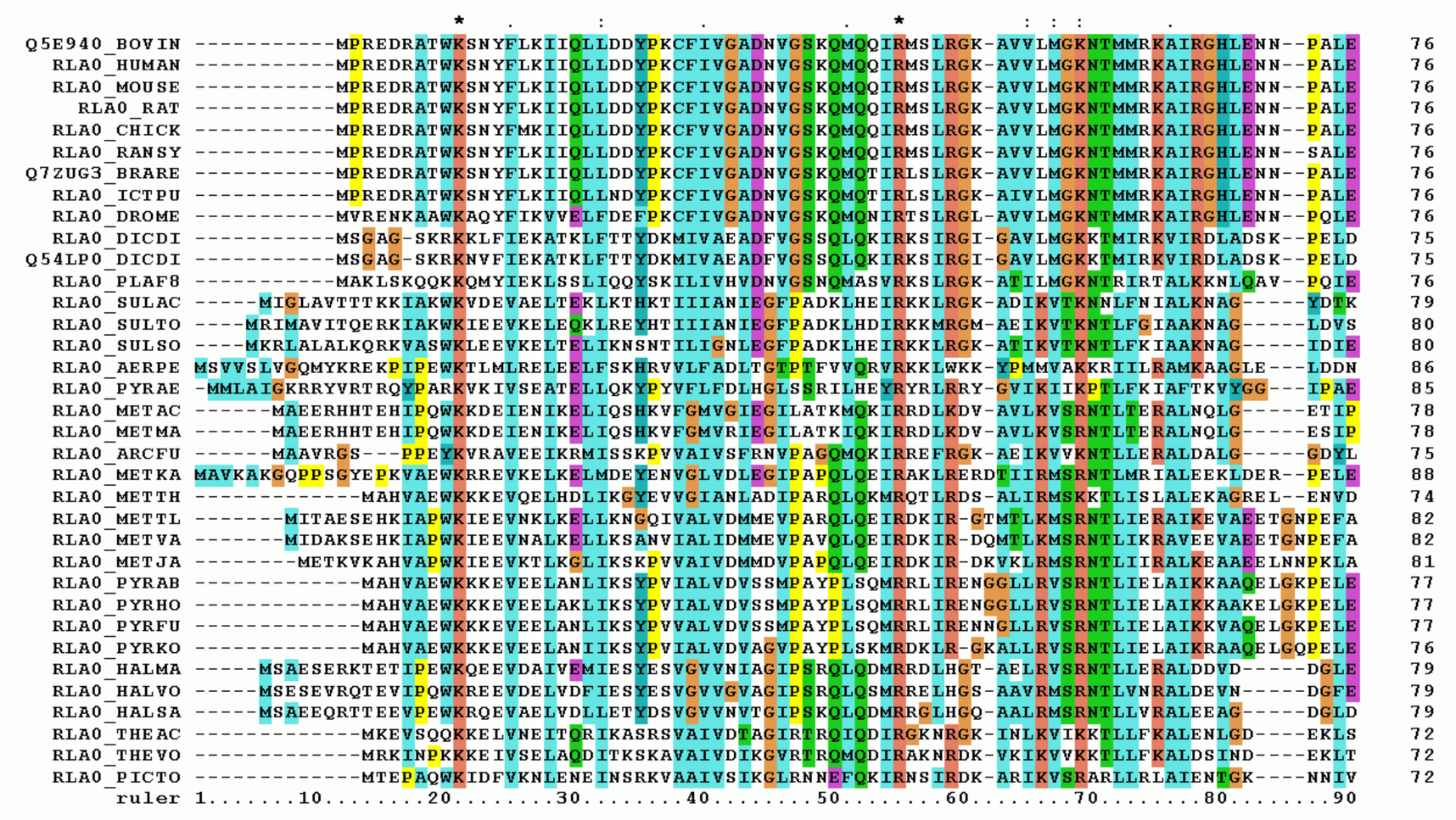
Multiple Sequence Alignment



Overview

Multiple Sequence Alignment

- introduction
- methods for Multiple Sequence Alignment (MSA)
- uses of MSA

Learning Objectives

- describe the uses of MSA
- explain how ClustalW performs MSA
- describe alternative methods for MSA (Muscle, ProbCons)
- understand how to interpret MSA results

ian.simpson@ed.ac.uk

Multiple Sequence Alignment (MSA)

- A multiple sequence alignment is a collection of >=3 nucleotide or protein sequences that are aligned
- These alignments can reveal domains of highly related sequence that in turn can can be used to define regions of interest
- For proteins that could be “domains” that define a protein family
- For nucleotides that could be conserved regions that define coding, regulatory or structural elements of a genome
- The assumption is that sequences that share similarity share function i.e. they are homologous
- MSAs can be used to “test” whether a novel sequence is a member of a family, contains a particular domain, or has some regulatory function in common with known regulators for example

ian.simpson@ed.ac.uk

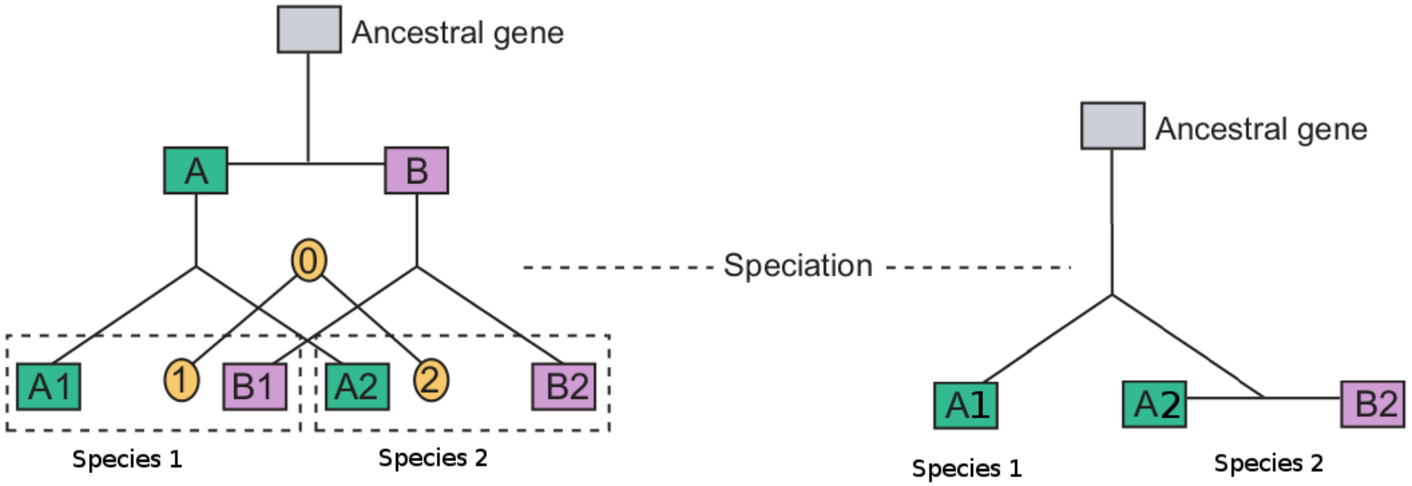
Homology (Reminder)

Orthologs are the result of a speciation event
Paralogs are the result of a duplication event

All are Homologs

Paralogs
A1 and B1
A1 and B2
A2 and B1
A2 and B2

Orthologs
A1 and A2
B1 and B2



All are Homologs

Paralogs
A2 and B2 are paralogs of each other

Orthologs
A2 and B2 are orthologs of A1

ian.simpson@ed.ac.uk

Multiple Sequence Alignment (MSA)

MSAs are easy to create for closely related sequences such as glyceraldehyde-3-phosphate dehydrogenase (GAPDH)

NP_002037.2	194	KLWRDGRGALQNIIPASTGAAKAVGKVIPELNGKLTGMAFRVPTANVSVV	243
XP_508955.1	194	KLWRDGRGALQNIIPASTGAAKAVGKVIPELNGKLTGMAFRVPTANVSVV	243
XP_001105471.1	194	KLWRDGRGALQNIIPASTGAAKAVGKVIPELNGKLTGMAFRVPTANVSVV	243
NP_001003142.1	192	KMWRDGRGAAQNIIPASTGAAKAVGKVIPELNGKLTGMAFRVPTPNVSVV	241
XP_003435697.1	192	KLWRDGRGAAQNIIPASTGAAKAVGKVIPELNGKLTGMAFRVPTPNVSVV	241
XP_003434435.1	192	KLWRDGRGAAQNIIPASTGAAKAVGKVIPELNGKLTGMAFCVPTPNVSVV	241
NP_001029206.1	192	KLWRDGRGAAQNIIPASTGAAKAVGKVIPELNGKLTGMAFRVPTPNVSVV	241
NP_032110.1	192	KLWRDGRGAAQNIIPASTGAAKAVGKVIPELNGKLTGMAFRVPTPNVSVV	241
XP_001476757.1	192	KLWRDGRGAAQNIIPASTGAAKAVGKVIPELNGKLTGMAFRVPTPNVSVV	241
NP_058704.1	192	KLWRDGRGAAQNIIPASTGAAKAVGKVIPELNGKLTGMAFRVPTPNVSVV	241
NP_989636.1	192	KLWRDGRGAAQNIIPASTGAAKAVGKVIPELNGKLTGMAFRVPTPNVSVV	241
NP_001108586.1	192	KLWRDGRGASQNIIPASTGAAKAVGKVIPELNGKLTGMAFRVPTPNVSVV	241
NP_001259584.1	191	KLWRDGRGAAQNIIPASTGAAKAVGKVIPALNGKLTGMAFRVPTPNVSVV	240
NP_525108.2	191	KLWRDGRGAAQNIIPAATGAAKAVGKVIPALNGKLTGMAFRVPTPNVSVV	240
XP_318655.2	191	KLWRDGRGAAQNIIPAATGAAKAVGKVIPALNGKLTGMAFRVPTPNVSVV	240
NP_496237.1	200	KLWRDGRGAGQNIIPASTGAAKAVGKVIPELNGKLTGMAFRVPTPDVSVV	249
NP_496192.1	200	KLWRDGRGAGQNIIPASTGAAKAVGKVIPELNGKLTGMAFRVPTPDVSVV	249
NP_508534.3	200	KLWRDGRGAGQNIIPASTGAAKAVGKVIPELNGKLTGMAFRVPTPDVSVV	249
NP_508535.1	200	KLWRDGRGAGQNIIPASTGAAKAVGKVIPELNGKLTGMAFRVPTPDVSVV	249
NP_012483.3	192	KDWRGGRTASGNIIPSSTGAAKAVGKVLPELQGKLTGMAFRVPTVDVSVV	241
NP_011708.3	192	KDWRGGRTASGNIIPSSTGAAKAVGKVLPELQGKLTGMAFRVPTVDVSVV	241
NP_012542.1	192	KDWRGGRTASGNIIPSSTGAAKAVGKVLPELQGKLTGMAFRVPTVDVSVV	241
XP_456022.1	191	KDWRGGRTASGNIIPSSTGAAKAVGKVLPELQGKLTGMAFRVPTVDVSVV	240
NP_596154.1	194	KDWRGGRGASANIIPSSTGAAKAVGKVIPALNGKLTGMAFRVPTPDVSVV	243
NP_595236.1	194	KDWRGGRGASANIIPSSTGAAKAVGKVIPALNGKLTGMAFRVPTPDVSVV	243
XP_003717853.1	192	KDWRGGRGAAQNIIPSSTGAAKAVGKVIPALNGKLTGMSMRVPTANVSVV	241
XP_956977.1	193	KDWRGGRTAAQNIIPSSTGAAKAVGKVIPDLNGKLTGMAMRVPTANVSVV	242
NP_001060897.1	196	KDWRGGRAASFNIIPSSTGAAKAVGKVLPDLNGKLTGMSFRVPTVDVSVV	245
NP_001004949.1	152	KLWRDGRGAGQNIIPASTGAAKAVGKVIPELNGKLTGMAFRVPTPNVSVV	201

ian.simpson@ed.ac.uk

Multiple Sequence Alignment (MSA)

MSAs can be challenging to create for more diverse members of a protein family such as the transcription factor Pax6 which is present in hugely divergent species

```
NP_001595.2          17   NGRPLPDSTRQKIVELAHSGARPCDISR---ILQTHADAKVQVLDNQNVS   63
XP_003954413.1       11   ---LYLIDSRELIAEVGTGWQGDEDALWYIFVLXTHADAKVQVLDNQNVS   57
XP_001084865.2       51   NGRPLPDSTRQKIVELAHSGARPCDISR---ILQTHADAKVQVLDNQNVS   97
NP_001035735.1       17   NGRPLPDSTRQKIVELAHSGARPCDISR---ILQ--------------VS   49
NP_001231127.1       17   NGRPLPDSTRQKIVELAHSGARPCDISR---ILQTHADAKVQVLDNENVS   63
NP_037133.1          17   NGRPLPDSTRQKIVELAHSGARPCDISR---ILQ--------------VS   49
NP_990397.1          17   NGRPLPDSTRQKIVELAHSGARPCDISR---ILQTHADAKVQVLDNQNVS   63
NP_571379.1          36   NGRPLPDSTRQKIVELAHSGARPCDISR---ILQTHADAKVQVLDNENVS   82
NP_524638.3          42   NGRPLPDSTRQKIVELAHSGARPCDISR---ILQ--------------VS   74
XP_311087.5          33   NGRPLPDSTRQKIVELAHSGARPCDISR---ILQ--------------VS   65
NP_001024570.1       18   NGRPLPDATRQRIVDLAHKGCRPCDISR---LLQ--------------VS   50
NP_001006763.1       17   NGRPLPDSTRQKIVELAHSGARPCDISR---ILQ--------------VS   49
```


ian.simpson@ed.ac.uk

Multiple Sequence Alignment (MSA)

Some features of proteins are revealed by MSAs

conserved structural features (for example)

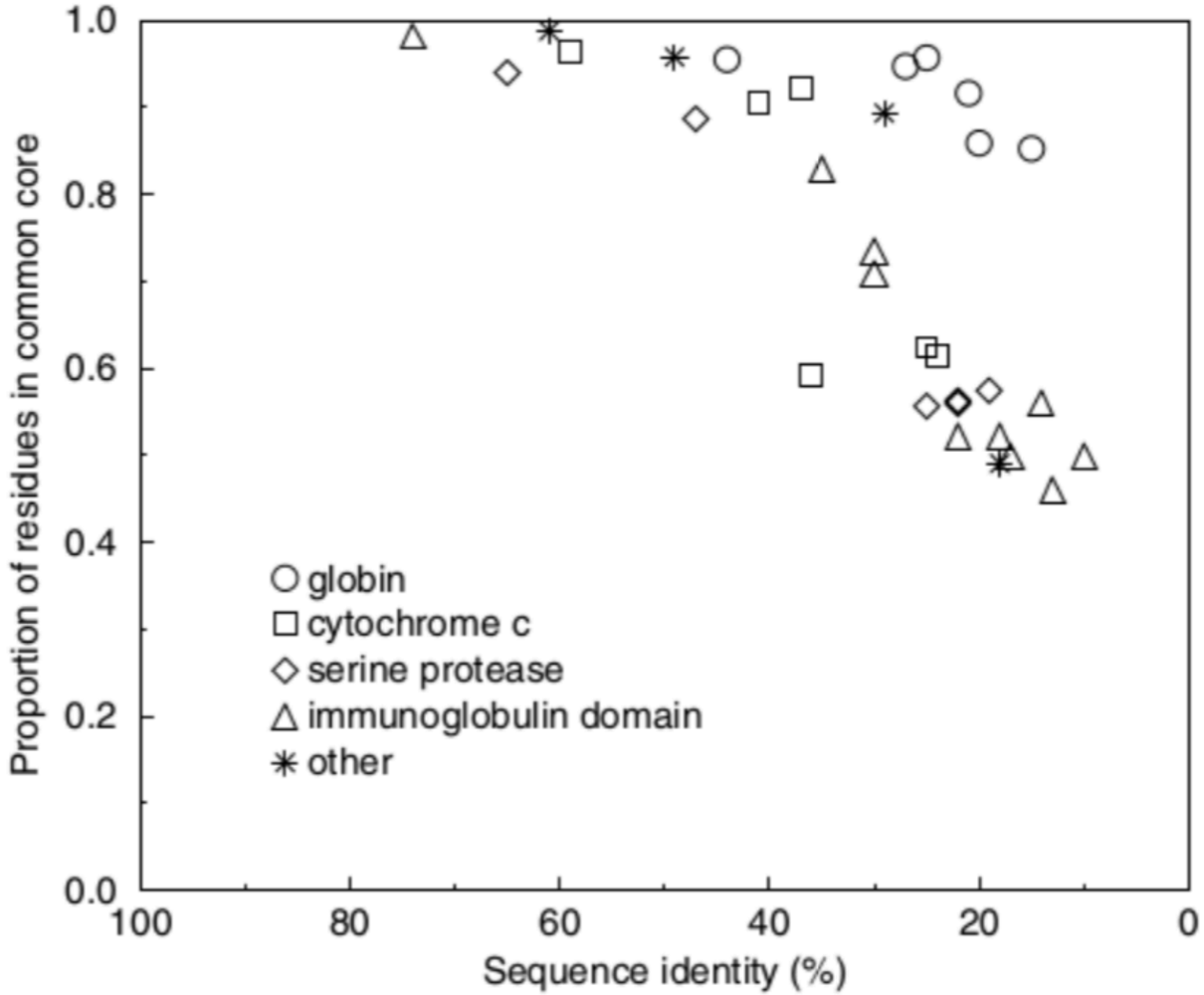
- alpha helices
- beta-sheets
- disulfide bridges
- active sites of enzymes
- transmembrane domains

Other parts of sequences may not play a functional/structural role, are not conserved and are not aligned well in MSAs

- linker/spacer regions of proteins
- non-coding, non-regulatory regions of genomic DNA
- dis-ordered parts of proteins

ian.simpson@ed.ac.uk

Sequence vs. Structure



For proteins with significant identity (>30%) where the structure is known most residues are super-imposable - there are many different sequences that can converge on similar core structures.

ian.simpson@ed.ac.uk

What can we use MSAs for?

- To infer function. If we have an MSA for a well defined protein family/domain and we can align an unknown sequence to it
- The vast majority of “functional” knowledge about genomes, genes and proteins is actually inferred from such alignments
- As greater numbers of genomes are being sequenced statistical models can be generated such as profile-HMMs and used to describe molecular patterns. These are built from multiple sequence alignments and are crucial to identify proteins present in novel genomes
- These profiles are far more sensitive that pairwise alignment methods such as BLAST. Variants of BLAST (such as DELTA-BLAST) exploit this (NB last week’s lecture!)

ian.simpson@ed.ac.uk

What can we use MSAs for?

- Finding the “functional” parts of sequences
  - such parts are commonly conserved and so align “better” in the MSA than “non functional” parts

- Prediction of potentially deleterious mutation
  - highly conserved residues in MSAs more commonly mark important positions in a regulatory element or protein

- Understanding the evolution of organisms, proteins, genomes
  - MSAs are the first step in performing phylogenetic inference.

- To find possible gene regulatory regions in genomes

ian.simpson@ed.ac.uk

Main Approaches to MSA

- Exact Approaches
- Progressive Sequence Alignment
- Iterative Approaches
- Consistency-Based Approaches
- Structure-Based Approaches

ian.simpson@ed.ac.uk

Exact MSA

In effect performing global pairwise alignment in N-dimensions (NW type Dynamic Programming)

memory complexity $O_m(L^N)$

time complexity $O_t(2^N L^N)$

e.g. N=10, L=1000: $O_m(10^{30})$, $O_t(10^{33})$

For any reasonably sized set of sequences this is computationally intractable so alternative approaches are needed

ian.simpson@ed.ac.uk

Progressive Sequence Alignment

Heuristic approach, not guaranteed to find the optimal alignment

Principle; align sequences pairwise then begin building the MSA starting with the closest pair and then sequentially adding sequences

ClustalW (Thompson et al. 1994) based on progressive alignment idea of Feng & Doolittle (1987,1990)

Step 1 - generate pairwise alignments of all sequences:

SeqA	Name	Len(aa)	SeqB	Name	Len(aa)	Score
1	beta_globin	147	2	myoglobin	154	25
1	beta_globin	147	3	neuroglobin	151	15
1	beta_globin	147	4	soybean	144	13
1	beta_globin	147	5	rice	166	21
2	myoglobin	154	3	neuroglobin	151	16
2	myoglobin	154	4	soybean	144	8
2	myoglobin	154	5	rice	166	12
3	neuroglobin	151	4	soybean	144	17
3	neuroglobin	151	5	rice	166	18
4	soybean	144	5	rice	166	43 ← 1

ian.simpson@ed.ac.uk

Progressive Sequence Alignment

Scoring the alignments

By default ClustalW just uses percentage identity, but better to calculate distance matrices as these then allow the construction of a "guide tree" to guide the alignment process (i.e. decide in which order to add sequences)

$$D = -\ln S_{eff}$$

$$S_{eff} = \frac{S_{\text{real}(ij)} - S_{\text{rand}(ij)}}{S_{\text{iden}(ij)} - S_{\text{rand}(ij)}} \times 100$$

where D is distance, S_{eff} is the a normalised similarity score. S_{eff} is calculated from $S_{real(i,j)}$ the actual similarity between sequences (i,j), $S_{iden(i,j)}$, the average score aligning i and j to themselves (in effect a form of length normalisation) and $S_{rand(i,j)}$ the mean alignment score for many random shuffles of the sequences to be aligned

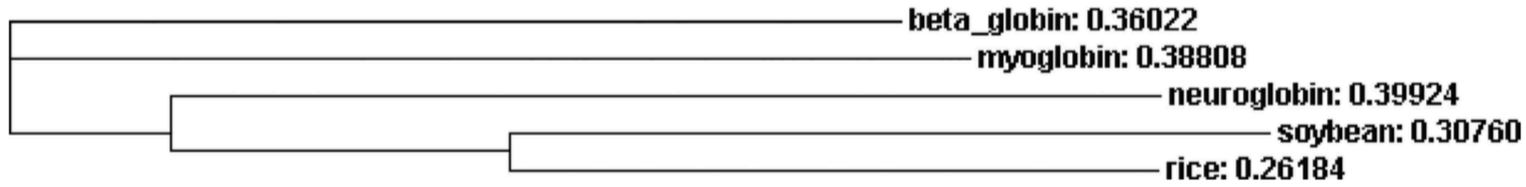
ian.simpson@ed.ac.uk

Progressive Sequence Alignment

The final distance score depends on S_{eff} so that if $S_{eff} = 0$, D = infinite and if $S_{eff} = 1$ (they have perfect similarity) D = 0.

Step 2 - Guide Tree

Using the Distance Matrix (pairwise distances) create a guide tree for the alignment, for example using the Unweighted Pair Group Method of Arithmetic Averages (UPGMA)



So order :- rice:soybean -> neuroglobin ->beta-globin -> myoglobin

ian.simpson@ed.ac.uk

Progressive Sequence Alignment

Step 3 - Creation of the MSA
Using the guide tree the closest pair are aligned by DP (e.g. NW) and then the next closest sequence is added to the alignment to produce a profile. This continues sequentially until all sequences have been added

When adding an additional sequence it is essentially a dynamic programming approach, but the gap extension/deletion penalties are dynamically calculated weighted by pre-existing gaps in the more closely related sequence alignments

There are several approaches to this ranging from ones that absolutely preserve gaps in earlier aligned sequences to those that moderate for those so as not to have highly related sequences biasing alignment of more distantly related ones

ian.simpson@ed.ac.uk

Progressive Alignment Example

For example, say your sequences are:

S_1	A	T	T	G	C	C	A	T	T	
S_2	A	T	G	G	C	C	A	T	T	
S_3	A	T	C	C	A	A	T	T	T	T
S_4	A	T	C	T	T	C	T	T		
S_5	A	C	T	G	A	C	C			

Perform the pairwise alignments

S_1	A	T	T	G	C	C	A	T	T	
S_2	A	T	G	G	C	C	A	T	T	

S_1	A	T	T	G	C	C	A	T	T	-	-
S_3	A	T	C	-	C	A	A	T	T	T	T

S_1	A	T	T	G	C	C	A	T	T	
S_4	A	T	C	T	T	C	-	T	T	

S_1	A	T	T	G	C	C	A	T	T	
S_5	A	C	T	G	A	C	C	-	-	

ian.simpson@ed.ac.uk

Progressive Alignment

Create a Distance matrix and then guide tree (for this simple example just a score matrix)

	S_1	S_2	S_3	S_4	S_5	
S_1	-	7	-2	0	-3	**2**
S_2	7	-	-2	0	-4	**1**
S_3	-2	-2	-	0	-7	**-11**
S_4	0	0	0	-	-3	**-3**
S_5	-3	-4	-7	-3	-	**-17**
	2	**1**	**-11**	**-3**	**-17**	

S_1 is the sequence most similar to the rest, and below are the best alignments between S_1 and the rest of the sequences.

ian.simpson@ed.ac.uk

Performing the Progressive Alignment

Let's use the alignment of S_1 and S_2.

S_1	A	T	T	G	C	C	A	T	T
S_2	A	T	G	G	C	C	A	T	T

S_1 and S_2 are aligned

Now, let's add S_3, using its alignment to S_1.

S_1	A	T	T	G	C	C	A	T	T	-	-
S_2	A	T	G	G	C	C	A	T	T	-	-
S_3	A	T	C	-	C	A	A	T	T	T	T

S_1, S_2, and S_3 are aligned

Then, let's add S_4, using its alignment to S_1.

S_1	A	T	T	G	C	C	A	T	T	-	-
S_2	A	T	G	G	C	C	A	T	T	-	-
S_3	A	T	C	-	C	A	A	T	T	T	T
S_4	A	T	C	T	T	C	-	T	T	-	-

S_1, S_2, S_3, and S_4 are aligned

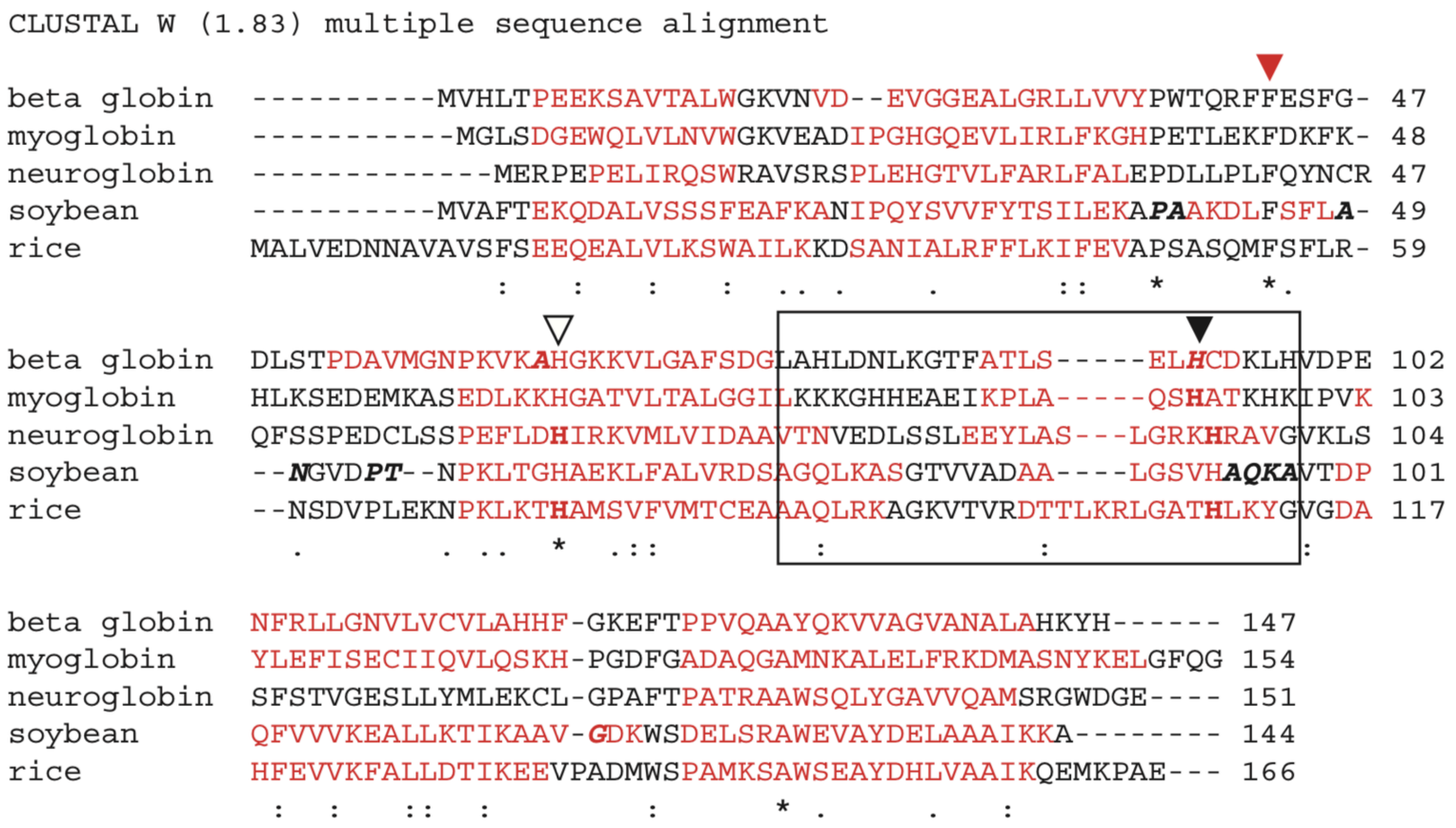
Finally, let's add S_5, using its alignment to S_1.

S_1	A	T	T	G	C	C	A	T	T	-	-
S_2	A	T	G	G	C	C	A	T	T	-	-
S_3	A	T	C	-	C	A	A	T	T	T	T
S_4	A	T	C	T	T	C	-	T	T	-	-
S_5	A	C	T	G	A	C	C	-	-	-	-

S_1, S_2, S_3, S_4 and S_5 are aligned

ian.simpson@ed.ac.uk

Progressive Sequence Alignment



* (100% conserved), : (conservative substitution), . (less conservative subst.), and arrowheads (highly conserved residues)

ian.simpson@ed.ac.uk

Iterative Alignment

Principle; create an initial progressive alignment and then optimise it using dynamic programming and a scoring function

Unlike progressive alignment itself, iterative alignment can correct errors made during the initial alignment. Poor guide tree, misplaced gaps etc.

Optimisation needs a scoring function for example the sum-of-pairs score:

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

For each column calculate the pairwise score for every combination. Then sum across all columns.

So for Column 3; P->R, P->-, P->P, R->-, R->P, - ->P

[given a simple scoring scheme of m=+1, mm=-1, gap=-2]

Pair Score = sum(-1, -2, +1, -2, -1, -2) = -6

Summing all pair scores across columns gives -20

ian.simpson@ed.ac.uk

Iterative Alignment

MUSCLE - 2004
High accuracy and exceptional speed; in the original paper 1000 protein sequences of 282aa were aligned in 21 seconds on a standard PC

Step 1 - Generate a draft progressive alignment, this can use identity or k-mer profiling of sequences; e.g. frequency fingerprints for k-mers between sequences (very fast)

ACGCTACGCATACGCATG

ACG TAC CAT CGC
CGC ACG ATA GCA
GCT CGC TAC CAT
CTA GCA ACG ATG

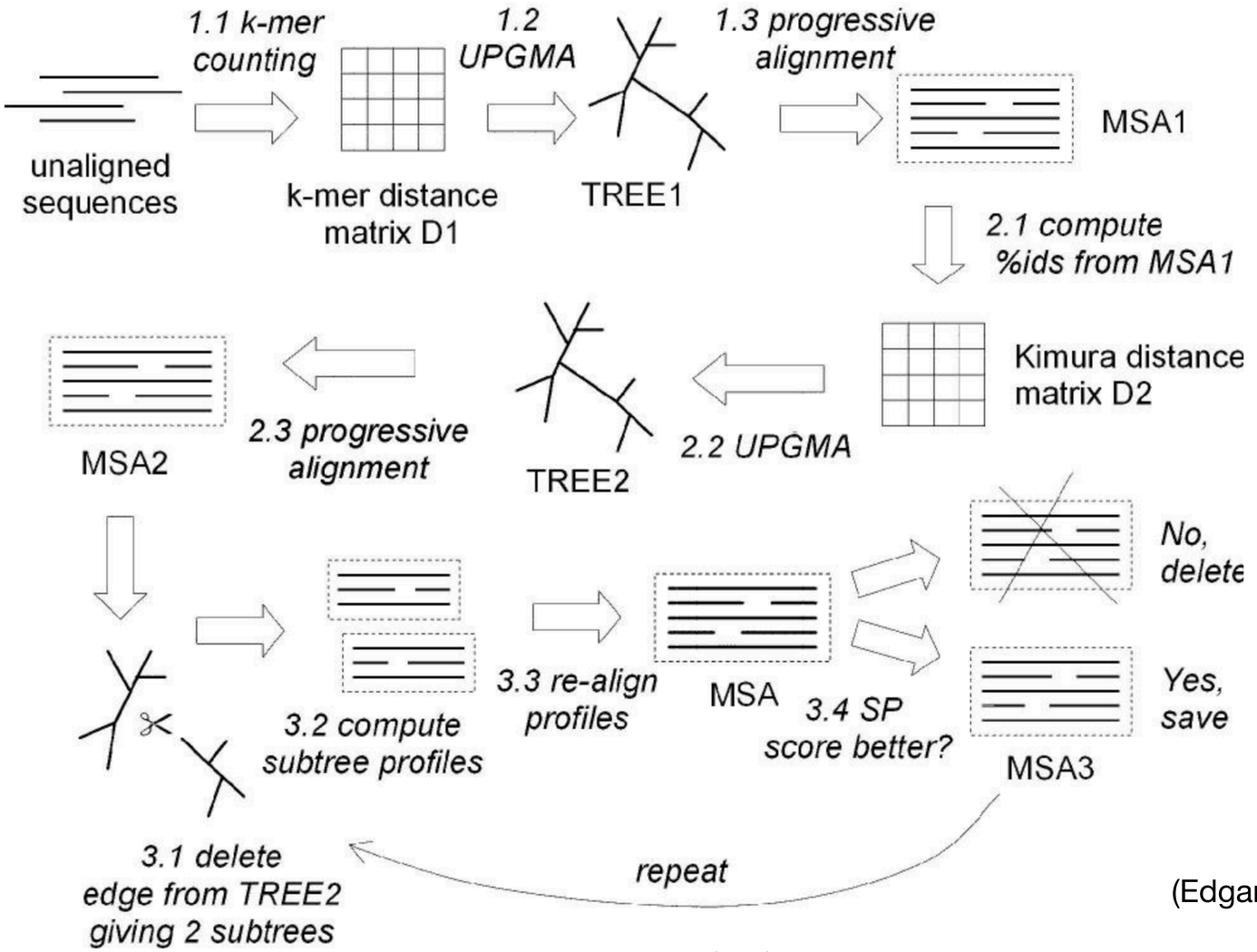
k-mer profiling (3-mers)

A guide tree is then built from either of these using for example UPGMA. Sequences are progressively added to the alignment as before

ian.simpson@ed.ac.uk

Iterative Alignment

Schematic of **MUSCLE** methodology



(Edgar 2004)

ian.simpson@ed.ac.uk

Iterative Alignment

Multiple Sequence Comparison by Log Expectation (**MUSCLE**) (Edgar 2004)

Step 2 - MUSCLE next improves the tree and builds a new progressive alignment. It does this by calculating a new distance matrix by using the Kimura distance (this attempts to model the possibility of multiple changes in sequence happening at the same position).

Using the new distance matrix a new tree is created and a new progressive alignment made.

Step 3 - Re-partitioning. Next for every branch in the tree MUSCLE creates sub-partitions and creates profiles (sub-alignments) of each and these two profiles are then aligned to create a new MSA. Consider this a set of local MSAs testing every possible bi-partition of the tree.

For every tree created this way the PSP is compared to the previous best. If it is worse the MSA from that bi-partition is deleted if better it replaces the MSA.

$$d_{\text{Kimura}} = -\log_e (1 - D - D^2/5)$$

ian.simpson@ed.ac.uk

MUSCLE profile-profile Alignment

In practice sum of pair (SP) scores are normally calculated as probabilistic scores and scoring functions use scoring matrices such as PAM, BLOSUM.

$$S_{ij} = log(p_{ij}/p_i p_j)$$

$$PSP^{xy} = \sum_i \sum_j f_i^x f_j^y S_{ij}$$

For amino acid types i and j, p_i is the background probability of i, p_{ij} is the joint probability of i and j being aligned, S_{ij} is the score from a substitution matrix, f_i^x is the observed frequency of i in column x of the first profile, PSP is a sequence-weighted sum of substitution matrix scores for each pair of letters (one from each column that is being aligned in a pairwise fashion). The PSP function maximises the sum-of-pairs objective score.

ian.simpson@ed.ac.uk

Iterative Alignment

MUSCLE uses a log expectation (LE) score that is in effect a PSP modified to score profile-to-profile alignments

$$LE^{xy} = (1 - f_G^x)(1 - f_G^y) \log \sum_i \sum_j f_i^x f_j^y \frac{p_{ij}}{p_i p_j}$$

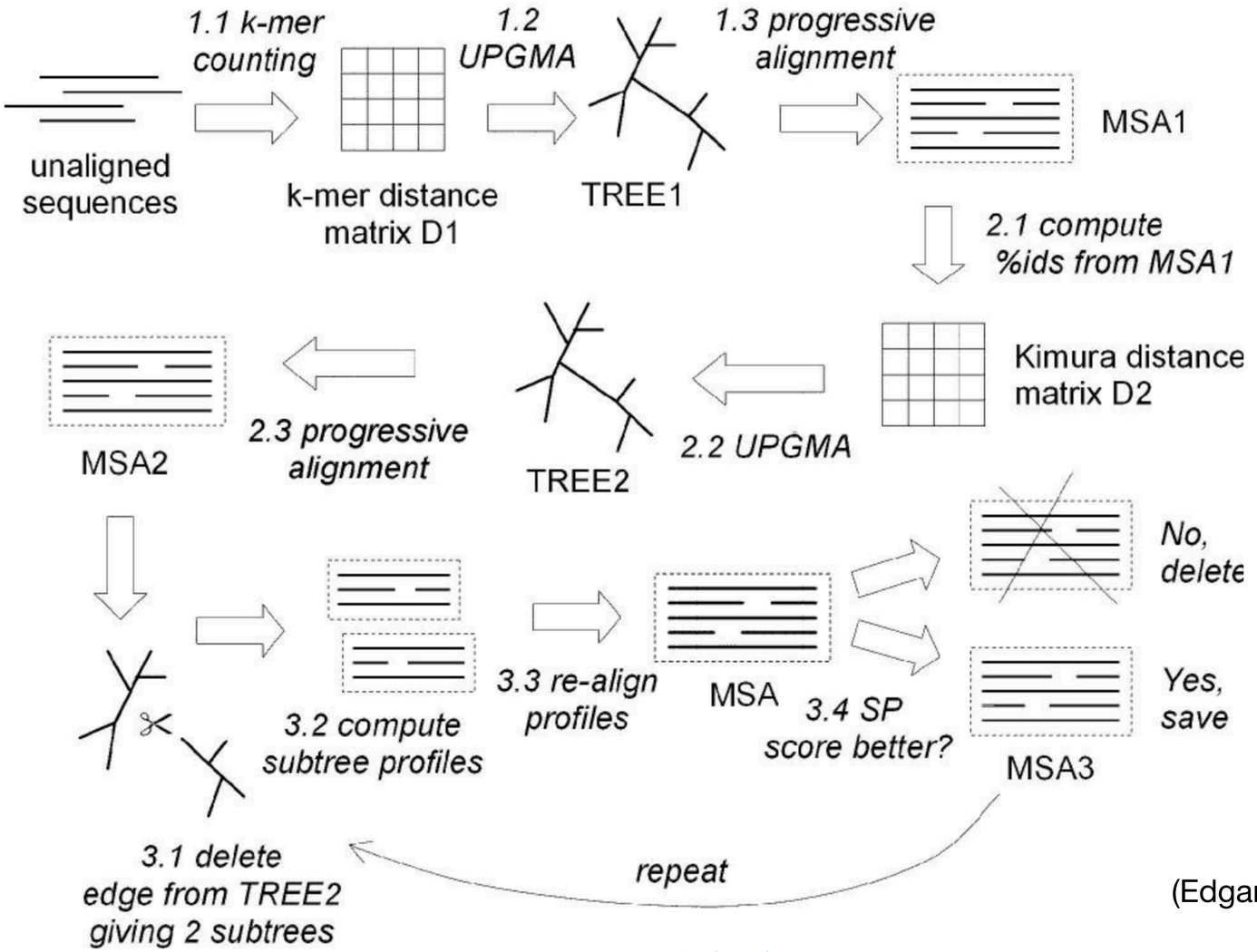
For profiles X and Y the LE is calculated. Columns are compared between profiles. (1 - f_G^x) is the occupancy of a column of profile X. Note the summed term from PSP is logged. This has been shown to increase accuracy of the alignment by rewarding columns that are full (better aligned) over those that have gaps.

MUSCLE uses PAM matrices in its PSP function to better refine the amino acid pair scoring.

ian.simpson@ed.ac.uk

Iterative Alignment

Schematic of **MUSCLE** methodology



(Edgar 2004)

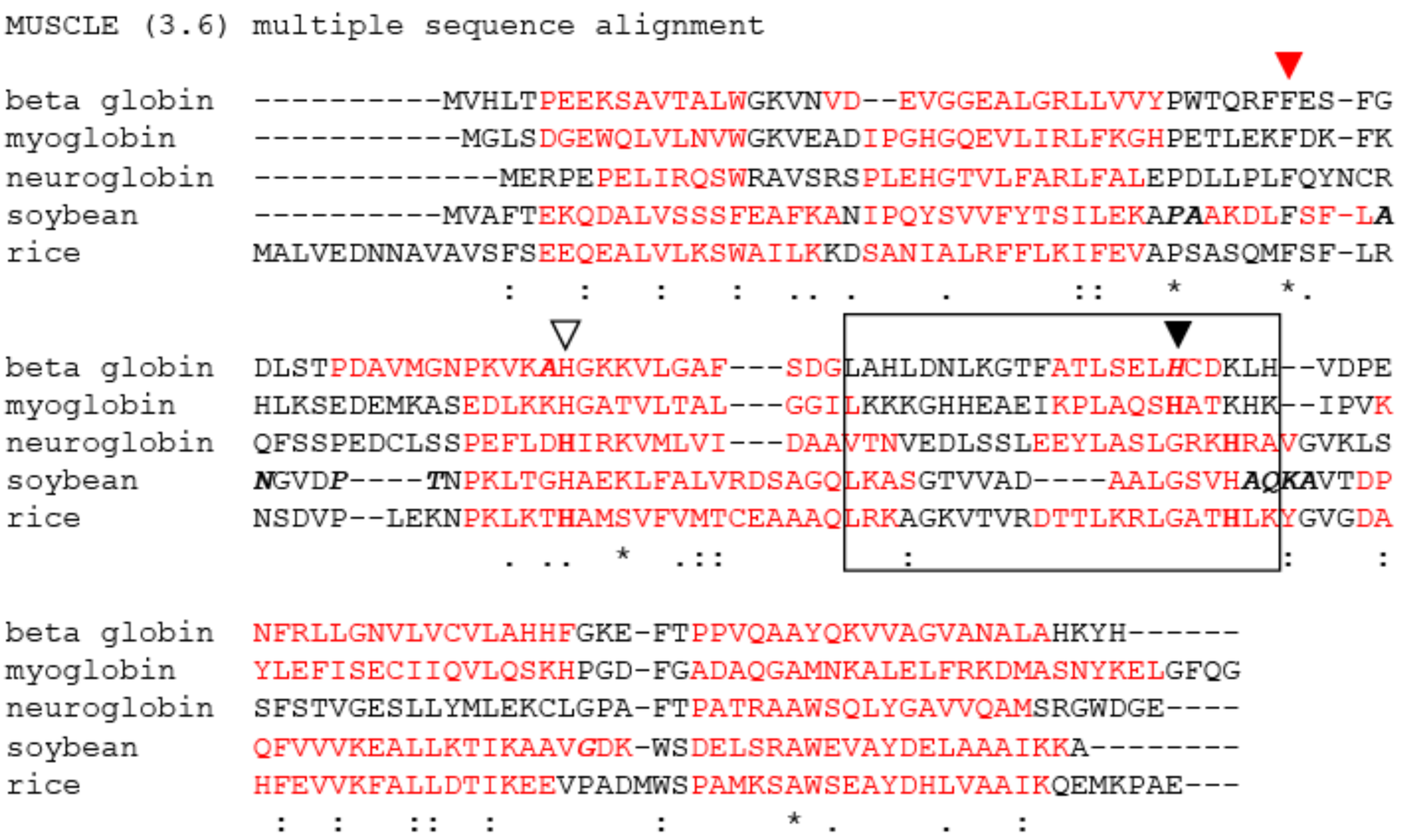
ian.simpson@ed.ac.uk

Progressive Sequence Alignment

* (100% conserved) : (conservative substitution) . (less conservative subst.), arrowheads (highly conserved residues)

ian.simpson@ed.ac.uk

Iterative Alignment



ian.simpson@ed.ac.uk

Consistency-Based Approaches

Principle; if for sequences x, y and z if amino acid x_i aligns with z_k and y_j aligns with z_k then x_i should align with y_j

This can be expressed as a conditional probability:-

$$P(x_i \sim y_j | \, x, y)$$

$$P(x_i \sim y_j | \, x, y, z) \approx \sum_k P(x_i \sim z_k \, | \, x, z) P(y_i \sim z_k \, | \, y, z)$$

It means that MSAs are adjusted as sequences are being added based on the "column context" of the alignment. This is a fundamentally different approach than we have seen so far.

ian.simpson@ed.ac.uk

Consistency-Based Approaches

Probabilistic Consistency-Based Alignment (**ProbCons**)

Step 1 - ProbCons first uses a paired-hidden Markov Model to calculate the posterior probability matrices for every pair of sequences to make up the MSA (i.e. a probabilistic scoring matrix for each alignment)

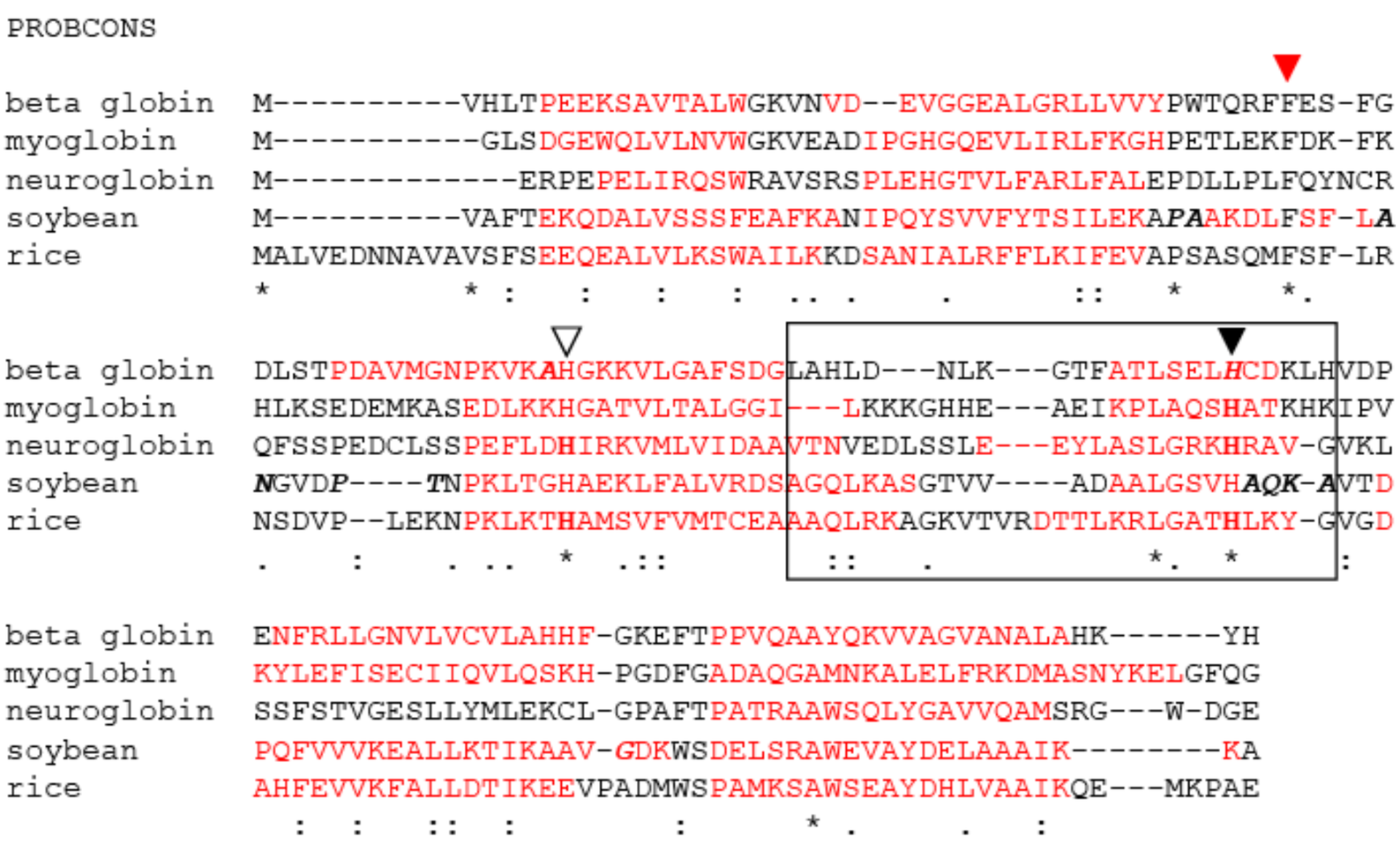
Step 2 - Using a NW type approach pairwise alignments are made using the posteriors from Step1, these are then refined using consistency information harvested from across all the pairwise alignments

Step 3 - A guide tree is then made and progressive alignment performed in much the same way as in previous methods.

Step 4 - This process can be iterated until the MSA scores have converged.

ian.simpson@ed.ac.uk

Consistency-Based Alignment



ian.simpson@ed.ac.uk

Structural-Based Approaches

Principle; incorporation of protein structural information can improve MSAs

Sequences to be aligned are first used in BLAST searches of the Protein Data Bank (PDB). Matches are then used to guide the creation of the multiple sequence alignment.

The Expresso tool from the T-COFFEE alignment suite does this and works on the principle that tertiary structure of proteins evolve much more slowly than lower order features.

Ultimately this allows for different amino acids producing similar higher order protein structures (which can commonly happen).These structural approaches try to capture that by incorporating actual crystal structure data.

Whilst powerful, this approach is limited by the available crystal structures for proteins. [NB new structure prediction tools such as AlphaFold2 are changing this]

Databases of MSAs

A lot of databases exist where MSAs for protein families have already been computed using some of the tools we have looked at:-

PFAM
http://pfam.xfam.org/
Protein domains/families identified by profile-HMM

SMART - Simple Modular Architecture Research Tool
http://smart.embl-heidelberg.de

CDD - Conserved Domain Database
https://www.ncbi.nlm.nih.gov/cdd

ian.simpson@ed.ac.uk

PFAM protein family MSA

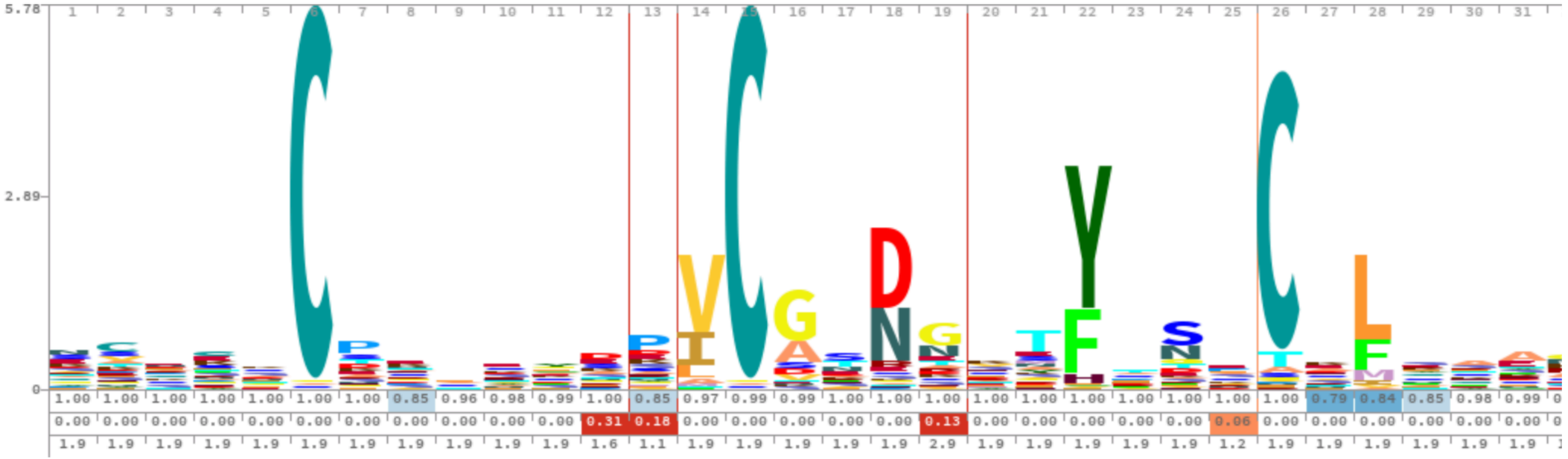
Kazal2 - Serine Protease Inhibitor

```
IACS_PIG/43-88              SHLFFCT.REMD..P.ICGTNG...KSYANP.CIFCSEKLGRNEKFD..........FGHWGHC
CO6_HUMAN/788-837           SPEEDCSHHSED....LCVFDTDSNDYFTSPACKFLAEKCLNNQQLH..........FLHIGSC
Q95011_CAEEL/325-376        STCITCPKDEKK.IP.ICDNRN...MTHPTL.CSFIQYNCEARNNED..EERV.....LVHIKSC
HTRA4_HUMAN/103-152         PSTCGCPTLGGA....VCGSDR...RTYPSM.CALRAENRAARRLGKVPAV.......PVQWGNC
CFAI_RAT/65-109             KLPYQCP..KAG.TP.VCATNG...RGYPTY.CHLKSFEC.LHPEIK..........FSNNGTC
Q9VUD9_DROME/293-342        KCNFQCPKASLS....ICASNGKCVVNFPGQ.CELSQWNCFNTKNVFH..........QVHDAEC
Q9VUD8_DROME/255-310        NCDELCTKFEYS..P.ICAHNGICIHEFANQ.CVMNTFNCKHRDLSFRAVDED.....VCRLGVC
AGRIN_RAT/92-137            CKKNACPATVAP....VCGSDA...STYSNE.CELQRAQCNQQRRIR..........LLRQGPC
HTRA3_MOUSE/82-132          CVRGVCR.CRWT..HTVCGTDG...HTYADV.CAL.QAASRRALQVSGTPVR......QLQKGAC
Q95011_CAEEL/133-183        DCNHNCTNTEFD..P.VCDTNG...SVYRNL.CVFQMRRCELQLESQR..IQL.....AEDRKFC
FSTL3_MOUSE/118-165         ECVPNCEGLPAG.FQ.VCGSDG...ATYRDE.CELRTARCRGHPDLR..........VMYRGRC
FST_PIG/117-164             VCAPDCSNITWK.GP.VCGLDG...KTYRNE.CALLKARCKEQPELE..........VQYQGKC
THBI_RHOPR/55-101           DVCQECDGDEYK..P.VCGSDD...ITYDNN.CRLECASISSSPGVE..........LKHEGPC
ISK5_HUMAN/106-151          DGDFICP.DYYE..A.VCGTDG...KTYDNR.CALCAENAKTGSQIG..........VKSEGEC
IOV7_CHICK/302-347          EAITACPFILQE....VCGTDG...VTYSND.CSLCAHNIELGTSVA..........KKHDGRC
F1PR78_CANLF/19-64          GSQIACP.RHLQ..P.ICGTDH...KTYSNE.CMFCALTLNKKFEVR..........KLQDTAC
IOVO_MELGA/81-126           KVMILCN.KALN..P.VCGTDG...VTYDNE.CVLCAHNLEQGTSVG..........KKHDGEC
Q9W271_DROME/465-512        SASCHCDYVHYA..P.VCSADN...ITFISA.C...HAGCSERTKDA..LGRT.....IYTGCEC
Q9W269_DROME/541-588        NSACSCDYVRYS..P.VCGENN...MTYISA.C...HAGCKKLLVNSE.GKK......IFYDCSC
SO2A1_MOUSE/444-493         RRDCLCPDSVFH..P.VCGDNG...VEYLSP.C...HAGCSSLNVSSAASKQP.....IYLNCSC
SO1B3_HUMAN/460-506         NSECNCDESQWE..P.VCGNNG...ITYLSP.C...LAGCKSSSGIK...KHT.....VFYNCSC
SO1B2_MOUSE/455-503         NSDCICDKNQWE..P.VCGENG...VTYISP.C...LAGCKSFRGDKKLMNI......EFYDCSC
SO1C1_HUMAN/477-523         NSRCKCSETKWE..P.MCGENG...ITYVSA.C...LAGCQTSNRSG...KNI.....IFYNCTC
SO1A1_RAT/440-486           NTRCSCSTNTWD..P.VCGDNG...VAYMSA.C...LAGCKKFVGTG...TNM.....VFQDCSC
SO1A2_HUMAN/440-486         NVDCNCPSKIWD..P.VCGNNG...LSYLSA.C...LAGCETSIGTG...INM.....VFQNCSC
SO4A1_HUMAN/505-553         NAACSCQPEHYS..P.VCGSDG...LMYFSL.C...HAGCPAATETNVDGQK......VYRDCSC
Q8SY02_DROME/481-527        NSNCGCSRTNYD..P.ICGVDG...VMYYSP.C...YAGCVQEEHAN...SLK.....RYHNCSC
Q9D5W6_MOUSE/481-527        NVPCGCTTSEYN..P.VCGRDE...TQYFSP.C...FAGCKATKKLR...KEK.....TYYNCSC
A0A0G2K9W9_RAT/488-534      NYHCACTTSLYS..S.VCGRDE...KEYFSP.C...FAGCSATKVQQ...NEK.....TYYNCSC
SO5A1_HUMAN/556-603         NVNCGCKIHEYE..P.VCGSDG...ITYFNP.C...LAGCVNSGNLST.GIR......NYTECTC
Q9VP16_DROME/514-563        YCEKICANVYDENDEIICGSDG...YMYTGE.TQLQCYSSCLNISVT..........IKSKGSC
TICN3_HUMAN/137-183         STCKQCPVVYPS..P.VCGSDG...HTYSFQ.CKLEYQACVLGKQIS..........VKCEGHC
TICN2_MOUSE/134-180         SVCKPCHMAQLA..S.VCGSDG...HTYSSV.CKLEQQACLSSKQLA..........VRCEGPC
TICN1_HUMAN/134-180         VKCKPCPVAQSA..M.VCGSDG...HSYTSK.CKLEFHACSTGKSLA..........TLCDGPC
FST_PIG/191-239             TCNRICPEPTSSEQY.LCGNDG...VTYSSA.CHLRKATCLLGRSIG..........LAYEGKC
FSTA_DANRE/272-319          VCAESCPESRSE.EA.VCASDN...TTYPSE.CAMKQAACSLGVLLE..........VKHSGSC
FSTL1_MOUSE/51-96           LCIEQCK.PHKR..P.VCGSNG...KTYLNH.CELHRDACLTGSKIQ..........VDYDGHC
SPRL1_MOUSE/443-495         QDPETCPPAKIL.DQ.ACGTDN...QTYASS.CHLFATKCRLEGTKKGHQLQ......LDYFGAC
SPRC_RABIT/98-151           QDPTSCPAPVGEFEK.VCSNDN...KTFDSS.CHFFATKCTLEGTKKGHKLH......LDYIGPC
SPRC_CAEEL/75-135           ECISKCPELDGDPMDKVCANNN...QTFTSL.CDLYRERCLCKRKSKECSKAFNAKVHLEYLGEC
```


ian.simpson@ed.ac.uk

PFAM family profile-HMM

Kazal2 - Serine Protease Inhibitor

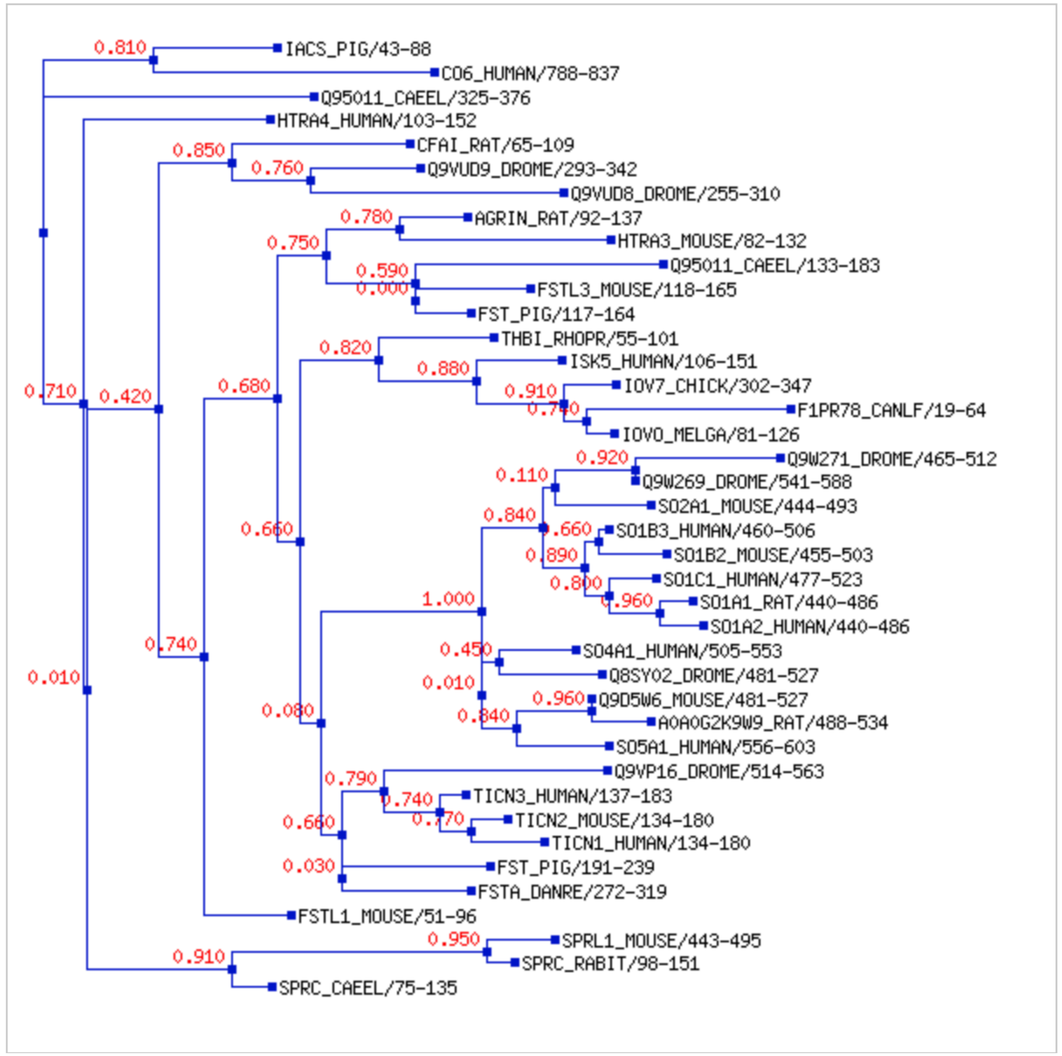


sequence logo (plotting information content per column)

ian.simpson@ed.ac.uk

PFAM Tree

Kazal2 - Serine Protease Inhibitor



ian.simpson@ed.ac.uk

Further Reading (optional)

MSA

- Feng, D. F., and Doolittle, R. F. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J. Mol. Evol. 25, 351–360 (1987).
- Doolittle, R. F. On the trail of protein sequences. Bioinformatics 16, 24–33 (2000).
- Higgins, D. G., Blackshields, G., and Wallace, I. M. Mind the gaps: Progress in progressive alignment. Proc. Natl. Acad. Sci. USA. 102, 10411–10412 (2005).
- Kumar, S., and Filipski, A. Multiple sequence alignment: in pur- suit of homologous DNA positions. Genome Res. 17, 127 – 135 (2007).

CLUSTALW

- Thompson, J. D., Higgins, D. G., and Gibson, T. J. CLUSTALW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680 (1994).

MUSCLE

- Edgar, R. C. MUSCLE: A multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5, 113 (2004a).
- Edgar, R. C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797 (2004b).

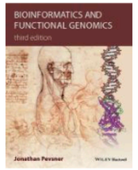
PFAM

- Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., Studholme, D. J., Yeats, C., and Eddy, S. R. The Pfam protein families database. Nucleic Acids Res. 32(Database issue), D138–D141 (2004).

ian.simpson@ed.ac.uk

Next Week

Week 5 - Exploring Biological Databases



This week you should browse BFG for examples of some of these databases in use, especially material about biological databases in Chapter 2, but you don't need to read the whole chapter.

Other useful things you might like to browse are some of the useful guides provided by some of the resources above such as:-

- NCBI Training Tutorials - https://www.ncbi.nlm.nih.gov/guide/training-tutorials/ (very good)
- PubMed User Guide - https://pubmed.ncbi.nlm.nih.gov/help/ (very good)
- Bioportal Help - https://www.bioontology.org/wiki/BioPortal_Help
- Biogrid Help - https://wiki.thebiogrid.org
- Reactome User Guide - https://reactome.org/userguide (very good)

ian.simpson@ed.ac.uk