

Introduction to Bioinformatics

Iosif Vaisman

Email: ivaisman@gmu.edu

Multiple alignment

```
VTISCTGSSSNIGAG-NHVKWYQQLPG
VTISCTGTSSNIGS--ITVNWYQQLPG
LRLSCSSSGFIFSS--YAMYWRQAPG
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, \dots, s_k , each sequence consisting of characters from an alphabet \mathcal{A}
multiple alignment is a rectangular array, consisting of characters from the alphabet \mathcal{A}' ($\mathcal{A} + \text{"-"}$), 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

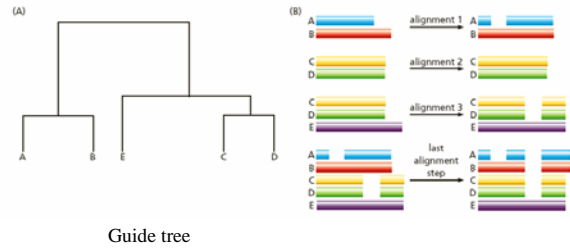
Consensus sequence - idealized sequence in which each position represents the amino acid most often found when many sequences are compared.

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

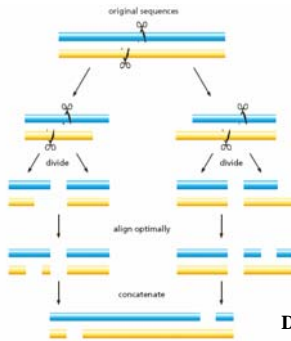


Guide tree

Progressive alignments

Zvelebil & Baum, 2007

Multiple alignment algorithm



Divide-and-conquer method

Zvelebil & Baum, 2007

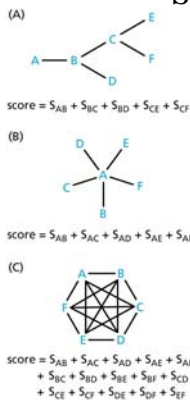
Scoring system (distances)

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

- $S_{real}(ij)$ - 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



Phylogenetic tree

Star tree

Sum of pairs

Multiple alignments comparison

```

(A) structural/functional alignment from BAIBase
T149  SHERKPPFNRK...
T151  ERKPPAFYFK...
T152  ERKPPAFYFK...
T153  ERKPPAFYFK...

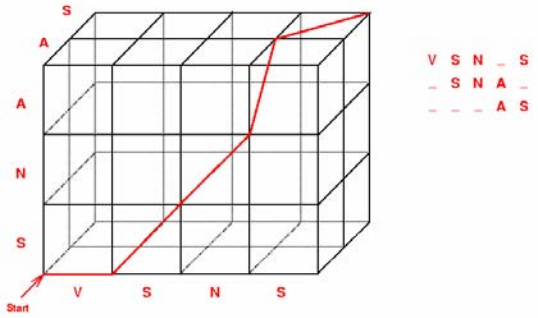
(B) DIALIGN multiple sequence alignment
T149  SHERKPPFNRK...
T151  ERKPPAFYFK...
T152  ERKPPAFYFK...
T153  ERKPPAFYFK...

(C) ClustalW multiple sequence alignment
T149  SHERKPPFNRK...
T151  ERKPPAFYFK...
T152  ERKPPAFYFK...
T153  ERKPPAFYFK...

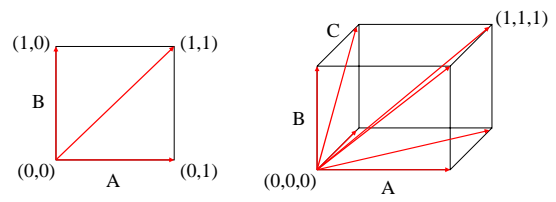
(D) divide-and-conquer multiple sequence alignment
T149  SHERKPPFNRK...
T151  ERKPPAFYFK...
T152  ERKPPAFYFK...
T153  ERKPPAFYFK...
    
```

Zvelebil & Baum, 2007

Multiple alignment

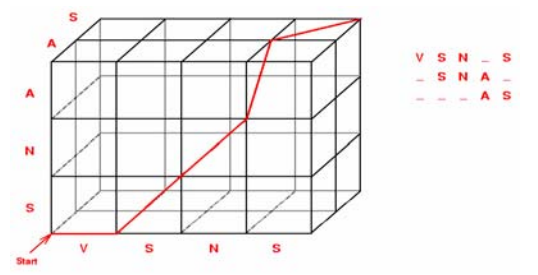


Multiple alignment



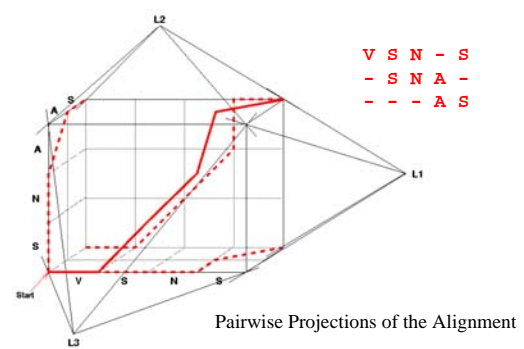
Segment - line joining two vertices
 Each unit m-dimensional cube in the lattice contains $2^m - 1$ segments

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

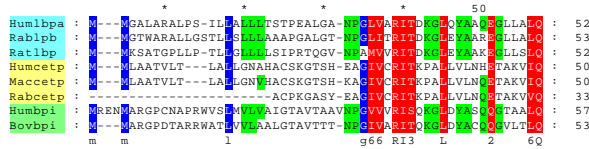
	Humlbpa	Rablpb	Ratlpb	Humcetp	Maccetp	Rabcetp	Humbpi	Bovbpi
	1	2	3	4	5	6	7	8
1	478	67%	65%	19%	19%	18%	42%	43%
	0	82%	80%	39%	39%	36%	64%	65%
	0	1%	0%	5%	5%	12%	2%	2%
2	327	483	58%	16%	16%	16%	39%	41%
	400	0	75%	38%	38%	35%	62%	63%
	5	0	0%	5%	5%	12%	1%	1%
3	318	284	482	18%	18%	17%	40%	43%
	390	367	0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

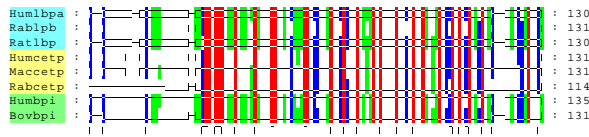
Alignment score

	Humlbpa	Rablpb	Ratlpb	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

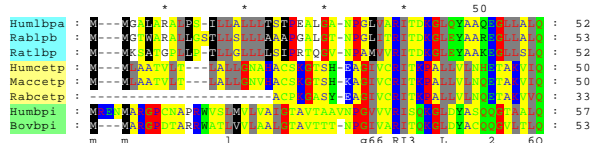


Identity

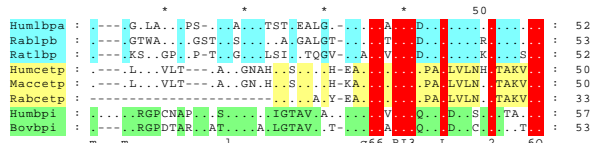


Summary view

Alignment visualization

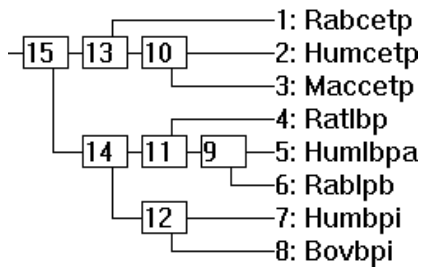


Physico-chemical properties



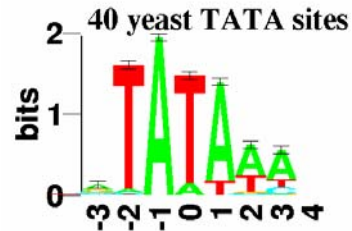
Differences mode

Alignment visualization (tree)



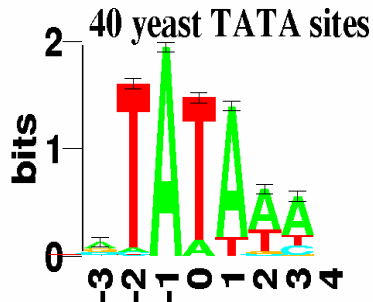
Sequence Logos:

a quantitative graphical display for binding sites and proteins

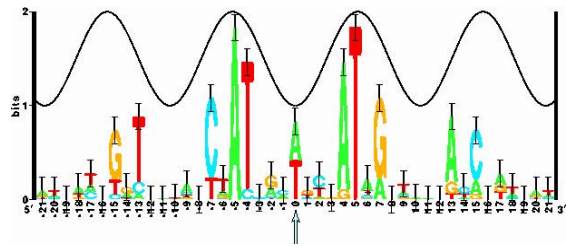


Reference: Schneider, T.D. *Meth. Enzym* 274:445, 1996

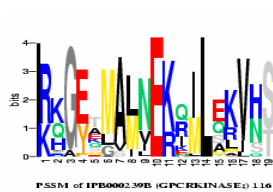
Sequence Logos



Sequence Logos



Sequence Logos



PSM1 of 1PB0002.39B (GPCR KINASE): 116 sequences

GRK6_BAT P97711	(225)	RKGEAMALNRKQILEKVN	11
GRK6_MOUSE O70293	(225)	RKGEAMALNRKQILEKVN	11
O70294	(225)	RKGEAMALNRKQILEKVN	11
Q13295	(225)	RKGEAMALNRKQILEKVN	11
O70296 O70296_MOUSE	(225)	RKGEAMALNRKQILEKVN	11
P97548 P97548_BAT	(225)	RKGEAMALNRKQILEKVN	11
P97549 P97549_BAT	(225)	RKGEAMALNRKQILEKVN	11
GRK5_HUMAN P34947	(225)	RKGEAMALNRKQILEKVN	12
GRK6_HUMAN P43250	(225)	RKGEAMALNRKQILEKVN	11
GRK541	(225)	RKGEAMALNRKQILEKVN	11
GRK5_BAT Q62833	(225)	RKGEAMALNRKQILEKVN	12
O70292	(225)	RKGEAMALNRKQILEKVN	12
O70297	(225)	RKGEAMALNRKQILEKVN	12
Q13293	(226)	RKGEAMALNRKQILEKVS	14
Q13294	(194)	RKGEAMALNRKQILEKVS	14
Q15314	(226)	RKGEAMALNRKQILEKVS	14
Q15316	(194)	RKGEAMALNRKQILEKVS	14
GRK5_BOVIN P43249	(225)	RKGEAMALNRKQILEKVN	14
Q15313	(226)	RKGEAMALNRKQILEKVS	14
Q15315	(194)	RKGEAMALNRKQILEKVS	14
GRK4_HUMAN P32298	(194)	RKGEAMALNRKQILEKVS	14
Q14853	(226)	RKGEAMALNRKQILEKVS	14
GRK4_MOUSE O70291	(225)	RKGEAMALNRKQILEKLS	21

Multiple Alignment Programs

- **Pileup (GCG)**: Needleman and Wunsch algorithm for pairwise alignment and UPGMA method for tree construction
- **CLUSTAL**: Wilbur and Lipman algorithm for pairwise alignment (*CABIOS* 8:189, 1992)
- **PIMA**: pattern-matching based algorithm (*PNAS* 87:118, 1990)
- **TreeAlign**: phylogenetic algorithm (*Meth. Enzymol.* 18:626, 1990)