

Multiple Sequence Alignment

Scott Walsmsley, 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?

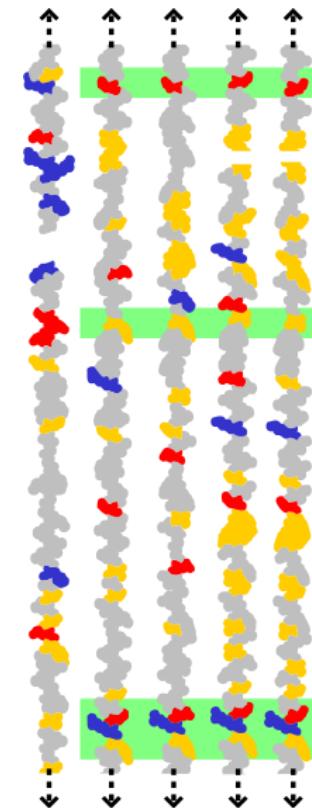
Alignment of 3 or (many) more sequences

- RNA / DNA
- Protein
- Structure

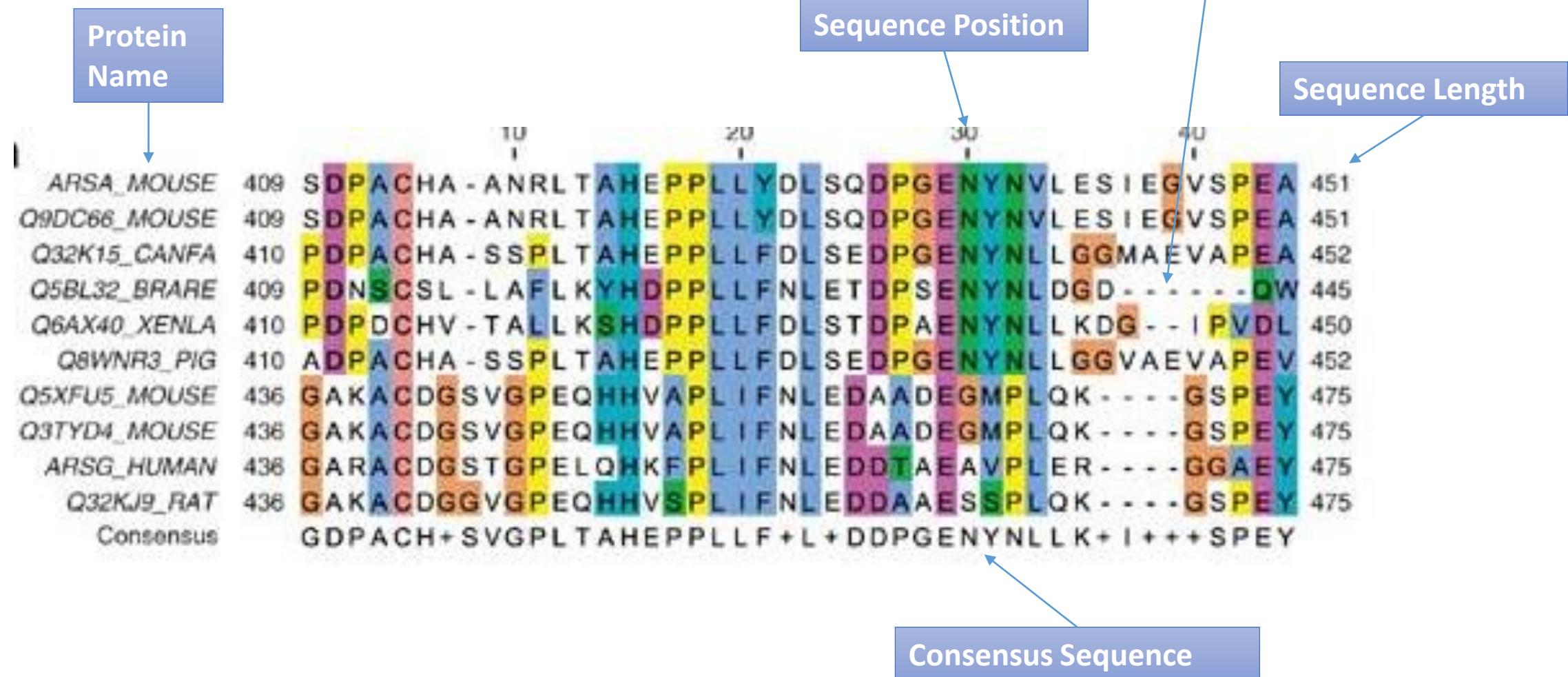
Global versus Local Alignments

- Whole sequence vs Local

Progressive versus Iterative versus others....



Anatomy of a MSA



Anatomy of an MSA:

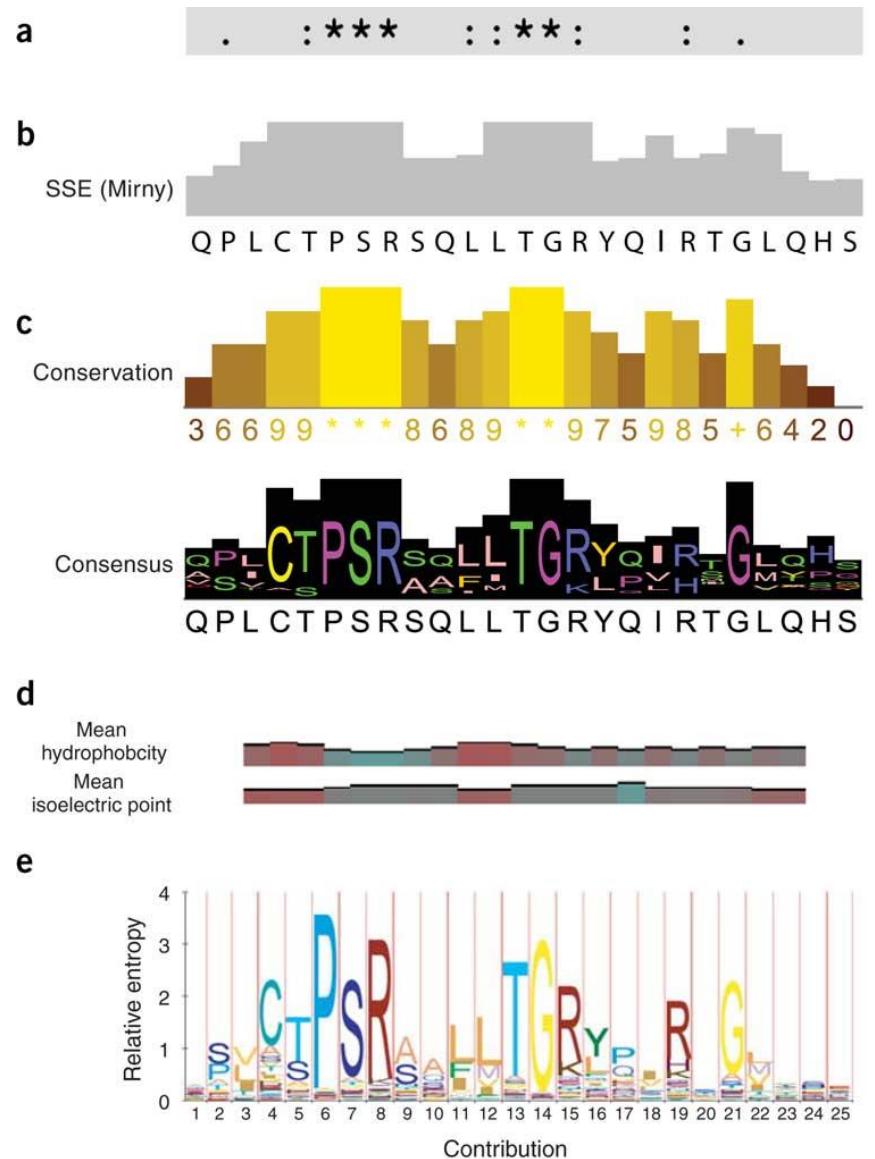
(a) ClustalW quality annotation from ClustalX

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

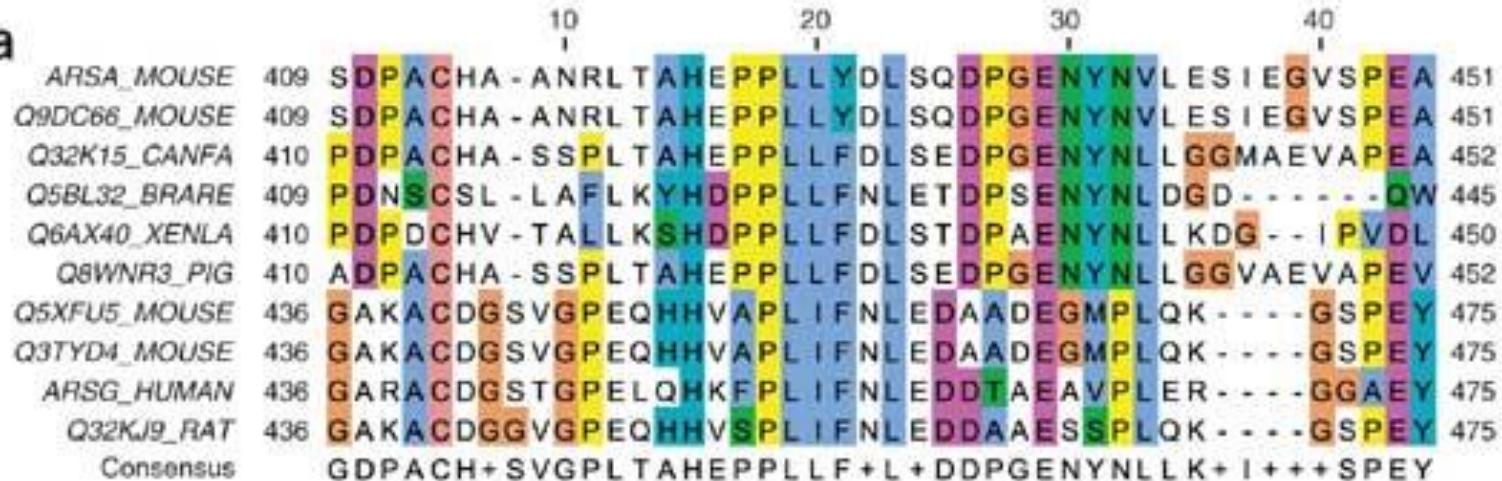
(c) Amino acid physicochemical property conservation, consensus and overlaid sequence logo from Jalview.

(d) Mean hydrophobicity and isoelectric point from Geneious.

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



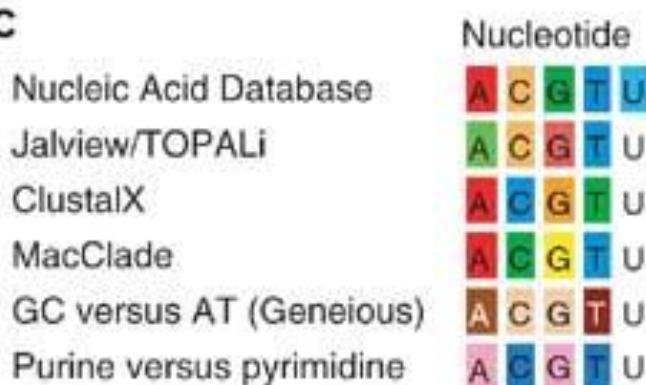
MSA: a



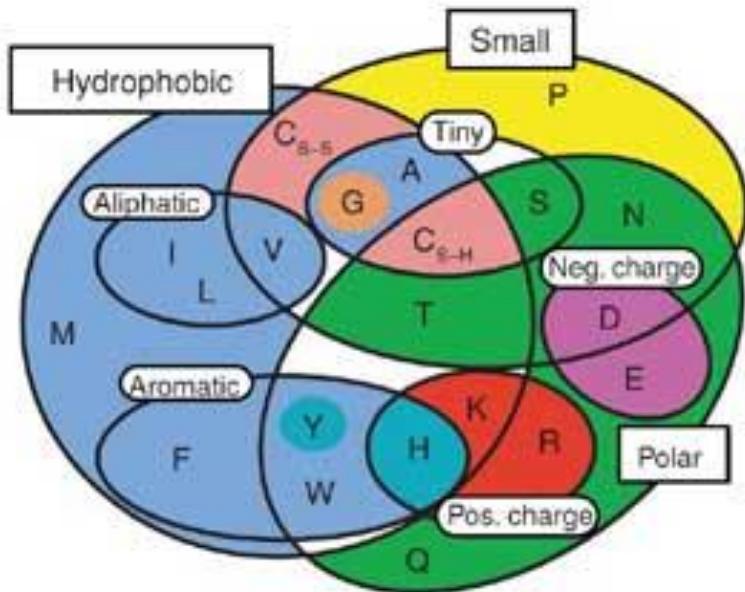
b



c



d



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”

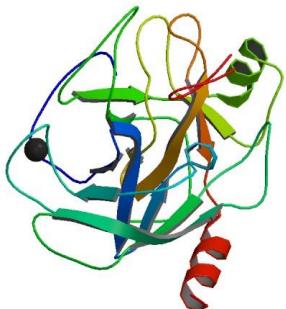
Valdar WS. Scoring residue conservation. Proteins. 2002 Aug 1;48(2):227-41. Review

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

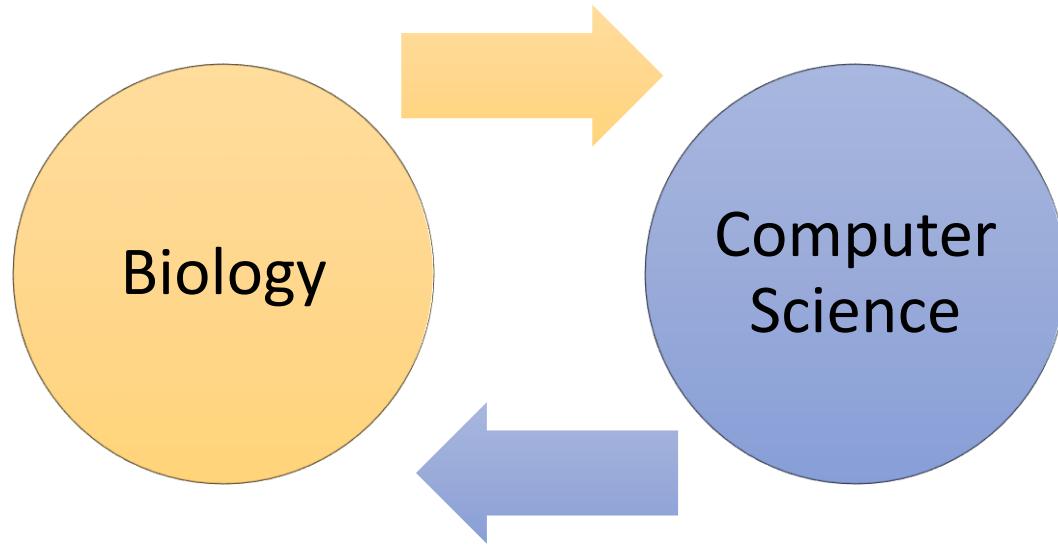
Q: What makes a good Multiple Sequence Alignment?



		1U	2U	3U	4U		
ARSA_MOUSE	409	S D P A C H A - A N R L T A H E P P L Y D L S Q D P G E N Y N V L E S I E G V S P E A					451
Q9DC68_MOUSE	409	S D P A C H A - A N R L T A H E P P L Y D L S Q D P G E N Y N V L E S I E G V S P E A					451
Q32K15_CANFA	410	P D P A C H A - S S P L T A H E P P L L F D L S E D P G E N Y N L L G G M A E V A P E A					452
Q5BL32_BRARE	409	P D N S C S L - L A F L K Y H D P P L L F N L E T D P S E N Y N L L D G D - - - - - Q W					445
Q6AX40_XENLA	410	P D P D C H V - T A L L K S H D P P L L F D L S T D P A E N Y N L L K D G - - - I P V D L					450
Q8WNR3_PIG	410	A D P A C H A - S S P L T A H E P P L L F D L S E D P G E N Y N L L G G V A E V A P E V					452
Q5XFU5_MOUSE	436	G A K A C D G S V G P E Q H H V A P L I F N L E D A A D E G M P L Q K - - - - G S P E Y					475
Q3TYD4_MOUSE	436	G A K A C D G S V G P E Q H H V A P L I F N L E D A A D E G M P L Q K - - - - G S P E Y					475
ARSG_HUMAN	436	G A R A C D G S T G P E L Q H K F P L I F N L E D D T A E A V P L E R - - - - G G A E Y					475
Q32KJ9_RAT	436	G A K A C D G G V G P E Q H H V S P L I F N L E D D A A E S S P L Q K - - - - G S P E Y					475
Consensus		G D P A C H + S V G P L T A H E P P L L F + L + D D P G E N Y N L L K + I + + + S P E Y					

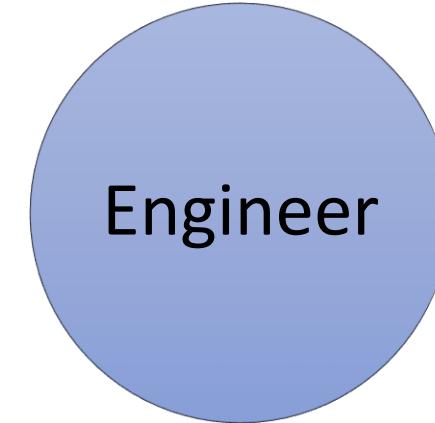
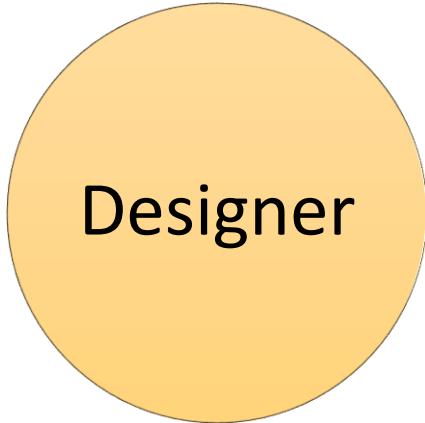
$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} nk} \quad k = 0, \dots, N-1.$$

Different perspectives on a good alignment?



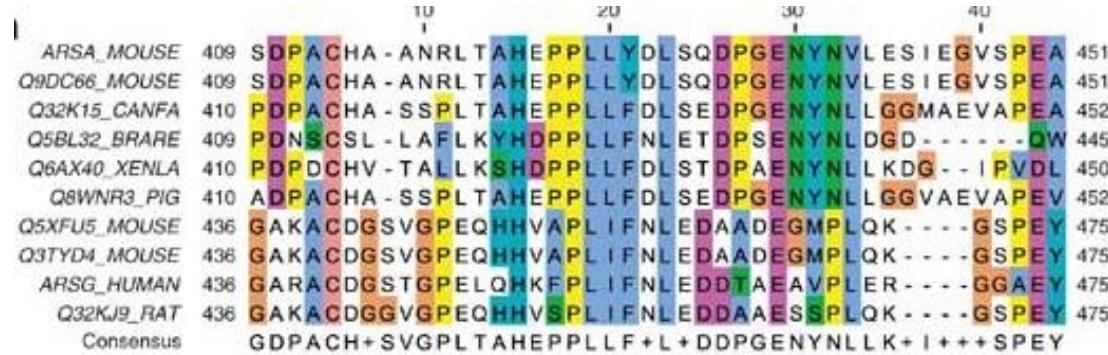
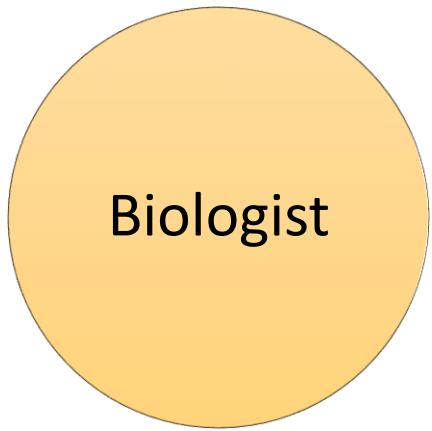
We have the same goal in mind: the optimum solution that makes sense...

Different perspectives on a good product:

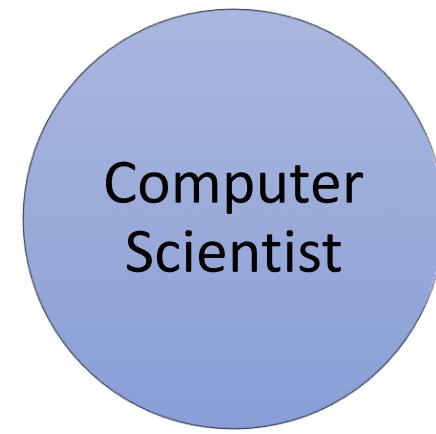


The interpretation of what makes it good is different.....

Different perspectives on a good alignment:



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:

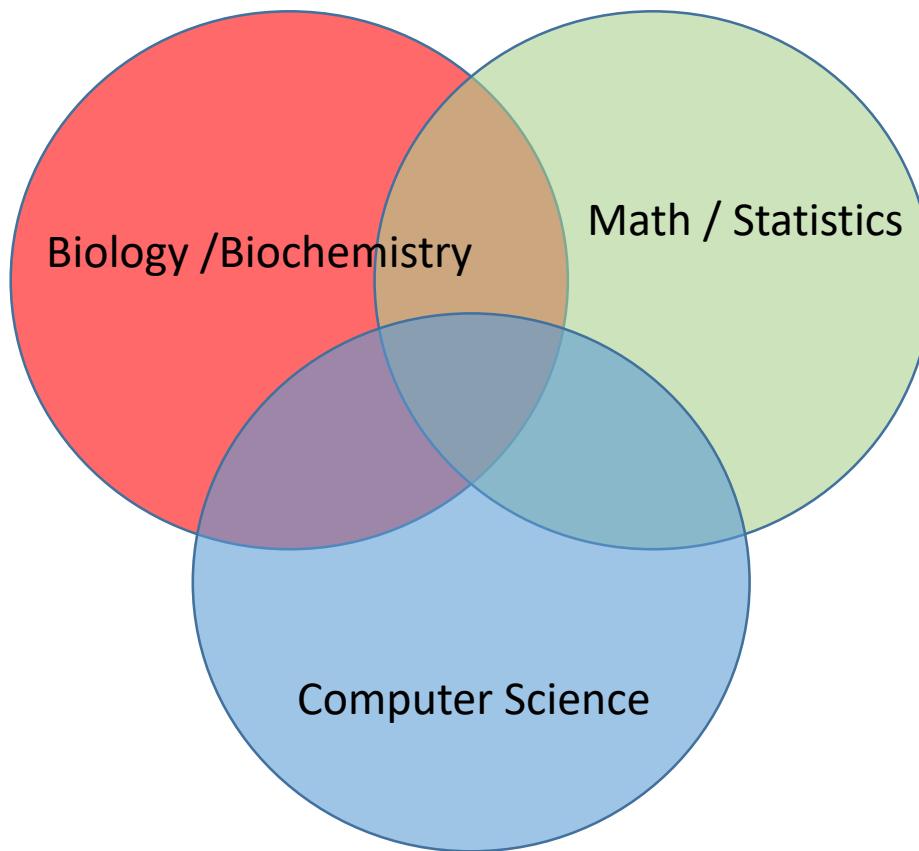
Computational / Math / Statistics

- Pairwise sequence alignment methods
- Substitution matrices
- Phylogenetic trees

Molecular Biology / Biochemistry

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

Pre-requisite knowledge

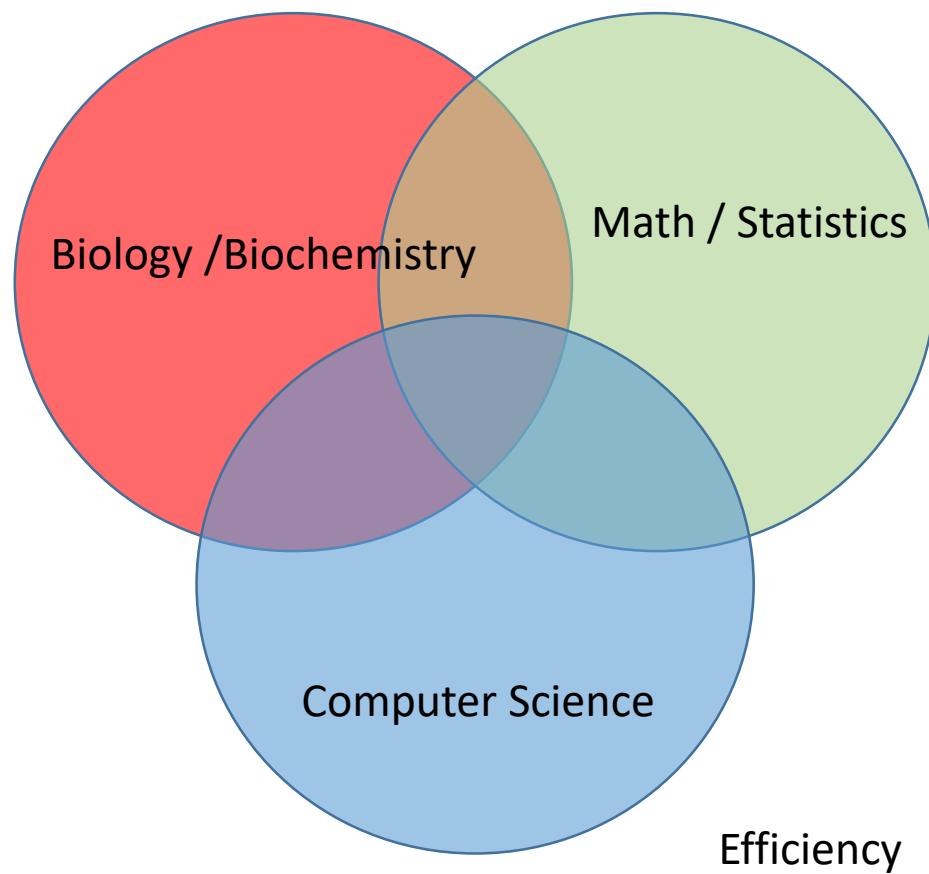


How specific in one field you want to go is up to you, but there are always others to collaborate with to complement your skillset.

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

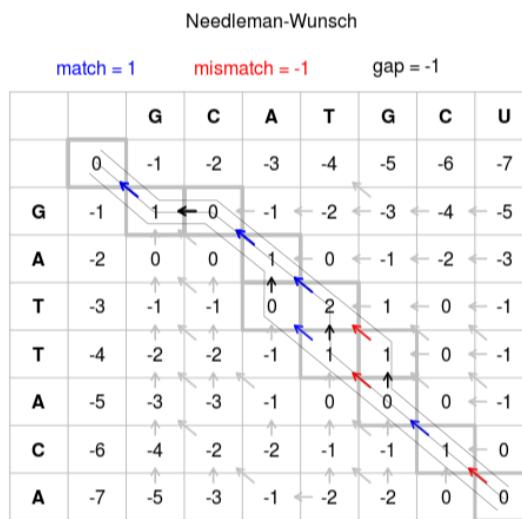
Pre-requisite knowledge

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)



Protein BLAST: search pro

BLAST® Basic Local Alignment Search Tool

Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
Or, upload file
Job Title
Align two or more sequences
Choose Search Set
Database: Non-redundant protein sequences (nr)
Organism: Optional
Exclude: Enter organism name or id—completions will be suggested
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.
Exclude: Models (XM/XP) Uncultured/environmental sample sequences
Optional
Entrez Query: Enter an Entrez query to limit search
Program Selection
Algorithm: blasp (protein-protein BLAST)
PSI-BLAST (Position-Specific Iterated BLAST)
PHI-BLAST (Pattern Hit Initiated BLAST)
DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
Choose a BLAST algorithm
BLAST

The figure shows the NCBI Protein BLAST search interface. It includes fields for entering a query sequence, choosing a database, selecting search parameters like matrix and gap cost, and choosing an algorithm. The 'Algorithm' section is expanded, showing options for blasp, PSI-BLAST, PHI-BLAST, and DELTA-BLAST.

Protein BLAST: search pro

BLAST® Basic Local Alignment Search Tool

Search database Non-redundant protein sequences (nr) using Blasp (protein-protein BLAST)

Algorithm parameters

General Parameters

Max target sequences: 100
Select the maximum number of aligned sequences to display

Short queries: Automatically adjust parameters for short input sequences

Expect threshold: 10

Word size: 6

Max matches in a query range: 0

Scoring Parameters

Matrix: BLOSUM62
PAM30
PAM70
PAM250
BLOSUM80
BLOSUM80
Gap Costs: Extension: 1
Compositional adjustments: BLOSUM45
BLOSUM50
BLOSUM90
BLOSUM90

Filters and Masking

Filter: Low complexity regions
Mask: Mask for lookup table only
Mask lower case letters

BLAST

Search database Non-redundant protein sequences (nr) using Blasp (protein-protein BLAST)

Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback

NCBI | NLM | NIH | DHHS

BLAST is a registered trademark of the National Library of Medicine.

The figure shows a detailed view of the NCBI Protein BLAST search interface. It includes sections for algorithm parameters, scoring parameters (matrix, gap costs, compositional adjustments), filters, and masking. The 'Algorithm parameters' section is expanded, showing specific settings for the BLAST search.

Pre-requisite knowledge

Computational / Math / Statistics

- Substitution matrices

PAM

Choice of mutation matrix can effect pairwise and subsequent MSA

BLOSUM

A good handle on how the choice effects your MSA might be based on how evolutionarily distant the sequences of interest are.

DYNAMIC

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	-1	-2	-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: Point Accepted Mutation

Pre-requisite knowledge

Computational / Math / Statistics

- Phylogenetic trees

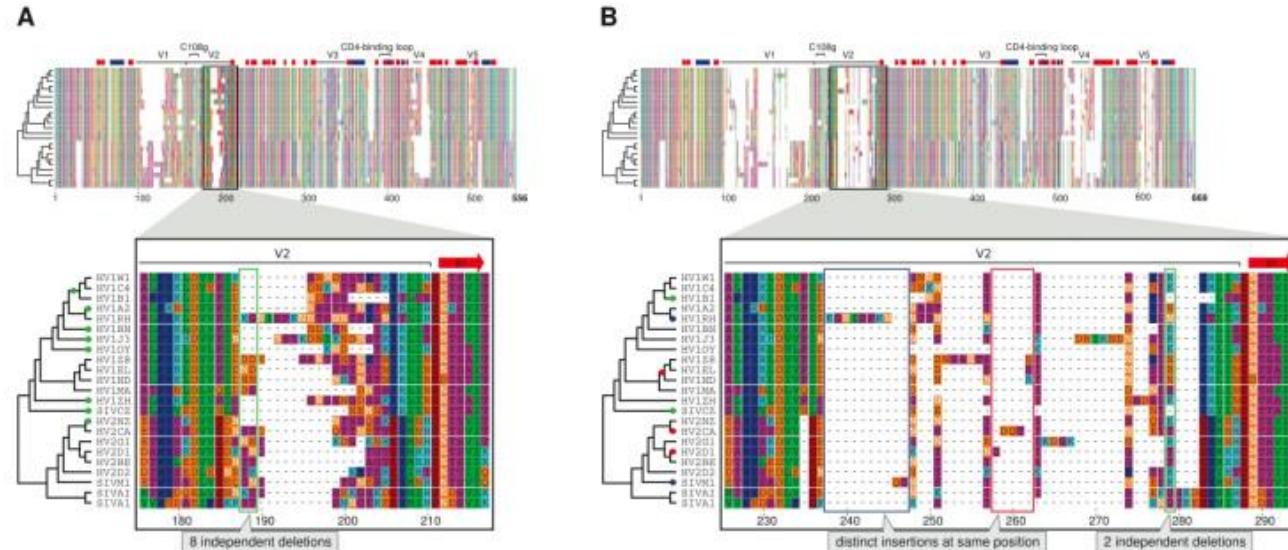


Fig. 1. Different sequence alignment approaches can give contradicting pictures of evolutionary mechanisms behind functional sequence changes. (A) (top) The CLUSTAL W (12) alignment of gp120 from different strains of human and simian immunodeficiency virus (5) represents a typical structural matching 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_{+P} (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).

Pre-requisite knowledge

Computational / Math / Statistics

- Alphabets

DNA (n= 4)

RNA (n = 4)

Amino Acids (n = 20)

		2. nucleotide					
		U	C	A	G		
		UUU } F	UCU }	UAU }	UGU }	U	C
U		UUC }	UCC }	UAC }	UGC }	A	A
UUA }		L	UCA }	UAA St, Q ₆	UGA St, W _{1,2,3,4,5}	G	G
UUG }		UCG }	UAG St, Q ₆	UGG W			
		CUU }	CCU }	CAU }	CGU }	U	C
C		CUC }	CCC }	CAC }	CGC }	A	A
CUA }		CCA }	CAA }	CGA }	CGA }	G	G
CUG }		L, T ₅ , S ₇	CCG }	CAG }	CGG }		
		AUU }	ACU }	AAU }	AGU }	U	C
A		AUC }	ACC }	AAC }	AGC }	A	A
AUA }		I, M _{2,3,4,5}	ACA }	AAA K, N ₁	AGA }	G	G
AUG }		M	ACG }	AAG K	AGG R, S _{1,2} , G ₃ , St ₄		
		GUU }	GCU }	GAU }	GGU }	U	C
G		GUC }	GCC }	GAC }	GGC }	A	A
GUA }		V	GCA }	GAA }	GGA }	G	G
GUG }			GCG }	GAG }	GGG }		
		3. nucleotide					

Pre-requisite knowledge

Computational / Math / Statistics & Biochemistry

- Alphabets

DNA (n= 4)

RNA (n = 4)

Amino Acids (n = 20)

What other alphabet exists?

		2. nucleotide									
		U	C	A	G						
		UUU } F	UCU }	UAU }	UGU }	U					
U		UUC	UCC }	UAC }	UGC }	C	C				
UUA }		UCA	S		UAA St, Q ₆	UGA St, W _{1,2,3,4,5}	A				
UUG }		UCG			UAG St, Q ₆	UGG W	G				
		CUU }	CCU }	CAU }	CGU }	U					
C		CUC }	CCC }	CAC }	CGC }	C	C				
CUA }		CCA	P		CGA }	CGA R	A				
CUG L, T ₅ , S ₇		CCG			CGG	G					
		AUU }	ACU }	AAU }	AGU }	U					
A		AUC }	ACC }	AAC }	AGC }	C	C				
AUA I, M _{2,3,4,5}		ACA }	T		AGA }	A	A				
AUG M		ACG			AGG R, S _{1,2} , G ₃ , St ₄	G	G				
		GUU }	GCU }	GAU }	GGU }	U					
G		GUC }	GCC }	GAC }	GGC }	C	C				
GUA V		GCA	A		GGA }	A	A				
GUG }		GCG			GGG	G					
1. nucleotide											
3. nucleotide											

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?

		2. nucleotide									
		U	C	A	G						
		UUU } F	UCU }	UAU }	UGU }	U					
U		UUC	UCC }	UAC }	UGC }	C	C				
UUA }		UCA	S		UAA St, Q ₆	UGA St, W _{1,2,3,4,5}	A				
UUG }		UCG			UAG St, Q ₆	UGG W	G				
		CUU }	CCU }	CAU }	CGU }	U					
C		CUC }	CCC }	CAC }	CGC }	C	C				
CUA }		CCA	P		CGA }	R	A				
CUG L, T ₅ , S ₇		CCG			CGG	G					
		AUU }	ACU }	AAU }	AGU }	U					
A		AUC }	ACC }	AAC }	AGC }	C	C				
AUA I, M _{2,3,4,5}		ACA }	T		AGA }	A	A				
AUG M		ACG			AGG R, S _{1,2} , G ₃ , St ₄	G	G				
		GUU }	GCU }	GAU }	GGU }	U					
G		GUC }	GCC }	GAC }	GGC }	C	C				
GUA V		GCA	A		GGA }	A	A				
GUG }		GCG			GGG	G					
1. nucleotide											
3. nucleotide											

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.

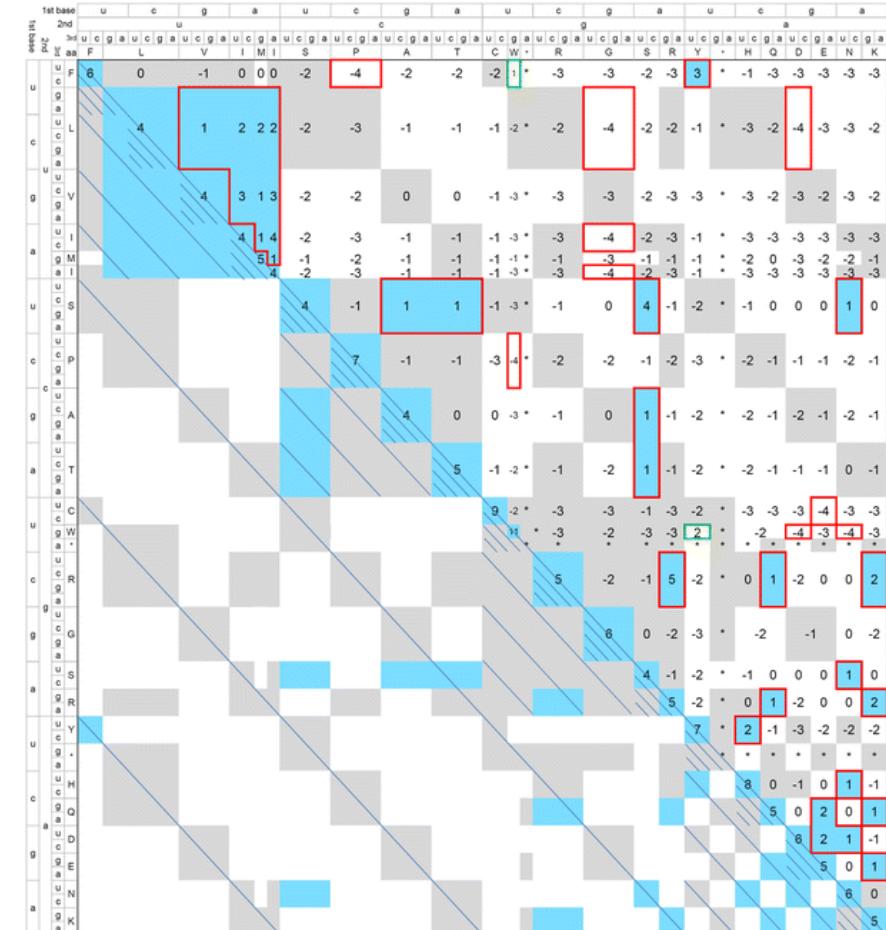
Amino acid (anticodon)	Codon group	Usage of codon ending in				Total	%
		A	C	G	T		
Ala (UGC)	GCN	94	108	4	88	294	7.53
Cys (GCA)	TGY	0	19	0	11	30	0.77
Asp (GUC)	GAY	0	37	0	39	76	1.95
Glu (UUC)	GAR	90	0	6	0	96	2.46
Phe (GAA)	TTY	0	120	0	127	247	6.33
Gly (UCC)	GGN	92	63	35	46	236	6.05
His (GUG)	CAY	0	66	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."



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*

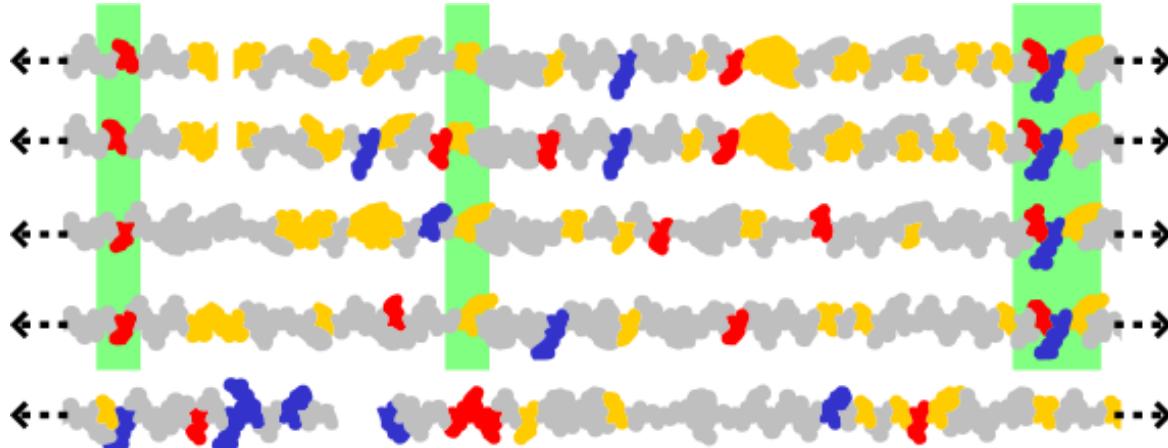
L = *Length of sequences*

Eg..... $F(x) = O(L^n)$

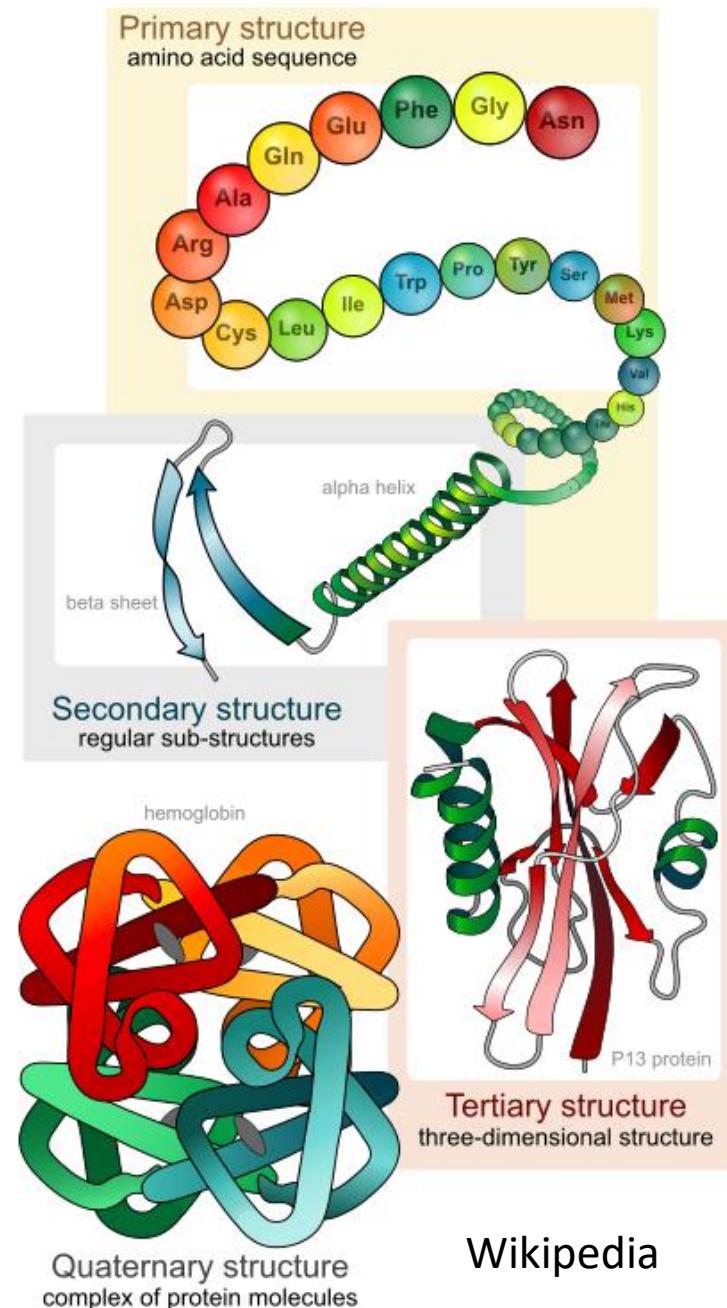
Pre-requisite knowledge

Biochemistry / Molecular Biology

- Mutation rates drive evolution
- Biophysical mechanisms produce mutation rates:
DNA / RNA Polymerase
- Insertion /Deletion : frameshift → altered CODON



Wikipedia



Wikipedia

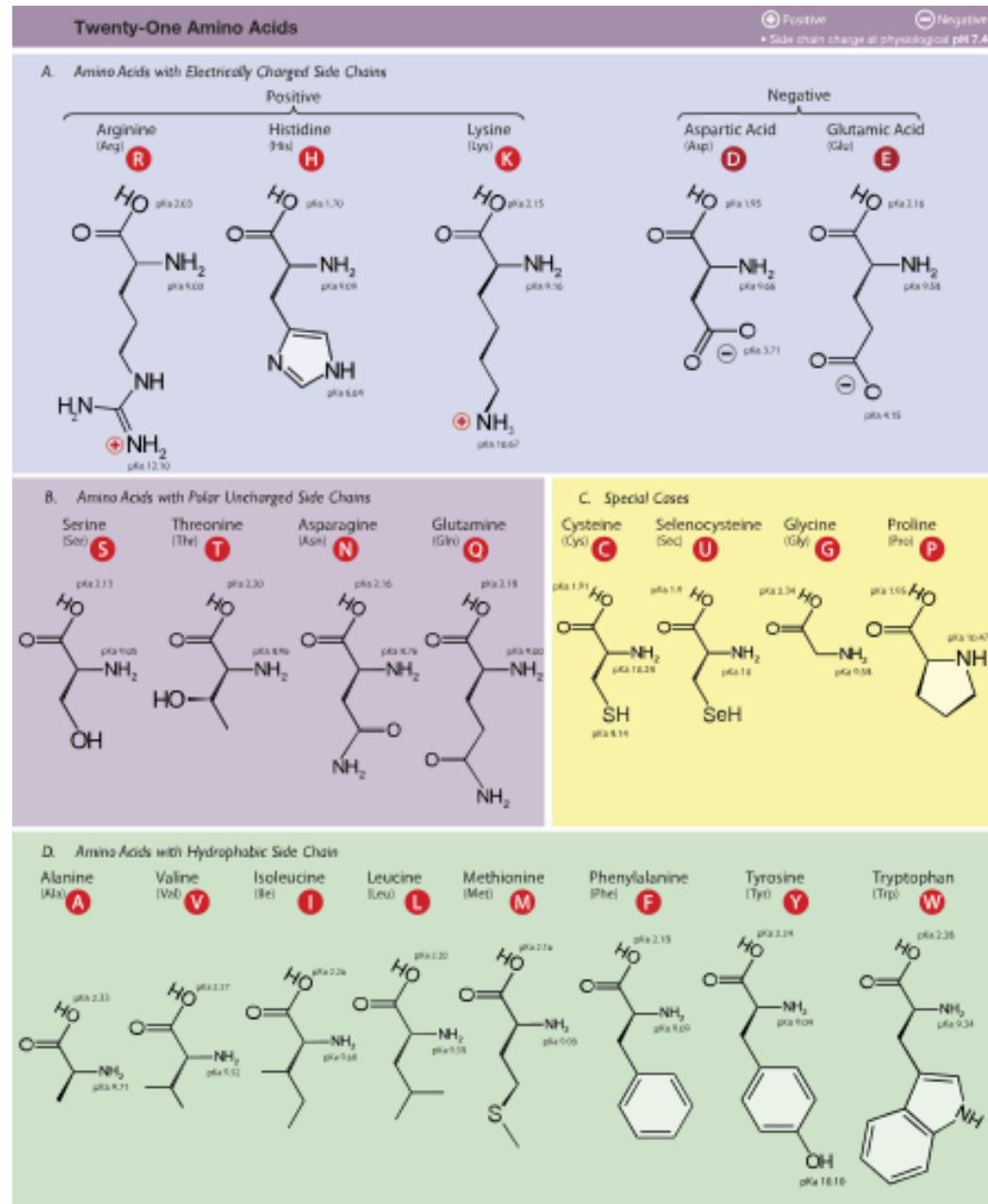
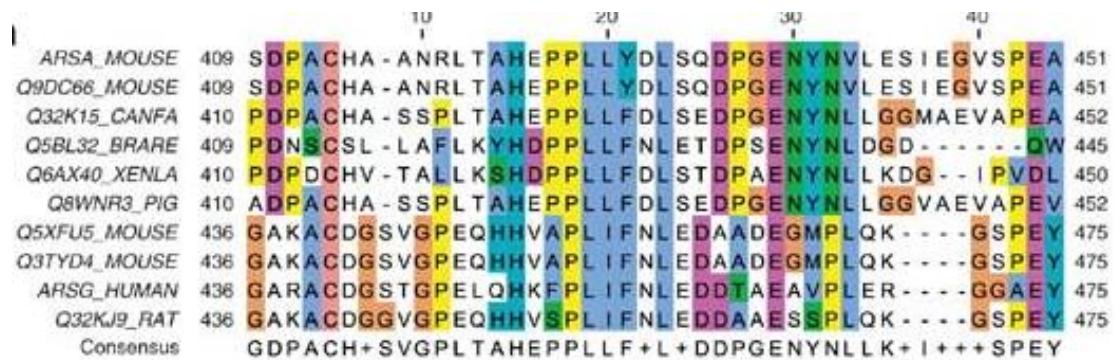
Pre-requisite knowledge

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

Examples of notable Mutations

		2nd base			
		U	C	A	G
3rd base in each row	1st base	UUU (Phe/F) Phenylalanine	UCU (Ser/S) Serine	UAU (Tyr/Y) Tyrosine	UGU (Cys/C) Cysteine
		UUC (Phe/F) Phenylalanine	UCC (Ser/S) Serine	UAC (Tyr/Y) Tyrosine	UGC (Cys/C) Cysteine
		UUA (Leu/L) Leucine	UCA (Ser/S) Serine	UAA Ochre (Stop)	UGA Opal (Stop)
		UUG (Leu/L) Leucine	UCG (Ser/S) Serine	UAG Amber (Stop)	UGG (Trp/W) Tryptophan
3rd base in each row	1st base	CUU (Leu/L) Leucine	CCU (Pro/P) Proline	CAU (His/H) Histidine	CGU (Arg/R) Arginine
		CUC (Leu/L) Leucine	CCC (Pro/P) Proline	CAC (His/H) Histidine	CGC (Arg/R) Arginine
		CUA (Leu/L) Leucine	- Myotonic dystrophy - SCA 8	CCA (Pro/P) Proline	CGA (Arg/R) Arginine
		CUG (Leu/L) Leucine	CCG (Pro/P) Proline	CAG (Gln/Q) Glutamine	CGG (Arg/R) Arginine
3rd base in each row	1st base	AUU (Ile/I) Isoleucine	ACU (Thr/T) Threonine	AAU (Asn/N) Asparagine	AGU (Ser/S) Serine
		AUC (Ile/I) Isoleucine	ACC (Thr/T) Threonine	AAC (Asn/N) Asparagine	AGC (Ser/S) Serine
		AUA (Ile/I) Isoleucine	ACA (Thr/T) Threonine	AAA (Lys/K) Lysine	AGA (Arg/R) Arginine
		AUG (Met/M) Methionine	ACG (Thr/T) Threonine	AAG (Lys/K) Lysine	AGG (Arg/R) Arginine
3rd base in each row	1st base	GUU (Val/V) Valine	Colorectal cancer		GGU (Gly/G) Glycine
		GUC (Val/V) Valine	GCU (Ala/A) Alanine	GAU (Asp/D) Aspartic acid	GGC (Gly/G) Glycine
		GUA (Val/V) Valine	GCC (Ala/A) Alanine	GAC (Asp/D) Aspartic acid	GGA (Gly/G) Glycine
		GUG (Val/V) Valine	GCA (Ala/A) Alanine	GAA (Glu/E) Glutamic acid	GGG (Gly/G) Glycine

Selection of notable mutations, ordered in a standard table of the genetic code of amino acids.

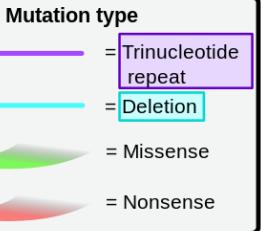
Clinically important missense mutations generally change the properties of the coded amino acid residue between being basic, acidic, polar or nonpolar, while nonsense mutations result in a stop codon.



Fragile X Syndrome

Polyglutamine (PolyQ) Diseases

- Huntington's disease
- Spinocerebellar ataxia (SCA) (most types)
- Spinobulbar muscular atrophy (Kennedy disease)
- Dentatorubral-pallidoluysian atrophy



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

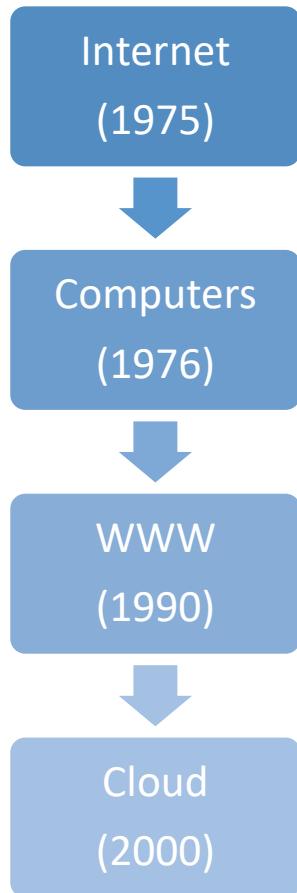
There are too many to discuss in one day.....

- 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

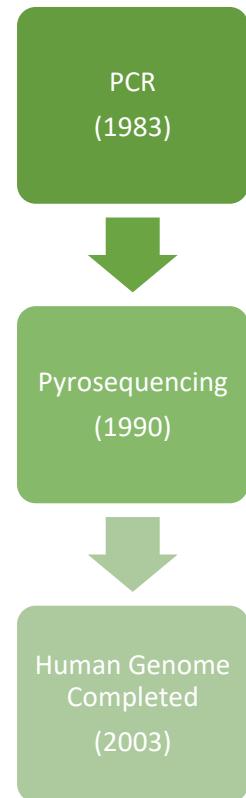
History

Computers, Information Exchange

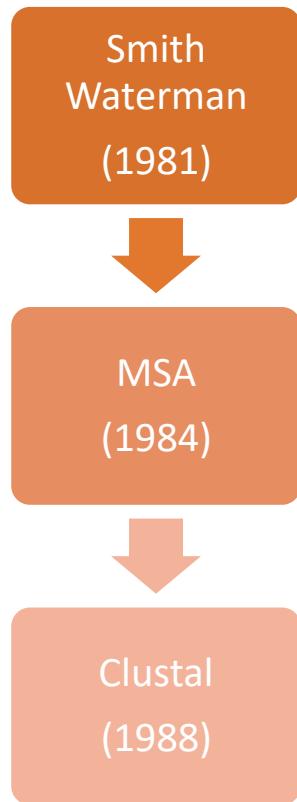
Co- evolution of technology



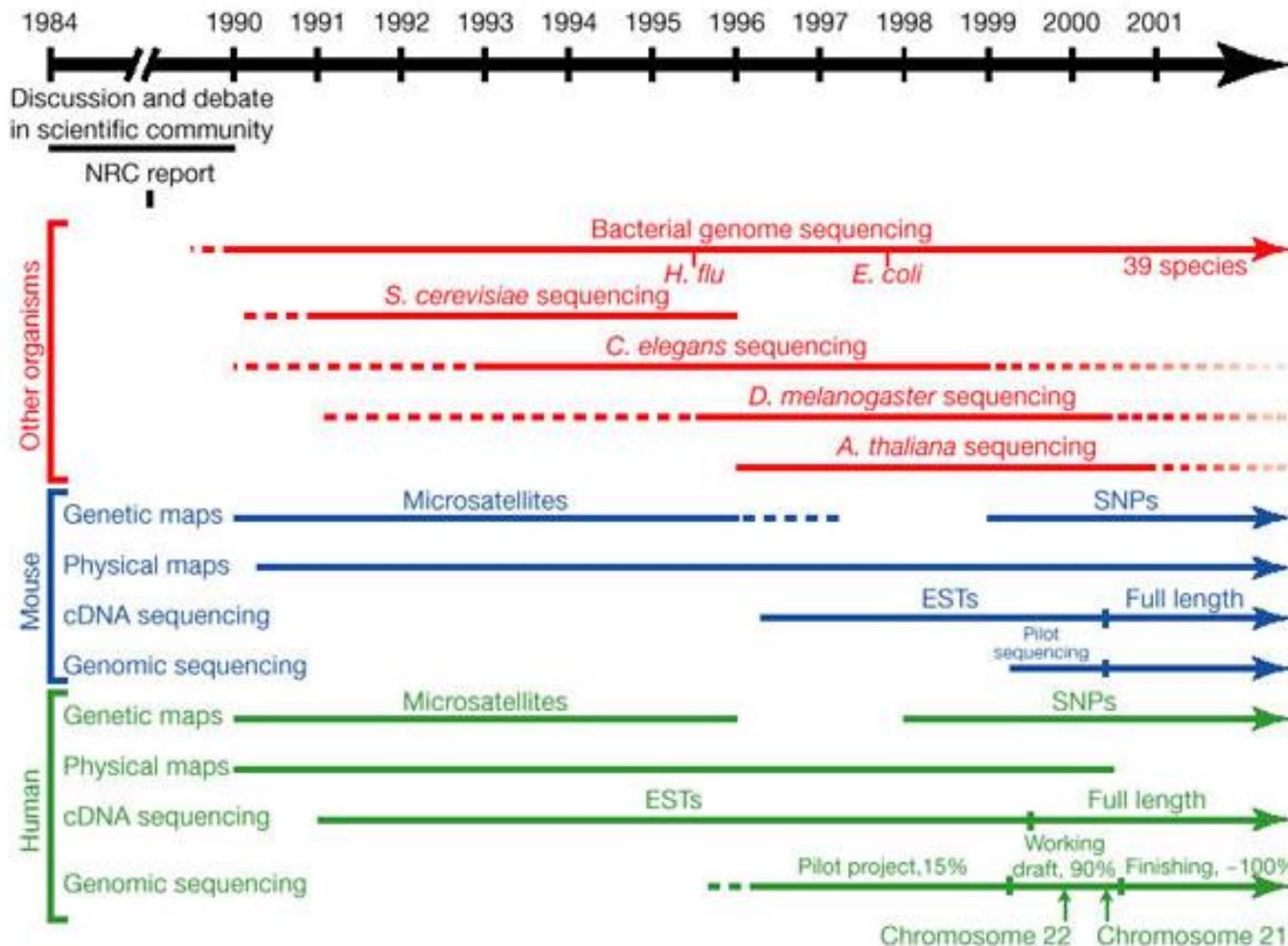
Physical Access to Genomic Information



Algorithmic Development

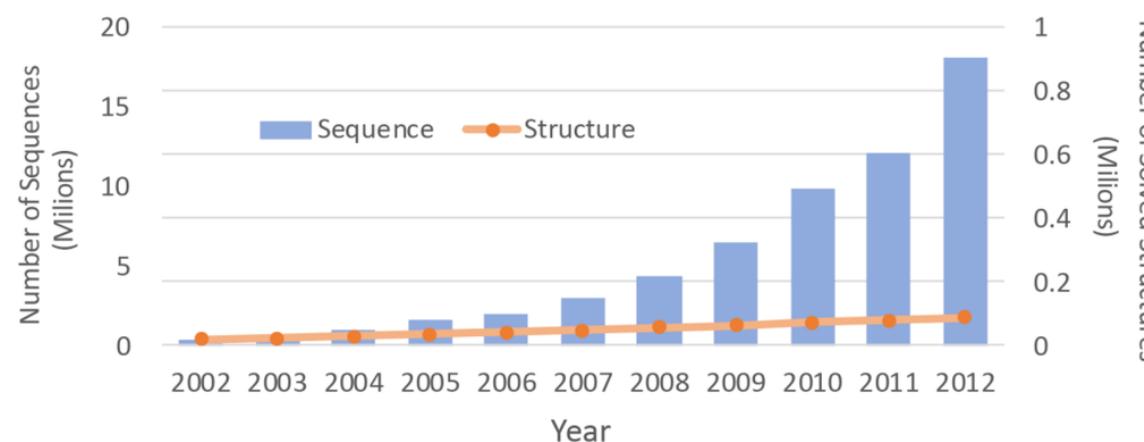
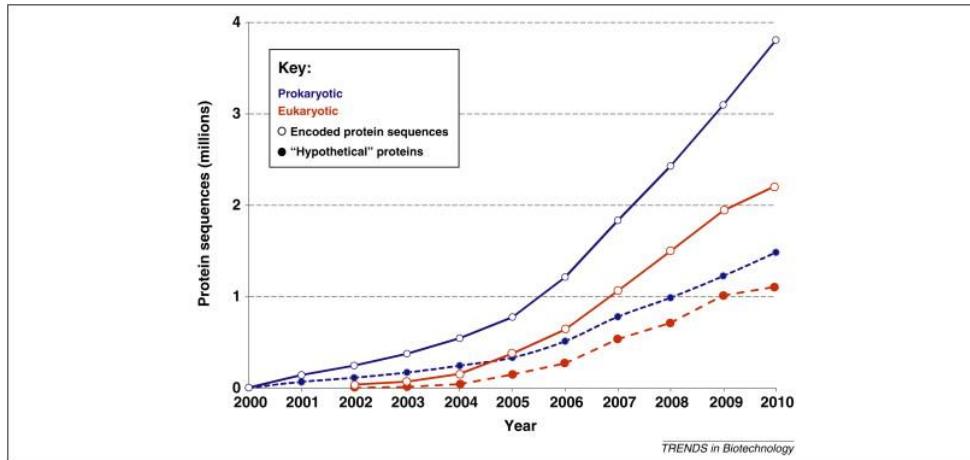


History: genomic sequencing



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

And Sequence Information is growing rapidly

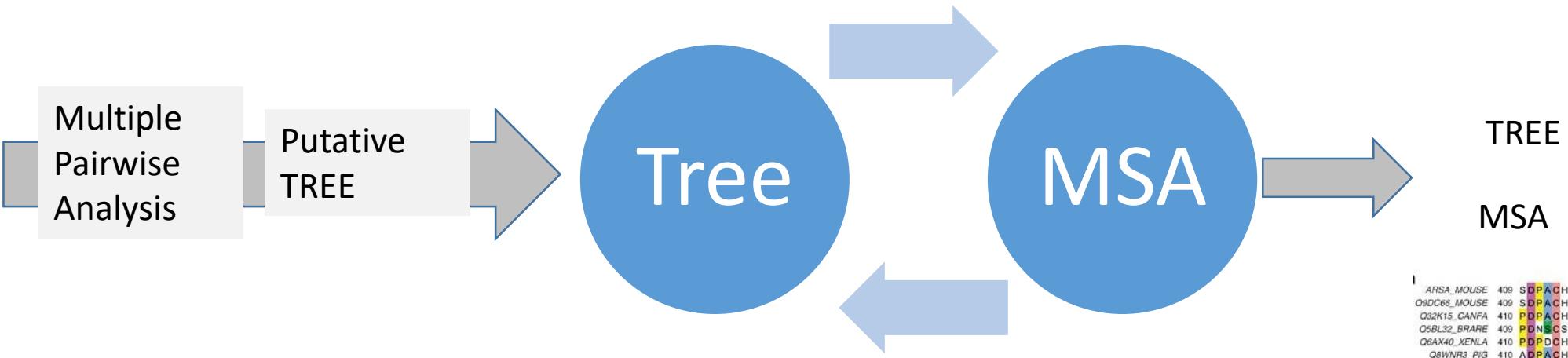


The screenshot shows the JGI Genome Portal homepage. At the top, there's a navigation bar with links to 'JGI HOME', 'GENOME PORTAL', and 'LOGIN'. Below the header is a search bar with a dropdown menu set to 'Keyword' and a 'Search' button. To the right of the search bar is a 'Show All Projects' button. The main content area features a large circular 'Tree of Life' diagram. The diagram is color-coded by domain: Archaea (light green), Eukarya (yellow-green), and Bacteria (light blue). Major phyla and genera are labeled, such as Crenarchaeota, Korarchaeota, Euryarchaeota, Firmicutes, Chlorobi, Actinobacteria, Cyanobacteria, Thermotogae, Spirochaetes, Synergistetes, Fusobacteria, Planctomyces, Dictyoglomi, Chlamydiae/Verrucomicrobia, Fibrobacteres, Thermodesulfobacteria, Deferribacteres, Alphaproteobacteria, Betaproteobacteria, Gammaproteobacteria, Deltaproteobacteria, Epsilonproteobacteria, and various amoebae and fungi. Below the tree, a yellow box highlights 'Metagenomes'. To the right of the tree, there's a section titled 'New on the portal: elviz' with a brief description and a 'Try it now with a sample project!' button. Further down, there's a 'genome Releases' section with links to 'Fungal Releases', 'Metagenomics Releases', 'Microbial Releases', and 'Plant Releases'. On the far right, there's a sidebar with a 'Learn More about Download Options' button and a 'New to Genome Portal? Explore what you can do here.' message.

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.]



	1U	2U	3U	4U	
ARSA_MOUSE	409 S D P A C H A - A N R L T A H E P P L Y D L S Q D P G E N Y N V L E S I E G V S P E A	451			
Q9DC66_MOUSE	409 S D P A C H A - A N R L T A H E P P L Y D L S Q D P G E N Y N V L E S I E G V S P E A	451			
Q32K15_CANFA	410 P D P A C H A - S S P L T A H E P P L Y D L S Q D P G E N Y N V L E S I E G V S P E A	452			
Q5BL32_BRARE	409 P D N A C S L - L A F L K Y H D P P L L F N L E T D P S E N Y N L L G G M A E V A P E A	445			
Q6AX40_XENLA	410 P D P D C H V - T A L L K S H D P P L L F D L S T D P A E N Y N L L K D G - - I P V D L	450			
Q5XFU5_MOUSE	436 G A K A C D G S V G P E Q H H V A P L I F N L E D A A D E G M P L Q K - - - G S P E Y	475			
Q3TYD4_MOUSE	436 G A K A C D G S V G P E Q H H V A P L I F N L E D A A D E G M P L Q K - - - G S P E Y	475			
ARSG_HUMAN	438 G A R A C D O S T G P E L O H K F P L I F N L E D D T A E A V P L E R - - - G G A E Y	475			
Q32K92_RAT	436 G A K A C D G G V G P E Q H H V S P L I F N L E D D A A S S P L Q K - - - G S P E Y	475			
Consensus	G D P A C H - S V G P L T A H E P P L L F + L + D D P G E N Y N L L K + + + S P E Y				

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?

Structural
Prediction

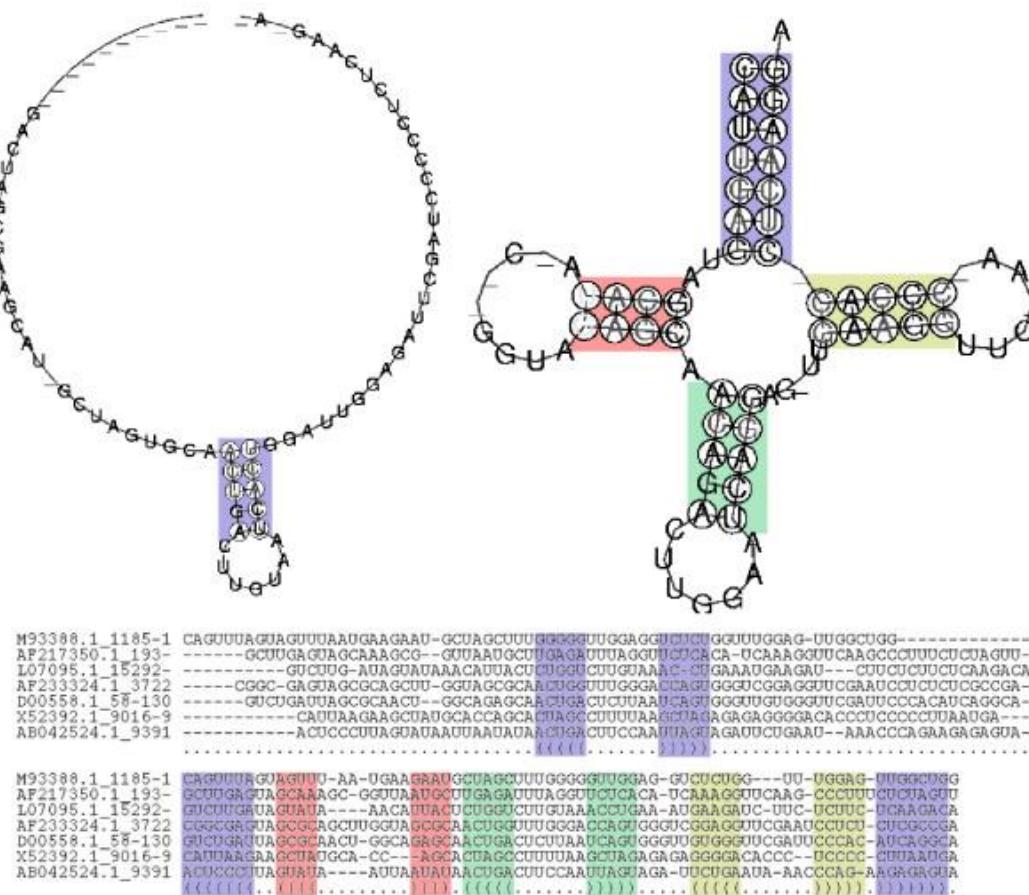
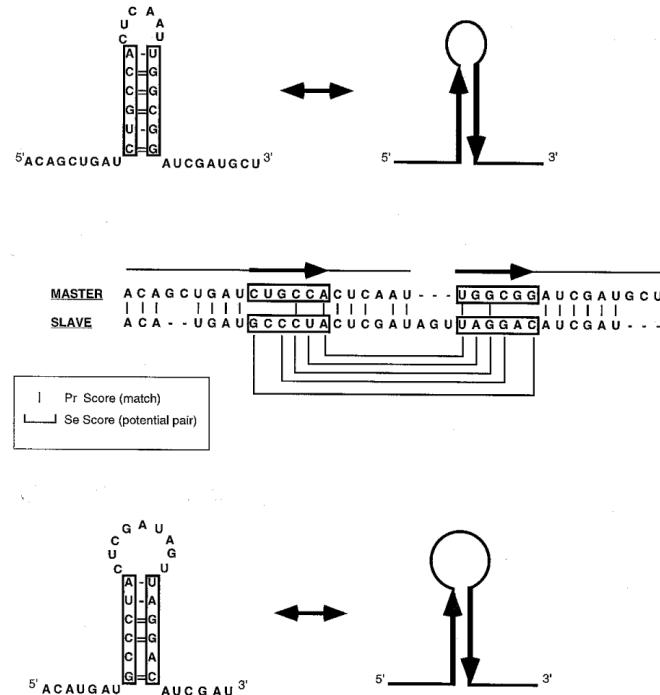
Phylogeny

Prediction
of Motifs

Functional
Prediction

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†} EMBL Outstation-The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK and²Department of Biochemistry, University College, Cork, Ireland
Received July 23, 1997; Revised and Accepted October 1, 1997

Application of MSA

PFAM

Conserved domains / protein clusters

NCBI Conserved Domain | www.ncbi.nlm.nih.gov/Structure/cdd/cdd_help.shtml#Hierarchy

cd00400 Sequence Cluster

The goal of the NCBI conserved domain curation project is to provide insights into how patterns of residue conservation and divergence in a family relate to functional properties, and to provide useful links to more detailed information that may help to understand those.

Click anywhere on the image to open the current, interactive record for the Voltage-Gated Chloride Channel domain model, cd00400, in the Conserved Domain Database (CDD). Note that the live web page may look different from the illustration shown here, because the Conserved Domain Database continues to evolve with the addition of new data; however, the concepts shown in the illustration remain stable.

What is a superfamily?

A superfamily cluster is a set of **conserved domain models** that generate overlapping annotation on the same protein sequences. These models are assumed to represent evolutionarily related domains and may be redundant with each other. A superfamily accession number begins with the prefix "cl" for "cluster". (Some superfamilies contain only a single conserved domain model (singleton), and these are not indexed in Entrez. Only superfamilies that contain two or more conserved domain models are indexed in Entrez and will therefore appear in search results.)

Clustering methodology:
Superfamily members are clustered through an automated process that involves the following steps:

Pfam 28.0 (May 2015, 16230 families)

The Pfam database is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs). [Less...](#)

Proteins are generally composed of one or more functional regions, commonly termed **domains**. Different combinations of domains give rise to the diverse range of proteins found in nature. The identification of domains that occur within proteins can therefore provide insights into their function.

Pfam also generates higher-level groupings of related families, known as **clans**. A clan is a collection of Pfam-A entries which are related by similarity of sequence, structure or profile-HMM.

QUICK LINKS

- SEQUENCE SEARCH**
- VIEW A PFAM FAMILY**
- VIEW A CLAN**
- VIEW A SEQUENCE**
- VIEW A STRUCTURE**
- KEYWORD SEARCH**
- JUMP TO**

YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...

- Analyze your protein sequence for Pfam matches
- View Pfam family annotation and alignments
- See groups of related families
- Look at the domain organisation of a protein sequence
- Find the domains on a PDB structure
- Query Pfam by keywords

enter any accession or ID

Enter any type of accession or ID to jump to the page for a Pfam family or clan, UniProt sequence, PDB structure, etc.

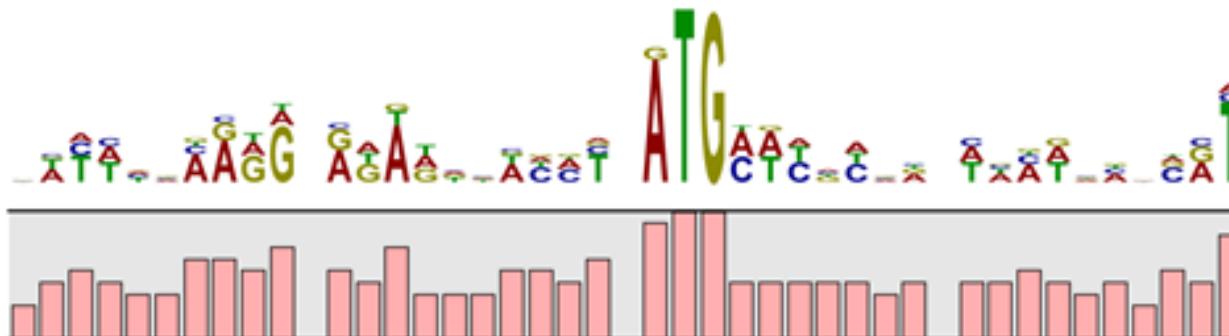
Or view the [help](#) pages for more information

Application

Prediction / conserved motifs

	-20	1	20	
talA	CTTTTCAAGG	AGTATTCCT	ATGAACGAGT	TAGACGGCAT
evgA	CATTGCAAAG	GGAATAATCT	ATGAACGCAA	TAATTATTGA
ypdI	CATTTTCAGG	ATAACTTTCT	ATGAAAGTAA	ACTTAATACT
nirB	GAAAAGAAAT	CGAGGCAAAA	ATGAGCAAAG	TCAGACTCGC
hmpA	TGCAAAAAAA	GGAAGACCAT	ATGCTTGACG	CTCAAACCAT
narQ	TTTTTGTGGA	GAAGACGCGT	GTGATTGTTA	AACGACCCGT
gltF	GTTATTAAAGG	ATATGTTCAT	ATGTTTTCA	AAAAGAACCT
intS	TACCCACCGG	ATTTTACCC	ATGCTCACCG	TTAAGCAGAT
yfdF	AATCAAAATG	GAATAAAATC	ATGCTACCAT	CTATTTCAT
dsdX	ATCACAGGGG	AAGGTGAGAT	ATGCACTCTC	AAATCTGGGT
suhB	ACATCCAGTG	AGAGAGACCG	ATGCATCCGA	TGCTGAACAT
Consensus	AATTTAAAGG	AGAATTACCT	ATGAACGCAA	TAATAAACAT

Sequence Logo



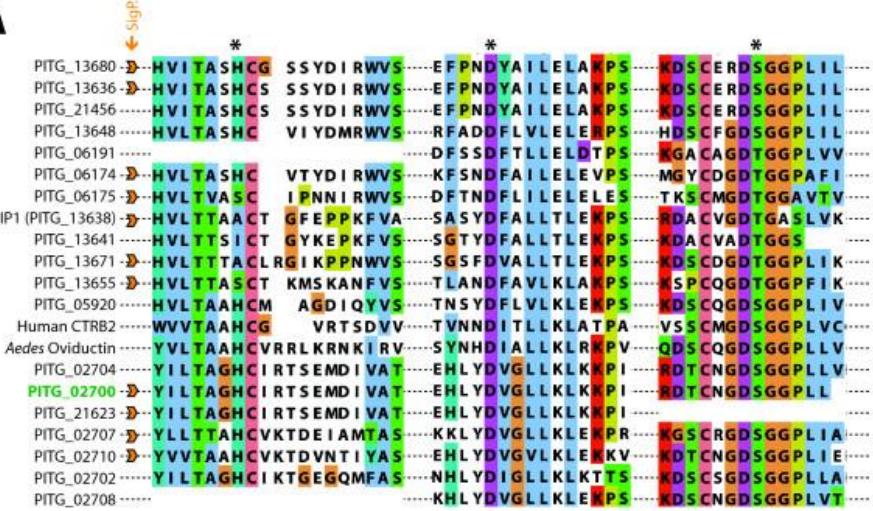
Conservation

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

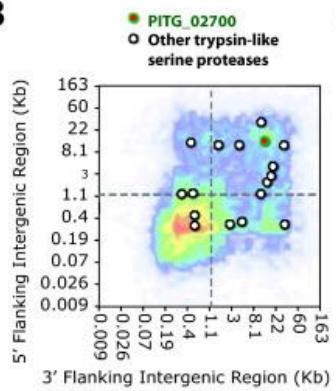
Application:

- Biochemistry: structural / functional

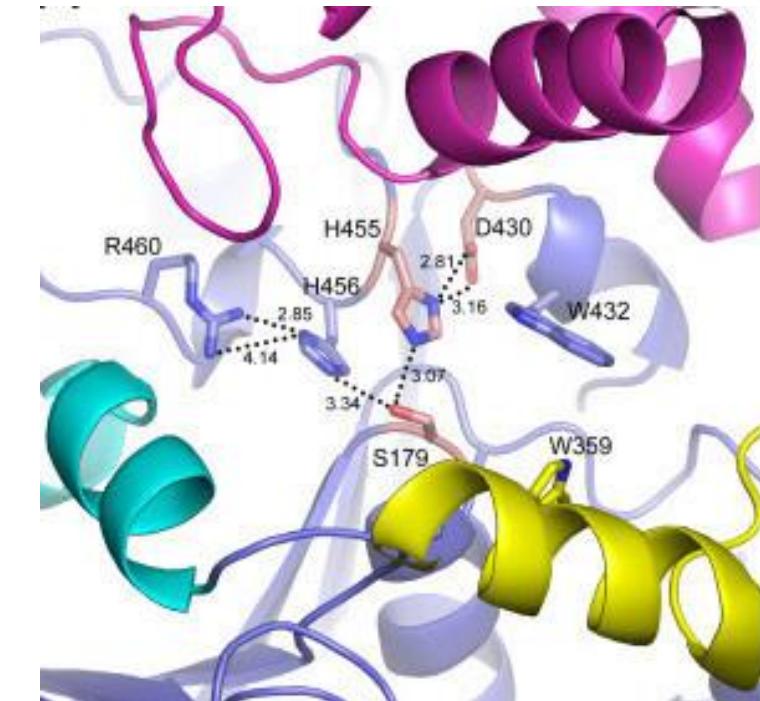
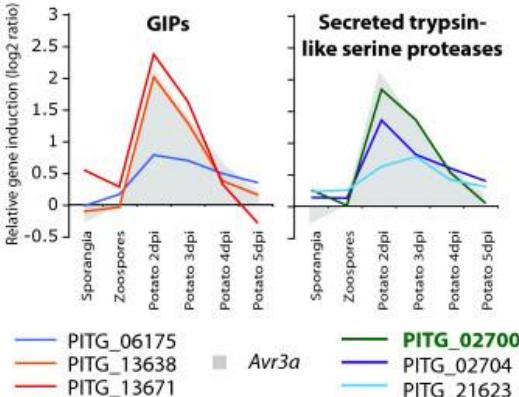
A



B



C



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

The screenshot shows the COBALT Constraint-based Multiple Protein Alignment Tool. At the top, there's a navigation bar with links for Home, Recent Results, Help, and a My NCBI sign-in option. Below the bar, the main title is "Cobalt Constraint-based Multiple Protein Alignment Tool". A prominent input field asks "Enter Query Sequences" and specifies "Enter at least 2 protein accessions, gis, or FASTA sequences". There's also a file upload section for "Or, upload FASTA file" and a "Job Title" input field. A large blue "Align" button is centered below these fields. At the bottom, there are links for Advanced parameters, Copyright, Disclaimer, Privacy, Accessibility, Contact, and Send feedback.

The screenshot shows the Google Genomics Store. The main heading is "Google Genomics is for ...". Below it, three user profiles are shown: "Bioinformaticians" (described as "Build what you want, not just what you need. Using open standards."), "Researchers" (described as "Speed up your research, ask new questions and share data in a secure, online environment."), and "IT" (described as "Rest easy knowing that you have the resources you need to meet computational demand, secure data and ensure system reliability."). At the bottom, there's a link to "Read the whitepaper for more information on how Google Genomics works".

The screenshot shows the EMBL-EBI Bioinformatics Tools for MSA page. At the top, a cookie consent notice is displayed. Below it, the EMBL-EBI logo is followed by a navigation menu with links for Services, Research, Training, and About us. The main title is "Multiple Sequence Alignment". A sub-section titled "Tools > Multiple Sequence Alignment" provides a brief introduction to MSA. It lists several tools: Clustal Omega, MUSCLE, Kalign, MAFFT, MView, and T-Coffee. Each tool entry includes a brief description, a "Launch" button, and a detailed description below it. At the bottom, there's a section for "WebPRANK" with a brief description and a "Try it out at WebPRANK" link.

Outline

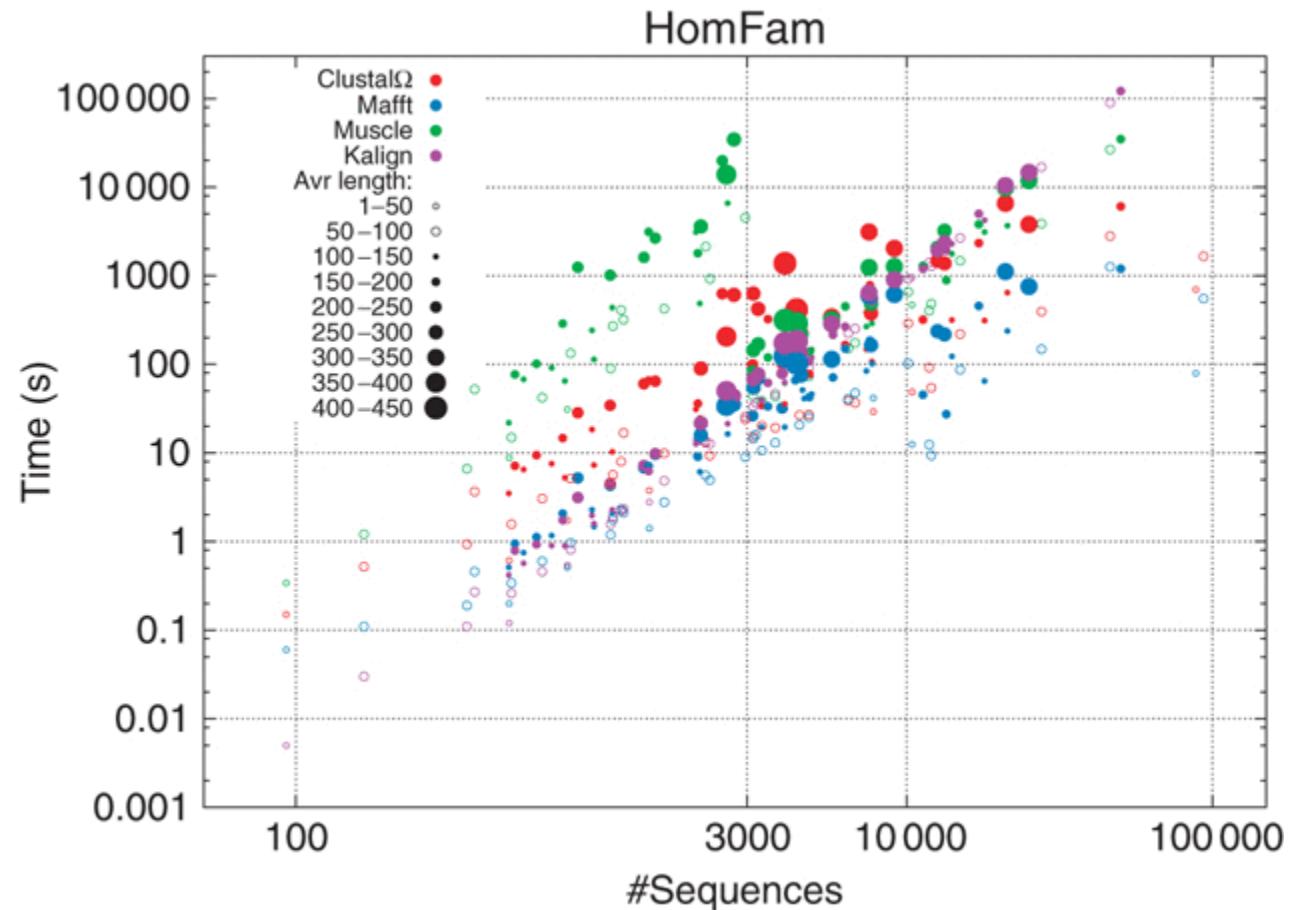
- What is and why perform Multiple Sequence Alignment (MSA)?
- Pre-requisite knowledge
- History of MSA
- Application – *post hoc* analysis – what can you do with it?
- Available Tools
- **Computational Methods**

Computational Methods

Methods

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

Efficiency / Speed / Accuracy



Clustal

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

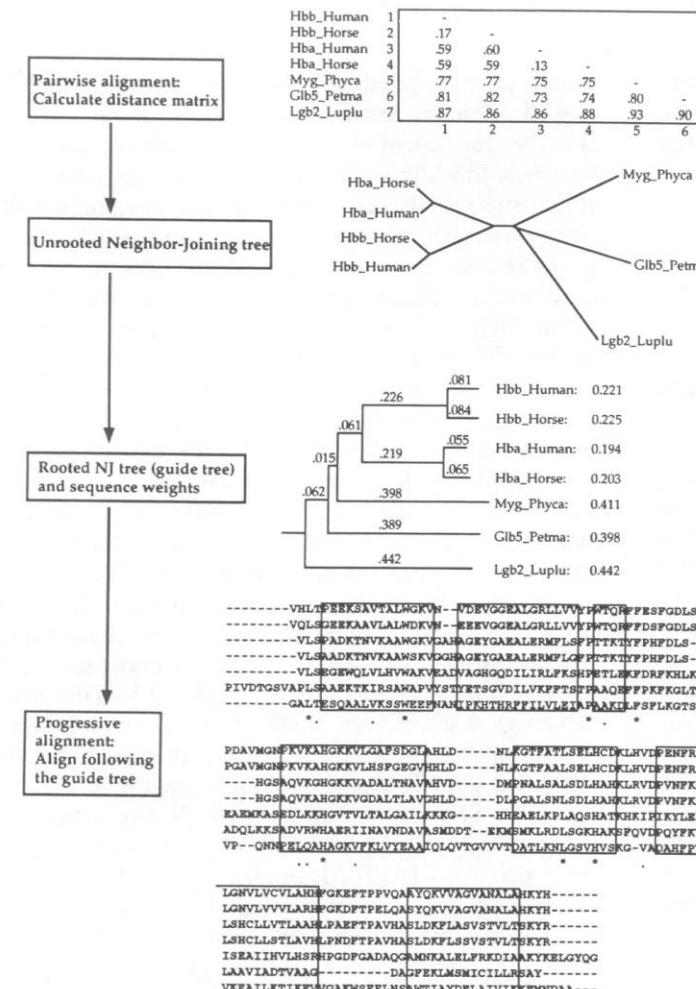
The screenshot shows a web browser window displaying the Clustal website at www.clustal.org. The page title is "Clustal: Multiple Sequence Alignment". It features the SFI logo on the left and the UCD Dublin logo on the right. Two large boxes are displayed side-by-side:

- Clustal Omega**: Features a yellow square icon with a black Greek letter Omega (Ω) and the word "CLUSTAL" below it. A bulleted list below the icon states:
 - Latest version of Clustal - fast and scalable (can align hundreds of thousands of sequences in hours), greater accuracy due to new HMM alignment engine
 - Command line/web server only (GUI public beta available soon)
- ClustalW/ClustalX**: Features a blue square icon with a black stylized 'W' and the word "CLUSTAL" below it. A bulleted list below the icon states:
 - "Classic Clustal"
 - GUI (ClustalX), command line (ClustalW), web server versions available

At the bottom of the page, a footer note reads: "Valid XHTML and Valid CSS | Viewable With Any Browser | Last modified on 08/30/2012 22:40:54".

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.

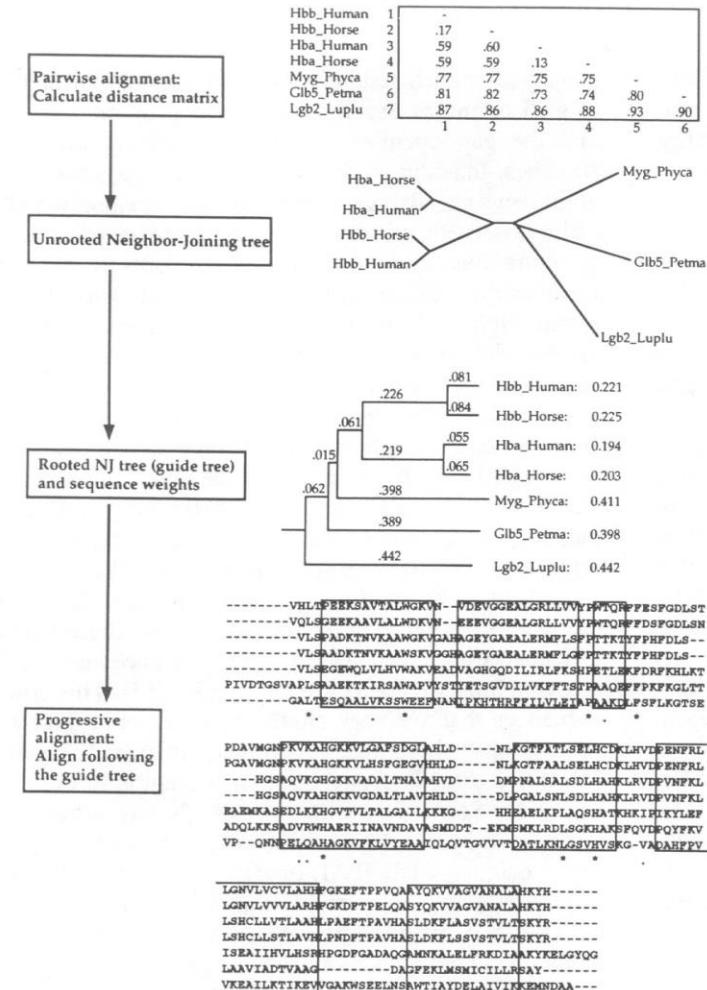


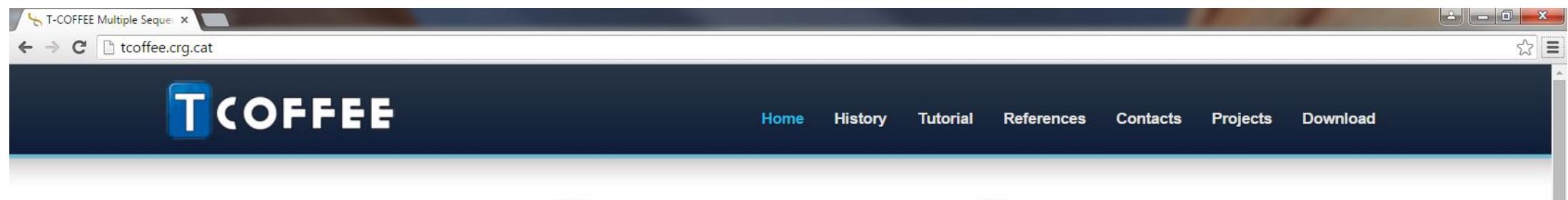
Weight matrix...
PAM / BLOSUM
Fixed
throughout the
alignment.

ClustalW

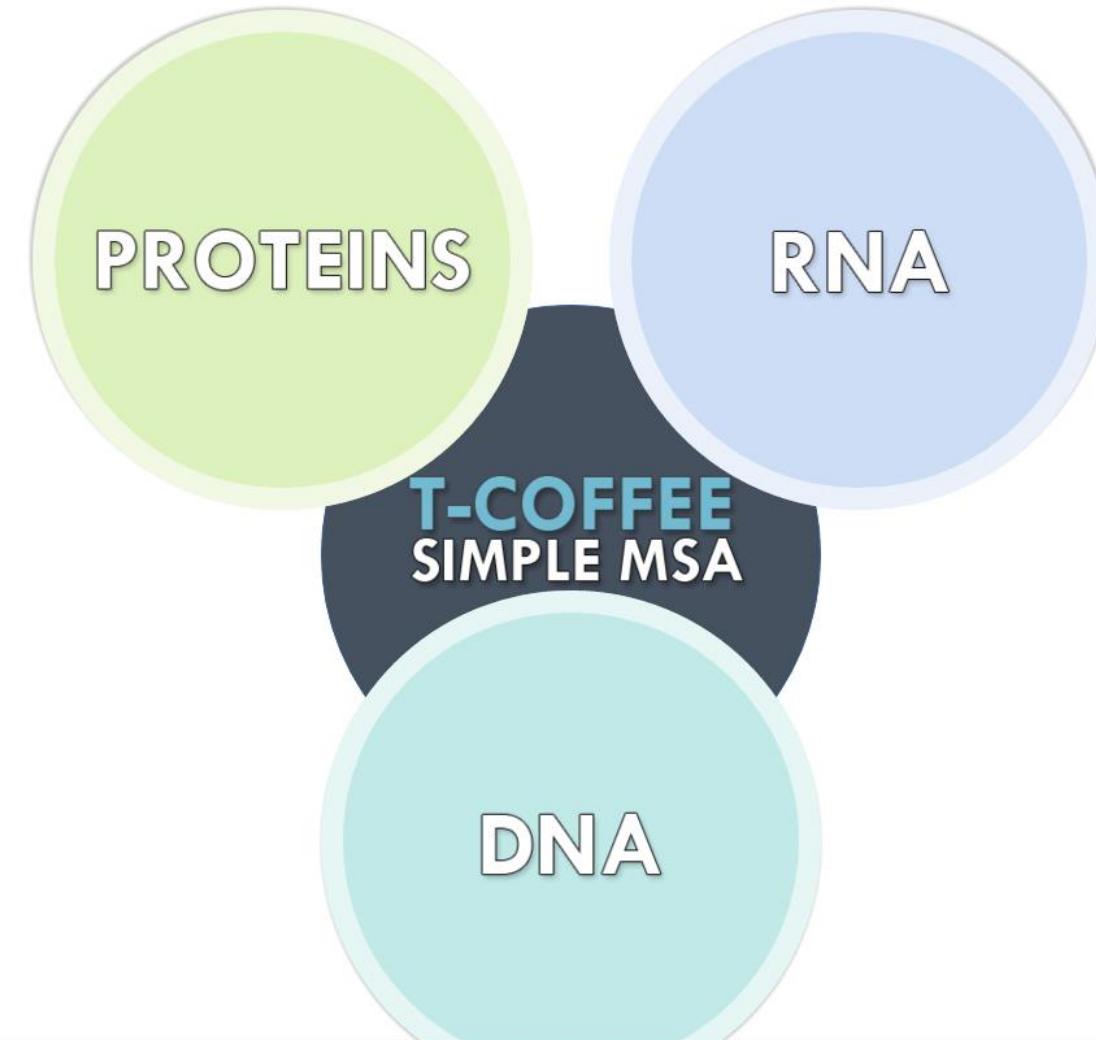
Overcomes several problems related to Progressive scoring:

- 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.





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.

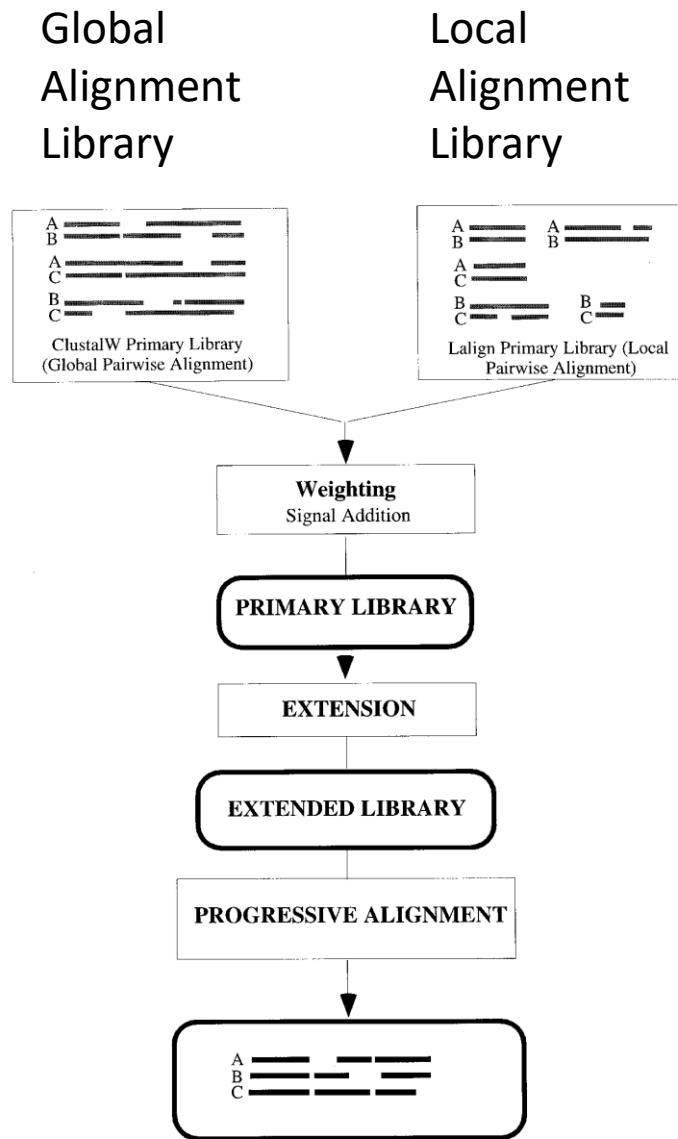


Figure 1. Layout of the T-Coffee strategy; the main steps required to compute a multiple sequence alignment using the T-Coffee method. Square blocks designate procedures while rounded blocks indicate data structures.

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.

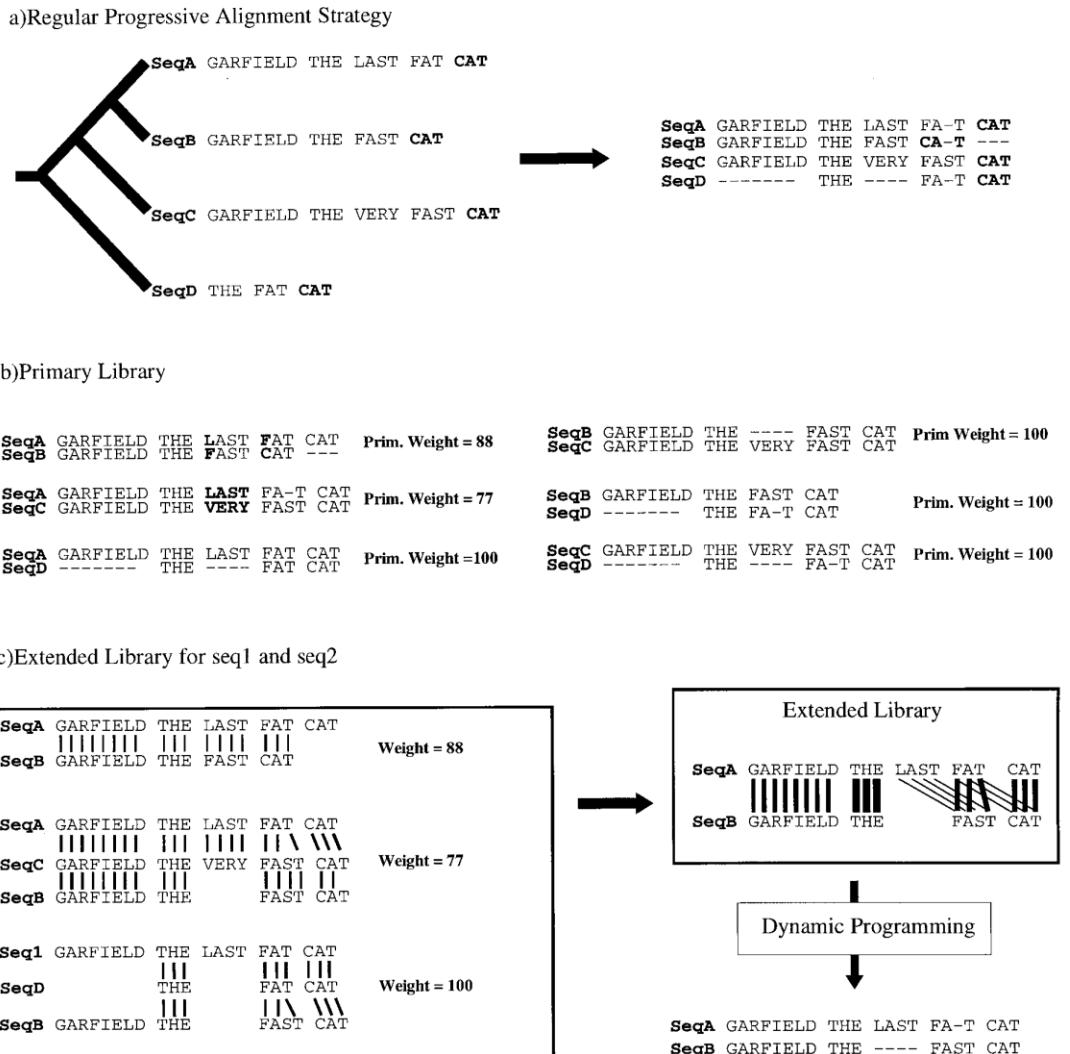
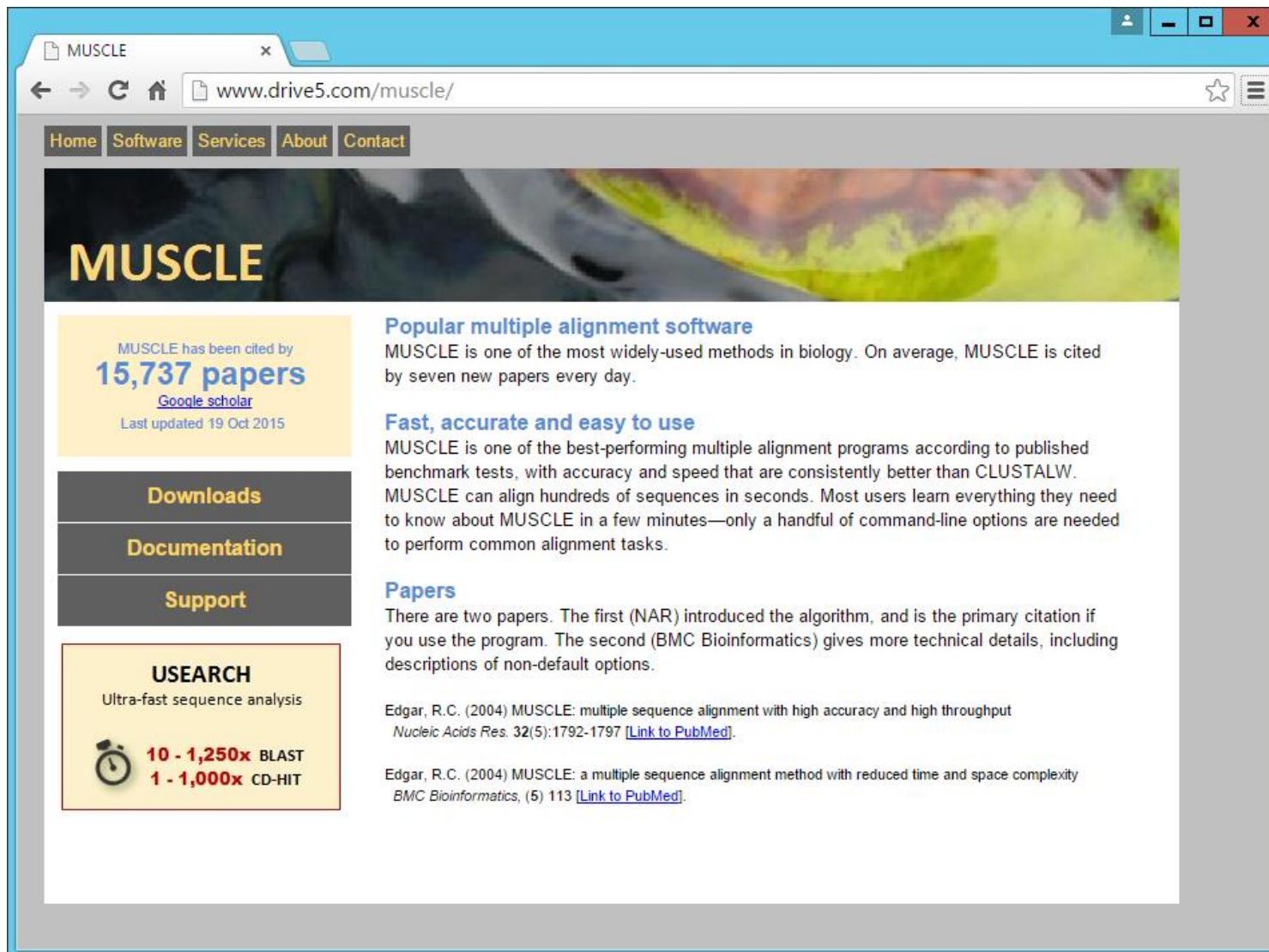


Figure 2. The library extension. (a) Progressive alignment. Four sequences have been designed. The tree indicates the order in which the sequences are aligned when using a progressive method such as ClustalW. The resulting alignment is shown, with the word CAT misaligned. (b) Primary library. Each pair 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 C, A and 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: MULTiple Sequence Comparison by Log-Expectation

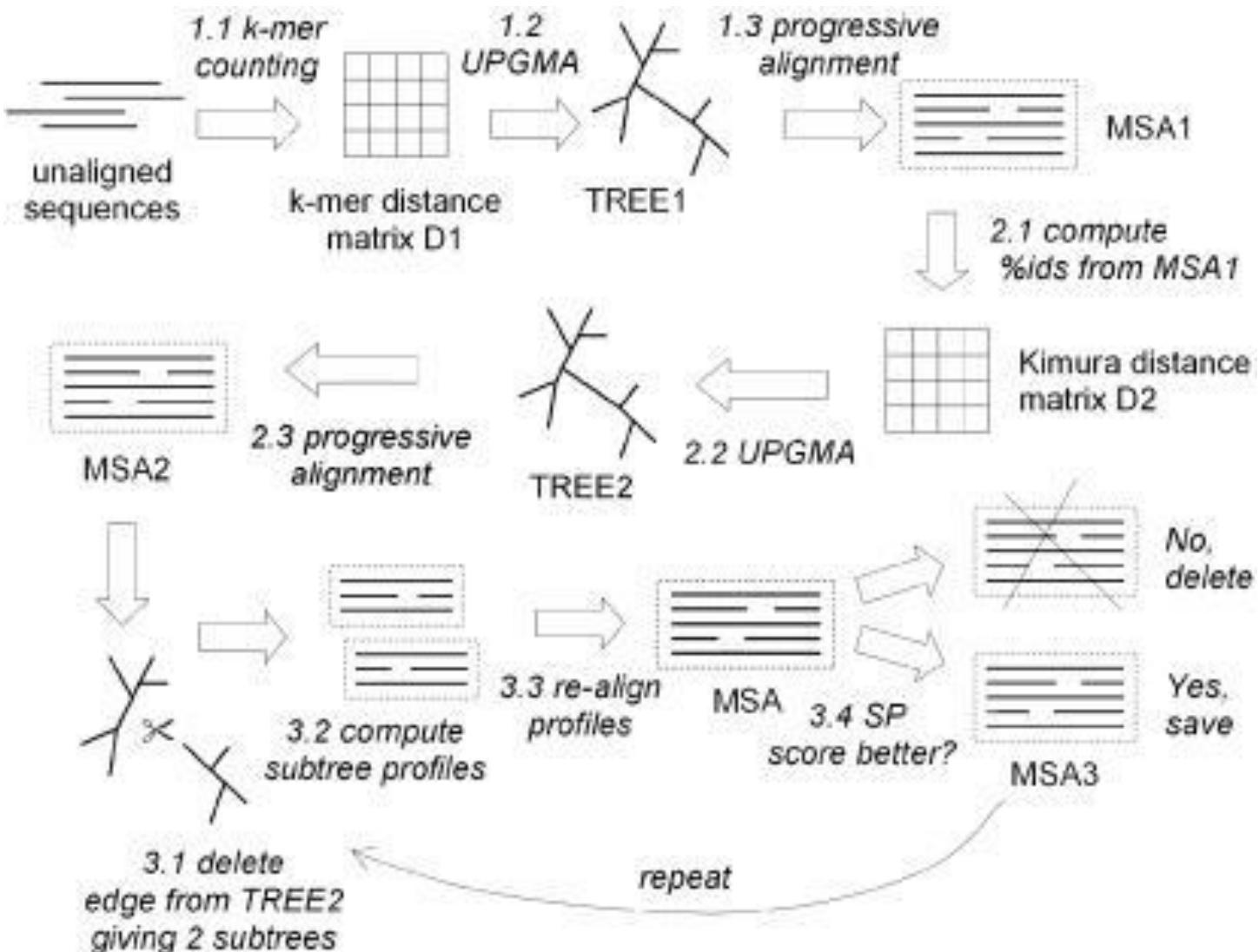


Unweighted Pair Group Method with Arithmetic Mean

MUSCLE

Yet another extension of progressive scoring with iterative progressive alignments.

K-mers are short identical sequence reads

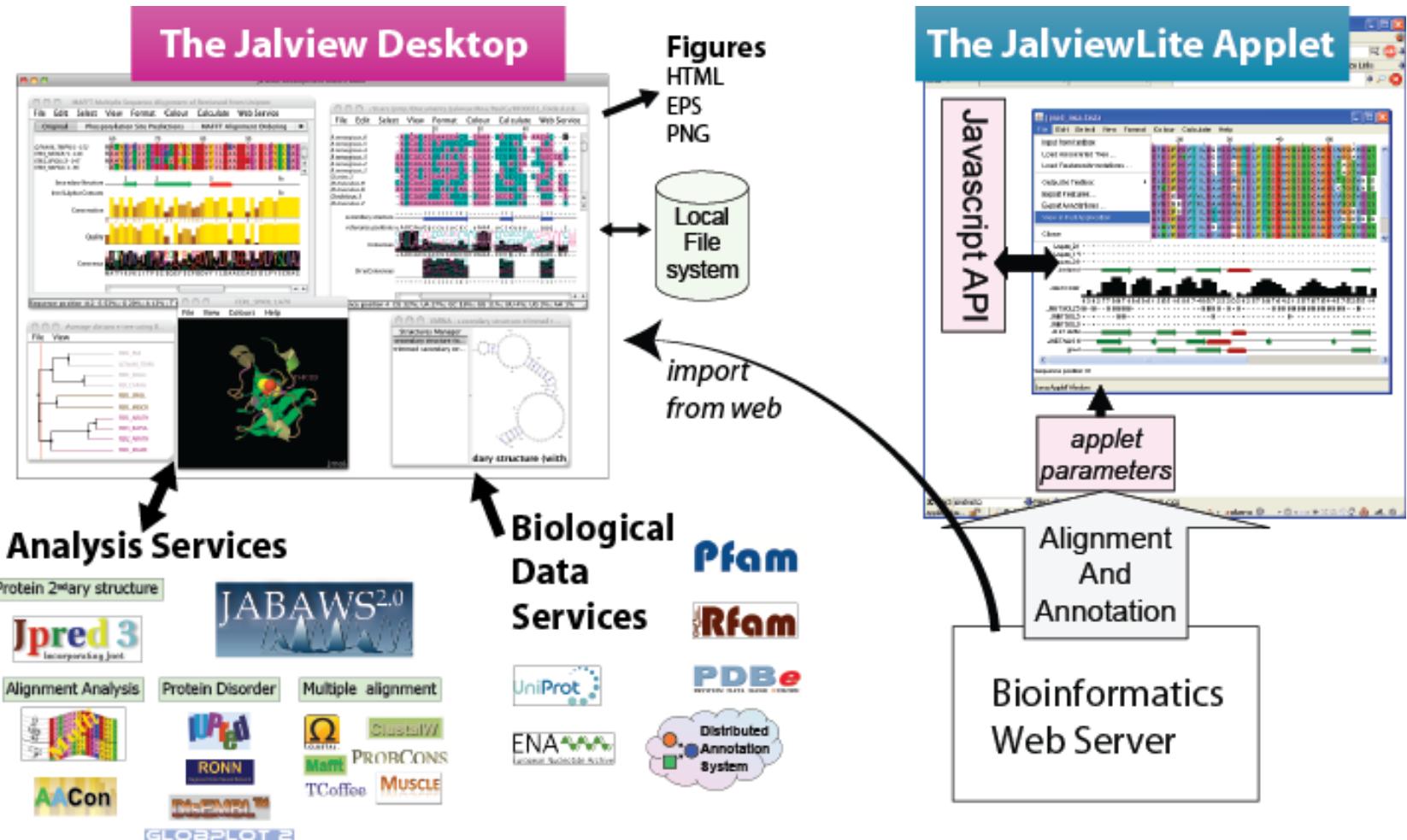


$$LE^{xy} = (1 - f^x_G)(1 - f^y_G) \log \sum_i \sum_j f^x_i f^y_j p_{ij}/p_i p_j$$

240 PAM VTML matrix

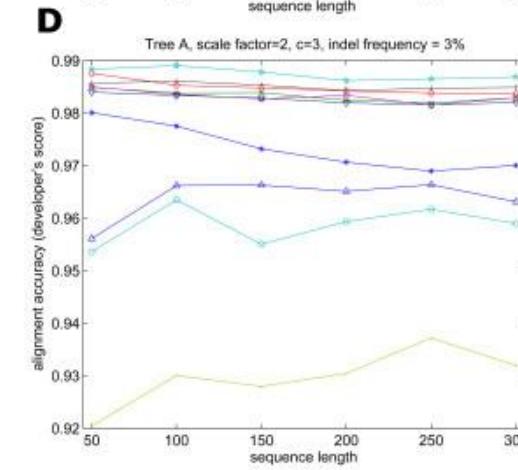
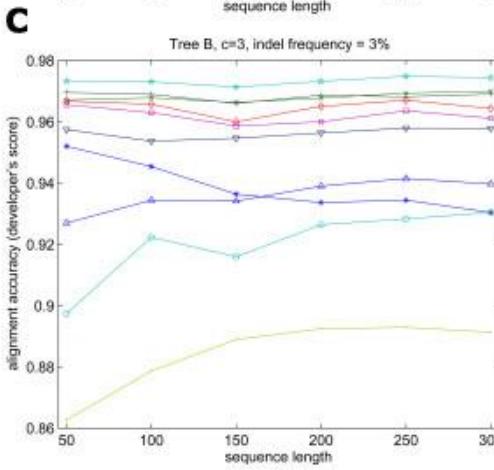
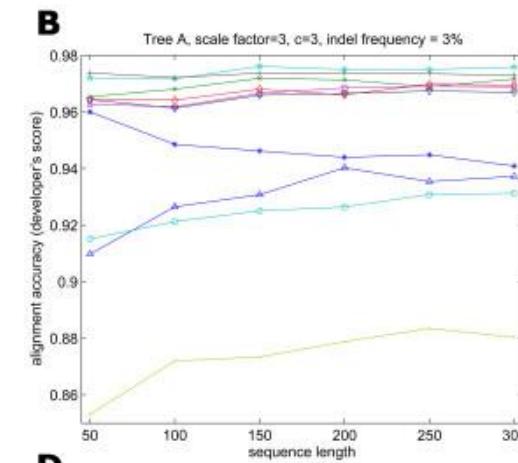
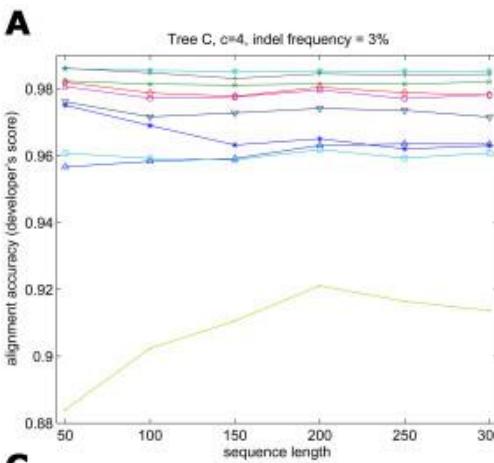
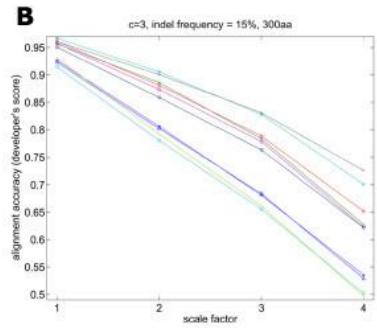
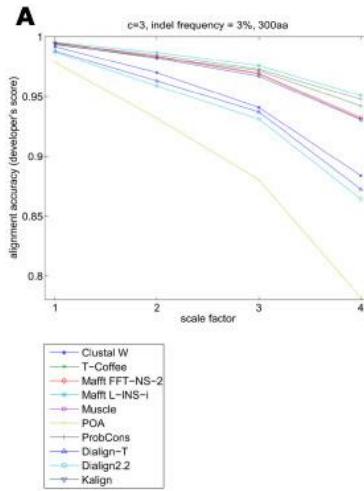
Do it all software:

- 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.

Legend:

- Clustal W
- T-Coffee
- Mafft FFT-NS-2
- Mafft L-INS-i
- Muscle
- POA
- ProbCons
- Dialign-T
- Dialign2.2
- Kalign

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