Multiple alignment

Orthologous genes vs. Parologous Genes

Orthologous genes: are homologous genes, present in different but related species, which encode for proteins that have similar functions and that have been separated not by a duplication event, but as a result of speciation (separation of species). (Homologous genes, different in different organisms derived from specialization of the species, encode the same function)

Example: α-globin proteins from human and mouse have diverged about 80 million years ago, when there was the division that gave rise to primates and rodents. The two genes are considered orthologous.

Paralogous genes: are genes originated from the duplication of a single gene in the same organism. (Different genes generated by gene duplication of an organism)

es. α-globin and ß-globin human started to diverge after the duplication of an ancestral globin gene. The two genes are considered paralogs.

Multiple alignment

The best way to know the characteristics of a particular proteins family is to align many proteins with similar function.

The functionally or structurally most relevant sites tend to remain unchanged in homologous proteins, while the less important sites may also change a lot.

Observe and study the conservations means understanding how the protein families work, what makes them different from each other, whether or not there exist inter- and intra-family phylogenetic relationships.

In this way it is possible to **identify the function of an unknown protein only by observing the sequence of its residues.**

The sequences to multi-align typically are obtained from the database search using the search systems for similarities such as BLAST and FASTA.

Since they come from an alignment (even if produced with heuristic methods) and that they take into account only sequences that have a high score (and a low E, expectation value), the alignment of these multiple DATASET will give satisfactory results.

In a multiple alignment there are taken into account the columns of residues, rather than the proteins to which they belong. Any residue is lined up to be implicitly considered as evolutionarily related, in some way.

Biological Meaning of the multiple alignment

The multiple alignment summarizes:

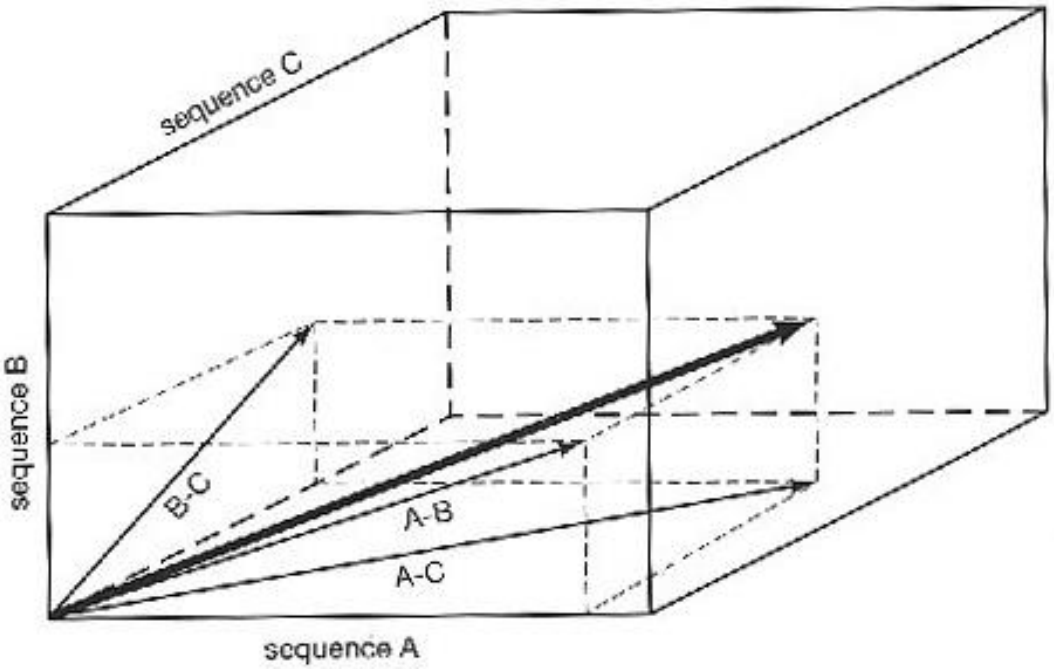
- The evolutionary history of a family of proteins.
- The conservation of function related residues.
- The conservation of the structure related residues.

Methods for the multiple alignment

- optimal alignment
  - –dynamic programming, MSA
- heuristic alignment
  - –Progressive
    - •Global (CLUSTALW, Pileup)
    - •Local (PIMA)
  - – Iterative
    - •Global (PRRP)
    - •Local (dialign)

Optimal Multi-alignment Complexity

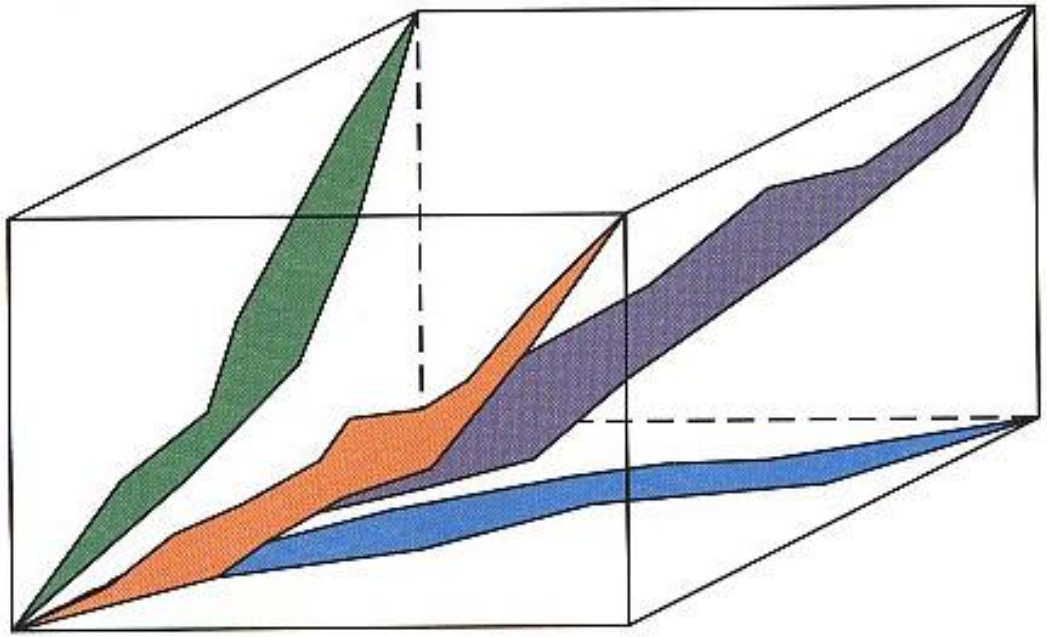
On the surface of the cube you will have the comparison matrices in pairs between the sequences A-B, B-C and A-C. The optimum alignment of three sequences (A-B-C) requires the filling of the cube and the evaluation of all possible moves within the cube.



The complexity of this algorithm is **$(O(L^N))$**, where L is the length and N is the number of sequences. For three sequences of 300 amino acids the number of comparisons is 2.7×10^7. A complete algorithm of dynamic programming is applicable only in the case of three sequences.

Optimal Alignment with dynamic programming

Carrillo & Lipmann, 1988



To find an optimal alignment between the three sequences is necessary to calculate the scores within the volume in gray: the volume is bounded by the projections of the areas on the faces of the cube.

This algorithm (implemented in the MSA program) can be used for a few (<10) short sequences.

Heuristic Algorithm: Progressive Alignment

Alignment of the three sequences A, B, C for the subsequent steps

A-B Alignment

```
DGEKFGPPQRSGQRSG  SeqA
|||.|||  ||||.|||
DGERFGP-QRSGNRSG  SeqB
```

Adding C to the former alignment

consensus ⟷ DHEKFGSSQRSGQRSG SeqC

Multiple Alignment

```
DGEKFGPPQRSGQRSG  SeqA
|||.|||  ||||.|||
DGERFGP-QRSGNRSG  SeqB
| |.||. .|||.|||
DHEKFGSSQRSGQRSG  SeqC
```

N sequences (dataset) arranged at random, not aligned

↓

Align all proteins with all proteins, in pairs: N (N-1) / 2 alignments

↓

Determine a guide tree based on the similarity scores of all pairs

↓

Starting with the most similar pair, determine the columns preserved, and align the next couple keeping these columns and calculating the overall score

↓

N sequences (dataset) aligned

Progressive Alignment

• Alignments in pairs [N (N-1) / 2 comparisons] of all sequences with dynamic programming or approximate methods (BLAST, FASTA). Calculate a diagonal matrix of distances.

• Build a tree (Neighbor-joining, UPGMA, etc.) on the basis of the matrix. The tree serves as a guide for subsequent alignments.

• Beginning with the first node added to the tree, which represents the closest two sequences, align gradually the other nodes (which may be two sequences, a sequence and an alignment or two alignments) until all sequences have been aligned

Distance Matrix Calculation

Objective: From the single alignment performed at the first step, we have the score representing the similarity between the aligned pairs. From it we compute the edit distance for each pair.

- now very important: scoring
- so far:
  - type 1: similarity (e.g., PAM)
  - type 2: distances $\hat{=}$ metric (edit distance)

    minimum entropy

- **however:** distances and similarities can be translated into each other.
- **often:** given similarity (PAM, BLOSUM), searched distance

- one of the first approaches: Feng-Doolittle
  - **given:** 2 sequences a,b and a similarity function $S(\cdot,\cdot)$
  - **wanted:** 2 sequences a,b and a distance function $D(\cdot,\cdot)$

Distance Matrix Calculation

Feng-Doolittle: Similarity $\leftrightarrow$ Distance

- **definitions:**
  - $S(a, b)$ $\rightarrow$ similarity of a, b
  - $S_{a,b}^{\max} = \frac{S(a,a) + S(b,b)}{2}$ $\rightarrow$ maximal possible similarity
  - S_{rand} $\rightarrow$ expected score for aligning two random sequences a', b' with same length and compositions $\Rightarrow$ a, b shuffled
- **then:** define $S_{a,b}^{\text{eff}} = \frac{S(a,b) - S_{rand}}{S_{a,b}^{max} - S_{rand}}$
- $S_{a,b}^{\text{eff}}$ normalised percentage similarity
  - values between 0 and 1
  - converges exponentially slow to 0 for increasing evolutionary distance
- **hence:** for an approx. linear distance define $D(a, b) = -\log S_{a,b}^{\text{eff}}$
- **comment:** $D(a, b)$ will be later used to build guide tree in "progressive MSA"

The guide Tree for the clusterization

- Give a score for each pair, it is possible to determine which is the best way to pass from a protein to an another following a specific order that can be considered as evolutionary progression.
- The (rapid) methods to build the tree, given a score set, are two:
  - Neighbour Joining (NJ)
  - Unweighted Pair Group Methos with Arithmetics Mean (UPGMA)
- These methods base their topological schema on the Distance Matrices that have different meaning with respect to the scores:
  - **Score**: high value => more similar sequences
  - **Distance** : high value => more distant sequences => less similar sequences

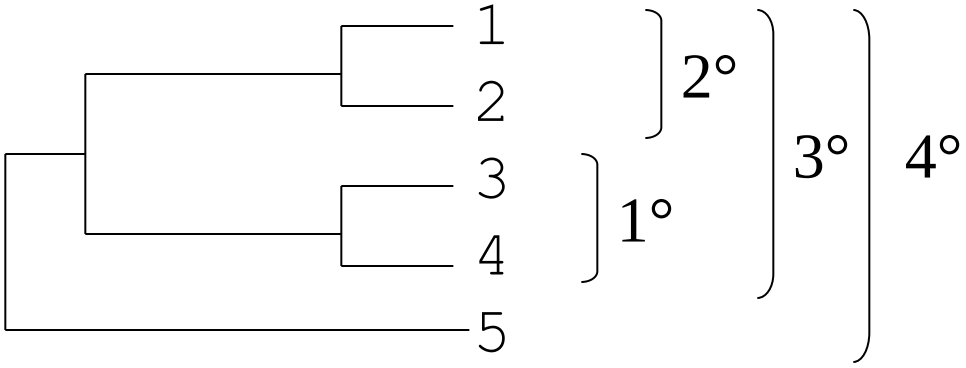
Clusterization and guide tree

1	Hbb_human	-	2°		
2	Hbb_horse	.17	-		
3	Hba_human	.59	.60	-	1°
4	Hba_horse	.59	.59	.13	-
5	Myg_whale	.77	.77	.75	.75

b_hu b_ho a_hu a_ho M_w

It is a matrix of distances, the lower the number, the greater the similarity

...

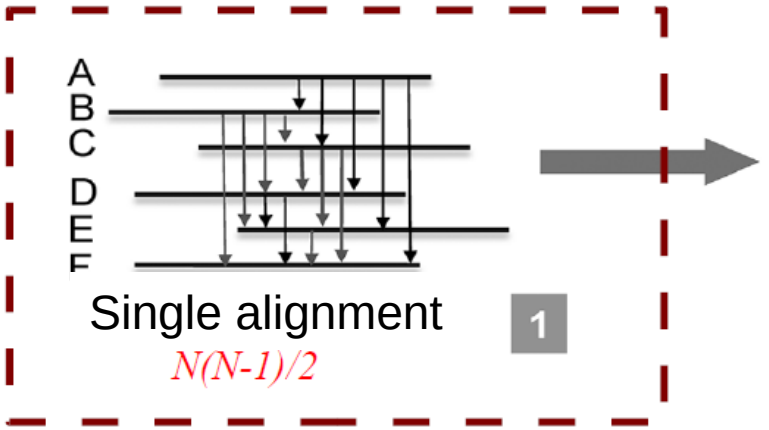


Guide Tree given by the clusterization order

```
PEEKSAVTALWGKVN--VDEVGG Hbb_human
GEEKAAVLALWDKVN--EEEVGG Hbb_horse
PADKTNVKAAWGKVGAHAGEYGA Hba_human
AADKTNVKAAWSKVGGHAGEYGA Hba_horse
EHEWQLVLHVWAKVEAGVAGHGQ Myg_whale
```

Final Alignment

Clustal: schema

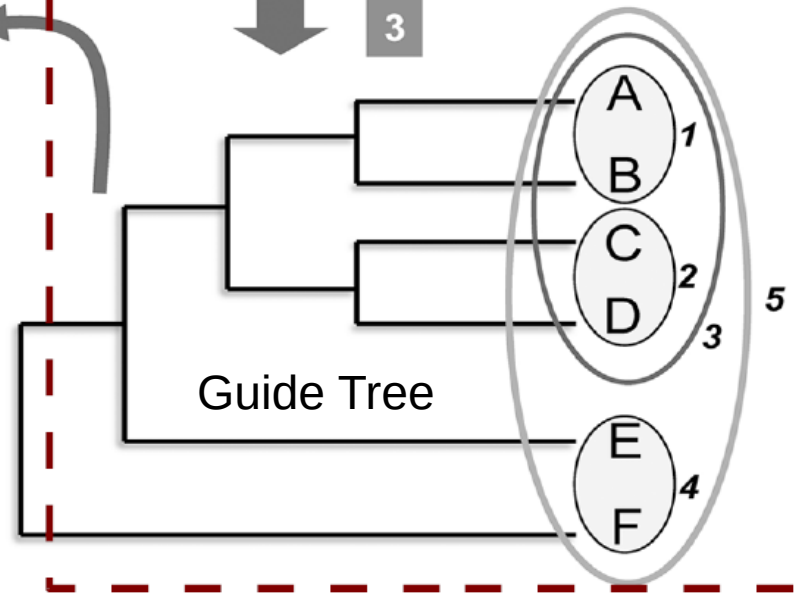


	A	B	C	D	E	F
A	0	D_{AB}	D_{AC}	D_{AD}	D_{AE}	D_{AF}
B		0	D_{BC}	D_{BD}	D_{BE}	D_{BF}
C			0	D_{CD}	D_{CE}	D_{CF}
D				0	D_{DE}	D_{DF}
E					0	D_{EF}
F						0

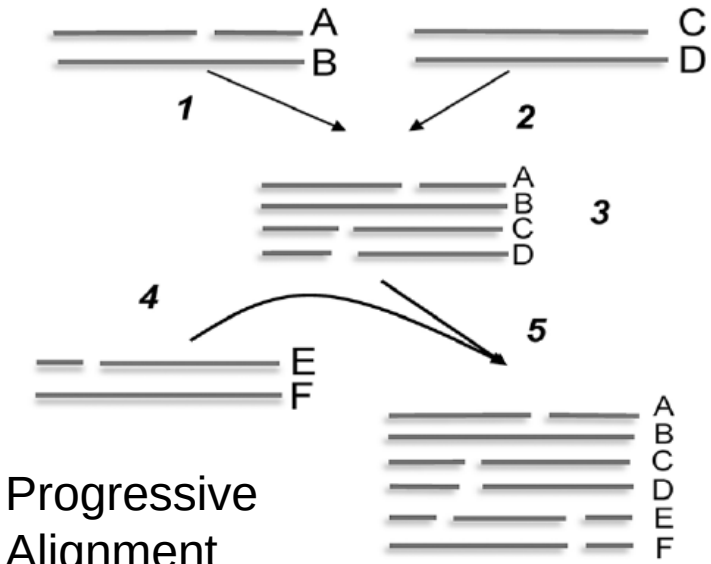
Distance Matrix

2

3



4



CLUSTAL: example

Higgins & Sharp 1988

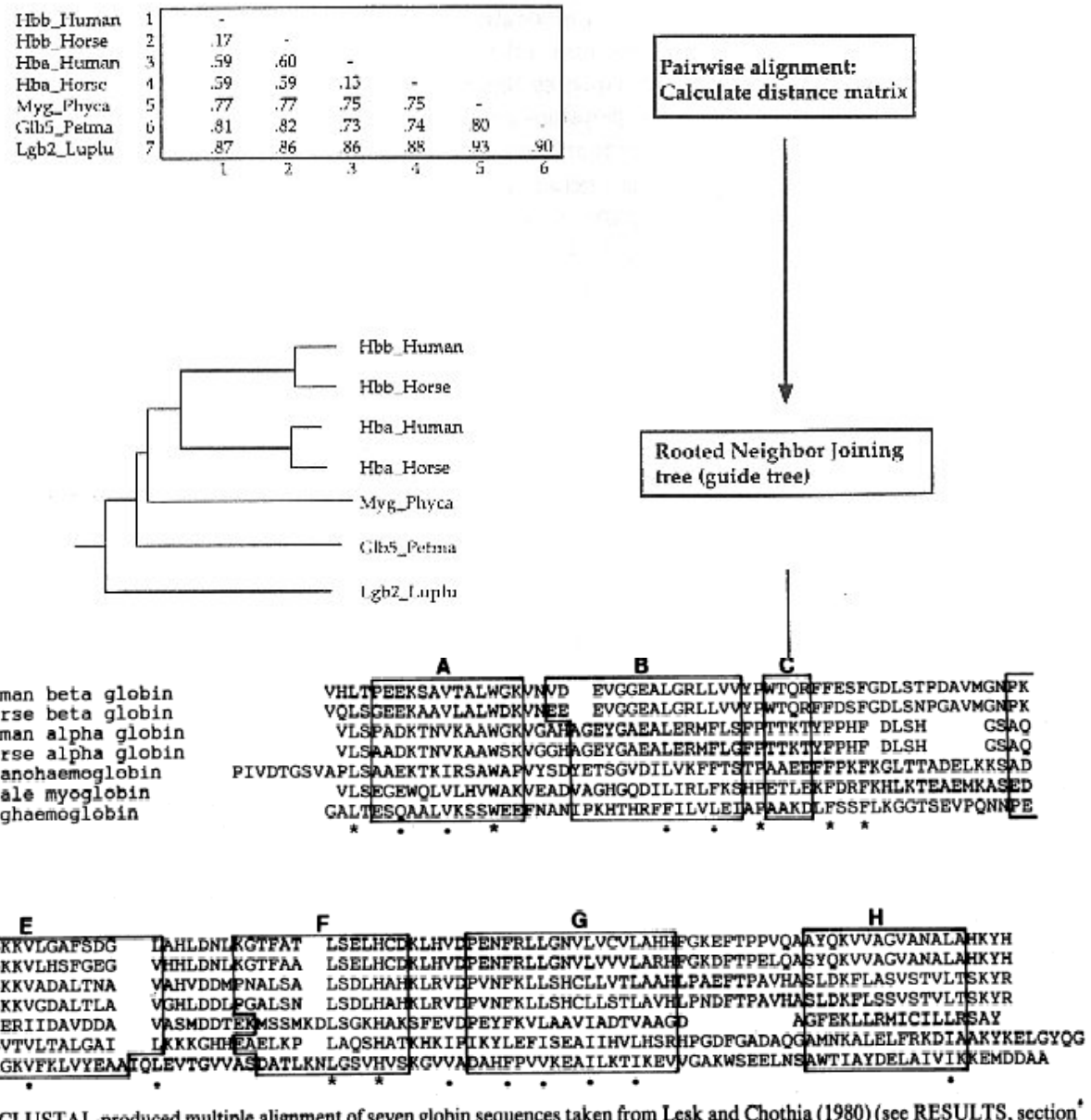
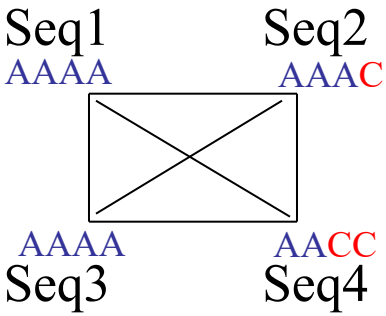


Fig. 3. CLUSTAL-produced multiple alignment of seven globin sequences taken from Lesk and Chothia (1980) (see RESULTS, section

Objective function for the multiple alignment

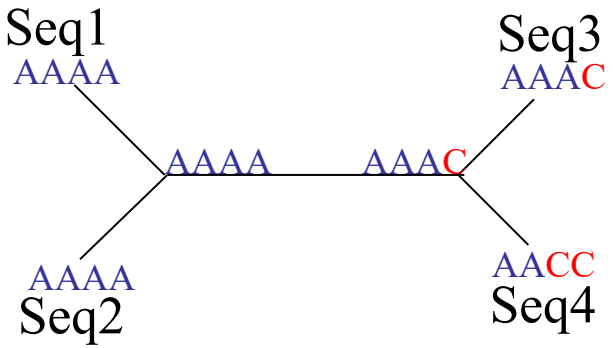
Seq1	A A A A
Seq2	A A A A
Seq3	A A A C
Seq4	A A C C

N(N-1)/2 comparisons



It is used the method of the sum of the score pairs in each column to determine the total alignment score.

This method does not take into account the history of the sequences and the fact that a same character in the column can be easily shared by very similar sequences for evolutionary reasons.



Possible phylogenetic relations of the sequences

To remove the bias introduced by proteins in the same family, it is assigned a weight to the sequences in order to increase the score in the comparisons of evolutionarily distant sequences and decrease it in comparisons between neighboring sequences.

Assess the quality of a multi-alignment

Generally it is added up all the scores of all possible pairs of aligned proteins, weighing the values based on the similarity in the same cluster to prevent some cluster prevail over others in the final count. The WSP (Weighted Sum of Pairs) is defined as:

$$WSP_{score} = \sum_{i=1}^{N-1} \sum_{j=1}^{N} W_{ij}\, \hat{S}(A_{ij})$$

Objective Function (OF)

N: sequence number i,j: sequence pair
$\hat{S}$: similarity score of the pair W: weight of the pair

The total value of the WSP depends on the scoring criteria used in the alignment rather than by biological considerations, but it is a valid criterion for all the alignments with the same parameters

CLUSTALW improvement

Thompson et al 1994

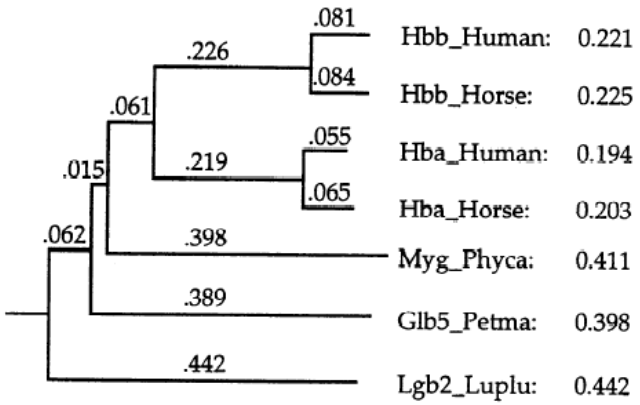


FIG. 3. Sequence weights for the seven globin sequences from Fig. 1. A rooted neighbor-joining tree is shown with branch lengths. The weights are shown for each sequence before normalization (the weights are normalized so as to make the largest equal to 1.0).

The most phylogenetically distant sequences receive a weight proportionally higher in the alignment

RESIDUE-SPECIFIC GAP OPENING PENALTY FACTORS"

Residue	Penalty	Residue	Penalty
A	1.13	M	1.29
C	1.13	N	0.63
D	0.96	P	0.74
E	1.31	Q	1.07
F	1.20	R	0.72
G	0.61	S	0.76
H	1.00	T	0.89
I	1.32	V	1.25
K	0.96	Y	1.00
L	1.21	W	1.23

The penalty to be assigned to the gap depends on the type of residue, such as observed in sequences of known structure (Pascarella & Argos)

The penalty also depends on the location. If there are gaps around, the penalty increases

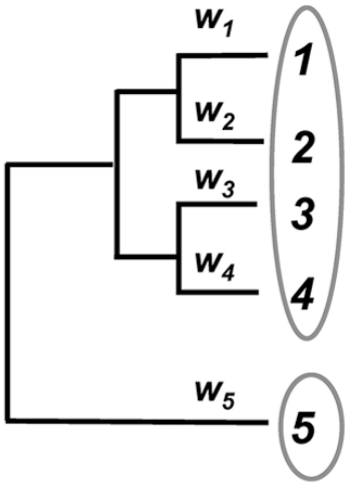
http://www.ebi.ac.uk/clustalw/ o **to download the software**

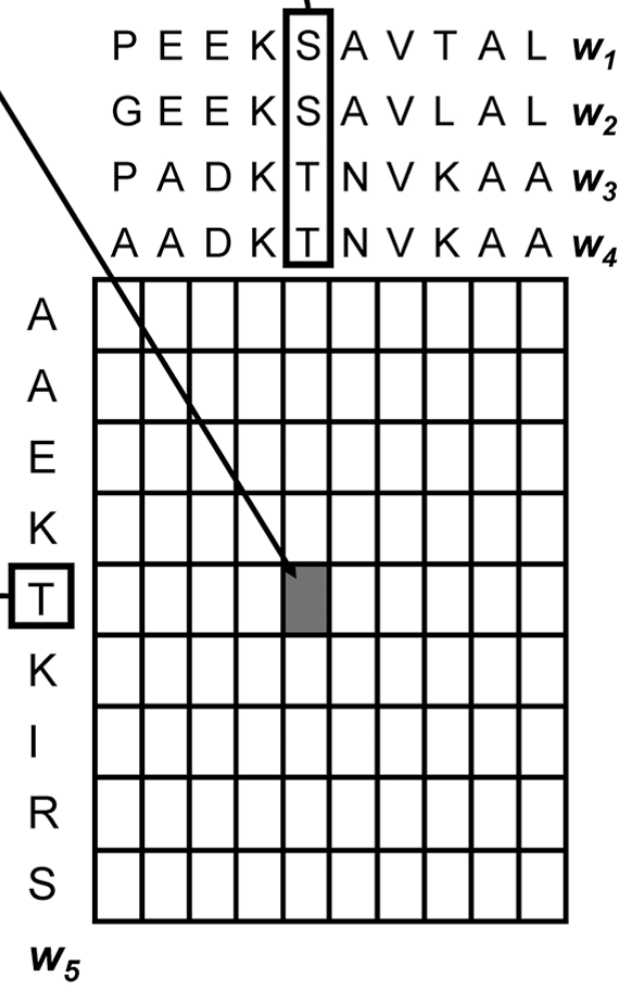
CLUSTALW improvement

Thompson et al 1994

$$S = [w_1 w_5 M(S \rightarrow T) + w_2 w_5 M(S \rightarrow T) + w_3 w_5 M(T \rightarrow T) + w_4 w_5 M(T \rightarrow T)]/4$$

The more distant sequences receive a weight (weight) greater in scoring

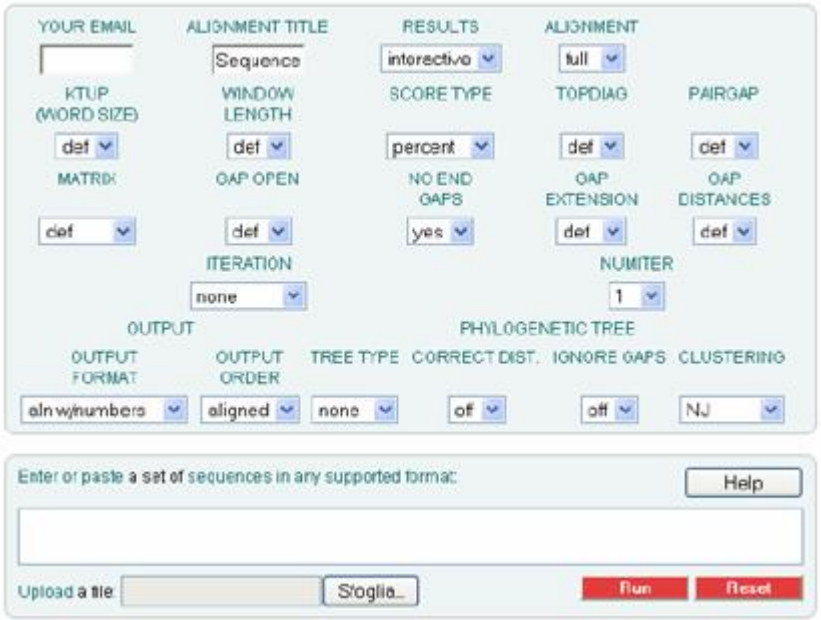




CLUSTALW improvement

ClustalW is the most famous and the most used software for the progressive multi-alignment

It is available on EBI.



ClustalX is a new version of ClustalW that provide an intuitive GUI.

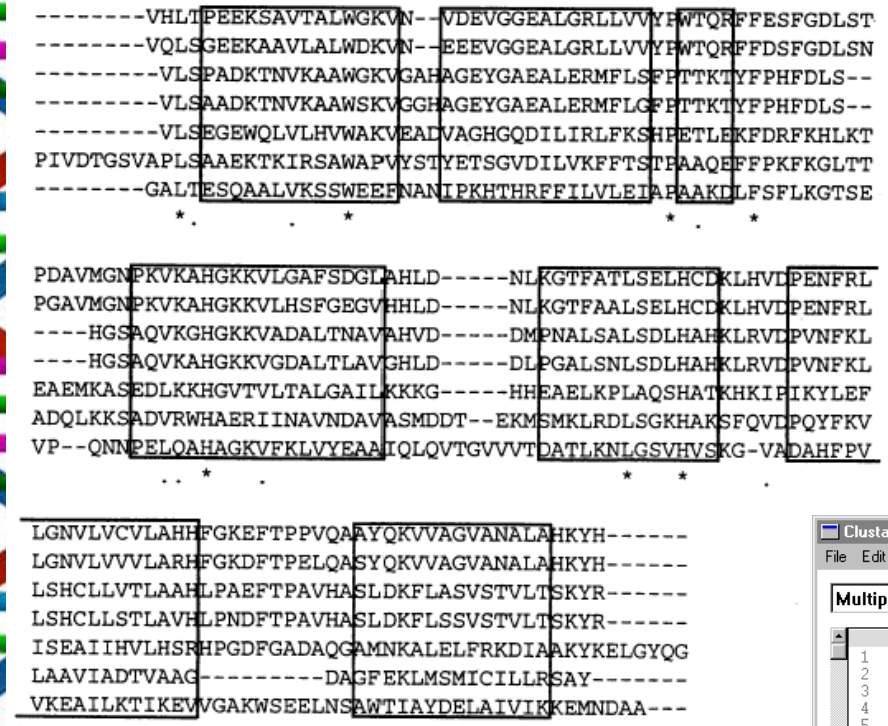
Characteristics

-The algorithm weights differently the sequences and progressive changes matrix when add in the alignment proteins less similar.

-It is possible to add sequences to the ones already aligned, using pre-existing alignments.

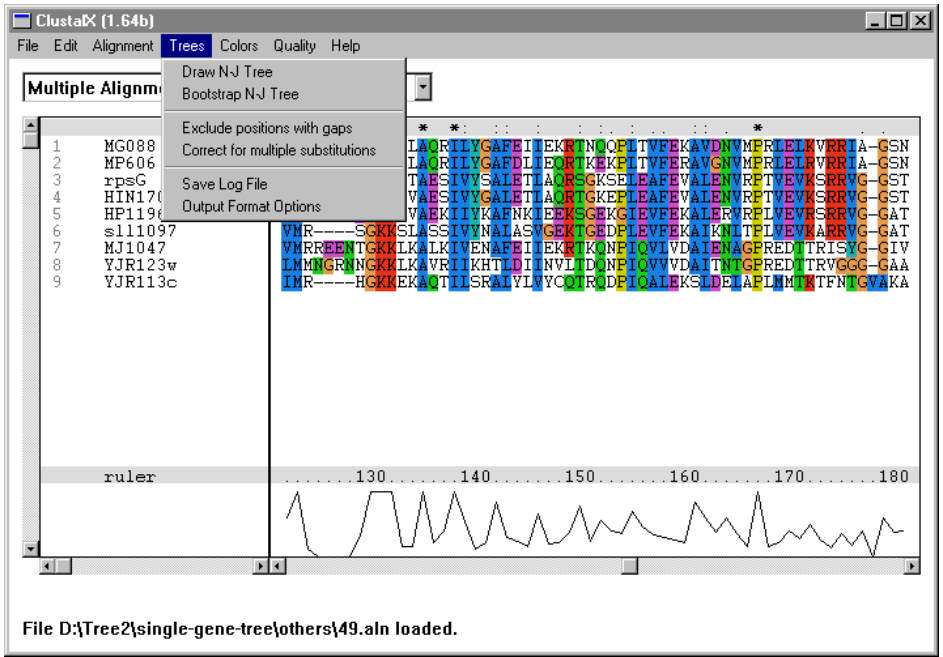
-It is possible to perform structural alignments

CLUSTALW vs CLUSTALX



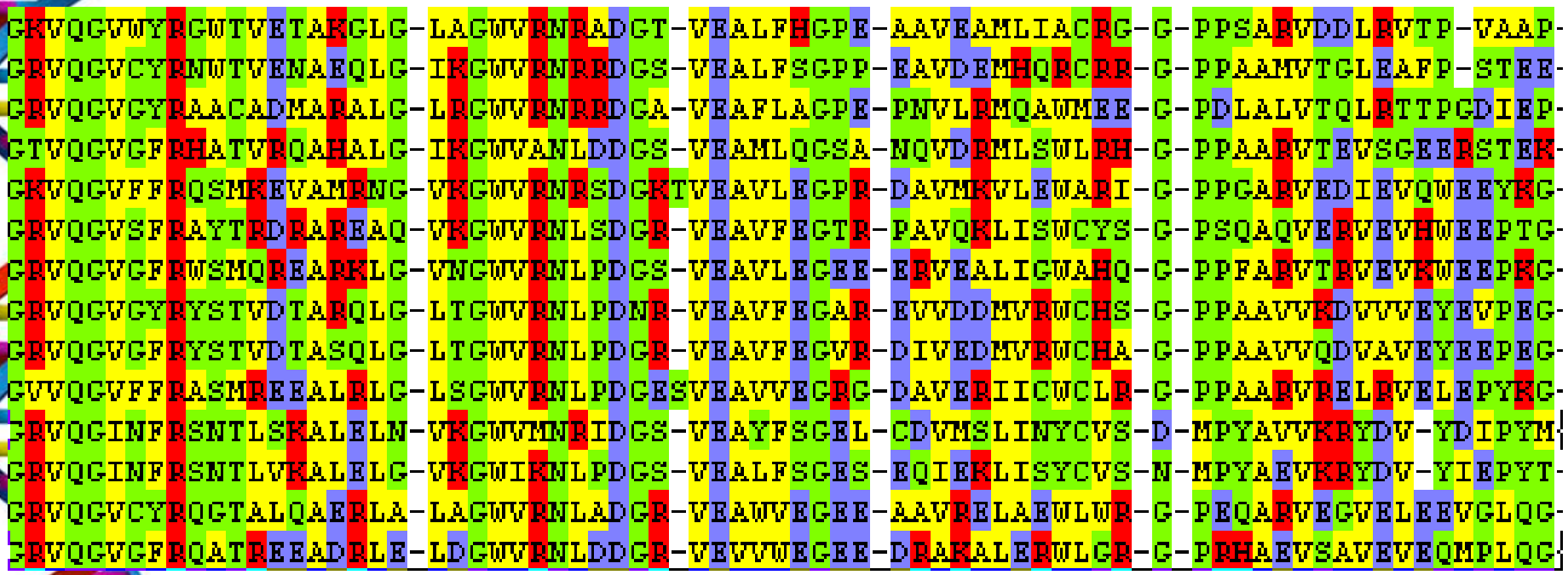
CLUSTALW

CLUSTALX.

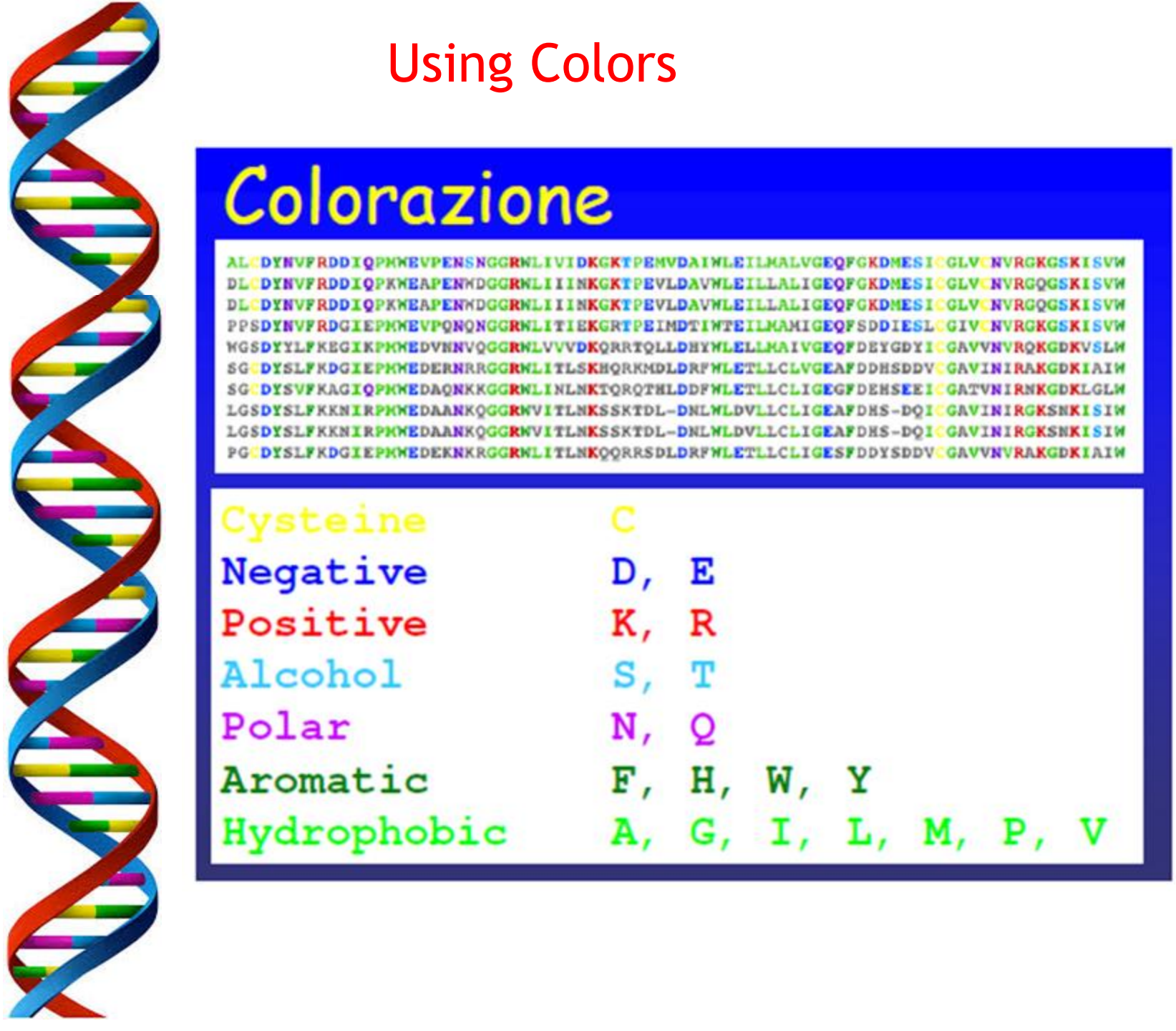


Colors in the alignment

The raw-text files can be used to display the columns, but you can associate different colors for different physical residues with chemical characteristics. This greatly facilitates the display of multi-alignment.



Using Colors



Raw Text files

```
                  *         20          *         40          *         60
TRY1_HUMAN : IVGGYNCEENSVPYQVSLN-------SGYHFCGGSLINEQWVVSAGHCYKSRIQVRLGEHNIEVLEGNE :  62
TRYP_PIG   : IVGGYTCAANSIPYQVSLN-------SGSHFCGGSLINSQWVVSAAHCYKSRIQVRLGEHNIDVLEGNE :  62
TRYU_DROME : IVGGADTSSYYTKYVVQLRRRSSSSSSYAQTCGGCILDAVTIATAAHCVYNREAENFLVVSGDDSRGGM :  69
TRYI_DROME : IIGGSDQLIRNAPWQVSIQ------ISARHECGGVIYSKEIIITAGHCLHERSVT-LMKVRVGAQNHNY :  62
TRY1_SALSA : IVGGYECKAYSQTHQVSLN-------SGYHFCGGSLVNENWVVSAAHCYKSRVEVRLGEHNIKVTEGSE :  62

           *         80          *         100         *         120         *         1
TRY1_HUMAN : --QFINAAKIIRHPQYDRKTLNNDIMLIKLSSRAVIN--ARVSTISLPTAPPATGTKCLISGWGNTASS : 127
TRYP_PIG   : --QFINAAKIITHPNFNGNTLDNDIMLIKLSSPATLN--SRVATVSLPRSCAAAGTECLISGWGNTKSS : 127
TRYU_DROME : YGVVVRVSQLIPHELYNSSTMDNDIALVVVDPPLPLDSFSTMEAIVIASEQPPVGVQATISGWGYTKEN : 138
TRYI_DROME : GGTLVPVAAYKVHEQFDSRFLHYDIAVLRLSTPLTFG--LSTRAINLASTSPSGGTTVTVTGWGHTDNG : 129
TRY1_SALSA : --QFISSSRVIRHPNYSSYNIDNDIMLIKLSKPATLN--TYVQPVALPTSCAPAGTMCTVSGWGNTMSS : 127

           40          *         160         *         180         *         200
TRY1_HUMAN : GADYPDELQCLDAPVLSQAKCEASYPG----KITSNMFCVGFLEGGKDSCQGDSGGPVVCNGQLQGVVS : 192
TRYP_PIG   : GSSYPSLLQCLKAPVLSDSSCKSSYPG----QITGNMICVGFLEGGKDSCQGDSGGPVVCNGQLQGIVS : 192
TRYU_DROME : GLSS-DQLQQVKVPIVDSEKCQEAYYWR---PISEGMLCAGLSEGGKDACQGDSGGPLVVANKLAGIVS : 203
TRYI_DROME : --ALSDSLQKAQLQIIDRGECASQKFGYGADFVGEETICAASTD--ADACTGDSGGPLVASSQLVGIVS : 194
TRY1_SALSA : TADS-NKLQCLNIPILSYSDCNNSYPG----MITNAMFCAGYLEGGKDSCQGDSGGPVVCNGELQGVVS : 191

             *         220         *
TRY1_HUMAN : WGDGCAQKNKPGVYTKVYNYVKWIKNTIAANS : 224
TRYP_PIG   : WGYGCAQKNKPGVYTKVCNYVNWIQQTIAAN- : 223
TRYU_DROME : WGEGCARPNYPGVYANVAYYKDWIAKQRTSYV : 235
TRYI_DROME : WGYRCADDNYPGVYADVAILRPWIVKAANAI- : 225
TRY1_SALSA : WGYGCAEPGNPGVYAKVCIFNDWLTSTMASY- : 222
```

Raw text files

```
                    *        20         *        40         *        60
TRY1_HUMAN : IVGGYNCEENSVPYQVSLN-------SGYHFCGGSLINEQWVVSAGHCYKSRIQVRLGEHNIEVLEGNE :  62
TRYP_PIG   : .....T.AA..I.......-------..S.........S......A................D...... :  62
TRYU_DROME : ....ADTSSYYTK.V.Q.RRRSSSSS.YAQT...CILDAVTIAT.A..VYN.EAENFLVVSGDDSR.GM :  69
TRYI_DROME : .I..SDQLIRNA.W...IQ------I.AR.E...VIYSKEIIIT....LHE.SVT-.MKVRVGAQNH.Y :  62
TRY1_SALSA : .....E.KAY.QTH.....-------..........V..N.....A......VE........K.T..S. :  62

               *        80         *       100         *       120         *       1
TRY1_HUMAN : --QFINAAKIIRHPQYDRKTLNNDIMLIKLSSRAVIN--ARVSTISLPTAPPATGTKCLISGWGNTASS : 127
TRYP_PIG   : --.........T..NFNGN..D..........P.TL.--S..A.V...RSCA.A..E.........K.. : 127
TRYU_DROME : YGVVVRVSQL.P.EL.NSS.MD...A.VVVDPPLPLDSFSTMEA.VIASEQ.PV.VQAT.....Y.KEN : 138
TRYI_DROME : GGTLVPV.AYKV.E.F.SRF.HY..AVLR..TPLTFG--LSTRA.N.ASTS.SG..TVTVT...H.DNG : 129
TRY1_SALSA : --...SSSRV....N.SSYNID.........KP.TL.--TY.QPVA...SCAPA..M.TV......M.. : 127

               40         *       160         *       180         *       200
TRY1_HUMAN : GADYPDELQCLDAPVLSQAKCEASYPG----KITSNMFCVGFLEGGKDSCQGDSGGPVVCNGQLQGVVS : 192
TRYP_PIG   : .SS..SL....K.....DSS.KS....----Q..G..I...........................I.. : 192
TRYU_DROME : .LSS-.Q..QVKV.IVDSE..QEA.YWR---P.SEG.L.A.LS.....A........L.VANK.A.I.. : 203
TRYI_DROME : --ALS.S..KAQLQIIDRGE.ASQKF.YGADFVGEETI.AASTD--A.A.T......L.ASS..V.I.. : 194
TRY1_SALSA : T..S-NK....NI.I..YSD.NN....----M..NA...A.Y...................E...... : 191

                 *       220         *
TRY1_HUMAN : WGDGCAQKNKPGVYTKVYNYVKWIKNTIAANS : 224
TRYP_PIG   : ..Y.............C...N..QQ.....- : 223
TRYU_DROME : ..E...RP.Y....AN.AY.KD..AKQRTSYV : 235
TRYI_DROME : ..YR..DD.Y....AD.AILRP..VKAAN.I- : 225
TRY1_SALSA : ..Y...EPGN....A..CIFND.LTS.M.SY- : 222
```

Consensus sequences

A consensus sequence is defined as a sequence derived by a multi-alignment that presets only most conserved residues for each position:

$\Rightarrow$ it **summarizes** a multi-alignment.

$\Rightarrow$ it **is different** to all proteins of the input dataset.

$\Rightarrow$ it is possible to define special symbols that specify not perfect conservation in some position.

$\Rightarrow$ it is possible to use a specific syntax that also allows to specify variants in a position, other than conservations.

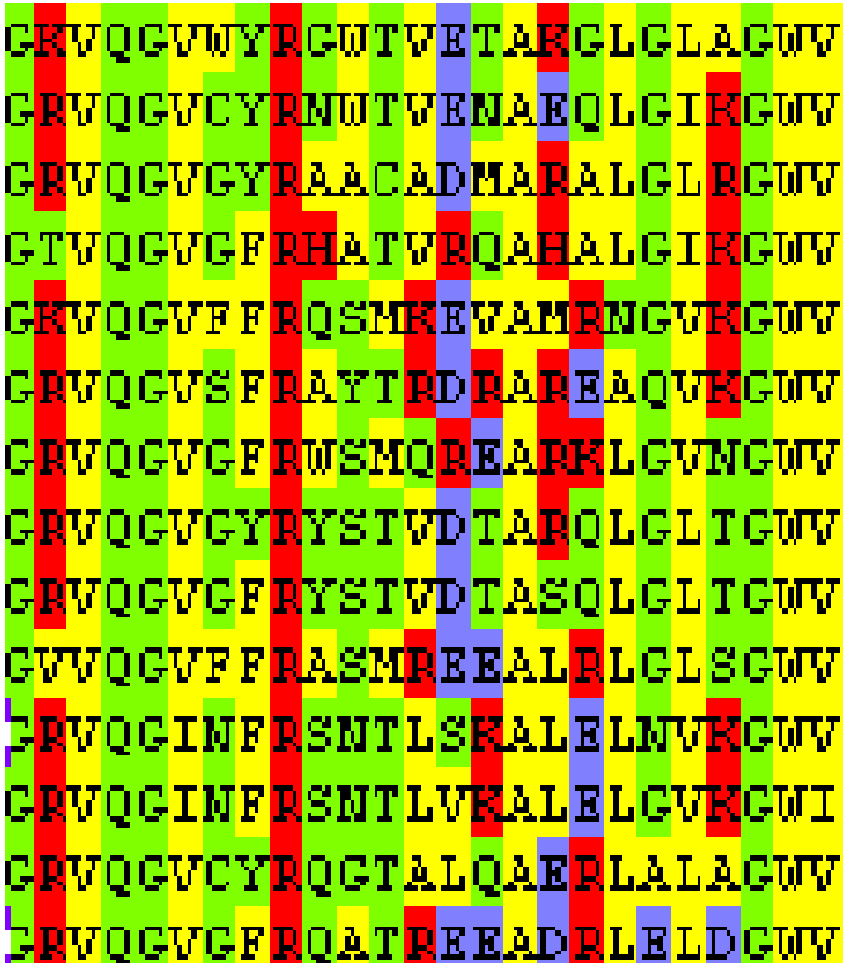
Consensus Sequences

To facilitate the reading of a multiple alignment and quickly highlight conserved amino acids, it can be represented the "consesus" sequence in the last line of a multi-alignment.

If all the sequences in a multiple alignment contains the same amino acid in a location than it is represented in the consensus sequence, in the same way if all the amino acids of a column belong to the same class, in the consensus line it is shown the symbol of class.

It can also be represented below the consensus sequence, the rows of the consensus that below a certain percentage (for example 90%, 80% or 70%). In these lines, to put a symbol it is sufficient that the amino acid is conserved in a population of amino acids of the upper column to the percentage of the consensus (rather than in all).

Example of consensus sequence representation



GRVQGV--R------A--LG--GWV

GRVQGh-aRvvvvvvAvvLGivGWV

GRVQG[VI]-[FY]R------A-L----GWY

GRVQGV--R-**6**A-LG--GWV

Exact consensus

Symbolic consensus

Consensu with variants

Consensus with repetitions

http://www.ebi.ac.uk/clustalw

```
PEEKSAVTALWGKVN--VDEVGG Hbb_human
GEEKAAVLALWDKVN--EEEVGG Hbb_horse
PADKTNVKAAWGKVGAHAGEYGA Hba_human
AADKTNVKAAWSKVGGHAGEYGA Hba_horse
EHEWQLVLHVWAKVEAGVAGHGQ Myg_whale
```