Introduction to Bioinformatics

Iosif Vaisman

Email: ivaisman@gmu.edu

Multiple alignment

```
VTISCTGSSSNIGAG-NHVKWYQQLPG
VTISCTGTSSNIGS--ITVNWYQQLPG
LRLSCSSSGFIFSS--YAMYWVRQAPG
LSLTCTVSGTSFDD--YYSTWVRQPPG
PEVTCVVVDVSHEDPQVKFNWYVDG--
ATLVCLISDFYPGA--VTVAWKADS--
AALGCLVKDYFPEP--VTVSWNSG---
VSLTCLVKGFYPSD--IAVEWESNG--
```

Column cost: the sum of costs for all possible pairs

Computational complexity

Alignment of protein sequences with 200 amino acid residues:

# of sequences	CPU time
2	1 sec
3	200 sec
10	200^8 sec

Multiple alignment

A correct multiple alignment corresponds to an evolutionary history:

no correct way to determine
practical way - to find an alignment with the maximum score

Multiple sequence alignment

Given k ($k > 2$) sequences, $s_1, \ldots, s_k$, each sequence consisting of characters from an alphabet $\mathcal{A}$ multiple alignment is a a rectangular array, consisting of characters from the alphabet $\mathcal{A}'$ ($\mathcal{A}$ + "-"), that satisfies the following 3 conditions:

1. There are exactly k rows.
2. Ignoring the gap character, row number i is exactly the sequence s_i.
3. Each column contains at least one character different from "-".

Consensus

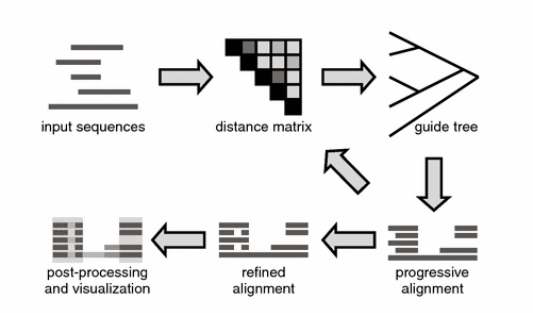
Consensus sequence - idealized sequence in which each position represents the amino acid most often found when many sequences are compared.

Plurality - minimum number of votes for a consensus
Threshold - scoring matrix value below which a symbol may not vote for a coalition.
Sensitivity - minimum score to select consensus
Profiles - blocks of prealigned sequences

Multiple alignment algorithm

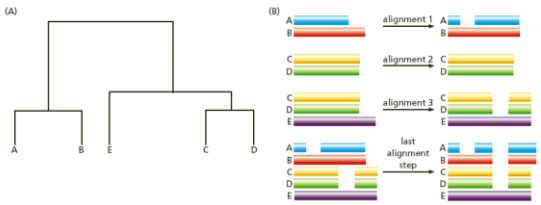
1. Pairwise alignments (progressive pairwise alignments)
2. Distance matrix calculation
3. Guide tree creation (hierarchical clustering)
4. New sequence addition

Multiple alignment algorithm



Do and Katoh, 2008

Multiple alignment algorithm

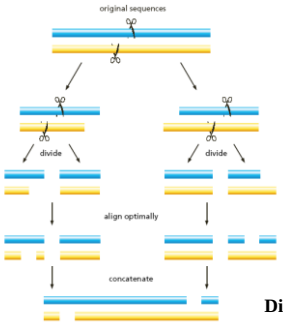


Guide tree

Progressive alignments

Zvelebil & Baum, 2007

Multiple alignment algorithm



Divide-and-conquer method

Zvelebil & Baum, 2007

Scoring system (distances)

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

$S_{real}(ij)$ - observed similarity score for two aligned sequences i and j

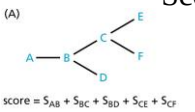
$S_{iden}(ij)$ - average of the two scores for each sequence aligned with itself

$S_{rand}(ij)$ - average score determined from 100 global randomizations of the two sequences

The distances D(ij) are used to generate the distance matrix from which the approximate guide tree is generated.

Scoring systems

(A)



Phylogenetic tree

score = $S_{AB} + S_{BC} + S_{BD} + S_{CE} + S_{CF}$

(B)

Star tree

score = $S_{AB} + S_{AC} + S_{AD} + S_{AE} + S_{AF}$

(C)

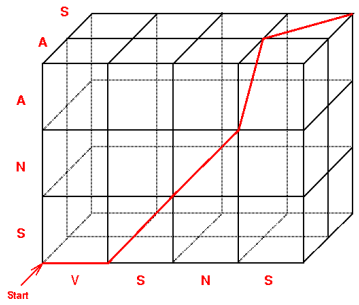
Sum of pairs

score = $S_{AB} + S_{AC} + S_{AD} + S_{AE} + S_{AF}$
$+ S_{BC} + S_{BD} + S_{BE} + S_{BF} + S_{CD}$
$+ S_{CE} + S_{CF} + S_{DE} + S_{DF} + S_{EF}$

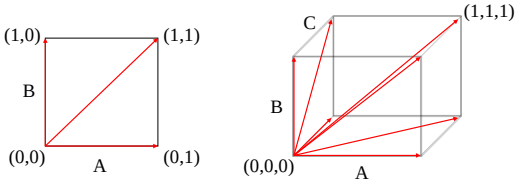
Multiple alignments comparison

Zvelebil & Baum, 2007

Multiple alignment



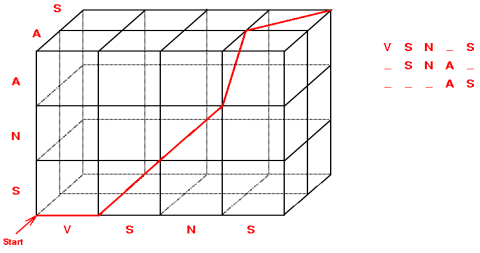
Multiple alignment



Segment - line joining two vertices

Each unit m-dimensional cube in the lattice contains 2^m -1 segments

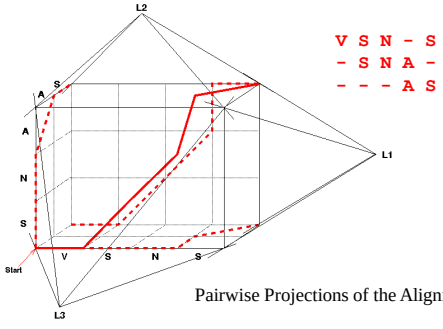
Multiple alignment



Alignment Path for 3 Sequences

(0,0,0), (1,0,0), (2,1,0), (3,2,0), (3,3,1), (4,3,2)

Multiple alignment



Pairwise Projections of the Alignment

Alignment statistics

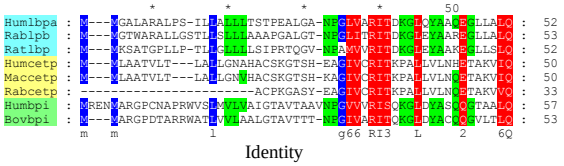
		Rablpb	Humcetp	Rabcetp	Bovbpi
	Humlbpa	Ratlbp	Maccetp	Humbpi	
	1	2	3	4	5	6	7	8
	478	67%	65%	19%	19%	18%	42%	43%
1	0	82%	80%	39%	36%	64%	65%
	0	1%	0%	5%	5%	12%	2%	2%
	327	483	58%	16%	16%	16%	39%	41%
2	400	0	75%	38%	38%	35%	62%	63%
	5	0	0%	5%	5%	12%	1%	1%
	318	284	482	18%	18%	17%	40%	43%
3	390	367	0	38%	38%	35%	64%	64%
	4	1	0	5%	5%	12%	1%	1%

1: The number of residues that match exactly (identical residues) between the two sequences.
2: The number of residues whose juxtaposition yields a greater than zero score in the current scoring table (similar residues or conservative substitutions).
3: The number of residues lined up with a gap character.

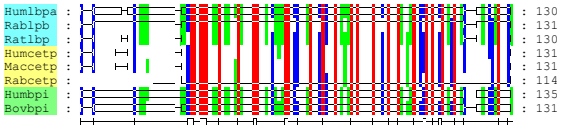
Alignment score

	Humlbpa	Rablpb	Ratlbp	Humcetp	Maccetp	Rabcetp	Humbpi	Bovbpi
	1	2	3	4	5	6	7	8
1	4077							
2	5358	4129						
3	5323	5650	4096					
4	8103	8229	8112	4210				
5	8109	8243	8118	4332	4219			
6	8535	8672	8575	5511	5519	4261		
7	6474	6531	6500	8103	8119	8572	4103	
8	6392	6434	6378	8033	8035	8520	5508	4083
	1	2	3	4	5	6	7	8

Alignment visualization

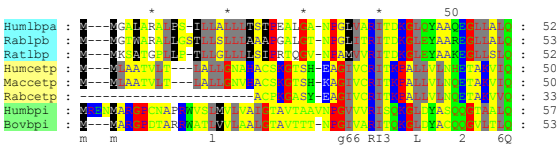


Identity

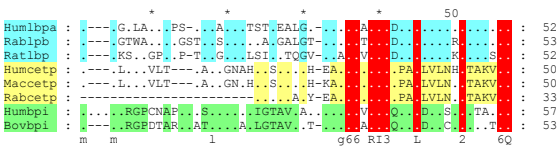


Summary view

Alignment visualization

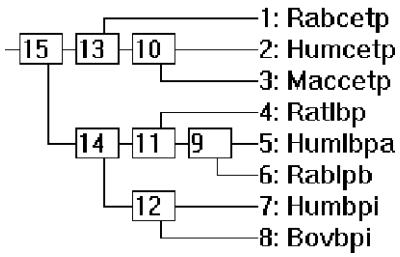


Physico-chemical properties



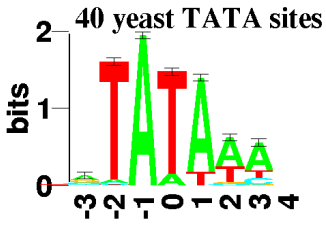
Differences mode

Alignment visualization (tree)



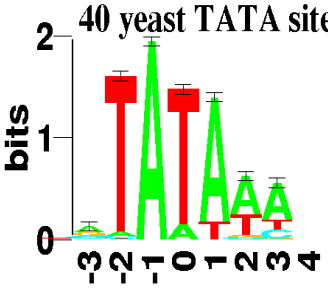
Sequence Logos:

a quantitative graphical display for binding sites and proteins

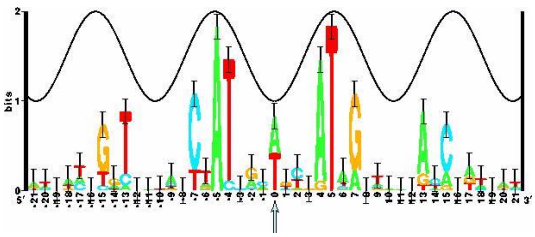


Reference: Schneider, T.D. *Meth. Enzym* **274**:445, 1996

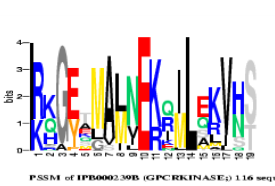
Sequence Logos



Sequence Logos



Sequence Logos




```
GRK6_RAT|P97711      ( 225) RKGEAMALNEKQILEKVNS  11
GRK6_MOUSE|O70293    ( 225) RKGEAMALNEKQILEKVNS  11
O70294               ( 225) RKGEAMALNEKQILEKVNS  11
O70295               ( 225) RKGEAMALNEKQILEKVNS  11
O70296|O70296_MOUSE  ( 225) RKGEAMALNEKQILEKVNS  11
P97548|P97548_RAT    ( 225) RKGEAMALNEKQILEKVNS  11
P97549|P97549_RAT    ( 225) RKGEAMALNEKQILEKVNS  11
GRK5_HUMAN|P34947    ( 225) RKGESMALNEKQILEKVNS  12
GRK6_HUMAN|P43250    ( 225) RKGEAMALNEKQILEKVNS  11
O60541               ( 225) RKGEAMALNEKQILEKVNS  11
GRK5_RAT|Q62833      ( 225) RKGESMALNEKQILEKVNS  12
O70292               ( 225) RKGESMALNEKQILEKVNS  12
O70297               ( 225) RKGESMALNEKQILEKVNS  12
Q13293               ( 226) RKGEAMALNERKILEKVQS  14
Q13294               ( 194) RKGEAMALNERKILEKVQS  14
Q15314               ( 226) RKGEAMALNERKILEKVQS  14
Q15316               ( 194) RKGEAMALNERKILEKVQS  14
GRK5_BOVIN|P43249    ( 225) RKGESMALNEKQILEKVNS  12
Q15395               ( 226) RKGEAMALNERKILEKVQS  14
Q15313               ( 226) RKGEAMALNERKILEKVQS  14
Q15315               ( 194) RKGEAMALNERKILEKVQS  14
GRK4_HUMAN|P32298    ( 194) RKGEAMALNERKILEKVQS  14
Q14453               ( 226) RKGEAMALNERKILEKVQS  14
GRK4_MOUSE|O70291    ( 225) RKGEAMALNEKQILEKLHS  21
```

Multiple Alignment Programs

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	—	—

C. Notredame, 2007

Multiple Alignment Programs

PROGRAM	ADVANTAGES	CAUTIONS
CLUSTALW	Uses less memory than other programs	Less accurate or scalable than modern programs
DIALIGN	Attempts to distinguish between alignable and non-alignable regions	Less accurate than CLUSTALW on global benchmarks
MAFFT, MUSCLE	Faster and more accurate than CLUSTALW; good trade-off of accuracy and computational cost.	For very large data sets (say, more than 1000 sequences) select time- and memory-saving options
PROBCONS	Highest accuracy score on several benchmarks	Computation time and memory usage is a limiting factor for large alignment problems (>100 sequences)
ProDA	Does not assume global alignability; allows repeated, shuffled and absent domains.	High computational cost and less accurate than CLUSTALW on global benchmarks
T-COFFEE	High accuracy and the ability to incorporate heterogeneous types of information	Computation time and memory usage is a limiting factor for large alignment problems (>100 sequences)

RC Edgar and S Batzoglou, 2006

Typical alignment tasks

Input data	Recommendations
2–100 sequences of typical protein length (maximum around 10,000 residues) that are approximately globally alignable	Use PROBCONS, T-COFFEE, and MAFFT or MUSCLE, compare the results using ALTAVIST. Regions of agreement are more likely to be correct. For sequences with low percent identity, PROBCONS is generally the most accurate, but incorporating structure information (where available) via 3DCoffee (a variant of T-COFFEE) can be extremely helpful.
100–500 sequences that are approximately globally alignable	Use MUSCLE or one of the MAFFT scripts with default options. Comparison using ALTAVIST is possible, but the results are hard to interpret with larger numbers of sequences unless they are highly similar.
>500 sequences that are approximately globally alignable	Use MUSCLE with a faster option (we recommend maxiters-2) or one of the faster MAFFT scripts
Large numbers of alignments, high-throughput pipeline.	Use MUSCLE with faster options (e.g. maxiters-1 or maxiters-2) or one of the faster MAFFT scripts

RC Edgar and S Batzoglou, 2006

Typical alignment tasks

Input data	Recommendations
2–100 sequences with conserved core regions surrounded by variable regions that are not alignable	Use DIALIGN
2–100 sequences with one or more common domains that may be shuffled, repeated or absent.	Use ProDA
A small number of unusually long sequences (say, >20,000 residues)	Use CLUSTALW. Other programs may run out of memory, causing an abort (e.g. a segmentation fault).

RC Edgar and S Batzoglou, 2006