Multiple Sequence Alignment

Inge Jonassen
Bioinformatics group
Dept. of Informatics, UoB
Inge.Jonassen@ii.uib.no

Overview

- Definition/examples
- Use of alignments
- The alignment problem
  - scoring alignments
  - finding good alignments
- Alignment algorithms
- Local alignment methods and Pattern discovery
- Conclusion

Definition

- A global alignment of a set of sequences is obtained by
  - inserting into each sequence gap characters '-'
- so that
  - the resulting sequences are of the same length
- and so that
  - no "column" has only gap characters

Example

Take the sequences

```
INDUSTRY
INTERESTING
IMPORTANT
```

One alignment is

```
IN-DU-STRY-
INTERESTING
IM-POR-TANT
```

Example

This is *not* an alignment:

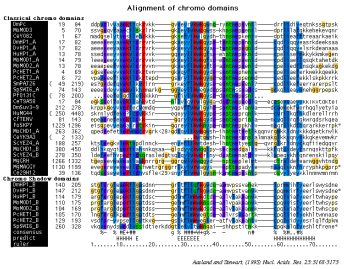
```
IN-DU--STRY-
INTERE-STING
IM-POR--TANT
```

Example

This is *not* an alignment:

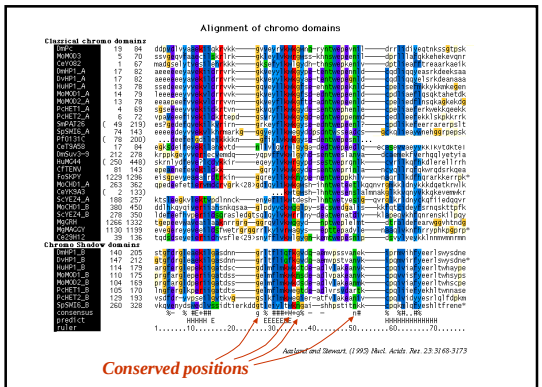
```
IN-DU--STRY-
INTERE-STING
IM-POR--TANT
```

Example: Chromo domains aligned



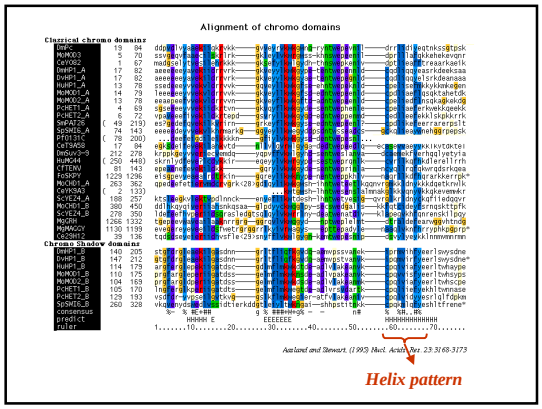
Use of alignments

- Predict features of aligned objects
  - conserved positions
    - structurally/functionally important



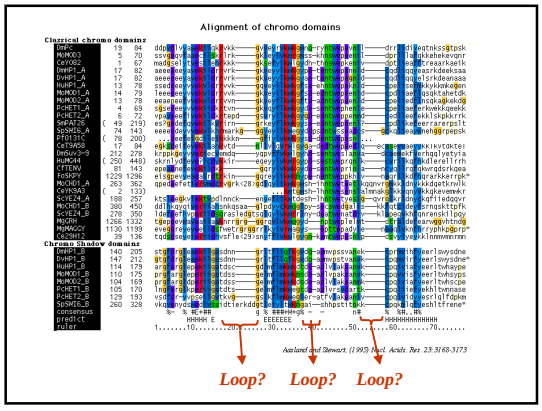
Use of alignments

- Predict features of aligned objects
  - conserved positions
    - structurally/functionally important
  - patterns of hydrophobicity/hydrophilicity
    - secondary structure elements



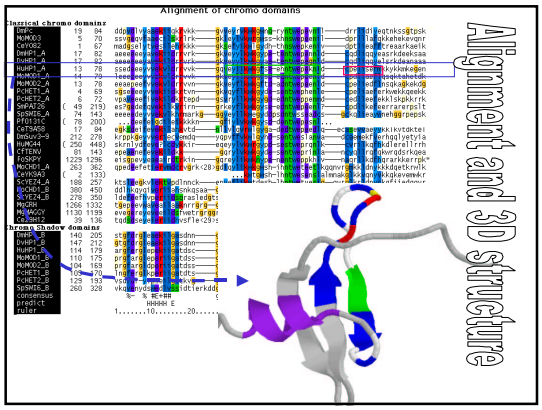
Use of alignments

- Predict features of aligned objects
  - conserved positions
    - structurally/functionally important
  - patterns of hydrophobicity/hydrophilicity
    - secondary structure elements
  - “gappy” regions
    - loops/variable regions



Use of alignments

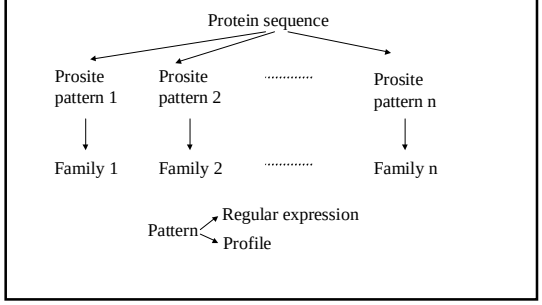
- Predict features of aligned objects
  - conserved positions
    - structurally/functionally important
  - patterns of hydrophobicity/hydrophilicity
    - secondary structure elements
  - "gappy" regions
    - loops/variable regions
  - covariation
    - structural proximity



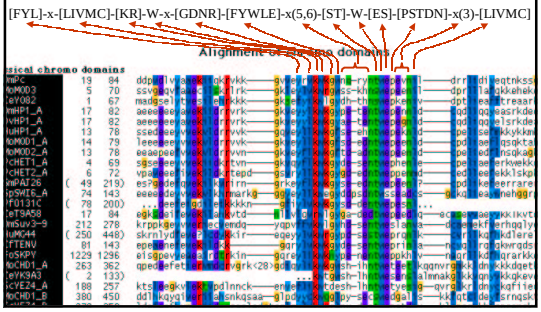
Use of Alignments - make patterns/profiles

- Can make a *profile* or a *pattern* that can be used to match against a sequence database and identify *new family members*
- Profiles/patterns can be used to predict family membership of *new* sequences
- *Databases* of profiles/patterns
  - PROSITE
  - PFAM
  - PRINTS
  - ...

Prosite: Motifs for classification



Pattern from alignment



Alignment problem

Given a set of sequences, produce a multiple alignment which corresponds as well as possible to the biological relationships between the corresponding bio-molecules

For homologous proteins

- *Two residues* should be aligned (on top of each other)
  - if they are *homologous* (evolved from the same residue in a common ancestor protein)
  - if they are *structurally equivalent*

Automatic approach

- Need a way of scoring alignments
  - *fitness function* which for an alignment quantifies its "goodness"
- Need an *algorithm* for finding alignments with good scores
- *Not all methods provide a scoring function for the final alignment!*

Analysis of fitness function

- One can test whether the alignments optimal under a given fitness function correspond well to the biological relationships between the sequences
- For example, if the structure of (some of) the proteins are known.

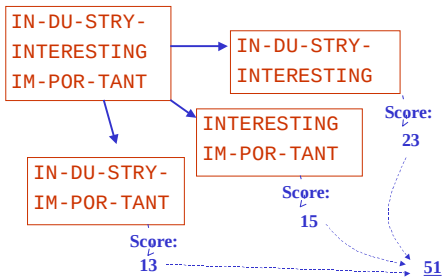
Alignment scores

- We can define the score of an alignment of two sequences
  - uses a scoring matrix (e.g., PAM, BLOSUM)
  - gap penalty (linear, affine)

Alignment scores: SP - sum-of-pairs

- A multiple alignment "implies" a pairwise alignment for each pair of sequences
- SP defines the score of the multiple alignment as the sum of scores of all implied pairwise alignments.

SP - example



SP - definition

- If $A_{i,j}$ is the score of the alignment implied for sequence pair *(i,j)*, then the total score is:

$$SP = \sum_{i,j} A_{i,j}$$

WSP - definition

- It is often useful to weight the sequence pairs

$$WSP = \sum_{i,j} w_{i,j} \cdot A_{i,j}$$

- There may be strong biases in the sequence set (e.g., a large number of nearly identical sequences - pairs including one of these can be given low weights to reduce their impact on the score)

Tree Alignment

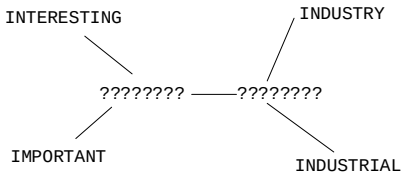
- It is assumed that an evolutionary tree for the sequences is known
- The sequences are leaves in the tree

Tree Alignment

- Problem:
  - assign sequences to interior nodes
    - scores can now be calculated for all edges in the tree

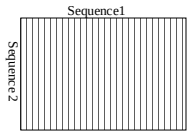
  so that
  - the score summed over all edges is maximal
- The sequence assignments giving the best score defines the best alignment according to this measure and for the given tree.

Tree alignment - example



Alignment Algorithms

- Given a set of *n* sequences of average length *l*, find a "good" alignment!
- For *n=2*, we have seen that dynamic programming can be used - time taken is proportional to $l^2 = l^n$



Dynamic programming for *n* sequences

- Assume we have *n* sequences of length *l*
- The table will have l^n entries
- For example, 10 sequences of length 100 gives
  - a table with 10^{20} entries which would take at least 100 million Terrabytes (one byte per entry) of memory
  - which would take about 3 million years to fill in if 1 million entries can be computed per second
- Not feasible for n>4 or 5

Progressive alignment

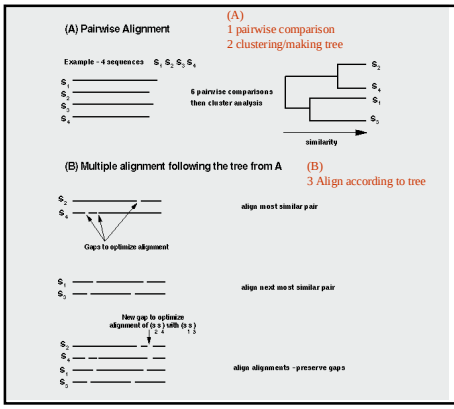
- **Observations:**
  - Align two sequences at a time - can be done using dynamic programming
  - The output of each pairwise alignment is an alignment
  - Pairs of alignment/alignment or alignment/sequence can be aligned - using dynamic programming

Progressive alignment

- **Strategy:**
  - Align first the most similar sequences
  - Progressively align more distant sequences until all sequences have been aligned
  - Use a rooted tree with the sequences at the leaves to decide the order of the alignments

The Clustal Algorithm

- Three steps:
  1. Compare all pairs of sequences to obtain a similarity matrix
  2. Based on the similarity matrix, make a guide tree relating all the sequences
  3. Perform progressive alignment where the order of the alignments is determined by the guide tree



ClustalW - Score of aligning two alignment columns

- sum the score matrix entry for all pairs of residues
- weight each pair by the sequences' weights

```
1:peeksavtal
2:geekaavlal

3:egewglvlhv
4:aaektkirsa
```

Score:
M(t,v)+M(t,i)+
M(l,v)+M(l,i)

ClustalW - Weighting sequences

- each sequence is given a weight
- groups of related sequences receive lower weight

```
1:peeksavtal
2:geekaavlal

3:egewglvlhv
4:aaektkirsa
```

Weighted score:
w1*w3*M(t,v)+
w1*s4*M(t,i)+
w2*w3*M(l,v)+
w2*w4*M(l,i)

ClustalW - Similarity matrix

- Distance between sequences - measure from the guide tree - determines which matrix to use
  - 80-100% seq-id -> use Blosum80
  - 60-80% seq-id -> Blosum60
  - 30-60% seq-id -> Blosum45
  - 0-30% seq-id -> Blosum30

ClustalW - Gap penalties

- Initial gap penalty
  - GOP
- Gap extension penalty
  - GEP

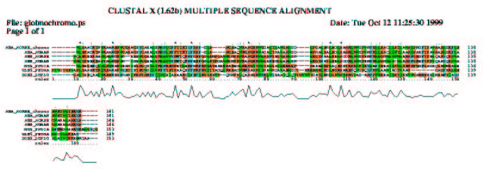
```
GTEAKLIVLMANE
GA---------KL
```

Penalty: GOP+8*GEP

ClustalW - Modifications of gap penalty

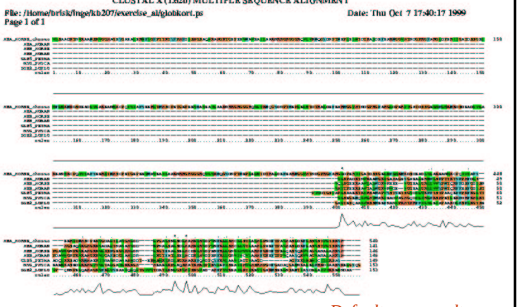
- Position specific penalty
  - gap at position
    - yes -> lower GOP
    - no, but gap within 8 residues -> increase GOP
  - hydrophilic residues
    - lower GOP

Globin alignment



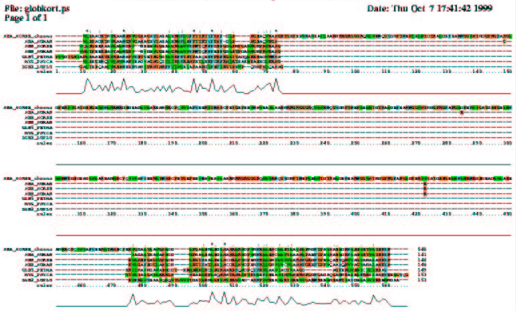
Default gap penalty
GEP=0.05

Globin alignment - with insert



Default gap penalty
GEP=0.05

Globin alignment - with insert



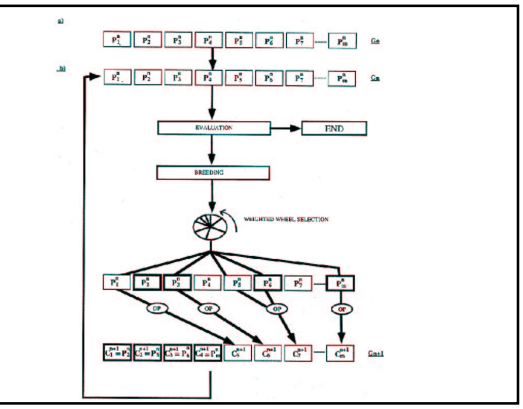
Lowered gap penalty

ClustalW - summary

- Does not use a score for the final alignment
- Each pairwise alignment is done using dynamic programming
- Heuristics (e.g., gap-penalty modifications) are used - tailored to globular proteins
- Graphical version: ClustalX

SAGA: Sequence Alignment by Genetic Algorithm

- An “objective function” is used to score the alignments
- An alignment is represented as a bit string
- A population of alignment is “evolved”
- Alignments can be combined (cross-over)
- Alignments can be mutated
- Alignments with higher score are more likely to be chosen for mating/survival



Local Multiple Alignment

- Take one (zero/several) segment(s) (fragment) from each sequence and align them
  - maximise similarity of aligned fragments
  - most methods do not allow for gaps in the local alignment
- Example method: MEME

MEME - Motif Elucidation by Multiple EM

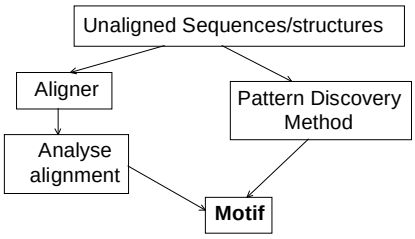
- EM= Expectation Maximisation
- Statistical method
- Builds a model of the local alignment
- Iteratively
  - refines the model
  - realigns the sequences to the model

Example MEME output

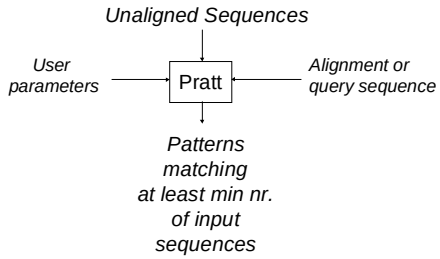
```
--------------------------------------------------------------------------------
   Possible examples of motif 1 in the training set
--------------------------------------------------------------------------------
Sequence name            Start  Score            Site
------------             -----  -----            ---------
2BHD_STREX                  81  28.80 VAYAREEFGS VDGLVNNAG ISTGMFLETE
3BHD_COMTE                  81  25.99 MAAVQRRLGT LNVLVNNAG ILLPGDMETG
ADH_DROME                   86  22.33 LKTIFAQLKT VDVLINGAG ILDDHQIERT
AP27_MOUSE                  77  24.36 TEKALGGIGP VDLLVNNAA LVIMQPFLEV
BA72_EUBSP                  86  26.39 VGQVAQKYGR LDVMINNAG ITSNNVFSRV
BDH_HUMAN                  138  23.46 PFEPEGPEKG MWGLVNNAG ISTFGEVEFT
BPHB_PSEPS                  79  18.60 ASRCVARFGK IDTLIPNAG IWDYSTALVD
BUDC_KLETE                  80  20.97 VEQARKALGG FNVIVNNAG IAPSTPIESI
DHES_HUMAN                  84  25.67 AARERVTEGR VDVLVCNAG LGLLGPLEAL
DHGB_BACME                  87  26.39 VQSAIKEFGK LDVMINNAG MENPVSSHEM
DHMA_FLAS1                 198  16.36 ILVNMIAPGP VDVTGNNTG YSEPRLAEQV
ENTA_ECOLI                  73  21.90 CQRLLAETER LDALVNAAG ILRMGATDQL
FIXR_BRAJA                 112  23.67 EVKKRLAGAP LHALVNNAG VSPKTPTGDR
GUTD_ECOLI                  82  17.17 SRGVDEIFGR VDLLVYSAG IAKAAFISDF
HDE_CANTR                   92  20.90 VETAVKNFGT VHVIINNAG ILRDASMKKM

...
```

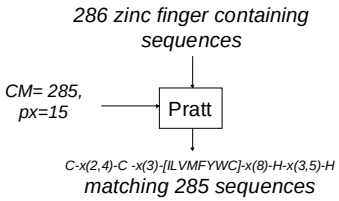
Motif Discovery



Pratt - functionality



Pratt - Example



Evaluation of Alignment Methods

- Align set of protein sequences where the structures are known (at least for some proteins)
- Align the protein structures
- Identify “motifs” from the structure alignment
- Check if sequence alignment has correctly aligned motifs

- McClure et al, 1994
- Thompson et al, 1999

Alignments are important

- Basis for other analyses
  - structure prediction
  - phylogeny
  - experiments
    - PCR primer identification
    - site directed mutagenesis
    - ...
  - identification of motifs

Open Problems - space for improvements!

- Good scoring function for alignments
  - identify well aligned regions
- Efficient algorithms
- Resolving repeat structure, domain movements etc.
- Incorporating external information

Future development

- More sequences
  - More families, but not so many
  - More densely populated families
  - “Easier” alignment problem
  - Identify more ancient relationships (superfamilies)
- More structures
  - more sequences can be “threaded”
  - alignments help