



1-month Practical Course Genome Analysis (Integrative Bioinformatics & Genomics)

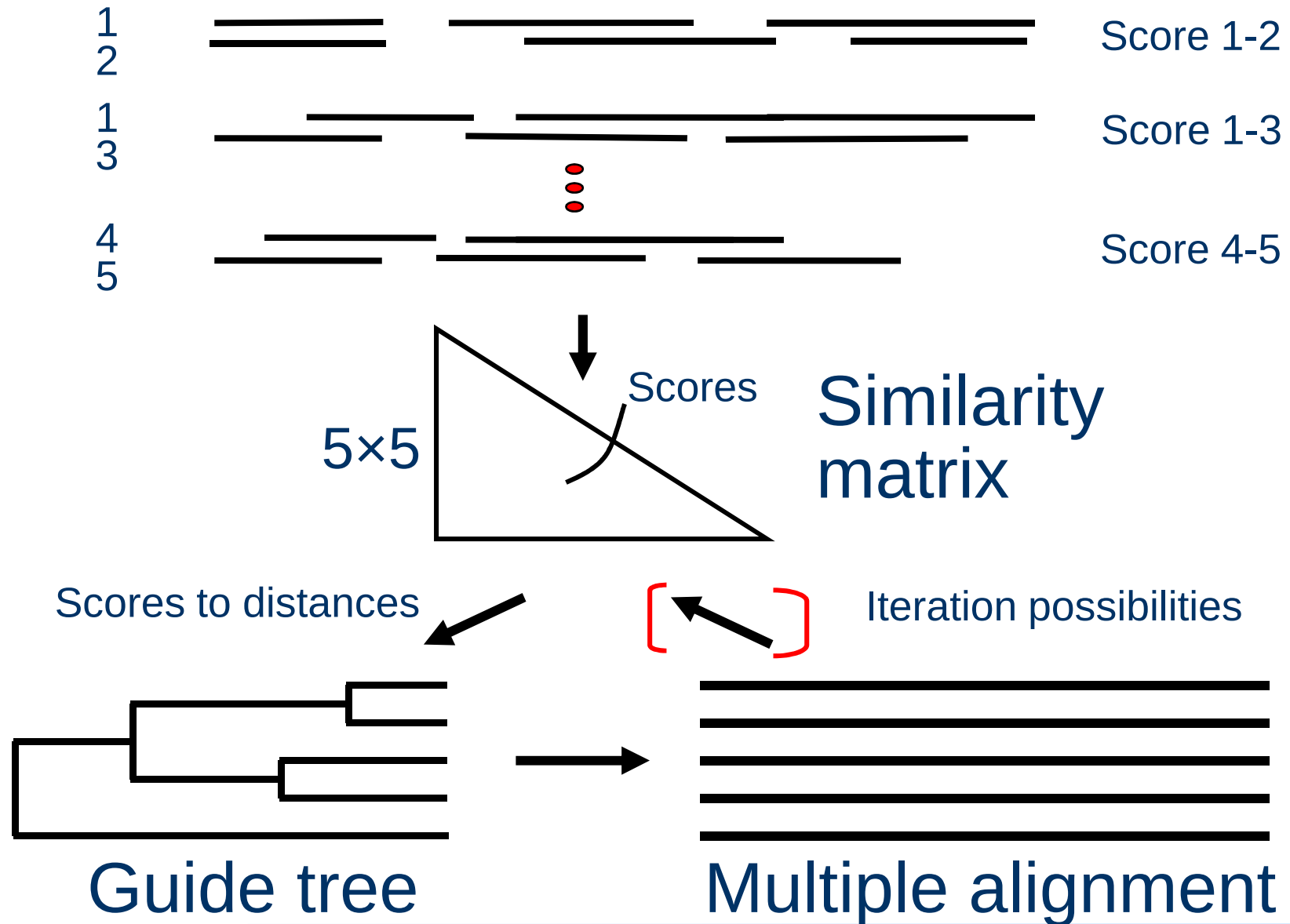
Lecture 5: Multiple sequence alignment (2)

Centre for Integrative Bioinformatics VU (IBIVU)
Vrije Universiteit Amsterdam
The Netherlands

ibivu.nl

heringa@cs.vu.nl

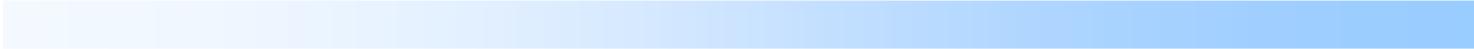
Progressive multiple alignment



Additional strategies for multiple sequence alignment

- Matrix extension (T-coffee)
- **Profile pre-processing (Praline)**
- Secondary structure-induced alignment

Objective: try to avoid (early) errors



PRALINE web-interface

PRALINE multiple sequence alignment



[Advanced Interface](#)

[Options Help](#)

[PRALINE sample output](#)

[References and FAQs](#)

PRALINE is a multiple sequence alignment program with many options to optimise the information for each of the input sequences; e.g. global or local preprocessing, predicted secondary structure information and iteration capabilities.

Paste in your sequences in FASTA format (MAX 500 sequences, length 2000):

Or Upload a FASTA file (MAX 500 sequences, length 2000):

Enter a name for your job

Options

Exchange weights matrix:

 [Help](#)

Associated gap penalties:

 Open Extension [Help](#)

Global progressive alignment strategy: [Help](#)

☐ Standard progressive strategy

☐ Pre-profile global processing Iterations

Score Cutt-off

☐ Pre-profile local processing Iterations

Score Cutt-off

☒ PSI-BLAST pre-profile processing (Homology-extended alignment) (new option)

PSI-BLAST Iterations: Start e-value Cutt-off at: DB:

Secondary structure prediction

 [Help](#)

DSSP-defined secondary structure search

☐ YES ☒ NO [Help](#)

Tree representation of the final alignment

☐ YES ☒ NO [Help](#)

Customize alignment representation colours

☐ YES ☒ NO [Help](#)

Final alignment file format

☐ NO FILE ☒ MSF ☐ FASTA [Help](#)

E-mail

If you would like to be notified when your job has completed, please **tick the box below** and enter the e-mail address the notification should be sent to:

☐ I want to be notified when my job is done at

[Submit](#)

Interface written by [Victor A. Simossis](#) and [Jaap Heringa](#)

The current colorscheme of the alignment is by **residue**. You can also view the alignment by

The colour assignments have been adapted from the defaults in CLUSTALX (Thompson *et al.*, 1997) [Abstract](#):

G, P, S, T **ORANGE** H, K, R **RED** F, W, V **BLUE** I, L, M, V **GREEN**

10 20 30 40 50 60 70

1fx1 - P K A V V V G S T T G N T E Y T A E T T A Q A A N A G V E D S R D A A S V E A G G T E G F D V L L G C S T W G D D S
FLAV_DESDE M S K V L I V F G S T T G N T E S T A Q L E E L A A G G H E V T L L N A A D A S A E N A D G V D A V L F g C S A W G M E D
FLAV_DESVH M P K A L I V V G S T T G N T E Y T A E T T A R E L A D A G V E D S R D A A S V E A G G T E G F D V L L g C S T W G D D S
FLAV_DESGI M P K A L I V V G S T T G N T E G V A E A A K T I N S E G M E T T V V N V A D V T A P G A E G V D V L L g C S T W G D D E
FLAV_DESSA M S K S L I V V G S T T G N T E T A A E Y V A E A F E N K E D V E L K N V T D V S V A D G N G V D I V L F g C S T W G E E E
4fxn - M K - - I V V W S G T G N T E K M A E L A K G T E S G K D V N T I N V S D V N I D E L N E - D I L I L G C S A M G D E V L
FLAV_MEGEL M V E - - I V V W S G T G N T E A M A N E I E A A V K A A G A D H E S V R F E D T N V D D V A S K D V I L L g C P A M G S E E
2fer - - K I G I F F S T T G N T E V A D F I G K T L G A - K A D A P I D V D D V T D P Q A T K D V D L F L G A P T W N T G A D
FLAV_ANASP S K K I G L F V G T Q T G K T E S V A E I I D E F G N - D V Y T L H D V S Q A E - V T D N D V Q Y L I l g C P T W N G E
FLAV_ECOLI - A T G I F F G S D T G N T E N I A K M I Q K Q L G K - D V A D V H D A K S S - K E D L E A V D I L L l g T P T W Y Y G E A
FLAV_AZOV I - A K I G L F F G S N T G K T R K V A K S I K R R D D E T M S D A L N V N R V S A E D F A Q V Q F I L l g T P T L G E G E P G L S
FLAV_ENTAG M A T I G I F F G S D T G Q T R K V A K S I H Q K D G - A D A P L D V R B A T H E Q F L S V P V L L l g T P T L G D G E P G V E
FLAV_CLOAB - M K I S I L V S S K T G K T E R V A K L I E E G V K R S G N E V K T M N D A V D K K F L Q E S E G I I f g T P T V Y A
3chy A D K E L K F L V V D D F S T M R R I V R N L K E L G F N N V E A E D G V D A N K L Q A G G V G F V I S D W N M

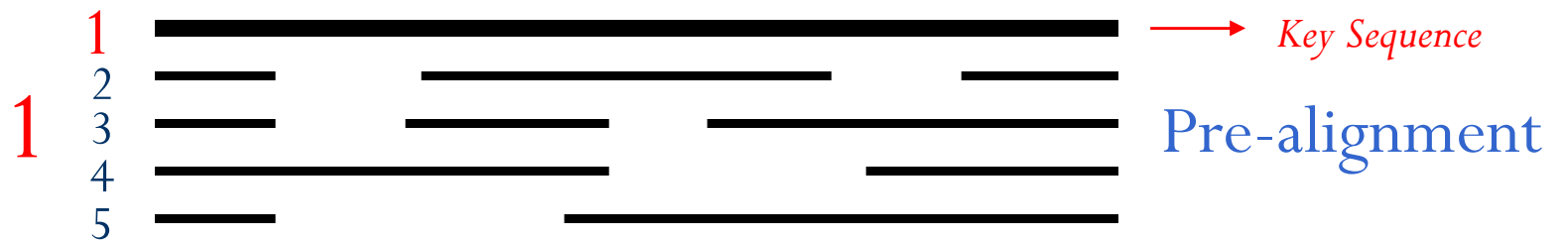
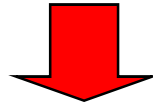
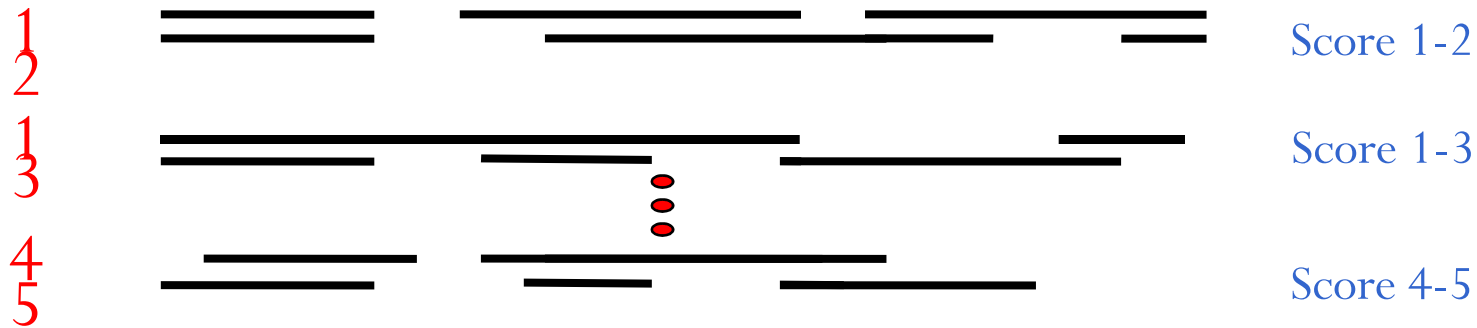
80 90 100 110 120 130 140

1fx1 - - E L Q D D F I P L F - D S L E E T G A O G H K V A C F G C G D S - S Y E V F C G A - V D A E E K L K N L G A E I V Q D
FLAV_DESDE - E M Q D D F L S L F - E E F N H F G L A G H K V A A A S G Q D - E Y E V F C G A - V P A E E R A K E L g A T I I A E
FLAV_DESVH - E L Q D D F I P L F - D S L E E T G A O G H K V A C F G C G D S - S Y E V F C G A - V D A E E K L K N L G A E I V Q D
FLAV_DESGI - E L Q E D F V P L Y - E D I D A G I G K K V G V I G C G D S - S Y T V F C G A - V D I E K A E E L g A T L V A S
FLAV_DESSA - E L Q D D F I P L Y - D S L E N A D I G K K V S V I G C G D S - D V T V F C G A - V D A I E E K L E K M g A V V I G D
4fxn - E E S E F E P F I - E E I S T K I S G K K V A F I G - - S V G W Q D G E W M D F E E I M N G V G C V V V E T
FLAV_MEGEL - E D S V V E P F F - T D L A P K I G K K V G I T G - - S V G W G S G E W M D A W K Q T E D T G A T V I G T
2fer T E H S G T S W D - E F L V D K P E V D M K D L P V A F I G L G A E G V P D N F C D A - E E I M D C F A K Q Q A K P V G F S N P D D V
FLAV_ANASP - - Q S D W E G L Y S E I D D V D N G K L V A Y I G T G D Q L G V A D N F Q D A - I G I L E E K I S O R g G K I V G H W S T D G V
FLAV_ECOLI - - Q C D W D D F F P T L E E D N G K L V A I T G C G D Q E D V A E V F C D A - L G T I M D I E P R g A T I V G H W P T A G Y
FLAV_AZOV I S D C E N E S W E - E F L P K I E G L D S G K T V A I G L G D Q V G Y P E N V L D A - L G E L Y S F F K D R g A K I V G S W S T D G Y
FLAV_ENTAG A G S Q V D S W Q - E F T N T L S E A D T G K T V A I G L G D Q N Y S K N F V S A - M B L I V D L V I A R g A C V V G N W P H E G Y
FLAV_CLOAB - - N I S W E M K K W D E S S E F N E G K L G A A L S T A N S A G G S D I - A - L T I L N H L M Y K g M I V V - S G G V
3chy - - - N I S W E M K K W D E S S E F N E G K L G A A L S T A N S A G G S D I - A - L T I L N H L M Y K g M I V V - S G G V

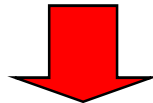
150 160 170 180 190

1fx1 G I R I D - G D P H A A D D V G W A N D V H G A
FLAV_DESDE G I K M E - G D A S N D P E A V A S I A E D V L X Q L
FLAV_DESVH G I R I D - G D P H A A D D V G W A N D V R G A
FLAV_DESGI S L K I D - G E P D - - S A E V I D W A R E V L A R V
FLAV_DESSA S L K I D - G D P E - - D E L Y S v G S G I A D K
4fxn P L I V Q - N E P D E A E Q D C I E F G K K L A N I
FLAV_MEGEL - A I V N - E M P D N A - P E C E I E G E A A A K A
2fer D V E E S K S V R - D G K F I G P L D M V N D Q P M E K R V A G W V E A V V S E T G V
FLAV_ANASP D F N D S K A I R - N G K F V G L A I D E D N Q S D L T D D R I K S V A Q L K S E F G
FLAV_ECOLI H F E A S K G L A D D H K F V G L A I D E D H Q P E L T A E R V E K V X Q S E E H D E I L N A
FLAV_AZOV I E S S E A - V V D G K F V G L A I D E D N Q S G K T D E R V A A L A Q A P E F G S E
FLAV_ENTAG K F S P S A A L E N N E F V G I P L D Q E N Q V D T E E R I D S W L E K L P A V
FLAV_CLOAB A T G K P K T H L G V V H N E I Q E N E D E N A B I G E R I A N V X Q I T
3chy G V V V - - - - K P F T A A T L E E K L N K I F E K L G M

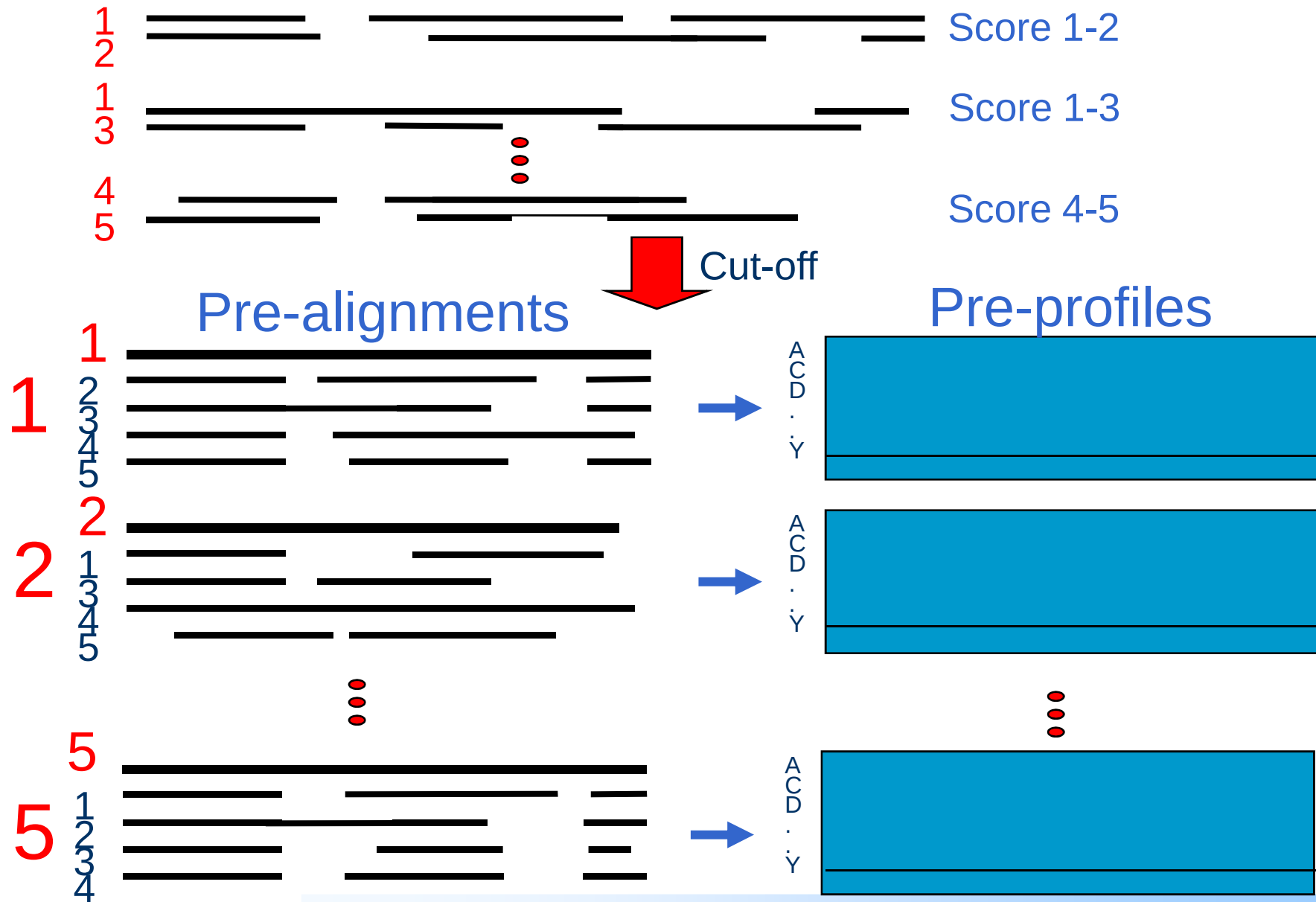
Profile pre-processing



*Master-slave (N-to-1)
alignment*

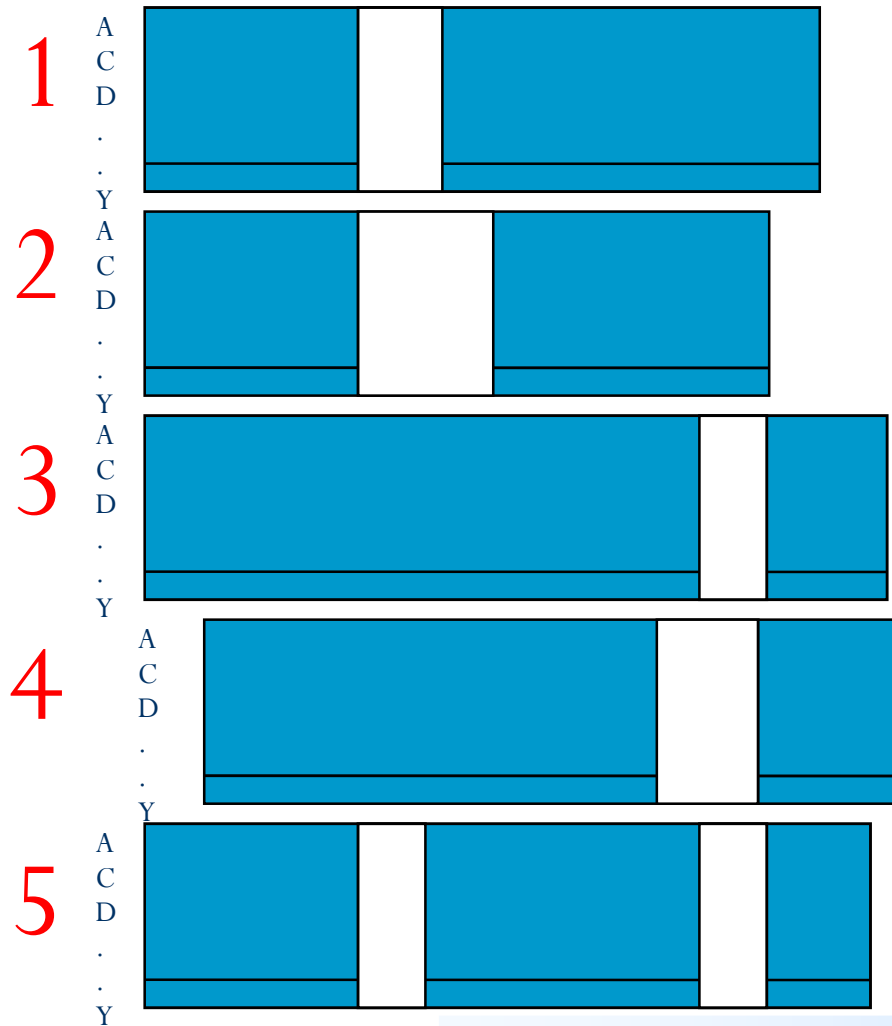


Pre-profile generation

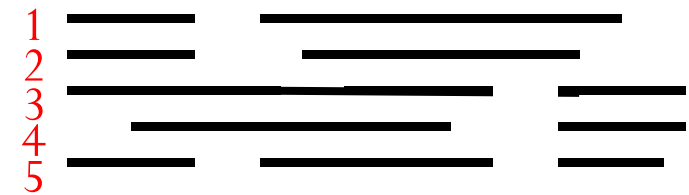


Pre-profile alignment

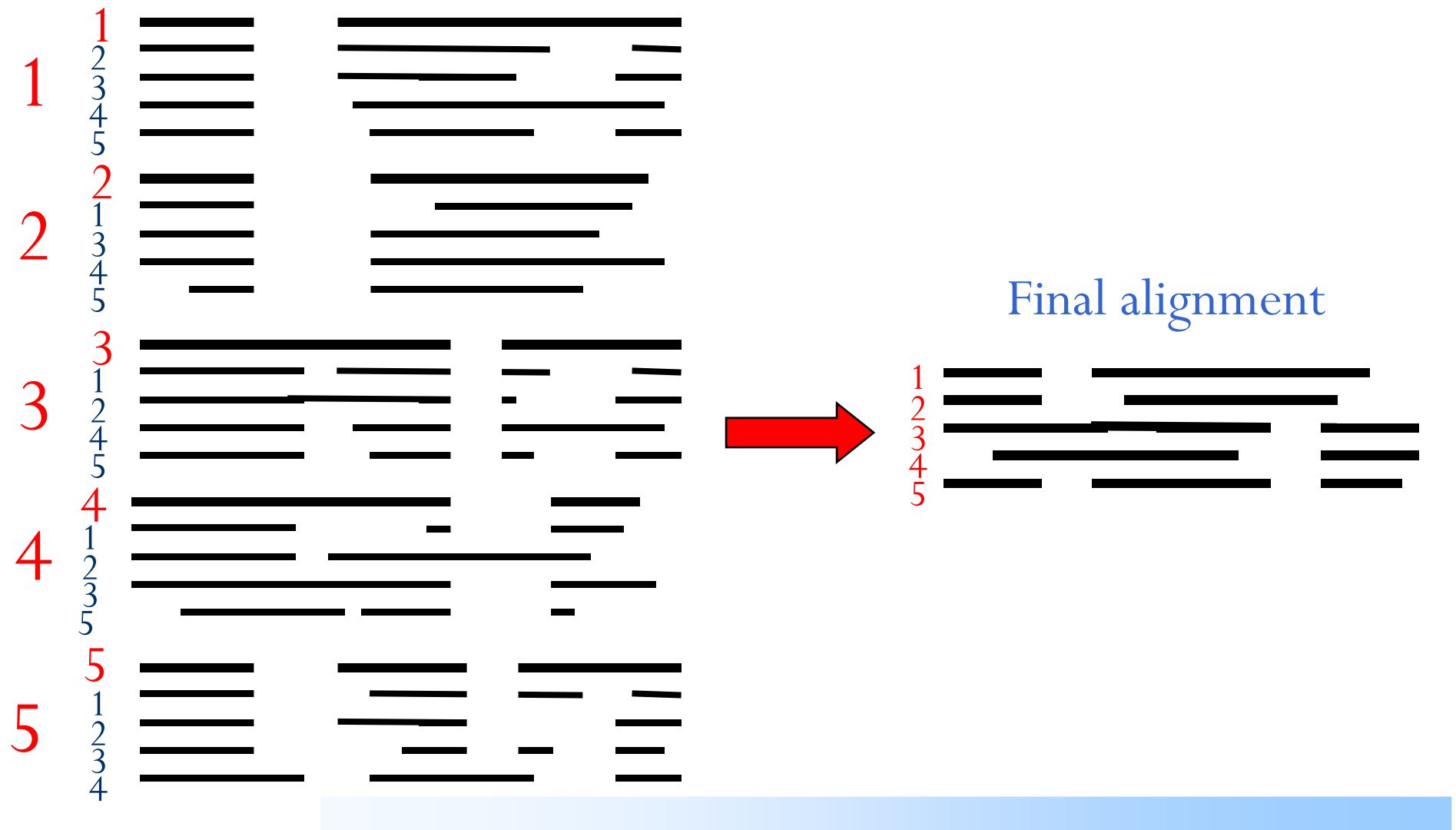
Pre-profiles



Final alignment

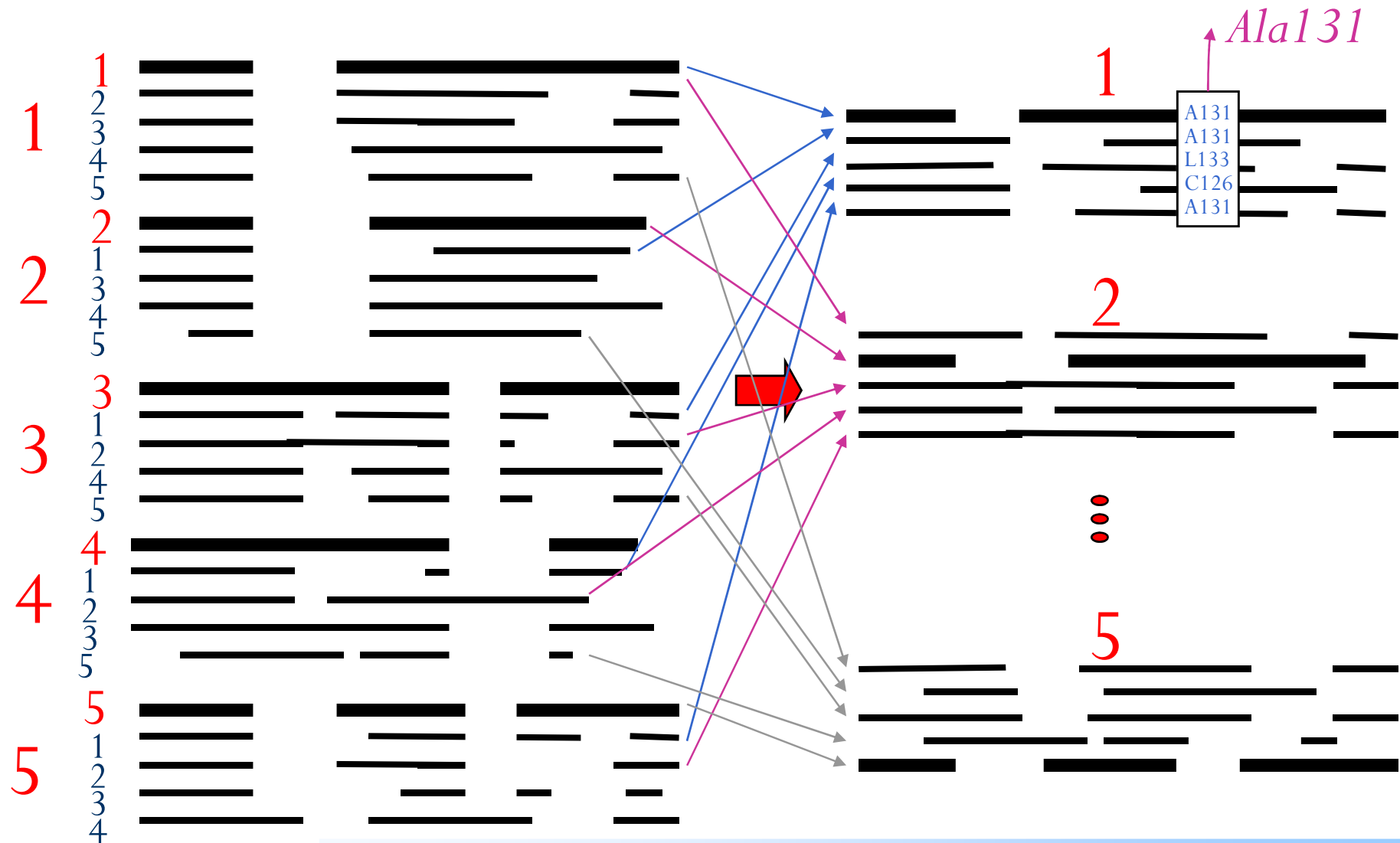


Pre-profile alignment



Pre-profile alignment

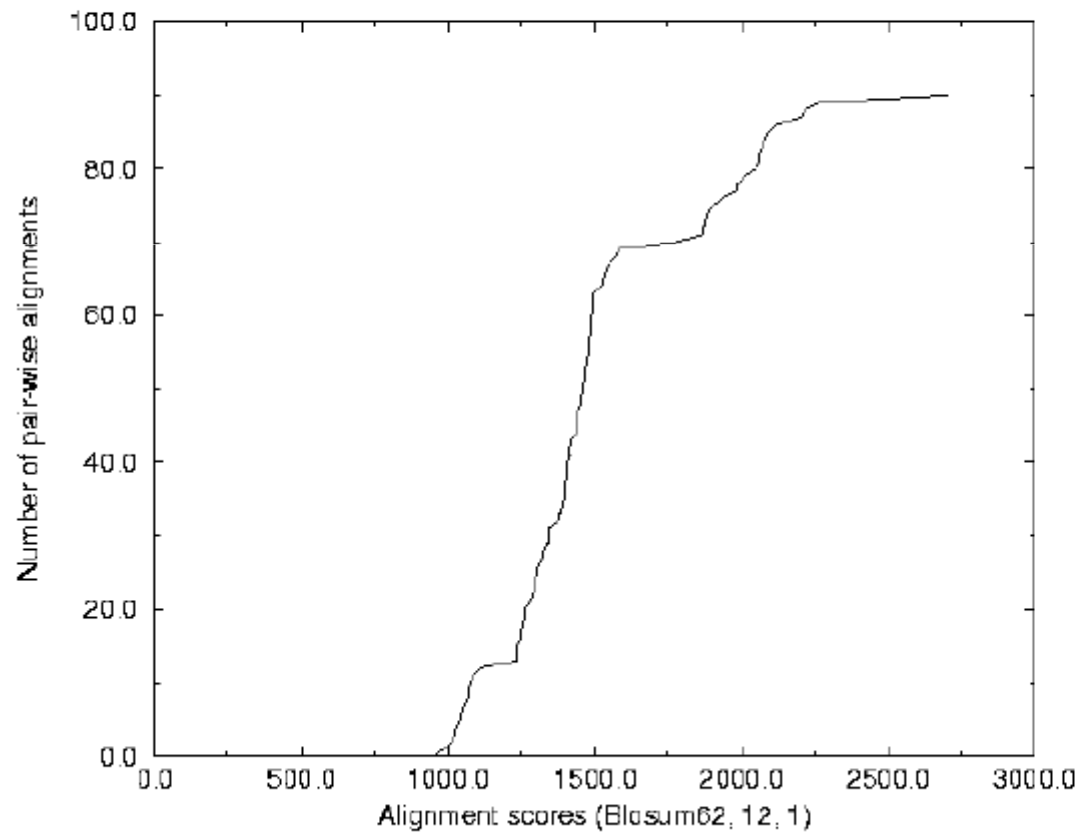
Alignment consistency



PRALINE pre-profile generation

- n Idea: use the information from all query sequences to make a pre-profile for each query sequence that contains information from other sequences
- n You can use all sequences in each pre-profile, or use only those sequences that will probably align 'correctly'. Incorrectly aligned sequences in the pre-profiles will increase the noise level.
- n Select using alignment score: only allow sequences in pre-profiles if their alignment with the score higher than a given threshold value. In PRALINE, this threshold is given as *prepro=1500* (alignment score threshold value is 1500 – see next two slides)

Reliable sequences for pre-profiles



The curve each time gives the number of pairwise alignments (y) scoring less than x. The range $1500 < x < 1800$ shows a flat section of the curve that can serve as a natural cut-off point for admitting sequences into the pre-alignment blocks

Global pre-processing

(prepro $\geq$ 0)

Preprocessed profile for sequence 2:

```
2fcr      KIGIFFSTSTGNTTEVADFIGKTLGAKADAPIDVDDVTDQPALKDYDLLFLGAPTWNTGADTERSGTSWDEFLYDKLPEVDMKDLPVAlFGLGDAEGYPD
1fx1      KALIVYGSTTGNTTEYTAETIARQL-ANAGYEVDSRDAASVEAFEGFDLVLLGCSTW--GDD--SIELQDDFLFDSLEETGAQGRKVACfGCGDS-SY-E
4fxn      -MKIVYWSGTGNTKMAELIAKGISGKDVNTINVSDVNIDELLNE-DILILGC--SAMDDEVLEESEFEPFIEEISTKISGKKVALGSYGWGDGKWMRD
FLAV_ANASP KIGLFYGTQTGKTESVaEIIRDEFNDVVLTHDVSEVTD--LNDYQYLIIGCPTWNIG--ELQ-SDW-EGLYSELDDVDFNGKLVAYfGTGDQIGYAD
FLAV_AZOVI KIGLFFGSNTGKTRKVaKSIKKRFDTMSDA-LNVNRVS-AEDFAQYQFLILgTPTLGPGLSSDCENESWEEFL-PKIEGLDFSGKTVALfGLGDQVGYPE
FLAV_CLOAB KISILYSSKTGKTERVaKLIIE--GVKRSGNIEVKDAVDKKFLQESEGIIIfgTPTYANISWEMK--KW----IDESSEFNLEGKLGAAfSTANAGGSDI
FLAV_DESDE KVLIVFGSSTGNTESiaQKLEELIAA-GGHEVTLNNAADASALADYDAVLfGCSAWGM-EDLEMQ---DDFLFEFNRFLAGRKVAafASGDQE-Y-E
FLAV_DESGI KALIVYGSTTGNTGEGVaEAIAKTLNSEGTTVVNVADVTAAPGLAEGYDVVLLgCSTW--GDDEIELQEDFVP-LYEDLDRAGLKDKKVGvfGCGDS-SY-T
FLAV_DESSA KSLIVYGSTTGNTETAaEYVAEAFENK-EIDVELKNVTDVSVANGYDIVLfgCSTW--G---EEEIELQDDFLYDSLENADLKGKKVSVfGCGDSD-Y-T
FLAV_DESVH KALIVYGSTTGNTTEYTaETIAREL-ADAGYEVDSRDAASVEAFEGFDLVLLgCSTW--GDD--SIELQDDFLFDSLEETGAQGRKVACfGCGDS-SY-E
FLAV_ECOLI AIGIFFGSDTGNTENiaKMIQKQLG--KDV-ADVHDISSKEDLEAYDILLgIPTWYIG---EAQCDWDDF-FPTLEEIDFNGKLVAlfGCGDQEDYAE
FLAV_ENTAG TIGIFFGSDTGQTRKVaKLIHQKLDGIADAPLDVRRATREQFL-SYPVLLLgTPTLGDGLPGVEAGSSWQEFT-NTLSEADLTGKTVALfGLGDQLNYSK
FLAV_MEGEL MVEIVYWSGTGNTTEAMaNEIEAAVAAGADVSVRFED-TNVDDVASKDVILLgCPA--MGSE-ELEDSVVEPFFTDLAPK--LKGKKVGLfGYGWGSG---
3chy      KELKFLVDDFSTRRIVRNLLKELGFNEEAEDGVDALNKLQA-GGYGfVI---SDWNM---PNMDGL---ELLKTIRADGAMSALPVLmv---TAEAKKE
```

```
2fcr      NFCDaIEEIHDCFAKQGAkPVGFsNPDDYDyEESKSVRDGKfLGLPLDMVNDQIPMEKRVAGWVEAVVSETGV
1fx1      YFCGAVDAIEEKLNlGA-----EIVQD----GLRID--GDPRAARDDIVGWAHDVRGAI--
4fxn      -FEERMNG-YGCVVVE--TPLIVQNEPD----EAE-----QDCIEFGKKIANI-----
FLAV_ANASP NFQDAIGILEEKISQRgGKTVGyWSTdGYDFNDsKALRNKGfVGLALDEdNQSDLTDDRIKSwaQLKSEfGL
FLAV_AZOVI NYLDALGELYSFFKDRgAKIVGSWSTdGYEFESSEAVVDGKFVGLALDLdNQSGKTDERVAawLAQIAPEfGL
FLAV_CLOAB ALLTILNHVKgMLVYSGG--VAFGKPKTHGYVHINEIQENE-----D-ENARI-fGERiANKvKQIF-----
FLAV_DESDE HFCGAVPAI-----EERAKELg-----ATIIAEG--LKMEGDASND--P--EAVASfAEDVLKQL--
FLAV_DESGI YFCGAVDVIEKKAEElgATLVA-----SSLKI-DGE-----PDSAeVLDwAREVLARV--
FLAV_DESSA YFCGAVDAIEEKLEKMgAVVIGDSLKIDGDPERDEIVSwGS--G-----IADKI-----
FLAV_DESVH YFCGAVDAIEEKLNlGa-----EIVQD----GLRID--GDPRAARDDIVGWAHDVRGAI--
FLAV_ECOLI YFCDALGTIRDIIePRgATIVGHwPTAGYHFEASKGLADdHFVGLAID--EDRQPTAERVEkwVKQISEELHL
FLAV_ENTAG NFVSAMRILYDLVIARgACVVGNwPREGYKfSfSAALENNEfVGLPLDQENQYDLTEERIDSwLEKL--KPAV
FLAV_MEGEL EWMDAWKQRTE---DTgATVIG-----TAIVNE-----MP-----DNAP-ECKElG--EAAAKA---
3chy      NIIAA-----AQAGAS--GY-----VVK--PFTAATLE-----EK-----LNKIFeKLGm
```

Iteration -1 SP= 127728.00 AvSP= 10.705 SId= 3764 AvSId= 0.315

Global pre-processing

(prepro≥0)

Preprocessed profile for sequence 3:

```
4fxn      MKIVYWSGTGNTTEKMAELIAKGIIESGKDVNTINVSDVNIDELLNEDILILGCSAMGDEVLEESEFEPFIEEISTKISGKKVALFGSYGWGDGKWMRDFE
1fx1      ALIVYGSTTGNTTEYTAETIARQLANAGYEVDSDAASVEAGGLFEGDLVLLGCSTWGDDSIQDDFIPLFDSLETGAQGRKVACFGSYEYFCGA-VDAIE
2fcr      IGIFSTSTGNTTEVADFICKTL--GAKADAPIDVDDVTPQALKDOLLFLGANTGADTERSGTSWDEFLYDKLPEVDMKDLPV-AIFGLGDAEGYPDFC
FLAV_ANASP IGLFYGTQTGKTESVaEIIRD---EFGNDVVTLDSQAQVTDLNDYQYLIIGCPTWNIGEL-QSDWEGLYSELDVDFNGKLVAYfGTIGYADNDAIGILE
FLAV_AZOVI IGLFFGSNTGKTRKVaKSIKKRFDETMS-DALNVNRVSAEDFAQYQFLILgTPTLGEGELENESWEEFLPKIGLDFSGKTVALfGQVGYPEGELYSFFK
FLAV_CLOAB MKILYSSKTGKTERVaKLIIEGVKRSNEVKTMNLDAVDKKFLQESEGIIFgTPTYANI--SWEMKKWIDESSENLEGKLGAAfSTAGGSDIALLTILN
FLAV_DESDE VLIVFGSSTGNTESIAQKLEELIAAGGHEVTLLNAADASAENLADYDAVLfGCSAWGMEDLEQDDFSLFEEFNRLAGRKVAaFAS--GDQEYVPAIE
FLAV_DESGI ALIVYGSTTGNTTEGVAEIAKTLNSEGMETTVMNVADVTAPGLAGYDVLLgCSTWGDDIEQEDFVPLYEDLDAGLKDKKVGvfGSYTYFCGA-VDVIE
FLAV_DESSA MSIVYGSTTGNTTETAaEYVAEAFENKEIDVELKNVTDVSVADLGNVDIVLfgCSTWGEEIEQDDFIPLYDSLADLKGKKVSVfGDYTYFCGA-VDAIE
FLAV_DESVH ALIVYGSTTGNTTEYTaETIARELADAGYEVDSDAASVEAGGLFEGDLVLLgCSTWGDDSIQDDFIPLFDSLETGAQGRKVACfGSYEYFCGA-VDAIE
FLAV_ECOLI TGIFFGSDTGNTENIaKMIQK---QLGKDVAADVDAKSSKEDLEAYDILLgIPTYGAEQCDWDDFFPTLEEID--FNGKLVALfGDYAFCDAGTIRDIE
FLAV_ENTAG IGIFFGSDTGQTRKVaKLIHQK-LDGIADA-PLDVRRATREQFLSYVLLLgTPTLGDELVEASQYDSWQEFNTDLDLTKTVALfGNYSKNFVSAMRILY
FLAV_MEGEL VEIVYWSGTGNTTEAMaNEIEAAVKAAGADVSVRFEDTNVDDVASKDVILLgCPAMGSEEELEDVVEPFFTDLAPKLKGKKVGLfGSYGWGSGEWMDAWK
3chy      DKELKFLVDDFSTMRIRVRNLLKELG--FNNVEEAEDGVD-ALNK-LQAGGYGVISDWNMPNMDGLELLKTI--RADGAMSALPVLMTAEAKKENIIA
```

```
4fxn      ERMNGYGCVVVETPLIVQNEPDEAEQDCIEFGKKIANI
1fx1      EKLNKGAEIVQDGLRIDGPRAARDDIVGWAHDVVRGA
2fcr      DAIEEHDCFAKQKPVGFSPDDSKNDQIPMEKRVAGW
FLAV_ANASP EKISGYGSKALRNGKFVGLALDEDNQDLTDDRIVKVAQL
FLAV_AZOVI DRTDGYEAVVVGALDLDNQSGKTDERVAaWLAQIAPE
FLAV_CLOAB HLMKgYGGVAFGKPYVHINEIQENEDENARfGERiANK
FLAV_DESDE ERAKELgATIIEAGLKMEDASNDPEAVASfAEDVLKQ
FLAV_DESGI KKAEElgATLVASSLKIDGEPDSAE--VLDwAREVARV
FLAV_DESSA EKLEKMgAVVIGDSLKIDGDPERDE--IVSwGSGIADI
FLAV_DESVH EKLNKGAEIVQDGLRIDGPRAARDDIVGWAHDVVRGA
FLAV_ECOLI PRTAGYGLAFVGLAIDEDRQPELTAERVEKwVKQISEE
FLAV_ENTAG DLVIARgCVVGNWPLENNEPDQENQDLTELEKKPAVL
FLAV_MEGEL QRTEDTgATVIGT-AIVNEMPDNA-PECKELGEAAAKA
3chy      AAQAGASGYVVK-PFTAATLEEKLKIFELKGM-----
```

Iteration -1 SP= 121196.00 AvSP= 10.075 SId= 3288 AvSId= 0.273

Reliable sequences for pre-profiles

```
Preprocessing: [ 1,  2]-> score 1459.00 compared to threshold 1500.00
Preprocessing: [ 1,  3]-> score 1470.00 compared to threshold 1500.00
Preprocessing: [ 1,  4]-> score 1399.00 compared to threshold 1500.00
Preprocessing: [ 1,  5]-> score 1323.00 compared to threshold 1500.00
Preprocessing: [ 1,  6]-> score 1214.00 compared to threshold 1500.00
Preprocessing: [ 1,  7]-> score 1800.00 compared to threshold 1500.00
Preprocessing: [ 1,  8]-> score 1952.00 compared to threshold 1500.00
Preprocessing: [ 1,  9]-> score 1944.00 compared to threshold 1500.00
Preprocessing: [ 1, 10]-> score 2594.00 compared to threshold 1500.00
Preprocessing: [ 1, 11]-> score 1458.00 compared to threshold 1500.00
Preprocessing: [ 1, 12]-> score 1294.00 compared to threshold 1500.00
Preprocessing: [ 1, 13]-> score 1254.00 compared to threshold 1500.00
Preprocessing: [ 1, 14]-> score 1040.00 compared to threshold 1500.00
Preprocessing: [ 2,  3]-> score 1250.00 compared to threshold 1500.00
Preprocessing: [ 2,  4]-> score 1880.00 compared to threshold 1500.00
Preprocessing: [ 2,  5]-> score 1907.00 compared to threshold 1500.00
Preprocessing: [ 2,  6]-> score 1319.00 compared to threshold 1500.00
Preprocessing: [ 2,  7]-> score 1329.00 compared to threshold 1500.00
Preprocessing: [ 2,  8]-> score 1388.00 compared to threshold 1500.00
Preprocessing: [ 2,  9]-> score 1406.00 compared to threshold 1500.00
Preprocessing: [ 2, 10]-> score 1360.00 compared to threshold 1500.00
Preprocessing: [ 2, 11]-> score 1768.00 compared to threshold 1500.00
Preprocessing: [ 2, 12]-> score 1773.00 compared to threshold 1500.00
Preprocessing: [ 2, 13]-> score 1165.00 compared to threshold 1500.00
Preprocessing: [ 2, 14]-> score 1025.00 compared to threshold 1500.00
Preprocessing: [ 3,  4]-> score 1242.00 compared to threshold 1500.00
Preprocessing: [ 3,  5]-> score 1266.00 compared to threshold 1500.00
Preprocessing: [ 3,  6]-> score 1238.00 compared to threshold 1500.00
Preprocessing: [ 3,  7]-> score 1312.00 compared to threshold 1500.00
Preprocessing: [ 3,  8]-> score 1407.00 compared to threshold 1500.00
Preprocessing: [ 3,  9]-> score 1394.00 compared to threshold 1500.00
Preprocessing: [ 3, 10]-> score 1388.00 compared to threshold 1500.00
Preprocessing: [ 3, 11]-> score 1243.00 compared to threshold 1500.00
Preprocessing: [ 3, 12]-> score 1242.00 compared to threshold 1500.00
Preprocessing: [ 3, 13]-> score 1690.00 compared to threshold 1500.00
Preprocessing: [ 3, 14]-> score 988.00 compared to threshold 1500.00
```



Pre-profiles (prepro $\geq$ 1500)

Preprocessed profile for sequence 1:

```
*      1 -      1. 1fk1
      2 -      7. FLAV_DESDE
      3 -      8. FLAV_DESGI
      4 -      9. FLAV_DESSA
      5 -     10. FLAV_DESVH
```

```
1fk1      PKALIVYGSTTGNTETAYTAETIARQLANAGYEVD SRDAASVEAGGLFEGFDLV LLGC STWG
FLAV_DESDE SKVLIVFGSSTGNTES I a QKLEELI AAGGHEVT LLNAADASAENLADGYDAV LFQC SAWG
FLAV_DESGI PKALIVYGSTTGNTETGV a EA IAKTLNSEGME TT VVNVADVTAPGLAEGYDVV LLQC STWG
FLAV_DESSA SKSLIVYGSTTGNTET A s EYVAEAFENKE IDVELKIVTDVSVADLGNGYDIV LFQC STWG
FLAV_DESVH PKALIVYGSTTGNTETAYTAETIARELADAGYEVD SRDAASVEAGGLFEGFDLV LLGC STWG
```

```
1fk1      DD SIELQDDF IP LFDSLEETGAQGRKVACFGCGDS SYEYFCGAVD AI EEKLNKLGAEI VQ
FLAV_DESDE MEDLEMQDDF LS LFE EFNRFGLAGRKVAA f a SGDQ EYEHFCGAVPAI EERAKELGATI IA
FLAV_DESGI DDEI ELQEDFVP LYED LDRAGLKDKKVG V f GCGDS SYTYFCGAVDVI EKKAEELGATLVA
FLAV_DESSA EEEI ELQDDF IP LYDSLENADLKGKKV SV f GCGDS DYTYFCGAVD AI EEKLEKMGAVVIG
FLAV_DESVH DD SIELQDDF IP LFDSLEETGAQGRKVAC f GCGDS SYEYFCGAVD AI EEKLNKLGAEI VQ
```

```
1fk1      DGLRIDGDPRAARDDI VGWAHDVVRGAI
FLAV_DESDE EGLKMEGDASNDPEAVAS f AEDVLKQL
FLAV_DESGI SSLKIDGEPDSA -- EVLDWAREVVLARV
FLAV_DESSA DSLKIDGDP -- RDEIVSWGSGIADKI
FLAV_DESVH DGLRIDGDPRAARDDI VGWAHDVVRGAI
```

Preprocessed profile for sequence 2:

```
*      1 -      2. 2fcr
      2 -      4. FLAV_ANASP
      3 -      5. FLAV_AZOVI
      4 -     11. FLAV_ECOLI
      5 -     12. FLAV_ENTAG
```

```
2fcr      KIGIFFSTSTGNTTEVADF IGKTLGAKADAP IDVDDVTD PQAL KD YD LLFLGAP T WNTGA
FLAV_ANASP KIGLFYGTQTGKTE SV a E I I RDEE GNDV VTL HDVSEVTD --- LNDYQYL I IgCPTWNI G-
FLAV_AZOVI KIGLFTGSGNTGKTRKV a K SIKKRFD TMSDA- LNVNRV S- AEDEAQYQFL I IgTPTLGPGL
FLAV_ECOLI AIGIFFGSDTGNTENI AKMI QKQLG -- KDV-ADVHDI SSKEDLEAYD I LLLgIPTWYYG-
FLAV_ENTAG TIGIFFGSDTGQTRKV a K LI HQKLDGI ADAP LDVRRATREQFL -SYPV LLLgTPTLGDGL
```

```
2fcr      DTERSGTSWDEF LYDKLP EVDIMKDL PVAI FG LGDAEGYPDNFCD AIEEI HDCEFAHQGAKP
FLAV_ANASP --ELQ- SDW- EGLY SE LDDVDF I NGK LVAY f GTGDQ IGYADNFQDA IGILEEK I SQRGKT
FLAV_AZOVI SSDCENESWEEFL -PKIEGLDF SGKTVAL f g LGDQ VGY PENY LDA LGELY SFFKDRGAKI
FLAV_ECOLI --EAQCDWDDF -FPT LEEI DF I NGK LVAL f GCGDQEDYA EYFCDA LGTIRDI IEPRGATI
FLAV_ENTAG PGVEAGSSWQEFT -NT LSEADLTGKTVAL f g LGDQ LNY SKNFV SAMR ILYDLVI ARGACV
```

```
2fcr      VGFSNPDDYDY EESKSVRDGKF LGLPLDMVNDQ IPMEKRVAGWVEAVVSETGV
FLAV_ANASP VGYWSTDG YDFND SKA LRNGKF VGLALDEDNQSDL TDDR I KSWVAQLKSEFGL
FLAV_AZOVI VGSWSTDGYEFE SSEAVVDGKF VGLALDLDNQSGKTDERVAAWLAQI APEFGL
FLAV_ECOLI VGHMPTAGYFEASKGLADD HFVGLAID- -EDRQPTAERVEKVKQI SEELHL
FLAV_ENTAG VGHMPREGYKFSFSAA LEHNEFVGLPLDQENQYDLTEER IDSWLEKL --KPAV
```

Pre-profiles (prepro $\geq$ 1500)

13

Preprocessed profile for sequence 13:

```
+      1 -   13. FLAV_MEGEL
      2 -   3. 4fk $\lambda$ 
```

```
FLAV_MEGEL      MVEIVYWSGTGNTTEAMANEIEAAVKAAGADVESVRFEDTINVDDVA SKDV ILLGCPAMGSE
4fk $\lambda$           -MKIVYWSGTGNTKMAELIANGIIESGKDVNTINVSDVNI DELLNEDI LILGC SAMGDE
```

```
FLAV_MEGEL      ELED SVVEPFFTDLAPKLLGKKVGLFGSYGWGSGEWMDAWKQRTEDTQATVIGTAVNIEM
4fk $\lambda$           VLEESEFEPTIEEI STKI SGKKVALEFGSYGWGDGKWMRD FEERMNGYGC VVVETLIVQNE
```

```
FLAV_MEGEL      PDIAPECKELGEAAAKA
4fk $\lambda$           PDEAQDCIEFGKKIANI
```

Preprocessed profile for sequence 14:

```
+      1 -   14. 3chy
```

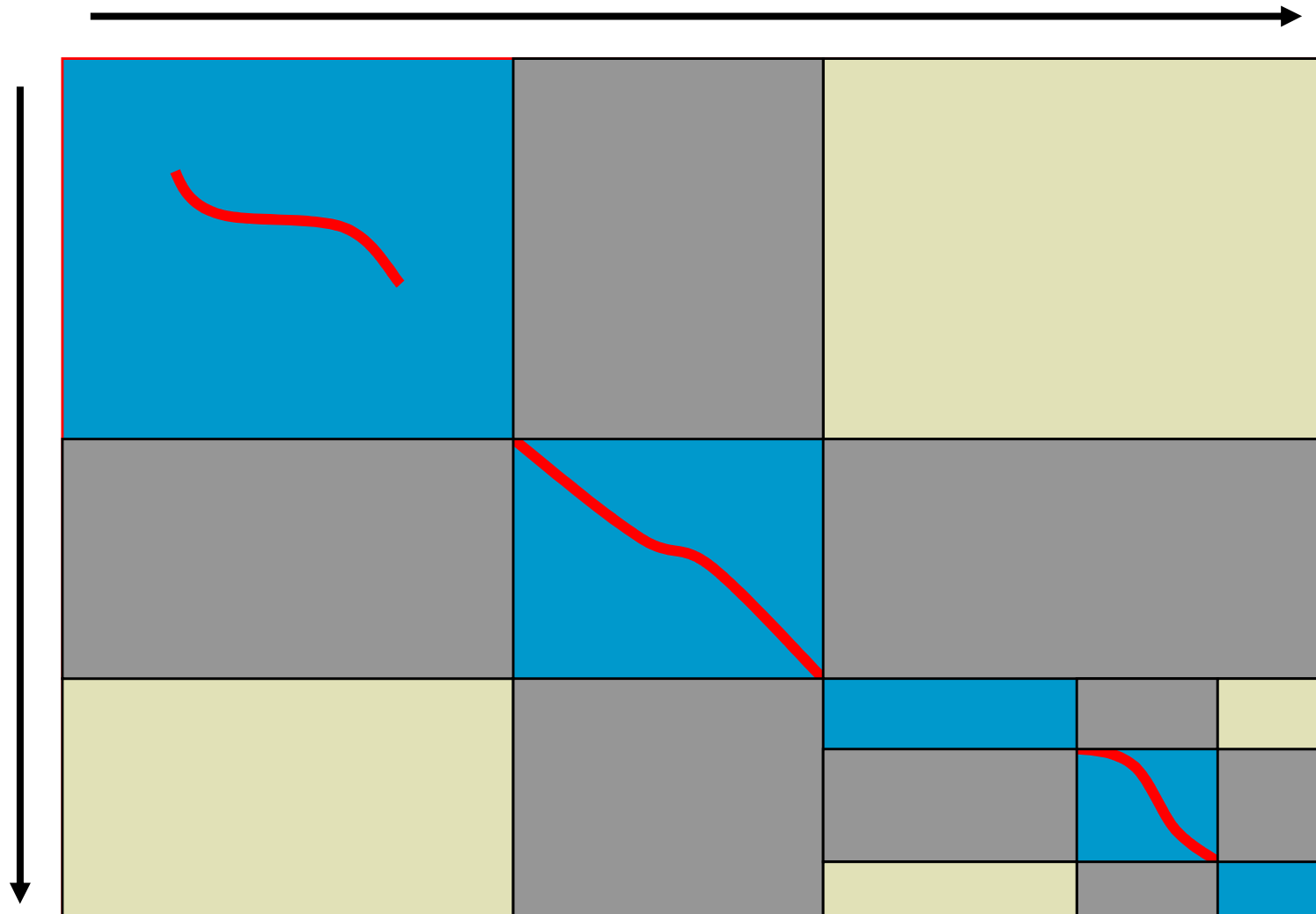
```
3chy          ADKELKFLVDDFSTMRIRVRNLLKELGFNNVEEAEDGVDA LNKLQAGGYGEVI SDWIMP
```

```
3chy          NMDGLELLKTIRADGAMSALPV LMTAEAKKENII AAAQAGASGYVVKPFTAATLEEK LK
```

```
3chy          KIFEKLGK
```

14

Local pre-processing



Local alignments are calculated from high to low scoring – each time the sequence parts corresponding to a selected local alignment are blocked such that a next local alignment has to emerge before or after the earlier selected one – this preserves co-linearity of the local alignments and associated sequence fragments in the pre-alignments

Local pre-processing

(locprepro≥0)

Preprocessed profile for sequence 2: 2fcr

```
2fcr      KIGIFFSTSTGNTTEVADFIGKTLGAKADAPIDVDDVTDPAQALKDYDLLFLGAPTWNITGADTERSGTSWDEFLYDKLPEVDMKDLPVIAIFGLGDAEGYPD
1fx1      ...IVYGSTTGNTTEYTAETIARQL---ANAGYEVDDAASVEAFEGFDLVLLGCSTW--GDDSELQ----DDFLFDSLEETGAQGRKVACFGCGDS-SY-E
4fxn      KI-VYWS-GTGNTTEKMAELIAKGIGKDVNT-INVSDVNIDELLNE-DILILGCSA--MGDEVEES--EFEPF----IEEISTKGKKVALFGWGDGKGYG-
FLAV_ANASP KIGLFYGTQTGKTESVaEIIIRDEFNDVTLHDVSEVTD---LNDYQYLIIGCPTWNIG---ELQ-SDW-EGLYSELDDVDFNGKLVAYfGTGDQIGYAD
FLAV_AZOVI KIGLFFGSNTGKTRKVaKSIKKT---SDA-LNVNRVS-AEDFAQYQFLILgTPTLGEGLDCENE--SWEEFL-PKIEGLDFSGKTVALfGLGDQVGYPE
FLAV_CLOAB KISILYSSKTGKTERVaKLIIE--GVKRSNGNIEVKDAVDKKFLQESGIIIfgTPTY-----YANISWEKWI-DESSEFNLEGKLGAAfSTANSAGGSD
FLAV_DESDE KVLIVFGSSTGNTESiAQKLEELIAAAADA--SAENLAD-----GYDAVLfGCSAWGM-EDLEMQ----DDFLFEEFNRFGLAGRKVAafASGDQE-Y-E
FLAV_DESGI ...IVYGSTTGNTTEGVAEIAKTLNSEGTTVVNVADVTAPGLAEGYDVVLLgCSTW--GDDIELQ----EDFLYEDLDRAGLKDKKVGvfGCGDS-SY-T
FLAV_DESSA ...IVYGSTTGNTTETAaEYVAEAFENK---EIDVENVTD-VSVADYDIVLfgCSTW--G---EEEIELQDDFLYDSLENADLKGGKVSvfGCGDSD-Y-T
FLAV_DESVH ...IVYGSTTGNTTEYTaETIAREL---ADAGYEVDDAASVEAFEGFDLVLLgCSTW--GDDSELQ----DDFLFDSLEETGAQGRKVACfGCGDS-SY-E
FLAV_ECOLI .GIFFGSDTGNTENiAKMIQKQLG-K-----DVADVHDKEDLEAYDILLlgIPTWYYG----EAQCDWDDF-FPTLEEIDFNGKLVAlfGCGDQEDYAE
FLAV_ENTAG .IGIFFGSDTGQTRKVaKLIHQKLDGIADAPLDVRRATREQFL-SYPVLLLgTPT--LG-DGELPGVSWQEFT-NTLSEADLTGKTVALfGLGDQLNYSK
FLAV_MEGEL .VEIVYWSGTGNTTEAMaNEIEKAAGADVESDTNVDDV----ASK--DVILLgCPA--MGSE-ELEDSSVVEPFFTDLAPK--LKGGKVGlfGYGWGSG---
3chy      .....ADKELKFLVVDDFIVRNL----LKEL-----GFNNVEEAED
```

```
2fcr      NFCDIAIEEIHDCFAKQGAQKPVGFSPDDYDYEESKSVRDGKFLGLPLDMVNDQIPMEKRVAGWVEAVVSETGV
1fx1      YFCDAIEE-----K--LKNLG-----AEIVQD----GLRID--GD--PRAARIVGWAHDV.....
4fxn      --CVVVE-----TPLIVQNPDE---AEQDCIEFGK.....
FLAV_ANASP NFQDAIGILEEKISQRgGKTVGYWSTDGyDFNDSKALRNGKFVGLALDEDNQSDLTDDRIKSwaQLKSEFGL
FLAV_AZOVI NYLDALGELYSFFKDRgAKIVGSWSTDGyEFESSEAVVDGKFVGLALDLNQSGKTDERVAawLAQIAPEFGL
FLAV_CLOAB ---IALLTIH-LMVKSGG--VAFGKPKTHGYVHINEIQENE-----D-ENARI-fGERiANKVKQI.....
FLAV_DESDE HFCGAVPAI-----EERAKELg-----ATIIEAGKMEG---DASND--P--EAVASfaEDVLKQ...
FLAV_DESGI YFCGAVDVIEKKAEElgATLVASSEPD-----SAEVLd.....
FLAV_DESSA YFCGAVDAIEEKLEKMgAVVIGDSLKIDGDPERDEIVSwGS--G-----IADKI.....
FLAV_DESVH YFCDAIEE-----K--LKNLG-----AEIVQD----GLRID--GD--PRAARIVGWAHDV.....
FLAV_ECOLI YFCDALGTIRDIIePRgATIVGHWPTAGYHFEASKGLADHDfVGLAID--EDRQPTAERVEKwVKQISEE...
FLAV_ENTAG NFVSAMRILYDLVIARgACVVG--NPEGYKFSFSAALENNEFVGLPLDQENQYDLTEERIDSwLEAVL.....
FLAV_MEGEL EWMDAWKQTED----TgATVIGTANPDN.....
3chy      G-VDALNKLQ-----AGGYGFSNMPNMDLELLKTIRDGAMSALPVLMTAEAKKENIAGYVAATLEE...
```

Local pre-processing

(locprepro $\geq$ 0)

Preprocessed profile for sequence 3: 4fxn

```
4fxn      MKIVYWSGTGNTKMAELIAKGIIESGKDVNTINVSDVNIDELLNEDILILGCSAMGDEVLEESEFEFPFIEEISTKISGKKVALFGSYGWGDGKWMRDFE
1fx1      ..IVYGSTTGNTTEYTAETIARQLANAGYEVDSDRAASVEAGGLFEGDLVLLGCSTWGDDSIQDDFIPLFDSLETGAQGRKVACFGC---GDSSYVDAIE
2fcr      .KIIFSSSTGNTTEVADFIGKTL---GAKADAIDVDDVTDQPALKD DLLFLGAPTTGADT-ERSSWDEFLPEVDMK--DLPVAIF---GLGDAE-----
FLAV_ANASP ..LFYGTQTGKTESVaEIIIRD---EFGNDVVTLDVSQAEVTDLNDYQYLIIgCPTIGE--L-QSDWEGLYSELDVDFNGKLVAYfGTIGYADGKWSTDFN
FLAV_AZOVI ..LFFGSNTGKTRKVAKSIKKRFDETMSD--ALNVNRVSAEDFAQYQFLILgTPTLGEGELNESEFLPKIEGLD--FSGKTVALfGQVGYGEGSWSTD--
FLAV_CLOAB MKILYSSKTGKTERVaKLIIEGVKRSNGNEVKTMNLDAVD-KKFLQEEGIIFgTPTMKKWIDESSEFN--LEAfSTANSgSDIALLGVAfGKPK-----
FLAV_DESDE ..IVFGSSTGNTKLEELIAAG---GHEVTLNNAADASAENLADYDAVLfGCSAWGMEDLEQDDFLSLFEEFNRLAGRKVAAfAS---GDQEY-EHFE
FLAV_DESGI ..IVYGSTTGNTGvAEIAKTLNSEGMEttVVNVADVtAPGLAGYDVLLgCSTWGDDIEQEDFVPLYEDLDAGLKDKKVGvFGC---GDSSYTYDIE
FLAV_DESSA ..IVYGSTTGNTETAAeEYVAEAFENKEIDVELKNVTDVSVADLGNyDIVLfgCSTWGEEEEIEQDDFIPLYDSLNaDLKGKKVSVfGC---GDS---DYE
FLAV_DESVH ..IVYGSTTGNTTEYtAEtIARELADAGYEVDSDRAASVEAGGLFEGDLVLLgCSTWGDDSIQDDFIPLFDSLETGAQGRKVACfGC---GDSSYVDAIE
FLAV_ECOLI ..IFFGSDTGNTENIaKMIQK---QLGKDV--ADVHDISKEDLEAYDILLlgIPTYGEAQCDWDDFFPTLEEID--FNGKLVALfGC---GD---QEDYA
FLAV_ENTAG ..IFFGSDTGQTRKVAKLIHQGIADAPLDVRR---ATREQFLSYPVLLLgTPTLGDELVEASQYDSWQEFTNTDLTGKTVALf---GLGDQNYSKNFV
FLAV_MEGEL VEIVYWSGTGNTTEAMaNEIEAAVKAAGADVesVRfEDTNVDDVASKDVILLgCPAMGSEEELEDSVVEPFfTDLAPKLKGKKVGLfGSYGWGSGEWMDAWK
3chy      .RIV.....N...LKEL---GFVEEAEDVDALNISDPNMDELLRADVLMVTAEAKKENIIAAAQVKPFLEEKLNKIFEK.....
```

```
4fxn      ERMNGYGCVVVETPLIVQNEPDEAEQDCIEFGKKIANI
1fx1      EKLNKGAEIVQDGLRIDGDPRAARDDIV.....
2fcr      ----GYPCDAIEKPVGFSN-PDDEESKSVRDGK....
FLAV_ANASP DSRNGVGLALDE----DNQSDLT-DRIEFG.....
FLAV_AZOVI ----GYEAVVVGLALDLDNQTDELAQIAPEFG.....
FLAV_CLOAB THL-GY----VHINEIQENEDENAR---I-fGERiAN.
FLAV_DESDE ERAKELgATIIEAGLKMENDP-EAAEDVLK.....
FLAV_DESGI KKAEElgATLVASSLKIDGEPDSAE--VLDwAREVARV
FLAV_DESSA EKLEKMgAVVIGDSLKIDGDPERDE--IVSwGSGIAD.
FLAV_DESVH EKLNKGAEIVQDGLRIDGDPRAARDDIV.....
FLAV_ECOLI E----YFCDALGTDII---EP.....
FLAV_ENTAG SAMRg-ACVVGWNWPLENNEPDQENQDLTE.....
FLAV_MEGEL QRTEDTgATVIGTAIV--NEPDNA-PECKEIGE....
3chy      .....
```

CLUSTAL X (1.64b) multiple sequence alignment Flavodoxin- cheY

```
1fx1      -PKALIVYGSTTGNT EYTAETIARQLANAG-Y-EVDSRD AASVEAGGLFEGFDLVLLGCSTWGDD SIE-----LQDDFIPLFD-SLEETGAQGRK
FLAV_DESVH MPKALIVYGSTTGNT EYTAETIARELADAG-Y-EVDSRD AASVEAGGLFEGFDLVLLGCSTWGDD SIE-----LQDDFIPLFD-SLEETGAQGRK
FLAV_DESGI MPKALIVYGSTTGNT EGVAEIAI AKTLNSEG-M-ETT VVNVDVTAPGLAEGYDVVLLGCSTWGDD EIE-----LQEDFVPLYE-DLDRAGLKDKK
FLAV_DESSA MSKSLIVYGSTTGNT ETAAEYVAEAFENKE-I-DVELKNVTDVSVADLNGYDIVLFGCSTWGE EEEIE-----LQDDFIPLYD-SLENADLKGGK
FLAV_DESDE MSKVLIVFGSSTGNT ESIAQKLEELIAAGG-H-EVTLLNAADASAENLADGYDAVLFGCSAWGMEDLE-----MQDDFLSLFE-EFNRFGLAGRK
FLAV_CLOAB -MKISILYSSKTGKTERVAKLIEEGVKRSGNI-EVKTMNLDAVDKKFLQE-SEGIIFGTPTYAN-----ISWEMKKWID-ESSEFNLEGKL
FLAV_MEGEL --MVEIVYWSGTGNT EAMANEIEAAVKAAG-A-DVESVRFEDTNVDDVAS-KDVILLGCPAMGSE--E-----LED SVVEPFF-TDLAPKLKGGK
4fxn      ---MKIVYWSGTGNT EKM AE LIAKG IIESG-K-DVNTIN VSDVNI DELLN-EDILILGCSAMGDE--V-----LEESEFEPFI-EEISTKISGKK
FLAV_ANASP SKKIGLFGYGTQGT KTESVAEIIRDEF GNDVVT----LHDVSQA E VTDLND-YQYLIIGCPTWNIGELQ---SD-----WEGLYS-ELDDVDFNGKL
FLAV_AZOVI -AKIGLFFGSNTGKTRK VAKSIKKRFDDETMSD---ALNVNRVSAEDFAQ-YQFLILGTPTLGEGELPGLSSDCENESWEEFLP-KIEGLDFSGKT
2fcr      --KIGIFFSTSTGNT E VADF IGKTLGAKADAP---IDVDDVTD P QAL KD-YDLLFLGAP TWNTGADTERSGT---S WDEFLYDKLPEVDMKDL P
FLAV_ENTAG MATIGIFFGSDTGQTRK VAKLIHQKLDGIADAP---LDVRRATREQFLS--YPVLLLGTPTLGDGELPGVEAGSQYDSWQEFTN-TLSEADLTGKT
FLAV_ECOLI -AITGIFFGSDTGNT ENIAKMIQKQLGKD VAD---VHDI AKSSKEDLEA-YDILLLGIPTWYYGEAQ-CD-----WDDFFP-TLEEIDFNGKL
3chy      --ADKELKFLVVD DFSTMRRIVRNL LKELG---FNNVEEAEDGV DALN-----KLQAGGYGFV--I-----SDWNMPNMDG-LELLKTIR---
```

. . . : . :

```
1fx1      VACFGCGDSSYEYF--CGAVDAIEEKLKNLGAEIVQDG-----LRIDGDPRAARD DI V GWAHDVRGAI-----
FLAV_DESVH VACFGCGDSSYEYF--CGAVDAIEEKLKNLGAEIVQDG-----LRIDGDPRAARD DI V GWAHDVRGAI-----
FLAV_DESGI VGVFGCGDSSYTYF--CGAVDVIEKKAEELGATLVASS-----LKIDGEPDSAE--VLDWAREVLARV-----
FLAV_DESSA VSVFGCGDSDYTYF--CGAVDAIEEKLKMGAVVIGDS-----LKIDGDPERDE--IVSWGSGIADKI-----
FLAV_DESDE VAAFASGDQEYEHF--CGAVPAIEERAKELGATIIAEG-----LKMEGDASNDPEAVASFAEDVLKQL-----
FLAV_CLOAB GAAFSTANSIAGGS--DIALLTILNHL MVK GMLVYSGGVA---FGKPKTHLGYVHINEIQENEDENARIFGERIANKVQIF-----
FLAV_MEGEL VGLFGSYGWGSGE----WMDAWKQRTEDTGATVIGTA-----IVN-EMP DNAPECKE-LGEAAKA-----
4fxn      VALFGSYGWGDGK----WMRDFEERMNGYGC V VETP-----LIVQNEPDEAEQDCIEFGKKIANI-----
FLAV_ANASP VAYFGTGDQIGYADNFQDAIGILEEKISQRGGKTVGYWSTDGYDFNDSKALR-NGKFVGLALDEDNQSDLTDDRISWVAQLKSEFGL-----
FLAV_AZOVI VALFGLGDQVGYPENYLDALGELYSFFKDRGAKIVGSWSTDGYEFESSEAVV-DGKFVGLALDLDNQSGKTDERVAAWLAQIAPEFGLSL----
2fcr      VAIFGLGDAEGYPDNFCD AIEEIHDCFAKQGA K PVGFSNPDDYDYEESKSVR-DGKFLGLPLDMVNDQIPMEKRVAGWVEAVVSETGV-----
FLAV_ENTAG VALFGLGDQLNYSKNFVSAMRILYDLVIARGACVVGWNPREGYKFSFSAALLENNEFVGLPLDQENQYDLTEERIDSWLEKLKPAVL-----
FLAV_ECOLI VALFGCGDQEDYAEYFCDALGTIRDIIEPRGATIVGHWPTAGYHFEASKGLADDDHFVGLAIDEDRQPELTAERVEKWKQISEELHLDEILNA
3chy      AD--GAMSALPVL----MVTAEAKKENIIAAQAGAS-----GYV-VKPFTAATLEEKL N KIFEKLG M-----
```

. . : . .

Flavodoxin-cheY: Pre-processing (prepro≥1500)

```

1fx1      -PKALIVYGSTTGNT-EYTAETIARQLANAG-YEVDSRDAASVEAGGLFEGFDLVLLGCSTWGDDSI-----ELQDDFIPLF-DSLEETGAQGRKVACF
FLAV_DESDE MSKVLIVFGSSTGNT-ESiAQKLEELIAAGG-HEVTLLNAADASAENLADGYDAVLFGCSAWGMEDL-----EMQDDFLSLF-EEFNRFGLAGRKVAAf
FLAV_DESVH MPKALIVYGSTTGNT-EYTaeETIARELADAG-YEVDSRDAASVEAGGLFEGFDLVLLgCSTWGDDSI-----ELQDDFIPLF-DSLEETGAQGRKVACf
FLAV_DESSA MSKSLIVYGSTTGNT-ETAaeYVAEAFENKE-IDVELKNVTDVSVADLNGYDIVLFgCSTWGEETI-----ELQDDFIPLY-DSLENADLKGGKVSf
FLAV_DESGI MPKALIVYGSTTGNT-EGVaEIAKTLNSEG-METTVVNADVTAAPGLAEGYDVVLLgCSTWGDDSI-----ELQEDFVPLY-EDLDRAGLKDKKVGf
2fcr      --KIGIFFSTSTGNT-TEVADFIgKTLGA---KADAPIDVDDVTDQPALKDYDLLFLGAPTWNtG---ADTERSGTSWDEFLYDKLPEVDMKDLPVAlF
FLAV_AZOVI -AKIGLFFGSNTGKT-RKVAKSIKKRFDDet-MSDA-LNVNRVS-AEDFAQYQFLILgTPTLGEGELPGLSSDCENESWEEFL-PKIEGLDFSGKTVALf
FLAV_ENTAG MATIGIFFGSDTGQT-RKVAKLIHQKLDG---IADAPLDVRRAT-REQFLSYPVLLLgTPTLGDELPGVEAGSQYDSWQEFT-NTLSEADLTGKTVALf
FLAV_ANASP SKKIGLFGYGTGTGKT-ESVaEIIRDEFgn---DVVTLHDVSQAE-VTDLNDYQYLIIGCPTWNIGEL-----QSDWEGLY-SELDDVDFNGKLVAYf
FLAV_ECOLI -AITGIFFGSDTGNT-ENiAKMIQKQLGK---DVADVHDIKSS-KEDLEAYDILLgIPTWYYGE-----AQCDWDDFF-PTLEEIDFNGKLVALf
4fxn      -MK--IVYWSGTGNT-EKMAELIAKGIIESG-KDVNTINVSDVINDELL-NEDILILGCSAMGDEVL-----EESEFEPFI-EEIS-TKISGKKVALF
FLAV_MEGEL MVE--IVYWSGTGNT-EAMaNEIEAAVKAAG-ADVesVRFEDTNVDDVA-SKDVILLgCPAMGSEEL-----EDSVVEPFF-TDLA-PKLKGKKVGLf
FLAV_CLOAB -MKISILYSSKTGKT-ERVAKLIEEGVKRSGNIEVKTMNLDAVD-KKFLQESEGIIIfgTPTYAN-----ISWEMKKWI-DESSEFNLEGLGAAf
3chy      ADKELKFLVVDfSTMRIRVRNLLKELGFN--NVEEAEDGVDALNKLQAGGYGFVI---SDWNMPNM-----DGLELL-KTIRADGAMSALPVLM

```

T

```

1fx1      GCGDS-SY-EYFCGA-VDAIEEKLNKLGAElVQD-----GLRIDGD--PRAARDDIVGWAHDVIRGAI-----
FLAV_DESDE ASGDQ-EY-EHFCGA-VPAIEERAKELgATIIAE-----GLKMEGD--ASNDPEAVASfAEDVLKQL-----
FLAV_DESVH GCGDS-SY-EYFCGA-VDAIEEKLNKLGAElVQD-----GLRIDGD--PRAARDDIVGWAHDVIRGAI-----
FLAV_DESSA GCGDS-DY-TYFCGA-VDAIEEKLEKMgAVVIGD-----SLKIDGD--PE--RDEIVSwGSGIADKI-----
FLAV_DESGI GCGDS-SY-TYFCGA-VDVIEKKAELgATLVAS-----SLKIDGE--PD--SAEVLdWAREVLARV-----
2fcr      GLGDAEGYPDNFCDA-IEEIHDCFAKQGAkPVGFSNPDDYDYESKS-VRDGKFLGLPLDMVNDQIPMEKRVAGWVEAVVSETGV-----
FLAV_AZOVI GLGDQVGYPENYLDA-LGELYSFFKDRgAKIVGSWSTDGYEFESSEA-VVDGKFVGLALDLNQSGKTDERVAawLAQIAPEFGLS--L--
FLAV_ENTAG GLGDQLNYSKNFVSA-MRILYDLVIARGAcVVGWNWPREGYKFSFSAALLENNEFVGLPLDQENQYDLTEERIDSwLEKLKPAV-L-----
FLAV_ANASP GTGDQIGYADNFQDA-IGILEEKISQRgGKTGVGYWSTDGYDFNDSKA-LRNGKFVGLALDEDNQSDLTDDRIKSwaQLKSEFGL-----
FLAV_ECOLI GCGDQEDYAEYFCDA-LGTIRDIIePRgATIVGHWPtagYHFEASKGLADDDHFVGLAIDEDRQPELTARVEKwVKQISEELHLDEILNA
4fxn      G-----SY-GWGDGKWMRDFEERMNGYGCVVVET-----PLIVQNE--PDEAEQDCIEFGKKIANI-----
FLAV_MEGEL G-----SY-GWGSGEWMDAWKQRTEDTgATVIGT-----AIVNEM--PDNA-PECKELGEAAAKA-----
FLAV_CLOAB STANSIAGGSDIA---LLTIILNHLMVKGMLVYSg---GVAFGKPKTHLGYVHINEIQENEDENARIfGERiANKVKQIF-----
3chy      VTAEAKK--ENIIAA-----AQAGAS-----GYVV-----KPFTAATLEEKLNKIFeKLGm-----

```

G

Flavodoxin-cheY: Local Pre-processing (locprepro $\geq$ 300)

```

1fx1      --PKALIVYGSTTGNTTEYTAETIARQLANAGYEVDSRDAASVEAGGLFEGFDLVLLGCSTWGDDSI-----ELQDDFIPL--FDSLEETGAQGRKVACF
FLAV_DESVH -MPKALIVYGSTTGNTEYTaETIARELADAGYEVDSRDAASVEAGGLFEGFDLVLLgCSTWGDDSI-----ELQDDFIPL--FDSLEETGAQGRKVACf
FLAV_DESSA -MSKSLIVYGSTTGNTETaEYVAEAFENKEIDVELKNVTDVSVADLGNGYDIVLFgCSTWGEEEEI-----ELQDDFIPL--YDSLENADLKGGKVSvf
FLAV_DESGI -MPKALIVYGSTTGNTEGVaEAIAKTLNSEGMETTVNVADVTAPGLAEGYDVVLLgCSTWGDEI-----ELQEDFVPL--YEDLDRAGLKDKKVGvf
FLAV_DESDE -MSKVLIVFGSSTGNTESiaQKLEELIAAGGHEVTLNAADASAENLADGYDAVLFgCSAWGMEDL-----EMQDDFSL--FEEFNRFGLAGRKVAAf
4fxn      --MK--IVYWSGTGNTTEKMAELIAKGIIESGKDVNTINVSDVNIDELLN-EDILILGCSAMGDEVL-----E-ESEFEPF--IEEIS-TKISGKKVALF
FLAV_MEGEL -MVE--IVYWSGTGNTEAMaNEIEAAVKAAGADVSVRFEDTNVDDVAS-KDVILLgCPAMGSEEL-----E-DSVVEPF--FTDLA-PKLKGKKVGLf
2fcr      ---KIGIFFSTSTGNTTEVADFIGKTLGAKADAPI--DVDDVTDPQALKDYDLLFLGAPTWNTGAD----TERSGTSWDEFL-YDKLPEVDMKDLPVAIF
FLAV_ANASP -SKKIGLFYGTQTGKTESVaEIIRDEFGNDVVTLH--DVSQAEV-TDLNDYQYLIIgCPTWNIGEL-----QSDWEGL--YSELDDVDFNGKLVAYf
FLAV_AZOVI --AKIGLFFGSNTGKTRKVaKSIKKRFDDETMSDA-LNVNRVSA-EDFAQYQFLILgTPTLGEGELPGLSSDCENESWEEF--LPKIEGLDFSGKTVALf
FLAV_ENTAG -MATIGIFFGSDTGQTRKVaKLIHQKLDG--IADAPLDVRRATR-EQFLSYPVLLLgTPTLGDGELPGVEAGSQYDSWQEF--TNTLSEADLTGKTVALf
FLAV_ECOLI --AITGIFFGSDTGNTENiaKMIQKQLGKDVADVH--DIAKSSK-EDLEAYDILLLgIPTWYYGEA-----QCDWDDF--FPTLEEIDFNGKLVAlf
FLAV_CLOAB --MKISILYSSKTGKTERVaKLIEEGVKRSGNIEVKTMNLDAVDKKFLQESEGIIFgTPTYA-----NISWEMKKWIDESSEFNLEGLGAAlf
3chy      ADKELKFLVVDDFSTMRRIVRNLLKELGFNNVEEAEDGVDALNKLQ-AGGYGFVI---SDWNMPNM-----DGLEL--LKTIRADGAMSALPVLM

```

```

1fx1      GCGDS--SY-EYFCGA-VD--AIEEKLKNLGAIVQD-----GLLRID--GDPRAARDDIVGWAHDVRGAI-----
FLAV_DESVH GCGDS--SY-EYFCGA-VD--AIEEKLKNLgAEIVQD-----GLRID--GDPRAARDDIVGwAHDVRGAI-----
FLAV_DESSA GCGDS--DY-TYFCGA-VD--AIEEKLEKMgAVVIGD-----SLKID--GDPE--RDEIVSwGSGIADKI-----
FLAV_DESGI GCGDS--SY-TYFCGA-VD--VIEKKAEELgATLVAS-----SLKID--GEPD--SAEVLdWAREVLARV-----
FLAV_DESDE ASGDQ--EY-EHFCGA-VP--AIEERAKELgATIIAE-----GLKME--GDASNDPEAVASfAEDVLKQL-----
4fxn      GS-----Y-GWGDGKWMR--DFEERMNGYGCVVET-----PLIVQ--NEPDEAEQDCIEFGKKIANI-----
FLAV_MEGEL GS-----Y-GWGSGEWMD--AWKQRTEDTgATVIGT-----AI-VN--EMPDNA-PECKElGEAAAKA-----
2fcr      GLGDAE-GYPDNFCDA-IE--EIHDCFAKQGAKPVGFSNPDDYDYEESKSVRD-GKFLGLPLDMVNDQIPMEKRVAGWVEAVVSETGV-----
FLAV_ANASP GTGDQI-GYADNFQDA-IG--ILEEKISQRgGKTVGYWSTDGYDFNDSKALRN-GKFVGLALDEDNQSDLTDDRIKSWVAQLKSEFGL-----
FLAV_AZOVI GLGDQV-GYPENYLDA-LG--ELYSFFKDRgAKIVGSWSTDGYEFESSEAVVD-GKFVGLALDLDNQSGKTDERVAAwLAQIAPEFGLS--L--
FLAV_ENTAG GLGDQL-NYSKNFVSA-MR--ILYDLVIARgACVVGNWPREGYKFSFSAALLENNEFVGLPLDQENQYDLTEERIDSWLEKLPAV-L-----
FLAV_ECOLI GCGDQE-DYAEYFCDA-LG--TIRDIIEPRgATIVGHWPTAGYHFEASKGLADDDHFVGLAIDEDRQPELTAERVEKWVKQISEELHLDEILNA
FLAV_CLOAB STANSIAGGSDIALLTILNHLMVKgMLVYSGGVAFGKPKTHLGYVH-----INEIQENEDENARIfGERiANKVKQIF-----
3chy      VTAEA---KKENIIAA-----AQAGAS-----GYVVK-----PFTAATLEEKLNKIFEKLGM-----

```

G

PSI-PRALINE

Multiple alignment of distant sequences using PSI-BLAST

- n Perform a PSI-BLAST search for each sequence

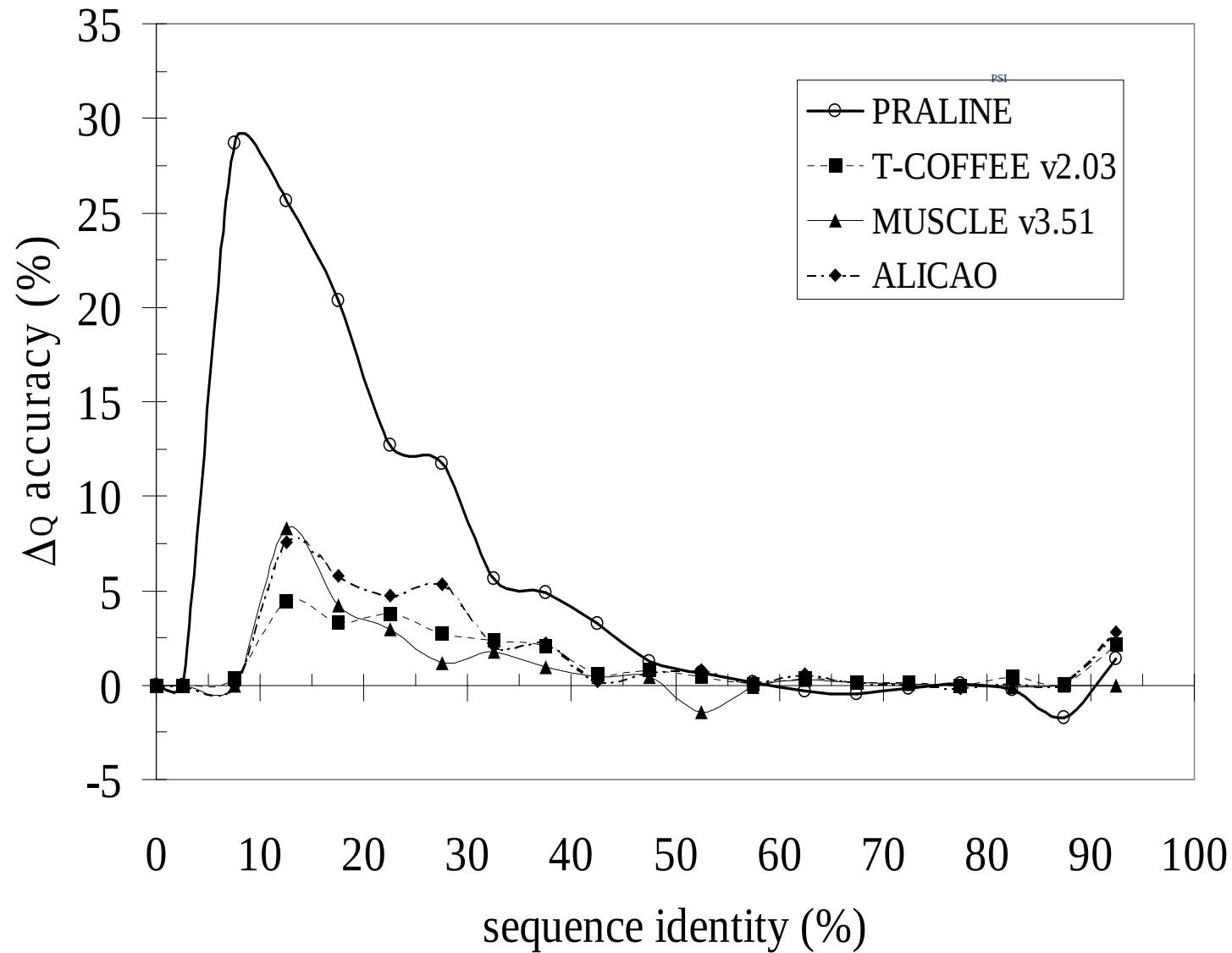
- n Keep putative homologs found as 'background' sequences

 - Make local pre-profile for each sequence

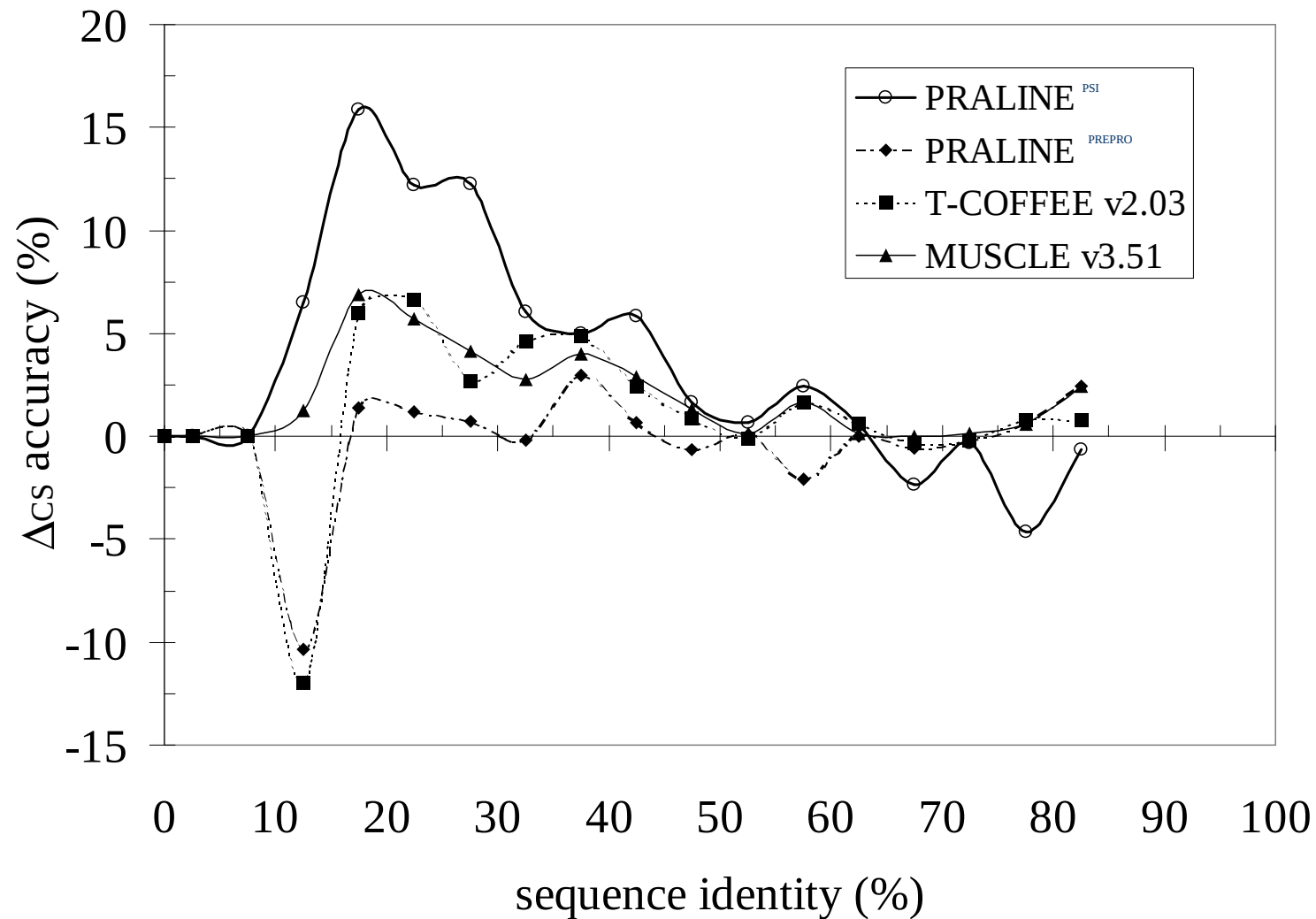
 - Align original sequences using extended information from homologous sequences



Pair-wise alignment



Multiple alignment



Example: methyltransferases

PRALINE BASIC (0.10)

```
1GZ0A SEIYG..... IHAVQALLER APERFQEVFI LKG.REDKRL LP....LIHA LESQGVVIQL ANRQY..... LDEKSDGAVH Q.....G IIARVK.PGR Q
1IPAA MRITSTANPR IKELARLLER KHRDSQRRFL IEGAREIERA LQAGIELEQA LVWEGGLNPE EQQVYAALLA LLEVSEAVLK KLSVRDNPAG LIALARMPER .
                                     ^ ^^^^^^
```

ALICAO (0.00)

```
1GZ0A .....SEIYG IHAVQALLER APERFQEVFI LKG.REDKRL LP....LIHA LESQGVVIQL ANRQYLDEKS DGAVHQGIIA RVKPGRQ... ..
1IPAA MRITSTANPR IKELARLLER KHRDSQRRFL IEGAREIERA LQAGIELEQA LVWEGGLNPE EQQVYAALLA LLEVSEAVLK KLSVRDNPAG LIALARMPER
```

MUSCLE v3.51 (0.15)

```
1GZ0A SEIYG..... IHAVQALLER APERFQEVFI LKG.REDKRL LP....LIHA LESQGVVIQL ANRQY..... LDEKSDGAVH Q.....G IIARVKPGRQ
1IPAA MRITSTANPR IKELARLLER KHRDSQRRFL IEGAREIERA LQAGIELEQA LVWEGGLNPE EQQVYAALLA LLEVSEAVLK KLSVRDNPAG LIALARMPER
                                     ^ ^^^^^^^^^^
```

T-COFFEE v2.03 (0.15)

```
1GZ0A .....SEIYG IHAVQALLER APERFQEVFI LKG.REDKRL LP....LIHA LESQGVVIQL ANRQY..... LDEKSDGAVH Q.....G IIARVKPGRQ
1IPAA MRITSTANPR IKELARLLER KHRDSQRRFL IEGAREIERA LQAGIELEQA LVWEGGLNPE EQQVYAALLA LLEVSEAVLK KLSVRDNPAG LIALARMPER
                                     ^ ^^^^^^^^^^
```

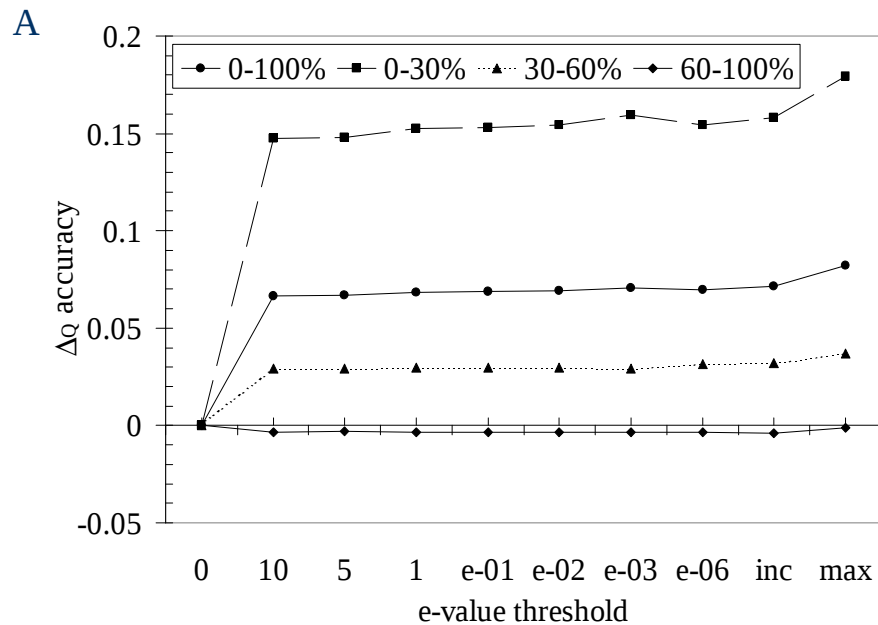
PSI-PRALINE (0.92)

```
1GZ0A .....SE IYGIHAVQAL LERAPERFQE VFILKGREDK RLLPLIHALE SQGVVIQLAN RQYLDEKSDG AVHQGIARV KPGRQ
1IPAA MRITSTANPR IKELARLLER KHRDSQRRFL IEGAREIERA LQAG.IELEQ ALVWEGGLNP EEQQVY... AALLALLEVS EAVLKKLSVR DNPAGLIALA RMPER
                                     ^^ ^^^^^^^^^^ ^^^^ ^^^ ^^^^^^^^^^ ^^^^^^ ^^^^^^^^^^ ^^^^^^^^^^ ^^^^^^
```

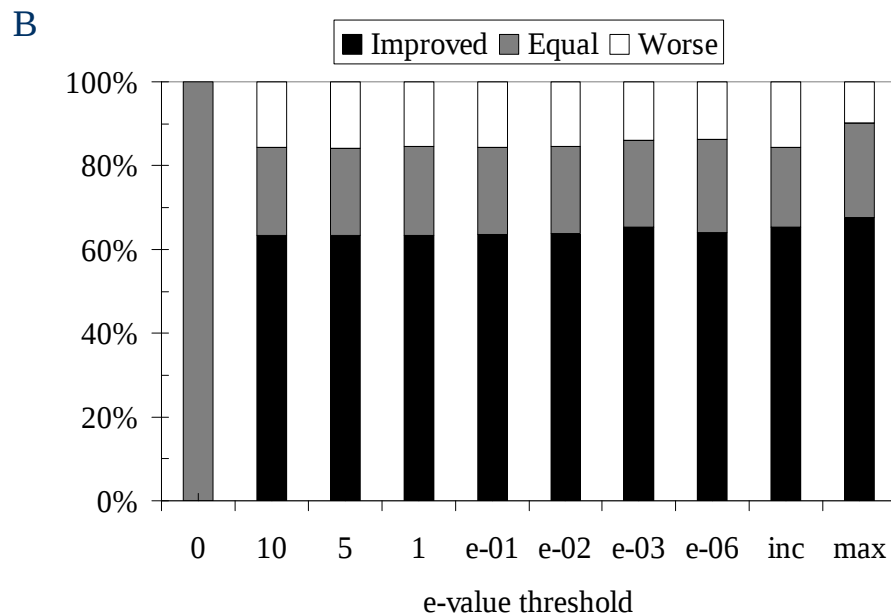
HOMSTRAD REFERENCE

```
1GZ0A .....SE IYGIHAVQAL LERAPERFQE VFILKGREDK RLLPLIHALE SQGVVIQLAN RQYLDEKSDG AVHQGIARV KPGRQ
DSSP ..... E EESHHHHHHH HHS GGGB E EEEESS T TTHHHHHHHH HHT EEEEE SHHHHHHTTS TTEEEEE
1IPAA MRITSTANPR IKELARLLER KHRDSQRRFL IEGAREIERA LQAGIE.LEQ ALVWEGGLNP EEQQVYAALL ...ALLEVS EAVLKKLSVR DNPAGLIALA RMPER
DSSP EE TTSHH HHHHHGGGSH HHHHHHTEEE EESHHHHHHH HHTT .B E EEEETT H HHHHHH ....EEEE HHHHHH S SS SEEEEE E
```

The effects of using E-value thresholds of increasing stringency in PRALINEPSI on the 624 HOMSTRAD pairwise alignments.



(A) The difference between the average Q scores of PRALINEPSI and the basic PRALINE method



(B) The distributions of improved, equal and worsened cases compared with the basic PRALINE method for each E-value threshold.

The 'inc' column is the PRALINEPSI incremental strategy starting from a threshold of 10^{-6} , and the 'max' column is PRALINEPSI's theoretical upper limit for the tested threshold range.

Strategies for multiple sequence alignment

- Profile pre-processing
- **Secondary structure-induced alignment (Praline-SS)**
- Globalised local alignment
- Matrix extension

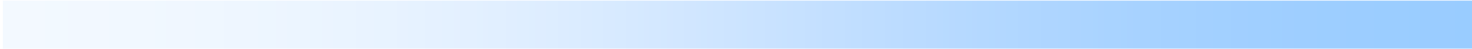
Objective: integrate secondary structure information to anchor alignments and avoid error



Additional strategies for multiple sequence alignment

- Matrix extension (T-coffee)
- Profile pre-processing (Praline)
- **Secondary structure-induced alignment**

Objective: try to avoid (early) errors

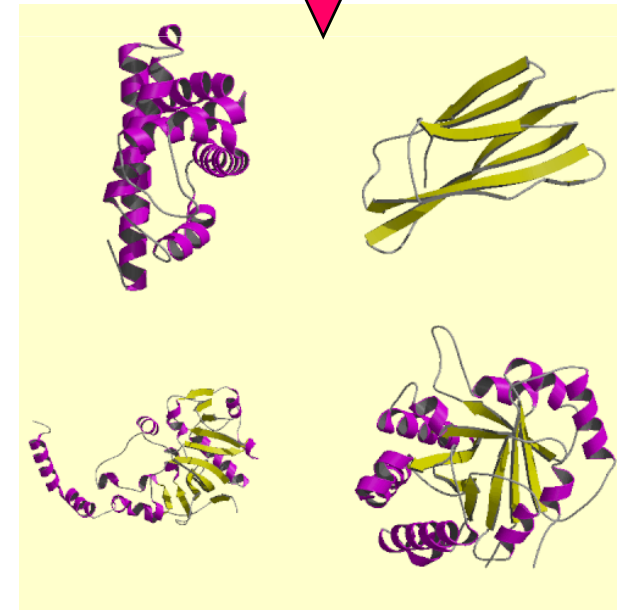
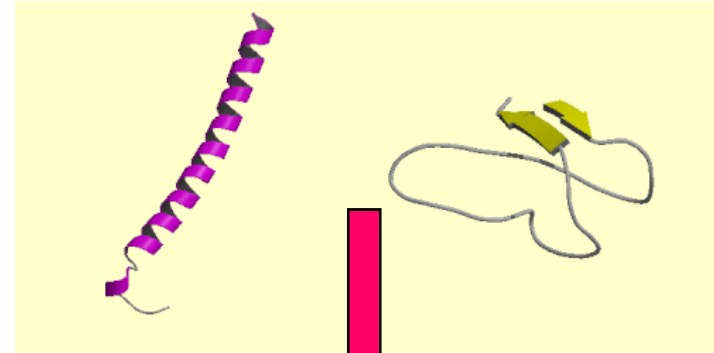


Protein structure hierarchical levels

PRIMARY STRUCTURE (amino acid sequence)

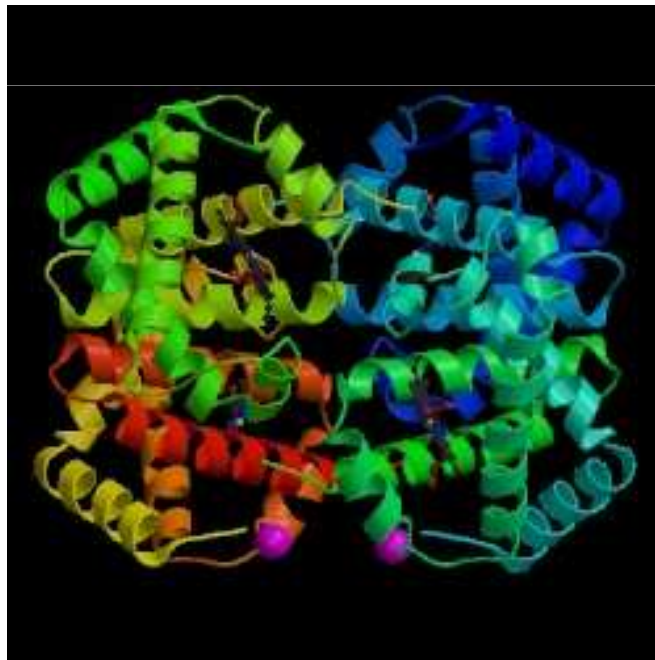
VHLTPEEKSAVTALWGKVNVDE
VGGEALGRLLVVYPWTQRFFE
SFGDLSTPDAMGNPKVKAHG
KKVLGAFSDGLAHLNLIKGTFA
TLSELHCDKLHVDPENFRLLGN
VLVCVLAHHFGKEFTPPVQAAY
QKVVAGVANALAHKYH

SECONDARY STRUCTURE (helices, strands)



TERTIARY STRUCTURE (fold)

QUATERNARY STRUCTURE (oligomers)



Why use (predicted) structural information

n **“Structure more conserved than sequence”**

- Many structural protein families (e.g. globins) have family members with very low sequence similarities. For example, globin sequences identities can be as low as 10% while still having an identical fold.

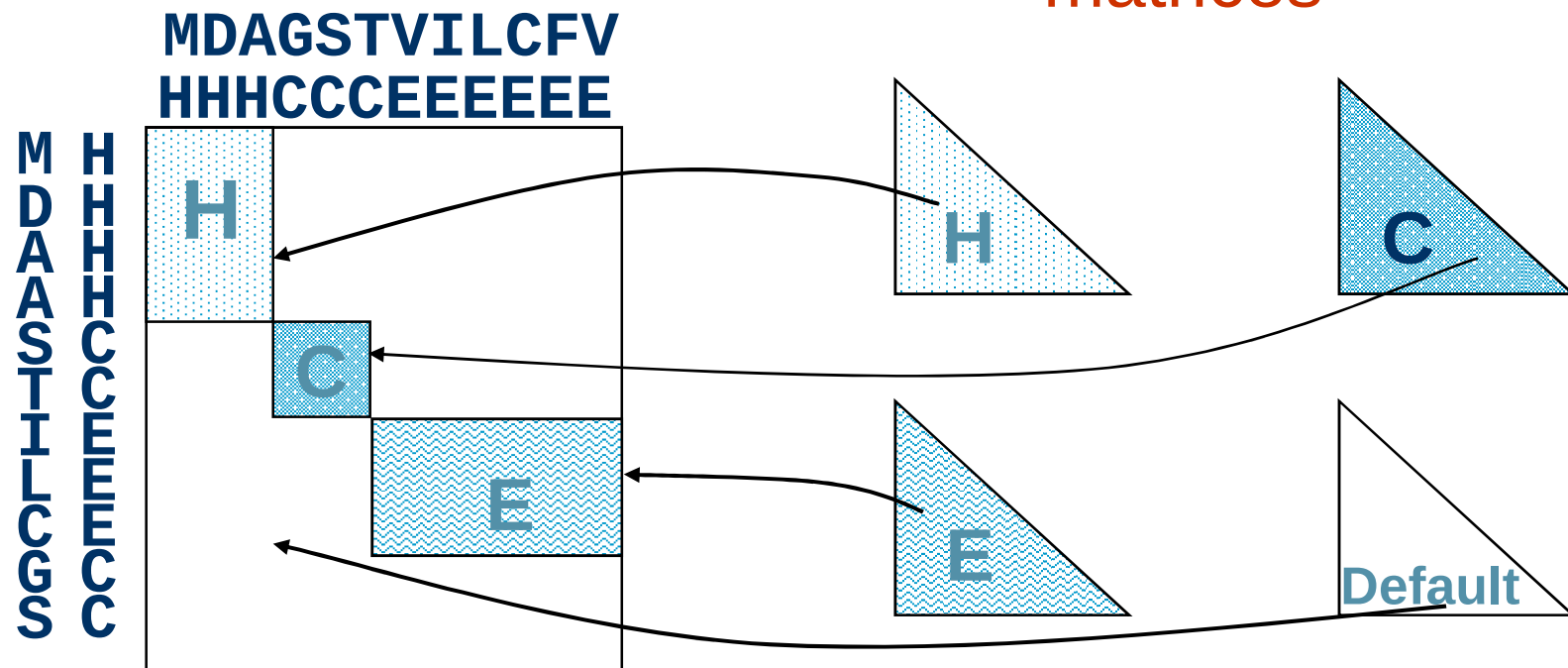
n This means that you can still observe equivalent secondary structures in homologous proteins even if sequence similarities are extremely low.

n But you are dependent on the quality of prediction methods. For example, secondary structure prediction is currently at 76% correctness. So, 1 out of 4 predicted amino acids is still incorrect.

How to combine secondary structure and amino acid information

Dynamic programming
search matrix

Amino acid
substitution
matrices

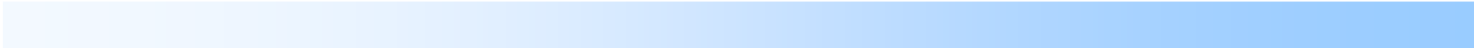


Using predicted secondary structure

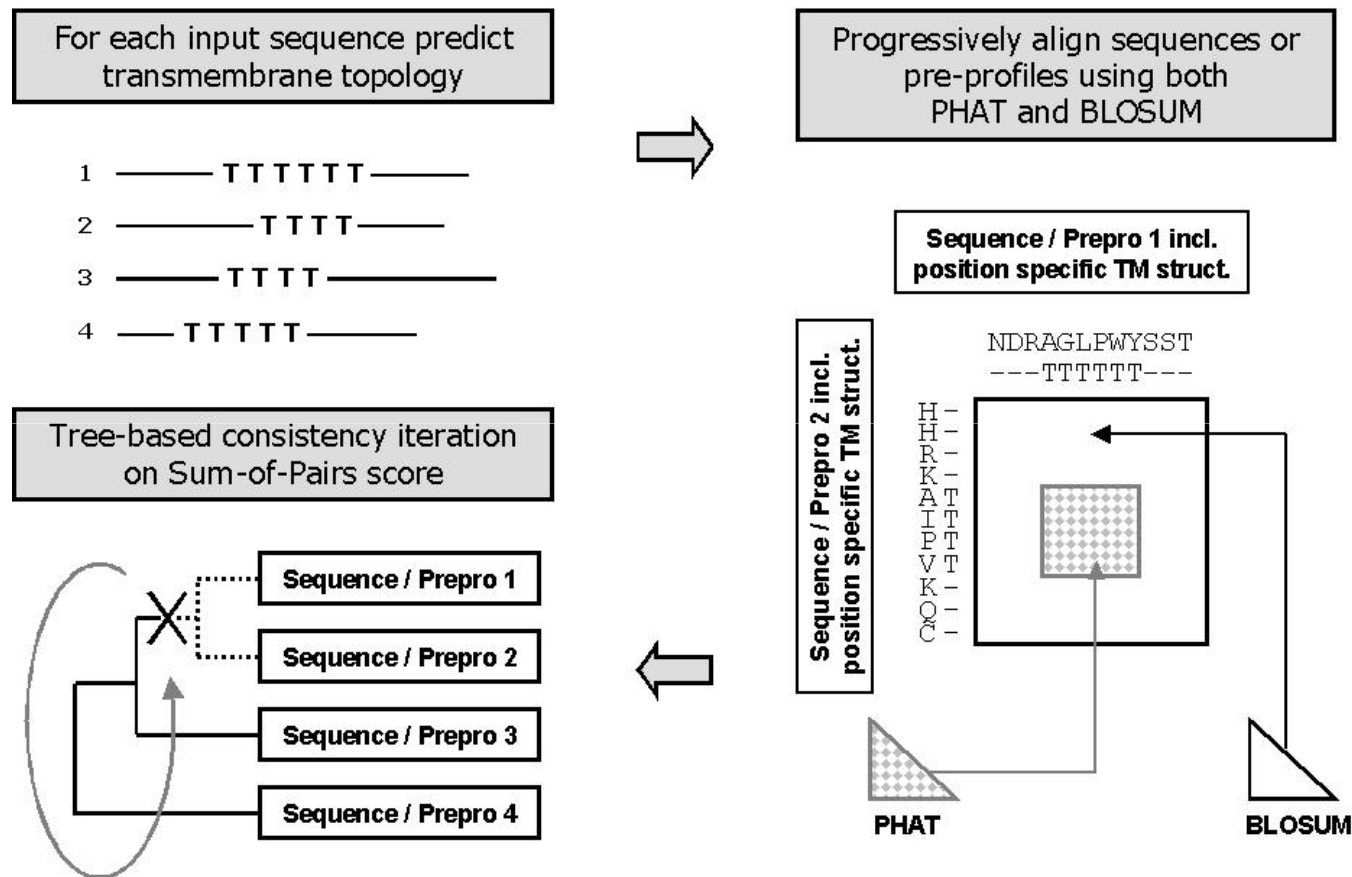
```
1fx1      -PK-ALIVYGSTTGNTTEYTAETIARQLANAG- YEVDSDRAASVEAGGLFEFGFDLVLLGCSTWGDSSI-----ELQDDFIPLFDS-LEETGAQGRKVACF
          e eeee b sssshhhhhhhhhhhhhhhh eeeee stt ttttt seeee b ee sss ee tthhhhhh ttss tt eeeee
FLAV_DESVH MPK-ALIVYGSTTGNTTEYTaETIARELADAG- YEVDSDRAASVEAGGLFEFGFDLVLLgCSTWGDSSI-----ELQDDFIPLFDS-LEETGAQGRKVACF
          e eeeee hhhhhhhhhhhhhhhh eeeee eeeee hhhhhh eeeee
FLAV_DESGI MPK-ALIVYGSTTGNTTEGVaEAIakTLNSEG- METTVVNVADVtAPGLAEGYDVVLLgCSTWGDDEI-----ELQEDFVPLYED-LDRAGLKDKKVGvf
          e eeeee hhhhhhhhhhhhhh eeeee hhhhhh eeeee hhhhhh eeeee
FLAV_DESSA MSK-SLIVYGSTTGNTETAaEYVAEAFENKE- IDVELKNVTDVSVADLGNgyDIVLFgCSTWGEIEI-----ELQDDFIPLYDS-LENADLKgKKVSVf
          eeeee hhhhhhhhhhhhhh eeeee eeeee hhhhhh h eeeee
FLAV_DESDE MSK-VLIVFGSSTGNTESiaQKLEELIAAGG- HEVTLNNAADASAENLADGYDAVLFgCSAWGMEDL-----EMQDDFSLFEE-FNRFLAGRKVAaf
          eeee hhhhhhhhhhhhhh eeeee hhhhhhhhhheeeee hhhhhh hh eeeee
2fcr      --K-IGIFFSTSTGNTTEVADFIGKTLGAK--ADAPIDVDVTPQALKDyDLLfLgAPTWNTGAD----TERSGTSWDEFLYDKLPEVDMKDLPAIf
          eeeee sssshhhhhhhhhhhhhggg b eeggg s gggggg seeeee stt s s s sthhhhhhhtggg tt eeeee
FLAV_ANASP SKK-IGLFYGTQTGKTESVaEIIRDEFgND--VVTl-HDVSQAE-VTDLNDYQYLIIgCPTWNIGEL-----QSDWEGLYSE-LDDVDfNGKLVAyf
          eeeee hhhhhhhhhhhhhh eee hhh hhhhhhhheeeee hhhhhhhh eeeee
FLAV_ECOLI -AI-TGIFFGSDTGNTENiaKMIQKQLGKD--VADV-HDIAKSS-KEDLEAYDILLlgIPTWYYGEA-----QCDWDDFFPT-LEEIDfNGKLVALf
          eee hhhhhhhhhhhh eee hhh hhhhhhhheeeee hhhhh eeeee
FLAV_AZOVI -AK-IGLFFGSNTGKTRKVaKSIKKRFDDet-MSDA-LNVNRVS-AEDFAQYQfLILgTPTLGEgELPGLSSDCENESWEEFLPK-IEGLDfSGKTVALf
          eee hhhhhhhhhhhh hhh hhhhhhhheeeee hhhhhhhh eeeee
FLAV_ENTAG MAT-IGIFFGSDTGQTRKVaKLIHQKLDG---IADAPLDVRRAT-REQFLSYPLVLLgTPTLGDGELPGVEAGSgYDSWQEFtNT-LSEADLTGKTVALf
          eeee hhhhhhhhhhhh hhh hhhhhhhheeeee hhhhh eeeee
4fxn      ----MKIVYWSGTGNTTEKMAELIAKGIIESG-KDVNTINVSdVNIdeLLNE-DILILGCSAMGDEVL-----E-ESEfEPfIEE-IST-KISGKKVALf
          eeeee sssshhhhhhhhhhhhhhhh eeeett stttt seeeee bttb tthhhhhh hst t tt eeeee
FLAV_MEGEL M---VEIVYWSGTGNTTEAMaNEIEAAVKAAG-ADVesVRfEDTnVDDVASK-DVILLgCPAMGSEEL-----E-DSVVEPFfTD-LAP-KLKgKKVGLf
          eeeee hhhhhhhhhhhh eeeee hhhhhhhh eeeee
FLAV_CLOAB M-K-ISILYSSKTGKTERVaKLIIEGVKRSgNIEVKTmNL-DAVDKKfLQESegIIfgTPTY-YANI-----SWEMKKWIDE-SSEfNLEgKLGAaf
          eee hhhhhhhhhhhh eeeee hhhhhhhh eee hhhhhhhh eeeee
3chy      ADKELKfLVDDfSTMRIRVNLLKELGfNN-VEEAEDGV-DALNKLQAGYGFVISD---WNMPNM-----DGLELLKtIRADGAMSALPVLmV
          tt eeee s hhhhhhhhhhhhhh eeeessh hhhhhhhh eeeee s sss hhhhhhhh tttt eeee

1fx1      GCGDS-SY-EYFCGAVDAIEEKLNKGAEIVQD-----GLRIDGD--PRAARDDIVGWAHDVRGAI-----
          eee s ss stthhhhhhhhhhhhhh ee s eeeee gggghhhhhhhhhhhhhh
FLAV_DESVH GCGDS-SY-EYFCGAVDAIEEKLNKGAEIVQD-----GLRIDGD--PRAARDDIVGWAHDVRGAI-----
          eee hhhhhhhhhhhh eeeee eeeee hhhhhhhhhhhhhh
FLAV_DESGI GCGDS-SY-TYFCGAVDVIIEKAEELgATLVAS-----SLKIDGE--P--DSAEVLdWAREVLARV-----
          eee hhhhhhhhhhhh eeeee hhhhhhhhhhhh
FLAV_DESSA GCGDS-DY-TYFCGAVDAIEEKLEKMGaVVIGD-----SLKIDGD--P--ERDEIVSwGSGIADKI-----
          hhhhhhhhhhhh eeeee e eee
FLAV_DESDE ASGDQ-EY-EHFCGAVPAIEERAKELgATIIE-----GLKMEGD--ASNDPEAVASfAEDVLKQL-----
          e hhhhhhhhhhhh eeeee ee hhhhhhhhhh
2fcr      GLGDAEGYPDNFCDAIEEIHDCfAKQgAKPVGfSNPDDYDYEEKSVRD-GKfLGLPLDMVNDQIPMEKRVAGWVEAVVSETGV-----
          eee ttt ttstthhhhhhhhhhhhhh eee b gggs s tteet teeseeeeett ss hhhhhhhhhhhhhhhh
FLAV_ANASP GTGDQIGYADNFQDAIGILEEKISQRgKTKVGYSTDGyDFNDsKALR-NGKFVGLALDEdNQSDLTDDRIKSwAQLKSEfGL-----
          hhhhhhhhhhhh eeee hhhhhhhhhhhhhh
FLAV_ECOLI GCGDQEDYAIEYfCDALGTIRDIIePRgATIVGHWPtAGYHfEASgKLADDDHfVGLAIDEDRQPELTARVEKwVKQISEELHLDEILNA
          hhhhhhhhhhhh eeee hhhhhhhhhhhhhh
FLAV_AZOVI GLGDQVGYPENYLDALGELYSfFKDRgAKIVGSWSTDGyEFESSEAVVD-GKFVGLALDLdNQSGKTDERVAawLAQIAPEfGLS--L--
          e hhhhhhhhhhhh eeeee hhhhhhhhhh
FLAV_ENTAG GLGDQLNYSKNfVFSAMRILYDLVIARgACVVGNWPREGYKfSfSAALLENNEfVGLPLDQENQYDLTEERIDSwLEKLKPAV-L-----
          hhhhhhhhhhhh eee hhhhhh hhhhhhhhhhhh
4fxn      G-----SYGWGDGKWMRDfEERMNGYGCvVvET-----PLIVQNE--PDEAEQDCIEfGKKIANI-----
          e eesss shhhhhhhhhhhhhh ee s eeeee gggghhhhhhhhhhhh
FLAV_MEGEL G-----SYGWGSGEWMDAWKQRTEDTgATVIGT-----AIVNEM--PDNAPE-CKEIGEAaAKA-----
          hhhhhhhhhh eeeee eee h hhhhhhhh
FLAV_CLOAB STANSIA-GGSdIALLTILNHLMVK-gMLVYSg---GVAfGKPKThLG----YVHINEI--QENEDENARIfGERiANKV--KQIF--
          hhhhhhhhhhhh eeeee hhh hhh hhhhhhhhhhhh h
3chy      -----TAEAKKENIIAAQAGASGY-----VVK---P-FTAATLEELKNKIfEKLGM-----
          ess hhhhhhhhhh see ees s hhhhhhhhhhhhhh
```

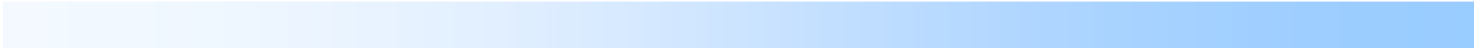
PRALINETM (Pirovano et al., 2008)

- n Membrane-bound proteins are a special class: different *hydrophobicity* patterns
 - n 20 – 30% of all ORFs are likely to be transmembrane (Wallin and Von Heijne, 1998)
 - n Less than 2% of all solved structures show a membrane topology (www.pdb.org)
- 

PRALINETM strategy



Substitution matrices

- n **JTT (Jones et al., 1994)**
polar residues are highly conserved,
hydrophobic residues more interchangeable.
 - n **PHAT (Ng et al., 2000)**
use background frequencies characteristic of
twilight zone rather than the amino acid
frequencies of the database.
- 

Transmembrane topology predictors

- n **HMMTOP** (Tusnády and Simon, 2001)
- n **TMHMM** (Krogh et al., 2001)
- n **PHOBIUS** (Käll et al., 2005)

However, not many techniques have been developed to improve alignment of transmembrane proteins

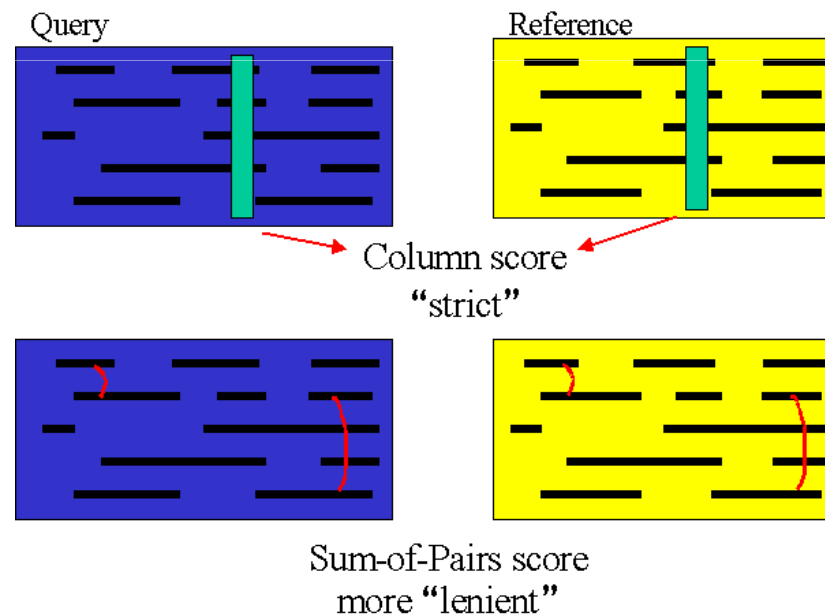
- n **STAM** (Shafrir and Guy, 2004)



Benchmark

- n **BALIBASE v2.0** transmembrane set:
435 aligned sequences – 8 families
av. seq len = 567 – from 2 to 14 TM helices

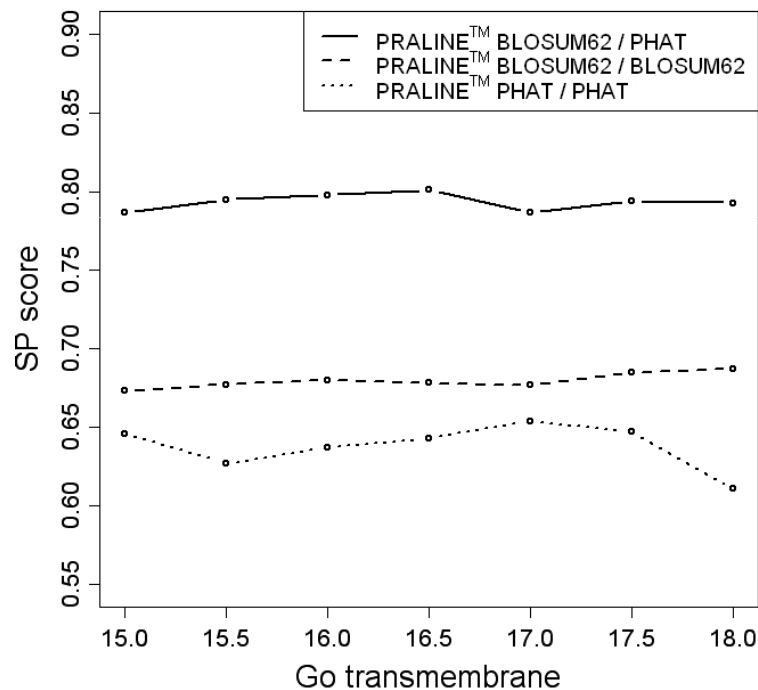
- n **Accuracy:**



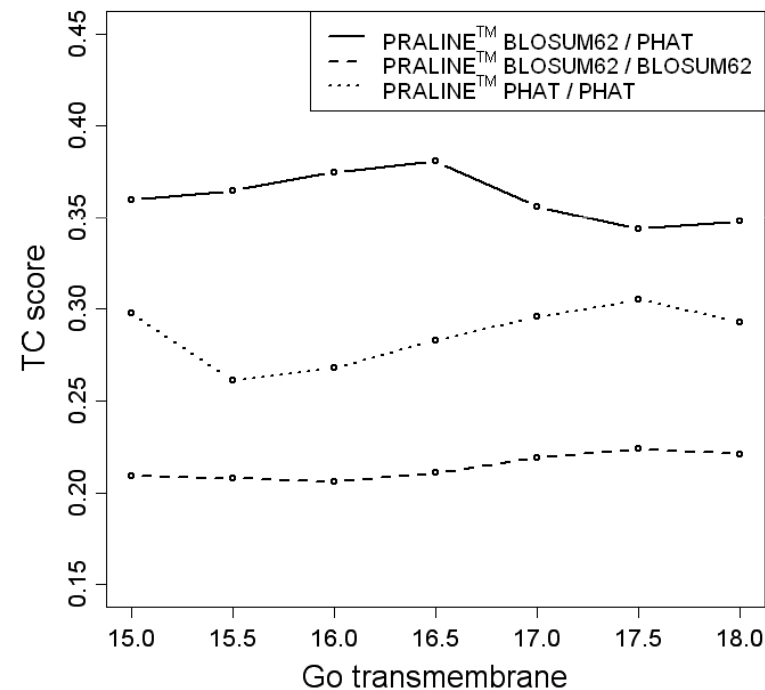
Independent contributions

PHAT matrix and gap values

a)



b)



Strategies for multiple sequence alignment

- Profile pre-processing
- Secondary structure-induced alignment
- **Matrix extension**

Objective: try to avoid (early) errors



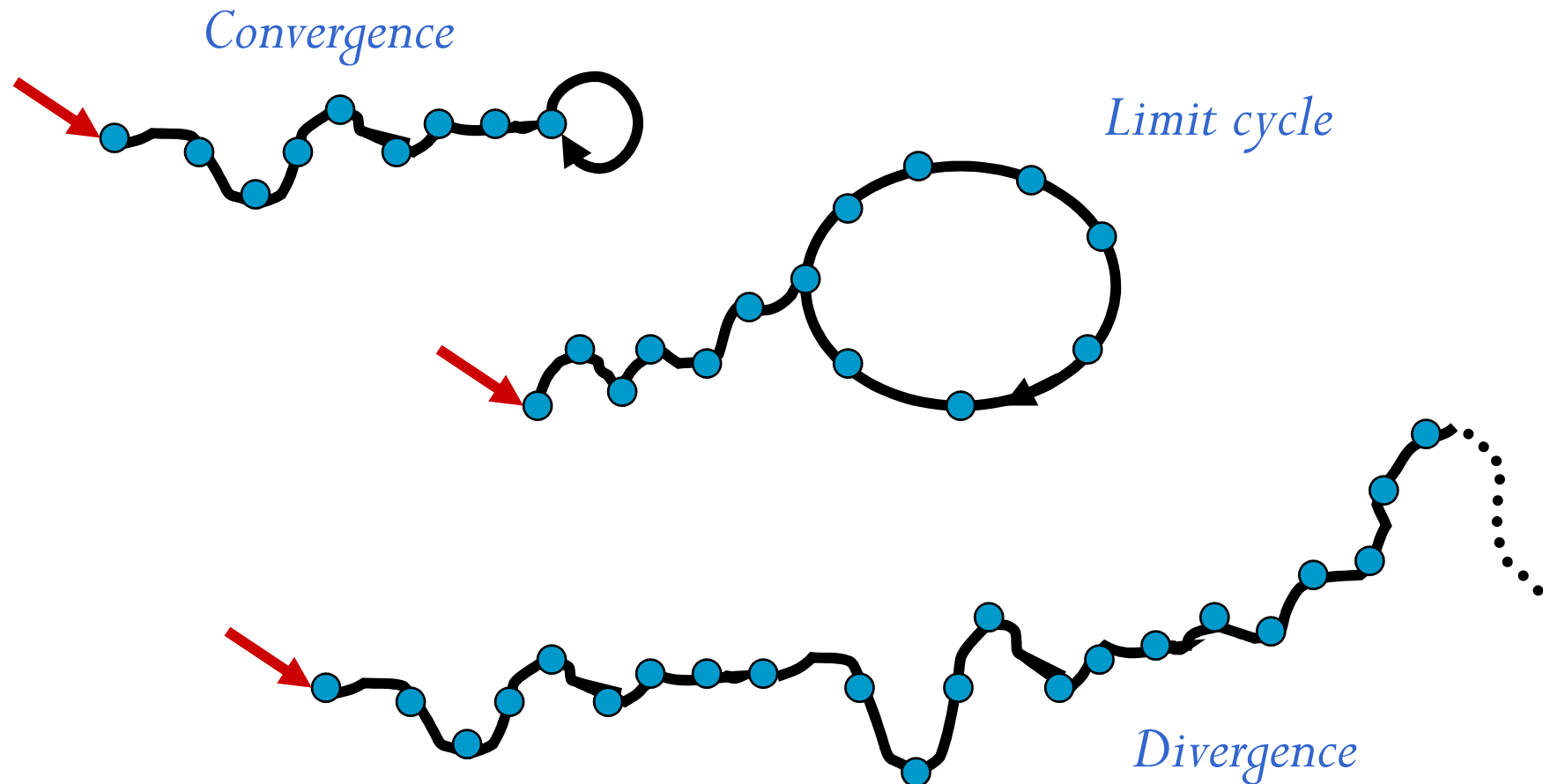
Multiple alignment methods

- q Multi-dimensional dynamic programming
 - > extension of pairwise sequence alignment.
- q Progressive alignment
 - > incorporates phylogenetic information to guide the alignment process
- q **Iterative alignment**
 - > correct for problems with progressive alignment by repeatedly realigning subgroups of sequence

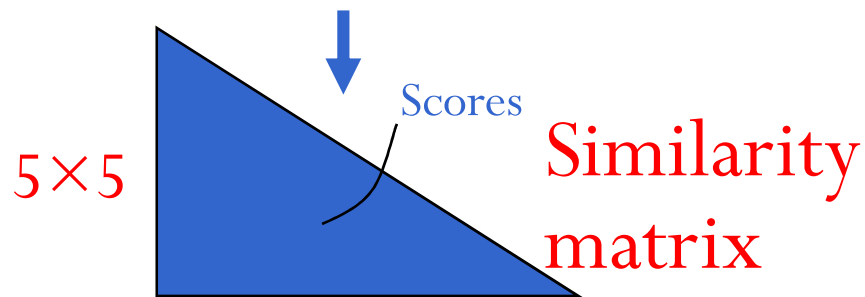
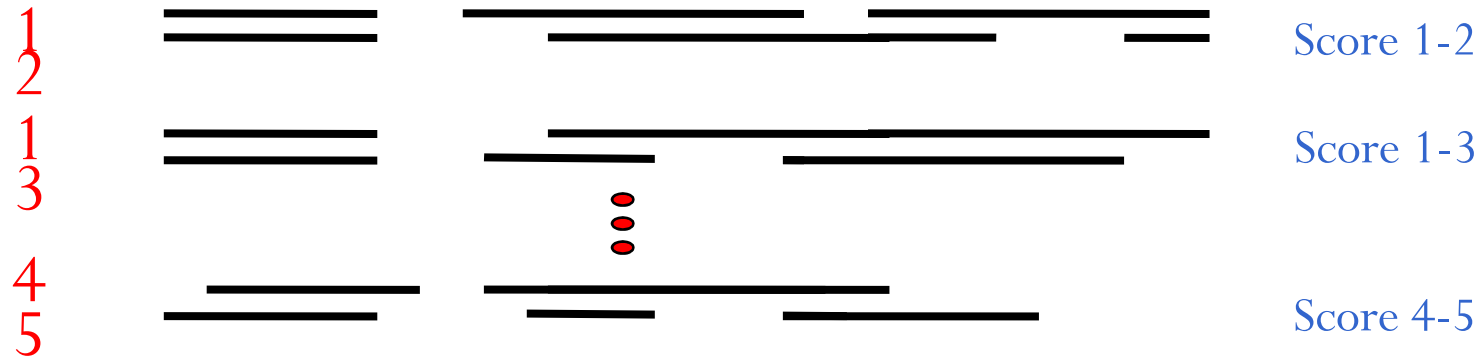


Iterative strategies

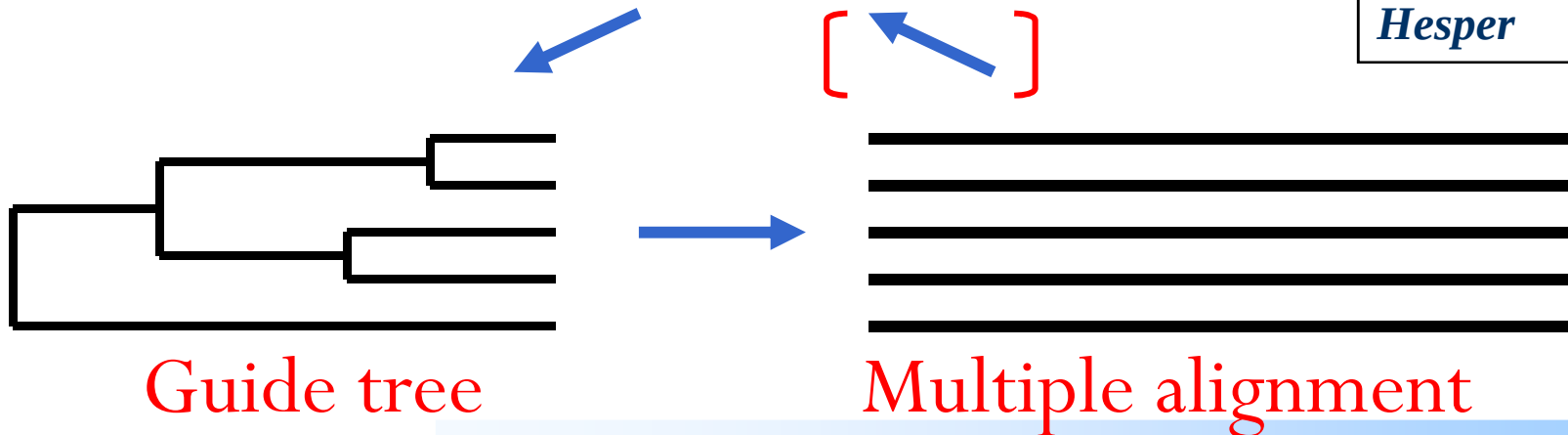
Iteration can help in cases where one can learn from the data produced in a preceding step, so that the next step can be taken in a 'more informed' way.



Iterate similarity matrix, guide tree and MSA

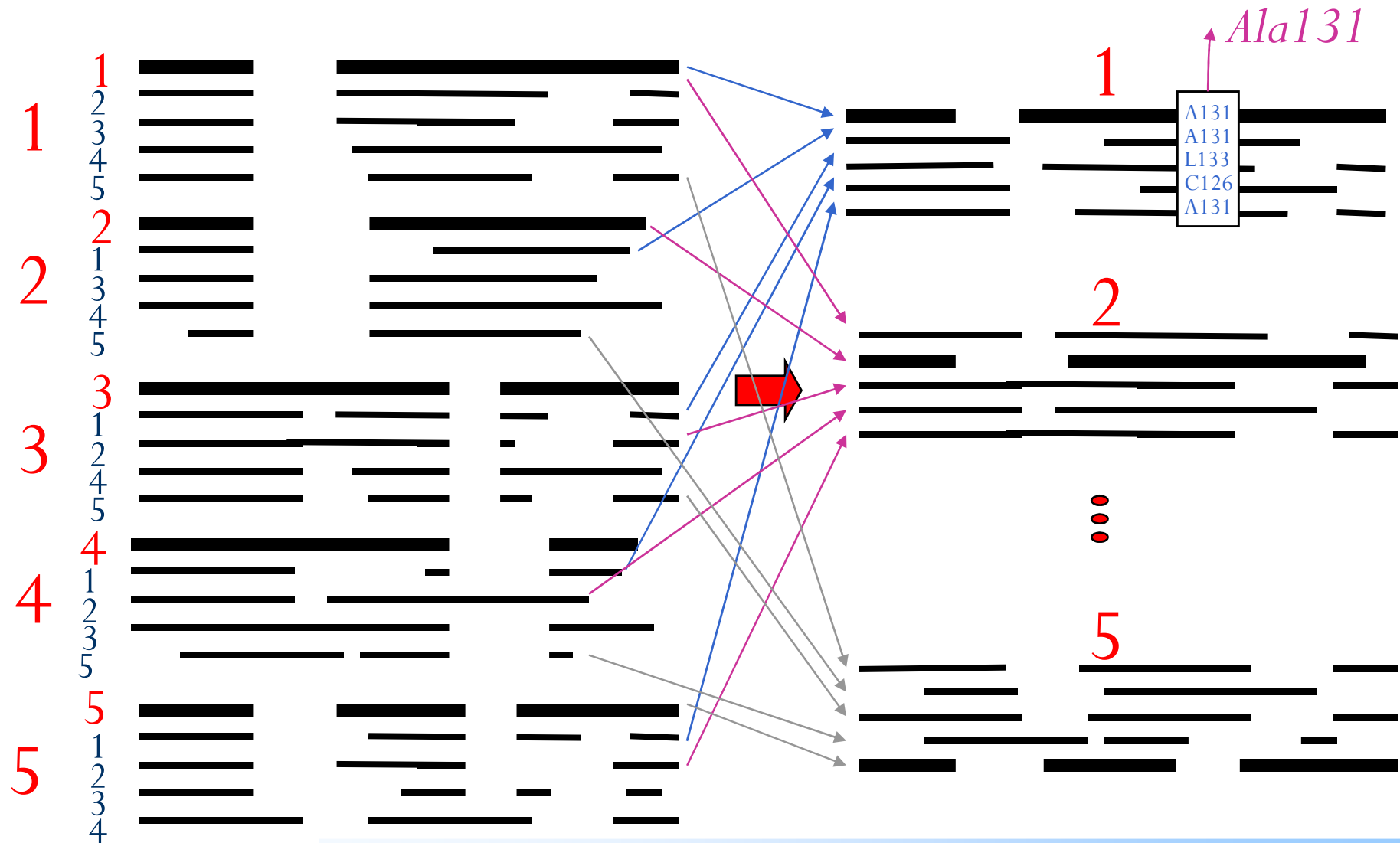


This way of iterating was already implemented in 1984 by Hogeweg and Hesper



Pre-profile alignment

Alignment consistency



Completely
consistently
aligned amino
acids

Avrg Consist	86665656444444466666666666666665666555556555555655565444443444443344455666666666666889999
Conservation	73663057433334163464534444*7467100000011010011000000010434744645443225474454448434301000000

Iteration 0 SP= 135136.00 AvSP= 10.473 SId= 3838 AvSId= 0.297

Consistency values are scored from 0 to 10; the value 10 is represented by the corresponding amino acid (red)

Flavodoxin-cheY consistency scores (PRALINE prepro=1500)

1fx1 -42444IVYGSTTGNTEYTAETIARQL88666666577777775667888DLVLLGCSTW7776 ---995476666769-77888788AQGRKVACF
FLAV_DESVH -34444IVYGSTTGNTEYTAETIAREL77666666577777775667888DLVLLGCSTW7776 ---995476666769-77888788AQGRKVACF
FLAV_DESSA -33444IVYGSTTGNTE9999988877765577766888899666686YDIVLFGCSTW7777 ---99646666779-88SL98ADLKGGKVSF
FLAV_DESGI -34444IVYGSTTGNTEGVA9999999997655567777788666678DVVLLGCSTW7777 ---99546666779-88887688888KKVGVF
FLAV_DESDE -44777IVFGSSTGNTE988777666655667777889999777777YDAVLFGCSAW8887 ---99758777779-8887766777GRKVAAF
4fxn -32222IVYWSGTGNTE888888887666778888888888NI8888586DILILGCSA88888 ---8-8888886-66665378ISGKKVALF
FLAV_MEGEL -12222IVYWSGTGNTEAMA888888888888888855555555555485DVILLGCPAMGSE77 ---572222288-8888755588GKKVGLF
2fcr -41456IFFSTSTGNTTEVA999998865432222765554443244779YDLLFLGAPT944411999-111112454441-8DKLPEVDMKDLPAIF
FLAV_ANASP -00456LFYGTQTGKTESVAEII987755323322427776666623589YQYLIIGCPTW55532 --999843678W988899988888888GKLVAYF
FLAV_AZOVI -42445LFFGSNTGKTRKVAKSIIK8777743433353666665467777YQFLILGTPTLGE86222222222355558-45666666888KTVALF
FLAV_ENTAG -266IGIFFGSDTGQTRKVAKLIHQKL6664664424DVRRATR888888SYPVLLLTPT8888864444444446WQEF8-8NTLSEADLTGKTVALF
FLAV_ECOLI -51114IFFGSDTGNTEINAKMI9877433111155555588355599YDILLGPT954431 ---88355225544-44666666779KTVALF
FLAV_CLOAB -63666ILYSSKTGKTERVAKLIE63333333333333333366LQESEIGFTPT93-6 ---66SWE3333333333333333GKLGAAF
3chy ADKELKFLVDDFSTMRIRVRLNLKELGFNNVEEAEDGVDALENKLO-AGGYGFVI ---SDWNMPNM ---DGLEL -LKTIRADGAMSALPVL

Avr	9334459999999999999999999877665555555666667756667889999999997676588887755555666689677776778899999999
Consist	0236428675848969746963946463344354312564565414344366588685675544550000003144654460055575345547747759
Conservation	

[illegible]

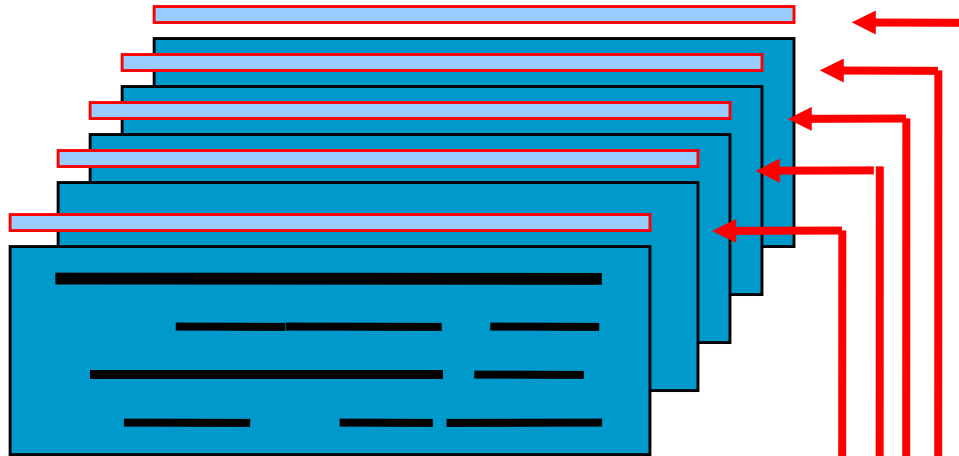
Avrg Consist 99887797877777777777997788888888888866777777777767766677777676667766655455577776666433355788788
Conservation 746640037154545706300354534444*745753000001010010000000010683760144442335574454448434301000000

Iteration 0 SP= 136702.00 AvSP= 10.654 SId= 3955 AvSId= 0.308

Consistency values are scored from 0 to 10; the value 10 is represented by the corresponding amino acid (red)

Consistency iteration

Pre-profiles

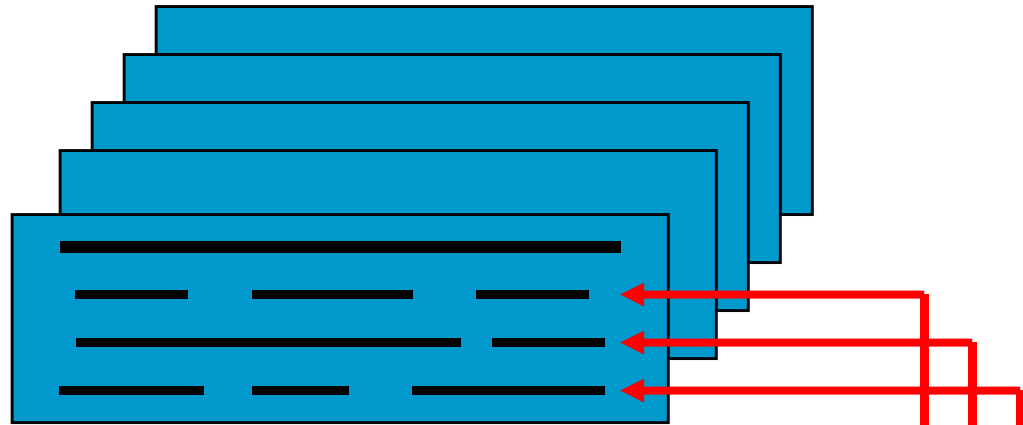


*Multiple
alignment
positional
consistency
scores*

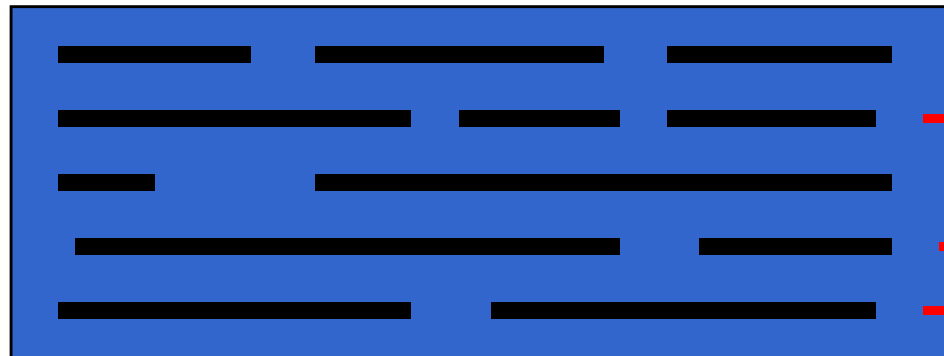


Pre-profile update iteration

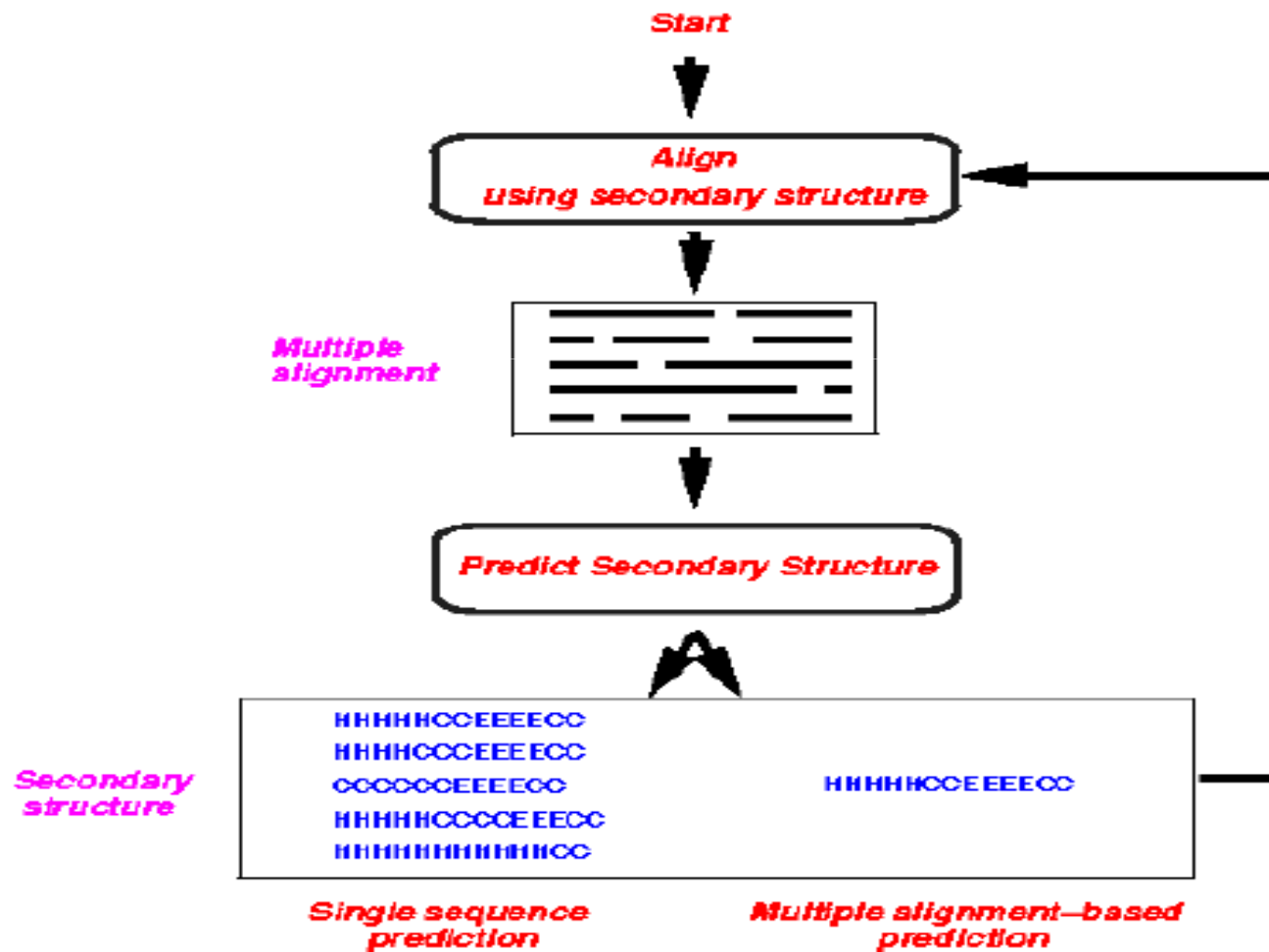
Pre-profiles



*Multiple
alignment*



Secondary structure-induced alignment

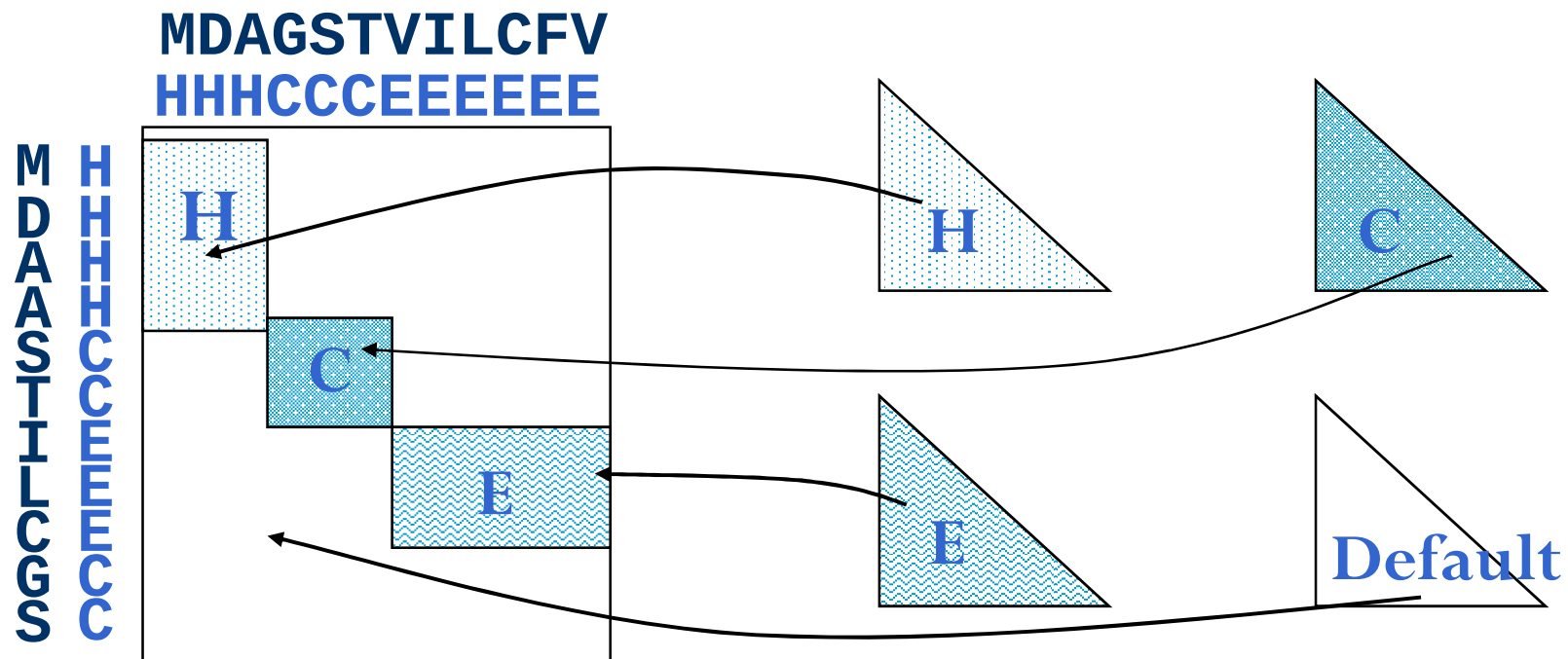


PRALINE

Using secondary structure for alignment

*Dynamic programming
search matrix*

*Amino acid exchange
weights matrices*



cheY SSEs

AA |ADKELKFLVVDDFSTMRRIVRNLLKELGFNNVEEAEDGVDALNKLQAGGYGFVISDWNMP|

PHD | EEEEEEE HHHHHHHHHHHHHHHHHHHH E HHHHHHHHHH HHHEEE |

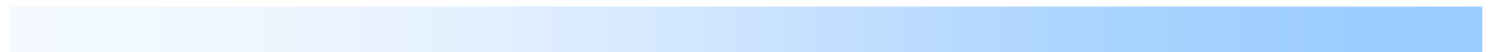
AA | NMDGLELLKTIRADGAMSALPVLMTAEAKKENIIAAAQAGASGYVVKPFTAATLEEKLNKIFEKLG |
PHD | HHHHHHBBBBB HHHHHHHHHHHHHHHHHHH HHHHHHHHHHHHHHHHH

Flavodoxin-cheY multiple alignment/ secondary structure iteration

cheY SSEs

3chy-AA SEQUENCE	AA		ADKELKFLVDDF	STMRRIVRNLLKELGFNNVEEAED	GVDALNKLQAGGYGFVISDWNMP	
3chy-ITERATION-0	PHD		EEEEEE	HHHHHHHHHHHHHHHHHH	E HHHHHHHHHH HHHEE	
3chy-ITERATION-1	PHD		EEEEEE	HHHHHHHHHHHHHHHHHH	HHHHHHH EEEEE	

3chy-AA SEQUENCE	AA		NMDGLELLKTIRADGAMSALPVLMTAEAK	KENIIAAAQAGASGYVVKPFT	AATLEEKLNKIFEKLG	M
3chy-ITERATION-0	PHD		HHHHHHEEEEE	HHHHHHHHHHHHHHHHHH	HHHHHHHHHHHHHHHH	
3chy-ITERATION-1	PHD		HHHHHHEEEEE	HHH HHHHHHHHHHHHHHHHH	EEE HHHHHHHHHHHHHHH	



Flavodoxin-cheY multiple alignment/
secondary structure iteration
cheY SSEs

3chy-AA SEQUENCE	AA	ADKELKFLVVDDFSTMRRIVRNLLKELGFNNVEEAEDGVDALNKLQAGGYGFVISDWNMP
3chy-ITERATION-0	PHD	EEEEEEE HHHHHHHHHHHHHHHHHHH E HHHHHHHHHH HHHEEE
3chy-ITERATION-1	PHD	EEEEEEE HHHHHHHHHHHHHHHHHH HHHHHHHH EEEEE
3chy-ITERATION-2	PHD	EEEEEEE HHHHHHHHHHHHHHHHH HHHHHHHH EEEEE

3chy-AA SEQUENCE	AA	NMDGLELLKTIADGAMSALPVLMTAEAKKENIIAAQAQASGYVVKPFTAATLEEKLNKIFEKLG													
3chy-ITERATION-0	PHD	HHHHHHHEEEEE				HHHHHHHHHHHHHHHHHH				HHHHHHHHHHHHHHHH					
3chy-ITERATION-1	PHD	HHHHHHHEEEEE				HHH		HHHHHHHHHHHHHHHHHH				EEE		HHHHHHHHHHHHHHHH	
3chy-ITERATION-2	PHD	HHHHHHHEEEEE				HHHHHHHHHHHHHHHHHH				EEE		HHHHHHHHHHHHHHHH			

Flavodoxin-cheY multiple alignment/ secondary structure iteration

cheY SSEs

3chy-AA SEQUENCE	AA	ADKELKFLVDDFSTMRRIVRNLLKELGFNNVEEAEDGVDALNKLQAGGYGFVISDWNMP
3chy-ITERATION-0	PHD	EEEEEEE HHHHHHHHHHHHHHHHHHH E HHHHHHHHHH HHHEEE
3chy-ITERATION-1	PHD	EEEEEEEE HHHHHHHHHHHHHHHHHHH HHHHHHHH EEEEE
3chy-ITERATION-2	PHD	EEEEEEEE HHHHHHHHHHHHHHHHHHH HHHHHHHH EEEEE
3chy-ITERATION-3	PHD	EEEEEEEE HHHHHHHHHHHHHHHHHH EEE HHHHHH EEEEE

3chy-AA SEQUENCE	AA	NMDGLELLKTIRADGAMSALPVLMTAEAKKENIIAAAQAGASGYVVKPFTAATLEEKLNKIFEKLG
3chy-ITERATION-0	PHD	HHHHHHEEEEEE HHHHHHHHHHHHHHHHHHH HHHHHHHHHHHHHHHHH
3chy-ITERATION-1	PHD	HHHHHHEEEEEE HHH HHHHHHHHHHHHHHHHHHH EEE HHHHHHHHHHHHHHHHH
3chy-ITERATION-2	PHD	HHHHHHEEEEEE HHHHHHHHHHHHHHHHHHH EEE HHHHHHHHHHHHHHHHH
3chy-ITERATION-3	PHD	HHHHHHHHHHHHH HHHHHHHHHHHHHHHHHHH EEE HHHHHHHHHHHHHHHHH

Flavodoxin-cheY multiple alignment/
secondary structure iteration
cheY SSEs

3chy-AA SEQUENCE	AA	ADKELKFLVVDDFSTMRRIVRNLLKELGFNNVEEAEDGVDALNKLQAGGYGFVISDWNMP
3chy-ITERATION-0	PHD	EEEEEEEEHHHHHHHHHHHHHHHHHHHEHHHHHHHHHHHHH
3chy-ITERATION-1	PHD	EEEEEEEEHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
3chy-ITERATION-2	PHD	EEEEEEEEHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
3chy-ITERATION-3	PHD	EEEEEEEEHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
3chy-ITERATION-4	PHD	EEEEEEEEHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH

3chy-AA SEQUENCE	AA	NMDGLELLKTI	RADGAMSALPVL	MVTAEAK	KENII	AAAQAGASGYV	VVKPFTAAT	LEEKL	NKIF	EKLGM
3chy-ITERATION-0	PHD	HHHHHHHEEEEE		HHHHHHHHHHHHHHHHHH			HHHHHHHHHHHHHHHH			
3chy-ITERATION-1	PHD	HHHHHHHEEEEE	HHH	HHHHHHHHHHHHHHHHHH		EEE	HHHHHHHHHHHHHHHH			
3chy-ITERATION-2	PHD	HHHHHHHEEEEE		HHHHHHHHHHHHHHHHHH		EEE	HHHHHHHHHHHHHHHH			
3chy-ITERATION-3	PHD	HHHHHHHHHHHHH		HHHHHHHHHHHHHHHHHH		EEE	HHHHHHHHHHHHHHHH			
3chy-ITERATION-4	PHD	HHHHH	EEEE	HHHHHHHHHHHHHHHHHH		EEE	HHHHHHHHHHHHHHHH			

Flavodoxin-cheY multiple alignment/
secondary structure iteration
cheY SSEs

3chy-AA SEQUENCE	AA	ADKELKFLVVDDFSTMRRIVRNLLKELGFNNVEEAEDGVDALNKLQAGGYGFVISDWNMP
3chy-ITERATION-0	PHD	EEEEEEEEHHHHHHHHHHHHHHHHHHHEHHHHHHHHHHHHHEEE
3chy-ITERATION-1	PHD	EEEEEEEEHHHHHHHHHHHHHHHHHHHHHHHHHHHHHEEEEE
3chy-ITERATION-2	PHD	EEEEEEEEHHHHHHHHHHHHHHHHHHHHHHHHHHHHHEEEEE
3chy-ITERATION-3	PHD	EEEEEEEEHHHHHHHHHHHHHHHHHHHHHHHHHHHEEEHHHHHHHEEEEE
3chy-ITERATION-4	PHD	EEEEEEEEHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHEEEEE
3chy-ITERATION-5	PHD	EEEEEEEEHHHHHHHHHHHHHHHHHHHHHHHHHHHEEEHHHHHHHEEEEE

3chy-AA SEQUENCE	AA	NMDGLELLKTI	RADGAMSALPVL	MVTAEAK	KENII	IAAAQAGASGYV	VKPFTAAT	LEEKL	NKIF	EKLGM
3chy-ITERATION-0	PHD	HHHHHHHEEEEE		HHHHHHHHHHHHHHHHHH			HHHHHHHHHHHHHHHH			
3chy-ITERATION-1	PHD	HHHHHHHEEEEE	HHH	HHHHHHHHHHHHHHHHHH		EEE	HHHHHHHHHHHHHHHH			
3chy-ITERATION-2	PHD	HHHHHHHEEEEE		HHHHHHHHHHHHHHHHHH		EEE	HHHHHHHHHHHHHHHH			
3chy-ITERATION-3	PHD	HHHHHHHHHHHH		HHHHHHHHHHHHHHHHHH		EEE	HHHHHHHHHHHHHHHH			
3chy-ITERATION-4	PHD	HHHHH	EEEE	HHHHHHHHHHHHHHHHHH		EEE	HHHHHHHHHHHHHHHH			
3chy-ITERATION-5	PHD	HHHHHHHH	EEEE	HHHHHHHHHHHHHHHHHH		EEE	HHHHHHHHHHHHHHHH			

Flavodoxin-cheY multiple alignment/ secondary structure iteration

cheY SSEs

3chy-AA SEQUENCE	AA	ADKELKFLVVDDFSTMRRIVRNLLKELGFNNVEEAEDGVDALNKLQAGGYGFVISDWNMP
3chy-ITERATION-0	PHD	EEEEEEE HHHHHHHHHHHHHHHHHHH E HHHHHHHHHH HHHEEE
3chy-ITERATION-1	PHD	EEEEEEEE HHHHHHHHHHHHHHHHHHH HHHHHHHH EEEEE
3chy-ITERATION-2	PHD	EEEEEEEE HHHHHHHHHHHHHHHHHHH HHHHHHHH EEEEE
3chy-ITERATION-3	PHD	EEEEEEEE HHHHHHHHHHHHHHHHHHH EEE HHHHHH EEEEE
3chy-ITERATION-4	PHD	EEEEEEEE HHHHHHHHHHHHHHHHHHH HHHHHHHH EEEEE
3chy-ITERATION-5	PHD	EEEEEEEE HHHHHHHHHHHHHHHHHHH EEE HHHHHH EEEEE
3chy-ITERATION-6	PHD	EEEEEEEE HHHHHHHHHHHHHHHHHHH HHHHHHHH EEEEE

3chy-AA SEQUENCE	AA	NMDGLELLKTIRADGAMSALPVLMTAEAKENIIAAAQAGASGYVVKPFTAATLEEKLNKIFEKLG
3chy-ITERATION-0	PHD	HHHHHHEEEEEE HHHHHHHHHHHHHHHHHHH HHHHHHHHHHHHHHHHH
3chy-ITERATION-1	PHD	HHHHHHEEEEEE HHH HHHHHHHHHHHHHHHHHHH EEE HHHHHHHHHHHHHHHHH
3chy-ITERATION-2	PHD	HHHHHHEEEEEE HHHHHHHHHHHHHHHHHHH EEE HHHHHHHHHHHHHHHHH
3chy-ITERATION-3	PHD	HHHHHHHHHHHHH HHHHHHHHHHHHHHHHHHH EEE HHHHHHHHHHHHHHHHH
3chy-ITERATION-4	PHD	HHHHH EEEEE HHHHHHHHHHHHHHHHHHH EEE HHHHHHHHHHHHHHHHH
3chy-ITERATION-5	PHD	HHHHHHHH EEEEE HHHHHHHHHHHHHHHHHHH EEE HHHHHHHHHHHHHHHHH
3chy-ITERATION-6	PHD	HHHHHHHH EEEEE HHHHHHHHHHHHHHHHHHH EEEE HHHHHHHHHHHHHHHHH

Flavodoxin-cheY multiple alignment/ secondary structure iteration

cheY SSEs

3chy-AA SEQUENCE	AA	ADKELKFLVVDDFSTMRRIVRNLLKELGFNNVEEAEDGVDALNKLQAGGYGFVISDWNMP
3chy-ITERATION-0	PHD	EEEEEEEE HHHHHHHHHHHHHHHHHHH E HHHHHHHHHH HHHEEE
3chy-ITERATION-1	PHD	EEEEEEEE HHHHHHHHHHHHHHHHHHH HHHHHHHH EEEEE
3chy-ITERATION-2	PHD	EEEEEEEE HHHHHHHHHHHHHHHHHHH HHHHHHHH EEEEE
3chy-ITERATION-3	PHD	EEEEEEEE HHHHHHHHHHHHHHHHHHH EEE HHHHHH EEEEE
3chy-ITERATION-4	PHD	EEEEEEEE HHHHHHHHHHHHHHHHHHH HHHHHHHH EEEEE
3chy-ITERATION-5	PHD	EEEEEEEE HHHHHHHHHHHHHHHHHHH EEE HHHHHH EEEEE
3chy-ITERATION-6	PHD	EEEEEEEE HHHHHHHHHHHHHHHHHHH HHHHHHHH EEEEE
3chy-ITERATION-7	PHD	EEEEEEEE HHHHHHHHHHHHHHHHHHH EEE HHHHHH EEEEE

3chy-AA SEQUENCE	AA	NMDGLELLKTIRADGAMSALPVLMTAEAKKENIIAAAQAGASGYVVKPFTAATLEEKLNKIFEKLG
3chy-ITERATION-0	PHD	HHHHHHEEEEEE HHHHHHHHHHHHHHHHHHH HHHHHHHHHHHHHHHHHHH
3chy-ITERATION-1	PHD	HHHHHHEEEEEE HHH HHHHHHHHHHHHHHHHHHH EEE HHHHHHHHHHHHHHHHHHH
3chy-ITERATION-2	PHD	HHHHHHEEEEEE HHHHHHHHHHHHHHHHHHH EEE HHHHHHHHHHHHHHHHHHH
3chy-ITERATION-3	PHD	HHHHHHHHHHHHHH HHHHHHHHHHHHHHHHHHH EEE HHHHHHHHHHHHHHHHHHH
3chy-ITERATION-4	PHD	HHHHH EEEEE HHHHHHHHHHHHHHHHHHH EEE HHHHHHHHHHHHHHHHHHH
3chy-ITERATION-5	PHD	HHHHHHHH EEEEE HHHHHHHHHHHHHHHHHHH EEE HHHHHHHHHHHHHHHHHHH
3chy-ITERATION-6	PHD	HHHHHHHH EEEEE HHHHHHHHHHHHHHHHHHH EEEE HHHHHHHHHHHHHHHHHHH
3chy-ITERATION-7	PHD	HHHHHHHH EEEEE HHHHHHHHHHHHHHHHHHH EEE HHHHHHHHHHHHHHHHHHH

Flavodoxin-cheY multiple alignment/ secondary structure iteration

cheY SSEs

3chy-AA SEQUENCE	AA	ADKELKFLVDDFSTMRRIVRNLLKELGFNNVEEAEDGVDALNKLQAGGYGFVISDWNMP
3chy-ITERATION-0	PHD	EEEEEEEE HHHHHHHHHHHHHHHHHHH E HHHHHHHHHH HHHEEE
3chy-ITERATION-1	PHD	EEEEEEEE HHHHHHHHHHHHHHHHHHH HHHHHHHH EEEEE
3chy-ITERATION-2	PHD	EEEEEEEE HHHHHHHHHHHHHHHHHHH HHHHHHHH EEEEE
3chy-ITERATION-3	PHD	EEEEEEEE HHHHHHHHHHHHHHHHHHH EEE HHHHHH EEEEE
3chy-ITERATION-4	PHD	EEEEEEEE HHHHHHHHHHHHHHHHHHH HHHHHHHH EEEEE
3chy-ITERATION-5	PHD	EEEEEEEE HHHHHHHHHHHHHHHHHHH EEE HHHHHH EEEEE
3chy-ITERATION-6	PHD	EEEEEEEE HHHHHHHHHHHHHHHHHHH HHHHHHHH EEEEE
3chy-ITERATION-7	PHD	EEEEEEEE HHHHHHHHHHHHHHHHHHH EEE HHHHHH EEEEE
3chy-ITERATION-8	PHD	EEEEEEEE HHHHHHHHHHHHHHHHHHH HHHHHHHH EEEEE

3chy-AA SEQUENCE	AA	NMDGLELLKTIRADGAMSALPVLMTAEAKENIIAAAQAGASGYVVKPFTAATLEEKLNKIFEKLG
3chy-ITERATION-0	PHD	HHHHHHEEEEEE HHHHHHHHHHHHHHHHHHH HHHHHHHHHHHHHHHHHHH
3chy-ITERATION-1	PHD	HHHHHHEEEEEE HHH HHHHHHHHHHHHHHHHHHH EEE HHHHHHHHHHHHHHHHHHH
3chy-ITERATION-2	PHD	HHHHHHEEEEEE HHHHHHHHHHHHHHHHHHH EEE HHHHHHHHHHHHHHHHHHH
3chy-ITERATION-3	PHD	HHHHHHHHHHHHHH HHHHHHHHHHHHHHHHHHH EEE HHHHHHHHHHHHHHHHHHH
3chy-ITERATION-4	PHD	HHHHH EEEEE HHHHHHHHHHHHHHHHHHH EEE HHHHHHHHHHHHHHHHHHH
3chy-ITERATION-5	PHD	HHHHHHHH EEEEE HHHHHHHHHHHHHHHHHHH EEE HHHHHHHHHHHHHHHHHHH
3chy-ITERATION-6	PHD	HHHHHHHH EEEEE HHHHHHHHHHHHHHHHHHH EEEE HHHHHHHHHHHHHHHHHHH
3chy-ITERATION-7	PHD	HHHHHHHH EEEEE HHHHHHHHHHHHHHHHHHH EEE HHHHHHHHHHHHHHHHHHH
3chy-ITERATION-8	PHD	HHHHHHHH EEEEE HHHHHHHHHHHHHHHHHHH EEE HHHHHHHHHHHHHHHHHHH

Flavodoxin-cheY multiple alignment/
secondary structure iteration
cheY SSEs

3chy-AA SEQUENCE	AA	ADKELKFLVVDDFSTMRRIVRNLLKELGFNNVEEAEDGVDALNKLQAGGYGVFISDWNMP
3chy-ITERATION-0	PHD	EEEEEEEEHHHHHHHHHHHHHHHHHHHHHEHHHHHHHHHHHHHHEEE
3chy-ITERATION-1	PHD	EEEEEEEEHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHEEE
3chy-ITERATION-2	PHD	EEEEEEEEHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHEEE
3chy-ITERATION-3	PHD	EEEEEEEEHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHEEE
3chy-ITERATION-4	PHD	EEEEEEEEHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHEEE
3chy-ITERATION-5	PHD	EEEEEEEEHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHEEE
3chy-ITERATION-6	PHD	EEEEEEEEHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHEEE
3chy-ITERATION-7	PHD	EEEEEEEEHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHEEE
3chy-ITERATION-8	PHD	EEEEEEEEHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHEEE
3chy-ITERATION-9	PHD	EEEEEEEEHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHEEE

3chy-AA SEQUENCE	AA		NMDGLELLKTI	RADGAMSALPVL	MVTAEAK	KENII	AAAQAGASGYV	VVKPFT	AATLEEKL	NKIF	EKLGM	
3chy-ITERATION-0	PHD		HHHHHHHEEEEE			HHHHHHHHHHHHHHHHHH			HHHHHHHHHHHHHHHH			
3chy-ITERATION-1	PHD		HHHHHHHEEEEE	HHH	HHHHHHHHHHHHHHHHHH		EEE		HHHHHHHHHHHHHHHH			
3chy-ITERATION-2	PHD		HHHHHHHEEEEE			HHHHHHHHHHHHHHHHHH		EEE		HHHHHHHHHHHHHHHH		
3chy-ITERATION-3	PHD		HHHHHHHHHHHHH			HHHHHHHHHHHHHHHHHH		EEE		HHHHHHHHHHHHHHHH		
3chy-ITERATION-4	PHD		HHHHH	EEEE	HHHHHHHHHHHHHHHHHH		EEE		HHHHHHHHHHHHHHHH			
3chy-ITERATION-5	PHD		HHHHHHHHH	EEEE	HHHHHHHHHHHHHHHHHH		EEE		HHHHHHHHHHHHHHHH			
3chy-ITERATION-6	PHD		HHHHHHHHH	EEEE	HHHHHHHHHHHHHHHHHH		EEEE		HHHHHHHHHHHHHHHH			
3chy-ITERATION-7	PHD		HHHHHHHHH	EEEE	HHHHHHHHHHHHHHHHHH		EEE		HHHHHHHHHHHHHHHH			
3chy-ITERATION-8	PHD		HHHHHHHHH	EEEE	HHHHHHHHHHHHHHHHHH		EEE		HHHHHHHHHHHHHHHH			
3chy-ITERATION-9	PHD		HHHHHHHHH	EEEE	HHHHHHHHHHHHHHHHHH		EEEE		HHHHHHHHHHHHHHHH			

Flavodoxin-cheY multiple alignment/ secondary structure iteration

cheY SSEs

3chy-AA SEQUENCE	AA	ADKELKFLVVDDFSTMRRIVRNLLKELGFNNVEEAEDGVDALNKLQAGGYGFVISDWNMP
3chy-ITERATION-0	PHD	EEEEEEEE HHHHHHHHHHHHHHHHHHH E HHHHHHHHHH HHHEEE
3chy-ITERATION-1	PHD	EEEEEEEE HHHHHHHHHHHHHHHHHHH HHHHHHHH EEEEE
3chy-ITERATION-2	PHD	EEEEEEEE HHHHHHHHHHHHHHHHHHH HHHHHHHH EEEEE
3chy-ITERATION-3	PHD	EEEEEEEE HHHHHHHHHHHHHHHHHHH EEE HHHHHH EEEEE
3chy-ITERATION-4	PHD	EEEEEEEE HHHHHHHHHHHHHHHHHHH HHHHHHHH EEEEE
3chy-ITERATION-5	PHD	EEEEEEEE HHHHHHHHHHHHHHHHHHH EEE HHHHHH EEEEE
3chy-ITERATION-6	PHD	EEEEEEEE HHHHHHHHHHHHHHHHHHH HHHHHHHH EEEEE
3chy-ITERATION-7	PHD	EEEEEEEE HHHHHHHHHHHHHHHHHHH EEE HHHHHH EEEEE
3chy-ITERATION-8	PHD	EEEEEEEE HHHHHHHHHHHHHHHHHHH HHHHHHHH EEEEE
3chy-ITERATION-9	PHD	EEEEEEEE HHHHHHHHHHHHHHHHHHH HHHHHHHHHH EEEEE

3chy-AA SEQUENCE	AA	NMDGLELLKTIRADGAMSALPVLMTAEAKKENIIAAAQAGASGYVVKPFTAATLEEKLNKIFEKLG
3chy-ITERATION-0	PHD	HHHHHHEEEEEE HHHHHHHHHHHHHHHHHHH HHHHHHHHHHHHHHHHH
3chy-ITERATION-1	PHD	HHHHHHEEEEEE HHH HHHHHHHHHHHHHHHHHHH EEE HHHHHHHHHHHHHHHHH
3chy-ITERATION-2	PHD	HHHHHHEEEEEE HHHHHHHHHHHHHHHHHHH EEE HHHHHHHHHHHHHHHHH
3chy-ITERATION-3	PHD	HHHHHHHHHHHH HHHHHHHHHHHHHHHHHHH EEE HHHHHHHHHHHHHHHHH
3chy-ITERATION-4	PHD	HHHHH EEEEE HHHHHHHHHHHHHHHHHHH EEE HHHHHHHHHHHHHHHHH
3chy-ITERATION-5	PHD	HHHHHHHH EEEEE HHHHHHHHHHHHHHHHHHH EEE HHHHHHHHHHHHHHHHH
3chy-ITERATION-6	PHD	HHHHHHHH EEEEE HHHHHHHHHHHHHHHHHHH EEEE HHHHHHHHHHHHHHHHH
3chy-ITERATION-7	PHD	HHHHHHHH EEEEE HHHHHHHHHHHHHHHHHHH EEE HHHHHHHHHHHHHHHHH
3chy-ITERATION-8	PHD	HHHHHHHH EEEEE HHHHHHHHHHHHHHHHHHH EEE HHHHHHHHHHHHHHHHH
3chy-ITERATION-9	PHD	HHHHHHHH EEEEE HHHHHHHHHHHHHHHHHHH EEEE HHHHHHHHHHHHHHHHH

Is the initial SS prediction good enough?

3chy-AA SEQUENCE|| AA |ADKELKFLVDDFSTMRRIVRNLLKELGFNNVEEAEDGVDALNKLQAGGYGFVISDWNMP|

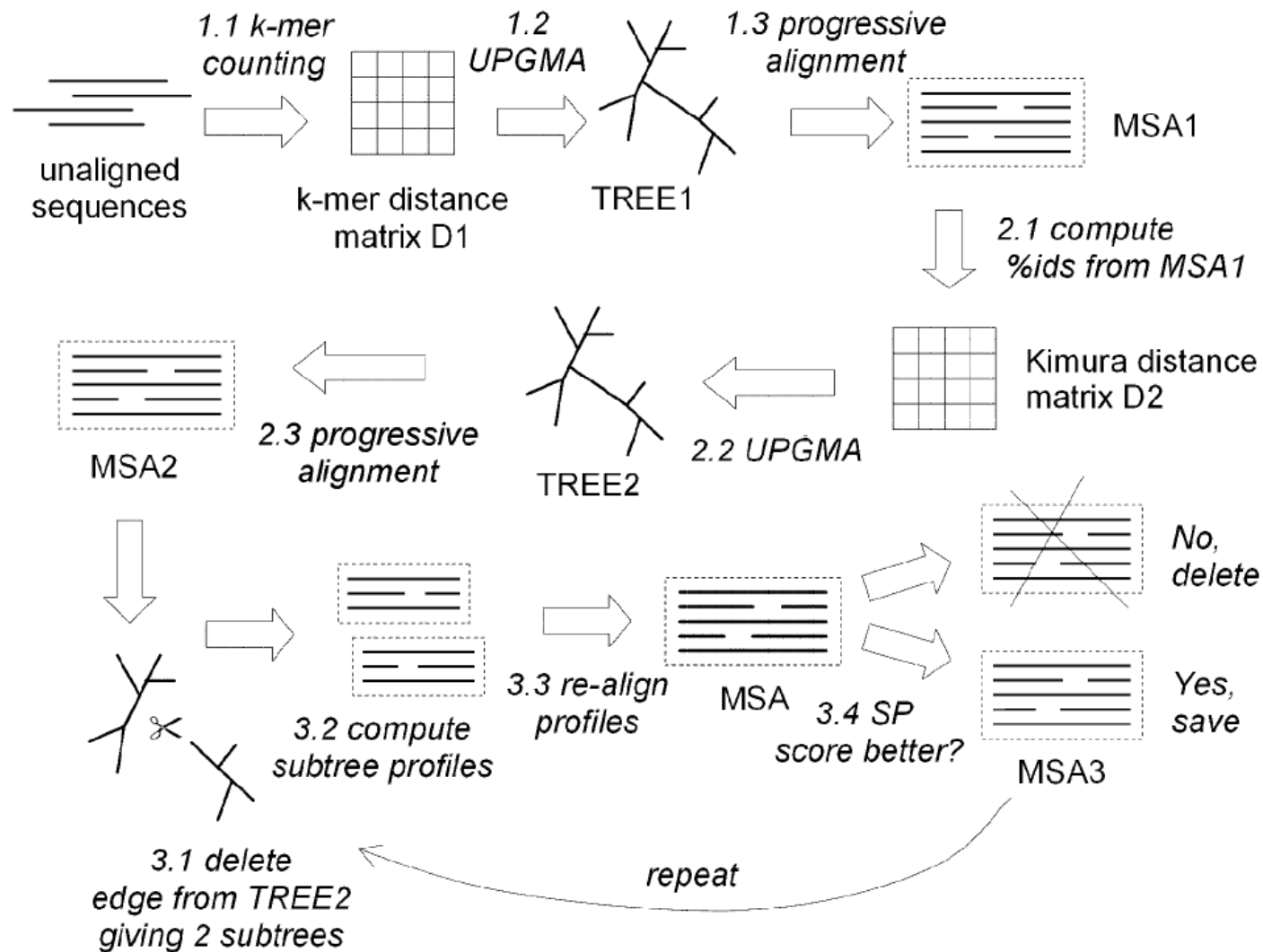
3chy-ITERATION-0	PHD	EEEEEEEE	HHHHHHHHHHHHHHHHHH	E	HHHHHHHHHH	HHHEEE	
3chy-ITERATION-1	PHD	EEEEEEEE	HHHHHHHHHHHHHHHHHH		HHHHHHHH	EEEEEE	
3chy-ITERATION-2	PHD	EEEEEEEE	HHHHHHHHHHHHHHHHHH		HHHHHHHH	EEEEEE	
3chy-ITERATION-3	PHD	EEEEEEEE	HHHHHHHHHHHHHHHHHH	EEE	HHHHHH	EEEE	
3chy-ITERATION-4	PHD	EEEEEEEE	HHHHHHHHHHHHHHHHHH		HHHHHH	EEEE	
3chy-ITERATION-5	PHD	EEEEEEEE	HHHHHHHHHHHHHHHHHH	EEE	HHHHHH	EEEE	
3chy-ITERATION-6	PHD	EEEEEEEE	HHHHHHHHHHHHHHHHHH		HHHHHH	EEEE	
3chy-ITERATION-7	PHD	EEEEEEEE	HHHHHHHHHHHHHHHHHH	EEE	HHHHHH	EEEE	
3chy-ITERATION-8	PHD	EEEEEEEE	HHHHHHHHHHHHHHHHHH		HHHHHH	EEEE	
3chy-ITERATION-9	PHD	EEEEEEEE	HHHHHHHHHHHHHHHHHH		HHHHHHHH	EEEE	

3chy-AA SEQUENCE|| AA |NMDGLELLKTIRADGAMSALPVLMTAEAKKENIIAAAQAGASGYVVKPFTAATLEEKLNKIFEKLG|

3chy-ITERATION-0	PHD	HHHHHHEEEEE	HHHHHHHHHHHHHHHHHH	HHHHHHHHHHHHHHHHHH	
3chy-ITERATION-1	PHD	HHHHHHEEEEE	HHH	HHHHHHHHHHHHHHHHHH	EEE HHHHHHHHHHHHHHHHH
3chy-ITERATION-2	PHD	HHHHHHEEEEE		HHHHHHHHHHHHHHHHHH	EEE HHHHHHHHHHHHHHHHH
3chy-ITERATION-3	PHD	HHHHHHHHHHHH		HHHHHHHHHHHHHHHHHH	EEE HHHHHHHHHHHHHHHHH
3chy-ITERATION-4	PHD	HHHHH	EEEE	HHHHHHHHHHHHHHHHHH	EEE HHHHHHHHHHHHHHHHH
3chy-ITERATION-5	PHD	HHHHHHHH	EEEE	HHHHHHHHHHHHHHHHHH	EEE HHHHHHHHHHHHHHHHH
3chy-ITERATION-6	PHD	HHHHHHHH	EEEE	HHHHHHHHHHHHHHHHHH	EEEE HHHHHHHHHHHHHHHHH
3chy-ITERATION-7	PHD	HHHHHHHH	EEEE	HHHHHHHHHHHHHHHHHH	EEE HHHHHHHHHHHHHHHHH
3chy-ITERATION-8	PHD	HHHHHHHH	EEEE	HHHHHHHHHHHHHHHHHH	EEE HHHHHHHHHHHHHHHHH
3chy-ITERATION-9	PHD	HHHHHHHH	EEEE	HHHHHHHHHHHHHHHHHH	EEEE HHHHHHHHHHHHHHHHH

MUSCLE

Edgar 2004



PRALINE and MUSCLE method

n **PRALINE** and **MUSCLE** use different formalisms to compare two profiles:

n **MUSCLE:**

$$LE^{xy} = (1 - f^x_G)(1 - f^y_G) \log \sum_i \sum_j f^x_i f^y_j \frac{p_{ij}}{p_i p_j}$$

n **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 (2004) calls the Muscle scoring scheme “Log-expectation scoring (LE)”

So what do we do ?

- n A single shot for a good alignment without thinking: **MUSCLE**, **T-COFFEE**, **PROBCONS** (maybe **POA**)
- n 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
- n There is no single method that always generates the best alignment
- n Therefore best is to use more than one method:
 - include **Dialign2** (local)
 - **PROBCONS** scores well in recent assessments

Recap

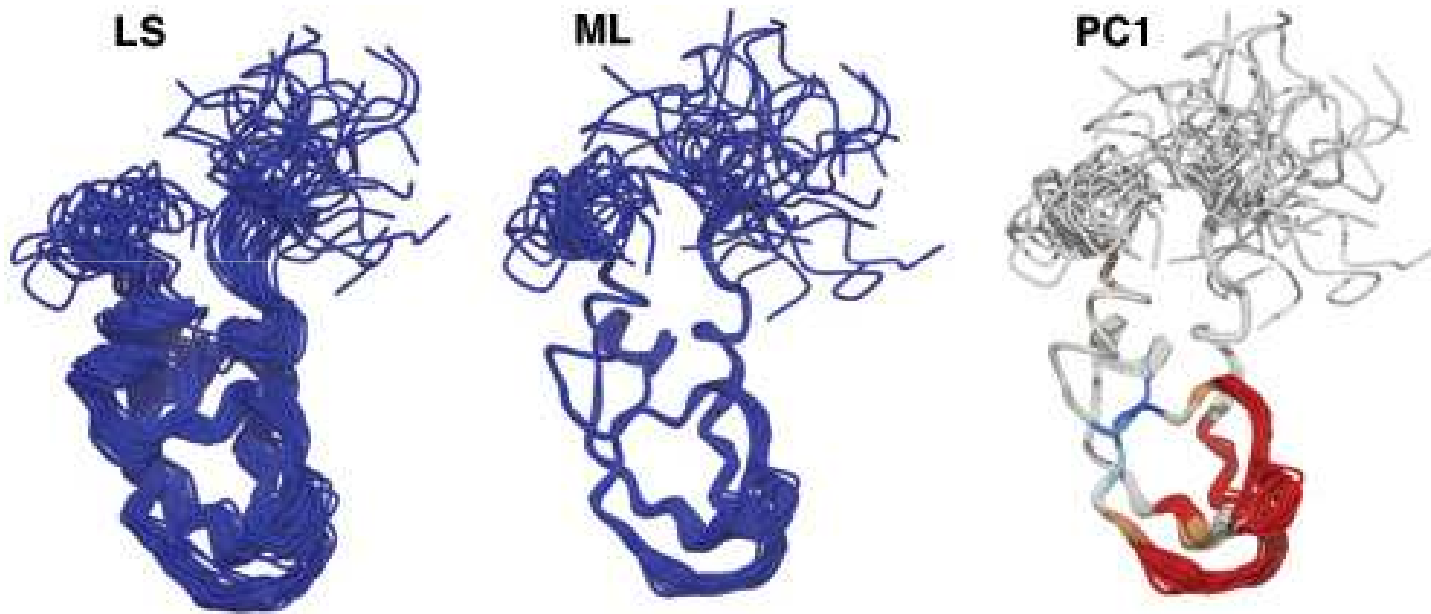
- n Pairwise alignment by Dynamic Programming
- n Weighting schemes to use information from all sequences right from the start during the progressive MSA protocol:
 - Profile pre-processing (global/local) ([PRALINE](#))
 - Matrix extension (well balanced scheme) ([T-Coffee](#))
- n Smoothing alignment signals:
 - Consistency based mixing of local and global alignment ([T-Coffee](#) and [PRALINE](#))
 - Homology-extended alignment ([PRALINE](#))
- n Using additional information:
 - secondary structure driven alignment ([PRALINE\(TM\)](#))
- n Iterative schemes to alleviate the ‘greediness’ of the progressive MSA protocol:
 - Profile pre-processing iteration ([PRALINE](#))
 - secondary structure driven iteration ([PRALINE](#))
 - Binary cutting of guide tree and realignment of groups ([MUSCLE](#))

Evaluating multiple alignments

- There are reference databases based on structural information: e.g. BAliBASE and HOMSTRAD
- Conflicting standards of truth
 - evolution
 - structure
 - function
- With orphan sequences no additional information
- Benchmarks depending on reference alignments
- Quality issue of available reference alignment databases
- Different ways to quantify agreement with reference alignment (sum-of-pairs, column score)
- “Charlie Chaplin” problem

Evaluating multiple alignments

- As a standard of truth, often a reference alignment based on structural superpositioning is taken



These superpositionings can be scored using the root-mean-

BAlibase benchmark alignments

Thompson et al. (1999) NAR 27, 2682.

8 categories:

- *cat. 1 - equidistant*
- *cat. 2 - orphan sequence*
- *cat. 3 - 2 distant groups*
- *cat. 4 – long overhangs*
- *cat. 5 - long insertions/deletions*
- *cat. 6 – repeats*
- *cat. 7 – transmembrane proteins*
- *cat. 8 – circular permutations*

BAlibase

BB11001	1aab_ref1	Ref1 V1 SHORT	high mobility group protein
BB11002	1aboA_ref1	Ref1 V1 SHORT	SH3
BB11003	1ad3_ref1	Ref1 V1 LONG	aldehyde dehydrogenase
BB11004	1adj_ref1	Ref1 V1 LONG	histidyl-trna synthetase
BB11005	1ajsA_ref1	Ref1 V1 LONG	aminotransferase
BB11006	1bbt3_ref1	Ref1 V1 MEDIUM	foot-and-mouth disease virus
BB11007	1cpt_ref1	Ref1 V1 LONG	cytochrome p450
BB11008	1csy_ref1	Ref1 V1 SHORT	SH2
BB11009	1dox_ref1	Ref1 V1 SHORT	ferredoxin [2fe-2s]



T-Coffee: correctly aligned Kinase nucleotide binding sites

```

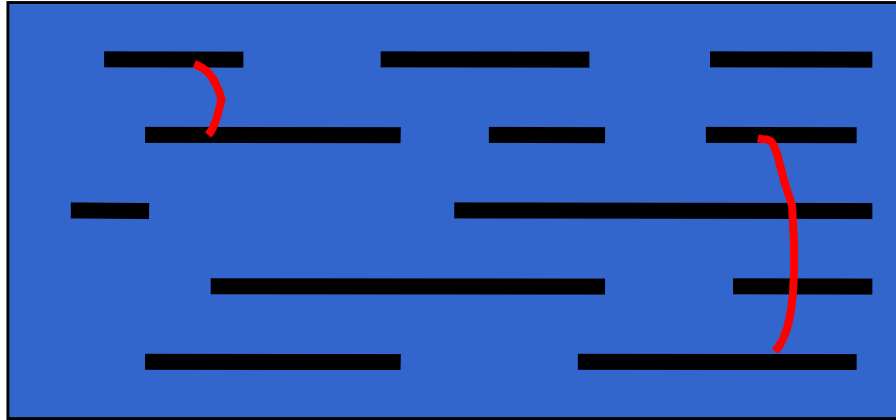
                                NBS
glla_oryza 1shff:llll:lgqgdigsvylsels---gte syfamKVMDK.as-----
lp08_human gndff:eieli gsggfqqvfkahhr---ldglt yviiKRVEY.as-----
gcaZ      --tll:rlaf sgqgafqgvvkaras---ldsr yvsiKKIKNTe-----
stil_yeast plawllbgac igsgsfqsvylgmas---htgelmasvKQVEI.kaaagvpt
lia3_yeast rseyqvlee igrgsfqsvrlvhi---ptll:lllvKDIKY.gh-----
nima_amei adlyevlel:igqgsfgilrlvrlr:---sdg filerKKIKY.il:-----
lia1_yeast lgdwe fvet vqgasmgl:vl:la:hr---ytaevcavKIVNR.at----la:f
koc1_yeast lll:lyvfglt lgagt fgvvrqslat---etgedwawKILIK.kas
ks02_human psqfel llvlgqgsfgl:vl:ll:isgsdarqlyamKVLIK.at-----
lp01_yeast ldsafvlllvlglgafgl:vl:ll:sl:---atdrlcasKVLIK.da-----
ypl:Z_yeast iddfdlllviglgsgfgl:vmqvr:rl:---dtq:liyaKALIK.ay-----
lrac_diodi vadfe llalwglbgsgfgl:viqvr:rl:---dtgev yamKVLIK.ch-----
lqp2_drome lldlrviat lgwggfgrvelvqt a--gdssrsfaKQIK.sq-----
kapa_mouse ldqfdrlt lgtgsfgrvmlvhl:---esgahyamKILIK.ql:-----
koca_drome leayit ravlgagsgfgrvmlvrl:---sgl:ay yamKMAIK.ed-----
arl:1_human madfsvhri igrggfgevygr:rl:---dtglmyamKCLIK.l:-----
dml:human rddfeillvlgrgafseavvllm:---qtgqv yamKIMIK.wd-----
dbf2_yeast ardfe mitqvgggggyqvylar:rl:---dt:evcasKILIK.l:-----
pim1_human esqyqvqgpllgsggfqsvysg:irv---sdalpvaiKHVEK.dr-----

glla_oryza -----lasrl:-ll:--lraqt
lp08_human -----elkar-----
gcaZ      -----e-ll:--stmis
stil_yeast aa:qaasdeaaeqeeqqel:iedvgavshpht:aqsihr:-mv--dalqh
lia3_yeast -----mas:-l:-er--qglia
nima_amei -----mst:-l:-er--egltas
lia1_yeast -----lh:aq-m-l--pppla
koc1_yeast -----lbgal:wgl--ealyd
ks02_human -----llvr--dr--vrtlm
lp01_yeast -----liqah-di--esara
ypl:Z_yeast -----ivsbo-ev--thtla
lrac_diodi -----iveha-ev--ehtls
lqp2_drome -----ivet r-qg--ghims
kapa_mouse -----vvl:ll:-ql--ehtla
koca_drome -----lvrl:q-v--ahvha
arl:1_human -----ilmhg-ge--tlala
dml:human -----ml:rg-ev--sfre
dbf2_yeast -----lfl:la-et--dhvlt
pim1_human -----isdwg-elpogtrvpm

                                NBS
glla_oryza EKEIL-----qcl dhp f--lpt lyt-----
lp08_human EVRAL-----akldhva--ivhyagwdgfd
gcaZ      EVMIL-----aslahqy--vrvyaaawleed
stil_yeast EMNLL-----lclhhea--ivt yyg-----
lia3_yeast ECSIL-----sql:hea--ive fya-w----
nima_amei EFNIL-----sslrhpa--ivayyh-r----
lia1_yeast EQDVLerql:l:lekeisrdl:rt ireaslgqilyhph--icrlfe--
koc1_yeast ELDIL-----qrlhha--ivafld-----
ks02_human ERDIL-----vevahp f--ivl:lhv-----
lp01_yeast EKEVF-----llatkt:chp f--ltalyo-----
ypl:Z_yeast ERTVL-----arvdep f--ivpl:l:f-----
lrac_diodi ERNIL-----ql:ishp f--lvslay-----
lqp2_drome EKEIM-----geaacq f--ivl:lfl:-----
kapa_mouse EKRIL-----qavafp f--lvl:lef-----
koca_drome EKHVL-----asarfp f--liylvd-----
arl:1_human ERIML-----slvstgdep f--ivcm sy-----
dml:human ERDVL-----vagdrw--itqlhf-----
dbf2_yeast ERDIL-----tttrsew--lv:llly-----
pim1_human EVVLL-----ll:vsqfsgvirllid-----

```

Scoring a single MSA with the Sum-of-pairs (SP) score



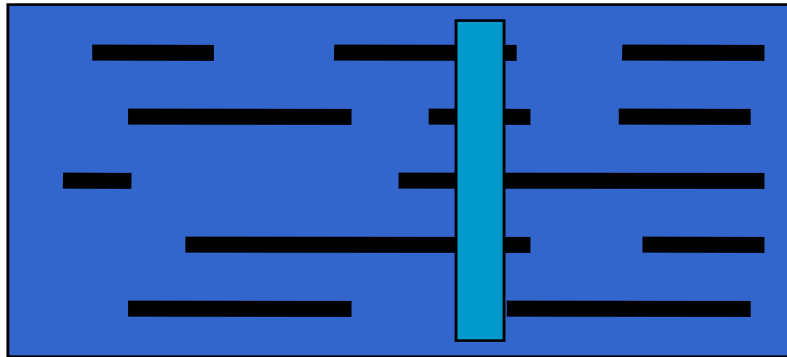
Good alignments should have a high SP score, but it is not always the case that the true biological alignment has the highest score.

Sum-of-Pairs score

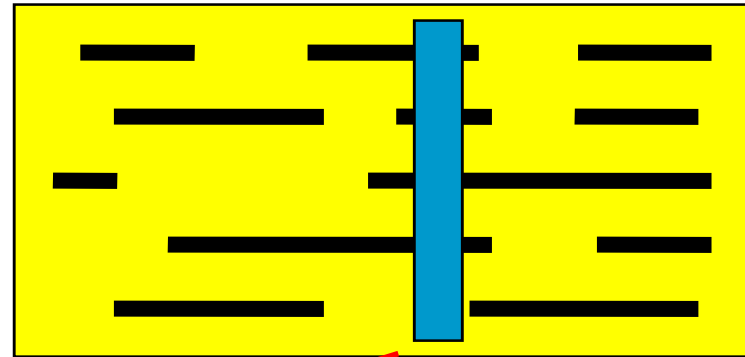
- Calculate the sum of all pairwise alignment scores
- This is equivalent to taking the sum of all matched a.a. pairs
- The latter can be done using gap penalties or not

Evaluation measures

Query

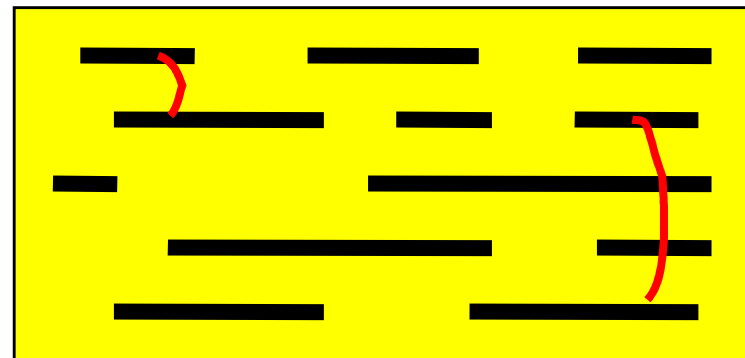
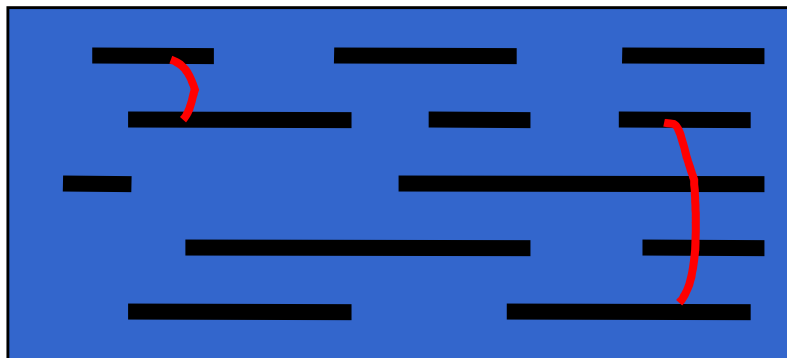


Reference



Column score

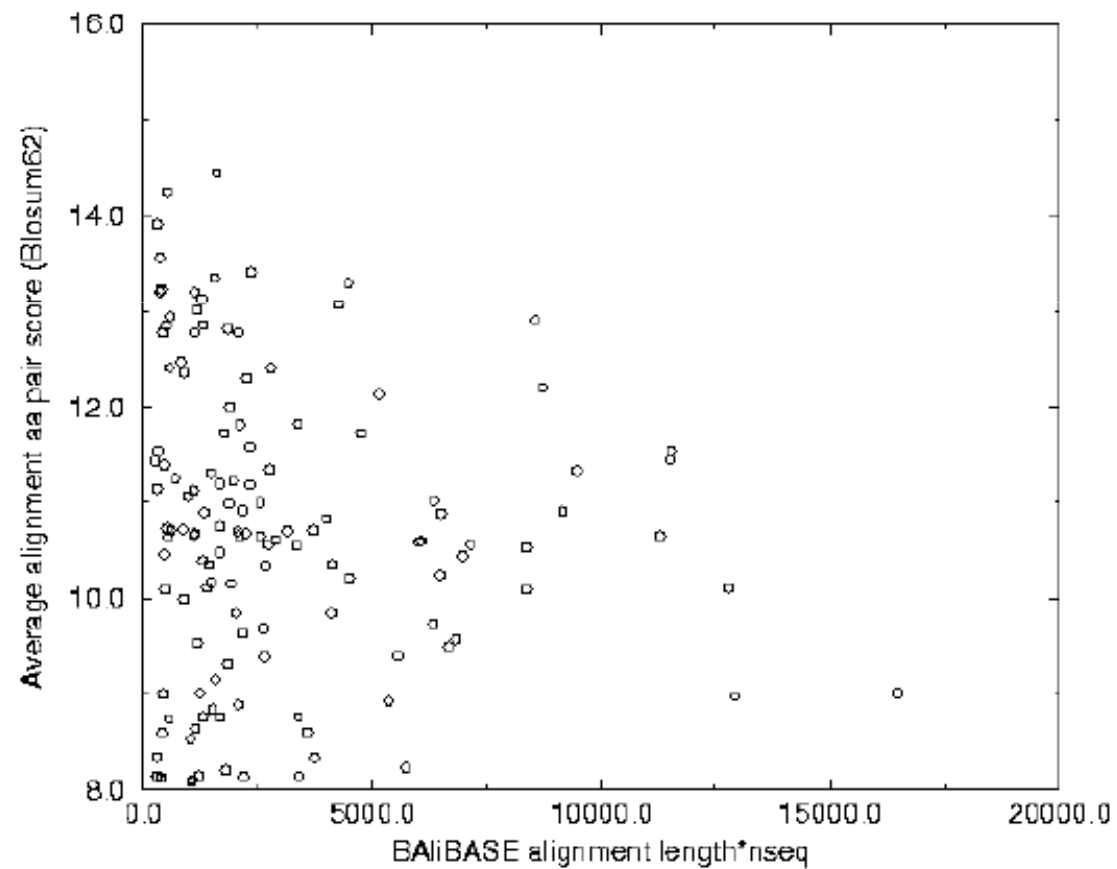
What fraction of the MSA columns in the reference alignment is reproduced by the computed alignment



Sum-of-Pairs score

What fraction of the matched amino acid pairs in the reference alignment is reproduced by the computed alignment

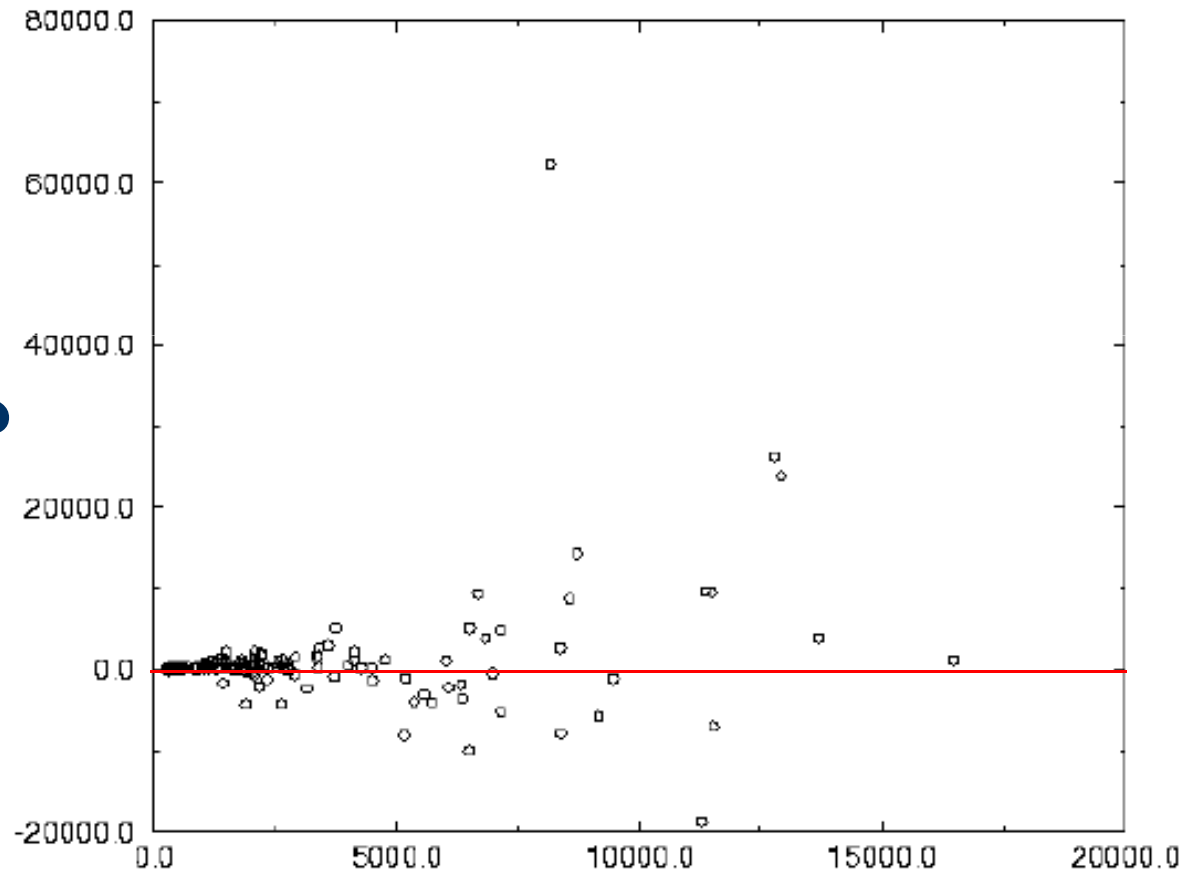
Evaluating multiple alignments



Evaluating multiple alignments

Charlie Chaplin problem

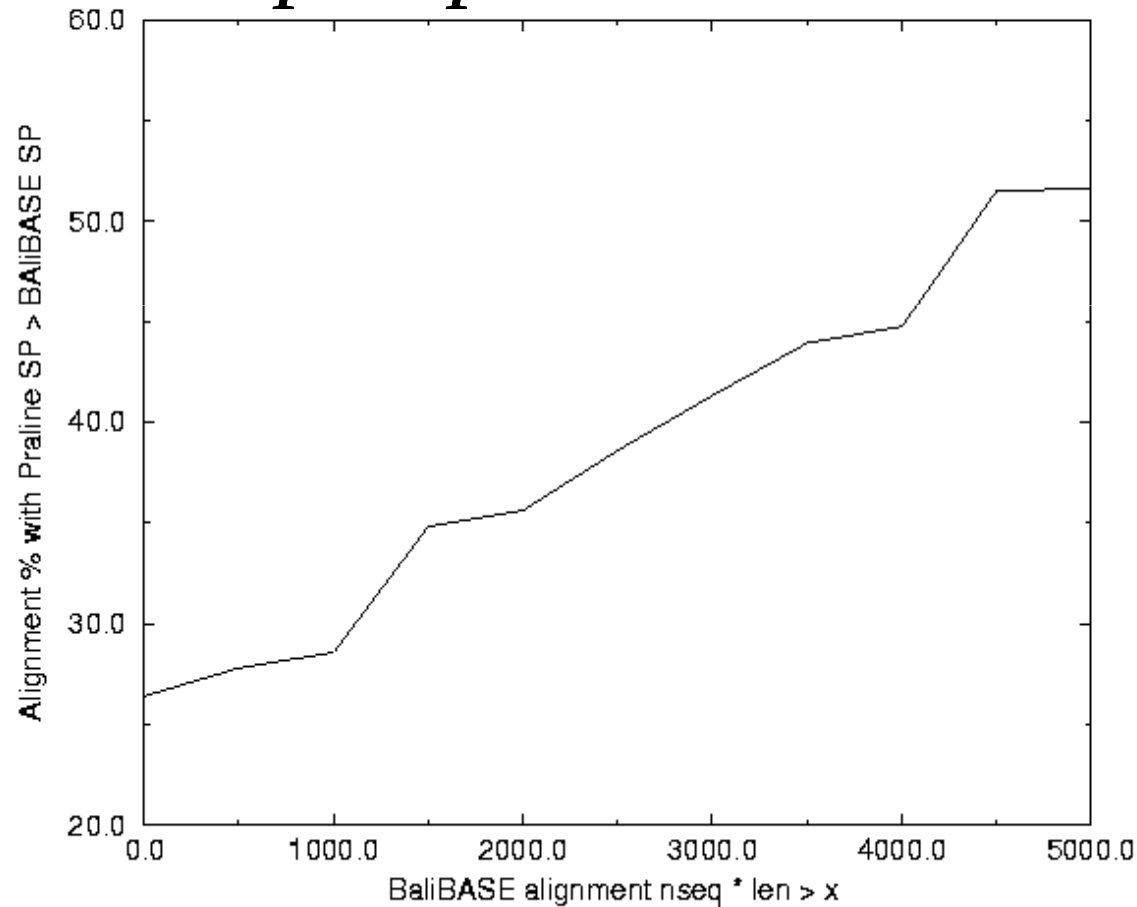
ΔSP



BAliBASE alignment nseq * len

Evaluating multiple alignments

Charlie Chaplin problem



T-coffee *global, local or both*

Table 1

Evaluation of different protocols with BaliBase.

Method				BaliBase Reference (Number of families)						CLE Outperformed by protocol (%)
				Cat 1 (82)	Cat 2 (23)	Cat 3 (12)	Cat 4 (13)	Cat 5 (11)	Tot (141)	
Name	Protocol									
C	ClustalW pw			70.6	26.7	43.0	56.0	60.0	58.9	7.8 **
CE	ClustalW pw		extend	77.1	33.6	47.6	64.8	75.9	66.3	17.7 **
L		Lalign pw		65.4	12.1	22.8	53.9	66.0	52.0	7.8 **
LE		Lalign pw	extend	72.6	25.6	47.2	77.5	85.5	64.2	16.3 **
CL	ClustalW pw	Lalign pw		76.2	32.0	48.3	76.2	74.6	66.5	12.1 **
CLE	ClustalW pw	Lalign pw	extend	<u>80.7</u>	<u>37.3</u>	<u>52.9</u>	<u>83.2</u>	<u>88.7</u>	<u>72.1</u>	

Protocol shows the way the library was created. *ClustalW pw* and *Lalign pw* show the pair-wise alignments computed with one of these programs, using default parameters. *Extend* indicates that the library was extended before progressive alignment. CLE uses a combination of ClustalW and Lalign alignments and library extension. Cat 1 to Cat 5 are the five reference categories of BaliBase, numbers in parenthesis indicate the number of alignments in a category. The average accuracy is then given for each protocol. The best accuracies in each column are shown in bold and underlined. Tot gives the average accuracy across all 141 test alignments. The last column shows the percentage of times that CLE is outperformed by each other protocol. The statistical significance of the improvement of CLE over each protocol is shown by * ($P < 0.01$) and ** ($P < 0.001$).

Comparing *T-coffee* with other methods

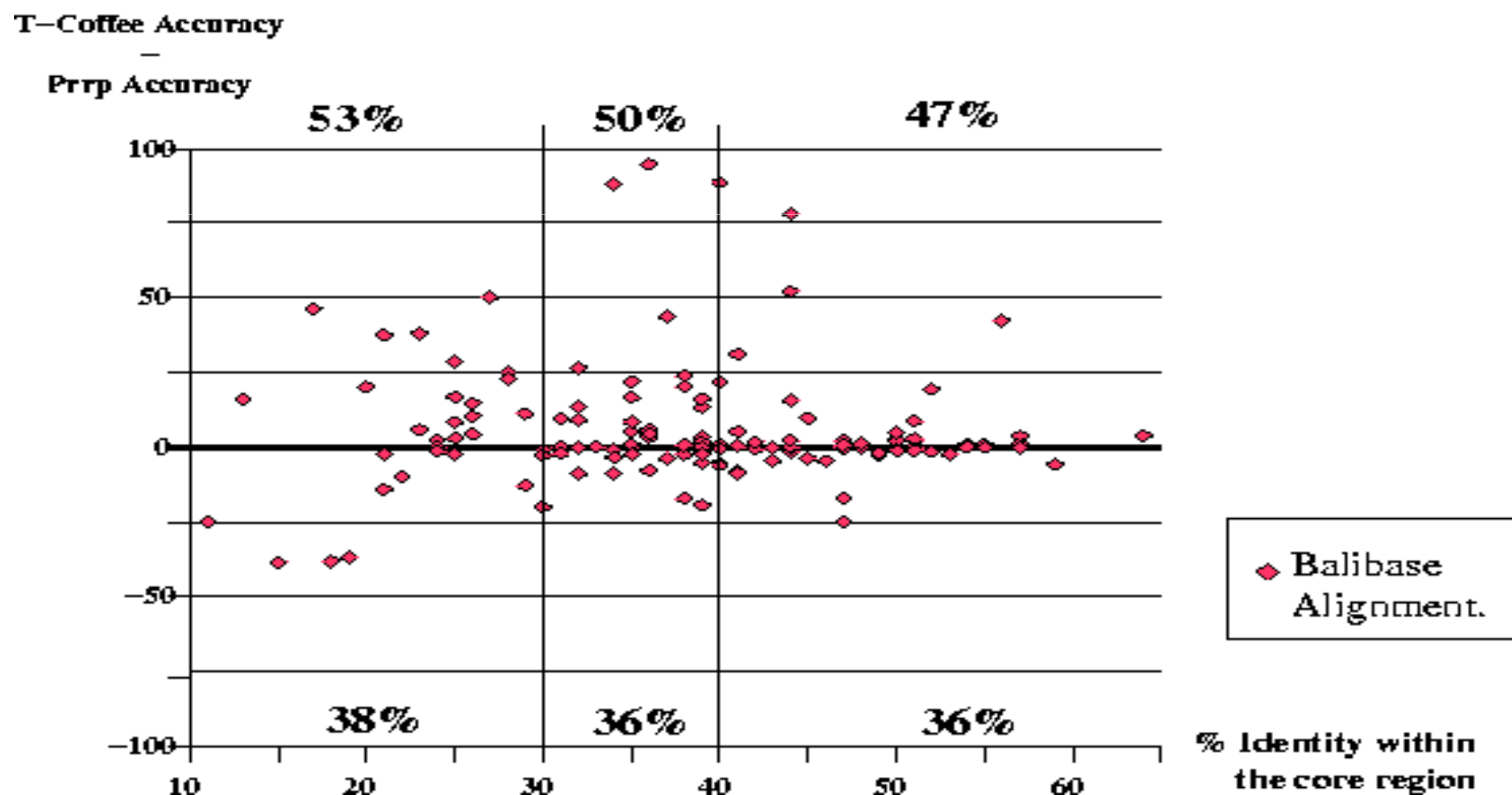
Table 2

Evaluation of four packages with BalBase

Method Name	BalBase Reference (Number of families)							T-Coffee Outperformed by method (%)
	Cat 1 (82)	Cat 2 (23)	Cat 3 (12)	Cat 4 (13)	Cat 5 (11)	Total 1 (141)	Total 2 (141)	
Dialign2	71.0	25.2	35.1	74.7	80.4	61.5	57.3	11.3**
ClustalW	78.5	32.2	42.5	65.7	74.3	66.4	58.6	26.2**
Prnp	78.6	32.5	50.2	51.1	82.7	66.4	59.0	36.9*
T-Coffee (CLE)	<u>80.7</u>	<u>37.3</u>	<u>52.9</u>	<u>83.2</u>	<u>88.7</u>	<u>72.1</u>	<u>68.7</u>	

BaliBASE benchmark alignments

Comparison of T-Coffee and Prnp



END

