MULTIPLE SEQUENCE ALIGNMENT

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

```
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

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 *A*

multiple alignment is a a rectangular array, consisting of characters from the alphabet *A'* (*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

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

Scoring system (distances)

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

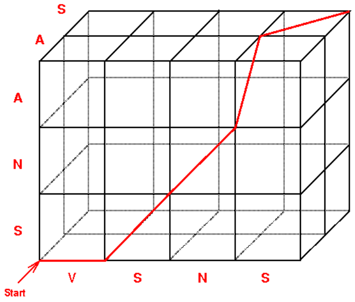
$S_{real}(ij)$ - observed similarity score for two aligned sequences i and j

$S_{iden}(ij)$ - 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

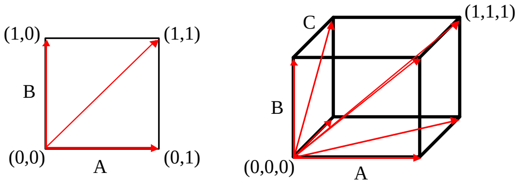
The distances D(ij) are used to generate the distance matrix from which the approximate guide tree is generated.

Multiple alignment



```
V S N _ S
_ S N A _
_ _ _ A S
```

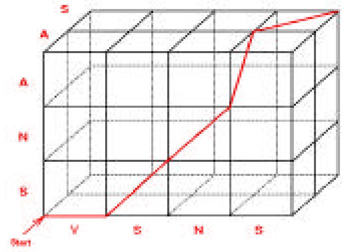
Multiple alignment



Segment - line joining two vertices

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

Multiple alignment

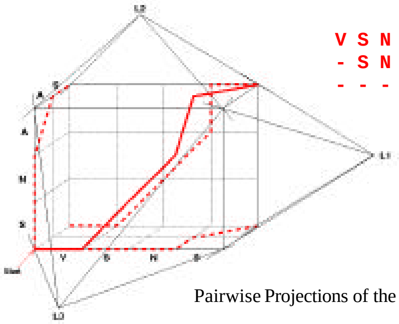


```
V S N _ S
_ S N A _
_ _ _ A S
```

Alignment Path for 3 Sequences

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

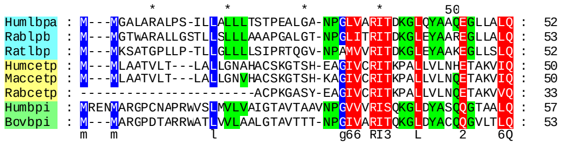
Multiple alignment



```
V S N - S
- S N A -
- - - A S
```

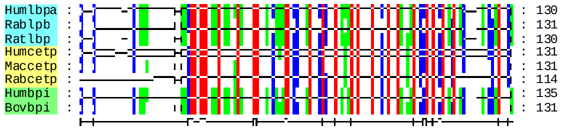
Pairwise Projections of the Alignment

Alignment visualization



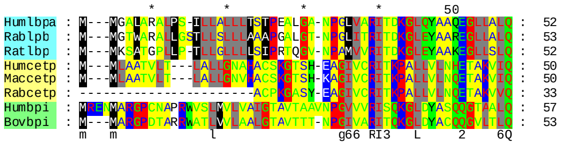
```
                 *         *         *         *        50
Humlbpa : M---MGALARALPS-ILLALLLTSTPEALGA-NPGLVARITDKGLQYAAQEGLLALQ :  52
Rablpb  : M---MGTWARALLGSTLLSLLLAAAPGALGT-NPGLITRITDKGLEYAAREGLLALQ :  53
Ratlbp  : M---MKSATGPLLP-TLLGLLLLSIPRTQGV-NPAMVVRITDKGLEYAAKEGLLSLQ :  52
Humcetp : M---MLAATVLT---LALLGNAHACSKGTSH-EAGIVCRITKPALLVLNHETAKVIQ :  50
Maccetp : M---MLAATVLT---LALLGNVHACSKGTSH-KAGIVCRITKPALLVLNQETAKVIQ :  50
Rabcetp : ------------------------ACPKGASY-EAGIVCRITKPALLVLNQETAKVVQ :  33
Humbpi  : MRENMARGPCNAPRWVSLMVLVAIGTAVTAAVNPGVVVRISQKGLDYASQQGTAALQ :  57
Bovbpi  : M---MARGPDTARRWATLVVLAALGTAVTTT-NPGIVARITQKGLDYACQQGVLTLQ :  53
          m   m        l                 g66 RI3   L     2    6Q
```

Identity



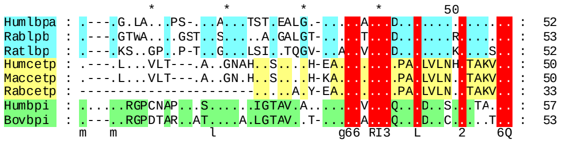
Summary view

Alignment visualization



```
                 *         *         *         *        50
Humlbpa : M---MGALARALPS-ILLALLLTSTPEALGA-NPGLVARITDKGLQYAAQEGLLALQ :  52
Rablpb  : M---MGTWARALLGSTLLSLLLAAAPGALGT-NPGLITRITDKGLEYAAREGLLALQ :  53
Ratlbp  : M---MKSATGPLLP-TLLGLLLLSIPRTQGV-NPAMVVRITDKGLEYAAKEGLLSLQ :  52
Humcetp : M---MLAATVLT---LALLGNAHACSKGTSH-EAGIVCRITKPALLVLNHETAKVIQ :  50
Maccetp : M---MLAATVLT---LALLGNVHACSKGTSH-KAGIVCRITKPALLVLNQETAKVIQ :  50
Rabcetp : ------------------------ACPKGASY-EAGIVCRITKPALLVLNQETAKVVQ :  33
Humbpi  : MRENMARGPCNAPRWVSLMVLVAIGTAVTAAVNPGVVVRISQKGLDYASQQGTAALQ :  57
Bovbpi  : M---MARGPDTARRWATLVVLAALGTAVTTT-NPGIVARITQKGLDYACQQGVLTLQ :  53
          m   m        l                 g66 RI3   L     2    6Q
```

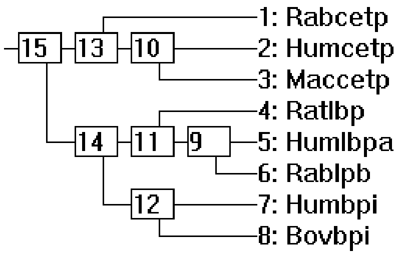
Physico-chemical properties



```
                 *         *         *         *        50
Humlbpa : .---.G.LA...PS-....A...TST.EALG.-.......A....D..........  :  52
Rablpb  : .---.GTWA....GST..S.....A.GALGT-.......T....D.......R.....  :  53
Ratlbp  : .---.KS..GP..P-T..G...LSI..TQGV-..A..V....D.......K....S.. :  52
Humcetp : .---.L...VLT---.A..GNAH..S....H-EA.........PA.LVLNH.TAKV.. :  50
Maccetp : .---.L...VLT---.A..GN.H..S....H-KA.........PA.LVLN..TAKV.. :  50
Rabcetp : -----------------------.......A.Y-EA.........PA.LVLN..TAKV.. :  33
Humbpi  : ........RGPCNAP...S......IGTAV.A........V....Q...D..S...TA... :  57
Bovbpi  : .---..RGPDTAR..AT....A.LGTAV..T-.......A....Q...D..C.....T.. :  53
          m   m        l                 g66 RI3   L     2    6Q
```

Differences mode

Alignment visualization (tree)



Alignment statistics

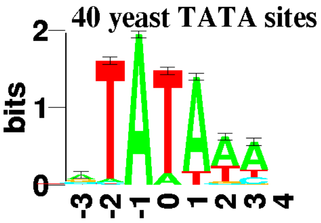
```
                  Rablpb          Humcetp          Rabcetp          Bovbpi
          Humlbpa          Ratlbp          Maccetp          Humbpi
                1        2        3        4        5        6        7        8

              478      67%      65%      19%      19%      18%      42%      43%
1               0      82%      80%      39%      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%

               96       84       95      494      95%      74%      20%      21%
4             198      192      194        0      98%      84%      40%      41%
               30       29       28        0       0%       7%       6%       5%
```

Alignment score

```
             Rablpb            Humcetp            Rabcetp            Bovbpi
   Humlbpa           Ratlbp            Maccetp            Humbpi
         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
```

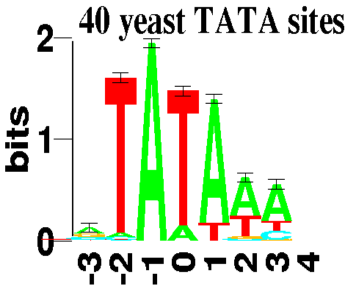
Sequence Logos:

a quantitative graphical display for binding sites and proteins

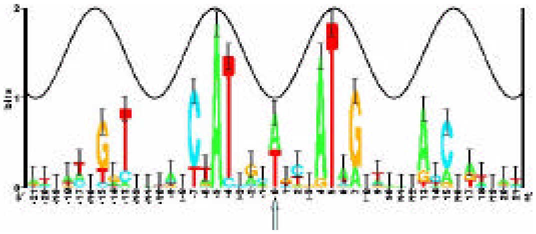


Reference: Schneider, T.D. *Meth. Enzym* **274**:445, 1996

Sequence Logos



Sequence Logos



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)