Multiple Sequence Alignment

Scott Walmsley, PhD

Research Instructor, Department Pharmaceutical Sciences

Skaggs School of Pharmacy

Outline

- What is and why perform Multiple Sequence Alignment (MSA)?
- Pre-requisite knowledge
- History of MSA
- Application – *post hoc* analysis – what can you do with it?
- Available Tools
- Computational Methods

Outline

- What is and why perform Multiple Sequence Alignment (MSA)?
- Pre-requisite knowledge
- History of MSA
- Application – *post hoc* analysis – what can you do with it?
- Available Tools
- Computational Methods

What is Multiple Sequence Alignment?

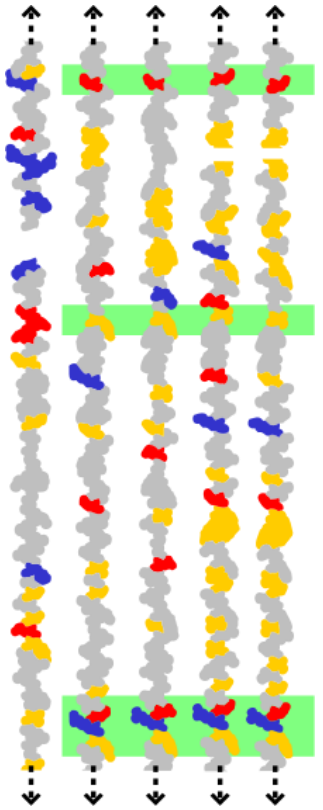
Alignment of 3 or (many) more sequences

- RNA / DNA
- **Protein**
- Structure

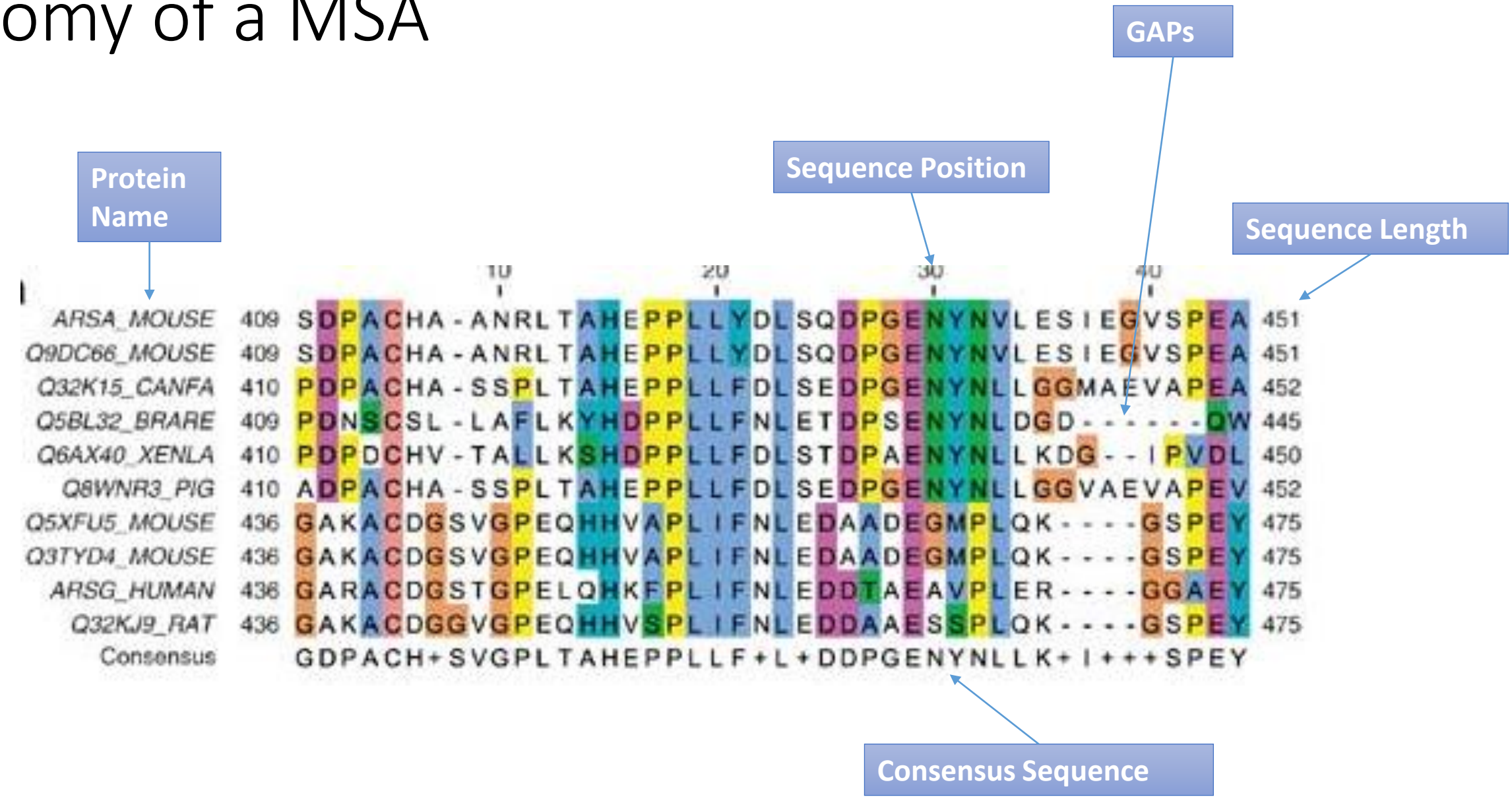
Global versus Local Alignments

- Whole sequence vs Local

Progressive versus Iterative versus others…..



Anatomy of a MSA



Procter JB, Thompson J, Letunic I, Creevey C, Jossinet F, Barton GJ. Visualization of multiple alignments, phylogenies and gene family evolution. Nat Methods. 2010 Mar;7(3 Suppl):S16-25.

Anatomy of an MSA:

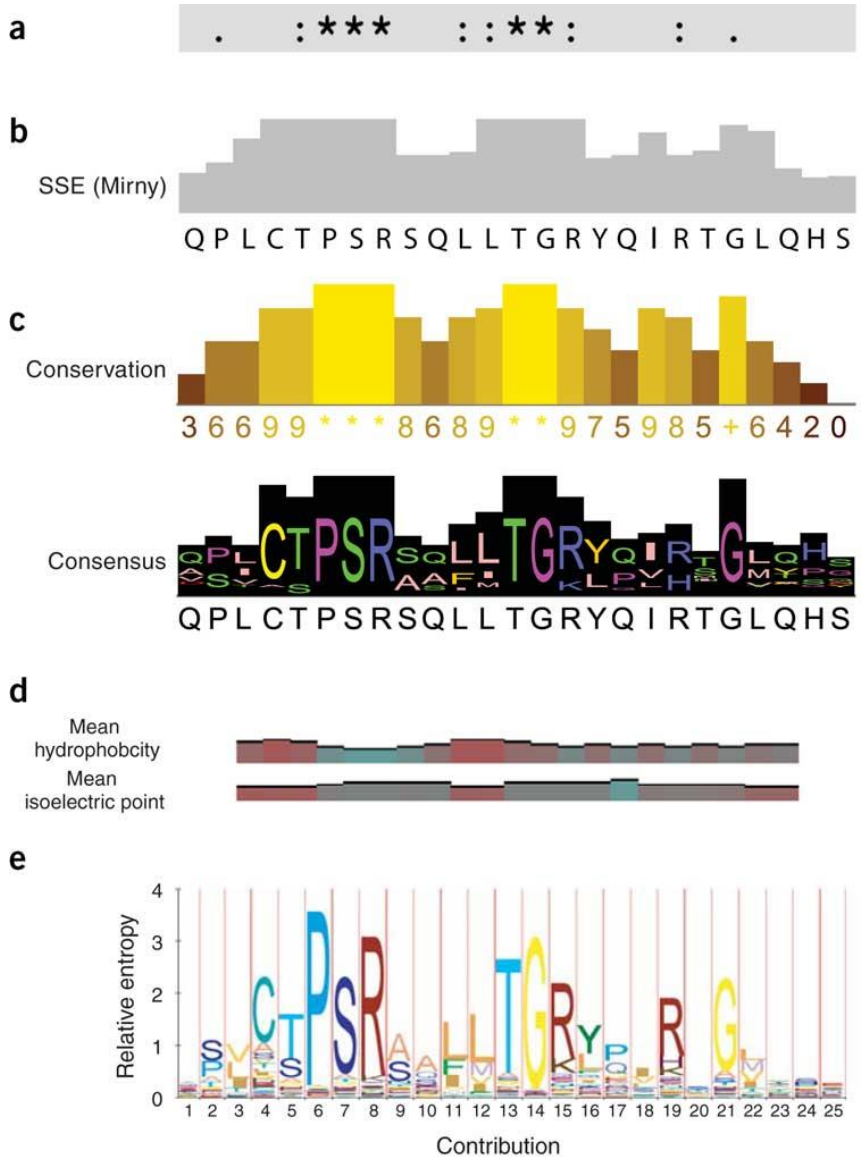
(a) ClustalW quality annotation from ClustalX

(**b**) Mirny conservation measure from PFAAT. Shannon entropy score is calculated for each column based on a reduced amino acid alphabet.

(**c**) Amino acid physicochemical property conservation, consensus and overlaid sequence logo from Jalview.

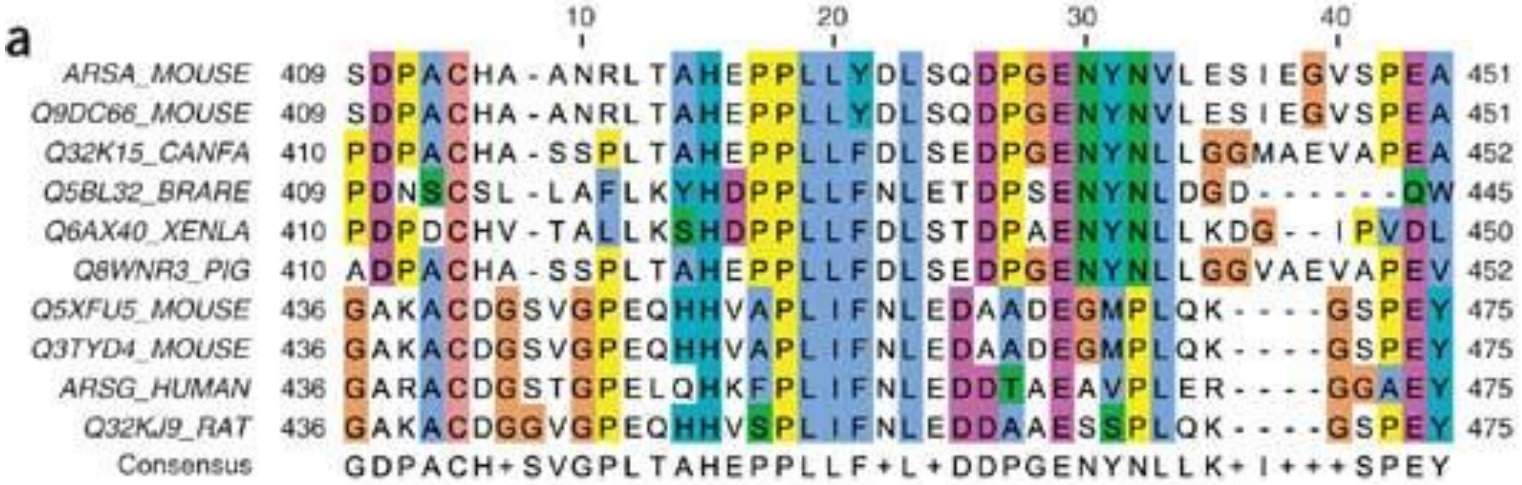
(**d**) Mean hydrophobicity and isoelectric point from Geneious.

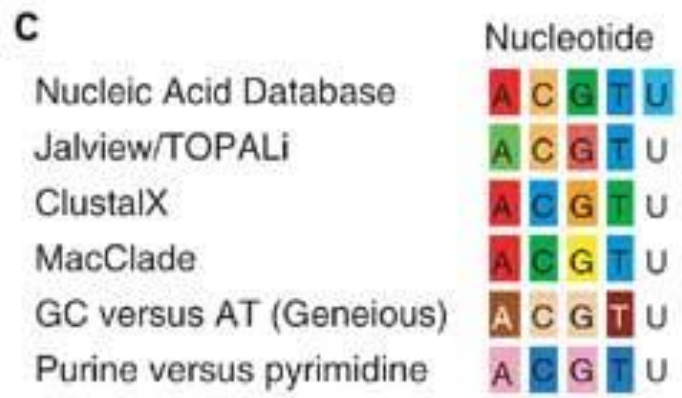
(**e**) HMMlogo visualization from Logomat-P using corresponding HMMER model.

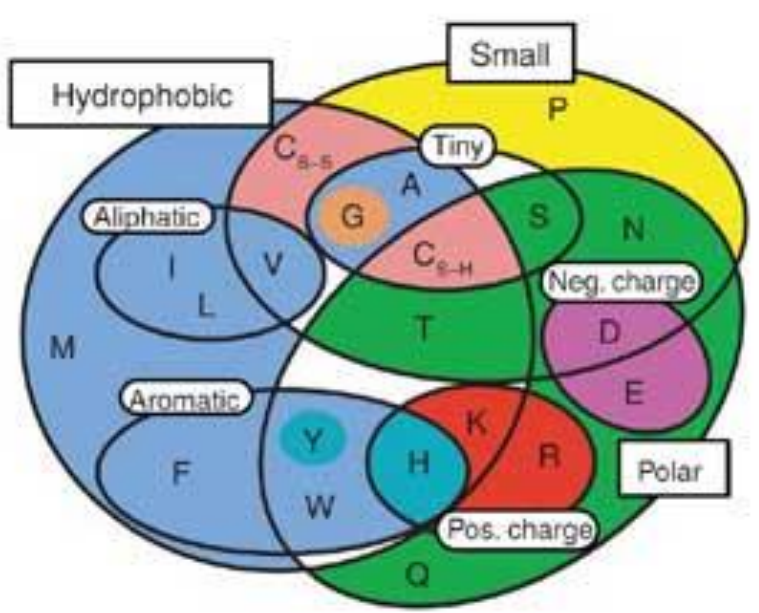


Procter JB, Thompson J, Letunic I, Creevey C, Jossinet F, Barton GJ. Visualization of multiple alignments, phylogenies and gene family evolution. Nat Methods. 2010 Mar;7(3 Suppl):S16-25.

MSA:







Procter JB, Thompson J, Letunic I, Creevey C, Jossinet F, Barton GJ. Visualization of multiple alignments, phylogenies and gene family evolution. Nat Methods. 2010 Mar;7(3 Suppl):S16-25.

.

Why MSA?

"Whether the ultimate aim is a ***phylogenetic*** analysis of several orthologues, the identification of a ***pattern*** for particular feature or motif, or the basis for ***structural modelling***, multiple sequence alignments allow the researcher to gather more biological information than a single sequence can offer"

"The importance of a residue for maintaining the structure and function of a protein can usually be inferred from how conserved it appears in a multiple sequence alignment of that protein and its homologues"

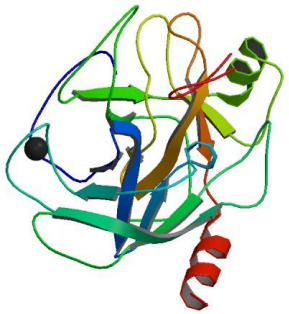
Valdar WS. Scoring residue conservation. Proteins. 2002 Aug 1;48(2):227-41. Review

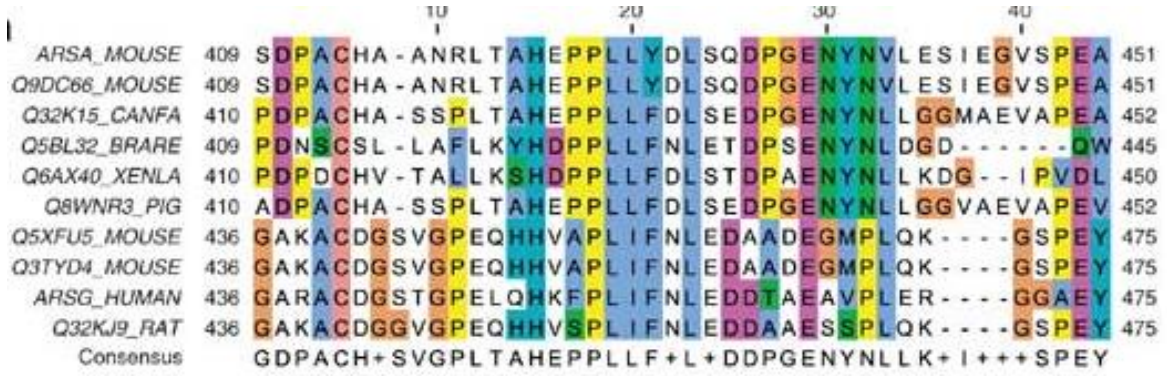
But by using MSA we proceed with caution:

"There is no rigorous mathematical test for judging a conservation measure, if there were one would use the test and not bother with an additional score"

Valdar WS. Scoring residue conservation. Proteins. 2002 Aug 1;48(2):227-41. Review

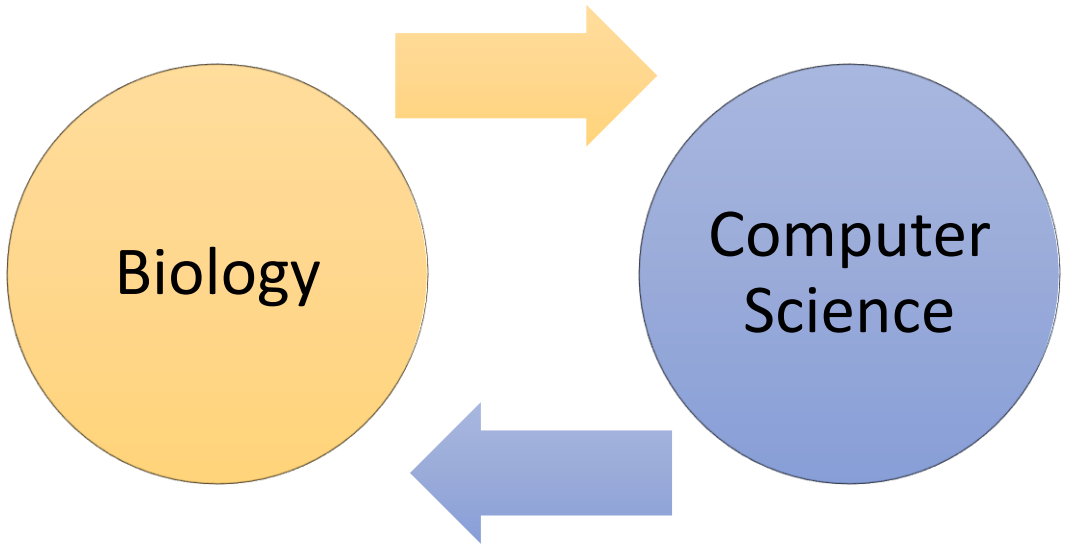
Q: What makes a good Multiple Sequence Alignment?





$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} nk} \qquad k = 0, \ldots, N-1.$$

Different perspectives on a good alignment?



We have the same goal in mind: the optimum solution that makes sense…

Different perspectives on a good product:

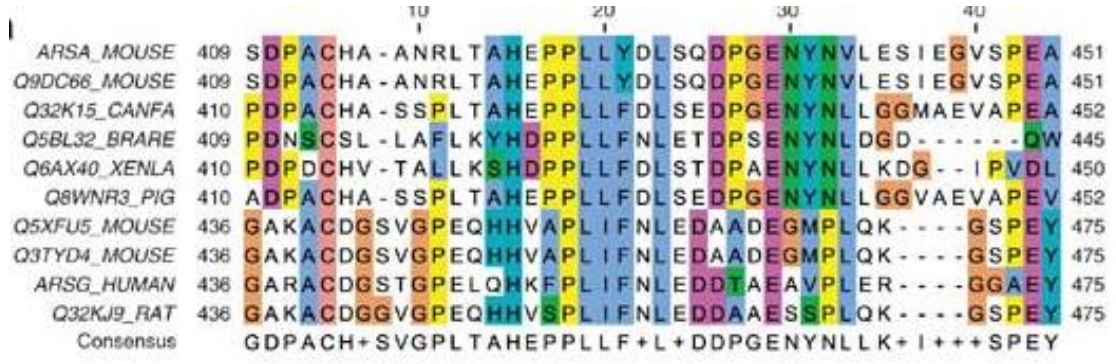
Designer

Engineer

The interpretation of what makes it good is different…..

Different perspectives on a good alignment:

Biologist



Computer Scientist

Structure / Function

Efficiency / Optimum Solution

Outline

- What is and why perform Multiple Sequence Alignment (MSA)?
- **Pre-requisite knowledge**
- History of MSA
- Application – *post hoc* analysis – what can you do with it?
- Available Tools
- Computational Methods

Pre-requisite knowledge

Knowledge of the following can help in your use of MSA:

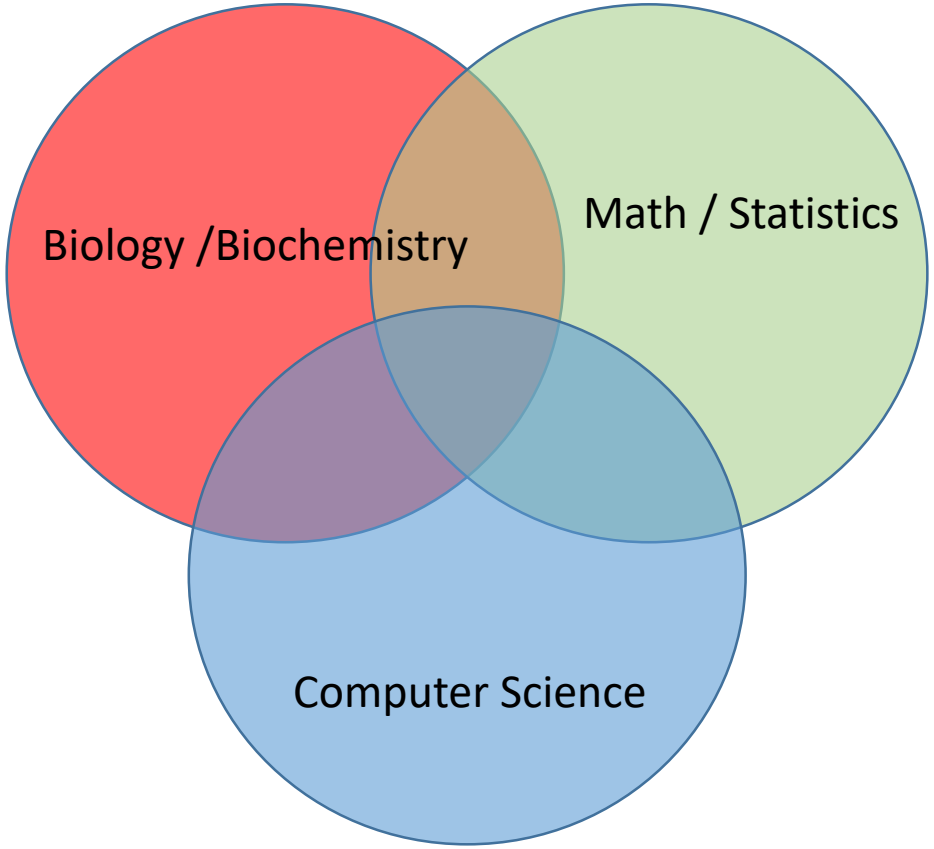
Computational / Math / Statistics

- Pairwise sequence alignment methods
- Substitution matrices
- Phylogenetic trees

Molecular Biology /Biochemistry

- Genetics / sequencing /evolution
- Structure – function
- Bio-chemistry

Pre-requisite knowledge

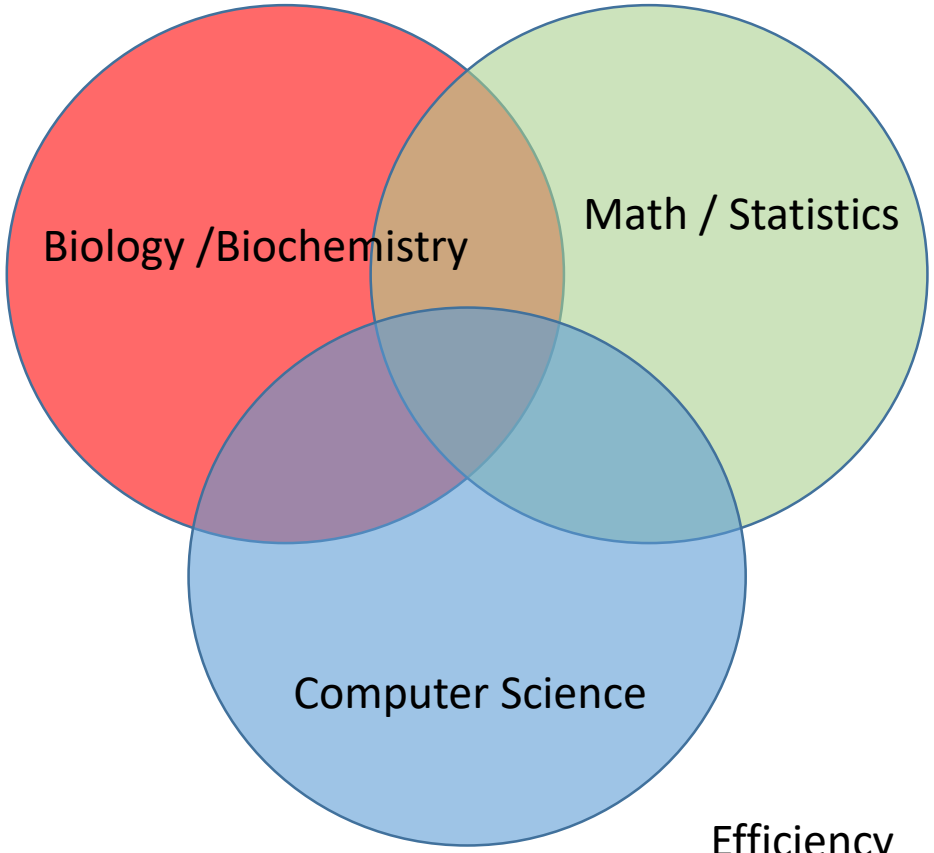


How specific in one field you want to go is up to you, but there are always others to collaborate with to complement your skillset.

Pre-requisite knowledge

Examples

Sequence / Structure /Function



Numerical methods / evaluation

Efficiency

Pre-requisite knowledge

Knowledge of the following can help in your use of MSA:

Computational / Math / Statistics

- Pairwise sequence alignment methods
- Substitution matrices
- Phylogenetic trees

Molecular Biology /Biochemistry

- Genetics / sequencing /evolution
- Structure – function
- Chemistry

Pre-requisite knowledge

Knowledge of the following can help in your use of MSA:

Computational / Math / Statistics

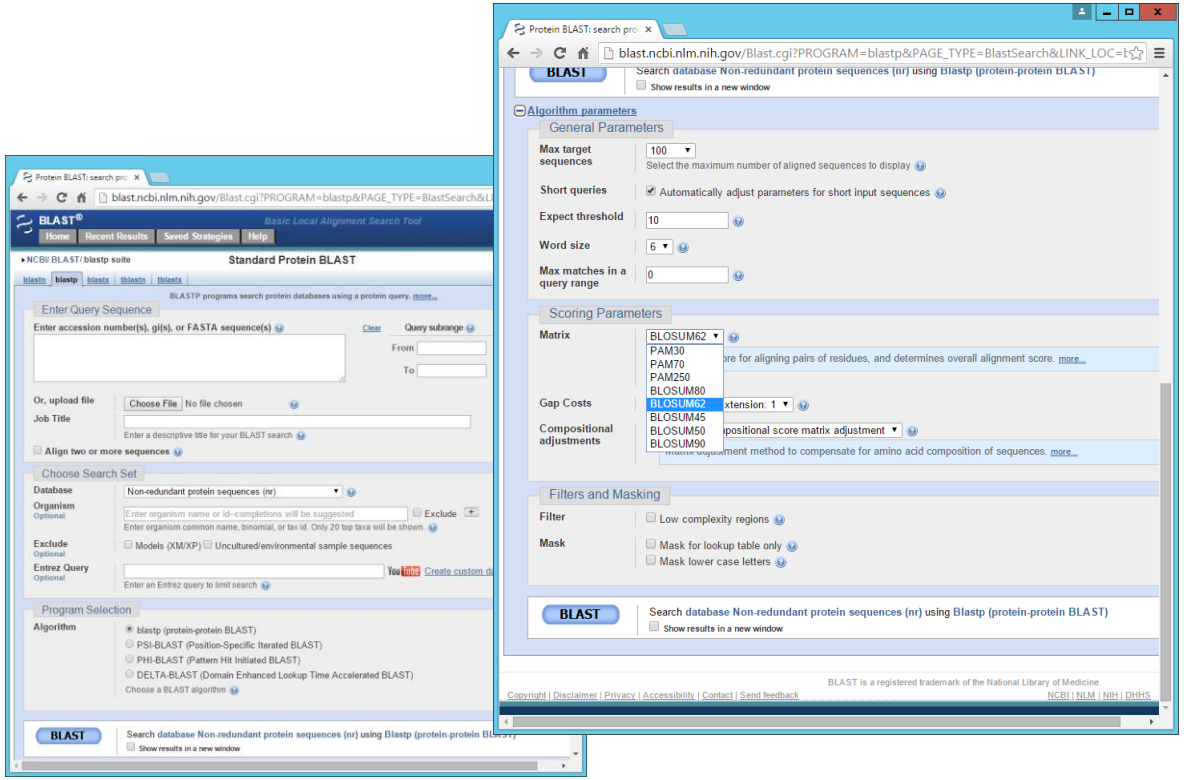
- Pairwise sequence alignment methods

Global (Needleman-Wunsch) vs. Local (Smith - Waterman) vs. Heuristic (BLAST)

Needleman-Wunsch

match = 1 mismatch = -1 gap = -1

		G	C	A	T	G	C	U
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	1	0	-1	-2	-3	-4	-5
A	-2	0	0	1	0	-1	-2	-3
T	-3	-1	-1	0	2	1	0	-1
T	-4	-2	-2	-1	1	1	0	-1
A	-5	-3	-3	-1	0	0	0	-1
C	-6	-4	-2	-2	-1	-1	1	0
A	-7	-5	-3	-1	-2	-2	0	0



Pre-requisite knowledge

Computational / Math / Statistics

- Substitution matrices

PAM

BLOSUM

DYNAMIC

Choice of mutation matrix can effect pairwise and subsequent MSA
A good handle on how the choice effects your MSA might be based on how evolutionarily distant the sequences of interest are.

BLOSUM62

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

PAM: **P**oint **A**ccepted **M**utation

Pre-requisite knowledge

Computational / Math / Statistics

- Phylogenetic trees

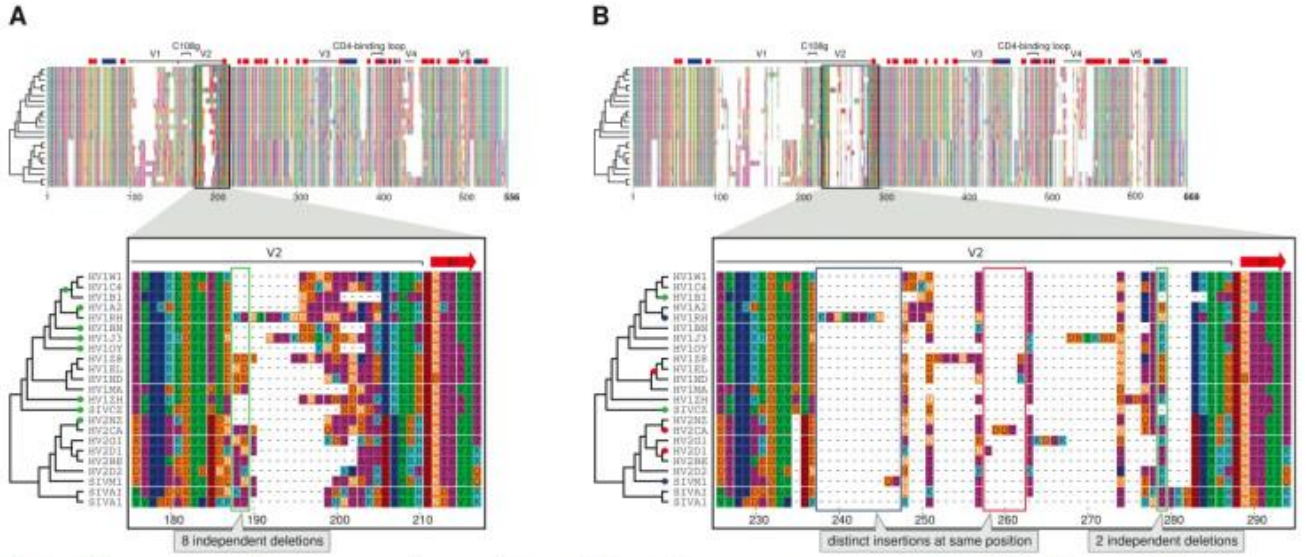


Fig. 1. Different sequence alignment approaches can give contradicting pictures of evolutionary mechanisms behind functional sequence changes. (**A**) (top) The CLUSTAL W (*12*) alignment of gp120 from different strains of human and simian immunodeficiency virus (*5*) represents a typical structural matching of protein sequences and clusters nearby alignment gaps in tight blocks. (bottom) The expanded fragment suggests that part of the V2 region evolves with a high rate of point substitutions and is shortened over evolutionary time by numerous overlapping deletions. For example, the pattern of gaps highlighted with a green box requires eight independent deletions, which have occurred in the lineages marked with green dots in the tree on the left. (**B**) (top) The PRANK$_{+F}$ (*5*) alignment of the same sequences suggests a markedly different evolutionary process dominated by short insertions and deletions. This phylogeny-aware algorithm separates distinct insertions that have taken place at the same positions (bottom: examples highlighted with blue and red boxes and dots) while permitting homologous sites to be deleted multiple times when the data support this (e.g., column highlighted in green). Alignment annotations indicate the V1 to V5 variable regions, C108g epitope, CD4-binding loop, N-linked glycosylation sites (residues N in bold white type), and α helices (blue blocks) and β strands (red blocks) of the known HIV-1 gp120 structure (*5*). Both alignments used the guide phylogeny generated by CLUSTAL W (left).

http://anthropology.net/2008/06/20/improving-multiple-sequence-alignments-with-a-phylogeny-aware-algorithm/

Pre-requisite knowledge

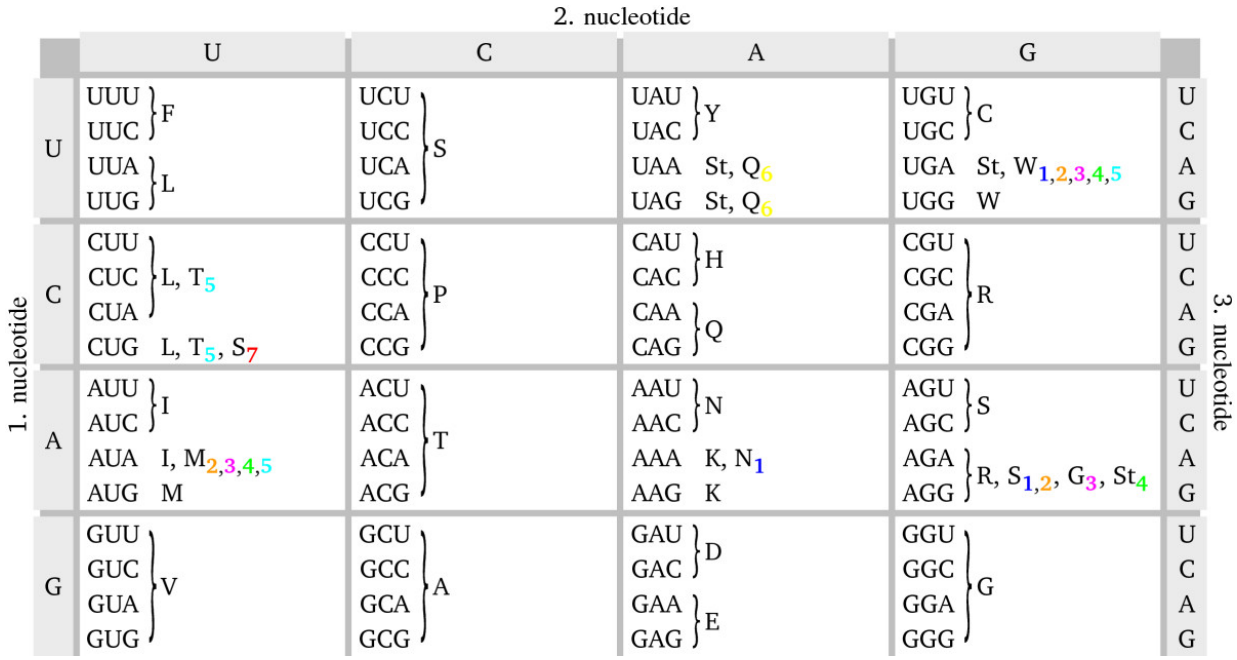
Computational / Math / Statistics

- Alphabets

  DNA (n= 4)

  RNA (n = 4)

  Amino Acids (n = 20)



Pre-requisite knowledge

Computational / Math / Statistics & Biochemistry

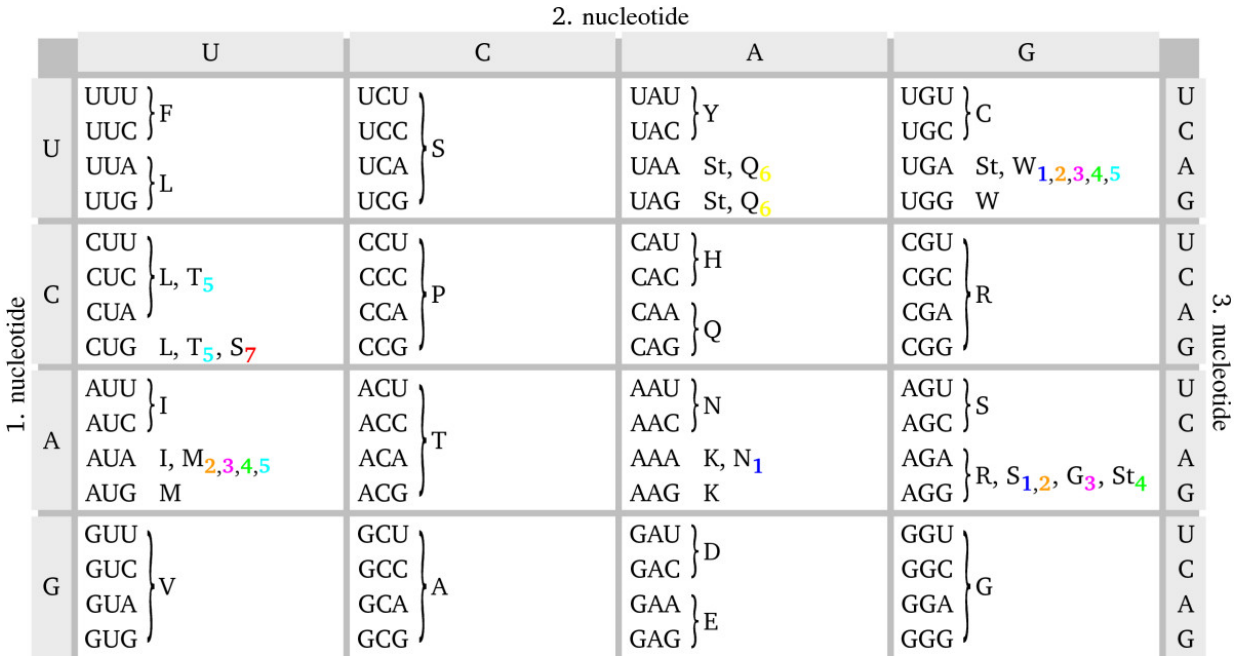
- Alphabets

  DNA (n= 4)

  RNA (n = 4)

  Amino Acids (n = 20)

What other alphabet exists?



1. nucleotide	2. nucleotide: U	C	A	G	3. nucleotide
U	UUU F	UCU S	UAU Y	UGU C	U
	UUC F	UCC S	UAC Y	UGC C	C
	UUA L	UCA S	UAA St, Q_6	UGA St, $W_{1,2,3,4,5}$	A
	UUG L	UCG S	UAG St, Q_6	UGG W	G
C	CUU L, T_5	CCU P	CAU H	CGU R	U
	CUC L, T_5	CCC P	CAC H	CGC R	C
	CUA L, T_5	CCA P	CAA Q	CGA R	A
	CUG L, T_5, S_7	CCG P	CAG Q	CGG R	G
A	AUU I	ACU T	AAU N	AGU S	U
	AUC I	ACC T	AAC N	AGC S	C
	AUA I, $M_{2,3,4,5}$	ACA T	AAA K, N_1	AGA R, $S_{1,2}$, G_3, St_4	A
	AUG M	ACG T	AAG K	AGG R, $S_{1,2}$, G_3, St_4	G
G	GUU V	GCU A	GAU D	GGU G	U
	GUC V	GCC A	GAC D	GGC G	C
	GUA V	GCA A	GAA E	GGA G	A
	GUG V	GCG A	GAG E	GGG G	G

Sammet SG, Bastolla U, Porto M. Comparison of translation loads for standard and alternative genetic codes. BMC Evol Biol. 2010 Jun 14;10:178. doi: 10.1186/1471-2148-10-178. PubMed PMID: 20546599

Pre-requisite knowledge

Computational / Math / Statistics & Biochemistry

- Alphabets

  DNA (n= 4)

  RNA (n = 4)

  Amino Acids (n = 20)

What other alphabet exists?

1. nucleotide	2. nucleotide: U	C	A	G	3. nucleotide
U	UUU, UUC: F	UCU, UCC, UCA, UCG: S	UAU, UAC: Y	UGU, UGC: C	U
	UUA, UUG: L		UAA: St, Q_6	UGA: St, $W_{1,2,3,4,5}$	C
			UAG: St, Q_6	UGG: W	A
					G
C	CUU, CUC, CUA: L, T_5	CCU, CCC, CCA, CCG: P	CAU, CAC: H	CGU, CGC, CGA, CGG: R	U
	CUG: L, T_5, S_7		CAA, CAG: Q		C
					A
					G
A	AUU, AUC: I	ACU, ACC, ACA, ACG: T	AAU, AAC: N	AGU, AGC: S	U
	AUA: I, $M_{2,3,4,5}$		AAA: K, N_1	AGA, AGG: R, $S_{1,2}$, G_3, St_4	C
	AUG: M		AAG: K		A
					G
G	GUU, GUC, GUA, GUG: V	GCU, GCC, GCA, GCG: A	GAU, GAC: D	GGU, GGC, GGA, GGG: G	U
			GAA, GAG: E		C
					A
					G

Sammet SG, Bastolla U, Porto M. Comparison of translation loads for standard and alternative genetic codes. BMC Evol Biol. 2010 Jun 14;10:178. doi: 10.1186/1471-2148-10-178. PubMed PMID: 20546599

Pre-requisite knowledge

Computational / Math / Statistics & Biochemistry

- Alphabets

  DNA (n= 4)

  RNA (n = 4)

  Amino Acids (n = 20)

CODON (n=64)

Table 3 - Codon usage of the *Arapaima gigas* mtDNA.

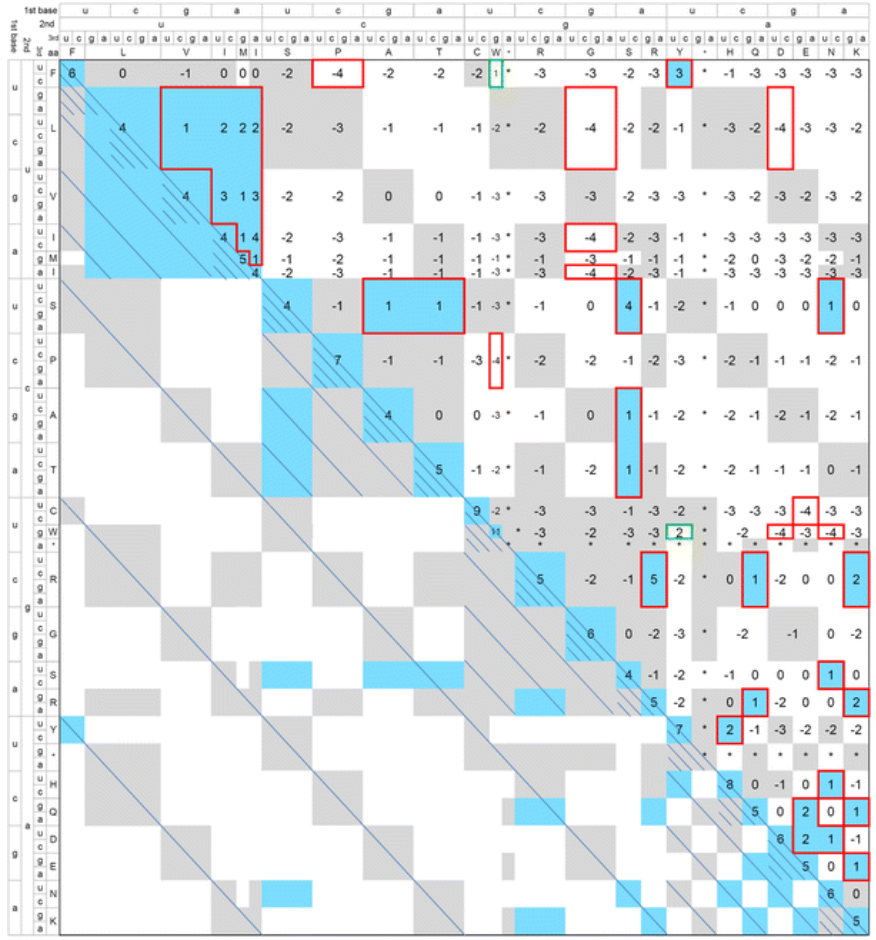
Amino acid (anticodon)	Codon group	Usage of codon ending in				Total	%
		A	C	G	T		
Ala (UGC)	GCN	94	108	4	88	294	7.53
Cys (GCA)	TGY	0	19	0	11	30	0.77
Asp (GUC)	GAY	0	37	0	39	76	1.95
Glu (UUC)	GAR	90	0	6	0	96	2.46
Phe (GAA)	TTY	0	120	0	127	247	6.33
Gly (UCC)	GGN	92	63	35	46	236	6.05
His (GUG)	CAY	0	66	0	45	111	2.84
Ile (GAU)	ATY	0	110	0	201	311	7.97
Lys (UUU)	AAR	82	0	5	0	87	2.23
Leu (UAG)	CTN+TTR	367	116	44	107	634	16.24
Met (CAU)	ATR	146	0	40	0	186	4.77
Asn (GUU)	AAY	0	70	0	64	134	3.43
Pro (UGG)	CCN	122	36	7	43	208	5.33
Gln (UUG)	CAR	94	0	94	0	188	4.82
Arg (UCG)	CGN	44	12	4	12	72	1.84
Ser (UGA)	TCN+AGY	89	99	4	60	252	6.46
Thr (UGU)	ACN	139	86	8	81	314	8.05
Val (UAC)	GTN	86	35	12	54	187	4.79
Trp (UCA)	TGR	111	0	9	0	120	3.07
Tyr (GUA)	TAY	0	49	0	64	113	2.90
Stop (UUA)	TAR	5	0	1	0	6	0.15
Stop (UCA)	TGR	1	0	0	0	1	0.03
Total		1562	1026	273	1042	3903	100.00

Pre-requisite knowledge

Computational / Math / Statistics & Biochemistry

CODON USAGE

“We suggest the codon table be brought up to date and, as a step, we present a novel superposition of the BLOSUM62 matrix and an allowed point mutation matrix. This superposition depicts an important aspect of the true genetic code—its ability to tolerate mutations and mistranslations.”



Cell Biochem Biophys. 2009;55(2):107-16. doi: 10.1007/s12013-009-9060-9. Epub 2009 Jul 29

Pre-requisite knowledge

Computational / Math / Statistics & Biochemistry

- Alphabets

  DNA (n= 4)

  RNA (n = 4)

  Amino Acids (n = 20)

Considerations for MSA performance:

n = Number of sequences

L = Length of sequences

Eg……..F(x) = O(L^n)

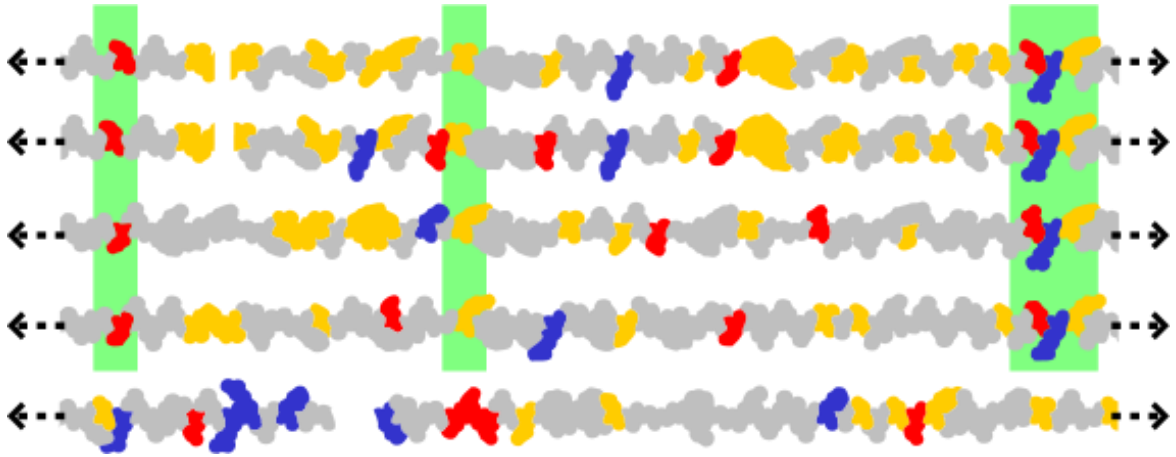
Pre-requisite knowledge

Biochemistry / Molecular Biology

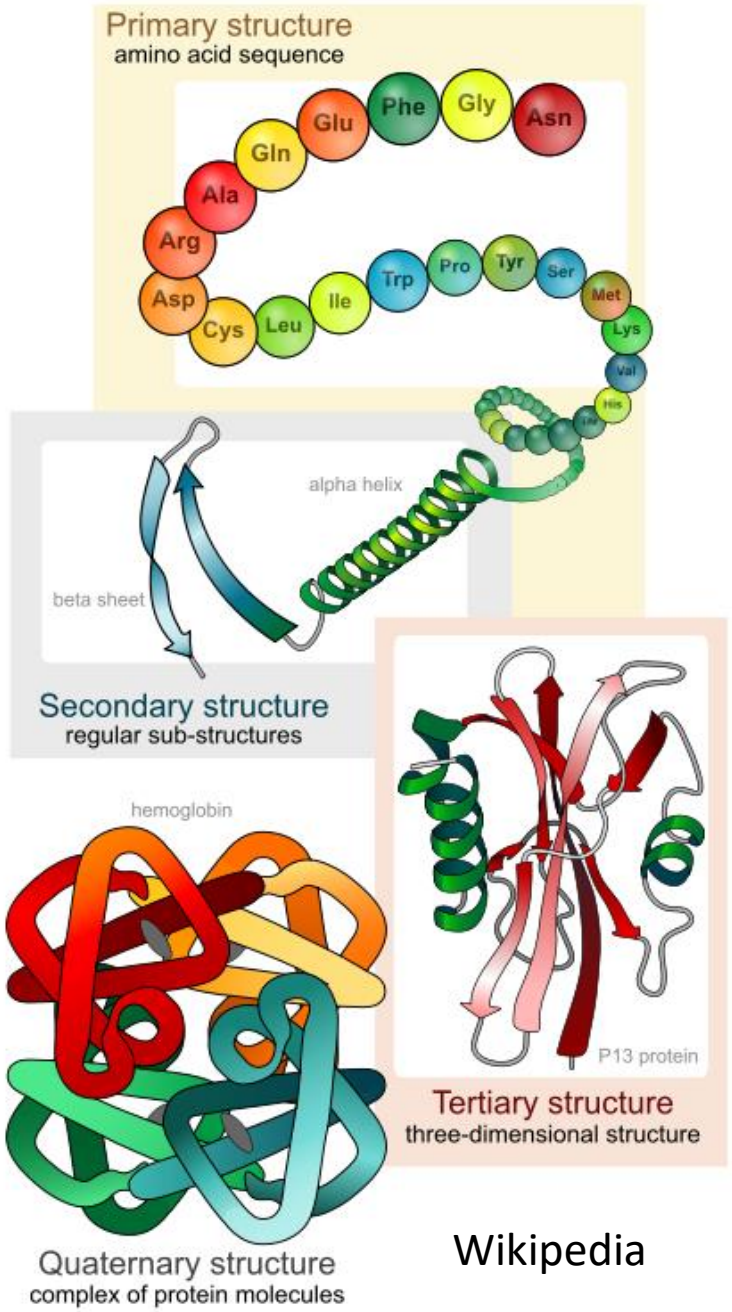
- Mutation rates drive evolution
- Biophysical mechanisms produce mutation rates:

  DNA / RNA Polymerase

- Insertion /Deletion : frameshift → altered CODON



Wikipedia



Wikipedia

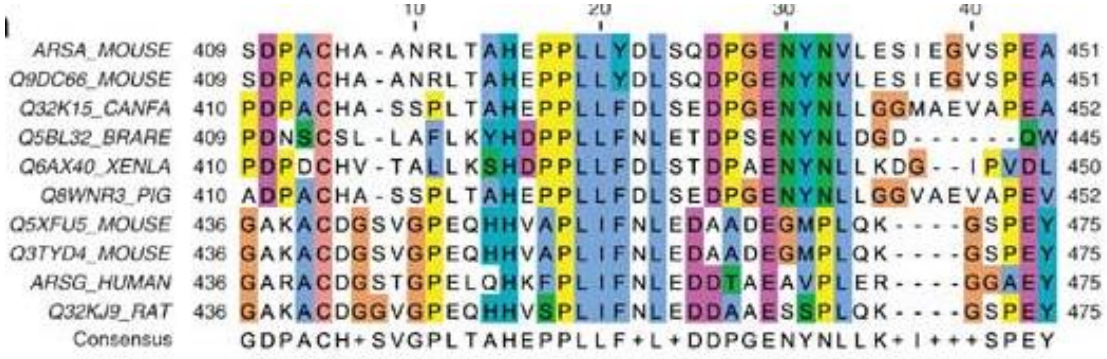
Pre-requisite knowledge

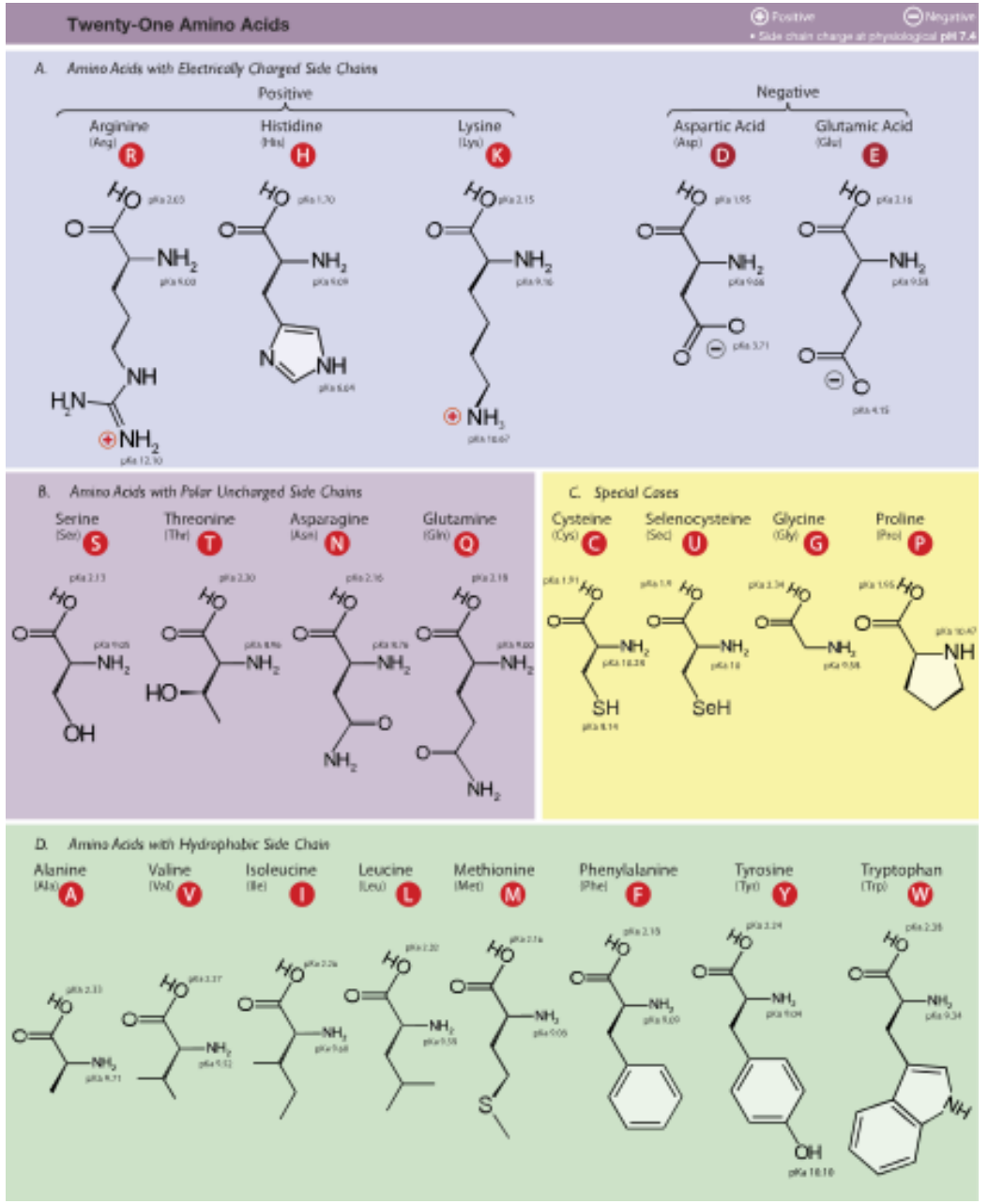
Biochemistry / Molecular Biology

Amino acids confer:

- Structure
- Function / catalysis
- Interaction

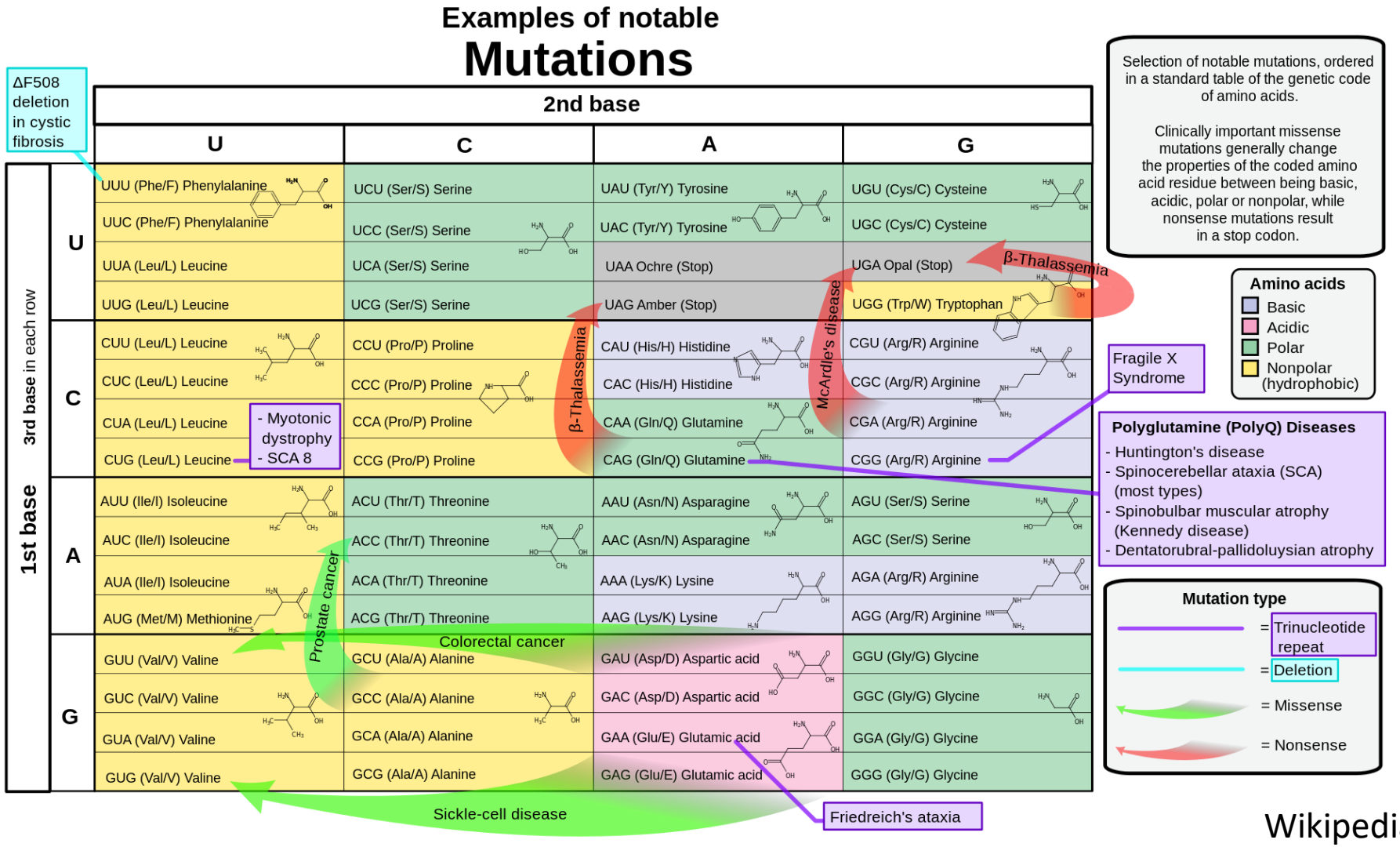
Conservation of sequence is related to maintenance of protein structure / function





Pre-requisite knowledge

Biochemistry / Molecular Biology



Wikipedia

Pre-requisite knowledge is:

- Required to make informed choice of MSA algorithms and the parameters.
- Allows you to make manual adjustments to alignments that make sense.
- Increases your cross cutting / collaborative capabilities
- All concepts support MSA which is central to many (most?) bioinformatics techniques

Outline

- What is and why perform Multiple Sequence Alignment (MSA)?
- Pre-requisite knowledge
- **History of MSA**
- Application – *post hoc* analysis – what can you do with it?
- Available Tools
- Computational Methods

History

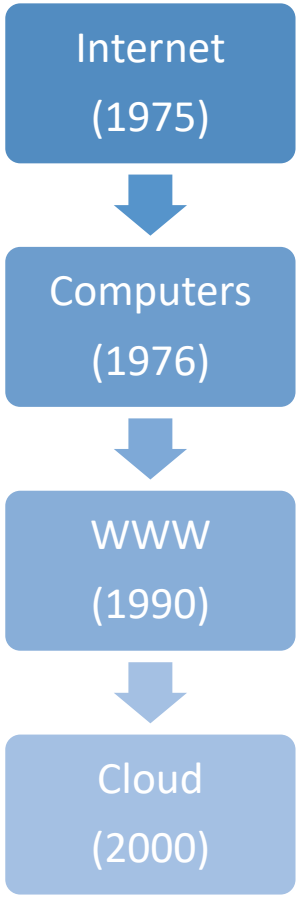
There are too many to discuss in one day……

- Hogeweg and Hesper (1983) -- Iterative
- Clustal (1988) -- Progressive alignment
- SAM (1994) -- Hidden Markov Model
- SAGA (1996) -- Genetic Algorithm
- T-Coffee (2000) -- Progressive
- MUSCLE (2004) -- Progressive / Iterative
- DECIPHER (2014) -- Progressive / Iterative

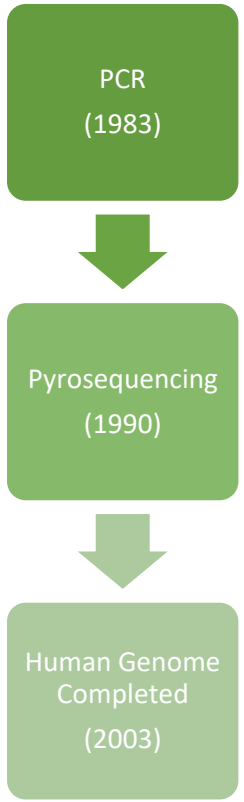
History

Co- evolution of technology

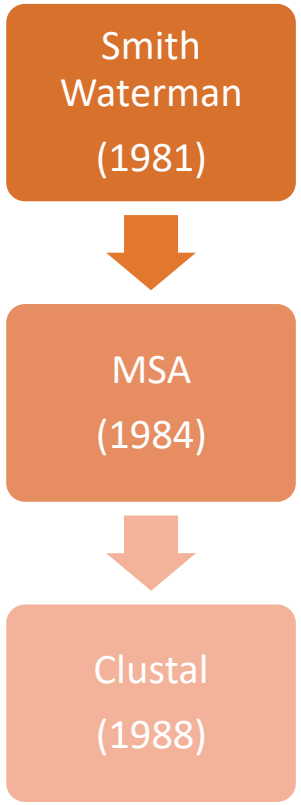
Computers, Information Exchange



Physical Access to Genomic Information

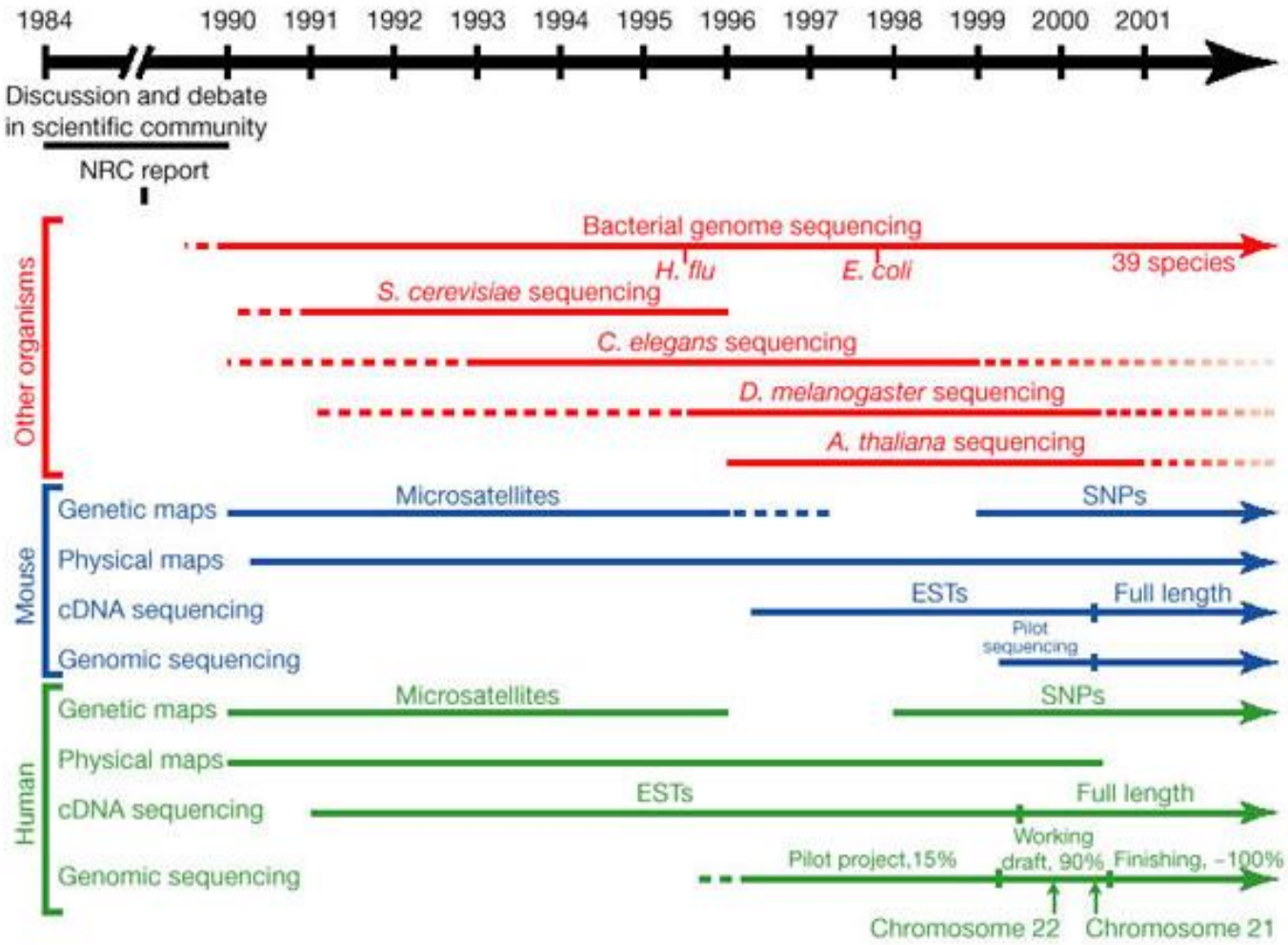


Algorithmic Development



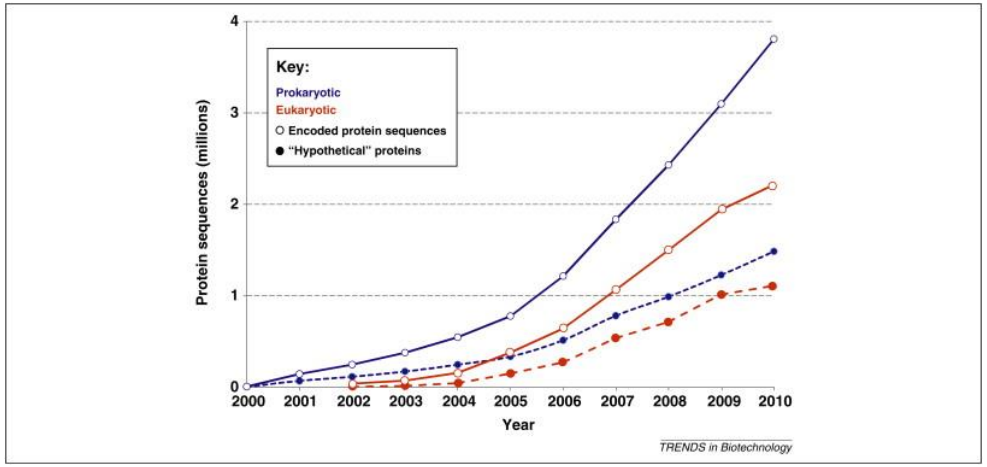
Nature **409**, 860-921 (15 February 2001) | doi:10.1038/35057062; Received 7 December 2000; Accepted 9 January 2001
Initial sequencing and analysis of the human genome

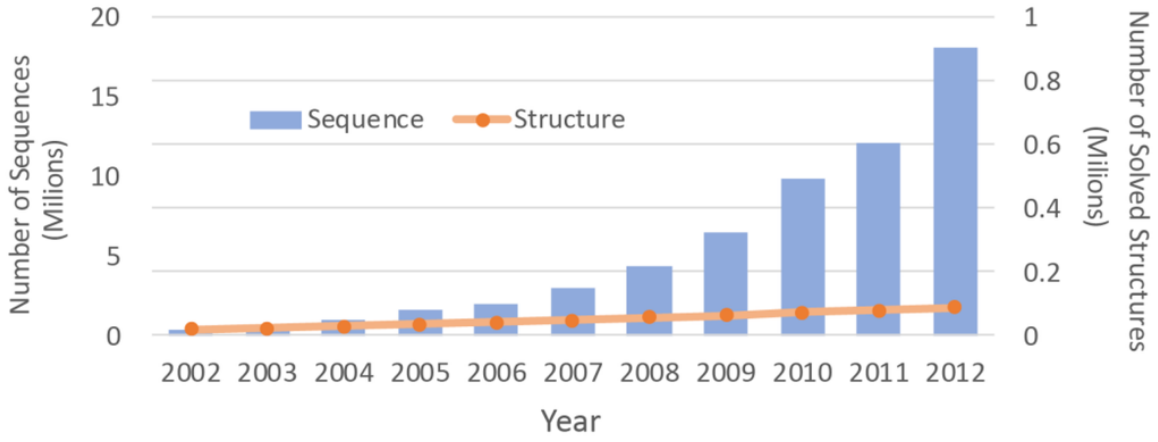
History: genomic sequencing

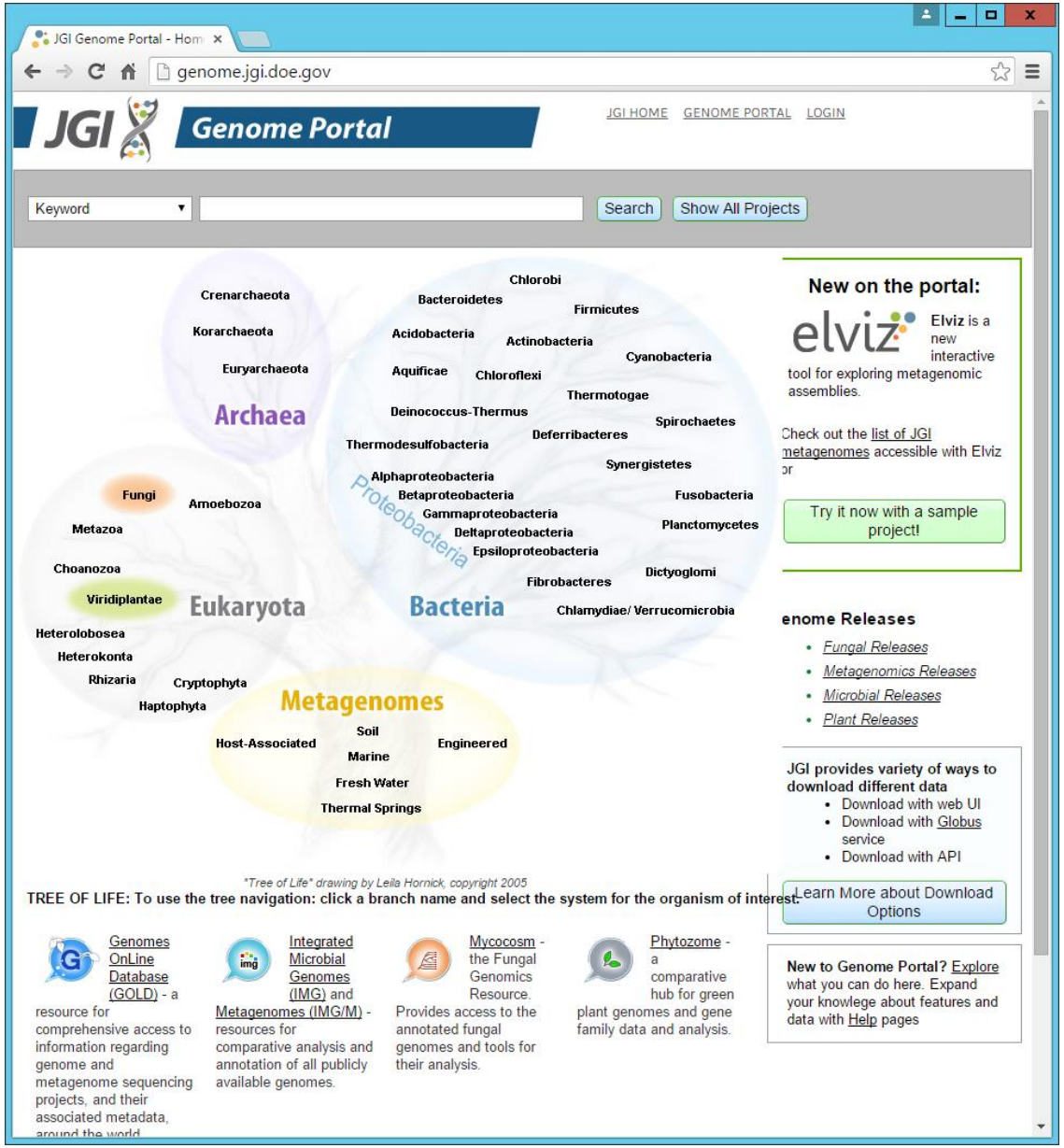


Technology has increased the rate in which data is acquired leading to more information to potentially align against.

And Sequence Information is growing rapidly





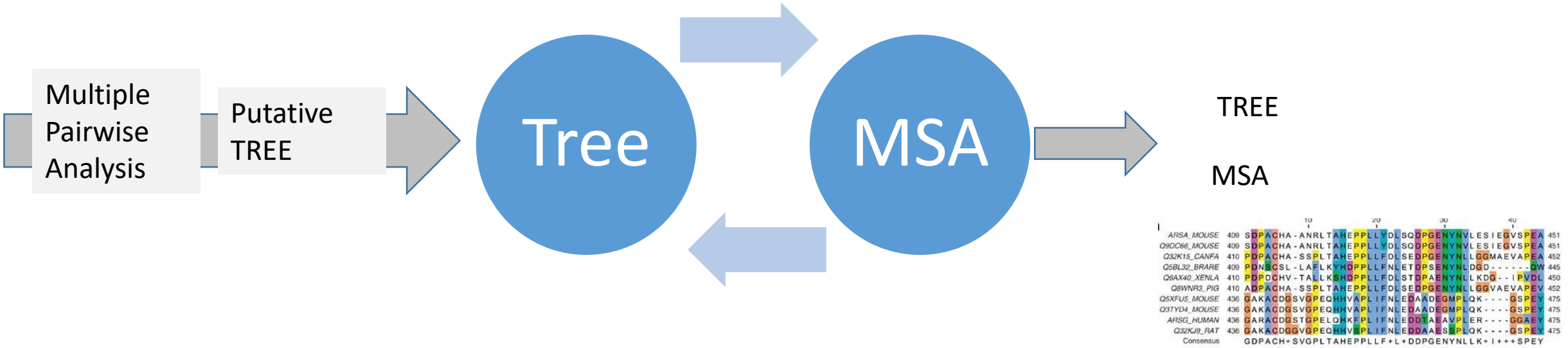


Pi J, Sael L. Mass Spectrometry Coupled Experiments and Protein Structure Modeling Methods. *International Journal of Molecular Sciences*. 2013;14(10):20635-20657. doi:10.3390/ijms141020635.

History

J Mol Evol. 1984;20(2):175-86. ***The alignment of sets of sequences and the construction of phyletic trees: an integrated method.*** Hogeweg P, Hesper B.

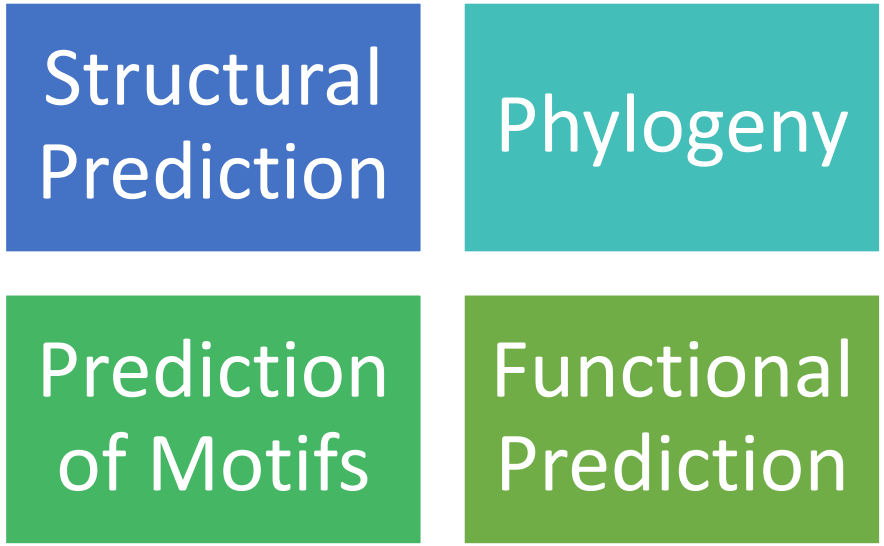
In this paper we argue that the alignment of sets of sequences and the construction of phyletic trees cannot be treated separately. The concept of 'good alignment' is meaningless without reference to a phyletic tree, and the construction of phyletic trees presupposes alignment of the sequences. We propose an integrated method that generates both an alignment of a set of sequences and a phyletic tree. In this method a putative tree is used to align the sequences and the alignment obtained is used to adjust the tree***; this process is iterated***. As a demonstration we apply the method to the analysis of the evolution of 5S rRNA sequences in prokaryotes.]



Outline

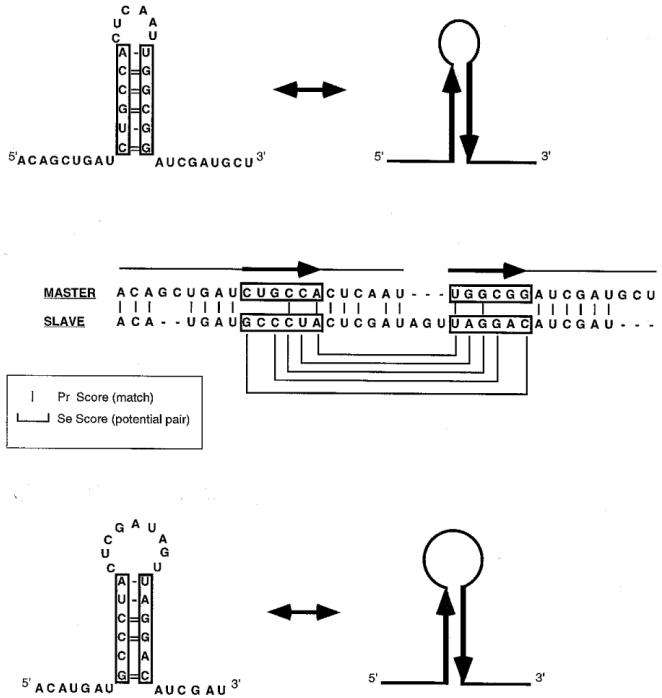
- What is and why perform Multiple Sequence Alignment (MSA)?
- Pre-requisite knowledge
- History of MSA
- **Application – *post hoc* analysis – what can you do with it?**
- Available Tools
- Computational Methods

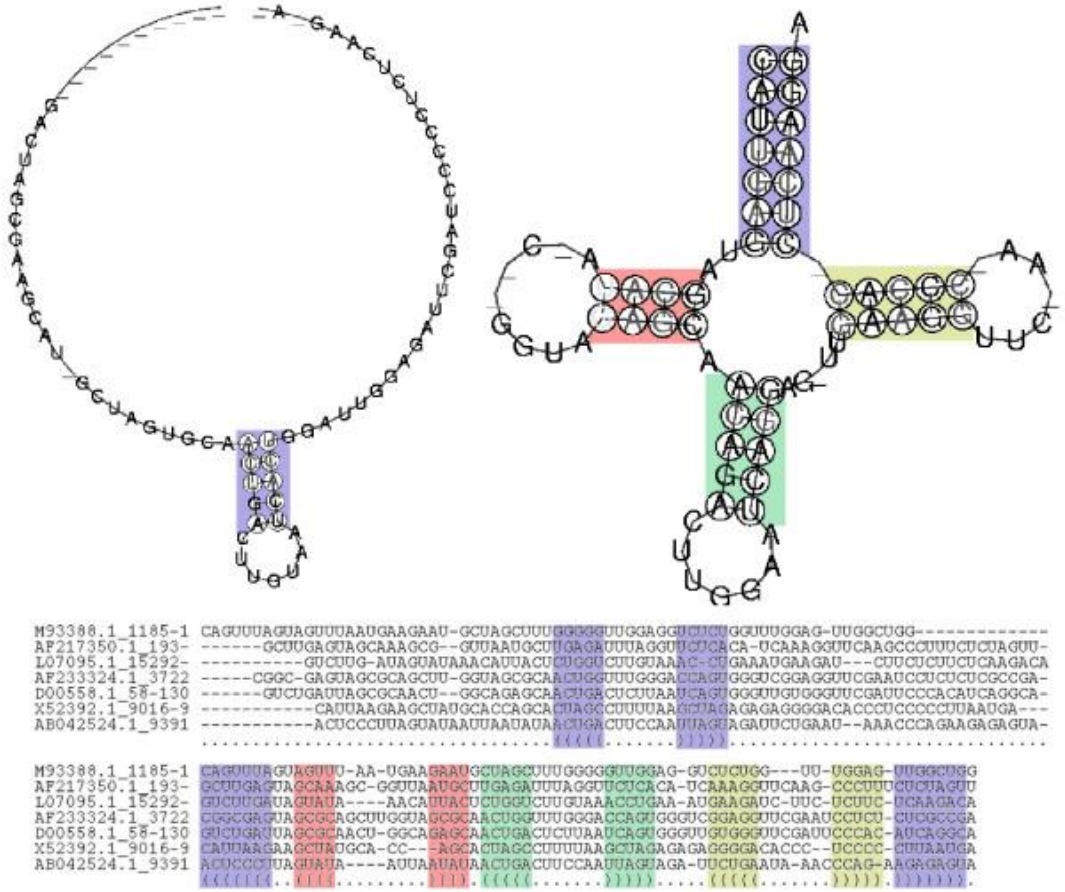
What can you do with MSA?



Application of MSA

RNA Structure Prediction





Bauer M, Klau GW, Reinert K. Accurate multiple sequence-structure alignment of RNA sequences using combinatorial optimization. BMC Bioinformatics. 2007 Jul 27;8:271

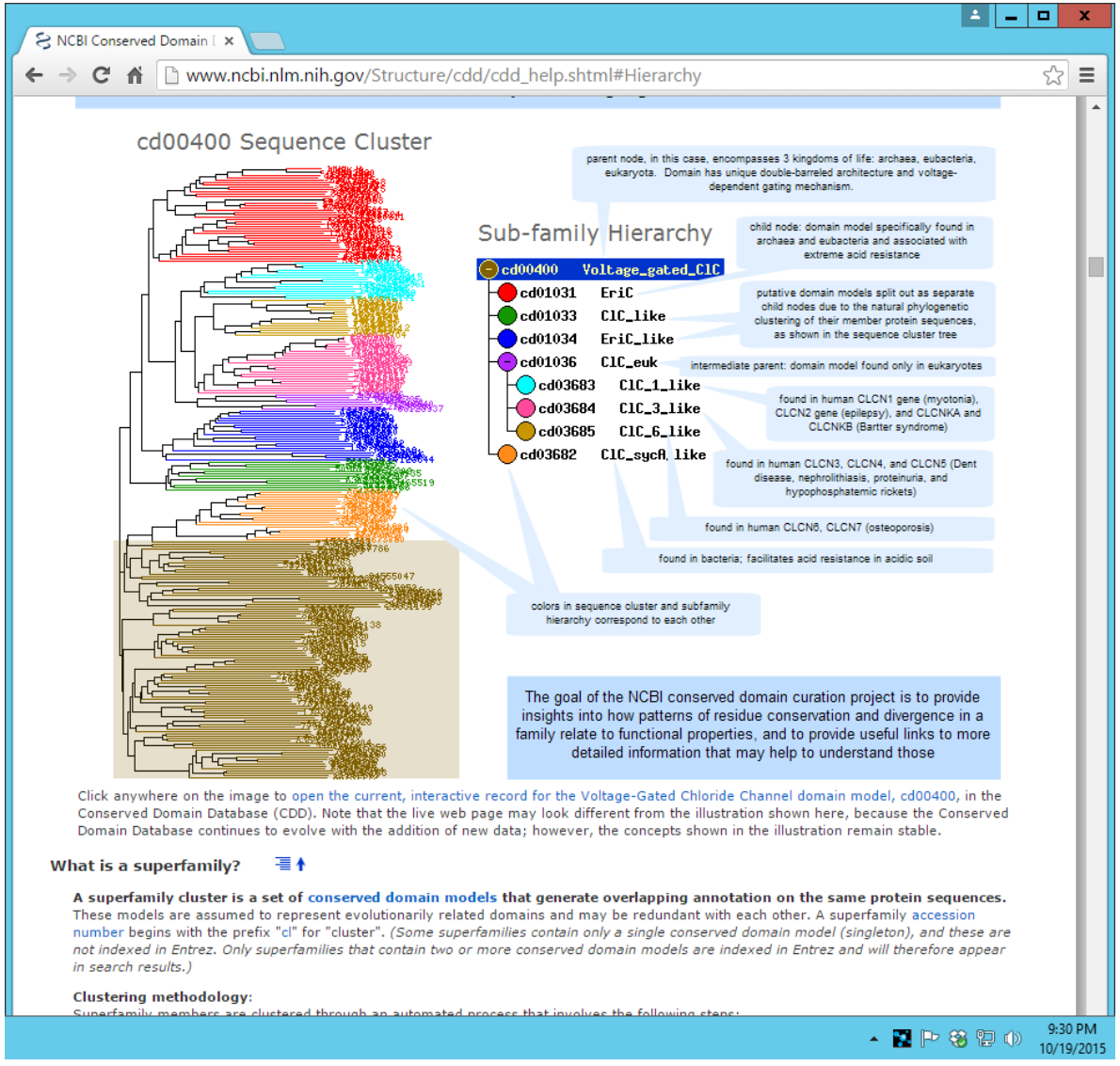
RAGA: RNA sequence alignment by genetic algorithm

Cédric Notredame[1,*], Emmet A. O'Brien[1,2] and Desmond G. Higgins[1,2][1]EMBL Outstation-The European Bioinformatics Institute, Welcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK and[2]Department of Biochemistry, University College, Cork, Ireland

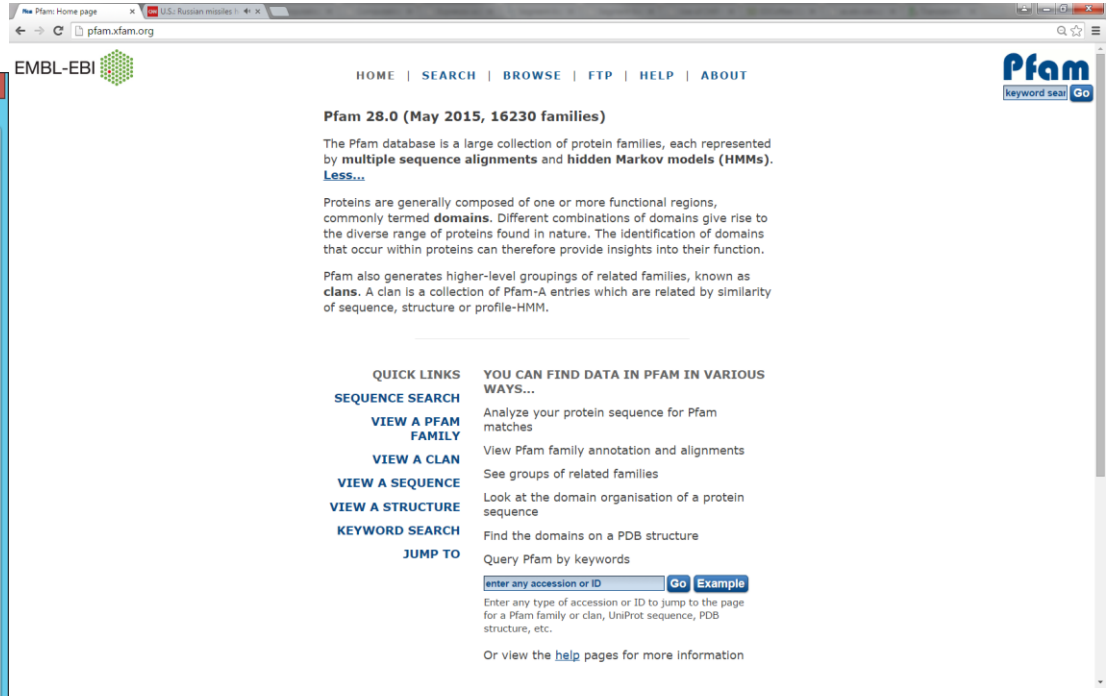
Received July 23, 1997; Revised and Accepted October 1, 1997

Application of MSA

Conserved domains / protein clusters

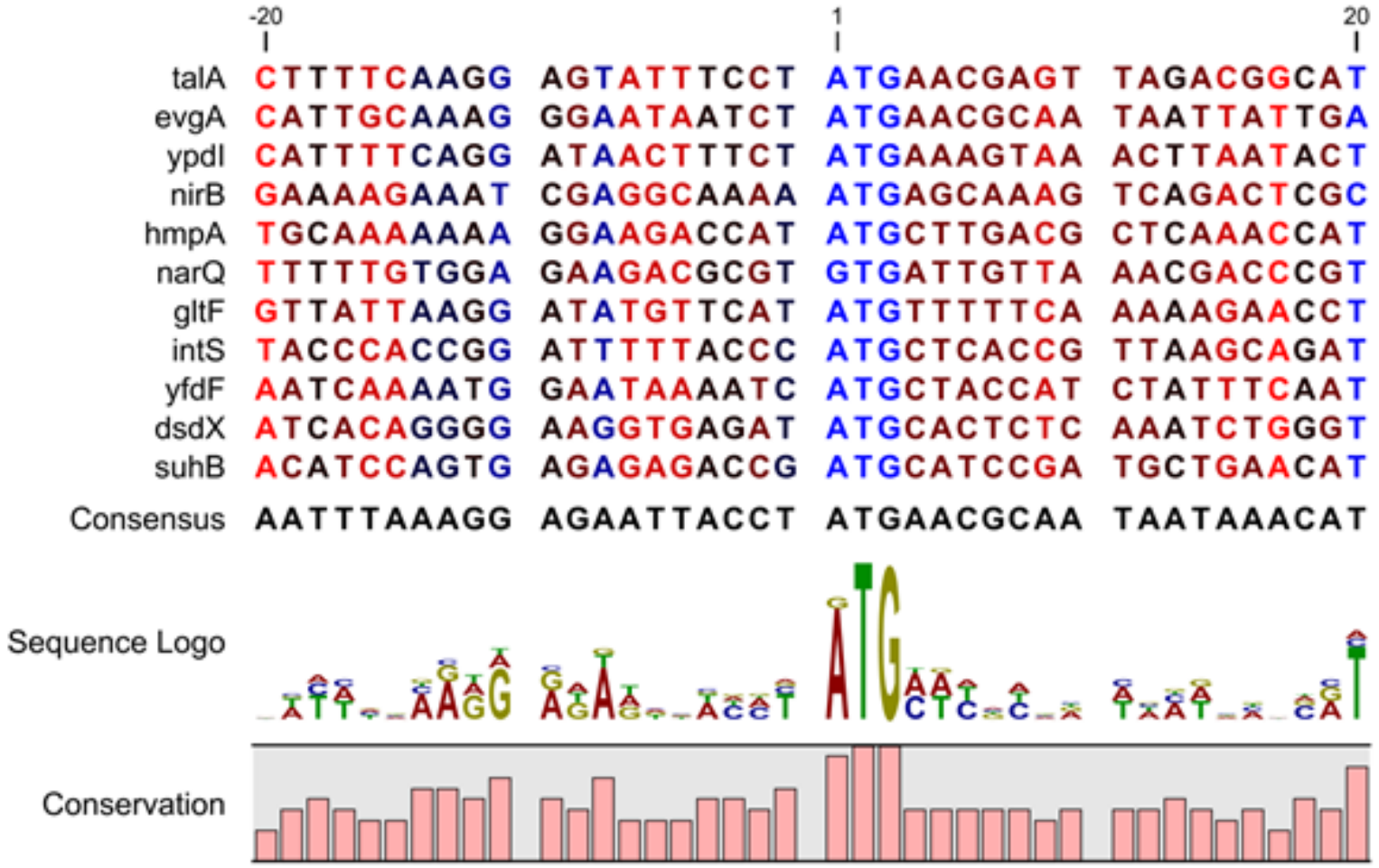


PFAM



Application

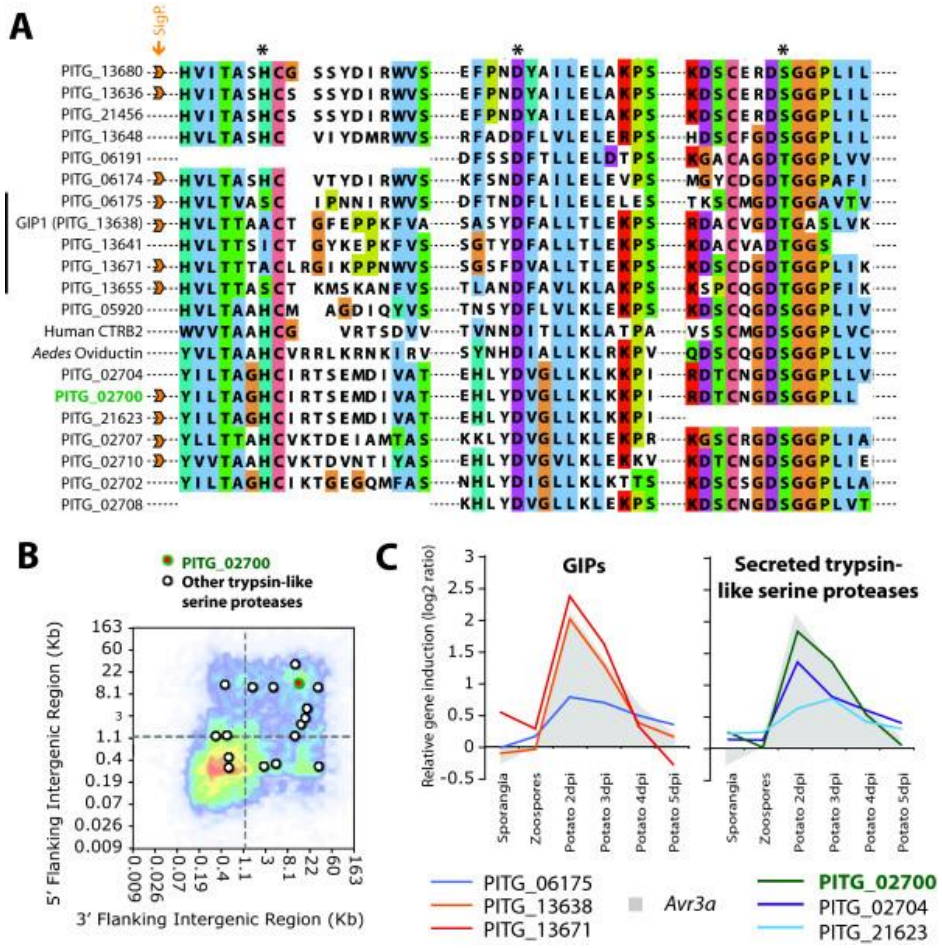
Prediction / conserved motifs

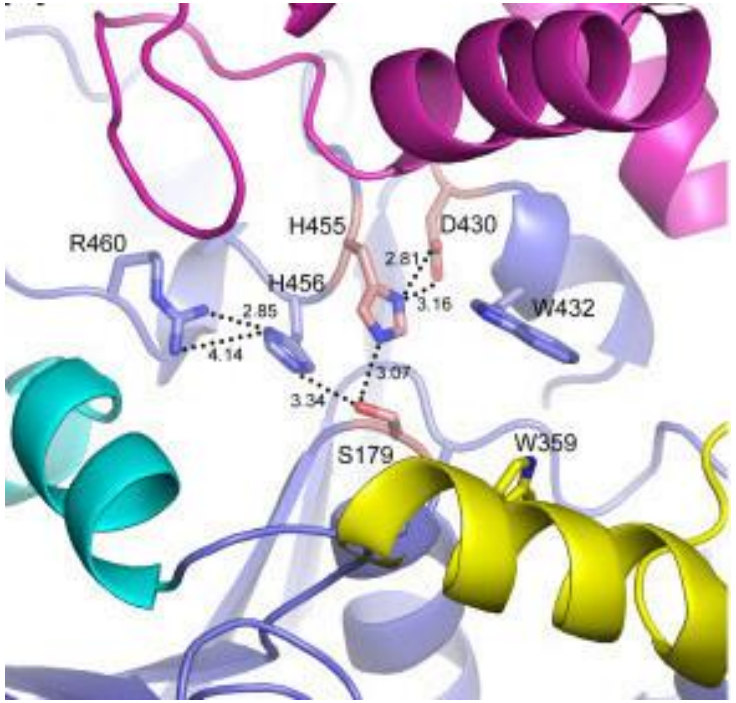


http://www.clcsupport.com/clcgenomicsworkbench/650/BE_Sequence_logo.html

Application:

- Biochemistry: structural / functional



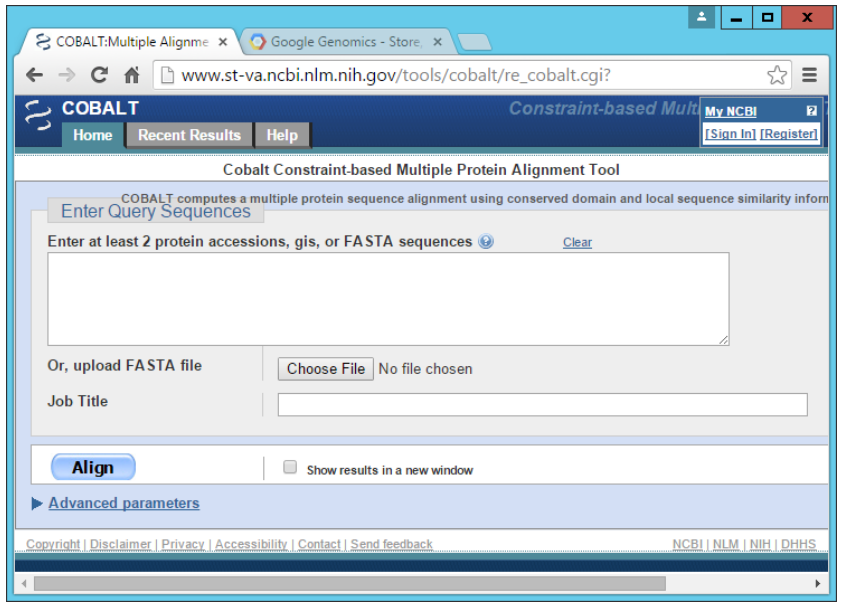


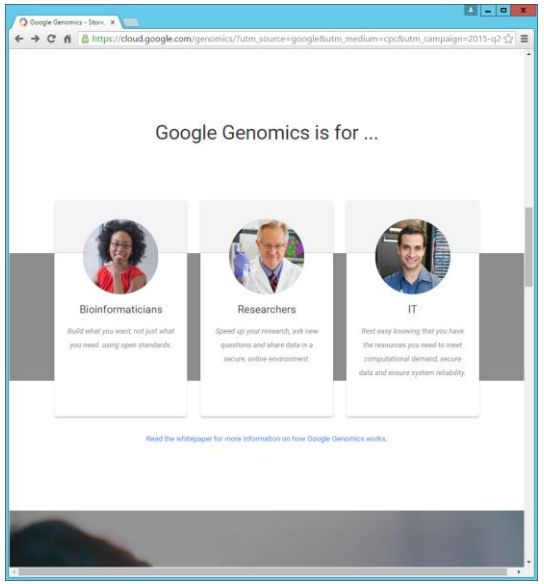
http://openi.nlm.nih.gov/detailedresult.php?img=2893137_1741-7007-8-87-1&req=4

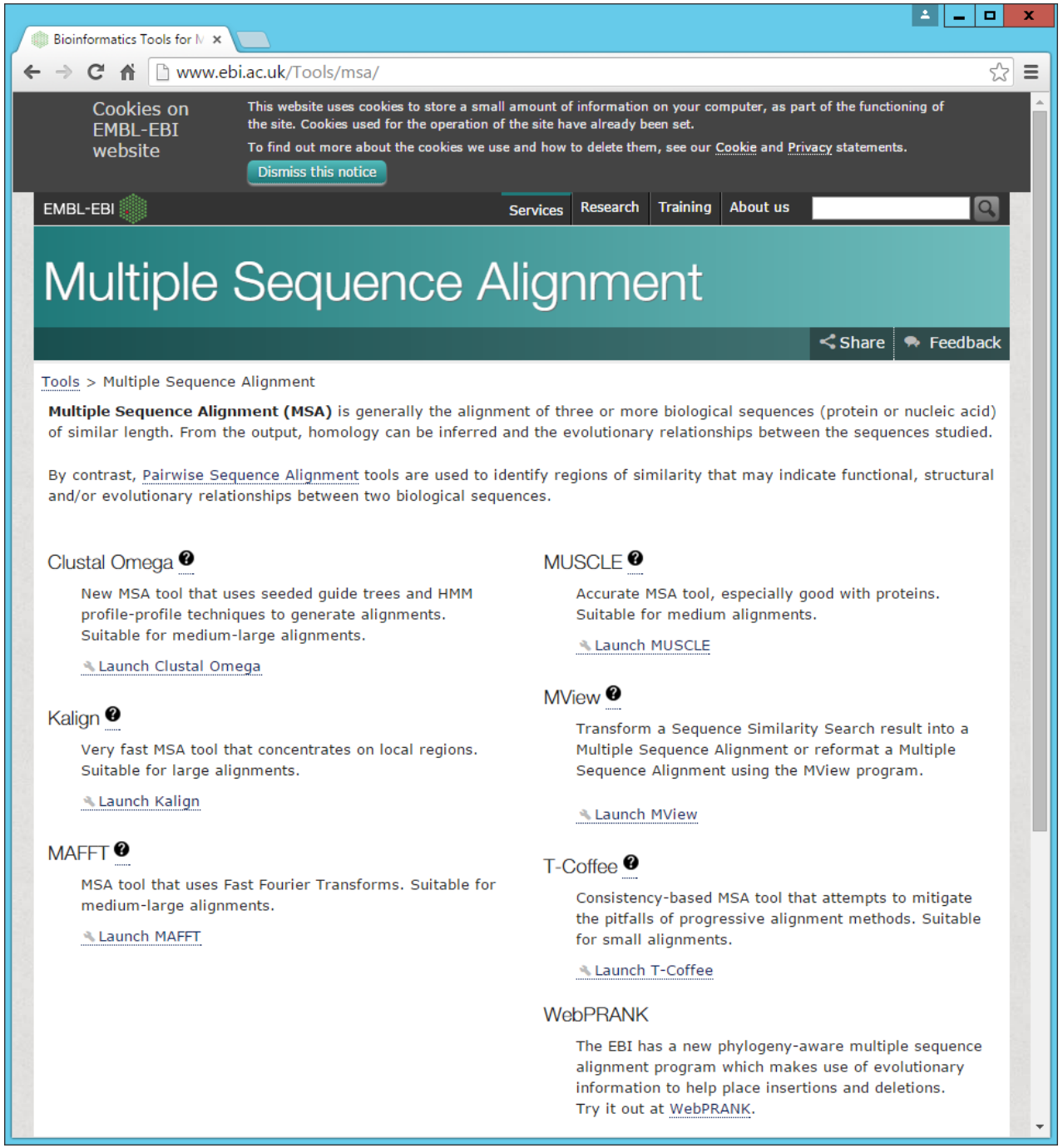
Outline

- What is and why perform Multiple Sequence Alignment (MSA)?
- Pre-requisite knowledge
- History of MSA
- Application – *post hoc* analysis – what can you do with it?
- **Available Tools**
- Computational Methods

Available Tools







Outline

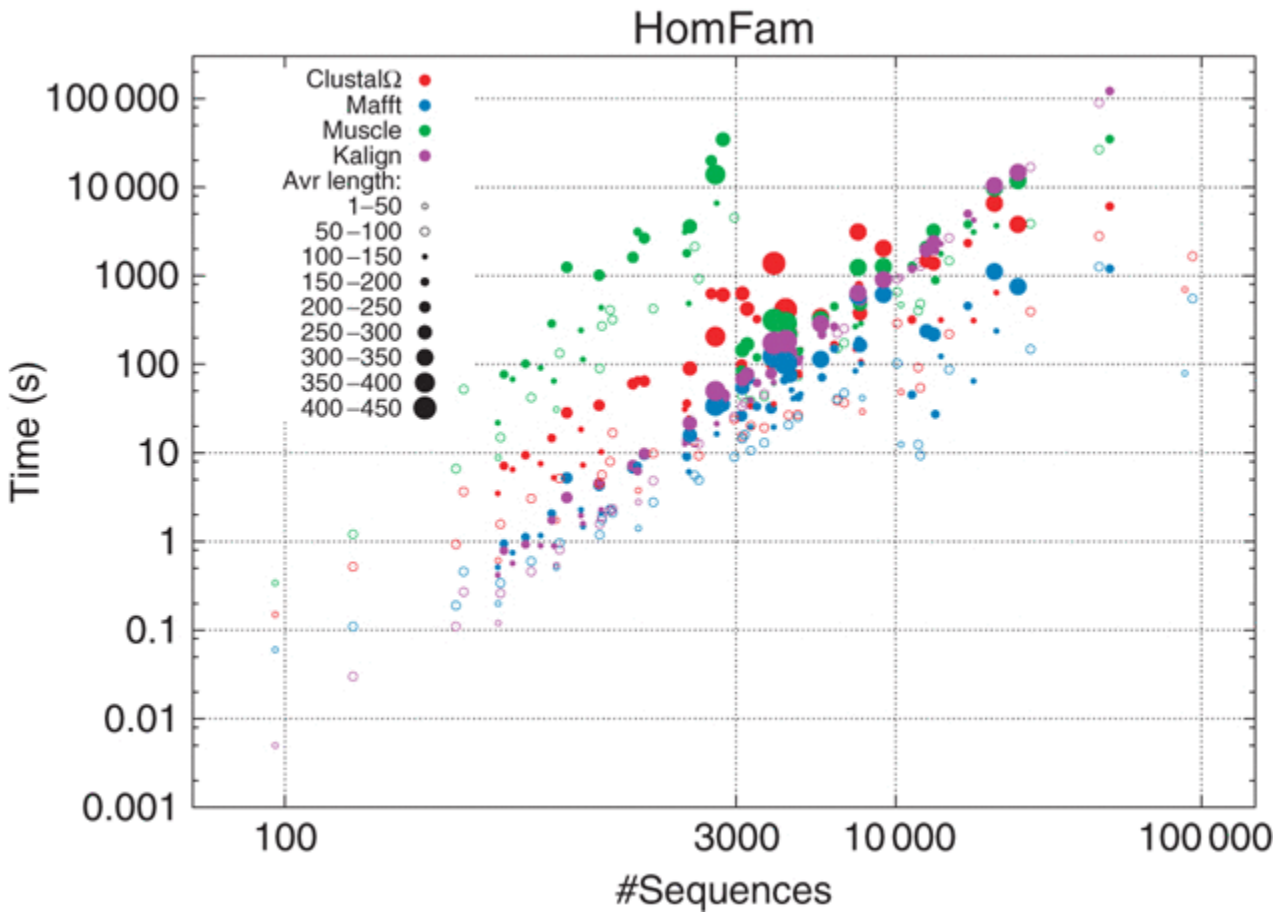
- What is and why perform Multiple Sequence Alignment (MSA)?
- Pre-requisite knowledge
- History of MSA
- Application – *post hoc* analysis – what can you do with it?
- Available Tools
- **Computational Methods**

Computational Methods

Methods

- Global versus Local….from pairwise analysis
- Progressive / Iterative
- Phylogeny Assistance
- Others….

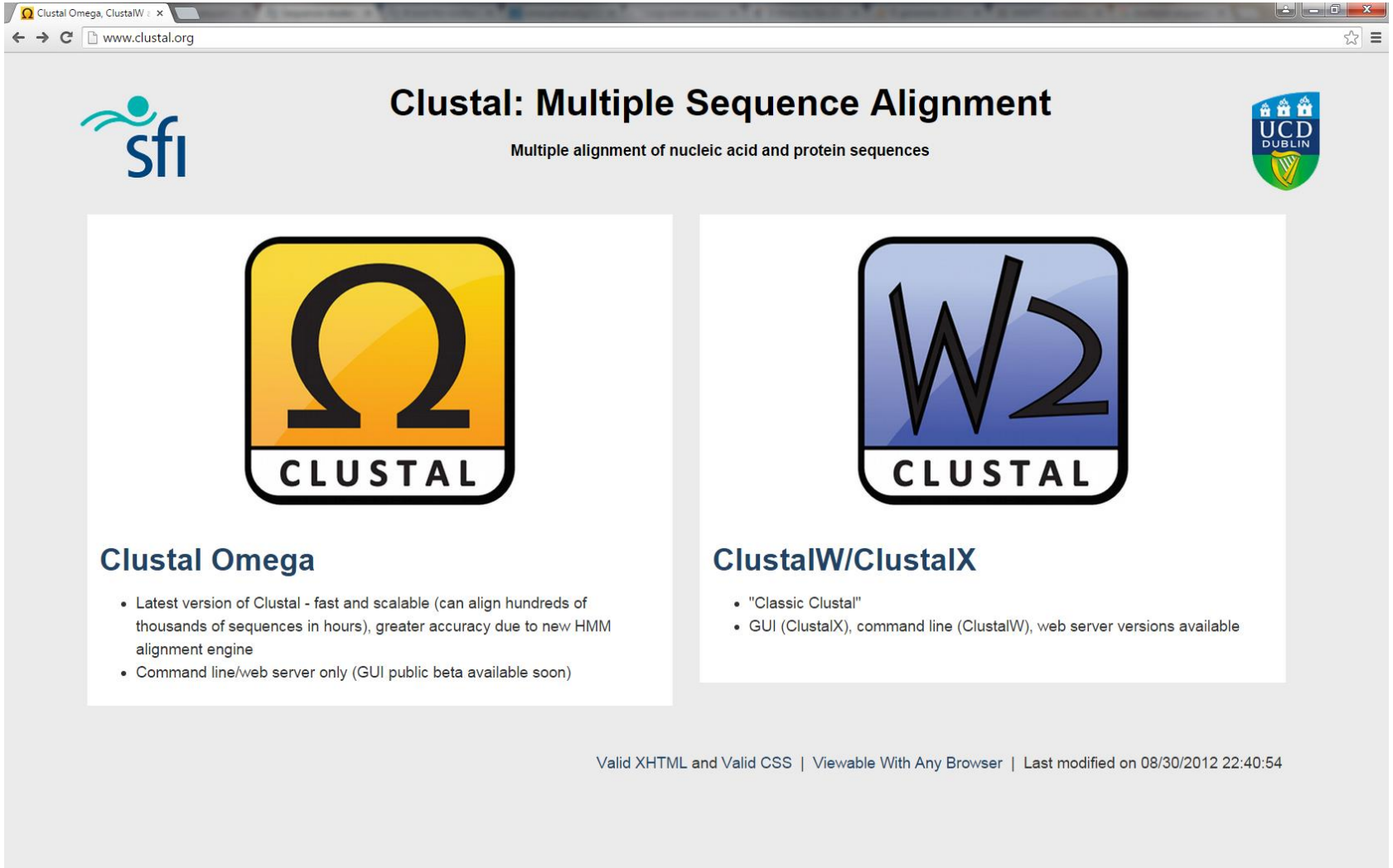
Efficiency / Speed /Accuracy



Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega
Fabian Sievers, Andreas Wilm, David Dineen, Toby J Gibson, Kevin Karplus, WeizhongLi, Rodrigo Lopez, Hamish McWilliam, Michael Remmert, Johannes Söding, Julie DThompson, Desmond G Higgins

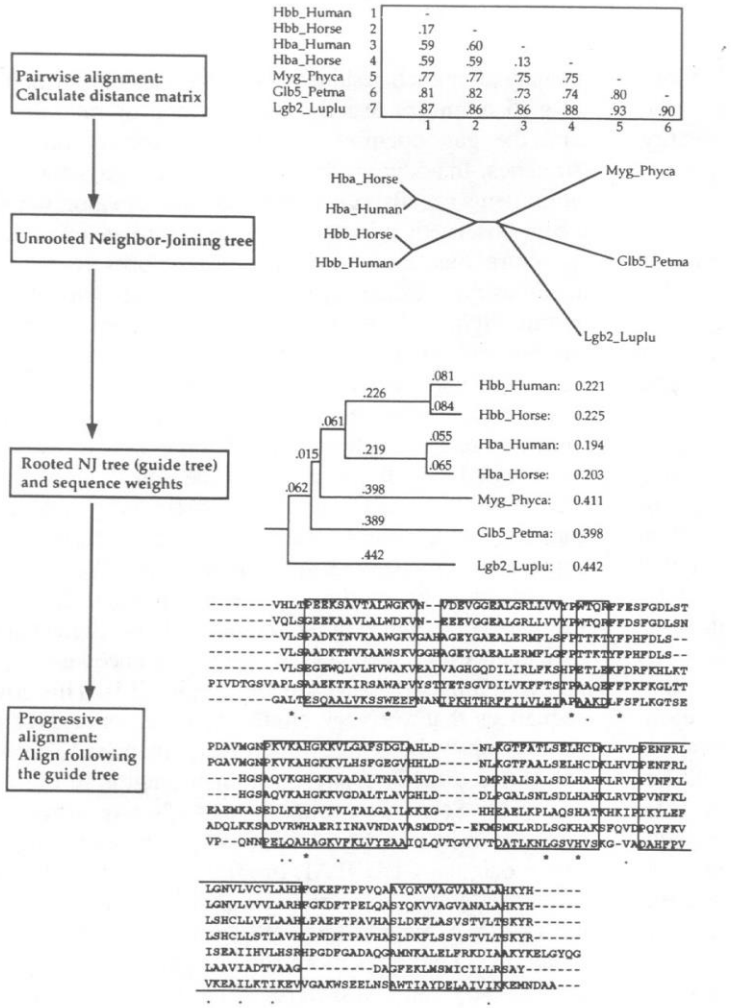
Clustal

Like many other MSA tools, Clustal has evolved to a couple of “flavors”



Progressive Scoring (Feng and Doolittle)

- All sequences are pairwise aligned and a score matrix is produced.
- A single “Guide” tree is constructed with branch length proportional to each pair score (ie…NJ method for tree construction).
- Closest pairs of sequences are aligned and more distant pairs are added according to the “Guide” tree.



Weight matrix…
PAM / BLOSUM
Fixed throughout the alignment.

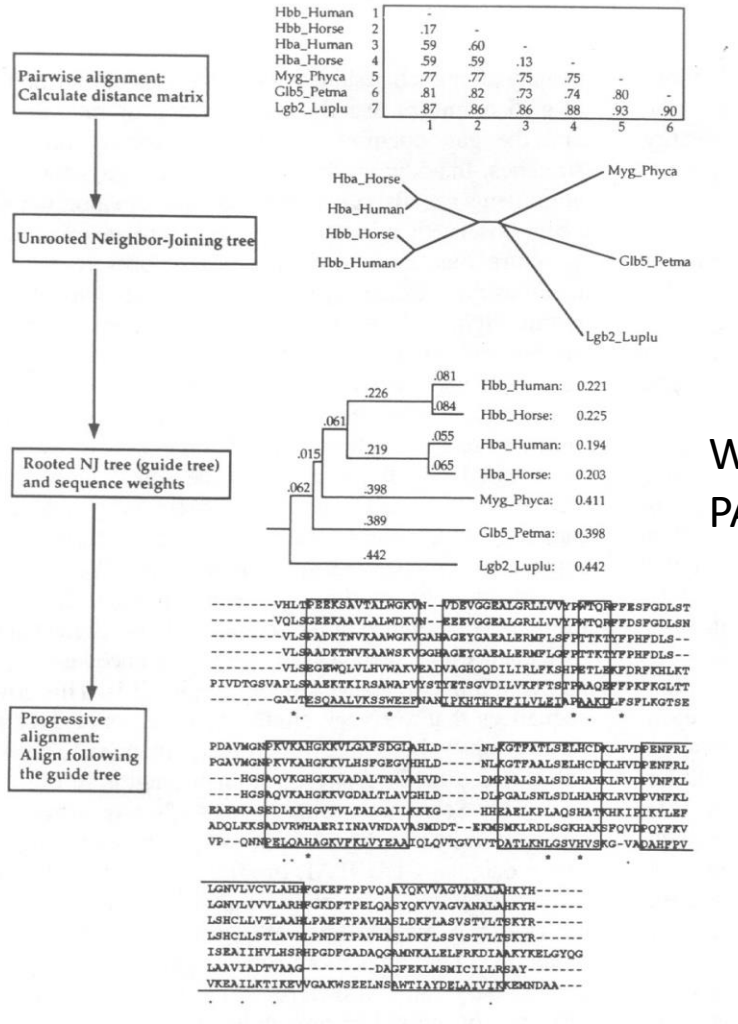
Feng DF, Doolittle RF. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J Mol Evol. 1987;25(4):351-60

ClustalW

Overcomes several problems related to Progressive scoring:

- Weighting substitution matrix of choice may not work for sequences of higher divergence….
- Gap penalties may vary with ranges of sequence divergence…
- Probabilities of a Gap occurring vary on the biochemistry of the aligned residues….eg…hydrophilic amino acids

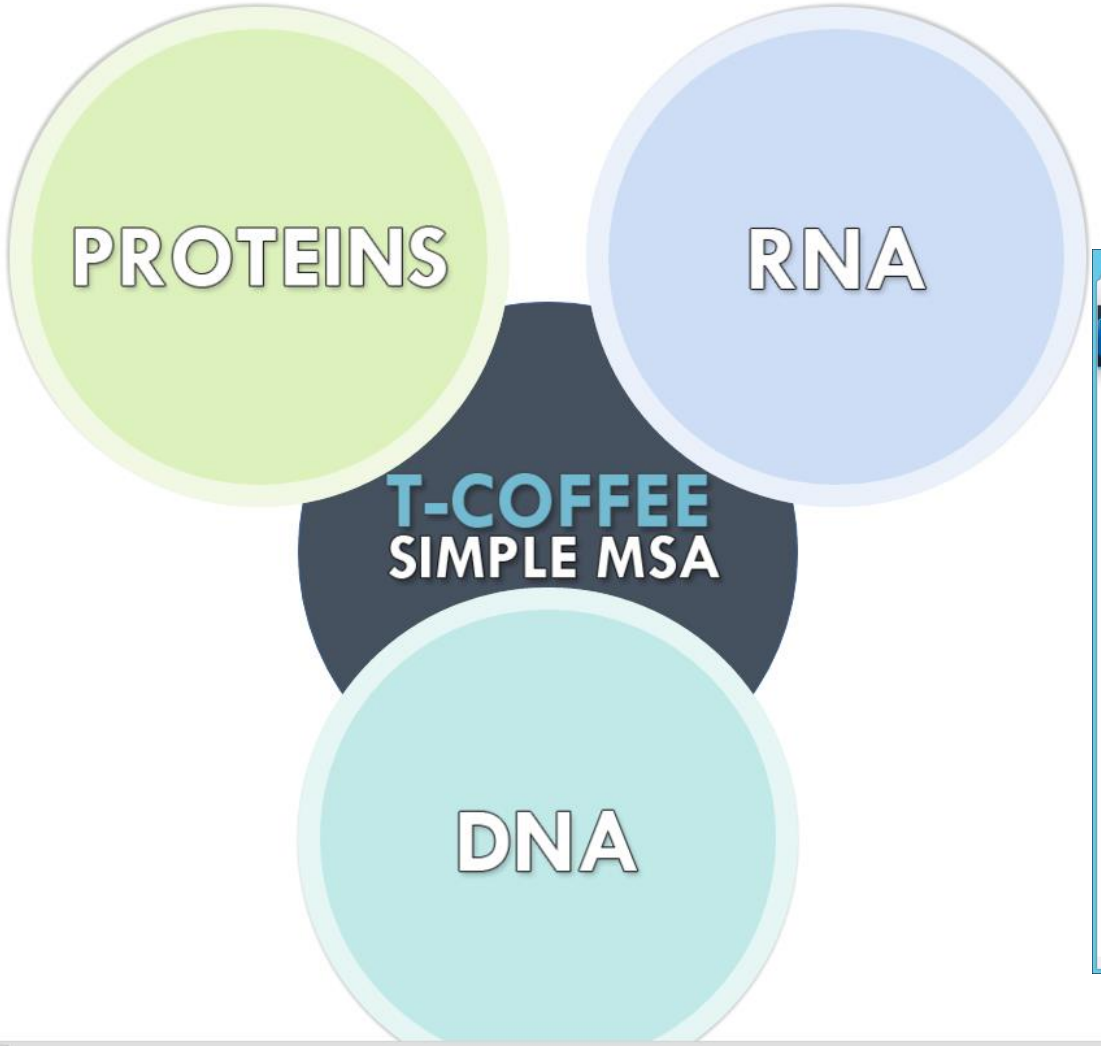
- CLUSTALW extends Progressive alignment by altering the gap penalties based on previous gaps, altering the weight matrix through the alignment, and then adding the most divergent sequences last.

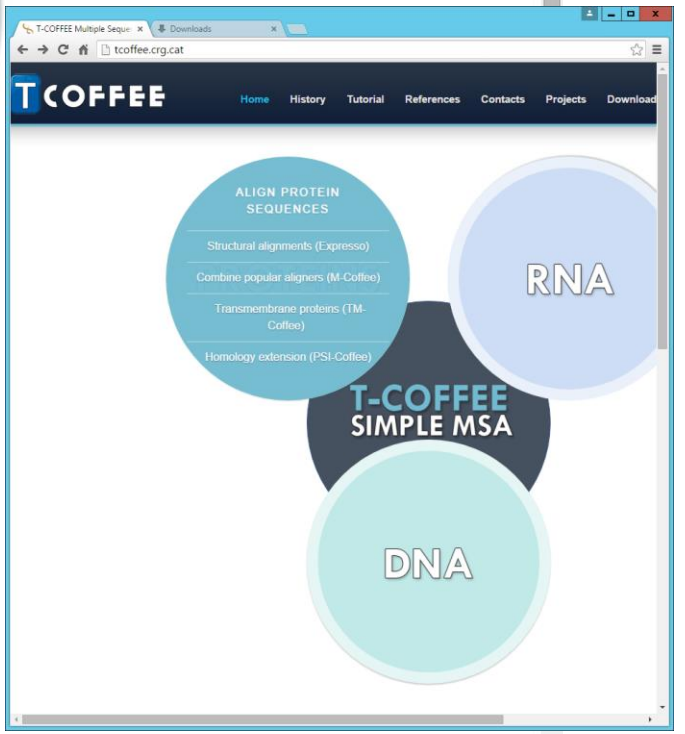


Weight matrix…
PAM / BLOSUM

Feng DF, Doolittle RF. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J Mol Evol. 1987;25(4):351-60

MANY “flavors”





MANY “flavors”

T-Coffee Method

- Progressive after pairwise library construction
- Libraries allow position specific weighting (no substitution matrices)
- Primary library weights are based on percent identity of the paired sequence.
- Extended libraries remove duplications to singletons and then sum weights.

Global Alignment Library

Local Alignment Library

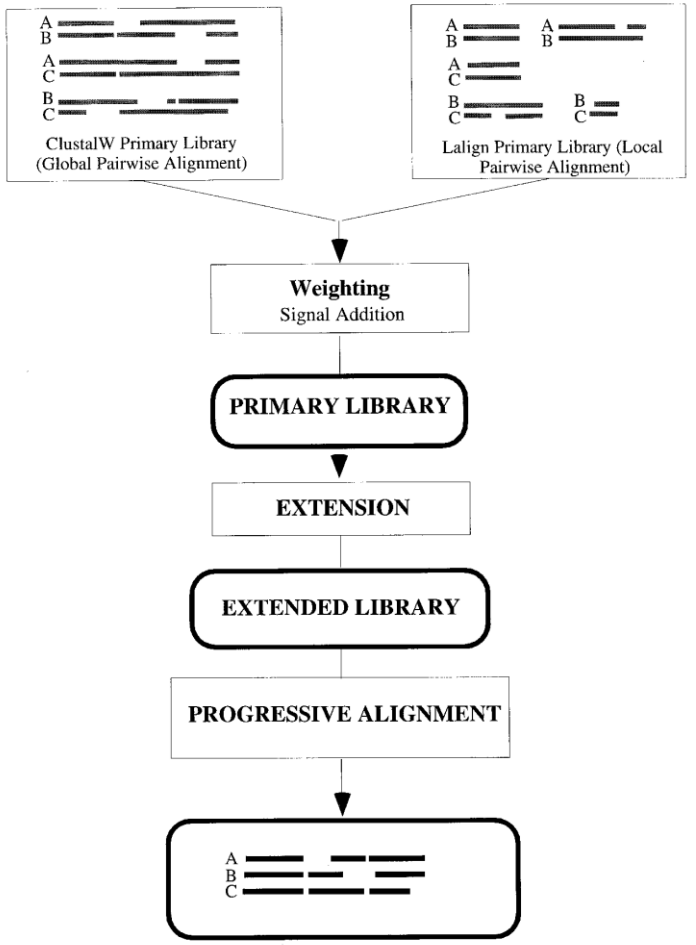


Figure 1. Layout of the T-Coffee strategy; the main steps required to compute a multiple sequence alignment using the T-Coffee method. Square blocks designate procedures while rounded blocks indicate data structures.

T-Coffee Method

- Progressive after pairwise library construction
- Libraries allow position specific weighting (no substitution matrices)
- Primary library weights are based on percent identity of the paired sequence.
- Extended libraries remove duplications to singletons and then sum weights.

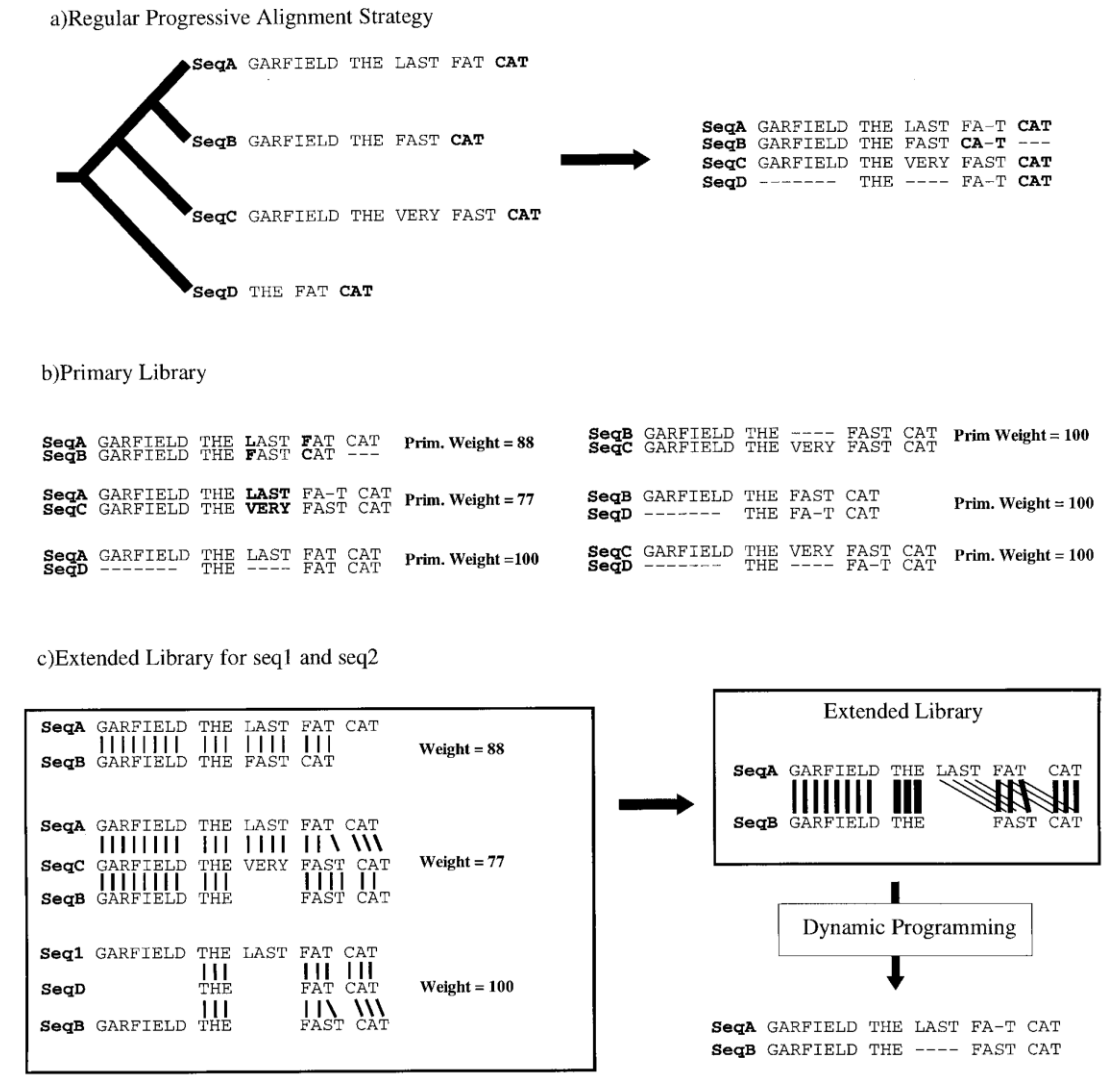
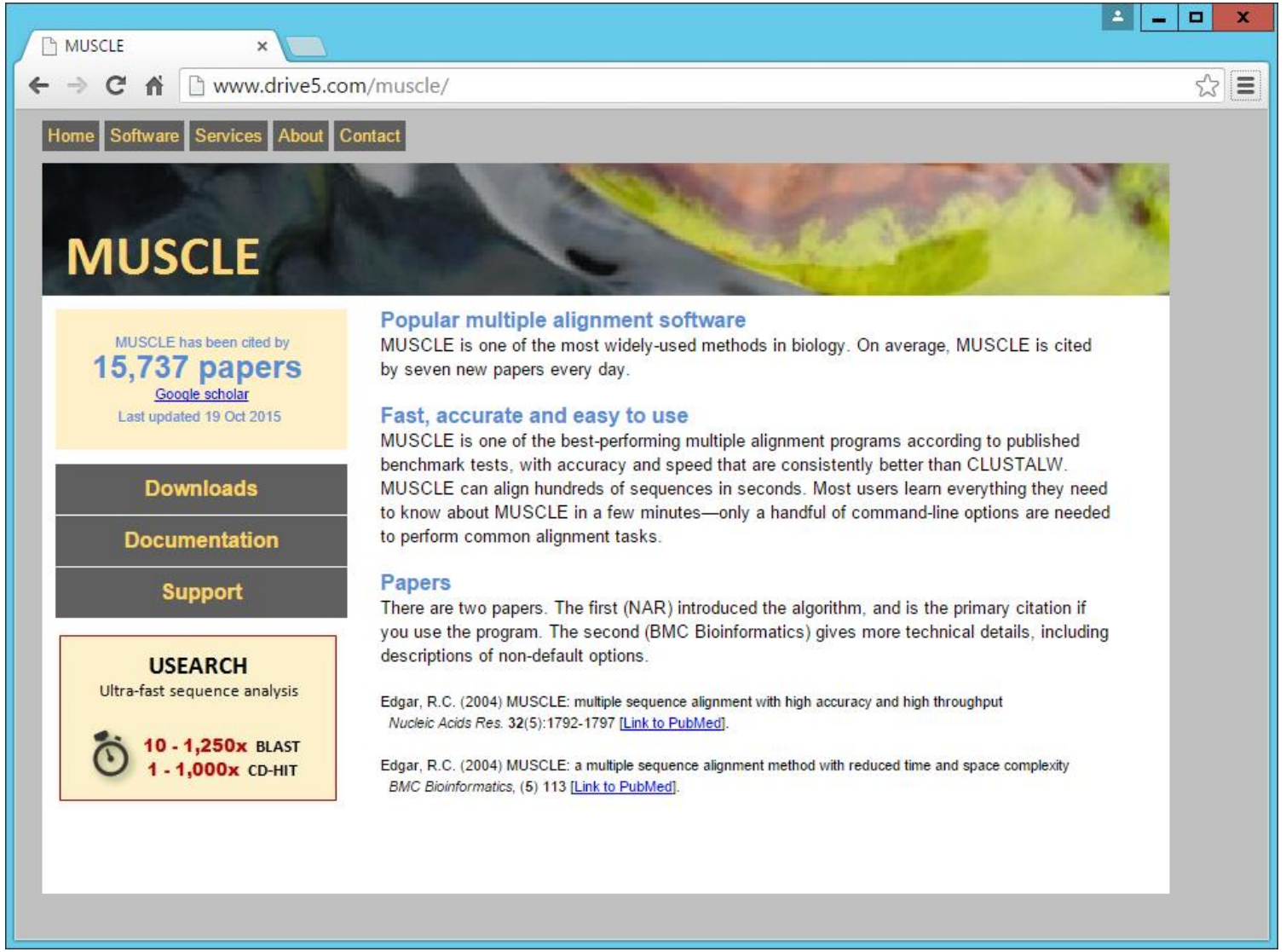


Figure 2. The library extension. (a) Progressive alignment. Four sequences have been designed. The tree indicates the order in which the sequences are aligned when using a progressive method such as ClustalW. The resulting alignment is shown, with the word CAT misaligned. (b) Primary library. Each pair of sequences is aligned using ClustalW. In these alignments, each pair of aligned residues is associated with a weight equal to the average identity among matched residues within the complete alignment (mismatches are indicated in bold type). (c) Library extension for a pair of sequences. The three possible alignments of sequence A and B are shown (A and B, A and B through C, A and B through D). These alignments are combined, as explained in the text, to produce the position-specific library. This library is resolved by dynamic programming to give the correct alignment. The thickness of the lines indicates the strength of the weight.

MUSCLE: **MU**ltiple **S**equence **C**omparison by **L**og- **E**xpectation

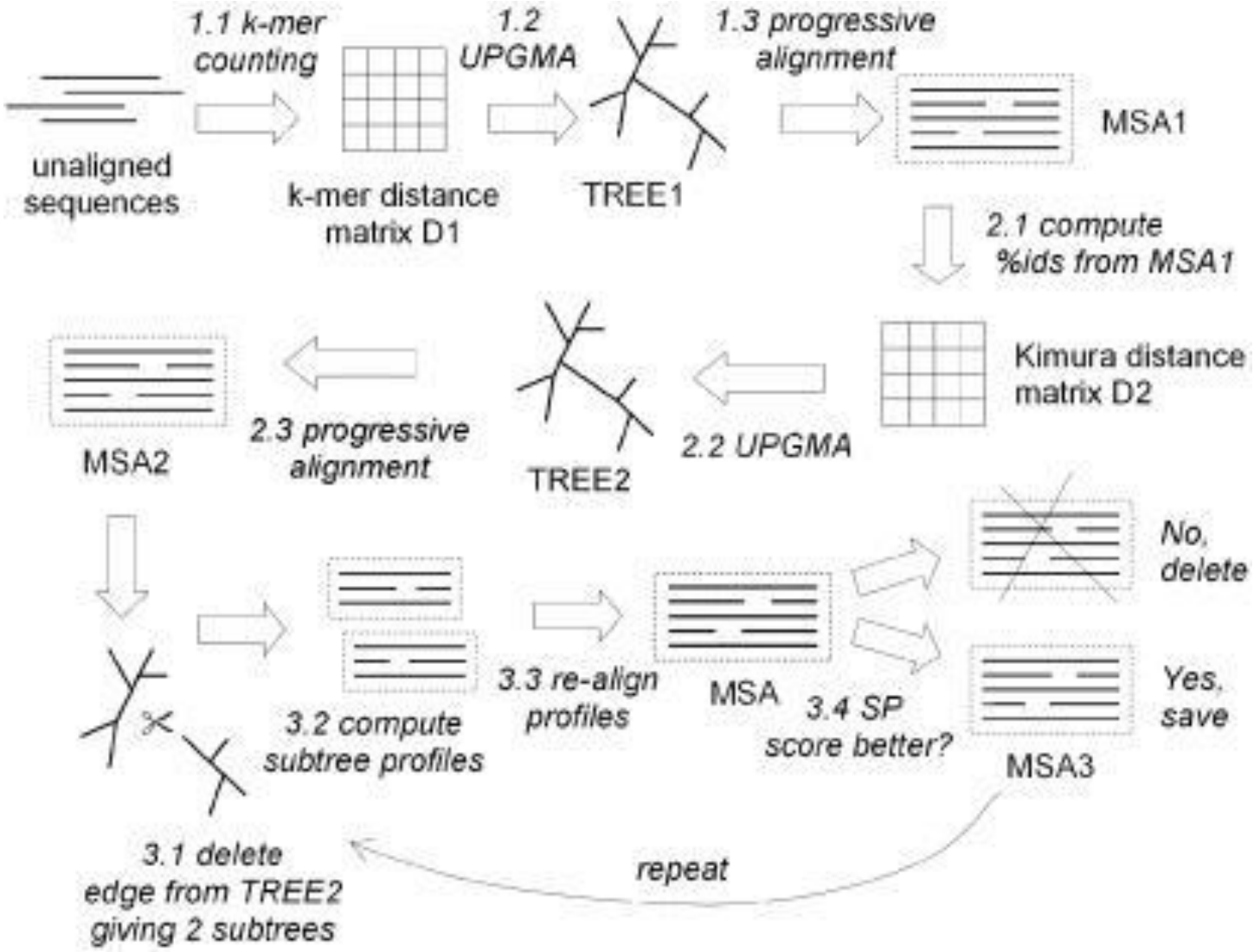


MUSCLE

Yet another extension of progressive scoring with interative progressive alignments.

K-mers are short identical sequence reads

Unweighted **P**air **G**roup **M**ethod with **A**rithmetic Mean



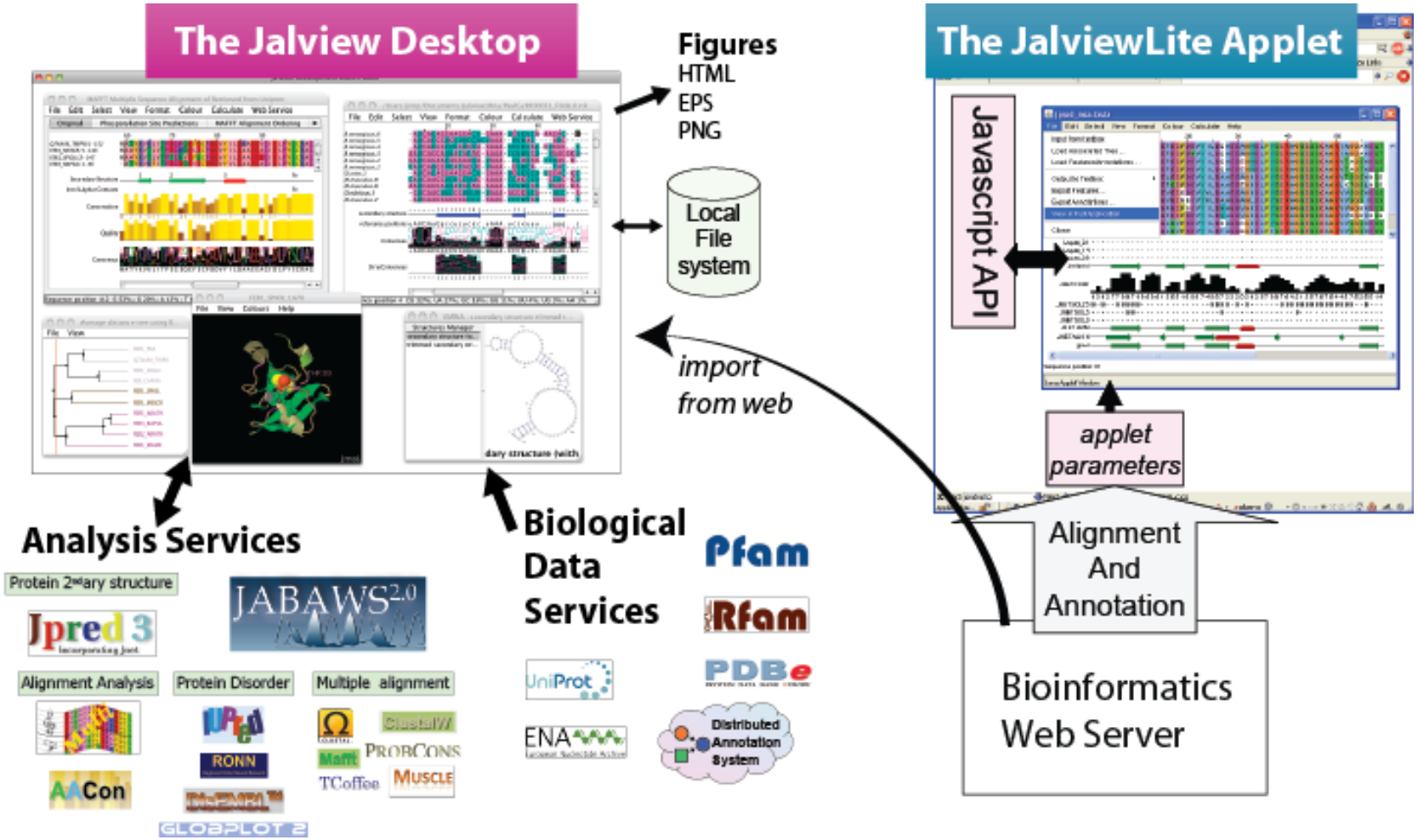
$$\mathrm{LE}^{xy} = (1 - f^{x}_{\mathrm{G}})\,(1 - f^{y}_{\mathrm{G}})\,\log \Sigma_{i}\,\Sigma_{j}\, f^{x}_{i}\, f^{y}_{j}\, p_{ij}/p_{i}\, p_{j}$$

240 PAM VTML matrix

Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004 Mar 19;32(5):1792-7. Print 2004.

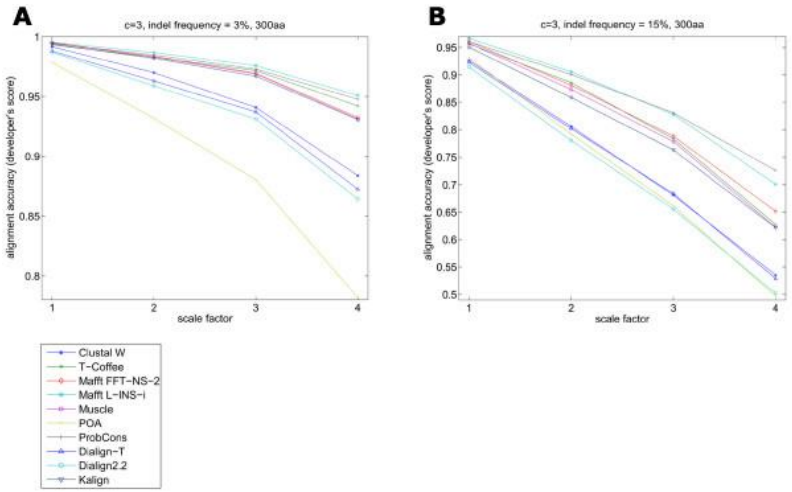
Do it all software:

- Eg..Jalview:

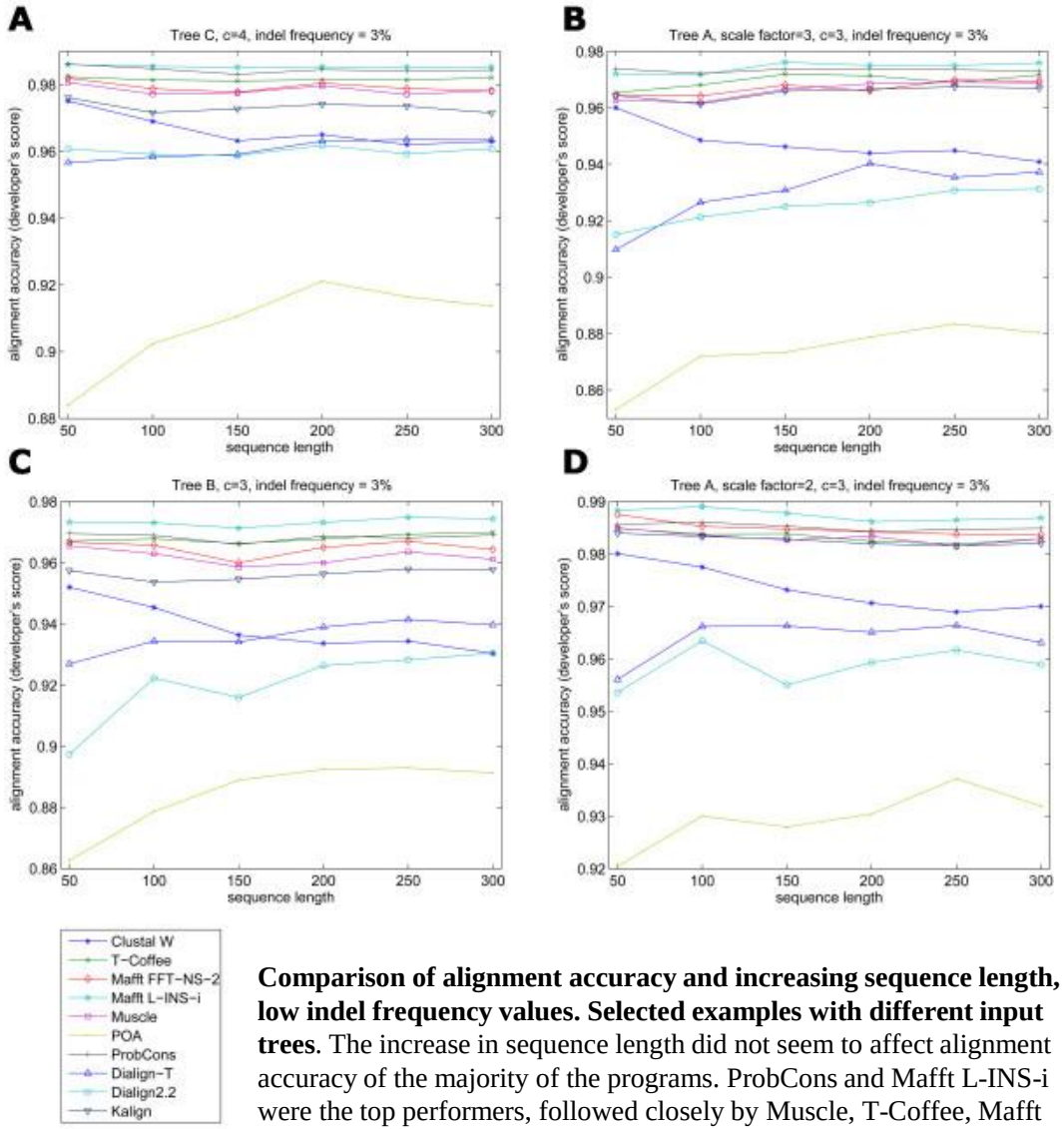


Benchmarking

Nuin PA, Wang Z, Tillier ER. The accuracy of several multiple sequence alignment programs for proteins. BMC Bioinformatics. 2006 Oct 24;7:471



Decrease in accuracy with an increase in the evolutionary scale factor of topology A. POA seemed to be the most affected by the increase of the scale factor applied to topology A from Figure 1. The top performers are again Mafft L-INS-i and ProbCons. An intermediary group formed by T-Coffee, Muscle, Mafft FFT-NS-2 and Kalign is followed by Dialign2.2, Dialign-T, Clustal W and POA that showed poor accuracy values as the scale factor increased.



Comparison of alignment accuracy and increasing sequence length, at low indel frequency values. Selected examples with different input trees. The increase in sequence length did not seem to affect alignment accuracy of the majority of the programs. ProbCons and Mafft L-INS-i were the top performers, followed closely by Muscle, T-Coffee, Mafft FFT-NS-2 and Kalign. Dialign2.2, Dialign-T and Clustal W presented a better accuracy than POA in most of the cases. Scale factor: value by which tree's branch lengths are multiplied, making them uniformly change; *c* is the Qian-Goldstein distribution value that determines average length of indels.

Other specific areas not discussed, but important:

- HMM , Genetic algorithms
- Benchmarking methods (BaliBase 3.0)

Thompson JD, Koehl P, Ripp R, Poch O. BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. Proteins. 2005 Oct 1;61(1):127-36.

Conclusion

- MSA requires pre-requisite knowledge to make informed choices about method choice
- MSA requires pre-requisite knowledge to make informed choices about interpretation of the output
- MSA is a core method for many bioinformatics studies
- MSA has improved with information gain and technological advances