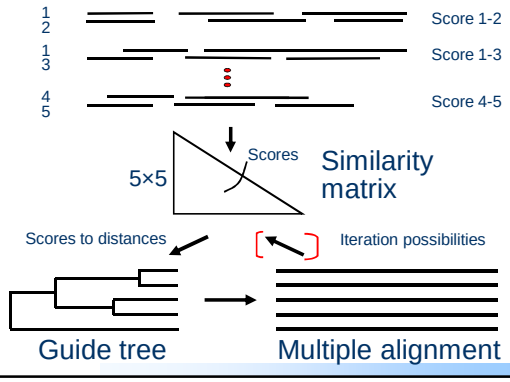
1-month Practical Course Genome Analysis (Integrative Bioinformatics & Genomics)

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

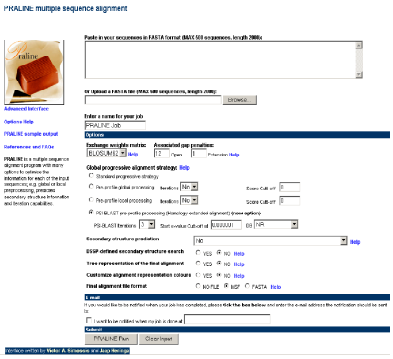


Additional strategies for multiple sequence alignment

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

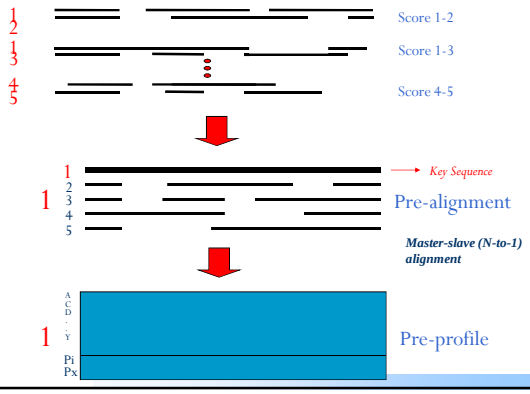
Objective: try to avoid (early) errors

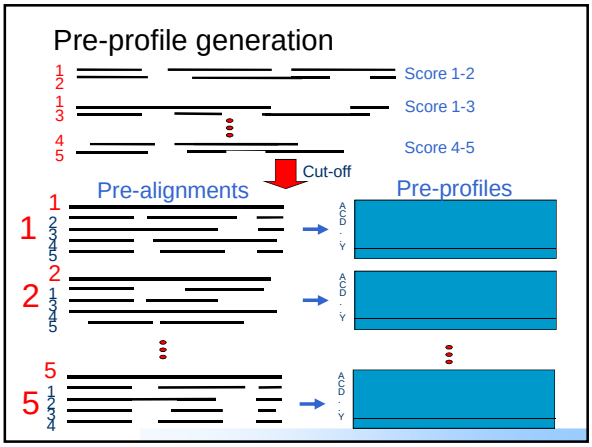
PRALINE web-interface

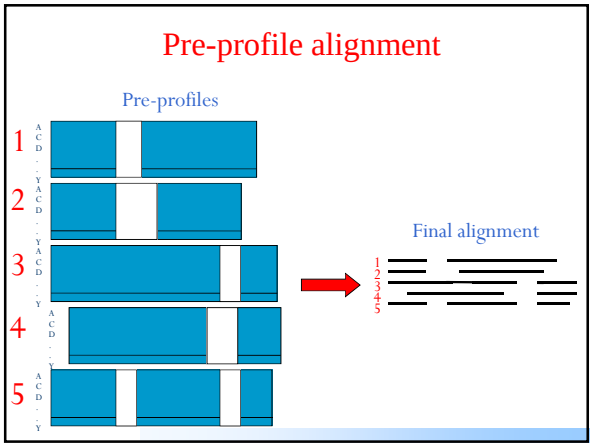


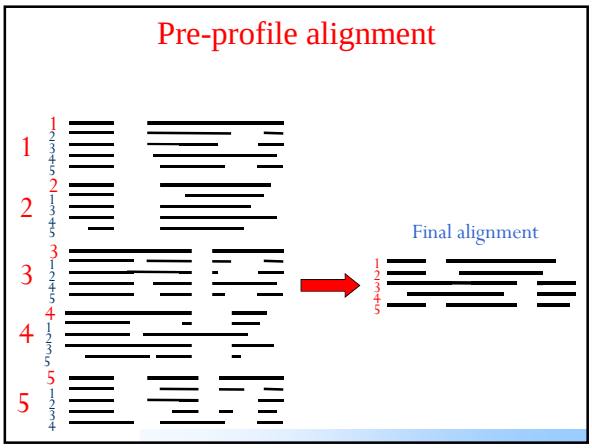
Flavodoxin-cheY: Pre-processing (prepro≥1500)

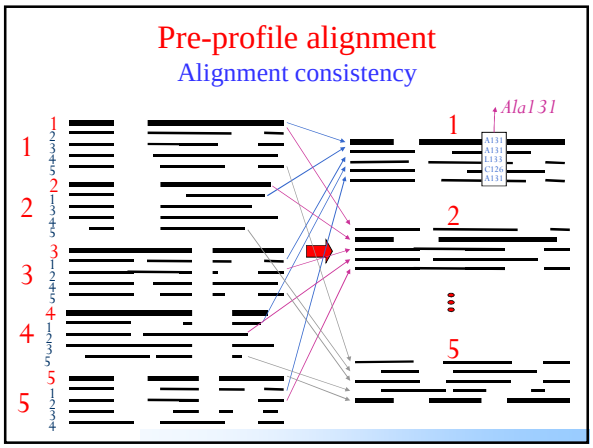
Profile pre-processing







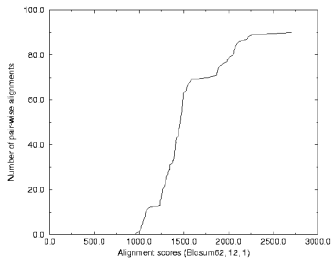




PRALINE pre-profile generation

- Idea: use the information from all query sequences to make a pre-profile for each query sequence that contains information from other sequences
- 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.
- 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≥0)

Preprocessed profile for sequence 2:

```
2fcr          KIGIFFSTSTGNTTEVADFIGKTLGAKADAPIDVDDVTDPQALKDYDLLFLGAPTWNTGADTERSGTSWDEFLYDKLPEVDMKDLPVAIFGLGDAEGYPD
1fx1          KALIVYGSTTGNTEYTAETIARQL-ANAGYEVDSRDAASVEAFEGFDLVLLGCSTW--GDD---SIELQDDFLFDSLEETGAQGRKVACFGCGDS-SY-E
4fxn          -MKIVYWSGTGNTEKMAELIAKGISGKDVNTINVSDVNIDELLNE-DILILGC---SAMGDEVLEESEFEPFIEEISTKISGKKVALGSYGWGDGKWMRD
FLAV_ANASP    KIGLFYGTQTGKTESVaEIIRDEFGNDVVTLHDVSEVTD---LNDYQYLIIgCPTWNIG---ELQ-SDW-EGLYSELDDVDFNGKLVAYfGTGDQIGYAD
FLAV_AZOVI    KIGLFFGSNTGKTRKVaKSIKKRFDTMSDA-LNVNRVS-AEDFAQYQFLILgTPTLGPGLSSDCENESWEEFL-PKIEGLDFSGKTVALfGLGDQVGYPE
FLAV_CLOAB    KISILYSSKTGKTERVaKLIEE--GVKRSGNIEVKDAVDKKFLQESEGIIFgTPTYYANISWEMK--KW----IDESSEFNLEGKLGAAfSTANAGGSDI
FLAV_DESDE    KVLIVFGSSTGNTESIaQKLEELIAA-GGHEVTLLNAADASALADYDAVLFgCSAWGM-EDLEMQ----DDFLFEEFNRFGLAGRKVAAfASGDQE-Y-E
FLAV_DESGI    KALIVYGSTTGNTEGVaEAIAKTLNSEGTTVVNVADVTAPGLAEGYDVVLLgCSTW--GDDEIELQEDFVP-LYEDLDRAGLKDKKVGVfGCGDS-SY-T
FLAV_DESSA    KSLIVYGSTTGNTETAaEYVAEAFENK-EIDVELKNVTDVSVANGYDIVLFgCSTW--G---EEEIELQDDFLYDSLENADLKGKKVSVfGCGDSD-Y-T
FLAV_DESVH    KALIVYGSTTGNTEYTaETIAREL-ADAGYEVDSRDAASVEAFEGFDLVLLgCSTW--GDD---SIELQDDFLFDSLEETGAQGRKVACfGCGDS-SY-E
FLAV_ECOLI    AIGIFFGSDTGNTENIaKMIQKQLG--KDV-ADVHDISSKEDLEAYDILLLgIPTWYYG----EAQCDWDDF-FPTLEEIDFNGKLVALfGCGDQEDYAE
FLAV_ENTAG    TIGIFFGSDTGQTRKVaKLIHQKLDGIADAPLDVRRATREQFL-SYPVLLLgTPTLGDGLPGVEAGSSWQEFT-NTLSEADLTGKTVALfGLGDQLNYSK
FLAV_MEGEL    MVEIVYWSGTGNTEAMaNEIEAAVAAGADVSVRFED-TNVDDVASKDVILLgCPA--MGSE-ELEDSVVEPFFTDLAPK--LKGKKVGLfGYGWGSG---
3chy          KELKFLVVDDFSTRRIVRNLLKELGFNEEAEDGVDALNKLQA-GGYGFVI---SDWNM---PNMDGL---ELLKTIRADGAMSALPVLMV---TAEAKKE

2fcr          NFCDAIEEIHDCFAKQGAKPVGFSNPDDYDYEESKSVRDGKFLGLPLDMVNDQIPMEKRVAGWVEAVVSETGV
1fx1          YFCGAVDAIEEKLKNLGA---------------EIVQD----GLRID--GDPRAARDDIVGWAHDVRGAI--
4fxn          -FEERMNG-YGCVVVE--TPLIVQNEPD----EAE--------------QDCIEFGKKIANI----------
FLAV_ANASP    NFQDAIGILEEKISQRgGKTVGYWSTDGYDFNDSKALRNGKFVGLALDEDNQSDLTDDRIKSwVAQLKSEFGL
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    YFCGAVDAIEEKLKNLgA---------------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          MKIVYWSGTGNTEKMAELIAKGIIESGKDVNTINVSDVNIDELLNEDILILGCSAMGDEVLEESEFEPFIEEISTKISGKKVALFGSYGWGDGKWMRDFE
1fx1          ALIVYGSTTGNTEYTAETIARQLANAGYEVDSRDAASVEAGGLFEGDLVLLGCSTWGDDSIEQDDFIPLFDSLETGAQGRKVACFGSYEYFCGA-VDAIE
2fcr          IGIFFSTSTGNTTEVADFIGKTL--GAKADAPIDVDDVTDPQALKDDLLFLGANTGADTERSGTSWDEFLYDKLPEVDMKDLPV-AIFGLGDAEGYPDFC
FLAV_ANASP    IGLFYGTQTGKTESVaEIIRD---EFGNDVVTLDVSQAEVTDLNDYQYLIIgCPTWNIGEL-QSDWEGLYSELDVDFNGKLVAYfGTIGYADNDAIGILE
FLAV_AZOVI    IGLFFGSNTGKTRKVaKSIKKRFDDETMS-DALNVNRVSAEDFAQYQFLILgTPTLGEGELENESWEEFLPKIGLDFSGKTVALfGQVGYPEGELYSFFK
FLAV_CLOAB    MKILYSSKTGKTERVaKLIEEGVKRSGNEVKTMNLDAVDKKFLQESEGIIFgTPTYYANI--SWEMKKWIDESSENLEGKLGAAfSTAGGSDIALLTILN
FLAV_DESDE    VLIVFGSSTGNTESIaQKLEELIAAGGHEVTLLNAADASAENLADYDAVLFgCSAWGMEDLEQDDFLSLFEEFNRGLAGRKVAAfAS---GDQEYVPAIE
FLAV_DESGI    ALIVYGSTTGNTEGVaEAIAKTLNSEGMETTVVNVADVTAPGLAGYDVVLLgCSTWGDDEIEQEDFVPLYEDLDAGLKDKKVGVfGSYTYFCGA-VDVIE
FLAV_DESSA    MSIVYGSTTGNTETAaEYVAEAFENKEIDVELKNVTDVSVADLGNYDIVLFgCSTWGEEEIEQDDFIPLYDSLNADLKGKKVSVfGDYTYFCGA-VDAIE
FLAV_DESVH    ALIVYGSTTGNTEYTaETIARELADAGYEVDSRDAASVEAGGLFEGDLVLLgCSTWGDDSIEQDDFIPLFDSLETGAQGRKVACfGSYEYFCGA-VDAIE
FLAV_ECOLI    TGIFFGSDTGNTENIaKMIQK---QLGKDVADVDIAKSSKEDLEAYDILLLgIPTYGEAQCDWDDFFPTLEEID--FNGKLVALfGDYAFCDAGTIRDIE
FLAV_ENTAG    IGIFFGSDTGQTRKVaKLIHQK-LDGIADA-PLDVRRATREQFLSYPVLLLgTPTLGDELVEASQYDSWQEFTNTDLTGKTVALfGNYSKNFVSAMRILY
FLAV_MEGEL    VEIVYWSGTGNTEAMaNEIEAAVKAAGADVESVRFEDTNVDDVASKDVILLgCPAMGSEELEDSVVEPFFTDLAPKLKGKKVGLfGSYGWGSGEWMDAWK
3chy          DKELKFLVVDDFSTMRRIVRNLLKELG--FNNVEEAEDGVD-ALNK-LQAGGYGVISDWNMPNMDGLELLKTI--RADGAMSALPVLMVTAEAKKENIIA

4fxn          ERMNGYGCVVVETPLIVQNEPDEAEQDCIEFGKKIANI
1fx1          EKLKNLGAEIVQDGLRIDGDPRAARDDIVGWAHDVRGA
2fcr          DAIEEHDCFAKQKPVGFSNPDDESKNDQIPMEKRVAGW
FLAV_ANASP    EKISGYGSKALRNGKFVGLALDEDNQDLTDDRIKVAQL
FLAV_AZOVI    DRTDGYEAVVVGLALDLDNQSGKTDERVAAwLAQIAPE
FLAV_CLOAB    HLMKgYGGVAFGKPYVHINEIQENEDENARfGERiANk
FLAV_DESDE    ERAKELgATIIAEGLKMEGDASNDPEAVASfAEDVLKQ
FLAV_DESGI    KKAEELgATLVASSLKIDGEPDSAE--VLDwAREVARV
FLAV_DESSA    EKLEKMgAVVIGDSLKIDGDPERDE--IVSwGSGIADI
FLAV_DESVH    EKLKNLgAEIVQDGLRIDGDPRAARDDIVGwAHDVRGA
FLAV_ECOLI    PRTAGYGLAFVGLAIDEDRQPELTAERVEKwVKQISEE
FLAV_ENTAG    DLVIARgCVVGNWPLLENNEPDQENQDLTELEKKPAVL
FLAV_MEGEL    QRTEDTgATVIGT-AIVNEMPDNA-PECKElGEAAAKA
3chy          AAQAGASGYVVK-PFTAATLEEKLNKIFEKLGM-----

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≥1500)

```
Preprocessed profile for sequence 1:

*     1 -    1. 1fx1
      2 -    7. FLAV_DESDE
      3 -    8. FLAV_DESGI
      4 -    9. FLAV_DESSA
      5 -   10. FLAV_DESVH

1fx1          PKALIVYGSTTGNTEYTAETIARQLANAGYEVDSRDAASVEAGGLFEGFDLVLLGCSTWG
FLAV_DESDE    SKVLIVFGSSTGNTESIaQKLEELIAAGGHEVTLLNAADASAENLADGYDAVLFgCSAWG
FLAV_DESGI    PKALIVYGSTTGNTEGVaEAIAKTLNSEGMETTVVNVADVTAPGLAEGYDVVLLgCSTWG
FLAV_DESSA    SKSLIVYGSTTGNTETAaEYVAEAFENKEIDVELKNVTDVSVADLGNGYDIVLFgCSTWG
FLAV_DESVH    PKALIVYGSTTGNTEYTaETIARELADAGYEVDSRDAASVEAGGLFEGFDLVLLgCSTWG

1fx1          DDSIELQDDFIPLFDSLEETGAQGRKVACFGCGDSSYEYFCGAVDAIEEKLKNLGAEIVQ
FLAV_DESDE    MEDLEMQDDFLSLFEEFNRFGLAGRKVAAfASGDQEYEHFCGAVPAIEERAKELgATIIA
FLAV_DESGI    DDEIELQEDFVPLYEDLDRAGLKDKKVGVfGCGDSSYTYFCGAVDVIEKKAEELgATLVA
FLAV_DESSA    EEEIELQDDFIPLYDSLENADLKGKKVSVfGCGDSDYTYFCGAVDAIEEKLEKMgAVVIG
FLAV_DESVH    DDSIELQDDFIPLFDSLEETGAQGRKVACfGCGDSSYEYFCGAVDAIEEKLKNLgAEIVQ

1fx1          DGLRIDGDPRAARDDIVGWAHDVRGAI
FLAV_DESDE    EGLKMEGDASNDPEAVASfAEDVLKQL
FLAV_DESGI    SSLKIDGEPDSA--EVLDwAREVLARV
FLAV_DESSA    DSLKIDGDPE--RDEIVSwGSGIADKI
FLAV_DESVH    DGLRIDGDPRAARDDIVGwAHDVRGAI
```

1

```
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          KIGIFFSTSTGNTTEVADFIGKTLGAKADAPIDVDDVTDPQALKDYDLLFLGAPTWNTGA
FLAV_ANASP    KIGLFYGTQTGKTESVaEIIRDEFGNDVVTLHDVSEVTD---LNDYQYLIIgCPTWNIG-
FLAV_AZOVI    KIGLFFGSNTGKTRKVaKSIKKRFDTMSDA-LNVNRVS-AEDFAQYQFLILgTPTLGPGL
FLAV_ECOLI    AIGIFFGSDTGNTENIaKMIQKQLG--KDV-ADVHDISSKEDLEAYDILLLgIPTWYYG-
FLAV_ENTAG    TIGIFFGSDTGQTRKVaKLIHQKLDGIADAPLDVRRATREQFL-SYPVLLLgTPTLGDGL

2fcr          DTERSGTSWDEFLYDKLPEVDMKDLPVAIFGLGDAEGYPDNFCDAIEEIHDCFAKQGAKP
FLAV_ANASP    --ELQ-SDW-EGLYSELDDVDFNGKLVAYfGTGDQIGYADNFQDAIGILEEKISQRgGKT
FLAV_AZOVI    SSDCENESWEEFL-PKIEGLDFSGKTVALfGLGDQVGYPENYLDALGELYSFFKDRgAKI
FLAV_ECOLI    ---EAQCDWDDF-FPTLEEIDFNGKLVALfGCGDQEDYAEYFCDALGTIRDIIEPRgATI
FLAV_ENTAG    PGVEAGSSWQEFT-NTLSEADLTGKTVALfGLGDQLNYSKNFVSAMRILYDLVIARgACV

2fcr          VGFSNPDDYDYEESKSVRDGKFLGLPLDMVNDQIPMEKRVAGWVEAVVSETGV
FLAV_ANASP    VGYWSTDGYDFNDSKALRNGKFVGLALDEDNQSDLTDDRIKSwVAQLKSEFGL
FLAV_AZOVI    VGSWSTDGYEFESSEAVVDGKFVGLALDLDNQSGKTDERVAAwLAQIAPEFGL
FLAV_ECOLI    VGHWPTAGYHFEASKGLADDHFVGLAID--EDRQPTAERVEKwVKQISEELHL
FLAV_ENTAG    VGNWPREGYKFSFSAALENNEFVGLPLDQENQYDLTEERIDSwLEKL--KPAV
```

2

Pre-profiles (prepro≥1500)

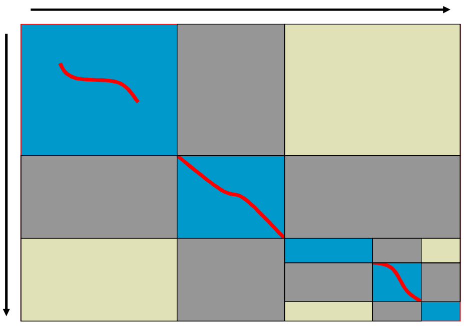
```
Preprocessed profile for sequence 13:

*     1 -   13. FLAV_MEGEL
      2 -    3. 4fxn

FLAV_MEGEL    MVEIVYWSGTGNTEAMaNEIEAAVKAAGADVESVRFEDTNVDDVASKDVILLgCPAMGSE
4fxn          -MKIVYWSGTGNTEKMAELIAKGIIESGKDVNTINVSDVNIDELLNEDILILGCSAMGDE

FLAV_MEGEL    ELEDSVVEPFFTDLAPKLKGKKVGLfGSYGWGSGEWMDAWKQRTEDTgATVIGTAIVNEM
4fxn          VLEESEFEPFIEEISTKISGKKVALFGSYGWGDGKWMRDFEERMNGYGCVVVETLIVQNE

FLAV_MEGEL    PDNAPECKElGEAAAKA
4fxn          PDEAQDCIEFGKKIANI
```

13

```
Preprocessed profile for sequence 14:

*     1 -   14. 3chy

3chy          ADKELKFLVVDDFSTMRRIVRNLLKELGFNNVEEAEDGVDALNKLQAGGYGFVISDWNMP

3chy          NMDGLELLKTIRADGAMSALPVLMVTAEAKKENIIAAAQAGASGYVVKPFTAATLEEKLN

3chy          KIFEKLGM
```

14

Local pre-processing



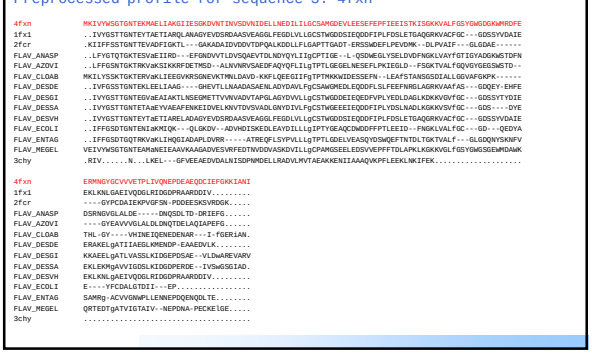
Local alignments are calculated from high to low scoring – each time the sequence parts corresponding to a selected local alignment are blocked such that a next local alignment has to emerge before or after the earlier selected one – this preserves co-linearity of the local alignments and assocaited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

Local pre-processing
(locprepro≥0)

Preprocessed profile for sequence 2: 2fcr

Local pre-processing
(locprepro≥0)

Preprocessed profile for sequence 3: 4fxn



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

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

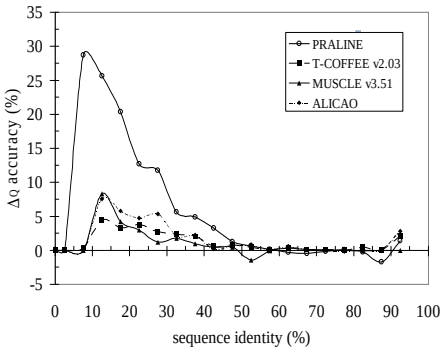
Flavodoxin-cheY: Local Pre-processing (locprepro≥300)

PSI-PRALINE

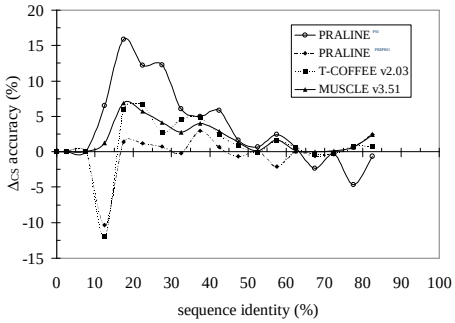
Multiple alignment of distant sequences using PSI-BLAST

- Perform a PSI-BLAST search for each sequence
- 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)

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

ALICAO (0.00)

```
1GZOA .....SEIYG IHAVQALLER APERFQEVFI LKG.REDKRL LP....LIHA LESQQVVIQL ANRQYLDEKS DGAVHQGIIA RVKPGRQ... ..........
1IPAA MRITSTANPR IKELARLLER KHRDSQRRFL IKGAREIERA LQAGIELEQA LVWEGGLNPE EQQVYAALLA LLEVSEAVLK KLSVRDNPAG LIALARMPER
```

MUSCLE v3.51 (0.15)

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

T-COFFEE v2.03 (0.15)

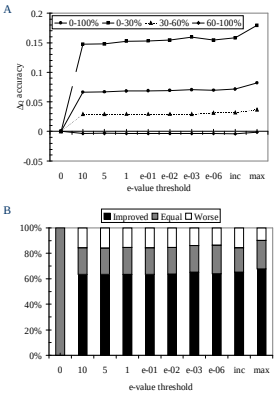
```
1GZOA .....SEIYG IHAVQALLER APERFQEVFI LKG.REDKRL LP....LIHA LESQQVVIQL ANRQY..... LDEKSDGAVH Q........G IIARVKPGRQ
1IPAA MRITSTANPR IKELARLLER KHRDSQRRFL IKGAREIERA LQAGIELEQA LVWEGGLNPE EQQVYAALLA LLEVSEAVLK KLSVRDNPAG LIALARMPER
                                                                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

PSI-PRALINE (0.92)

```
1GZOA ..... ..... ..... .....SE IYGIHAVQAL LEKAPERFQE VPILKGREDK RLLPLIHALE SQQVVIQLAN RQYLDEKSDG AVHQGIIARV KPGRQ
1IPAA MRITSTANPR IKELARLLER KHRDSQRRFL IKGAREIERA LQAG.IELEQ ALVWEGGLNP EEQQVVAALL ....AALLALLEVS EAVLKKLSVR DNPAGLIALA RMPER
                  ^^ ^^^^^^^^^^ ^^^^   ^^^ ^^^^^^^^^^ ^^^^^^          ^^^^^^ ^^^^^^^^^^ ^^^^^^^^^^ ^^^^^
```

HOMSTRAD REFERENCE

```
1GZOA ..... ..... ..... .....SE IYGIHAVQAL LEKAPERFQE VPILKGREDK RLLPLIHALE SQQVVIQLAN RQYLDEKSDG AVHQGIIARV KPGRQ
DSSP  ..... ..... ..... ..... E EESHHHHHHH HHS GGGB E EEEESS   T TTHHHHHHHH HHT EEEES  SHHHHHTTTS    TTEEEEE
1IPAA MRITSTANPR IKELARLLER KHRDSQRRFL IKGAREIERA LQAGIE.LEQ ALVWEGGLNP EEQQVVAALL ....ALLEVS EAVLKKLSVR DNPAGLIALA RMPER
DSSP   EE  TTSHH HHHHHGGGSH HHHHHHTTEE EESHHHHHHH HHTT  .B E EEEETT   H HHHHHHH     ....EEEEE  HHHHHHH  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

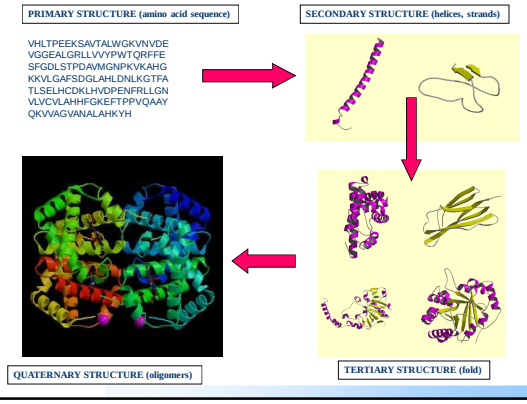
Objective: integrate secondary structure information to anchor alignments and avoid error

Additional strategies for multiple sequence alignment

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

Objective: try to avoid (early) errors

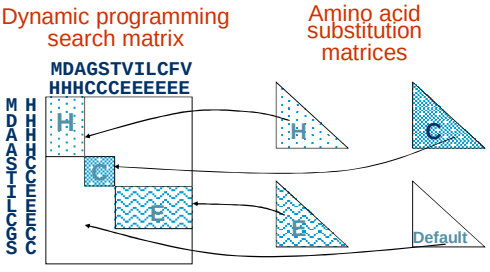
Protein structure hierarchical levels



Why use (predicted) structural information

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

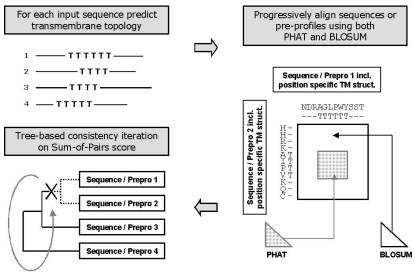


Using predicted secondary structure

PRALINE™ (Pirovano et al., 2008)

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

PRALINE™ strategy



Substitution matrices

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

Transmembrane topology predictors

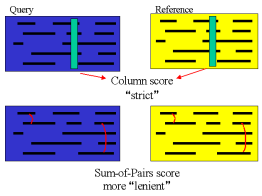
- **HMMTOP** (Tusnády and Simon, 2001)
- **TMHMM** (Krogh et al., 2001)
- **PHOBIUS** (Käll et al., 2005)

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

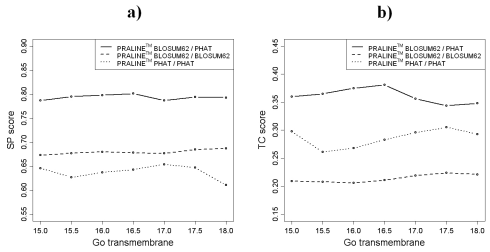
- **STAM** (Shafrir and Guy, 2004)

Benchmark

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



Independent contributions PHAT matrix and gap values



Strategies for multiple sequence alignment

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

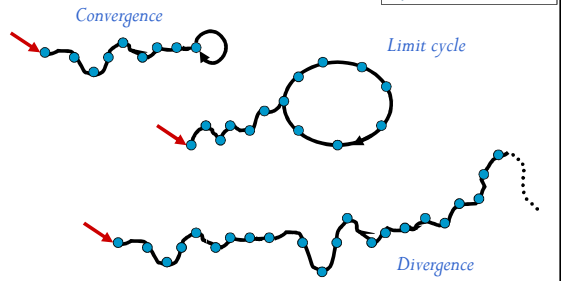
Objective: try to avoid (early) errors

Multiple alignment methods

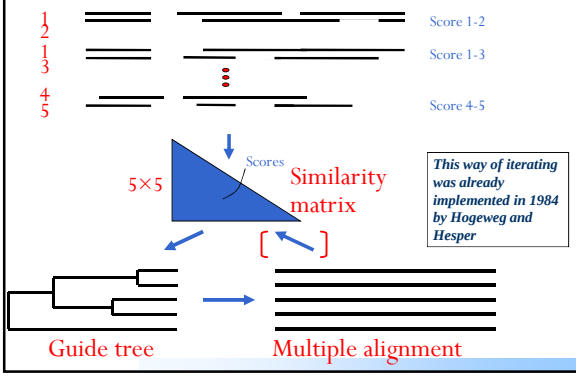
- Multi-dimensional dynamic programming
  > extension of pairwise sequence alignment.
- Progressive alignment
  > incorporates phylogenetic information to guide the alignment process
- **Iterative alignment**
  > correct for problems with progressive alignment by repeatedly realigning subgroups of sequence

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.



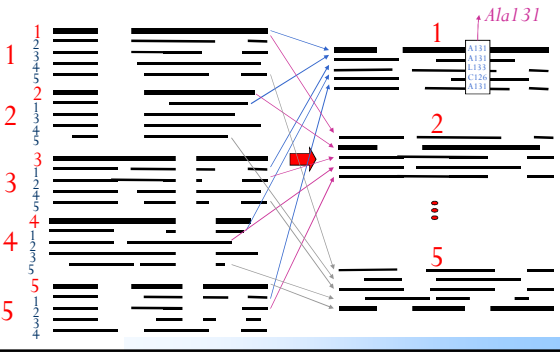
Iterate similarity matrix, guide tree and MSA



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

Pre-profile alignment

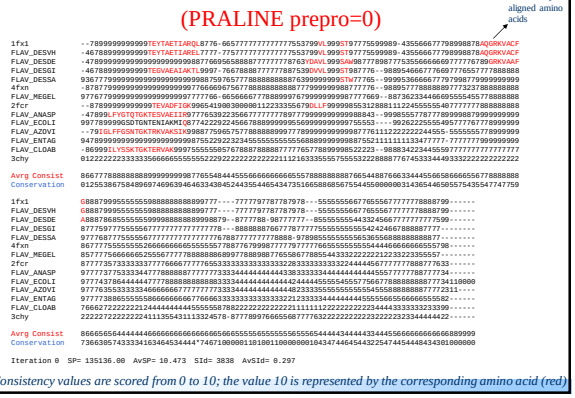
Alignment consistency



Flavodoxin-cheY consistency scores

(PRALINE prepro=0)

Completely consistently aligned amino acids



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

Flavodoxin-cheY consistency scores

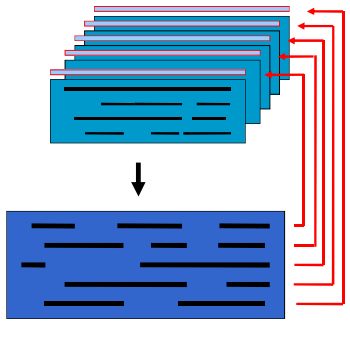
(PRALINE prepro=1500)

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

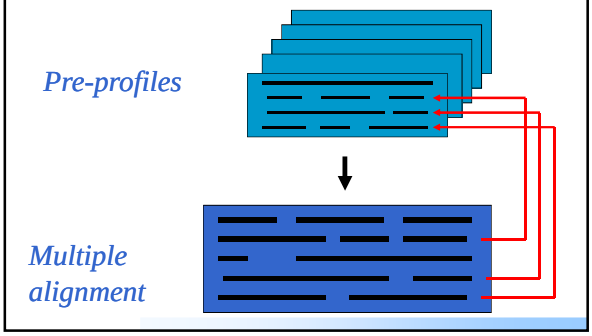
Consistency iteration

Pre-profiles

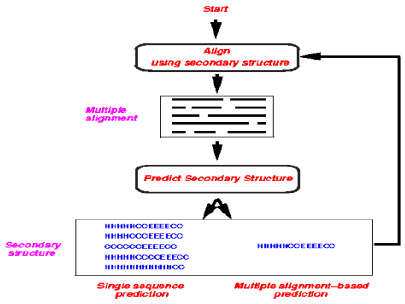
Multiple alignment positional consistency scores



Pre-profile update iteration



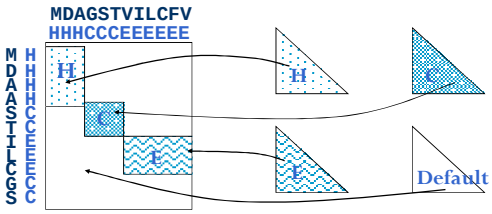
Secondary structure-induced alignment



PRALINE
Using secondary structure for alignment

Dynamic programming search matrix

Amino acid exchange weights matrices



Flavodoxin-cheY multiple alignment/ secondary structure iteration
cheY SSEs

```
3chy-AA SEQUENCE||        AA  |ADKELKFLVVDDFSTMRRIVRNLLKELGFNNVEEAEDGVDALNKLQAGGYGFVISDWNMP|
3chy-ITERATION-0||        PHD |  EEEEEEE   HHHHHHHHHHHHHHHHH      E  HHHHHHHHH HHHEEE    |

3chy-AA SEQUENCE||        AA  |NMDGLELLKTIRADGAMSALPVLMVTAEAKKENIIAAAQAGASGYVVKPFTAATLEEKLNKIFEKLGM|
3chy-ITERATION-0||        PHD |  HHHHHHEEEEEE        HHHHHHHHHHHHHHHHH            HHHHHHHHHHHHHH  |
```

Flavodoxin-cheY multiple alignment/ secondary structure iteration
cheY SSEs

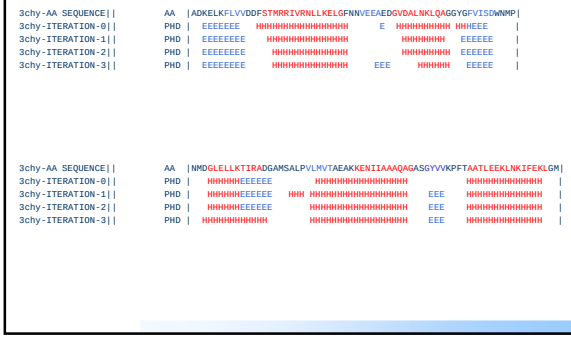
```
3chy-AA SEQUENCE||        AA  |ADKELKFLVVDDFSTMRRIVRNLLKELGFNNVEEAEDGVDALNKLQAGGYGFVISDWNMP|
3chy-ITERATION-0||        PHD |  EEEEEEE   HHHHHHHHHHHHHHHHH      E  HHHHHHHHH HHHEEE    |
3chy-ITERATION-1||        PHD |  EEEEEEE    HHHHHHHHHHHHHHH          HHHHHHHH  EEEEEE   |

3chy-AA SEQUENCE||        AA  |NMDGLELLKTIRADGAMSALPVLMVTAEAKKENIIAAAQAGASGYVVKPFTAATLEEKLNKIFEKLGM|
3chy-ITERATION-0||        PHD |  HHHHHHEEEEEE        HHHHHHHHHHHHHHHHH            HHHHHHHHHHHHHH  |
3chy-ITERATION-1||        PHD |  HHHHHHEEEEEE   HHH HHHHHHHHHHHHHHHHH    EEE    HHHHHHHHHHHHHH  |
```

Flavodoxin-cheY multiple alignment/ secondary structure iteration
cheY SSEs

```
3chy-AA SEQUENCE||        AA  |ADKELKFLVVDDFSTMRRIVRNLLKELGFNNVEEAEDGVDALNKLQAGGYGFVISDWNMP|
3chy-ITERATION-0||        PHD |  EEEEEEE   HHHHHHHHHHHHHHHHH      E  HHHHHHHHH HHHEEE    |
3chy-ITERATION-1||        PHD |  EEEEEEE    HHHHHHHHHHHHHHH          HHHHHHHH  EEEEEE   |
3chy-ITERATION-2||        PHD |  EEEEEEE    HHHHHHHHHHHHHHH          HHHHHHHH  EEEEEE   |

3chy-AA SEQUENCE||        AA  |NMDGLELLKTIRADGAMSALPVLMVTAEAKKENIIAAAQAGASGYVVKPFTAATLEEKLNKIFEKLGM|
3chy-ITERATION-0||        PHD |  HHHHHHEEEEEE        HHHHHHHHHHHHHHHHH            HHHHHHHHHHHHHH  |
3chy-ITERATION-1||        PHD |  HHHHHHEEEEEE   HHH HHHHHHHHHHHHHHHHH    EEE    HHHHHHHHHHHHHH  |
3chy-ITERATION-2||        PHD |  HHHHHHEEEEEE        HHHHHHHHHHHHHHHHH    EEE    HHHHHHHHHHHHHH  |
```

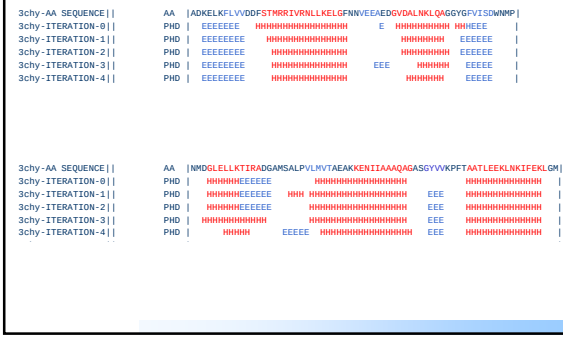
Flavodoxin-cheY multiple alignment/ secondary structure iteration

cheY SSEs



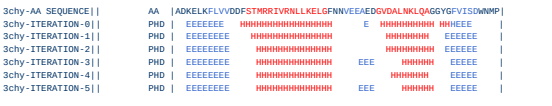
Flavodoxin-cheY multiple alignment/ secondary structure iteration

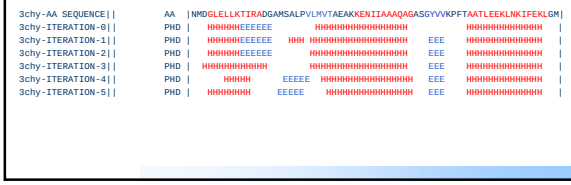
cheY SSEs



Flavodoxin-cheY multiple alignment/ secondary structure iteration

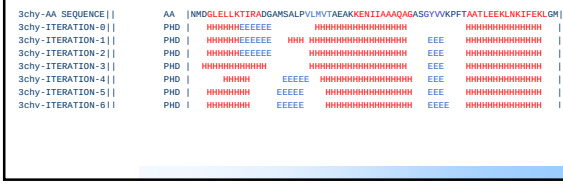
cheY SSEs





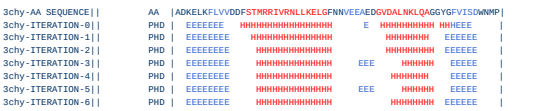
Flavodoxin-cheY multiple alignment/ secondary structure iteration

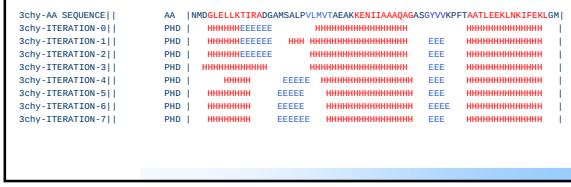
cheY SSEs



Flavodoxin-cheY multiple alignment/ secondary structure iteration

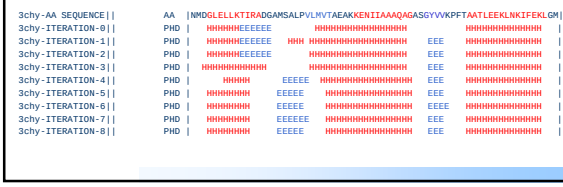
cheY SSEs





Flavodoxin-cheY multiple alignment/ secondary structure iteration

cheY SSEs



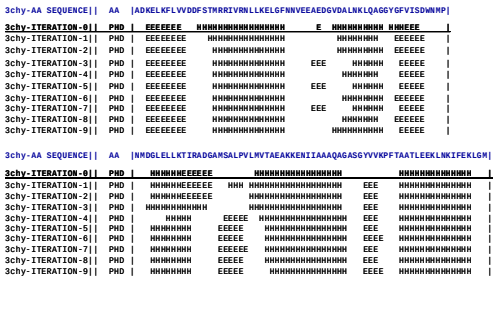
Flavodoxin-cheY multiple alignment/ secondary structure iteration

cheY SSEs

Flavodoxin-cheY multiple alignment/ secondary structure iteration

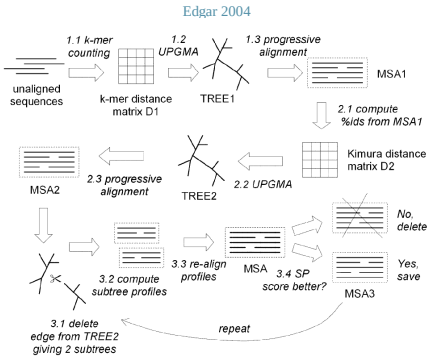
cheY SSEs

Is the initial SS prediction good enough?



MUSCLE

Edgar 2004



PRALINE and MUSCLE method

- **PRALINE and MUSCLE use different formalisms 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 (2004) calls the Muscle scoring scheme “Log-expectation scoring (LE)”

So what do we do ?

- **A single shot for a good alignment without thinking: MUSCLE, T-COFFEE, PROBCONS (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:**
  - **include Dialign2 (local)**
  - **PROBCONS scores well in recent assessments**

Recap

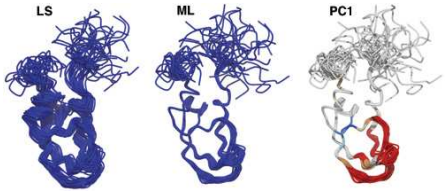
- Pairwise alignment by Dynamic Programming
- Weighting schemes to use information from all sequences right from the start during the progressive MSA protocol:
  - Profile pre-processing (global/local) (PRALINE)
  - Matrix extension (well balanced scheme) (T-Coffee)
- Smoothing alignment signals:
  - Consistency based mixing of local and global alignment (T-Coffee and PRALINE)
  - Homology-extended alignment (PRALINE)
- Using additional information:
  - secondary structure driven alignment (PRALINE(TM))
- 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.

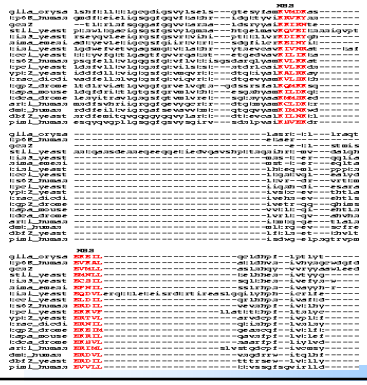
8 categories:

- *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



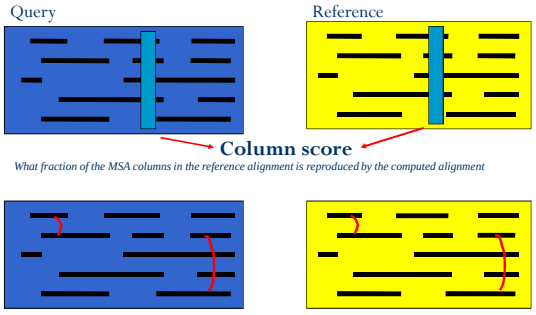
Scoring a single MSA with the Sum-of-pairs (SP) score

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

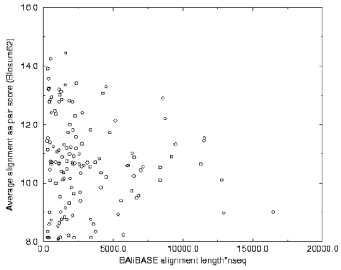
Sum-of-Pairs score

- 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

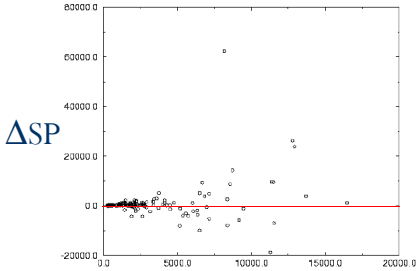


Evaluating multiple alignments



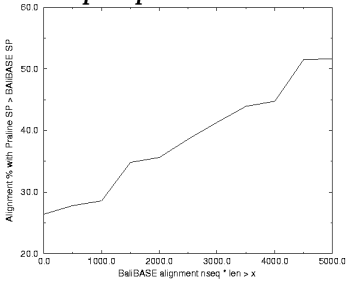
Evaluating multiple alignments

Charlie Chaplin problem



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 (%)
Name	Protocol			Cat 1 (82)	Cat 2 (23)	Cat 3 (12)	Cat 4 (13)	Cat 5 (11)	Tot (141)	
C	ClustalW pw			70.6	36.7	43.0	36.0	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	**80.7**	**37.3**	**52.9**	**83.2**	**88.7**	**72.1**	

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).

25

Comparing T-coffee with other methods

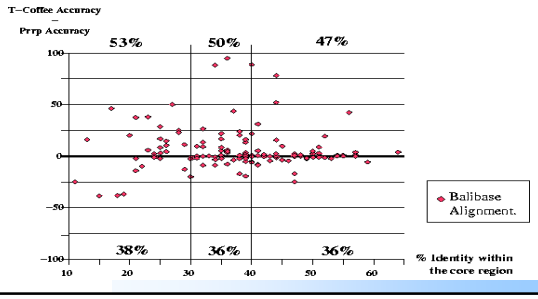
Table 2

Evaluation of four packages with BaliBase

Method Name	BaliBase 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**
Prrp	78.6	32.5	50.2	51.1	82.7	66.4	59.0	36.9*
T-Coffee (CLE)	**80.7**	**37.3**	**52.9**	**83.2**	**88.7**	**72.1**	**68.7**	

BAliBASE benchmark alignments

Comparison of T–Coffee and Prrp



END