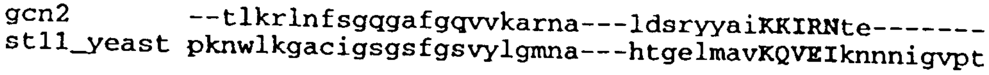
New Approaches to Multiple Sequence Alignment

Featuring: T Coffee & POA

Discussed by Catherine S. Grasso

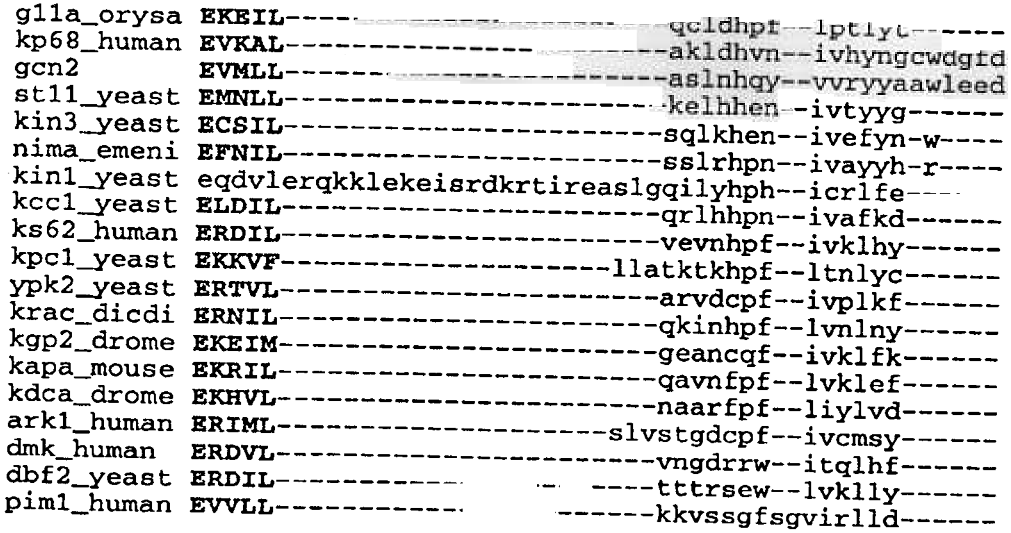
What Is Multiple Sequence Alignment?

- Pairwise Sequence Alignment:



```
gcn2        --tlkrlnfsgqgafgqvvkarna---ldsryyaiKKIRNte-------
st11_yeast  pknwlkgacigsgsfgsvylgmna---htgelmavKQVEIknnnigvpt
```

- Multiple Sequence Alignment:



```
g11a_orysa  EKEIL----------------------qcldhpf--lptlyc------
kp68_human  EVKAL----------------------akldhvn--ivhyngcwdgfd
gcn2        EVMLL----------------------aslnhqy--vvryyaawleed
st11_yeast  EMNLL----------------------kelhhen--ivtyyg------
kin3_yeast  ECSIL----------------------sqlkhen--ivefyn-w----
nima_emeni  EFNIL----------------------sslrhpn--ivayyh-r----
kin1_yeast  eqdvlerqkklekeisrdkrtireaslgqilyhph--icrlfe------
kcc1_yeast  ELDIL----------------------qrlhhpn--ivafkd------
ks62_human  ERDIL----------------------vevnhpf--ivklhy------
kpc1_yeast  EKKVF--------------------llatktkhpf--ltnlyc------
ypk2_yeast  ERTVL----------------------arvdcpf--ivplkf------
krac_dicdi  ERNIL----------------------qkinhpf--lvnlny------
kgp2_drome  EKEIM----------------------geancqf--ivklfk------
kapa_mouse  EKRIL----------------------qavnfpf--lvklef------
kdca_drome  EKHVL----------------------naarfpf--liylvd------
ark1_human  ERIML--------------------slvstgdcpf--ivcmsy------
dmk_human   ERDVL----------------------vngdrrw--itqlhf------
dbf2_yeast  ERDIL----------------------tttrsew--lvklly------
pim1_human  EVVLL----------------------kkvssgfsgvirlld------
```

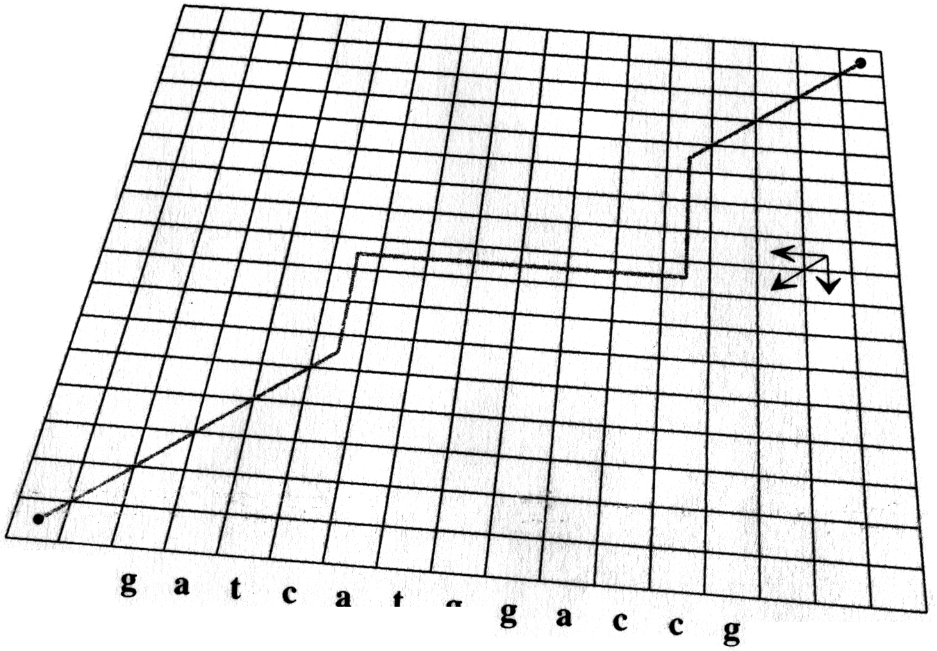
Why Do Multiple Sequence Alignment?

- Pairwise sequence alignment models an evolutionary relationship between two sequences. It models the process of insertion, deletion, and mutation by which the two sequences diverged from each other.

- Multiple sequence alignment models the evolutionary relationship between N sequences. It models the process of insertion, deletion, and mutation by which the N sequences diverged from a common ancestor.

- Since sequence similarities between proteins reflect structural and functional similarities, we can use a multiple sequence alignment to infer these relationships.

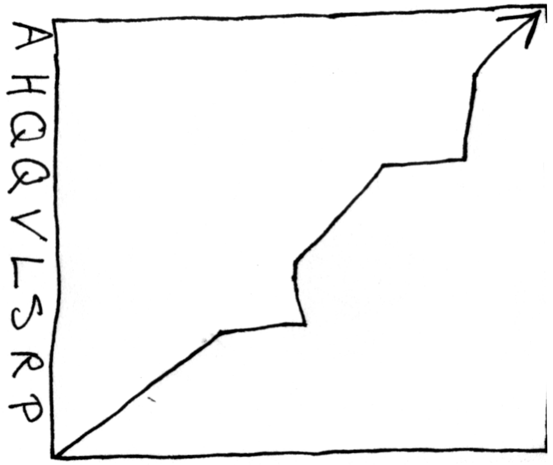
PSA With Dynam c Programming

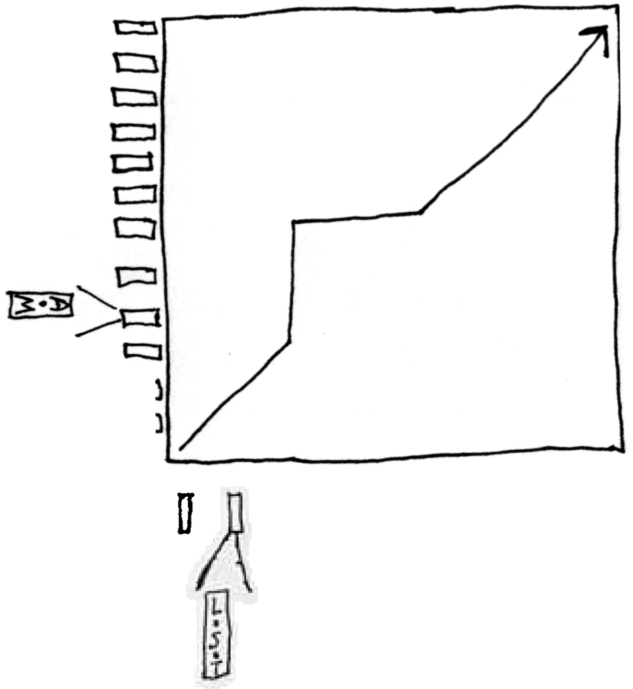


Finding a PSA   Finding a path through a 2 Dim matrix   Is O(L

* L   the   equence   ength

PSA of eaf nodes . branch nodes



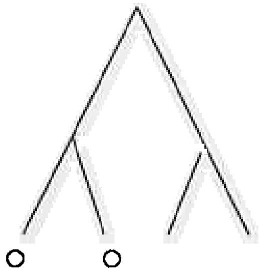


- PS ode

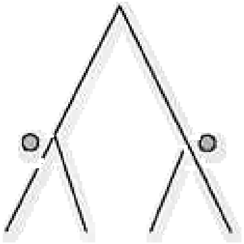
- PSA of branch nodes.

Align N sequences using guide tree:

1. Use standard PSA to align leaf sequence.



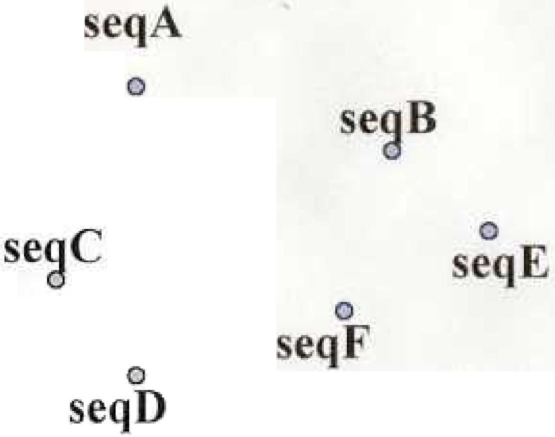
2. Profile multiple sequence alignments at branch nodes.
3. Use standard PSA on profiles.



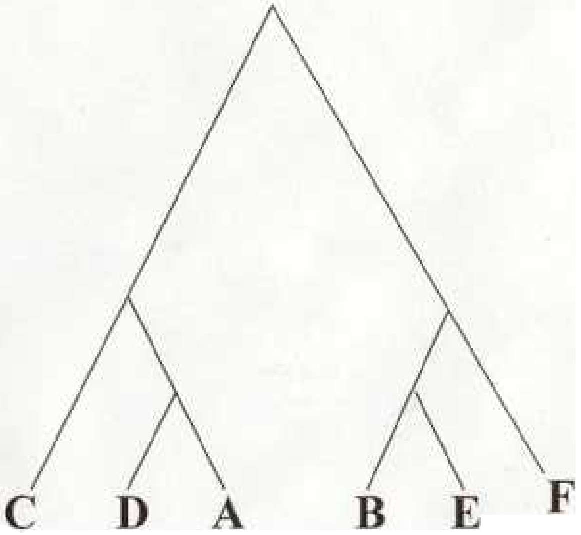
4. Recurse.

Optimal MSA Is Not Possible!
What's done instead?
Progressive Alignment (CLUSTAL)

Compute PSAs of all N sequences.

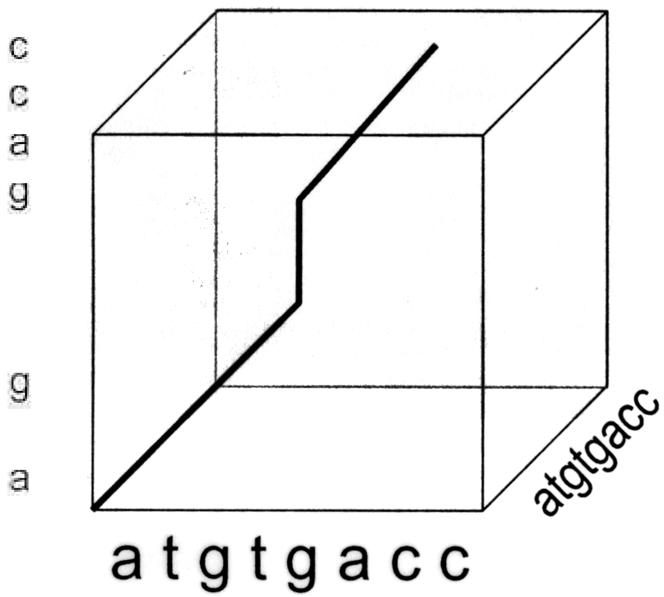


Build Guide Tree



MSA With Dynamic Programming

High dimensional MSA



Finding an MSA   Finding a path through an N-Dim matrix   is O L^N)

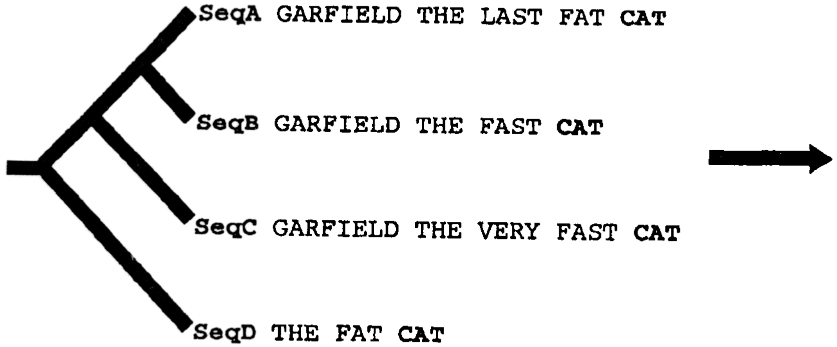
* N is the number of sequences and L is the sequence length

Problems With Progressive Alignment

- Greedy Algorithm results in local minimum. → T-Coffee
- Artifact Gaps. → POA

Local Minimum Problem

a)Regular Progressive Alignment Strategy



SeqA GARFIELD THE LAST FAT CAT

SeqB GARFIELD THE FAST CAT

SeqC GARFIELD THE VERY FAST CAT

SeqD THE FAT CAT

```
SeqA GARFIELD THE LAST FA-T CAT
SeqB GARFIELD THE FAST CA-T ---
SeqC GARFIELD THE VERY FAST CAT
SeqD -------  THE ---- FA-T CAT
```

T-Coffee Conclusions

- Despite its being somewhat slower than CLUSTALW T-Coffee is being used by more and more bioinformaticists Their method being fairly heuristic is clearly not the last word on the local minimum problem in MSA.

T-Coffee Compared With Other MSA Methods on Balibase Set

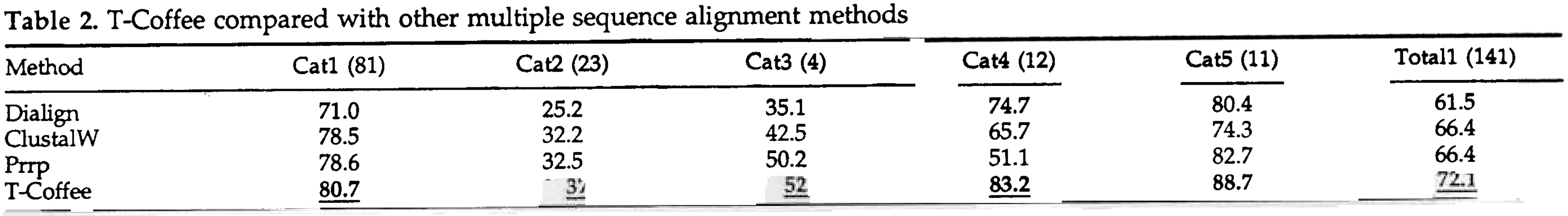


Table 2. T-Coffee compared with other multiple sequence alignment methods

Method	Cat1 (81)	Cat2 (23)	Cat3 (4)	Cat4 (12)	Cat5 (11)	Total1 (141)
Dialign	71.0	25.2	35.1	74.7	80.4	61.5
ClustalW	78.5	32.2	42.5	65.7	74.3	66.4
Prrp	78.6	32.5	50.2	51.1	82.7	66.4
T-Coffee	80.7	37	52	83.2	88.7	72.1

is percent average accuracy

Other Methods:

Dialign 2 segment based method constructs MSAs by assembling collection of high scoring segments in a sequence independent progressive fashion. (Morgenstern, 1999)

ClustALW progressive alignment (Thompson, et al 1994)

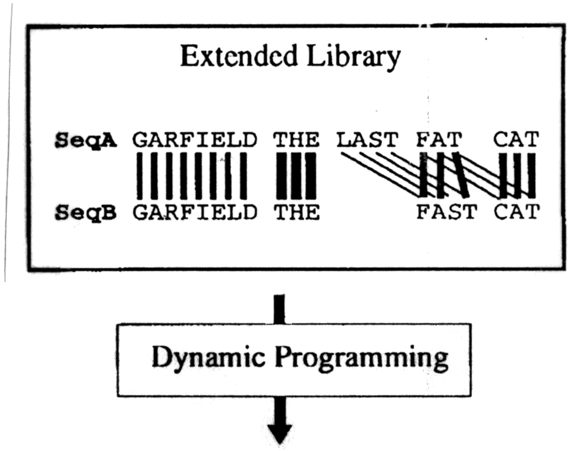
Prrp attempts to simultaneously align all sequences in an iterative manner (Gotoh, 1996)

Balibase MSA Test Set

- Contains 141 protein alignments. Constructed by manual structure comparison, validated using SSAP or DALI.

- Five Categories: a) phylogenetically equidistant members, b) one orphan with group of closely related, c) two distant groups, d) long terminal insertions, e) long internal insertions.

Progressive Alignment



```
SeqA GARFIELD THE LAST FA-T CAT
SeqB GARFIELD THE ---- FAST CAT
```

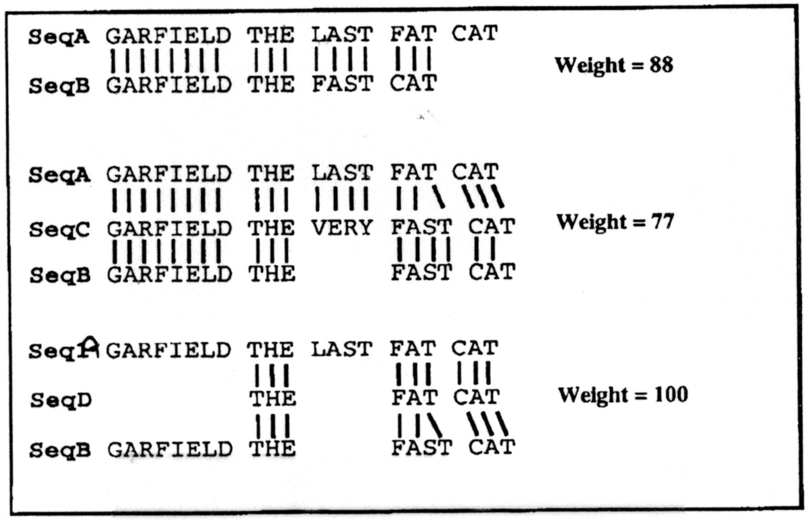
→ Extended to PS, i.e. profile MSAs.

Branch Nodes

Pairwise Alignment of leaf Nodes

Extended Library Construction

c)Extended Library for seq1 and seq2



```
SeqA GARFIELD THE LAST FAT CAT
     |||||||| ||| |||| |||
SeqB GARFIELD THE FAST CAT              Weight = 88


SeqA GARFIELD THE LAST FAT CAT
     |||||||| ||| |||| ||\ \\\
SeqC GARFIELD THE VERY FAST CAT         Weight = 77
     |||||||| |||      |||| ||
SeqB GARFIELD THE      FAST CAT


SeqA GARFIELD THE LAST FAT CAT
              |||      ||| |||
SeqD          THE      FAT CAT          Weight = 100
              |||      ||\ \\\
SeqB GARFIELD THE      FAST CAT
```

Consider G in Garfield:

Let S(G) be the G of sequence S.

$W_1(A(G), B(G)) = 88$ in primary library

$W_2(A(G), B(G)) = \min(W_1(A(G), C(G)), W_1(C(G), B(G))) = 77$

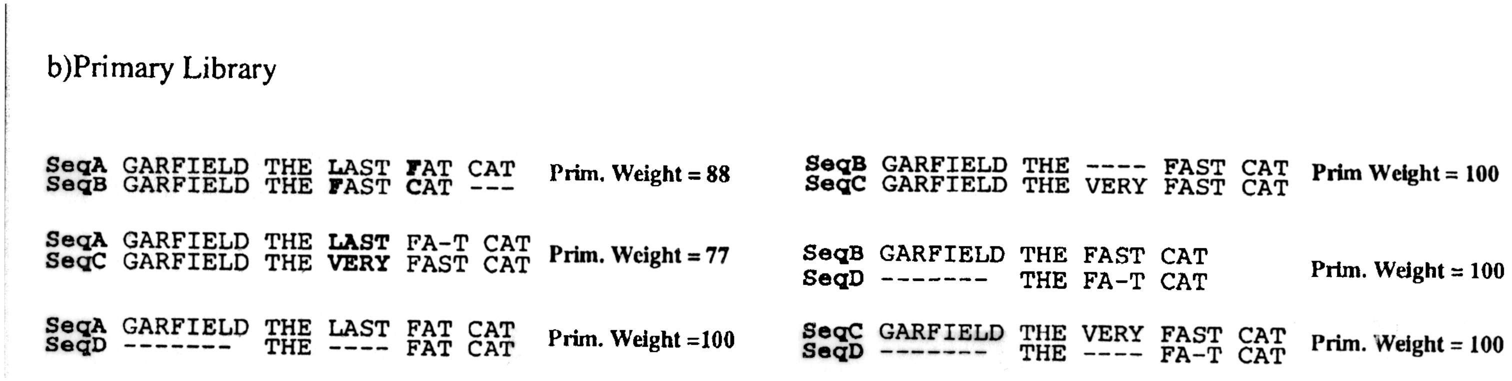
in extended library

where $W_1(A(G), C(G)) = 77$

& $W_1(C(G), B(G)) = 100$

Not all triplets bring information. D does not contain any information relative to A(G) or B(G).

Primary Library Construction



b)Primary Library

```
SeqA GARFIELD THE LAST FAT CAT    Prim. Weight = 88     SeqB GARFIELD THE ---- FAST CAT  Prim Weight = 100
SeqB GARFIELD THE FAST CAT ---                          SeqC GARFIELD THE VERY FAST CAT

SeqA GARFIELD THE LAST FA-T CAT   Prim. Weight = 77     SeqB GARFIELD THE FAST CAT
SeqC GARFIELD THE VERY FAST CAT                         SeqD -------- THE FA-T CAT       Prim. Weight = 100

SeqA GARFIELD THE LAST FAT CAT    Prim. Weight =100     SeqC GARFIELD THE VERY FAST CAT  Prim. Weight = 100
SeqD -------- THE ---- FAT CAT                          SeqD -------- THE ---- FA-T CAT
```

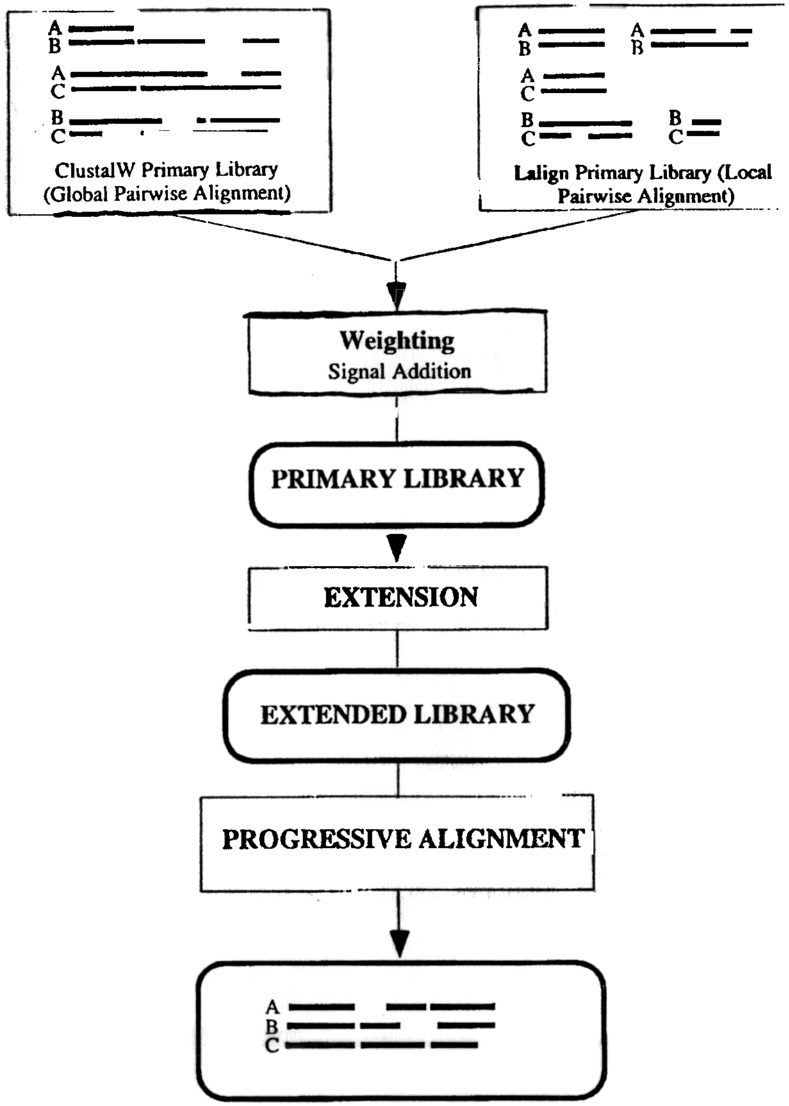
Consists of weighted pairs of residues. Weight used is percent d in alignment pair derives from. Reflects reliability of pair.

Examples:

W, (Seq A: A, 3, SeqC: E, 13) = 77

W, (Seq B: T, 15, Seq D: T, 6) = 100

T-Coffee Strategy



T-Coffee Objective:

Use as much **information** from **pre-alignment** to not only **guide** the **order of** sequence **alignment**, but the **alignments** themselves

Gap Artifact Problem

Alignment A :

```
..   ACATGTCGAT
TGCAC.....TCGAT
```

(S = TGCACTCGAT)

Alignment A'

```
ACATG.....TCGAT
  ...TGCACTCGAT
```

S = TGCACTCGAT

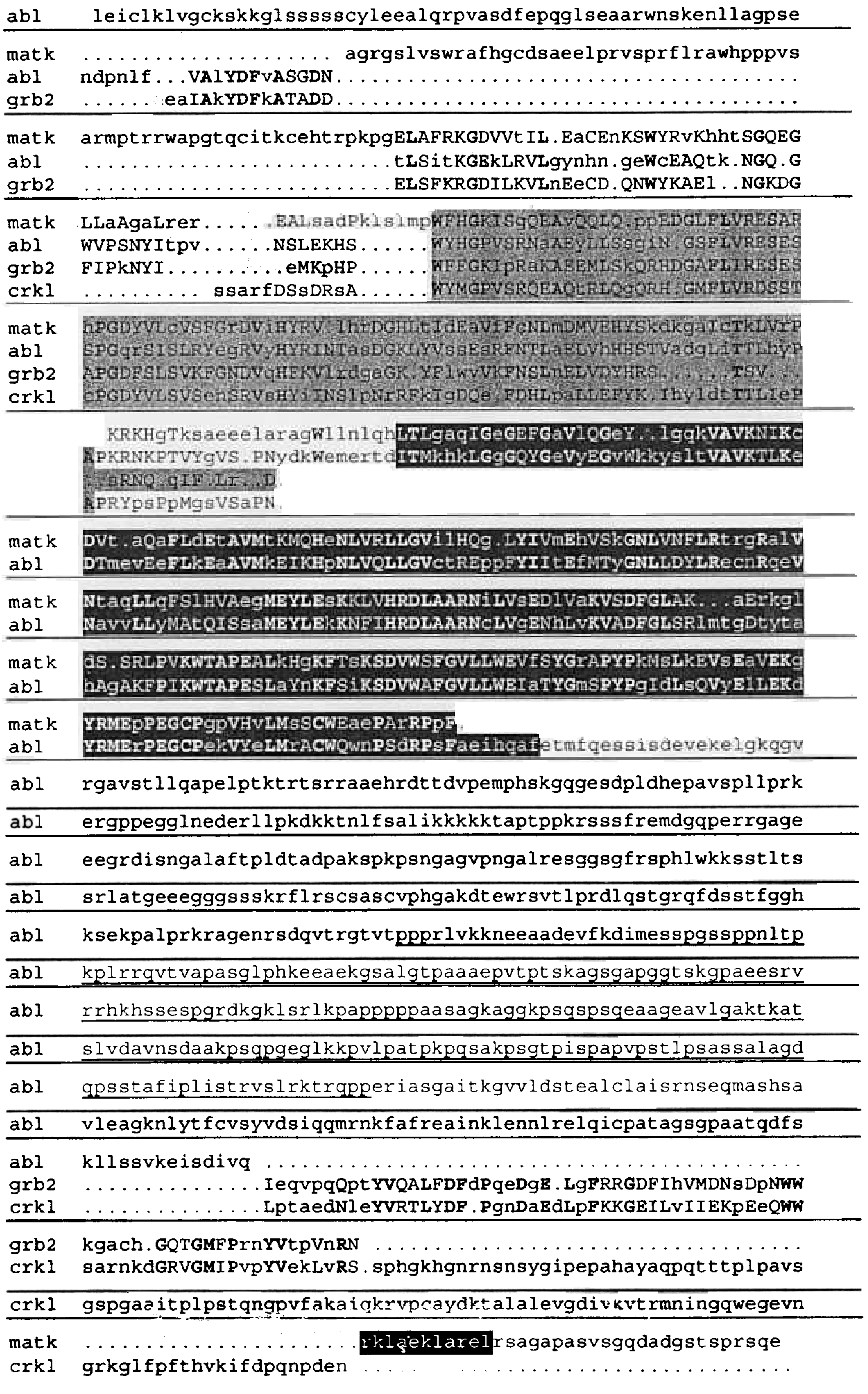
$A = A'$

What do we really want to know about an MSA?

1. The order of letters within a sequence. 5’ to 3’ or N-terminal to C-terminal.
2. Which letters are aligned between sequences.

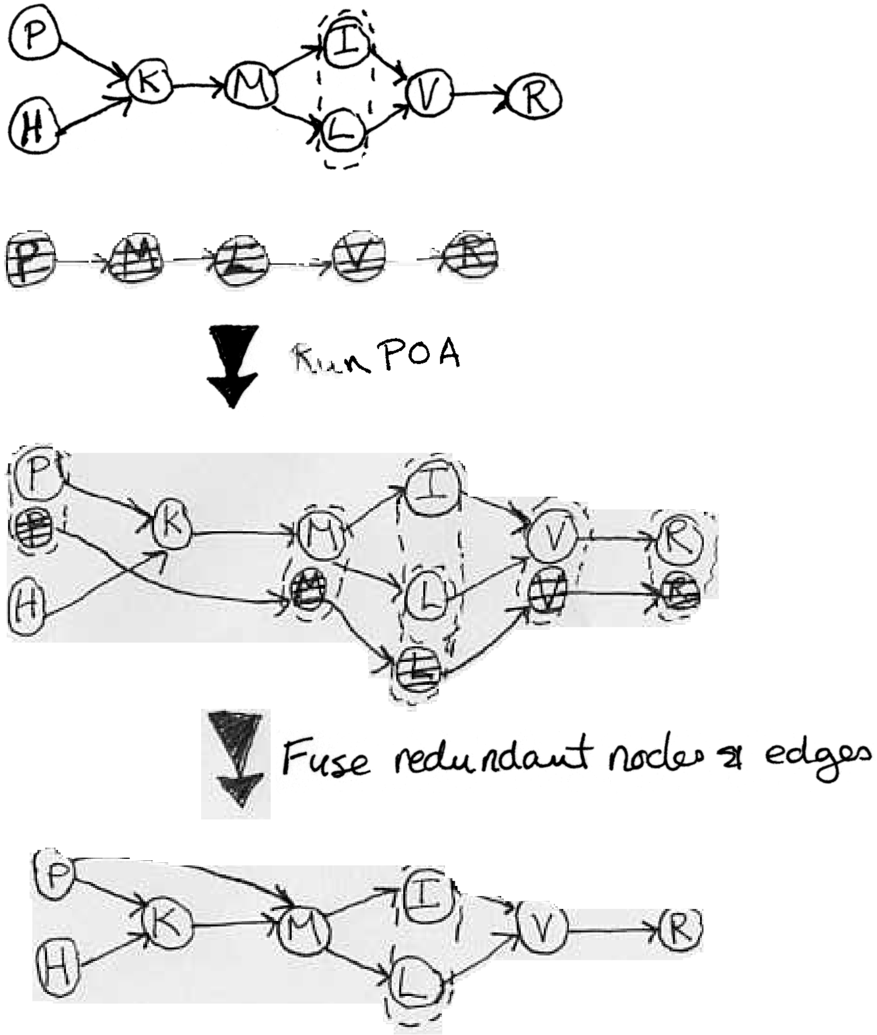
Ordering can be imposed by one sequence on another sequence only through alignment.

RC-MSA of Four Proteins

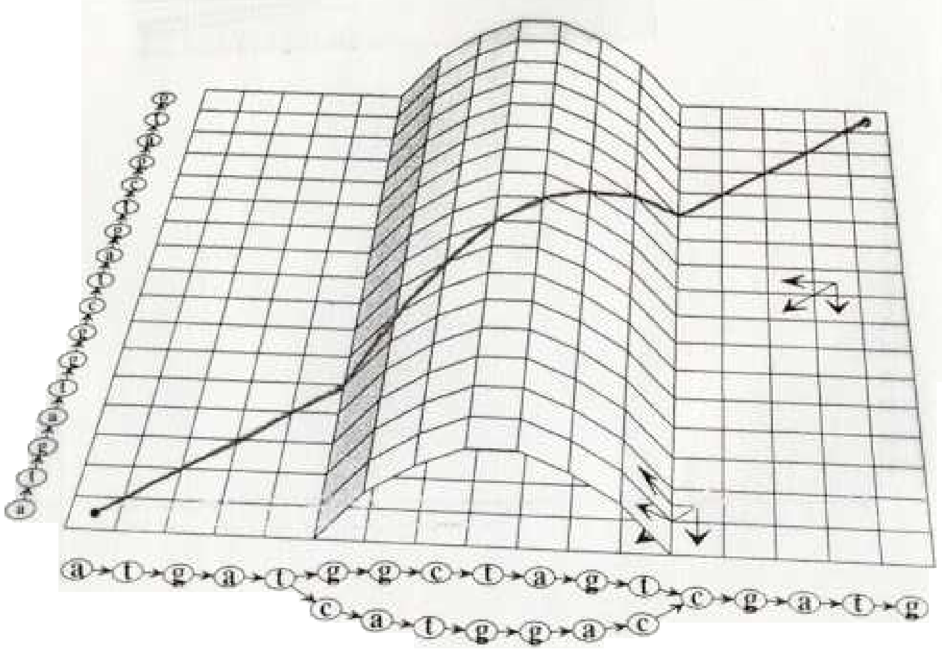


```
abl     leiclklvgckskkglssssscyleealqrpvasdfepqglseaarwnskenllagpse

matk    .....................agrgslvswrafhgcdsaeelprvsprflrawhpppvs
abl     ndpnlf...VAlYDFvASGDN......................................
grb2    ......eaIAkYDFkATADD.......................................

matk    armptrrwapgtqcitkcehtrpkpgELAFRKGDVVtIL.EaCEnKSWYRvKhhtSGQEG
abl     ..........................tLSitKGEkLRVLgynhn.geWcEAQtk.NGQ.G
grb2    ..........................ELSFKRGDILKVLnEeCD.QNWYKAEl..NGKDG

matk    LLaAgaLrer......EALsadPklslmpWFHGKISgQEAvQQLQ.ppEDGLFLVRESAR
abl     WVPSNYItpv......NSLEKHS......WYHGPVSRNaAEyLLSsgiN.GSFLVRESES
grb2    FIPkNYI..........eMKpHP......WFFGKIpRaKAEEMLSkQRHDGAFLIRESES
crkl    .......... ssarfDSsDRsA......WYMGPVSRQEAQtRLQgQRH.GMFLVRDSST

matk    hPGDYVLcVSFGrDVIHYRV.lhrDGHLtIdEaVfFcNLmDMVEHYSkdkgaIcTkLVrP
abl     SPGqrSISLRYegRVyHYRINTasDGKLYVssEsRFNTLaELVhHHSTVadgLiTTLhyP
grb2    APGDFSLSVKFGNDVqHFKVlrdgaGK.YFlwvVKFNSLnELVDYHRS......TSV...
crkl    ePGDYVLSVSenSRVsHYiINSlpNrRFkIgDQe.FDHLpaLLEFYK.IhyldtTTLieP

          KRKHgTksaeeelaragWllnlqhLTLgaqIGeGEFGaVlQGeY..lggkVAVKNIKc
        APKRNKPTVYgVS.PNydkWemertdITMkhkLGgGQYGeVyEGvWkkysltVAVKTLKe
        ..sRNQ.qIF.Lr..D.
        APRYpsPpMgsVSaPN.

matk    DVt.aQaFLdEtAVMtKMQHeNLVRLLGVilHQg.LYIVmEhVSkGNLVNFLRtrgRalV
abl     DTmevEeFLkEaAVMkEIKHpNLVQLLGVctREppFYIItEfMTyGNLLDYLRecnRqeV

matk    NtaqLLqFSlHVAegMEYLEsKKLVHRDLAARNiLVsEDlVaKVSDFGLAK...aErkgl
abl     NavvLLyMAtQISsaMEYLEkKNFIHRDLAARNcLVgENhLvKVADFGLSRlmtgDtyta

matk    dS.SRLPVKWTAPEALkHgKFTsKSDVWSFGVLLWEVfSYGrAPYPkMsLkEVsEaVEKg
abl     hAgAKFPIKWTAPESLaYnKFSiKSDVWAFGVLLWEIaTYGmSPYPgIdLsQVyElLEKd

matk    YRMEpPEGCPgpVHvLMsSCWEaePArRPpF.
abl     YRMErPEGCPekVYeLMrACWQwnPSdRPsFaeihqafetmfqessisdevekelgkqgv

abl     rgavstllqapelptktrtsrraaehrdttdvpemphskgqgesdpldhepavspllprk

abl     ergppegglnederllpkdkktnlfsalikkkkktaptppkrsssfremdgqperrgage

abl     eegrdisngalaftpldtadpakspkpsngagvpngalresggsgfrsphlwkksstlts

abl     srlatgeeegggssskrflrscsascvphgakdtewrsvtlprdlqstgrqfdsstfggh

abl     ksekpalprkragenrsdqvtrgtvtppprlvkkneeaadevfkdimesspgssppnltp

abl     kplrrqvtvapasglphkeeaekgsalgtpaaaepvtptskagsgapggtskgpaeesrv

abl     rrhkhssespgrdkgklsrlkpappppaasagkaggkpsqspsqeaageavlgaktkat

abl     slvdavnsdaakpsqpgeglkkpvlpatpkpqsakpsgtpispapvpstlpsassalagd

abl     qpsstafiplistrvslrktrqpperiasgaitkgvvldstealclaisrnseqmashsa

abl     vleagknlytfcvsyvdsiqqmrnkfafreainklennlrelqicpatagsgpaatqdfs

abl     kllssvkeisdivq..............................................
grb2    ................IeqvpqQptYVQALFDFdPqeDgE.LgFRRGDFIhVMDNsDpNWW
crkl    ................LptaedNleYVRTLYDF.PgnDaEdLpFKKGEILvIIEKpEeQWW

grb2    kgach.GQTGMFPrnYVtpVnRN.....................................
crkl    sarnkdGRVGMIPvpYVekLvRS.sphgkhgnrnsnsygipepahayaqpqtttplpavs

crkl    gspgaaitplpstqngpvfakaiqkrvpcaydktalalevgdivkvtrmningqwegevn

matk    ........................rklaeklarelrsagapasvsgqdadgstsprsqe
crkl    grkglfpfthvkifdpqnpden.......................................
```

Construction of Resulting PO-MSA



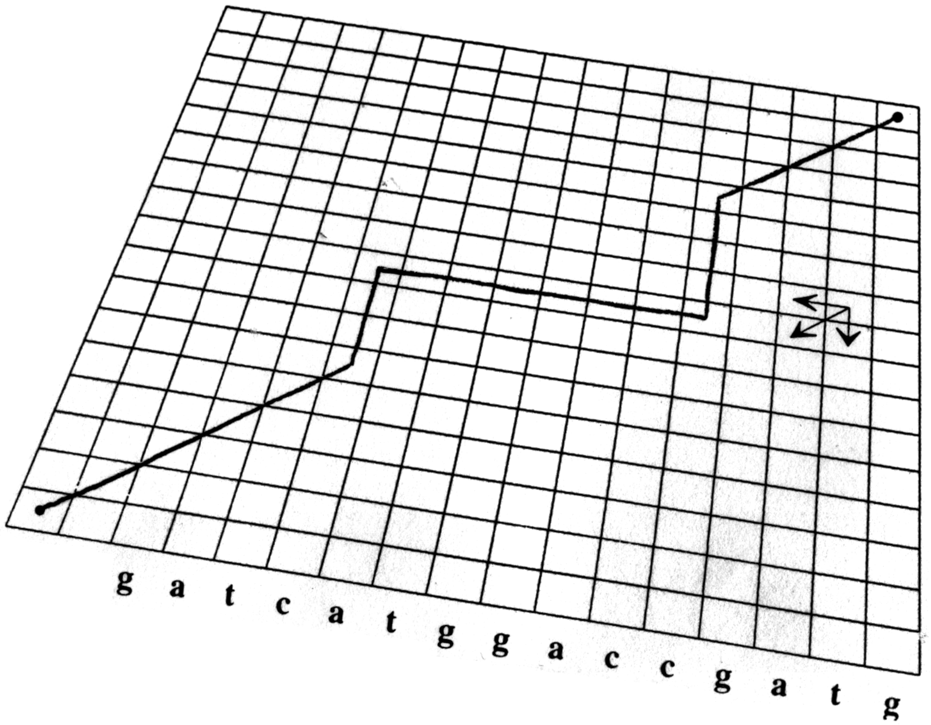
Sequence Alignment Using PO-MSA Representation



$$S(n,m) = \max \begin{cases} S(p,m-1) + s(n,m) \\ S(p,m) + \Delta(m) \\ S(n,m-1) + \Delta(n) \end{cases},$$

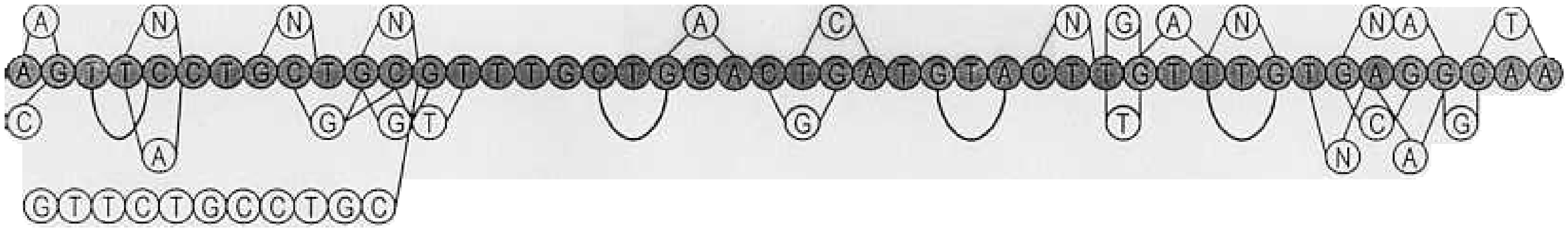
considering all predecessor nodes p that have a directed edge from $p \rightarrow n$.

Sequence Alignment Using RC-MSA Representation



$$S(n \qquad \max \left\{ \begin{array}{l} S(n-1, m-1) + s(n, m \\ S(n-1, m) + \Delta(m) \\ S(n, m-1) + \Delta(n) \end{array} \right.$$

PO-MSA of Human EST Cluster

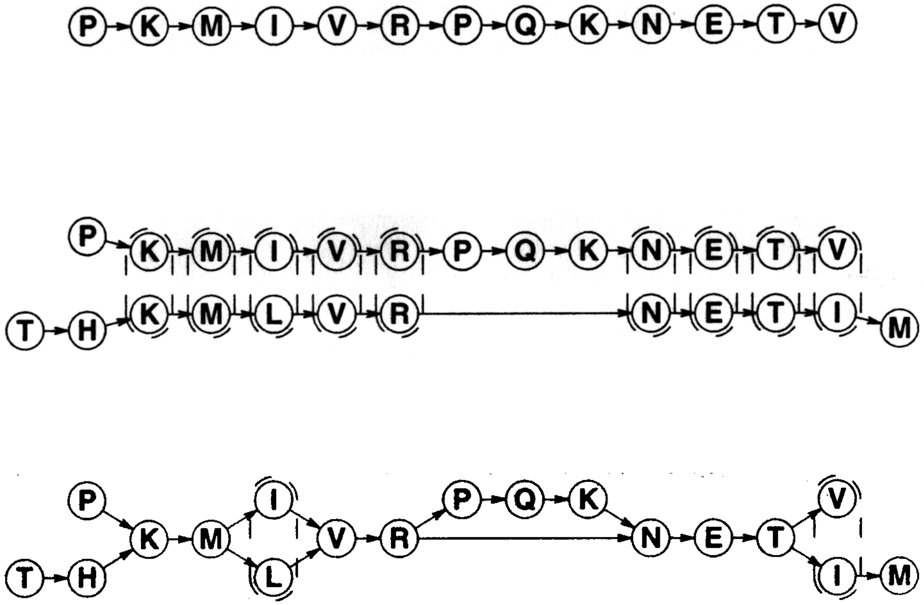


A N N N A C N G A N N A T
A G T T C C T G C T G C G T T T G C T G G A C T G A T G T A C T T G T T T G T G A G G C A A
C G G T G T C G
A N A
G T T C T G C C T G C

```
CONSENS1     ......................................TGTACNT.GTTTGTGAGG.CTA
CONSENS0     A.GTTCCTGCTGC...........GTTTGCTGGACTGATGTACTT.GTTTGTGAGG.CAA
Hs#S663801   A.GTTCCTGCTGC...........GTTTGCTGGACTTATGTACTT.GTTTGTGAGG.CAA
Hs#S337687   AAGTTCCTGCTGC...........GTTTGCTGGACTGATGTACTTGGTTTGTGNAGGCAA
Hs#S629177   A.GTTCCTGCTGC...........GTTTGCTGGACTGATGTACTT.GTTTGTNAGG.CAA
Hs#S672957   A.GTTCCTGCTGC...........GTTTGCT.............................
Hs#S672182   A.GTTCCTGCTGC...........GTTTGCTGGACTGATGTACTT.GTTT..........
Hs#S674099   A.GTTCCTGCTGC...........GTTTGCTGGACTGATGTACTT.GTTTGTGAGG.CAA
Hs#S196113   A.GTTNCTGNTGN...........GTTTGCTGGACTGATGTACTT.GTTTGTGAGG.CAA
Hs#S994400   .......................................GTACNT.GTTTGTGAGG.CTA
Hs#S550772   A.GTTCCTGCTGC...........GTTTGCTGGACTGATGTACTT.GTTTGTGAGG.CAA
Hs#S80460    A.GTTCCTGCTGC...........GTTTGCTGGACTGATGTACTT.GTTTGTGAGG.CAA
Hs#S39701    A.GTTCCTGCTGC...........GTTTGCTGGACTGATGTACTT.GTTTGTGAGG.CAA
Hs#S1988018  A.GTTCCTGCTGC...........TTTTGCTGGACTGATGTACTT.GATTGTGAGG.CAA
Hs#S341915   A.GTTCCTGCTGC...........GTTTGCTGGACTGATGTACTT.GTTTGTGAGG.CAA
Hs#S1794113  A.GTTCCTGCTGC...........GCTTGCTGGACTGATGTACTT.GTTTGTGAGG.CAA
Hs#S4698     A.GTTCCTGCTGC...........GTTTGCTGGACTGATGTACTT.GTTTGTGCGG.CAA
Hs#S813765   A.GT.CCTGCG.C...........GTTTGC.GGACGGATGTACTT.GTT.GTGAGG.CAA
Hs#S1184845  .......................................................G.CAA
Hs#S1577463  ......................................................GG.CAA
Hs#S914987   ..................................CTGATGTACTT.GTT.GTGAGGGCAA
Hs#S1985364  A.GTTCCTGCTGC......     GTTTGCTGGACTGATGTACTT.GTTTGTGAGG.CAA
Hs#S1465644  .............GTTCTGCCTGCGTTTGCTGAACTGATGTACTT.GTTAGT.AAG.CAA
Hs#S1850471  C.GTTACTGCGGG...........GTTTGCTGGACTCATG.ACTTTGTTNGT.AGG.CAA
```

New Data-Structure for MSA: PO-MSA

```
    P K M I V R P Q K N E T V
T H   K M L V R       N E T I M
```


P K M I V R P Q K N E T V
P K M I V R P Q K N E T V
T H K M L V R N E T I M
P T H K M I L V R P Q K N E T V I M

Note:
Sequence indices and residue position indices stored on each node.

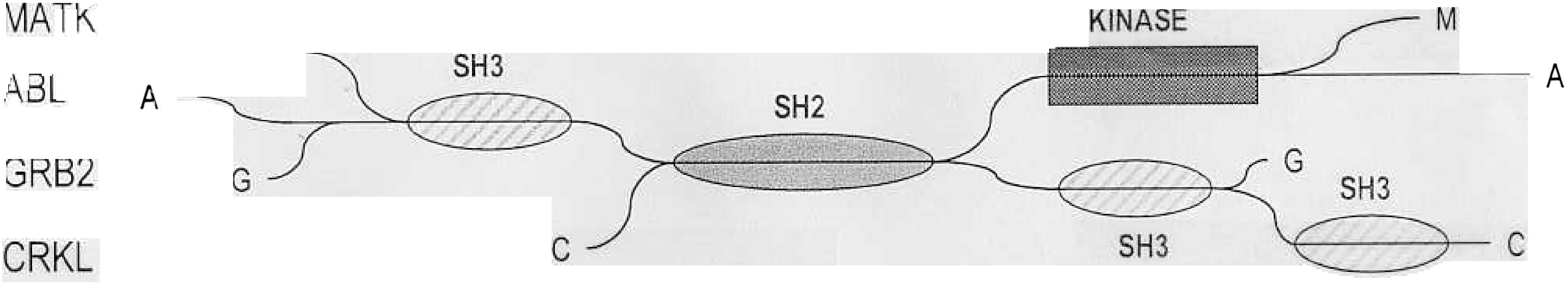
Using PSA on 1D Profiles

- Each column is treated in isolation.
- But interpreting what's a true gap requires looking outside of column.
- We can try to solve this problem by adjusting the scoring process.
- This results in a non-local scoring function, which violates dynamic programming.
- Instead, we can try a new MSA representation.

What do we want to do with our MSA?

- We want to use it as an object in progressive multiple sequence alignment
- We want to analyze it for biologically interesting features

PO-MSA of Four Proteins



Advantages of POA

- Can align up to 5,000 sequences at once
  An order of magnitude speed up from CLUSTALW, T-Coffee, and Phrap
- Can look for biologically nteresting features in a PO-MSA using graph algorithms
- Can be used with T-Coffee BLAST, etc
- Can be used to easily add additiona data nto a sequence alignment.

POA Conclusions

- POA is just now being made publicly available. While it has been very useful in the Lee lab, it is untested by other researchers. It does not yet address issues of local minima and scoring functions. However, it does introduce the possibility of entirely new algorithms for bioinformatics.