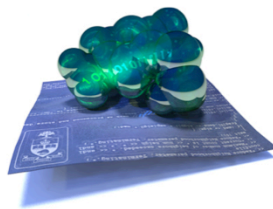


A
BIOINFORMATICS
COURSE

MULTIPLE SEQUENCE ALIGNMENT

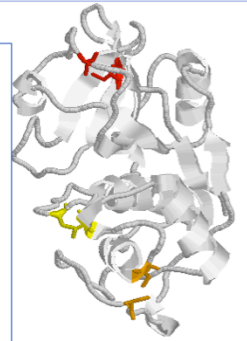


BORIS STEIPE

*DEPARTMENT OF BIOCHEMISTRY – DEPARTMENT OF MOLECULAR GENETICS
UNIVERSITY OF TORONTO*

ADDED VALUE OF MSAS

- Phylogenetic relationships
- Conservation patterns
 - Mutable regions
 - Conserved residues
 - Conserved properties
 - Conserved sequence patterns
 - Domain boundaries
 - [...]



	56	63	96						
NLVD	CYSE	..ND	GC	GGGY	... SCM	...			
ELVD	CER	...RSH	GC	XGGY	... TOR	...			
ELVE	CSTNG	.QNS	GC	NGGL	... KCD	...			
CYSE	ELVES	... ELVD	CDR	.S	YNE	GC	DGGL	... VCD	...
CYSP	HEMSP	... ELVD	CDKEE	..NQ	GC	NGGL	... TCD	...	
CATL	DROME	... NLVD	CS	T	KYGNN	GC	NGGL	... SCH	...
CATJ	RAT	... NLLD	TKSE	...GI	GL	FWGT	... PCR	...	
ALEU	HORVU	... QLVD	CAG	.GFNNF	GC	NGGL	... VCH	...	
BROM	EUM1	CPR5	CYS1						

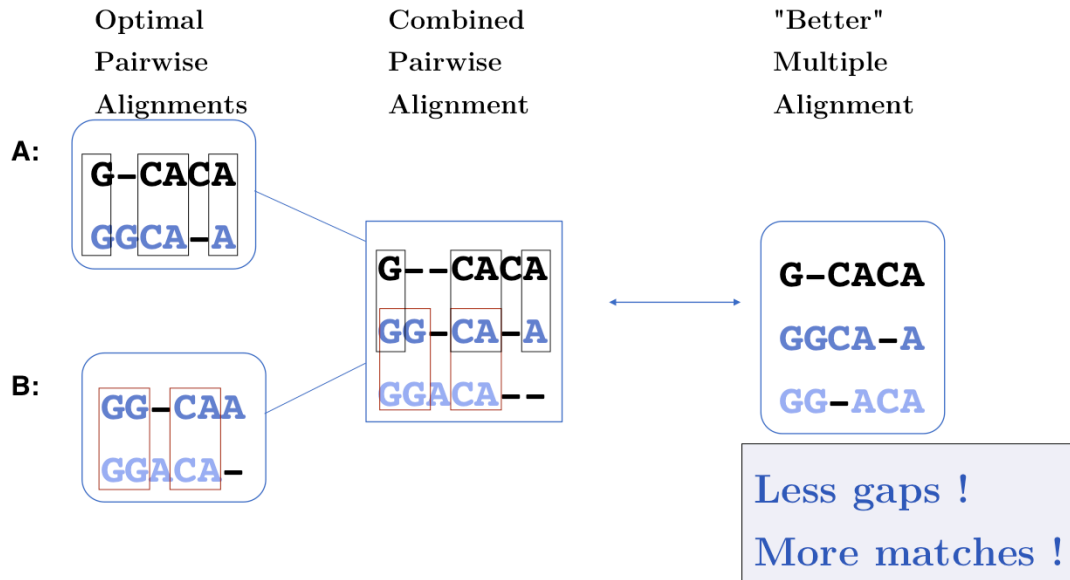
- Homology modelling
- Secondary structure prediction
- Phylogenetic reconstruction
- Sensitive homology searches
- [...]

Multiple Sequence Alignments show conservation patterns.

Multiple sequence alignments don't just match residues. They also give information on how strongly a residue is conserved, what it can be replaced with, which species share particular sequence patterns, and where in the sequence indels can be tolerated. An analysis of conservation even allows to distinguish between structurally and functionally conserved residues!

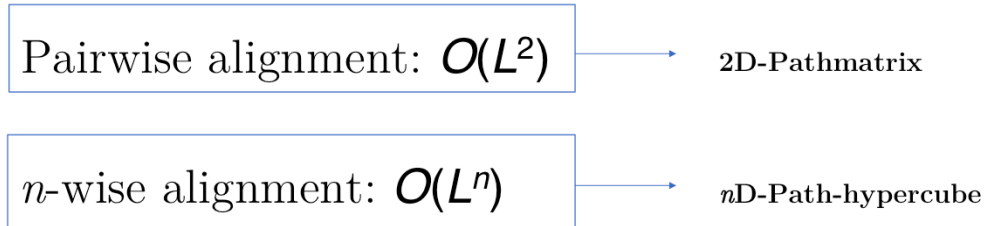
- Multiple sequence alignments are more accurate than pairwise alignments, thus they are the method of choice for starting **homology modeling** projects;
- Combined information from numerous sequences is invaluable for **secondary structure prediction, predicting domain boundaries, and sensitive sequence database searches**;
- They contain the information needed for inferences about **evolutionary relationships**, i.e. the order in which particular sequence changes occurred.

Optimal pairwise alignment \neq best multiple alignment



Multiple alignments cannot necessarily be **constructed** by merging pairwise alignments. Moreover, it may be actually be impossible to merge three mutually pairwise alignments into a non-contradicting multiple alignment. However the inverse is always possible: a multiple alignment can be **decomposed** into pairwise alignments – which is likely the most informative alignment available, but not necessarily an optimal alignment considering mutation data matrix pairscores.

Optimal multiple alignment is intractable for more than about ten sequences



E.g. 20 APSES domains: $L = 100$, $n = 20$: 10^{21} operations

Despite more than thirty years of efforts, multiple sequence alignment is still a topic of current research in bioinformatics.

Besides being intractable, it is questionable how meaningful the objective function of *optimal sequence alignments* is for *multiple alignments*. This “optimal” alignment score maximizes the score derived from a mutation data matrix, for pairs of aligned residues. But – for example – the pair score does not optimize the pattern of indel placements, or whether a particular motif is well-conserved.

Maximizing amino acid similarity in an MSA is not necessarily biological meaningful. Mutation data matrices were developed as the best representation of an **average** case – MSAs allow us to consider a **specific** protein.

"Objective function" of MSAs:
the score that MSA algorithms try to maximize.

- Many alternatives have been proposed
 - Most common: sum of scores of all pairwise alignments
 - Scores are not comparable across different alignments and across different algorithms
-

How to define an objective function that identifies the biologically most meaningful MSA?

If we want an algorithm to optimize anything at all, we first must define how we can measure the quality of the result. This metric defines the **target function** or **objective function**.

(Note that "objective" here is not used in the sense of "unbiased" but in the sense of being a "target", or "goal".)

Biologically motivated objectives for multiple alignments:

- Minimize number of indels, not length
- Minimize number of sites at which indels are tolerated
- Maximize sequence similarity
- Retain conserved motifs and patterns
- Recapitulate phylogeny
- Concentrate on alignable regions not on gapped regions
- [...]

Alignment objectives are based on the **biological models** we apply to multiple alignments - they attempt to capture constraints on sequence similarity that go beyond optimizing a pairwise alignment score (which is based on a **context-independent** mutation data matrix and on an **empirical model** for the probability of indels).

Reasonable alignment metrics are based on models of how evolution has shaped a family of related sequences.

Each of the reasonable biological objectives suggests a different alignment strategy! The most modern algorithms currently available attempt to satisfy these heuristics simultaneously. Note that these are *heuristics*, they are not the result of some rigorously applied theory, but reflect the complex relationship between protein sequence, structure, evolution and selection.

STRATEGIES

Computational strategies for multiple alignment algorithms are suggested from the diverse objective functions. The objective function should reflect biological heuristics.

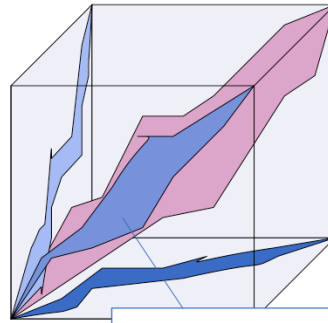
Objective	Algorithm	Type of alignment algorithm
Maximize similarity, Minimize gaps	Bounded optimal solution search	Exact
Align according to phylogeny	Align most similar first, then merge together	Progressive
Retain conserved regions	Conserved regions guide alignment	Consistency based
Maximize similarity to model	Create a model, align each sequence to that	Probabilistic
"Learn" about important regions and extend the alignment from secure seeds	Improve alignment from draft alignments	Iterated

Alignment algorithms that operate according to one or more of these principles are easily accessible online *via* the EBI:

<http://www.ebi.ac.uk/Tools/msa/>

MSA:

Multidimensional dynamic programming; search space in n-dimensional path matrix can be bounded by optimal pairwise alignments. Optimizes sum-of-pairs score which may not be the most biologically meaningful score anyway. More accurate than e.g. progressive methods but compute intensive. Practical limit ~ 10 seq. of $L = 200$.



Pairwise alignments bound the search volume in n-space!

Novel developments:

DCA (divide and conquer: split into subsequences, then recombine),
OMA (iterated improvement of splits).

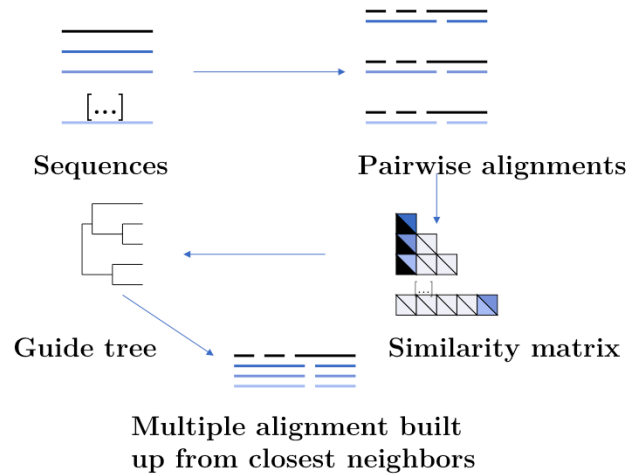
Exact methods certainly have their place where it comes to analyzing and improving algorithms; they are especially of interest to computer science because high-dimensional optimal alignment is a difficult problem. However they cannot compete in terms of result-quality with modern heuristic methods. This is not only because they really don't scale to current genome-scale questions or even modest sized protein families, but also because optimizing the score derived from a pair-score mutation data matrix plus an empirical affine gap model is not a really a very good objective for MSAs that inform about biology in the first place.

Clustal W, X:

Considered by many to be *The Standard* method.

1. Build pairwise similarity matrix
- 2. Build guide tree.**
3. join neighbors into profiles according to guide tree.
4. Align according to tree.

Key limitation: early errors persist! Best performance is for globally alignable, gap-poor sequence sets. Performance progressively worse for multidomain proteins and distant similarities.



Bottom line: Don't use CLUSTAL W. There are much better and equally convenient algorithms available.

Progressive alignment is one of the fundamental algorithmic approaches to MSA. Pure progressive alignment algorithms are only of historical interest today, since they suffer from unacceptable degradation of accuracy for sequences below ~30% ID due to the fact that early alignment errors cannot be corrected.

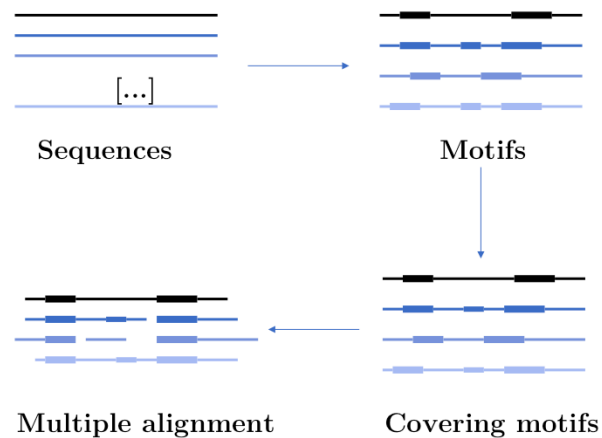
Consistency based multiple alignment: alignment based on motifs and patterns

MUSCA:

- (I) TEIRESIAS motif discovery.
- (II) Use set covering algorithm to select motifs that are common to sequence set.

MEME:

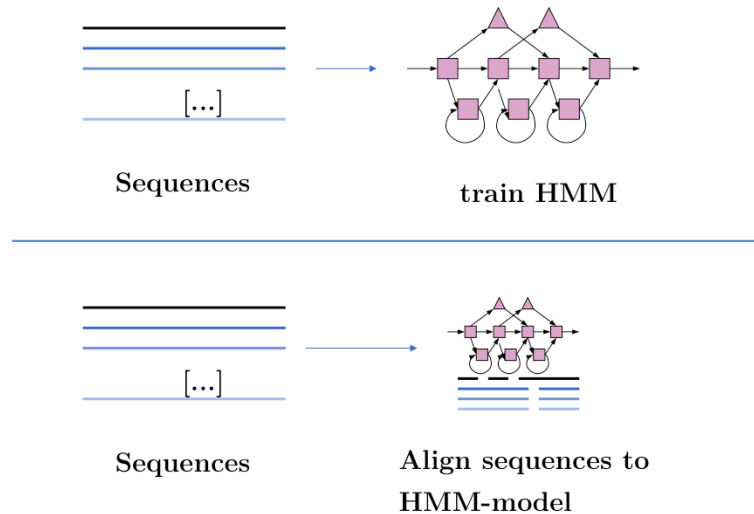
Postulate motif, compare residue composition of motif with background, choose motifs to maximize composition difference, output.



Consistency based multiple alignment is one of the fundamental algorithmic approaches to MSA. Many modern algorithms have a consistency based step included, however none of them relies solely on consistency, since problems from spurious local similarity can corrupt the alignment.

SAM:

Build HMM from
input sequences.
Align sequences to
HMM.



Probabilistic multiple alignment is one of three fundamental algorithmic approaches to MSA.

A statistical model of the sequences is built, then the alignment can be generated by aligning the sequences to the model. Of course, aligning sequences to a profile is a special case of this procedure: PSI BLAST can thus be used as an alignment algorithm. The most widely used algorithm is Sean Eddy's **HMMER**¹, a profile hidden Markov model tool, which is also used in the generation of the **Pfam** domain database².

¹ <http://hmmer.janelia.org/>

² <http://pfam.sanger.ac.uk/Pfam>

PROFILE-BASED MSA

The MSA derived from aligning sequences to a profile in a **PSI-BLAST** search is also a model based alignment.

PSI-BLAST

1. Begin with BLAST search
2. Identify significant hits
3. Align to query
4. Compile into position specific scoring matrix (PSSM or "sequence profile")
5. Repeat search with profile
6. Add new aligned hits to PSSM
7. Iterate until no new sequences can be added

Results can be displayed as an MSA.

Choose formatting option:

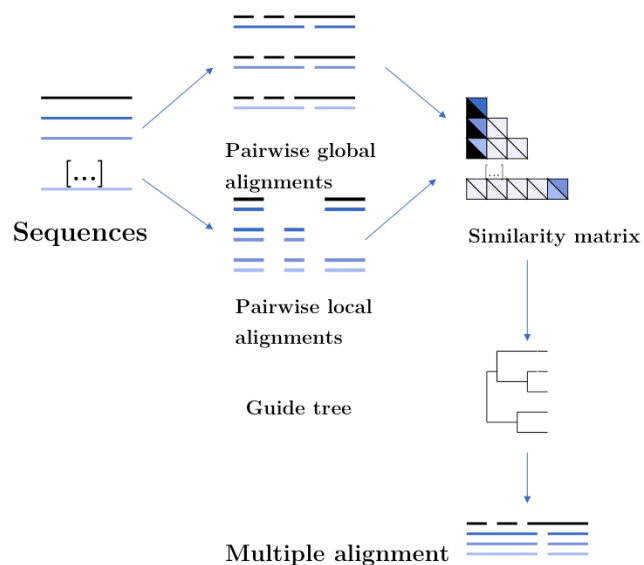
"Flat query-anchored with letters for identities"

```
NCBI Blast:publ[18M] (99 letters)
http://www.ncbi.nlm.nih.gov/BLAST/BLAST.cgi

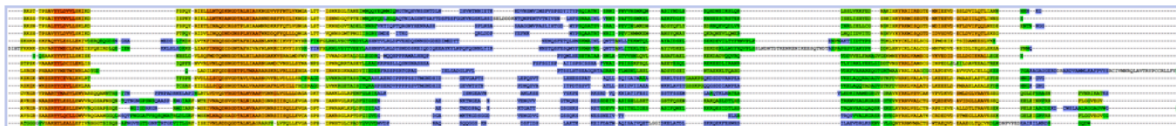
Hit #1: U00001.1 (100%)
Hit #2: U00002.1 (100%)
Hit #3: U00003.1 (100%)
Hit #4: U00004.1 (100%)
Hit #5: U00005.1 (100%)
Hit #6: U00006.1 (100%)
Hit #7: U00007.1 (100%)
Hit #8: U00008.1 (100%)
Hit #9: U00009.1 (100%)
Hit #10: U00010.1 (100%)
Hit #11: U00011.1 (100%)
Hit #12: U00012.1 (100%)
Hit #13: U00013.1 (100%)
Hit #14: U00014.1 (100%)
Hit #15: U00015.1 (100%)
Hit #16: U00016.1 (100%)
Hit #17: U00017.1 (100%)
Hit #18: U00018.1 (100%)
Hit #19: U00019.1 (100%)
Hit #20: U00020.1 (100%)
Hit #21: U00021.1 (100%)
Hit #22: U00022.1 (100%)
Hit #23: U00023.1 (100%)
Hit #24: U00024.1 (100%)
Hit #25: U00025.1 (100%)
Hit #26: U00026.1 (100%)
Hit #27: U00027.1 (100%)
Hit #28: U00028.1 (100%)
Hit #29: U00029.1 (100%)
Hit #30: U00030.1 (100%)
Hit #31: U00031.1 (100%)
Hit #32: U00032.1 (100%)
Hit #33: U00033.1 (100%)
Hit #34: U00034.1 (100%)
Hit #35: U00035.1 (100%)
Hit #36: U00036.1 (100%)
Hit #37: U00037.1 (100%)
Hit #38: U00038.1 (100%)
Hit #39: U00039.1 (100%)
Hit #40: U00040.1 (100%)
Hit #41: U00041.1 (100%)
Hit #42: U00042.1 (100%)
Hit #43: U00043.1 (100%)
Hit #44: U00044.1 (100%)
Hit #45: U00045.1 (100%)
Hit #46: U00046.1 (100%)
Hit #47: U00047.1 (100%)
Hit #48: U00048.1 (100%)
Hit #49: U00049.1 (100%)
Hit #50: U00050.1 (100%)
Hit #51: U00051.1 (100%)
Hit #52: U00052.1 (100%)
Hit #53: U00053.1 (100%)
Hit #54: U00054.1 (100%)
Hit #55: U00055.1 (100%)
Hit #56: U00056.1 (100%)
Hit #57: U00057.1 (100%)
Hit #58: U00058.1 (100%)
Hit #59: U00059.1 (100%)
Hit #60: U00060.1 (100%)
Hit #61: U00061.1 (100%)
Hit #62: U00062.1 (100%)
Hit #63: U00063.1 (100%)
Hit #64: U00064.1 (100%)
Hit #65: U00065.1 (100%)
Hit #66: U00066.1 (100%)
Hit #67: U00067.1 (100%)
Hit #68: U00068.1 (100%)
Hit #69: U00069.1 (100%)
Hit #70: U00070.1 (100%)
Hit #71: U00071.1 (100%)
Hit #72: U00072.1 (100%)
Hit #73: U00073.1 (100%)
Hit #74: U00074.1 (100%)
Hit #75: U00075.1 (100%)
Hit #76: U00076.1 (100%)
Hit #77: U00077.1 (100%)
Hit #78: U00078.1 (100%)
Hit #79: U00079.1 (100%)
Hit #80: U00080.1 (100%)
Hit #81: U00081.1 (100%)
Hit #82: U00082.1 (100%)
Hit #83: U00083.1 (100%)
Hit #84: U00084.1 (100%)
Hit #85: U00085.1 (100%)
Hit #86: U00086.1 (100%)
Hit #87: U00087.1 (100%)
Hit #88: U00088.1 (100%)
Hit #89: U00089.1 (100%)
Hit #90: U00090.1 (100%)
Hit #91: U00091.1 (100%)
Hit #92: U00092.1 (100%)
Hit #93: U00093.1 (100%)
Hit #94: U00094.1 (100%)
Hit #95: U00095.1 (100%)
Hit #96: U00096.1 (100%)
Hit #97: U00097.1 (100%)
Hit #98: U00098.1 (100%)
Hit #99: U00099.1 (100%)
Hit #100: U00100.1 (100%)
```

TCoffee - a hybrid algorithm significantly improves performance:

1. Compute **global** pairwise similarity matrix.
2. Compute top 10 non-intersecting **local** alignments.
3. Combine by looking at triplets of sequences.
4. Build guide tree.
5. Align according to tree.



Very good results.

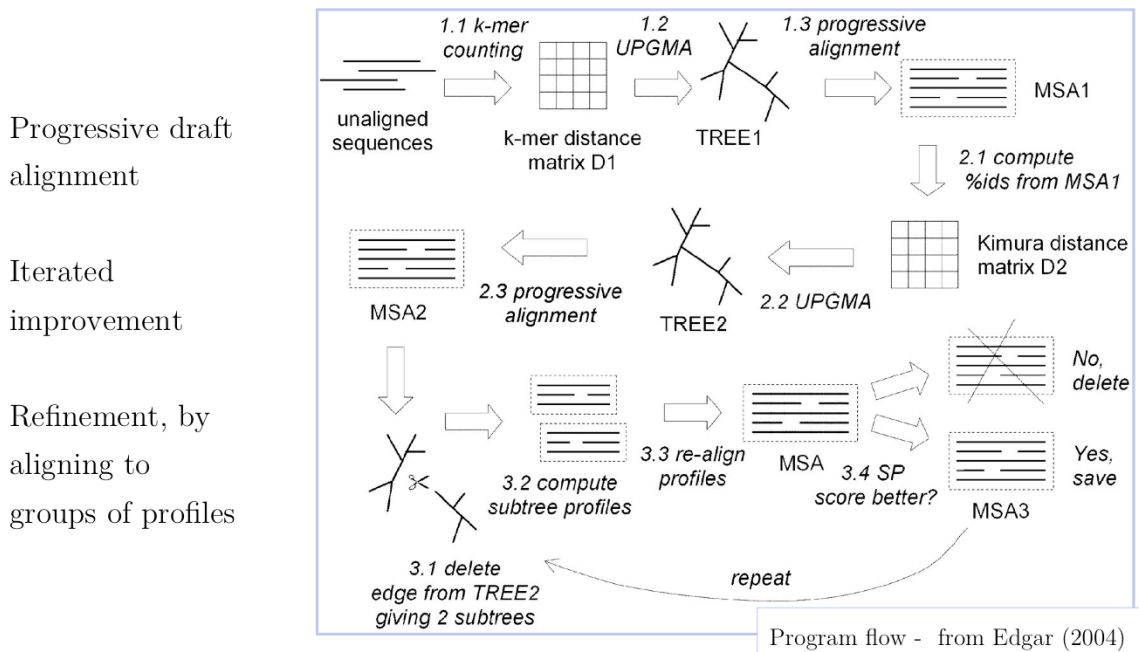


I personally rate TCOFFEE as one of the most useful and useable tools that are currently available. It is robust, fast, and gives reasonable results for many cases. Usually it is **very** noticeably better than CLUSTAL W and I would reject any result based on CLUSTAL W.

Run TCOFFEE via the EBI **TCoffee** server which is very easy to use (although alignment size is limited;). Source code can be obtained and a local installation on UNIX machines is straightforward. The TCOFFEE Web page¹ links to another Web server and also offers 3DCoffee, a variant that automatically fetches related structures and incorporates structural alignments for increased accuracy.

The inset image shows one of the useful features of TCOFFEE: an alignment output in which sequence is coloured according to the local quality of the alignment. This makes reliable and unreliable regions easy to spot, and immediately highlights outliers that could for example be due to sequence errors, such as frameshifts in exons. (MSA taken from the Mbp1 full-length protein alignment).

MUSCLE MSA



Better than CLUSTAL, faster than ProbCons - one of the general purpose algorithms of first choice with the capacity to align thousands of sequences in one run.

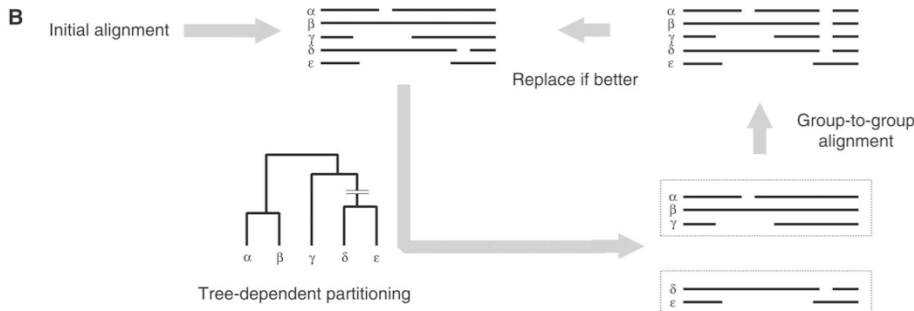
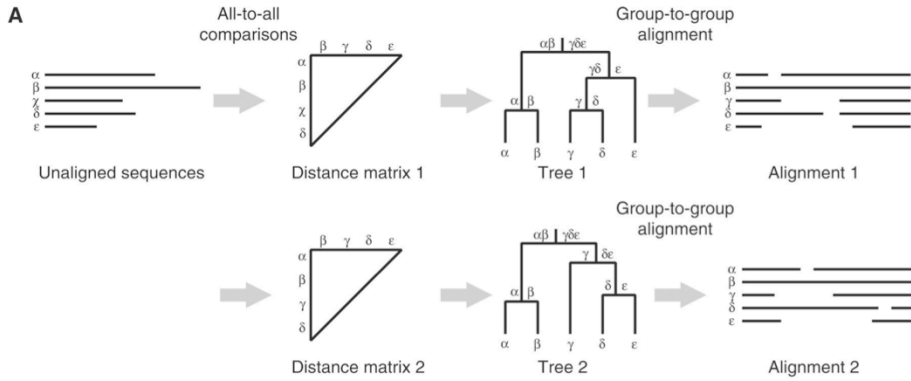
Run MUSCLE MSAs via the EBI MSA server, which is very easy to use, or via the Berkeley **MUSCLE server**¹, courtesy of Kimmen Sjolander's lab. Source code and compiled code can be obtained from the MUSCLE homepage² and a local installation on UNIX and Windows machines is straightforward. That site also hosts the PREFAB multiple alignment benchmark.

MUSCLE is one of the algorithms provided in the R package **msa**.

¹ http://phylogenomics.berkeley.edu/cgi-bin/muscle/input_muscle.py

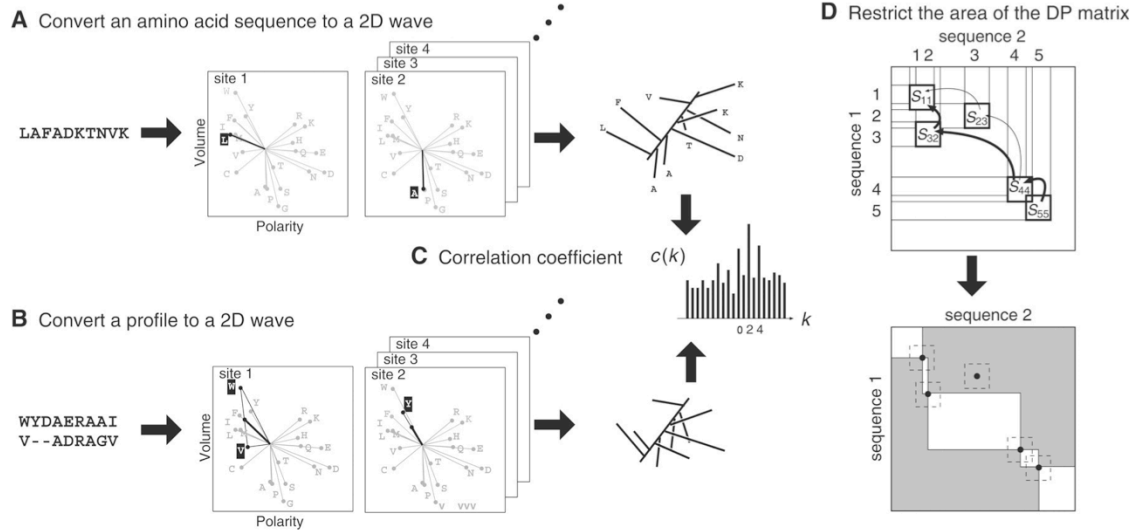
² <http://www.drive5.com/muscle/> Muscle

MAFFT MSA



Kazutaka Katoh & Hiroyuki Toh (2008)
Recent developments in the MAFFT multiple sequence alignment program
Briefings in Bioinformatics 9(4): 286-298.

MAFFT MSA



Kazutaka Katoh & Hiroyuki Toh (2008)

Recent developments in the MAFFT multiple sequence alignment program

Briefings in Bioinformatics 9(4):286-298.

As novel methods are being developed, we need to make informed choices.

Pairwise sequence alignment is a solved problem.

MSA is not.

New algorithms come online practically every month; **important** new algorithms, i.e. algorithms that provide significant benefits: about one or two each year.

How do we choose which algorithm to use? How do we discover new solutions?

CHOOSING THE “BEST”

Which algorithm is “**the best**” depends on the biological **question** you would like to answer.

Conservation patterns in highly conserved sequence are a simple task for alignment. Any algorithm will do.

To define **domain boundaries**, the correct placement of indels is the most important factor.

For **homology modeling**, it is important to maximize the number of correctly aligned pairs between template and target.

Both tasks are much more demanding, especially for more highly diverged sequences. Use the best algorithm available! Use more than one algorithm and compare the results.

What is the "Standard of Truth" ?

(Evolutionary, structural, functional criteria ...)

What is the reference alignment set ?

(Quality, errors ...)

Which alignment score to use?

(Sum-of-pair score, weighted SP, column score ...)

Lassmann T, Sonnhammer
EL. (2002)
Quality assessment of
multiple alignment programs.
FEBS Lett. **529**:126-30.

Dialign (Consistency based: best for low similarity)
T-Coffee (Consistency-based progressive: best for higher similarity)
POA (Partial Order: A dynamic Programming variant: fast)
ClustalW (Progressive: least good)

How do we know that a new algorithm is better than a previous one? Benchmarks, or Gold Standards are an essential part of scientific hygiene. We as users must demand objective comparisons to existing methods, as referees we must require them for publication, as members of the research community we must participate in defining them and provide raw data for their construction. But we must also realize that an "arms-race" of sorts may be ensuing: as developers use the benchmarks as a training set, artificially high performance scores may be generated and performance on novel problems may degrade.

MSA BENCHMARKS:
BALiBASE

Banque d'alignements multiples de référence

Structurally supported, manually curated reference alignments in several groups: **basic, orphans, sub-families, extensions, insertions, repeats, transmembrane, circular permutation**, each representative of a particular alignment difficulty.
Average sequence identity: **31.5%**

Thompson J. *et al.*, (1999) *Bioinformatics* **15**:87-88.

Name	SH3
Number of sequences	5
Alignment Length	80
Longest Sequence	80
Shortest Sequence	49
Average Percent Identity	15
Maximum Percent Identity	22
Minimum Percent Identity	6

Sequence Name	SWISSPROT	Accession
laboA		P00520
lycsB		P04637
lpht		P27986
lihvA		P00383
lvie		P12497

Family	laboA	lycsB	lpht
Family	lihvA		
Family	lvie		

laboA	1	. <u>N</u> L <u>F</u> V <u>A</u> L <u>Y</u> D <u>f</u> v <u>s</u> g <u>d</u> n <u>t</u> l <u>s</u> i <u>t</u> k <u>G</u> E <u>K</u> L <u>R</u> V <u>L</u> g <u>y</u> n <u>h</u> n.....g <u>E</u>
lycsB	1	k <u>G</u> V <u>I</u> Y <u>A</u> L <u>W</u> D <u>y</u> e <p>q<u>n</u>d<u>d</u>e<u>l</u>p<u>m</u>k<u>e</u><u>G</u>D<u>C</u>M<u>T</u>I<u>I</u>h<u>r</u>e<u>d</u>e.....d<u>e</u>i<u>E</u></p>
lpht	1	g <u>Y</u> Q <u>Y</u> R <u>A</u> L <u>Y</u> D <u>y</u> k <u>k</u> e <u>r</u> e <u>e</u> d <u>i</u> d <u>l</u> h <u>l</u> <u>G</u> D <u>I</u> L <u>T</u> V <u>N</u> k <u>g</u> s <u>l</u> v <u>a</u> l <u>g</u> f <u>s</u> d <u>g</u> <u>q</u> e <u>a</u> r <u>p</u> e <u>e</u> i <u>G</u>
lihvA	1	. <u>N</u> <u>F</u> R <u>V</u> Y <u>Y</u> R <u>D</u> s <u>r</u> d.....p <u>v</u> w <u>k</u> <u>G</u> P <u>A</u> K <u>L</u> L <u>W</u> k.....e <u>G</u>
lvie	1	. <u>d</u> r <u>v</u> r <u>k</u> k <u>s</u> g <u>a</u>a <u>w</u> q <u>G</u> O <u>I</u> V <u>G</u> W <u>Y</u> ctnlt.....p <u>e</u> G

laboA	36	<u>W</u> C <u>E</u> A <u>Q</u> t.. <u>k</u> n <u>g</u> q <u>G</u> W <u>V</u> P <u>S</u> N <u>Y</u> I <u>T</u> P <u>V</u> N.....
lycsB	39	<u>W</u> W <u>W</u> A <u>R</u> l.. <u>n</u> d <u>k</u> e <u>G</u> Y <u>V</u> P <u>R</u> N <u>L</u> L <u>G</u> L <u>Y</u> P.....
lpht	51	<u>W</u> L <u>N</u> G <u>Y</u> n <u>e</u> t <u>t</u> g <u>e</u> r <u>G</u> D <u>F</u> P <u>G</u> T <u>Y</u> V <u>E</u> Y <u>I</u> G <u>r</u> k <u>k</u> i <u>s</u> p
lihvA	27	<u>A</u> V <u>V</u> I <u>Q</u> d.. <u>n</u> s <u>d</u> i <u>K</u> V <u>V</u> P <u>R</u> R <u>K</u> A <u>K</u> I <u>I</u> Rd.....
lvie	28	<u>Y</u> A <u>V</u> E <u>S</u> e <u>a</u> h <u>p</u> g <u>s</u> v <u>Q</u> I <u>Y</u> P <u>V</u> A <u>A</u> L <u>E</u> R <u>I</u> N.....

alpha helix **RED**
beta strand **BLUE**
core blocks **UNDERSCORE**

Only "core blocks" are regions in which the superposition suggests one-to-one correspondence of residues

Central to BALiBASE is the concept of "'core blocks'" of alignable regions in which a pairwise correspondence of residues can be defined; outside these regions an alignment is not possible since the structural differences are too large.

As important as such benchmarks are, BALiBASE is not without criticism. Developers have remarked on the lack of structurally founded standard of truth (see.: http://www.drive5.com/muscle/manual/balibase_problems.html).

Sequence Alignment Benchmark

Single domains, based on SCOP classification of entire fold space. Reference constructed from consensus 3D-superpositions of two program.

Contains:

"Superfamily": 23% average ID: HARD DATASET and

"Twilight Zone" (BLAST e-value > 1.0) VERY HARD DATASET.

In addition, decoys are available: similar sequences with a BLAST e-value greater than at least one of the family members, but having a different 3D-fold, i.e. the structures cannot be superimposed.

Van Walle *et al.* (2005) *Bioinformatics* **21**:1267-1268

SABmark may be most representative of real-world alignment tasks.

PREFAB implements a fully automated procedure!

- 1: Pairs of protein structures are automatically superimposed with a procedure that does not use sequence similarity.
- 2: corresponding residues are defined as being “correct”.
- 3: proteins with similar sequences to either of the two are collected from the sequence database, using PSI-BLAST. <24 each randomly selected to keep < 50 sequences in the set.
- 4: To score alignment accuracy, a MSA is performed for the complete set, then the score is calculated only for the known, correctly aligned residues of the first two proteins.

Average sequence identity in PREFAB 21%: HARD DATASET

PREFAB was built by the author of MUSCLE.

Table 1. Evaluation of alignment methods on SABmark and PREFAB benchmarks

Method	SABmark - twi(209/7.7)	SABmark - sup(425/8.3)	PREFAB (1682/45.2)
PROMALS	0.391	0.665	0.790
SPEM	0.326	0.628	0.774
MUMMALS	0.196	0.522	0.731
ProbCons	0.166	0.485	0.716
MAFFT-linsi	0.184	0.510	0.722
MUSCLE	0.136	0.433	0.680
ClustalW	0.127	0.390	0.617

Average Q-scores of two SABmark data sets ('twi' for 'twilight zone' set, 'sup' for 'superfamily' set) and the PREFAB 4.0 data set are shown. Q-score is the number of correctly aligned residue pairs in the test alignment divided by the total number of aligned residue pairs in the reference alignment. For each data set, the two numbers in the parentheses separated by a slash are the number of alignments tested and the average number of sequences per alignment, respectively. [...] PROMALS and SPEM use secondary structure prediction and database homologs in alignment process, while the other five methods only utilize the input sequences.

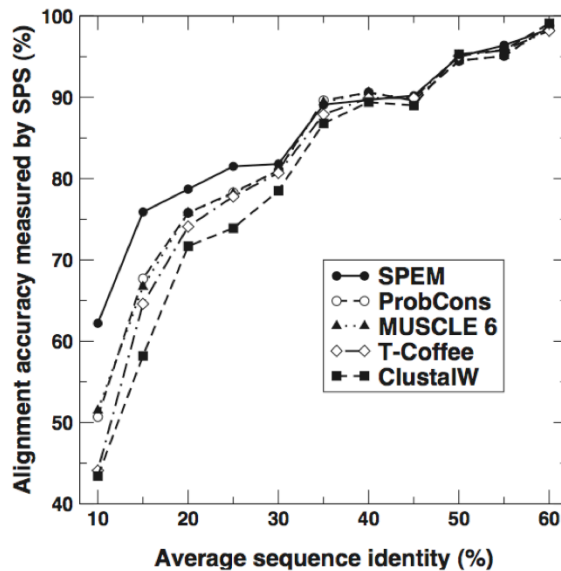
... as performed by the PROMALS authors.

One of the best currently available MSA algorithms is PROMALS. But what does this mean, relative to e.g. CLUSTAL?

For one, we can see a clear leap in performance through the inclusion of database information and consensus structure predictions (SPEM and PROMALS).

On the other hand, regarding the SABmark superfamily dataset (which is probably the benchmark that is most characteristic of "typical" alignment problems, containing alignments with recognizable, but low identity), PROMALS achieves a **50% improvement relative to CLUSTAL**, a **30% improvement relative to MUSCLE and ProbCons**. This is much more than just statistical noise.

<http://prodata.swmed.edu/promals/promals.php>



*... as performed
by the SPEM
authors.*

Fig. 3. Alignment accuracies (measured by SPS) as a function of average sequence identity given by methods SPEM, ProbCons, MUSCLE 6.0, T-Coffee and ClustalW, shown as labeled. Each point is represented by the lower bound of sequence identity at each bin.

... from the SPEM paper (Zhou & Zhou, 2005). Above $\sim 35\%$ pairwise sequence identity, all algorithms get it more or less right. Below $\sim 20\%$ pairwise sequence identity the differences are dramatic with the methods that rely on the sequences only scoring more than 20% better than CLUSTAL and SPEM outperforming CLUSTAL by about 40%.

FINDING THE “BEST” METHOD

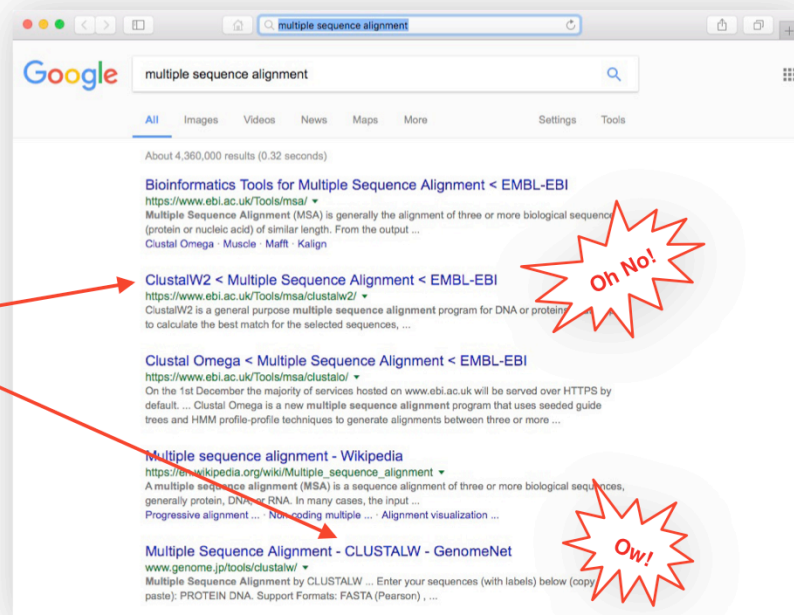
MSA is an unsolved problem in bioinformatics, and new approaches are being published every year. From the many available algorithms, how do we choose the “best”?

Can you Google for the best algorithm?
The most widely referenced algorithm - CLUSTAL - is by a wide margin the “least best” and this has been known for decades now!

(Search results: October 2017)

The question is:

How do we keep abreast of new developments?



Using CLUSTAL for anything but the simplest alignment problems is *Cargo Cult Bioinformatics*. You are doing something that may look good to the non-expert, but you can't get good results. Benchmark results have identified **significant** progress in the field! In fact, CLUSTAL has been retired from the (excellent) suite of MSA methods offered by the EBI.

"Relevance" for Google may not be the same as relevance for your work. For some applications, novelty is more important than cross-references and page-hits. For a more curated view, you can try the **Wikipedia** page on Multiple Sequence Alignment¹.

¹ http://en.wikipedia.org/wiki/Multiple_sequence_alignment

Identifying new developments via PubMed: an expert review on MSA methods

Method	Score	Templates	Validation Values	
			PreFab	HOMSTRAD
ClustalW [14]	Matrix	—	61.80 [12]	—
Kalign	Matrix	—	63.00 [18]	—
MUSCLE [6]	Matrix	—	68.00 [16]	45.0 [9]
T-Coffee [10]	Consistency	—	69.97 [12]	44.0 [9]
ProbCons [7]	Consistency	—	70.54 [12]	—
MAFFT [8]	Consistency	—	72.20 [12]	—
M-Coffee [12]	Consistency	—	72.91 [12]	—
MUMMALS [16]	Consistency	—	73.10 [16]	—
DbClustal [24]	Profiles	—	—	—
PRALINE [9]	Matrix	Profiles	—	50.2 [9]
PROMALS [16]	Consistency	Profiles	79.00 [16]	—
SPEM [28]	Matrix	Profiles	77.00 [28]	—
Expresso [13]	Consistency	Structures	—	71.9 [11] ^a
T-Lara [29]	Consistency	Structures	—	—

Cédric Notredame (2007) Recent evolutions of multiple sequence alignment algorithms. PLoS Comp Biol

The obvious first approach is to search for a recent review. For recent sequence alignment literature in PubMed search:

```
("multiple protein sequence alignment"[ti] OR "multiple sequence alignment"[ti] OR "multiple alignment"[ti]) AND (server OR algorithm) AND "last 2 years"[dp]
```

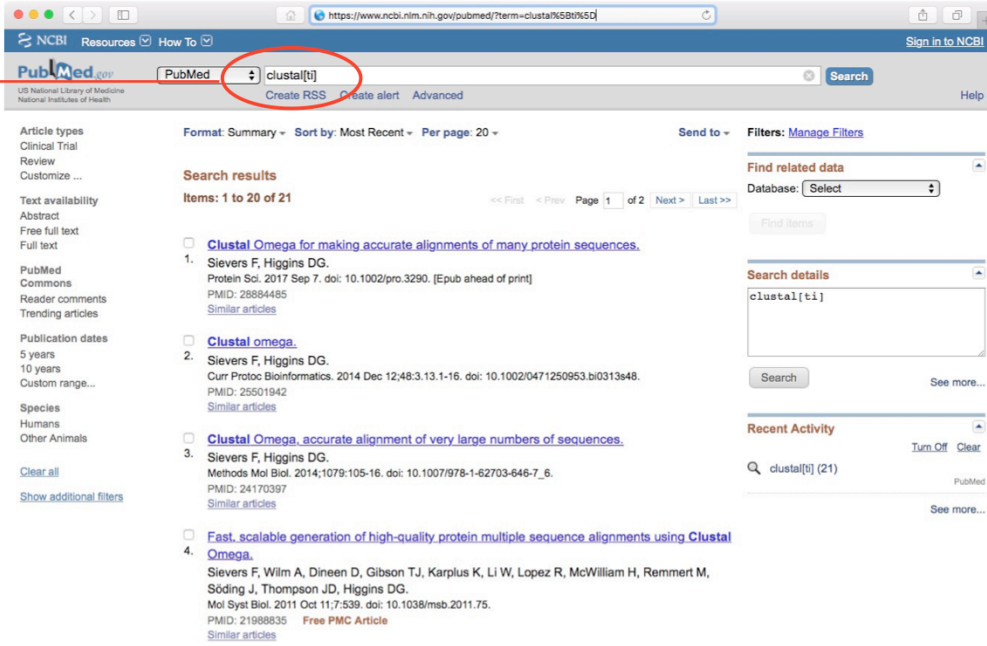
Edgar and Batzoglou's MSA review (2005¹), by the authors of MUSCLE and ProbCons, is a readable and comprehensive introduction to modern methods.

¹ <http://www.ncbi.nlm.nih.gov/pubmed/16679011>

THE [TI] FILTER

Identifying new developments via PubMed: filtering by **title** keywords

clustal[ti]



The screenshot shows a PubMed search results page. The search bar contains the query 'clustal[ti]'. The results are sorted by 'Most Recent' and show 21 items. The first four items are listed:

- Clustal Omega for making accurate alignments of many protein sequences.**
Sievers F, Higgins DG. *Protein Sci.* 2017 Sep 7. doi: 10.1002/pro.3290. [Epub ahead of print]
PMID: 28884485
[Similar articles](#)
- Clustal omega.**
Sievers F, Higgins DG. *Curr Protoc Bioinformatics.* 2014 Dec 12;48:3.13.1-16. doi: 10.1002/0471250953.bi0313e48.
PMID: 25501942
[Similar articles](#)
- Clustal Omega, accurate alignment of very large numbers of sequences.**
Sievers F, Higgins DG. *Methods Mol Biol.* 2014;1079:105-16. doi: 10.1007/978-1-62703-646-7_6.
PMID: 24170397
[Similar articles](#)
- Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega.**
Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG. *Mol Syst Biol.* 2011 Oct 11;7:539. doi: 10.1038/msb.2011.75.
PMID: 21988835 [Free PMC Article](#)
[Similar articles](#)

The 'Recent Activity' section shows the search history: clustal[ti] (21).

An alternative and more exploratory approach is to choose a recent "highly relevant" article, then to use the NCBI's "Related Articles" service. This search strategy allows you to search "forward" in time from a particular publication. In the above example, a search for **clustal[ti]** yielded a modern publication ...

RELATED ARTICLES

Identifying new developments via PubMed:
from **ANY** relevant article, find **ALL** others too.

The screenshot shows a PubMed search results page for the query "clustal[tit]". The search results are sorted by "Most Recent" and show 21 items. The first four items are listed, each with a checkbox and a "Similar articles" link. The "Find related data" dropdown menu is open, showing a list of databases and search options. The "PubMed" option is highlighted in blue. The "Find related data" option is circled in red.

Search results for "clustal[tit]":

- [Clustal Omega for making accurate alignments of many protein sequences.](#)
1. Sievers F, Higgins DG. *Protein Sci.* 2017 Sep 7. doi: 10.1002/pro.3290. [Epub ahead of print]
PMID: 28884485
[Similar articles](#)
- [Clustal omega.](#)
2. Sievers F, Higgins DG. *Curr Protoc Bioinformatics.* 2014 Dec 12;48:3.13.1-16. doi: 10.1002/0471250953.bi0313e48.
PMID: 25501942
[Similar articles](#)
- [Clustal Omega, accurate alignment of very large numbers of sequences.](#)
3. Sievers F, Higgins DG. *Methods Mol Biol.* 2014;1079:105-16. doi: 10.1007/978-1-62703-646-7_6.
PMID: 24170397
[Similar articles](#)
- [Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega.](#)
4. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG. *Mol Syst Biol.* 2011 Oct 11;7:539. doi: 10.1038/msb.2011.75.
PMID: 21988835 **Free PMC Article**
[Similar articles](#)

Find related data dropdown menu options:

- Assembly
- BioProject
- BioSample
- BioSystems
- Books
- ClinVar
- Conserved Domains
- dbGaP
- dbVar
- EST
- Gene
- Genome
- GEO DataSets
- GEO Profiles
- GSS
- HomoloGene
- MedGen
- Nucleotide
- OMIM
- PMC
- PopSet
- Probe
- Protein
- Protein Clusters
- PubChem BioAssay
- PubChem Compound
- PubChem Substance
- PubMed**
- SNP
- STW
- Structure
- Taxonomy
- UniGene

... and the “**Find related data**” option articles finds all modern relevant methods (although you will need to wade through the list).

Practical considerations for MSAs:

Principle: try to use sequences that are well distributed on the evolutionary tree. If a group of sequences biases the alignment, then not all information contributes equally to the result.

Include sequences with known structure wherever possible.

You may include more sequences than the ones that you are actually interested in comparing.

Ensure that your selection of sequences is appropriate to your question, e.g. clearly distinguish between orthologues and paralogues when trying to study function.

How do we work with MSAs in practice?

Spend some time and thought **before** you run the MSA to review the sequences that you are planning to align. Including un-alignable sequence **will** lead the algorithms astray and has the potential to degrade the entire alignment.

The requirement not to align **non-homologous** sequence should really be extended not to align (or at least: not to evaluate) sequence segments that have evolved in different context, such as in different local structural environments after insertions or deletions have occurred. The reason is: if the structural environment is not conserved, the mutation data matrix scores are irrelevant for the residues that are paired up. They may be "aligned" by the algorithm, but they are really not equivalent in structure or function, thus whether they have a good or poor similarity score is meaningless.

Rule #1 of pairwise alignment applies:

Don't align non-homologous sequence!

Here are some heuristics for preparing the sequences for MSAs:

- Remove unalignable (non-homologous) domains and N- and C-terminal extensions.
- Consider what the alignment is supposed to show! Include no more sequence than that which is needed.
- Be cautious when including widely divergent outlier sequences - better to align those to a profile derived from the result, than to create errors in the sensitive early stages of progressive alignments.
- If individual sequences do not align well, reconsider the evidence for their homology: the alignment may try to tell you that they are not homologous after all and should be removed from the set.
- If sequence segments behave as outliers, they may indicate sequencing frameshifts, skipped exons or erroneously translated introns: review the evidence for the gene-model that was used to define the sequence.
- If domain annotations exist: align domains separately.

Data formats for MSA: multi FASTA .mfa

```

>0918 CANAL XP_710918.256..352
-VIWDYETGWVHLTGIWKASLTIDGSNVSPSHLKADIVKLESTPKE----Y--QQYIKR
IRGGFLKIQ---GTWLPYKLCILARRFCYLRYSLIP-IFGTFDPDS
>9773 DEBHA XP_459773.187..274
-IIWDYETGVVHLTGIWKASIND--EVNTHRNKADIVKLESTPKQ----Y--HQHIKR
IRGGFLKIQ---GTWLPFDLCKMLAKRFCYHIRFQLIP-IF-----
>MBP1 SACCE NP_010227.024..107
SIMKRKDDWVNATHILKAANF-----AKAKRTRILEKEV-----L--KETHEK
VQGGFGKYQ---GTWVPLNIAKQLAEKFSVY--DQLKP-LFDFQTQTDG
>2599 ASPTE XP_001212599.130..218
-IMWDYNIGLVRTTFLFRS-----QNYSKTTPAKVLDANPGL--REISHS
ITGGAIVAQDKPGYWIPFEAAKAVAATFCWRIRYALTP-IFGLDFPSQ
>3510 ASPFU XP_753510.089..163
-LMRRSKDGYVSATGMFKIAPW--AKLEEEKAEREYLKTRGTSSEDEIAG-----
-----NIWVSPLLALELAKEY-----QMYDWVRRALD---
>7766 ASPNI XP_657766.089..163
-LMRRSKDGYVSATGMFKIAPW--AKLEEERSEREYLKTRPETSSEDEIAG-----
-----NVWISPVLALELAEEY-----KMYDWVRRALD---
>2267 NEUCR XP_962267.085..162
-LMRRSQDGYISATGMFKATFPY--ASQEEEAERKYIKSIPTTSSEETAG-----
-----NVWIPPEQALILAEY-----QITPWIRALLDPSD
>3762 MAGGR XP_363762.084..161
-LMRRSSDGYVSATGMFKATFPY--ADADEEAERNYIKSLPATSKEETAG-----
-----NVWISPDQALALAEY-----SIATWIRALLDPTD
>5459 GIBZE XP_385459.077..154
-LMRRSYDGFVSATGMFKASFPY--AEASDEDAERKYIKSLPTTSHEETAG-----
-----NVWIPPEQALILAEY-----KISPWIRALLDPTD
>3412 CANAL XP_723412.087..178
-VLRRVQDSFVNVVTLFQLIKL--EVLPTSQVDNYFDNEILSNLKYFGSSSNTQPQLDL
RKHQNIYLO---GIWIPYDKAVNLALKFD-----IYEITKKLF----
>9901 DEBHA XP_459901.067..158
-ILRRVQDSYINISQLFSILLKI--GHLSEAQLTNFLNNEILTNTQYLSGGSNPQFNDL
RNHEVRDLR---GLWIPYDRAVSLALKFD-----IYELAKSLF----

```

Three common formats exist for MSA results.

An **aligned** multi FASTA file contains FASTA formatted sequences into which gap characters have been inserted.

Of course, multi FASTA files can also be unaligned and they are the most common way of formatting **input files** for MSAs. Note that the example above contains hyphens, not just amino acid codes, and the hyphens ensure that the aligned sequences match up.

Data formats for MSA: CLUSTAL .aln

```

CLUSTAL FORMAT for T-COFFEE Version_5.05 SCORE=36, Nseq=11, Len=108

0918 CANAL      -VIWDYETGWVHLTGIWKASLTIDGSNVSPSHLKADIVKLESTPK-----Y--QQYIKR
9773 DEBHA      -IIWDYETGFVHLTGIWKASIND--EVNTHRNKADIVKLESTPKQ-----Y--HQHIKR
MBP1 SACCE      SIMKRKKDDWVNATHILKAANF-----AKAKRTRILEKEV-----L--KETHEK
2599 ASPTE      -IMWDYNIGLVRTTPLFRS-----QNYSKTTPAKVLDANPGL--REISHS
3510 ASPFU      -LMRRSKDGYVSATGMFKIAPFW--AKLEEEKAEREYLKTRREGTSEDEIAG-----
7766 ASPNI      -LMRRSKDGYVSATGMFKIAPFW--AKLEEESEEREYLKTRPETSEDEIAG-----
2267 NEUCR      -LMRRSQDGYISATGMFKATFPY--ASQEEEAERKYIKSIPPTSSEETAG-----
3762 MAGGR      -LMRRSDGYVSATGMFKATFPY--ADADEEAERNYIKSLPATSKEETAG-----
5459 GIBZE      -LMRRSYDGFVSATGMFKASFPY--AEASDEDAERKYIKSLPTSSEETAG-----
3412 CANAL      -VLRVQDSFVNVTQLFQILIKL--EVLPTSQVDNYFDNEILSNLKYFGSSSNTPOYLDL
9901 DEBHA      -ILRRVQDSYINISQLFSILLKI--GHLSEAQLTNFLNNEILTNTQYLSSGGSNPQFNDL
      ::      . : : :

0918 CANAL      IRGGFLKIQ---GTWLPYKLCILARRFCYLYRSLIP-IFGTDFFPDS
9773 DEBHA      IRGGFLKIQ---GTWLPPDLCKMLAKRFCYHIRFQLIP-IF-----
MBP1 SACCE      VQGGFGKYQ---GTWVPLNIAKQLAEKFSVY--DQLKP-LFDFQTQTDG
2599 ASPTE      ITGGAIVAQDKPGYWIPFEAAKAVAAATFCWRIRYALTP-IFGLDFPSQ
3510 ASPFU      -----NIWVSPVLALELAKEY-----QMYDWVRALLD---
7766 ASPNI      -----NVWISPVLALELAKEY-----KMYDWVRALLD---
2267 NEUCR      -----NVWIPPEQALILAKEY-----QITPWIRALLDPSD
3762 MAGGR      -----NVWISPDQALALAKEY-----SIATWIRALLDPTD
5459 GIBZE      -----NVWIPPEQALILAKEY-----KISPWIRALLDPTP
3412 CANAL      RKHQNIYLO---GIWIPYDKAVNLALKFD-----IYEITKKLF---
9901 DEBHA      RNHEVRDLR---GLWIPYDRAVSLALKFD-----IYELAKSLF---
      . * : . . : * : :

```

Three common formats exist for MSA results.

The CLUSTAL format is not the same as the CLUSTAL algorithm. A CLUSTAL formatted alignment is probably the most common way to print alignment data, because it shows the aligned columns.

Take care when formatting input FASTA files to ensure the **first 10 characters in your input file are unique** and contain **no special characters!** These are the characters that are usually used for the sequence names of the .aln files. I have seen programs break if they contain blanks, hyphens and | (the pipe character). The latter is especially annoying, since the | character is used in NCBI FASTA files to separate the database identifier from the accession number.

Data formats for MSA: MSF .msa

```

MSF: 108 Type: P Check: 3302 ..
Name: 0918_CANAL oo Len: 108 Check: 8295 Weight: 1.000
Name: 9773_DEBHA oo Len: 108 Check: 3489 Weight: 1.000
Name: MBP1_SACCE oo Len: 108 Check: 808 Weight: 1.000
Name: 2599_ASPTE oo Len: 108 Check: 241 Weight: 1.000
Name: 3510_ASPFU oo Len: 108 Check: 9082 Weight: 1.000
Name: 7766_ASPNI oo Len: 108 Check: 9952 Weight: 1.000
Name: 2267_NEUCR oo Len: 108 Check: 5383 Weight: 1.000
Name: 3762_MAGGR oo Len: 108 Check: 3063 Weight: 1.000
Name: 5459_GIBZE oo Len: 108 Check: 4901 Weight: 1.000
Name: 3412_CANAL oo Len: 108 Check: 5134 Weight: 1.000
Name: 9901_DEBHA oo Len: 108 Check: 2955 Weight: 1.000

//

0918 CANAL .VIWDYETGW VHLTGIWKAS LTIDGSNVSP SHLKADIVKL LESTPKE... .Y..OOYIKR
9773 DEBHA .IIWDYETGF VHLTGIWKAS IND..EVNTH RNLKADIVKL LESTPKQ... .Y..HOHIKR
MBP1 SACCE SIMKRKDDW VNATHILKAA NF..... .AKAKRTRI LEKEV... .L..KETHEK
2599 ASPTE .IMWDYNI GL VRTTPLFRS . . . . . ONYSKT TPAKVLDA NP GL..REISHS
3510 ASPFU .LMRRSKDGY VSATGMFKIA FPW..AKLEE EKAEREYLKT REGTSEDEIA G.....
7766 ASPNI .LMRRSKDGY VSATGMFKIA FPW..AKLEE ERSEREYLKT RPTSEDEIA G.....
2267 NEUCR .LMRRSDGY ISATGMFKAT FPY..ASOE EBAERKYIKS IPTSSEETA G.....
3762 MAGGR .LMRRSDGY VSATGMFKAT FPY..ADAED EBAERKYIKS LPATSKETA G.....
5459 GIBZE .LMRRSYDGF VSATGMFKAS FPY..AEASD EDAERKYIKS LPTTSHEETA G.....
3412 CANAL .VLRVVQDSF VNVTLFOIL IKL..EVLPT SQVDNYFDNE ILSNLKYFGS SSNTPOYLDL
9901 DEBHA .ILRRVQDSY INISQLFSTL LRI..GHLSE AQLTNFLNNE ILTNTQYLSS GGSNPQFNDL

0918 CANAL IRGGFLKIO. .GTWLPYKL CKILARRFCY YLRYSLIP. I FGTDFPDS
9773 DEBHA IRGGFLKIO. .GTWL PFDL CKMLAKRFCY HIRFOLIP. I F.....
MBP1 SACCE VOGGFGKYO. .GTWVPLNI AKQLAEKFSV Y..DOLKP.L FDFQTQTDG
2599 ASPTE ITGGAIVAQD KPGYWIPFEA AKAVAATFCW RIRYALTP. I FGLDFPSQ
3510 ASPFU . . . . . NIWVSPLL ALELAKEY. . . . . QMYDWV RALLD...
7766 ASPNI . . . . . NVWISPV L ALELAEEY. . . . . KMYDWV RALLD...
2267 NEUCR . . . . . NVWIPPEO ALILAEY. . . . . OTPTWI RALLDPSD
3762 MAGGR . . . . . NVWISPDO ALALAEY. . . . . SIATWI RALLDPTD
5459 GIBZE . . . . . NVWIPPEO ALILAEY. . . . . KISPWI RALLDPTP
3412 CANAL RKHQNIYLO. .GIWIPYDK AVNLALKFD. . . . . IYELIT KKLK...
9901 DEBHA RNHEVRDLR. .GLWIPYDR AVSLALKFD. . . . . IYELA KSLF...

```

Three common formats exist for MSA results.

MSF is a legacy format from the GCG package of sequence alignments, also produced by the EMBOSS tool EMMA, and supported as a valid input format for many programs. Gaps are denoted by periods and checksums are calculated for the sequences and for the alignment.

Manual editing can improve *columnwise* *homology*, and conserved structural and functional motifs.

It is oK to manually edit alignments to improve them, **if** you have additional knowledge (e.g. about the protein's biology) that can inform a better alignment.

MANUAL EDITING

Manual editing can often improve MSAs because you can take context into account whereas algorithms typically calculate scores based on objective functions that are computed over columns only.

(Context: e.g. where are indels placed, which residues are part of a functional site ...)

- Move sequences of different length (thus having obviously different structure) for dissimilar families into separate columns. Try not to include non-alignable residues in the same column and create the impression they are alignable.
- Move indels into regions adjacent to secondary structure elements, even if this creates a lower sequence alignment score. This is where they are commonly **accommodated** in structure, even if they have been generated elsewhere.
- Move two/four/six-residue insertions equally to both sides of a conserved beta-turn, rather than postulating only a single insertion on one side.
- Adjust gap positions to minimize the number of indel **sites**.
- Consider the evolutionary tree to minimize number of indel **events**.
- Conserve hydrophobic patterns in regions of secondary structure, e.g. alternating hydrophobic residues in beta-strands and $i \rightarrow i+4$ patterns in alpha-helices.
- Preserve conservation especially in binding sites and functional motifs, even at the cost of larger indels.

It is common and perfectly permissible to manually edit a MSA with some biologically motivated heuristic in mind **as long as you document what you have done!** In the early days of MSAs, editing was always required since the results were usually obviously inadequate. In all cases in which the algorithm uses only the input sequences for the alignment, this is still often true. However, regarding the more modern template-based procedures (e.g. SPEM, PROMALS or PRALINE) editing may be actually ignore/discard the additional information the algorithm has used.

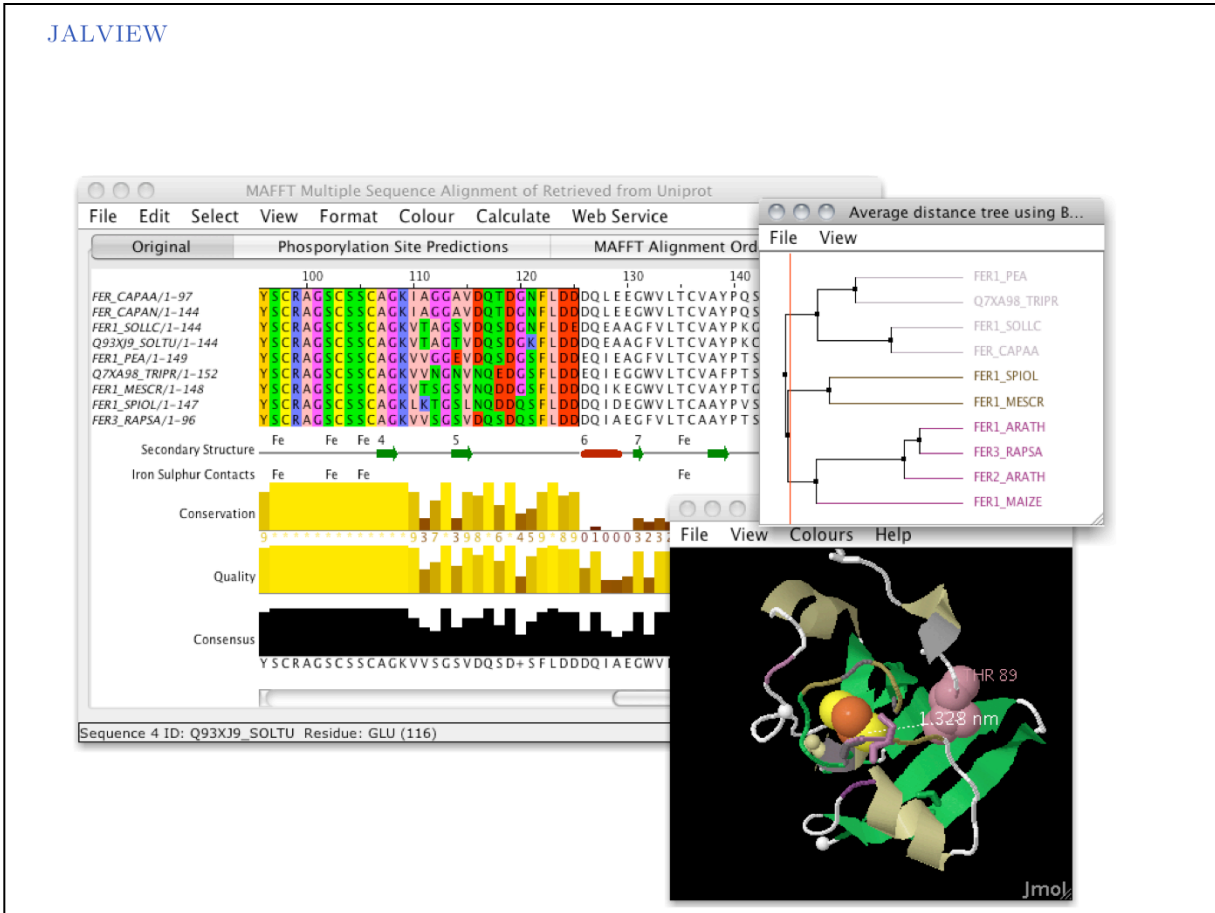
MS WORD?

You **can** use a text-editor to edit MSA's, (e.g. MS Word) ...

1. Format your sequences in a **fixed-width** font (e.g. **Courier New**).
2. Display the entire sequence in one line: small font, landscape -page setup, legal- or custom wide paper size, small margins, magnified view.
3. Press the **<ALT>** key
4. You can now select **columns** and **blocks** of text.
5. You can **format** your selection; you can **delete**, **copy**, **cut**...
6. You can also **paste** entire columns or column ranges into the alignment.

... but you should avoid it, because **much better tools are freely available**.

JALVIEW



JALVIEW is a well engineered, very functional, free MSA editor and analysis program from Geoff Barton's lab in Dundee. If an MSA is going to be published by you, you will probably find its functions very useful. Good Website with documentation and tutorials.

<http://www.jalview.org/>

SUMMARY

Not selecting your input sequences with the utmost care **is cargo cult science.**

Not constructing the best alignment that is possible **is cargo cult science.**

Use not one, but several algorithms

Identify areas of consensus between algorithms and areas that are sensitive to parameters and algorithms.

Don't be afraid to edit your alignment.

BEST PRACTICE

Use more than one alignment method,
but above all, use your common sense.

Representation of multiple alignments:

```
DIVMTQSPSSL...  
DIVMSQSPSSL...  
DIVMTQSPTFL...  
DVVMTQTPTLL...  
NIVLTQSPASL...  
DIQMTQSTSSL...  
  
DIVMTQSPSSL...
```

"Consensus sequence"

Consensus sequences are lossy.
consider e.g. the prokaryotic TATAAT
box:

TAT...
49% 58% 54%

Only ~5% of bacterial promoters
actually have the consensus sequence
TATAAT.

How do we represent the information from an MSA?


```
DIVMTQSPSSL...  
DIVMSQSPSSL...  
DIVMTQSPFTL...  
DVVMTQTPTLL...  
NIVLTQSPASL...  
DIQMTQSTSSL...
```

1. DIVMTQSPSSL...
2. NVQLS TTTFL
3. AL

Ranked characters are better ...
but still lossy.

Even better: report explicit
frequencies ...

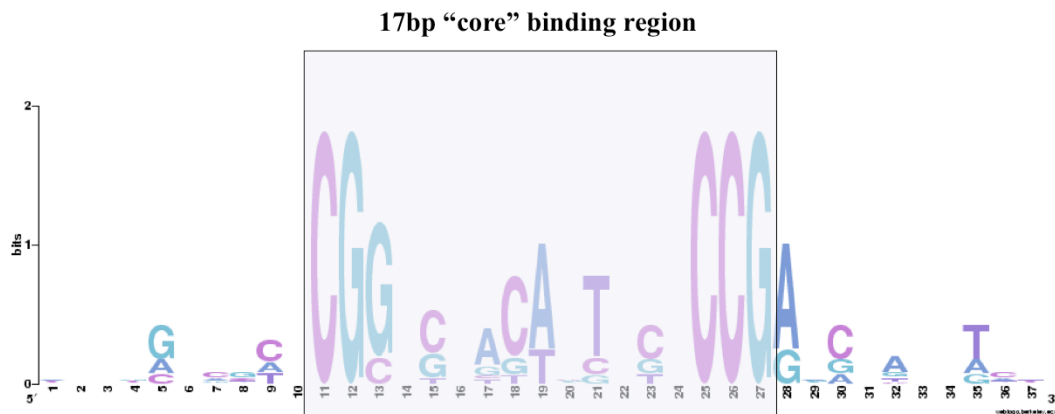
... or profiles. However profiles are
not very readable

REPRESENTING MSAs: SEQUENCE LOGO

“Sequence logos” concentrate the following information into a single graphic:

1. The general consensus of the sequences.
2. The order of predominance of the residues at every position.
3. The relative frequencies of every residue at every position.
3. The amount of information present at every position in the sequence, measured in bits.
5. An initiation point, cut point, or other significant location (if appropriate)."

(Schneider, Tom S (1990) Sequence logos: a new way to display consensus sequences *NAR* **18**:6097)



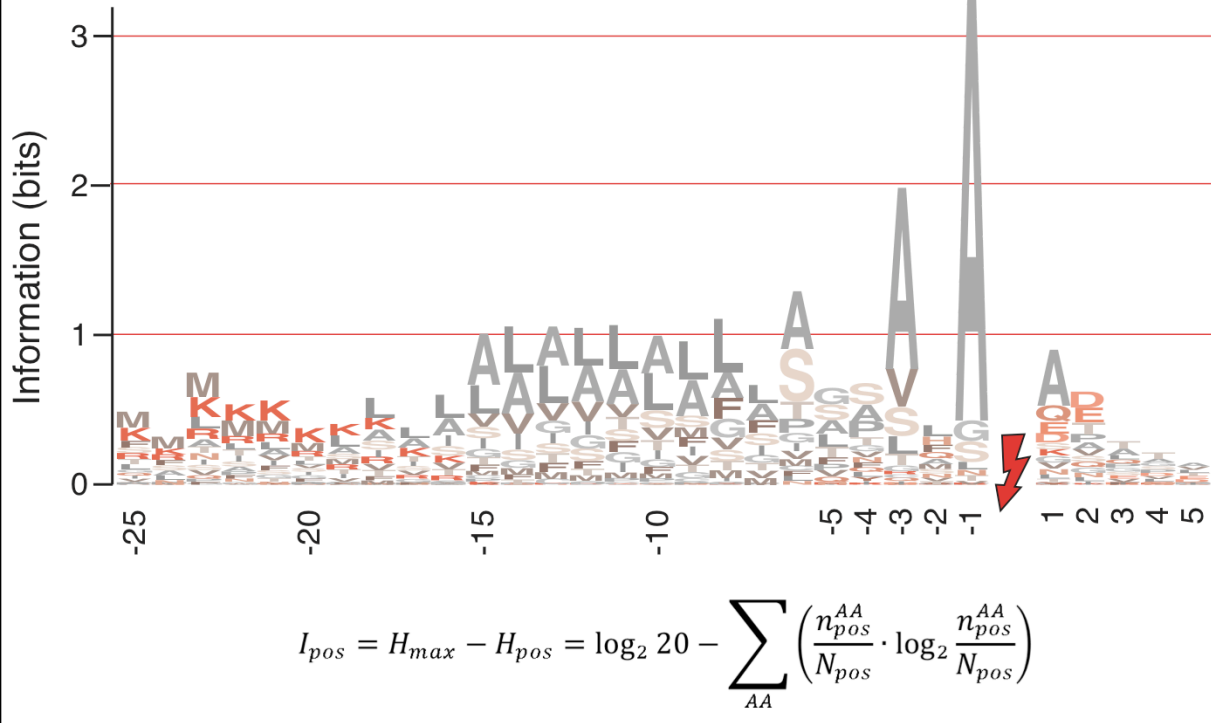
Pro: Much better indication of conservation/propensity of characters.

Con: Small samples have artificially high information scores.

Sequence logo of Gal4 binding sites with 10 nucleotides flanking bases. Created with **WebLogo** (<http://weblogo.berkeley.edu/>).

A Sequence Logo is a graphical representation of aligned sequences where at each position the height of a column is proportional to the (Shannon) information of that position and the relative size of each character is proportional to its frequency within the column. Sequence Logos were pioneered by Tom Schneider who maintains an informative Website about their use and theoretical foundations (<https://schneider.ncifcrf.gov/>). Note that there is considerable additional information in the flanking sequences that are not included in the published description of the core binding pattern; it is advantageous if you are able to run such analyses yourself, rather than rely on someone else's opinion.

Sequence logo example: bacterial signal sequences



Example for common features in gram-negative signal-peptide sequences in a Sequence Logo.

Sequences were aligned on the signal-peptidase cleavage site. Their common features include a positively charged N-terminus (K, R), a hydrophobic helical stretch (A, L, V) and a small residue that precedes the actual cleavage site (A). The information is calculated by comparing the entropy of 20 equiprobable amino acids and the entropy actually observed amino acids in each position.

Yet another question once we have produced an MSA is how to analyse it quantitatively: can we define a meaningful score for each column in the alignment?

This ...

... could be based on the substitution matrix; (variation score)

... could be based on observed frequencies; (information score)

... could be based on explicit phylogenies. (conservation score)

Variation Score

Sum the substitution matrix values for every unique pair.

Easy to calculate ...

Easy to interpret ...

$$S = \sum_{i=1}^n \sum_{j=i+1}^n \text{MDM}(C_i, C_j)$$

$i=4$ — CTT**CG**GATCACGG
 AATGAGCCTTCGC
 TATTGAAGTACGG
 AG**CC**GCCGAGCGG
 AGATGTGCCTCGC
 ACC**CC**CACGTTCCG
 AAA**A**CTCGCACGG
 TCG**GG**AAGCTCGG
 $j=9$ — CTT**C**ATTTACCGG
 CTGGGCGCCGCGG
 MDM(C,C) = 1 G**CC**GACAATCGG
 CCG**GG**TCGCCCGG

But these sums reflect a particular model that may not be adequate for the question at hand.

Variation score based measures are derived from average properties of the mutation data matrix, they are probably not accurate for the particular conserved positions of specific sequences.

Information Score

Evaluate the frequencies of every amino acid at a position.

Easy to calculate ...

Easy to interpret ...

But require a reference distribution
for interpretation.

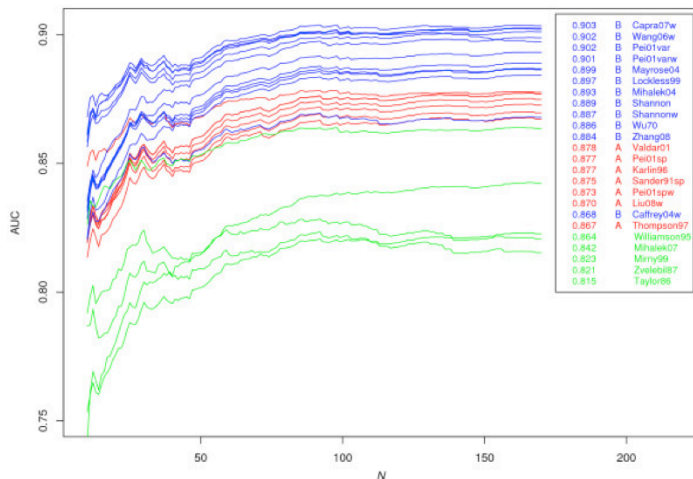
Many possible scoring systems exist:

```
CTTCGGATCACGG
AATGAGCCTTCGC
TATTGAAGTACGG
AGCCGCCGAGCGG
AGATGTGCCTCGC
ACCCACGTTTCGG
AAAACTCGCACGG
TCGGGAAGCTCGG
CTTCATTTACCGG
CTGGGCGCCGCGG
GGCGAACAATCGG
CCGGGTCGCCCGG
```

A=1, C=4, G=5, T=2

Johansson & Toh found 2010 that:

“When using a score to predict catalytic sites, frequency based scores that also consider a background distribution are most successful.”



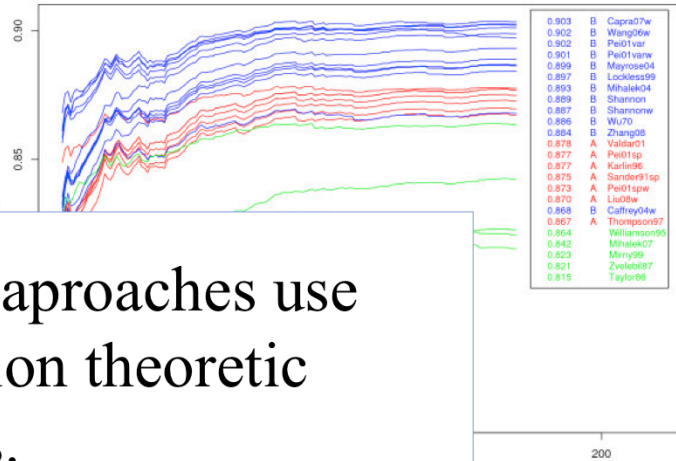
Scores evaluated by comparisons of alignments over the Catalytic Site Atlas.

Johansson, F & Toh, H (2010) A comparative study of conservation and variation scores. *BMC Bioinformatics* 11:388

Johansson & Toh found
2010 that:

“When using a score to
predict c
frequen
that also
backgro
are mos

The best approaches use
information theoretic
measures.



Scores evaluated by comparisons of alignments over the Catalytic Site Atlas.

Johansson, F & Toh, H (2010) A comparative study of conservation and variation scores. *BMC Bioinformatics* 11:388

<http://steipe.biochemistry.utoronto.ca/abc>

B O R I S . S T E I P E @ U T O R O N T O . C A

DEPARTMENT OF BIOCHEMISTRY & DEPARTMENT OF MOLECULAR GENETICS
UNIVERSITY OF TORONTO, CANADA