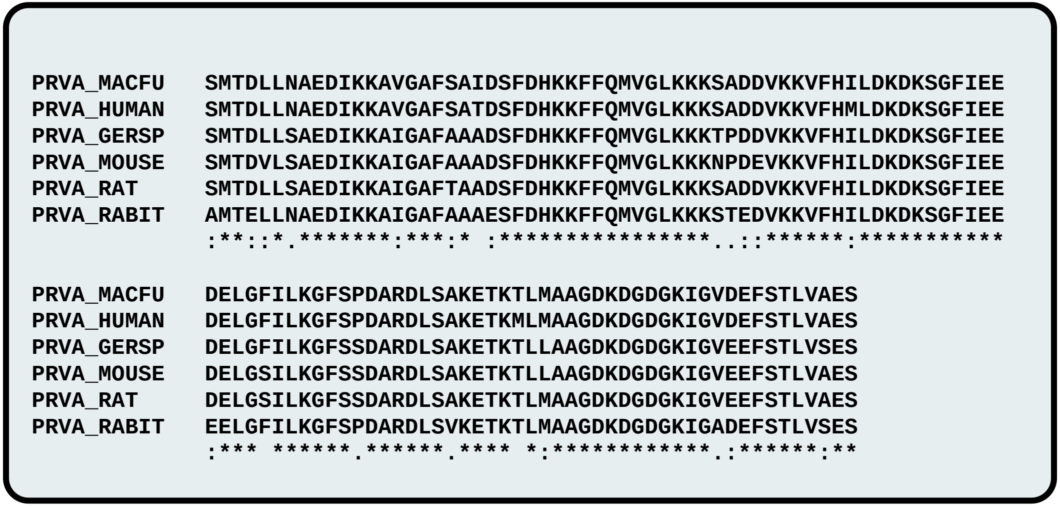
⇔

Structure Constraint



Common mistakes (I)

Choose sequences too closely related



```
PRVA_MACFU   SMTDLLNAEDIKKAVGAFSAIDSFDHKKFFQMVGLKKKSADDVKKVFHILDKDSGFIEE
PRVA_HUMAN   SMTDLLNAEDIKKAVGAFSATDSFDHKKFFQMVGLKKKSADDVKKVFHMLDKDSGFIEE
PRVA_GERSP   SMTDLLSAEDIKKAIGAFAAADSFDHKKFFQMVGLKKKTPDDVKKVFHILDKDSGFIEE
PRVA_MOUSE   SMTDVLSAEDIKKAIGAFAAADSFDHKKFFQMVGLKKKNPDEVKKVFHILDKDSGFIEE
PRVA_RAT     SMTDLLSAEDIKKAIGAFTAADSFDHKKFFQMVGLKKKSADDVKKVFHILDKDSGFIEE
PRVA_RABIT   AMTELLNAEDIKKAIGAFAAAESFDHKKFFQMVGLKKKSTEDVKKVFHILDKDSGFIEE
             :**::*.*******:***:* :****************..::******:**********

PRVA_MACFU   DELGFILKGFSPDARDLSAKETKTLMAAGDKDGDGKIGVDEFSTLVAES
PRVA_HUMAN   DELGFILKGFSPDARDLSAKETKMLMAAGDKDGDGKIGVDEFSTLVAES
PRVA_GERSP   DELGFILKGFSSDARDLSAKETKTLLAAGDKDGDGKIGVEEFSTLVSES
PRVA_MOUSE   DELGSILKGFSSDARDLSAKETKTLLAAGDKDGDGKIGVEEFSTLVAES
PRVA_RAT     DELGSILKGFSSDARDLSAKETKTLMAAGDKDGDGKIGVEEFSTLVAES
PRVA_RABIT   EELGFILKGFSPDARDLSVKETKTLMAAGDKDGDGKIGADEFSTLVSES
             :*** ******.******.**** *:************.:******:**
```

Identical sequences bring no information for the multiple sequence alignment

Multiple sequence alignments thrive on diversity…

Entropy

How much information is in column *i* ?

Shannon entropy or **information content** (H(i))

$$H(i) = -\sum_{x} p_x(i) \log_b p_x(i)$$

H(i): 0 → no information; all amino acids are the same

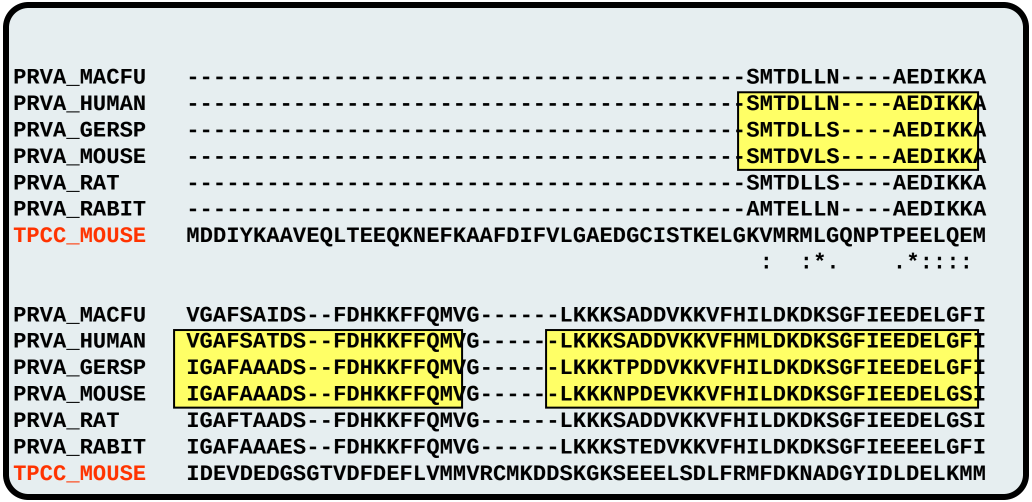
H(i): 1 → all amino acids are equally frequent

(i): frequency of amino acid x in column i

b: 2 (tosses in a coin)
20 (possible amino acids)

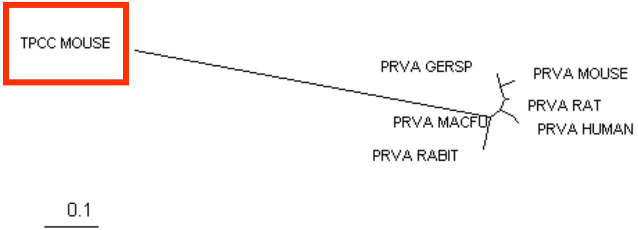
1 aa:	$p_x = 1$	$H = -(1 \times \log_{20} 1 = 0$
2 aa:	$p_x = 0.5$	$H = -2 \times (.5 \times -.23) = 0.22$
20 aa:	$p_x = 0.05$	$H = -20 \times (.05 \times -1) = 1$

Common mistakes (I)



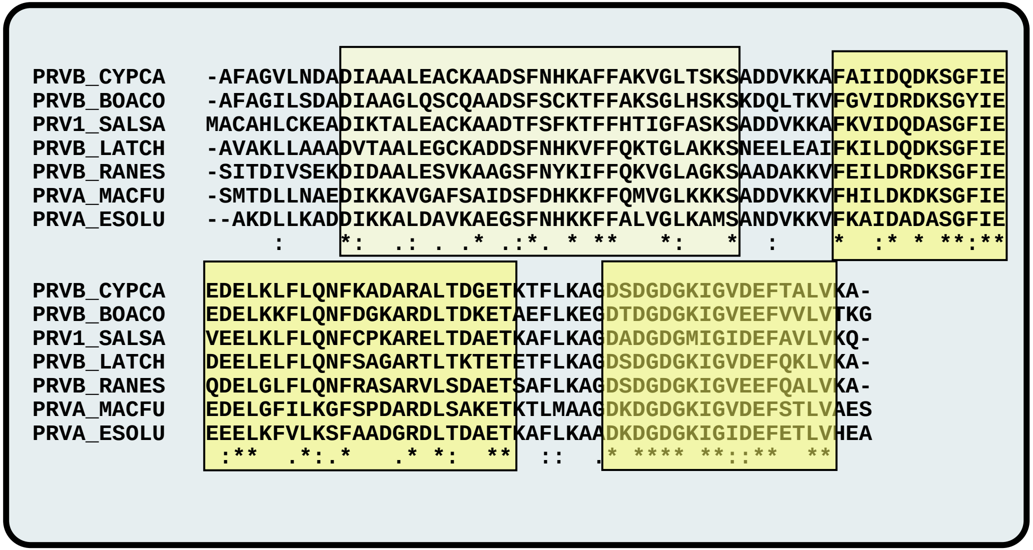
```
PRVA_MACFU     ------------------------------------------SMTDLLN----AEDIKKA
PRVA_HUMAN     ------------------------------------------SMTDLLN----AEDIKKA
PRVA_GERSP     ------------------------------------------SMTDLLS----AEDIKKA
PRVA_MOUSE     ------------------------------------------SMTDVLS----AEDIKKA
PRVA_RAT       ------------------------------------------SMTDLLS----AEDIKKA
PRVA_RABIT     ------------------------------------------AMTELLN----AEDIKKA
TPCC_MOUSE     MDDIYKAAVEQLTEEQKNEFKAAFDIFVLGAEDGCISTKELGKVMRMLGQNPTPEELQEM
                                                           :  :*.    .*::::

PRVA_MACFU     VGAFSAIDS--FDHKKFFQMVG------LKKKSADDVKKVFHILDKDKSGFIEEDELGFI
PRVA_HUMAN     VGAFSATDS--FDHKKFFQMVG------LKKKSADDVKKVFHMLDKDKSGFIEEDELGFI
PRVA_GERSP     IGAFAAADS--FDHKKFFQMVG------LKKKTPDDVKKVFHILDKDKSGFIEEDELGFI
PRVA_MOUSE     IGAFAAADS--FDHKKFFQMVG------LKKKNPDEVKKVFHILDKDKSGFIEEDELGSI
PRVA_RAT       IGAFTAADS--FDHKKFFQMVG------LKKKSADDVKKVFHILDKDKSGFIEEDELGSI
PRVA_RABIT     IGAFAAAES--FDHKKFFQMVG------LKKKSTEDVKKVFHILDKDKSGFIEEEELGFI
TPCC_MOUSE     IDEVDEDGSGTVDFDEFLVMMVRCMKDDSKGKSEEELSDLFRMFDKNADGYIDLDELKMM
```

This alignment Is not Informative about the relation between TPCC MOUSE and the rest of the sequences.

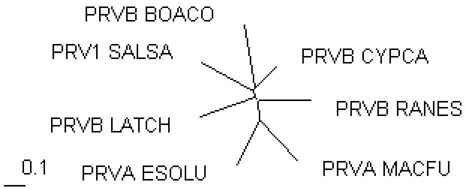


A better spread of the sequences is needed

A more reasonable model: picking diverse sequences



```
PRVB_CYPCA    -AFAGVLNDADIAAALEACKAADSFNHKAFFAKVGLTSKSADDVKKAFAIIDQDKSGFIE
PRVB_BOACO    -AFAGILSDADIAAGLQSCQAADSFSCKTFFAKSGLHSKSKDQLTKVFGVIDRDKSGYIE
PRV1_SALSA    MACAHLCKEADIKTALEACKAADTFSFKTFFHTIGFASKSADDVKKAFKVIDQDASGFIE
PRVB_LATCH    -AVAKLLAAADVTAALEGCKADDSFNHKVFFQKTGLAKKSNEELEAIFKILDQDKSGFIE
PRVB_RANES    -SITDIVSEKDIDAALESVKAAGSFNYKIFFQKVGLAGKSAADAKKVFEILDRDKSGFIE
PRVA_MACFU    -SMTDLLNAEDIKKAVGAFSAIDSFDHKKFFQMVGLKKKSADDVKKVFHILDKDKSGFIE
PRVA_ESOLU    --AKDLLKADDIKKALDAVKAEGSFNHKKFFALVGLKAMSANDVKKVFKAIDADASGFIE
                   :    *:  .: . .* .:*. * **   *:   *  :    *  :* * **:**

PRVB_CYPCA    EDELKLFLQNFKADARALTDGETKTFLKAGDSDGDGKIGVDEFTALVKA-
PRVB_BOACO    EDELKKFLQNFDGKARDLTDKETAEFLKEGDTDGDGKIGVEEFVVLVTKG
PRV1_SALSA    VEELKLFLQNFCPKARELTDAETKAFLKAGDADGDGMIGIDEFAVLVKQ-
PRVB_LATCH    DEELELFLQNFSAGARTLTKTETETFLKAGDSDGDGKIGVDEFQKLVKA-
PRVB_RANES    QDELGLFLQNFRASARVLSDAETSAFLKAGDSDGDGKIGVEEFQALVKA-
PRVA_MACFU    EDELGFILKGFSPDARDLSAKETKTLMAAGDKDGDGKIGVDEFSTLVAES
PRVA_ESOLU    EEELKFVLKSFAADGRDLTDAETKAFLKAADKDGDGKIGIDEFETLVHEA
               :**  .*:.*   .* *:  **  ::  .* **** **::**  **
```



Going Further: adding remote homologues

```
PRVA_MACFU   ------------------------------------------SMTDLLNA----EDIKKA
PRVA_ESOLU   -------------------------------------------AKDLLKA----DDIKKA
PRVB_CYPCA   ------------------------------------------AFAGVLND----ADIAAA
PRVB_BOACO   ------------------------------------------AFAGILSD----ADIAAG
PRV1_SALSA   -----------------------------------------MACAHLCKE----ADIKTA
PRVB_LATCH   ------------------------------------------AVAKLLAA----ADVTAA
PRVB_RANES   ------------------------------------------SITDIVSE----KDIDAA
TPCS_RABIT   -TDQQAEARSYLSEEMIAEFKAAFDMFDADGG-GDISVKELGTVMRMLGQTPTKEELDAI
TPCS_PIG     -TDQQAEARSYLSEEMIAEFKAAFDMFDADGG-GDISVKELGTVMRMLGQTPTKEELDAI
TPCC_MOUSE   MDDIYKAAVEQLTEEQKNEFKAAFDIFVLGAEDGCISTKELGKVMRMLGQNPTPEELQEM
                                                              :        ::

PRVA_MACFU   VGAFSAIDS--FDHKKFFQMVG------LKKKSADDVKKVFHILDKDKSGFIEEDELGFI
PRVA_ESOLU   LDAVKAEGS--FNHKKFFALVG------LKAMSANDVKKVFKAIDADASGFIEEEELKFV
PRVB_CYPCA   LEACKAADS--FNHKAFFAKVG------LTSKSADDVKKAFAIIDQDKSGFIEEDELKLF
PRVB_BOACO   LQSCQAADS--FSCKTFFAKSG------LHSKSKDQLTKVFGVIDRDKSGYIEEDELKKF
PRV1_SALSA   LEACKAADT--FSFKTFFHTIG------FASKSADDVKKAFKVIDQDASGFIEVEELKLF
PRVB_LATCH   LEGCKADDS--FNHKVFFQKTG------LAKKSNEELEAIFKILDQDKSGFIEDEELELF
PRVB_RANES   LESVKAAGS--FNYKIFFQKVG------LAGKSAADAKKVFEILDRDKSGFIEQDELGLF
TPCS_RABIT   IEEVDEDGSGTIDFEEFLVMMVRQMKEDAKGKSEEELAECFRIFDRNADGYIDAEELAEI
TPCS_PIG     IEEVDEDGSGTIDFEEFLVMMVRQMKEDAKGKSEEELAECFRIFDRNMDGYIDAEELAEI
TPCC_MOUSE   IDEVDEDGSGTVDFDEFLVMMVRCMKDDSKGKSEEELSDLFRMFDKNADGYIDLDELKMM
             :   .  .:  .. . *:              *  :    *  :* : .*:*: :**  .

PRVA_MACFU      LKGFSPDARDLSAKETKTLMAAGDKDGDGKIGVDEFSTLVAES-
PRVA_ESOLU      LKSFAADGRDLTDAETKAFLKAADKDGDGKIGIDEFETLVHEA-
PRVB_CYPCA      LQNFKADARALTDGETKTFLKAGDSDGDGKIGVDEFTALVKA--
PRVB_BOACO      LQNFDGKARDLTDKETAEFLKEGDTDGDGKIGVEEFVVLVTKG-
PRV1_SALSA      LQNFCPKARELTDAETKAFLKAGDADGDGMIGIDEFAVLVKQ--
PRVB_LATCH      LQNFSAGARTLTKTETETFLKAGDSDGDGKIGVDEFQKLVKA--
PRVB_RANES      LQNFRASARVLSDAETSAFLKAGDSDGDGKIGVEEFQALVKA--
TPCS_RABIT      FR---ASGEHVTDEEIESLMKDGDKNNDGRIDFDEFLKMMEGVQ
TPCS_PIG        FR---ASGEHVTDEEIESIMKDGDKNNDGRIDFDEFLKMMEGVQ
TPCC_MOUSE      LQ---ATGETITEDDIEELMKDGDKNNDGRIDYDEFLEFMKGVE
                ::     .. ::  :   ::  .* :.** *. :**  ::
```

What makes a good alignment…

- The more divergent the sequences, the better

- The fewer (blocks of) indels, the better

- Nice ungapped blocks separated with indels

- Different classes of residues within a block:
  - completely conserved (*)
  - conserved for size *and* hydropathy (:)
  - conserved for size *or* hydropathy (.)

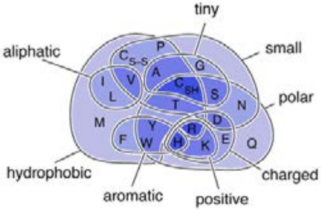
- The ultimate evaluation is a matter of personal judgment and knowledge.

- The BEST alignment method:
  - Your brain
  - The right data

Patterns of conservation in multiple sequence alignments

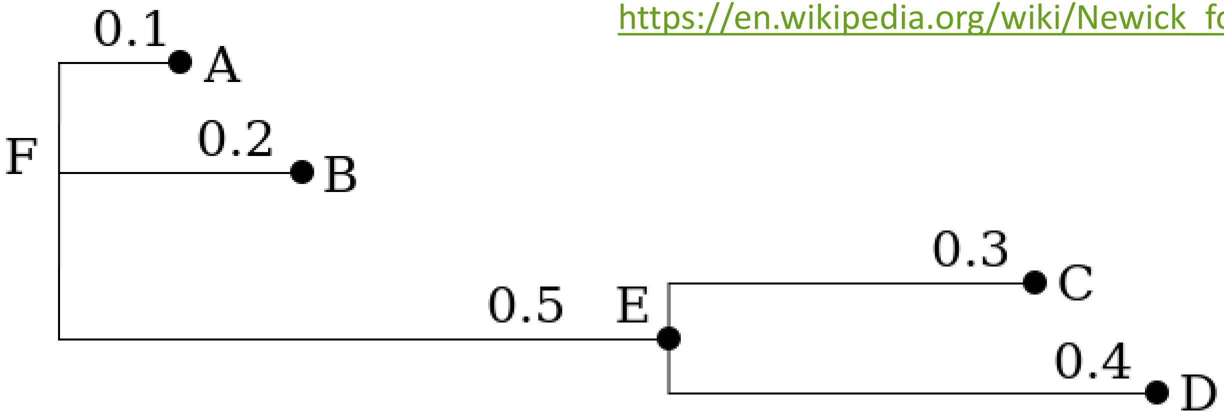
Amino Acid	*Characteristic*
W	It is common to find conserved Tryptophans. Tryptophan is a large hydrophobic residue that sits deep in the core of proteins. It plays an important role in their stability and is therefore difficult to mutate. When tryptophan mutates, it is usually replaced by another aromatic amino acid like phenylalanine or tyrosine. Patterns of conserved aromatic amino acids constitute the most common signatures for recognizing protein domains.
G,P	It is common to find conserved columns with a Glycine or a Proline in a multiple alignment. These two amino acids often coincide with the extremity of well-structured beta strands or alpha helices (see Chapter 13).
C	Cysteines are famous for making C-C (disulphide) bridges. Conserved columns of cysteines are rather common and usually indicate such bridges. Columns of conserved cysteines with a specific distance provide a useful signature for recognizing protein domains and folds.
H,S	Histidine and Serine are often involved in catalytic sites, especially those of proteases. Conserved Histidine or a conserved Serine are good candidates for being part of an active site.
K,R,D,E	These charged amino-acids are often involved in ligand binding. Highly conserved columns can also indicate a salt bridge inside the core of the protein.
L	Leucines are rarely very conserved unless they are involved in protein-protein interactions like leucine zipper.

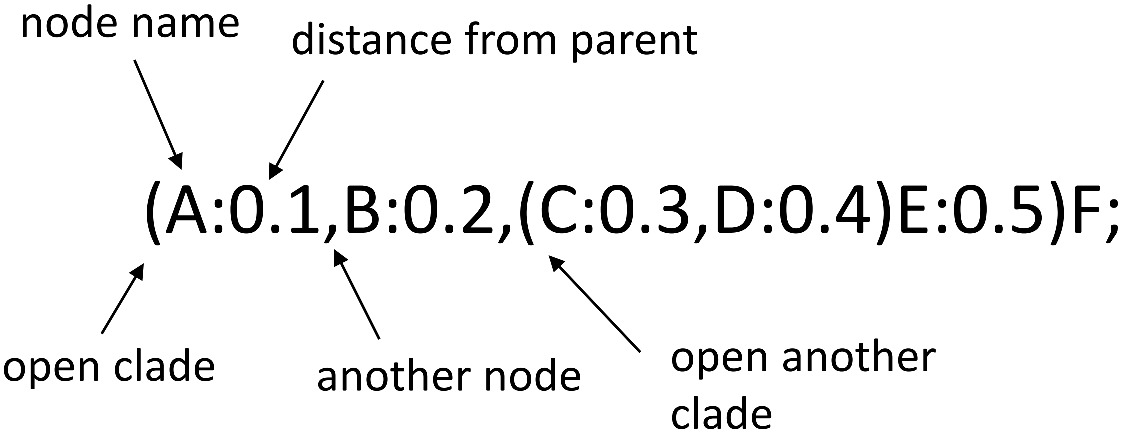
from Bioinformatics For Dummies - Claverie & Notredame



Newick format

https://en.wikipedia.org/wiki/Newick_format





node name

distance from parent

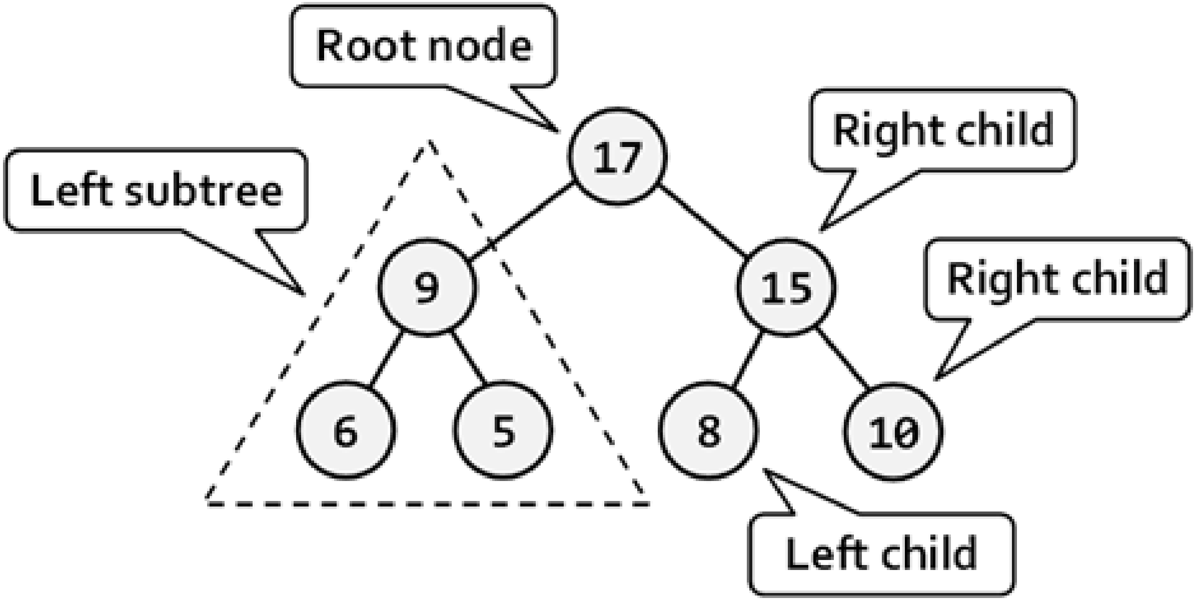
(A:0.1,B:0.2,(C:0.3,D:0.4)E:0.5)F;

open clade

another node

open another clade

Binary trees


Root node
Right child
Left subtree
17
9
15
Right child
6
5
8
10
Left child