Multiple sequence alignment 2

Optimizing progressive multiple alignment methods

Sequence analysis 2006

Lecture 7 - 20/11/06
Pair-wise alignment quality \textit{versus} sequence identity

- Vogt et al., JMB 249, 816-831, 1995
Remember the **greediness** of the progressive alignment algorithm?
Strategies for progressive alignment optimization

- Heuristics
- Profile pre-processing
- Secondary structure-guided alignment
- Globalised local alignment
- Matrix extension

**Objective**: try to avoid (early) errors
Clustal, ClustalW, ClustalX
Higgins et al. 1994

- Probably most well-known progressive alignment program
- Neighbour Joining (NJ) algorithm (Saitou and Nei, 1984) to construct guide tree
- NJ does not require that all lineages have diverged by equal amounts
- Sequence blocks are represented by profiles
- Individual sequences are additionally weighted according to the branch lengths in the NJ tree
Clustal, ClustalW, ClustalX

• Further carefully crafted heuristics include:
  • local gap penalties
  • automatic selection of the amino acid substitution matrix
  • automatic gap penalty adjustment
  • mechanism to delay alignment of sequences that appear to be distant at the time they are considered.

• **Clustal** (W/X) does not allow iteration
  (Hogeweg and Hesper, 1984; Corpet, 1988, Gotoh, 1996; Heringa, 1999, 2002)

• Still limited by “greedy” nature of progressive algorithm
ClustalW web-interface

ClustalW is a general purpose multiple sequence alignment program for DNA or proteins. It produces biologically meaningful multiple sequence alignments of divergent sequences. It calculates the best match for the selected sequences, and lists them up so that the identities, similarities and differences can be seen. Evolutionary relationships can be seen via viewing Cladograms or Phylogenies. New users, please read the FAQ.

Protein structure hierarchical levels

(a) Primary structure
- Ala–Glu–Val–Thr–Asp–Pro–Gly–

(b) Secondary structure
- α helix
- β sheet

(c) Tertiary structure (fold)

(d) Quartenary structure (oligomers)
Flavodoxin fold: helix-beta-helix
Flavodoxin cheY NJ tree
<table>
<thead>
<tr>
<th>Sequence</th>
<th>Alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1fx1</td>
<td>PKALIVYGSTGNTETAYTAQSRANAG-Y-EVDSRDASAASVEAGLFLFDVFLLGLGCSTWWGDSIE-------LQDDFIPLFD-SLEETGAGQRK</td>
</tr>
<tr>
<td>FLAV_DESVH</td>
<td>MPKALIVYGSTGNTETAYTAIARELADAG-Y-EVDSRDASAASVEAGLFLFDVFLLGLGCSTWWGDSIE-------LQDDFIPLFD-SLEETGAGQRK</td>
</tr>
<tr>
<td>FLAV_DESGI</td>
<td>MPKALIVYGSTGNTETGAIAAKTNSEG-M-ETTVNYADVTAPGMAEGYDVDVGGCSTWGGDEIE -------LQEDFVPLYE-DLLRAYLAAK</td>
</tr>
<tr>
<td>FLAV_DESSA</td>
<td>MSKALIVFGGTSGTNESCIAQKLEELIAAGG-H-EVTLILAAADAAENLADGYAVLFGCSCWMEDEIE---MQDFLSILFE-EFRFLAGRRK</td>
</tr>
<tr>
<td>FLAV_CLOAB</td>
<td>-MKSILYYSKTGTAKVAKLIEEGVRGSM-EVTVKMLDADVKVKFLDK-------ISWEMASKWID-ESSFNELEK</td>
</tr>
<tr>
<td>FLAV_MEGEL</td>
<td>--MVEIVYWSGTNGTEAMANIAEANRAAG-A-DVESRFETDNNVDDVAS-KDVILLGCPAMGE--E-------LEDVSVEPPFD-TDLAPKLK</td>
</tr>
</tbody>
</table>
T-COFFEE
Notredame, Higgins and Heringa 2000

• Performs progressive alignment as in Clustal
• But there are many differences …
  • Integrating different pair-wise alignment techniques (NW, SW, ..)
  • Combining different multiple alignment methods (consensus multiple alignment)
  • Combining sequence alignment methods with structural alignment techniques
  • Plug in user knowledge
T-COFFEE
Notredame, Higgins and Heringa 2000

• **Step 1:**
  - Global alignment: Clustal
  - Local alignment: Lalign (pick top ten)

• **Step 2:**
  - Put collection of local & global in a library

• **Step 3:**
  - Extend the library by aligning each pairwise alignment with a possible third sequence
    *(Search matrix or library extension)*
Search matrix extension
Search matrix extension

a) Signal Addition: Computation of the primary weights.

<table>
<thead>
<tr>
<th>Primary Library</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Res. 1 Res. 2</td>
<td>Weight</td>
<td></td>
</tr>
<tr>
<td>A(6) B(3)</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>A(6) B(14)</td>
<td>25</td>
<td></td>
</tr>
</tbody>
</table>

b) Extension: Alignment of A and B through sequence C.

<table>
<thead>
<tr>
<th>Res. 1 Res. 2</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>A(6) B(2)</td>
<td>10 mm(10,15)</td>
</tr>
<tr>
<td>A(6) B(6)</td>
<td>10 mm(10,20)</td>
</tr>
<tr>
<td>A(6) B(14)</td>
<td>35 mm(40,55)</td>
</tr>
<tr>
<td>A(6) B(19)</td>
<td>20 mm(40,20)</td>
</tr>
</tbody>
</table>

C) Combination: Position Specific Scoring Scheme for A and B using information from sequence A, B and C.
Search matrix extension

a) Regular Progressive Alignment Strategy

b) Primary Library

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Prim. Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>SeqA GARFIELD THE LAST P A T C A T</td>
<td>88</td>
</tr>
<tr>
<td>SeqB GARFIELD THE P A S T C A T</td>
<td>77</td>
</tr>
<tr>
<td>SeqC GARFIELD THE V E R Y F A S T C A T</td>
<td>100</td>
</tr>
<tr>
<td>SeqD ----- THE ----- P A - T C A T</td>
<td>100</td>
</tr>
</tbody>
</table>

Extended Library for seq1 and seq2

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>SeqA GARFIELD THE LAST P A T C A T</td>
<td>88</td>
</tr>
<tr>
<td>SeqB GARFIELD THE P A S T C A T</td>
<td>77</td>
</tr>
<tr>
<td>SeqC GARFIELD THE V E R Y F A S T C A T</td>
<td>100</td>
</tr>
<tr>
<td>SeqD ----- THE ----- P A - T C A T</td>
<td>100</td>
</tr>
</tbody>
</table>

Dynamic Programming

Extended Library

- SeqA GARFIELD THE LAST P A T C A T
- SeqB GARFIELD THE P A S T C A T
- SeqC GARFIELD THE V E R Y F A S T C A T
- SeqD ----- THE ----- P A - T C A T

So for T-COFFEE ...

• Original pairwise alignments are refined based on the consistency with other “intermediate” sequences

• Distance matrix and guide tree are built based on the refined alignments
The T-COFFEE effect

- Although a direct alignment might choose one path, using information from the other sequences (T-COFFEE) finds an alternate and usually better one.

- Minimize errors in the early stages of alignment assembly.

- Extra computation time for the calculation of the consistency scores.
but ... **T-COFFEE**(v1.23) Flavodoxin-cheY

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1fx1</td>
<td>PKALIVYGSTTGTNTEYTAETIARQLANAG-YEVSREAASVE-AGGLFEGFDLVLLGCSTWGDDEIE-LQDDFIPL-FDSLEEETGQAQGRK</td>
</tr>
<tr>
<td>FLAV_DESVH</td>
<td>MPKALIVYGSTTGTNTEYTAETIARQLANAG-YEVSREAASVE-AGGLFEGFDLVLLGCSTWGDDEIE-LQDDFIPL-FDSLEEETGQAQGRK</td>
</tr>
<tr>
<td>FLAV_DESGI</td>
<td>MPKALIVYGSTTGTNTEVAIAKTSLNGEM-METTVNVAODV-APGLAEYDVVLGCSTWGDDEIE-LQDDFIPL-FDSLEEETGQAQGRK</td>
</tr>
<tr>
<td>FLAV_DESSA</td>
<td>MSKLVIFGSGSTGTAEIAQKLEELIAAGG-HEVTLNADAS-AENLADGVDVLGCSTWGDDEIE-MQDDFLSL-FEEFNRFLGVRK</td>
</tr>
<tr>
<td>4fxn</td>
<td>MNIYVWSTQGNTKAEIAKGIIEEG-KDVNTVSDVN-DDEE-NDELIILGCSTWGDDEIE-ESEFEPF-IEEIS-TKISGK</td>
</tr>
<tr>
<td>FLAV_MEGEL</td>
<td>MSEIIYWSGTNTEMAEIAIAAVKAG-ADVESRVFEDT-VDVA-SKAVILCPSMGSEEE-DSEVFEPF-FDYOFL-KPKGK</td>
</tr>
<tr>
<td>FLAV_CLOAB</td>
<td>MKEISLYSKSTGKTERRVAKLIEEGKRSNIKVEKMLADV-KKFLIQ-ESEGIIIFGTPTYAN-ISMWEIKW-IDSSENFLGKL</td>
</tr>
<tr>
<td>2fcr</td>
<td>KIGIFFSTSGNTTEVADFIGTLGAKA-DAPIDVDDVDTPQAAL-KDYLFLGAPTWNNTG-IDTERSGTSNWDELFDLPEVDKMLP</td>
</tr>
<tr>
<td>FLAV_ENTAG</td>
<td>MATIGIFGSQDTGQTRKVAHKLQKDLDAG-AAPLLIGAVGDPEA-NGTSQEPF-NTNLSEADLTGK</td>
</tr>
<tr>
<td>FLAV_ANASP</td>
<td>SKKIGLFYQTFQTEASVAILREDGFN-METTVVNVADVT-APGLAEGYDVVLGCSTWGDDEIE-LQEDFLVL-FFEFFFFFLGVRK</td>
</tr>
<tr>
<td>FLAV_AZOVI</td>
<td>AKIGLFFGSNTKVRKAKQPST-REDLVLAAV-EDAYLDFGELPGLSGCENESWEEF-LEPK1EDLDFGVRK</td>
</tr>
<tr>
<td>FLAV_ECOLI</td>
<td>MSKSLIVYQSTTGTNTEAAPEYVAEAFKENG-HEVTVNVIADVT-APGLAEGYDVVLGCSTWGDDEIE-LQEDFLVL-FFEFFFFGLGVRK</td>
</tr>
<tr>
<td>3chy</td>
<td>ADKELKFLVD-DFSTMRRIVRNLKKEQLFN-NEV-AEDGVDNALKQL-AGQYVFYTSDWMPNMNDG</td>
</tr>
</tbody>
</table>

T-COFFEE web-interface
3D-COFFEE

- Computes structural based alignments
- Structures related to the sequences are retrieved
- More accurate … but for many (many) proteins we do not have the structure!
PRALINE
Heringa 1999

• Again a progressive alignment method
• Like in T-COFFEE the idea is minimise error-propagation by including prior knowledge about the other sequences when making pairwise alignments

• **PRALINE** does this by making **pre-profiles**: the sequences to be aligned are represented by pre-profiles instead of single sequences
Pre-profile generation

Pre-alignments

Score 1-2
Score 1-3
Score 4-5

Cut-off

Pre-profiles

## PRALINE Flavodoxin-cheY

### Pre-processing (cut-off ≥1500)

| 1fx1      | PKALIVYGSTGNT-EYTAETIARQLANAG-YEVSDRDAASVEAGGLFEGFDLVLLGCSTWDSSI-ELQDDFIPLF-DSLEEETGAGQRKVACF |
| FLAV_DESCD | MSDLIVYGFSTGNT-ETLqKLEELLAAAGG-HEVTLINADASAENLAGYDAVLFYGCSAMGMDL-EMQDDFLSLF-EERFNGFLAGKRVKAF |
| FLAV_DESCDH | MPKALIVYGSTGNT-EYTAETIARELADAG-YEVSDRDAASVEAGGLFEGFDLVLLGCSTWDSSI-ELQDDFIPLF-DSLEEETGAGQRKVACF |
| FLAV_DESCDI | MSDLIVYGFSTGNT-EYTAETIARELADAG-YEVSDRDAASVEAGGLFEGFDLVLLGCSTWDSSI-ELQDDFIPLF-DSLEEETGAGQRKVACF |
| 2fcr      | --KIGIFFTSTGNT-TEVADFKTGLA--KADAPIVDDVTDQALKLMYLFLLGAPTWNAG--ADTERTGSENDEFLYLKDPEVMKDLPVAVIF |
| FLAV_AZOVI | -AKIGLFFGSGTGRKVEIKKRFDDEKMSDA-IINNKRVS-ADDFAQYQYFLIGTFILGEGFEHGELSICENCEEFSFL-PKIEGLDFSGKTVLAF |
| FLAV_ENTAG | MATIGIFFGSDTQGTKAKVAEKHAELKSADVLMHADLVADLDVQ---LLNFSDQYKAWQKAKLAPLM |
| FLAV_ANASP | SKKIGLFYGTQGKTSKRAESFSLIKQLEQGKVQ---LLTVVLAEDVFLLGPTWYYGEL---LSDWEDFLY-DKSLKGKKVGAf |
| FLAV_ECOLI | -AITGIFFGSDTGNT-ENIKMIQKQLGK---DVADVHDIAEKSS-KEDLEAYIYLLNLPTWYYGE--------AQCDWDDFF-PTLEEIDFNGKLVALf |
| 4fxn      | GLGDAEGYPDNFCDA-LGELYSFFKDRgAGTIVGNSLGFYSTDGWDFNDSQKARNGKFVGLALDEGPASELTSDDRIKWSAQLISFISF |
| FLAV_MEGEL | MVE--IVYYSGTGNT-EAMaELIHEEGVRSGNIEVKTMNLDAVDVKNFLQESEGIIFgTPTYYAN---------IVSEMKWIKI-DSEFNLGKAVAF |
| FLAV_CLOAB | -MKISILYSSKTGRT-EKVAELIGEGVRSNEIVMDMDV---KDQIGYADNSQP--ECFQLQGRWSAEFLV-GLFCDKCVLAf |
| 3chy      | ADKELKFLVVDFTSTRMRRVLNLKELGFN--NEEAEEDGVIDAIKNIQLAGGGYVFI--SDWNPMP---DGLELL-KTIRADQAMSAPVLM |

---

Using structural information

- Structure is more conserved than sequence

- CASP (Critical Assessment of Techniques for Protein Structure Prediction)

- Tertiary structure prediction: *de novo* prediction is not accurate at all! (especially for larger folds)
Using secondary structure

- 10 years SS prediction method development:
  Accuracy $\pm 5\%$
- 10 years MSA method development:
  Accuracy can be $\pm 40\%$

- Amino acid patterns
- Secondary structure patterns
- Super-secondary structure patterns
- Alternate matrices with associated gap penalties according to region
How to combine ss and aa info

Dynamic programming search matrix

Amino acid substitution matrices

[MDAGSTVILCFV]
[HHHCCCEEEEE]

H
C
E

MDAAS
H
H
H
H
C
C
C
C
E
E
E
E
C

Default
In terms of scoring ...

• So how would you score a profile using this extra information?
  • Same formula as in lecture 6, but you can use sec. Structure-specific substitution scores in various combinations.

• Where does it fit in?
  • Very important: structure is always more conserved than sequence so it can help with the insertion(or not) of gaps.
SS-PRALINE

Sequences to be aligned

Predict secondary structure or DSSP

HHHHCCCEECCCCCEEECCCHH
CCCCCCCCCCCCCCCCCCCCCHH
HHHHCCCEECCCCCEEEHHH
HHHHHCCCCEEEEECCCEEEEE
HHHHHHHHHHHHHHHCCCEEEE

Align sequences using secondary structure

Multiple alignment
Using predicted secondary structure

1fx1

- PKALIVYGSTTTAEAYTARQLANAGYEVDSRGLFVLLGCSTGWDDISILQDDFIPLFDSLEETGAQGRKVAFC

FLAV DESVH MPK-ALIVYGSTTTAEAYTARTQALANAGYEVDSRGLFVLLGCSTGWDDISILQDDFIPLFDSLEETGAQGRKVAFC

FLAV DESGI MPRALIVYGSTTTAEAYTARTQALANAGYEVDSRGLFVLLGCSTGWDDISILQDDFIPLFDSLEETGAQGRKVAFC

FLAV DESSA MKSLIVYGSTTTAEAYTARTQALANAGYEVDSRGLFVLLGCSTGWDDISILQDDFIPLFDSLEETGAQGRKVAFC

FLAV DESDE MKVSLIVYGSTTTAEAYTARTQALANAGYEVDSRGLFVLLGCSTGWDDISILQDDFIPLFDSLEETGAQGRKVAFC

2fcr

- PKALIVYGSTTTAEAYTARQALANAGYEVDSRGLFVLLGCSTGWDDISILQDDFIPLFDSLEETGAQGRKVAFC

FLAV DESVH MPKALIVYGSTTTAEAYTARTQALANAGYEVDSRGLFVLLGCSTGWDDISILQDDFIPLFDSLEETGAQGRKVAFC

FLAV DESGI MPRLIVYGSTTTAEAYTARTQALANAGYEVDSRGLFVLLGCSTGWDDISILQDDFIPLFDSLEETGAQGRKVAFC

FLAV DESSA MKSLIVYGSTTTAEAYTARTQALANAGYEVDSRGLFVLLGCSTGWDDISILQDDFIPLFDSLEETGAQGRKVAFC

FLAV DESDE MKVSLIVYGSTTTAEAYTARTQALANAGYEVDSRGLFVLLGCSTGWDDISILQDDFIPLFDSLEETGAQGRKVAFC

3chy

- PKALIVYGSTTTAEAYTARQALANAGYEVDSRGLFVLLGCSTGWDDISILQDDFIPLFDSLEETGAQGRKVAFC

FLAV DESVH MPKALIVYGSTTTAEAYTARTQALANAGYEVDSRGLFVLLGCSTGWDDISILQDDFIPLFDSLEETGAQGRKVAFC

FLAV DESGI MPRLIVYGSTTTAEAYTARTQALANAGYEVDSRGLFVLLGCSTGWDDISILQDDFIPLFDSLEETGAQGRKVAFC

FLAV DESSA MKSLIVYGSTTTAEAYTARTQALANAGYEVDSRGLFVLLGCSTGWDDISILQDDFIPLFDSLEETGAQGRKVAFC

FLAV DESDE MKVSLIVYGSTTTAEAYTARTQALANAGYEVDSRGLFVLLGCSTGWDDISILQDDFIPLFDSLEETGAQGRKVAFC

4fxn

- PKALIVYGSTTTAEAYTARQALANAGYEVDSRGLFVLLGCSTGWDDISILQDDFIPLFDSLEETGAQGRKVAFC

FLAV DESVH MPKALIVYGSTTTAEAYTARTQALANAGYEVDSRGLFVLLGCSTGWDDISILQDDFIPLFDSLEETGAQGRKVAFC

FLAV DESGI MPRLIVYGSTTTAEAYTARTQALANAGYEVDSRGLFVLLGCSTGWDDISILQDDFIPLFDSLEETGAQGRKVAFC

FLAV DESSA MKSLIVYGSTTTAEAYTARTQALANAGYEVDSRGLFVLLGCSTGWDDISILQDDFIPLFDSLEETGAQGRKVAFC

FLAV DESDE MKVSLIVYGSTTTAEAYTARTQALANAGYEVDSRGLFVLLGCSTGWDDISILQDDFIPLFDSLEETGAQGRKVAFC

8000

**Sequence analysis 2006**
PRALINE web-interface

PRALINE multiple sequence alignment

[Image]

Enter a name for your job

Options

Exchange weights metric:

BLAST-like Help

Global progressive alignment strategy:

- Standard progressive strategy
- Pre-profile local processing
- Pre-profile local processing
- PSIBLAST pre-profile processing + secondary structure information + extension

Secondary structure prediction

DSSP-defined secondary structure search

Tree representation of the final alignment

Customize alignment representation colours

Final alignment file format

[More options]

E-mail

Submit

[Options, Help]

Advanced Interface

References and FAQs

PRALINE is a multiple sequence alignment program with many options to optimize the information for each of the input sequences, e.g., global or local progressive alignment, secondary structure information, and iteration capabilities.
Multiple alignment methods

- Multi-dimensional dynamic programming
- Progressive alignment
- **Iterative alignment**
  - correct for problems with progressive alignment by repeatedly realigning subgroups of sequence

From here on refer to the PRALINE paper for help
(further reading section online)
Iterative alignment

• **Idea**: an optimal might be found by repeatedly modifying existing suboptimal solutions

• Start with producing low-quality alignment and gradually improve by iterative refinement

• Continue until no more improvements in the alignment scores can be achieved
Iteration fates

- Convergence
- Limit cycle
- Divergence
Consistency scoring
Consistency iteration

Pre-profiles

Multiple alignment positional consistency scores

Multiple sequence alignment 2 – Sequence analysis 2006
Pre-profile update iteration

Pre-profiles

Multiple alignment
Is the initial SS prediction good enough?

<table>
<thead>
<tr>
<th>3chy-AA SEQUENCE</th>
<th>AA</th>
<th>ADKELKFLVDFSTMRRIVRNLLKELGFNNVEEAEDGVDALNKLQAGGYGFVISDWNMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>3chy-ITERATION-0</td>
<td>PHD</td>
<td>EEEEEE EHHHHHHHHHHHHHHHH HH EHHHHH EEEEEE</td>
</tr>
<tr>
<td>3chy-ITERATION-1</td>
<td>PHD</td>
<td>EEEEEE EHHHHHHHHHHHHHHHH HH EEEEEE</td>
</tr>
<tr>
<td>3chy-ITERATION-2</td>
<td>PHD</td>
<td>EEEEEE EHHHHHHHHHHHHHHHH HH EEEEEE</td>
</tr>
<tr>
<td>3chy-ITERATION-3</td>
<td>PHD</td>
<td>EEEEEE EHHHHHHHHHHHHHHHH HH EEEEEE</td>
</tr>
<tr>
<td>3chy-ITERATION-4</td>
<td>PHD</td>
<td>EEEEEE EHHHHHHHHHHHHHHHH HH EEEEEE</td>
</tr>
<tr>
<td>3chy-ITERATION-5</td>
<td>PHD</td>
<td>EEEEEE EHHHHHHHHHHHHHHHH HH EEEEEE</td>
</tr>
<tr>
<td>3chy-ITERATION-6</td>
<td>PHD</td>
<td>EEEEEE EHHHHHHHHHHHHHHHH HH EEEEEE</td>
</tr>
<tr>
<td>3chy-ITERATION-7</td>
<td>PHD</td>
<td>EEEEEE EHHHHHHHHHHHHHHHH HH EEEEEE</td>
</tr>
<tr>
<td>3chy-ITERATION-8</td>
<td>PHD</td>
<td>EEEEEE EHHHHHHHHHHHHHHHH HH EEEEEE</td>
</tr>
<tr>
<td>3chy-ITERATION-9</td>
<td>PHD</td>
<td>EEEEEE EHHHHHHHHHHHHHHHH HH EEEEEE</td>
</tr>
</tbody>
</table>

| 3chy-ITERATION-0 | PHD| HHHHHHEEEEEE HH HHHHHHHHHHHHHHHHH HEE EHHHHHHHHHHHHHH |
| 3chy-ITERATION-1 | PHD| HHHHHHEEEEEE HH HHHHHHHHHHHHHHHHH HEE EHHHHHHHHHHHHHH |
| 3chy-ITERATION-2 | PHD| HHHHHHEEEEEE HH HHHHHHHHHHHHHHHHH HEE EHHHHHHHHHHHHHH |
| 3chy-ITERATION-3 | PHD| HHHHHHEEEEEE HH HHHHHHHHHHHHHHHHH HEE EHHHHHHHHHHHHHH |
| 3chy-ITERATION-4 | PHD| HHHHHHEEEEEE HH HHHHHHHHHHHHHHHHH HEE EHHHHHHHHHHHHHH |
| 3chy-ITERATION-5 | PHD| HHHHHHEEEEEE HH HHHHHHHHHHHHHHHHH HEE EHHHHHHHHHHHHHH |
| 3chy-ITERATION-6 | PHD| HHHHHHEEEEEE HH HHHHHHHHHHHHHHHHH HEE EHHHHHHHHHHHHHH |
| 3chy-ITERATION-7 | PHD| HHHHHHEEEEEE HH HHHHHHHHHHHHHHHHH HEE EHHHHHHHHHHHHHH |
| 3chy-ITERATION-8 | PHD| HHHHHHEEEEEE HH HHHHHHHHHHHHHHHHH HEE EHHHHHHHHHHHHHH |
| 3chy-ITERATION-9 | PHD| HHHHHHEEEEEE HH HHHHHHHHHHHHHHHHH HEE EHHHHHHHHHHHHHH |
SS iteration

Sequences to be aligned

Align using secondary structure

Multiple alignment

Predict secondary structure

Secondary structure

Single Sequence  MA-based

HHHHHCCCCCCEEECCCEEEECHH
CCCCCCCCCCCCCCCCCCCCCCCCHH
HHHHCCCCCCCCCCCCCHHHH
HHHHCCCCCCCCCCCCCCCCCC
HHHHHHHHHHHHHHHHHHHHHHH
HHHHHCCCCCCCCCCCCCCCCCC
HHHHHCCCCCCCCCCCCCCC
HHHHHCCCCCCCCCCCC
HHHHHCCCCCCCCCC
HHHHHCCCCC
HHHHHCCCC
HHHHHCC
HHHHHCC
HHHHH
HHHH
HHH
HH
H

PRALINE and MUSCLE method

- PRALINE and MUSCLE use almost the same formalism to compare two profiles:

  - MUSCLE:  
    \[ LE^{xy} = (1 - f_x^G)(1 - f_y^G) \log \sum_i \sum_j f_x^i f_y^j \frac{p_{ij}}{p_i p_j} \]

  - PRALINE:  
    \[ score^{xy} = (1 - f_x^G)(1 - f_y^G) \sum_i \sum_j f_x^i f_y^j \log \frac{p_{ij}}{p_i p_j} \]

  The difference is the position of the log in the above equations:
  Edgar calls the Muscle scoring scheme “Log-expectation score (LE)”
So what do we do?

- A single shot for a good alignment without thinking: **MUSCLE, T-COFFEE** (maybe POA)
- If you want to experiment with making alignments for a given sequence set: **PRALINE**
  - Profile pre-processing
  - Iteration
  - Secondary structure-induced alignment
  - Globalised local alignment
- There is no single method that always generates the best alignment
- Therefore best is to use more than one method: e.g. include **Dialign2 (local)**
Summary

- Weighting schemes simulating simultaneous multiple alignment
  - Profile pre-processing (global/local)
  - Matrix extension (well balanced scheme)

- Using additional information
  - secondary structure driven alignment

- Schemes strike balance between speed and sensitivity