



# Multiple Sequence Alignment Construction, Visualization, and Analysis Using Partial Order Graphs

Catherine S. Grasso  
Christopher J. Lee

# Overview of Talk

The title 'Overview of Talk' is positioned at the top left. To its right, there are six circles arranged in a horizontal line. The first circle is solid light purple. The second circle is a light purple outline. The third circle is solid light purple. The fourth circle is a light purple outline. The fifth circle is solid light purple. The sixth circle is a light purple outline.

- Intro to Partial Order Multiple Sequence Alignment Representation
- Multiple Sequence Alignment Construction Using Partial Order Graphs
- Conclusions



Q: Why Do Multiple Sequence Alignment?

A: To model the process which constructed a set of sequences from a common source sequence.

# A multiple sequence alignment allows biologists to infer:

- Protein Structure
- Protein Function
- Protein Domains
- Protein Active Sites
- Splice Sites
- Regulatory Motifs
- Single Nucleotide Polymorphisms
- mRNA Isoforms

For example, protein sequences that are >30% identical often have the same structure and function.

# Row Column Multiple Sequence Alignment

## RC-MSA

```
CONSENS1      .....TGTACNT.GTTTGTGAGG.CTA
CONSENS0      A.GTTCCTGC.TGCGTTTGCTGGACTGATGACTT.GTTTGTGAGG.CAA
Hs#S663801    A.GTTCCTGC.TGCGTTTGCTGGACTTATGACTT.GTTTGTGAGG.CAA
Hs#S337687    AAGTTCCTGC.TGCGTTTGCTGGACTGATGACTTGGTTTGTGNAGGCAA
Hs#S629177    A.GTTCCTGC.TGCGTTTGCTGGACTGATGACTT.GTTTGTNAGG.CAA
Hs#S672957    A.GTTCCTGC.TGCGTTTGCT.....
Hs#S672182    A.GTTCCTGC.TGCGTTTGCTGGACTGATGACTT.GTTT.....
Hs#S674099    A.GTTCCTGC.TGCGTTTGCTGGACTGATGACTT.GTTTGTGAGG.CAA
Hs#S196113    A.GTTNCTGN TGN GTTTTGCTGGACTGATGACTT.GTTTGTGAGG.CAA
Hs#S994400    .....GTACNT.GTTTGTGAGG.CTA
Hs#S550772    A.GTTCCTGC.TGCGTTTGCTGGACTGATGACTT.GTTTGTGAGG.CAA
Hs#S80460     A.GTTCCTGC.TGCGTTTGCTGGACTGATGACTT.GTTTGTGAGG.CAA
Hs#S39701     A.GTTCCTGC.TGCGTTTGCTGGACTGATGACTT.GTTTGTGAGG.CAA
Hs#S1988018   A.GTTCCTGC.TGCTTTTGCTGGACTGATGACTT.GATTGTGAGG.CAA
Hs#S341915    A.GTTCCTGC.TGCGTTTGCTGGACTGATGACTT.GTTTGTGAGG.CAA
Hs#S1794113   A.GTTCCTGC.TGCGCTTGCTGGACTGATGACTT.GTTTGTGAGG.CAA
Hs#S4698      A.GTTCCTGC.TGCGTTTGCTGGACTGATGACTT.GTTTGTGCGG.CAA
Hs#S813765    A.GT.CCTGC.GCGTTTGC.GGACCGATGACTT.GTT.GTGAGG.CAA
Hs#S1184845   .....G.CAA
Hs#S1577463   .....GG.CAA
Hs#S914987    .....CTGATGACTT.GTT.GTGAGGCAA
Hs#S1985364   A.GTTCCTGC.TGCGTTTGCTGGACTGATGACTT.GTTTGTGAGG.CAA
Hs#S1465644   ..GTTT.TGCGCTGCGTTTGCTGAAGTACTGATGACTT.GTTAGT.AAG.CAA
Hs#S1850471   C.GTTACTGC.CCGTTTGCTGGACTCATG.ACTTGTNGT.AGG.CAA
```

# RC-MSA Representation Does Not Reveal Large Scale Features

While it is easy to interpret single residue changes in this format,

Large scale changes are not easy to interpret.

```
abl mleiclk lvgck skkgl sssss cylee alq rpvas dfepq glsea arwns kenll agpse
matk ..... magrgsl vswra fhgcd saeel prvsp rfl rawhp ppvs
abl hdpnlf...VALYDFvASGDN.....
grb2 ..... eeaIAkYDFkATADD.....
matk armpt rwapg tqcit kcehtrpk pgELAFRKGdVvTIL.EaCEn KSWYRvKhh tSQQEG
abl ..... tLSitkGEkLRVLgynhn. geWceAQtk .NGQ.G
grb2 ..... ELSFKRGDILKVLnEeCD. QNWKAEI. .NGKDG
matk LLaAga Lrer. .... EALs adPk lslmp WPHGKISgQE AvQQLQ.ppED GLFLVRESAR
abl WPSNY Itpv. .... NSLEKHS. .... WYHGpVSRNa AEyLLSsgin. GSFLVRES
grb2 FIPkNY I..... eMKp HP..... WFFGKI pRaK AEEMLSkQRHD GAFLIRES
crkl ..... mssarfD SsDRs A..... WYMGpVSRQE AQtRLQgQRH. GMFLVRDSS
matk hPGDYVLcV SFGrD ViHYRV.lh rDGHl tIdEa VffCnLmDMVEH YSkdk gaIcTklVrP
abl SPGqrSISL RYegR VyHYRINTa sDGKL YVssE sRFNTLaELVhH HSTVa dgLiTTLhyP
grb2 APGDFSLSV KFGND VgHFkVlrd gaGK. Yflwv VKFNSLnELVDY HRS... .i.TSV...
crkl ePGDYVLSV SensR VsHYiINSL pNRRF kIgdQ e.FDHLpaLLEF YK.Ih vldtTLLieP
matk ..KRKHgT ksaeel aragwllnlgh tLgaaqIGeGEFGaVlQGeY..lg qkVAVKNIkC
abl APKRnkPT VYgVS .PNyd kWemer tD TMkhhLGGgQYGeVyEGvWkkys ltVAVKTLkE
grb2 ..sRNQ .qIF. Lr. .D.
crkl APRYpsPp MgsVS aPN.
matk DVt.aQa FLdEtAVMtkMQHeNLVRLlGVilHQg. LYIVMEHVSkGNLVNfLrt rgralV
abl DtmEvEe FLkEaAVMkEIKHpNLVQLLGVctREpp FYIItE fMTyGNLLDYlRecnRqEV
matk NtaqLLqFSLHV AegMEYLESKKLVHRDLAA RNiLVsEDlVaKVSDfGLAK. .a ErkgI
abl NavvLLyMaTqL SsaMEYLEkKNfIHRDLAA RNcLVgEnhLVKVADfGLSRlmtg Dtyta
matk ds.SRL PVKWTAPe AlkHgKFTsKSDVWSFGVLLW EVfSYgrAPYpKmsLkEvsEaVEKkG
abl HAgAKF PIKWTAPe SLaYnKFSiKSDVWAFGVLLW EiaTYGmSPYPgIdLsQVYElLEKd
matk YRMEpPEGCPgpVHVLMsSCWEaePARPpP.....
abl YRMErPEGCPeKVYeLmRACWQwnFSdRPsFaeihqa fctmfqe ssisd eveke lgkqg v
abl rgavst llqap elptk trtsrr aaehr dtttdv pemph skgqg esd pldhe pavsp llprk
abl ergppe gglne derll pkdkk tnlfS ali kkkkk taapt pkrss sfrem dgqpe rrag e
abl eegrdis ngal aftpl dtadp akspk psn gagvp ngalr esggs gfrsp hlwkk sstlt s
abl srlatg eeeegg gsssk rflrs csasc vph gakdt ewrsv tIprd lqstg rqfds stfgg h
abl ksekpa lprkr agenrsd qvtrg tvtppprlv kknee aadev fkdime ssp pnltp
abl kplrrqvtvap asglp hkeea ekgsa lgt paaae pvtpt skags gapgg tskgp aeer v
abl rrrhks sespg rdkgk lsrlk papppp ppa asagk aggkp sqsps qeaag eavlg aktka t
abl slvdav nsdaa kpsqp geglk kvlp atp kqsa kpsgt pisp pvpst lpsas salag d
abl qpssta fipli strvs lrktr qpperiasg aitkg vvlDs tealc laisr nse qwash sa
abl vleagk nlytf cvsyv dsiqq mrnkf afr eaink lennl relqi cpata gsgpa atgdf s
abl kllssv keisd ivgr.....
grb2 .....I eqvpgQptYVQALFDfdPgeDgE.LgFRRGDFiHVM DNsDp NWW
crkl .....Lpt aedNleYVRTLYDF.PgnDaEdLpFKKGEILvII EKpEe QWW
grb2 kgach. GQTGMFPrnYVtpVnRNv.....
crkl sarnkd GRVGMIPvpYVEkLVRs.sphg khgnr nsnsy gipep ahaya qpq tttpl pavS
crkl gspgaa itplp stqhgpvfa kaiqk rvpca ydkta lalev gdi kvvtr mning qwege vn
matk ..... fklaklarel rsagap asvsg qdadg stspr sqep
crkl grkglf pfthv kifdp qnpende.....
```

# The Scale of Features of Interest Should Inform MSA Representation

- Features from single residue changes can be easily seen in RC-MSA Representation:
  - **Regulatory Motifs**
  - **Single Nucleotide Polymorphisms**
  - **Promoter Binding Sites**
- Features from large scale changes cannot:
  - **Protein Domain Differences**
  - **Alternative Splicing**
  - **Genome Duplications**

# Degeneracy of RC-MSA Representation

Alignment A is biologically equivalent to alignment A'.

**A:**   .....ACATGTCGAT.....AGGTG  
TGCAC.....TCGATACATAAGGTG

**A':**   ACATG.....TCGAT.....AGGTG  
.....TGCACTCGATACATAAGGTG

However, they look different solely due to representation degeneracy.

We'd like a representation that is not degenerate.





## 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.

One sequence can impose its order on another sequence only through alignment.

## What do we really want to do with an MSA?

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

# Partial Order Multiple Sequence Alignment PO-MSA

Conventional Format  
*(RC-MSA)*

A

```

. . 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
    
```

Draw each sequence  
as a directed graph:  
node for each letter,  
connect by directed  
edges

B

```

graph LR
  P1((P)) --> K1((K))
  K1 --> M1((M))
  M1 --> I1((I))
  I1 --> V1((V))
  V1 --> R1((R))
  R1 --> P2((P))
  P2 --> Q1((Q))
  Q1 --> K2((K))
  K2 --> N1((N))
  N1 --> E1((E))
  E1 --> T1((T))
  T1 --> V2((V))
    
```

C

```

graph LR
  T2((T)) --> H2((H))
  H2 --> K3((K))
  K3 --> M2((M))
  M2 --> L2((L))
  L2 --> V2((V))
  V2 --> R2((R))
  R2 --> N2((N))
  N2 --> E2((E))
  E2 --> T2_2((T))
  T2_2 --> I2((I))
  I2 --> M2_2((M))
    
```

Fuse aligned, identical  
letters *(PO-MSA)*

D

```

graph LR
  P1((P)) --> K1((K))
  K1 --> M1((M))
  M1 --> I1((I))
  I1 --> V1((V))
  V1 --> R1((R))
  R1 --> P2((P))
  P2 --> Q1((Q))
  Q1 --> K2((K))
  K2 --> N1((N))
  N1 --> E1((E))
  E1 --> T1((T))
  T1 --> V2((V))
  T2((T)) --> H2((H))
  H2 --> K3((K))
  K3 --> M2((M))
  M2 --> L2((L))
  L2 --> V2((V))
  V2 --> R2((R))
  R2 --> N2((N))
  N2 --> E2((E))
  E2 --> T2_2((T))
  T2_2 --> I2((I))
  I2 --> M2_2((M))
    
```

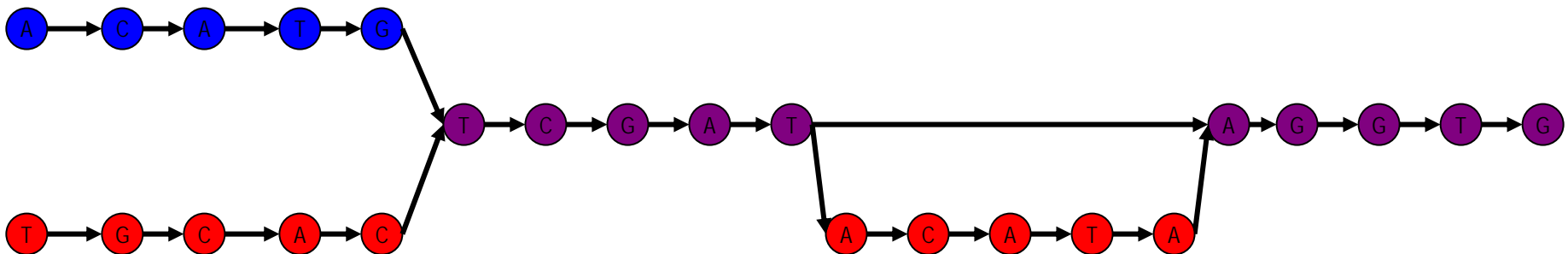
Returning to the previous example...

In the PO-MSA format, both *A* and *A'*

*A*: .....ACATGTCGAT.....AGGTG  
TGCAC.....TCGATACATAAGGTG

*A'*: ACATG.....TCGAT.....AGGTG  
.....TGCACTCGATACATAAGGTG

Can be represented as



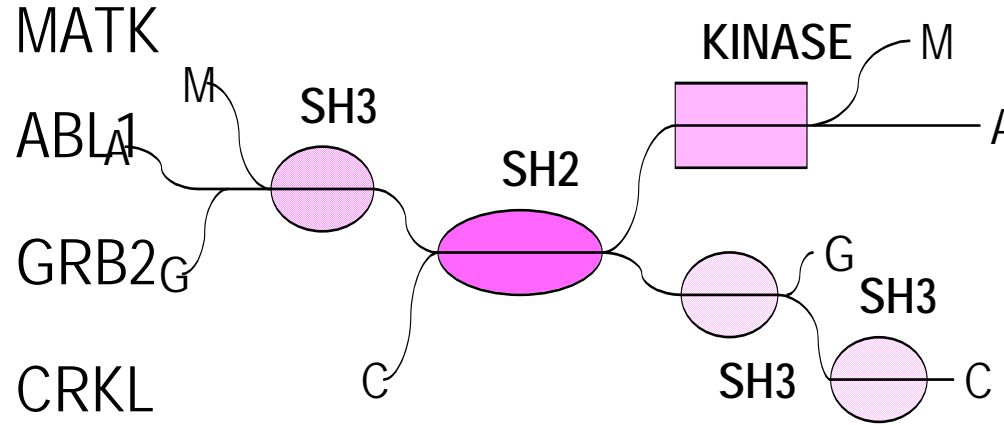
# Real Example: Human SH2 Domain Containing Proteins

```

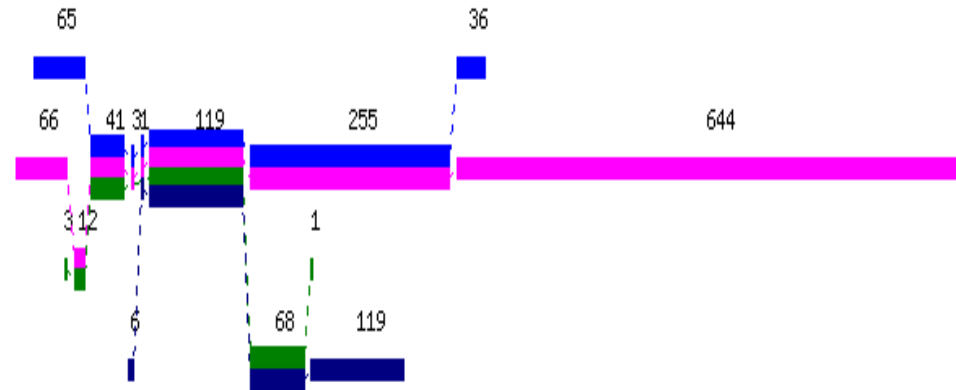
abl  mLeiclkIvlgckskkglssssscyleaalqrpvdsfepqglseearwnskennllagpse
matk  .....magrgslsvsrafhngcdsaeelprvsprflrawhpppvs
abl  hdpnlf...VALYDFVASGDM.....
grb2  .....heaIAkYDFkATADD.....
matk  armptrrwapggtqcitkcehtrpkpgELAFRKGDVVtIL.EaCENkSWYRvKhhtSGQEG
abl  .....tLSitKGEKLRVLgynhn.gwCEAQtk.NGO.G
grb2  .....ELSFKRGDILKVLnEeCD.QNWYKAEI..NGKDG
matk  LLaAgaLrer.....EALsaaPklslmpWFHGKISgQEAvQQLQ.ppEDGLFLVRESAR
abl  WVPsNYItpv.....NSLEKHS.....WYHGpVSRNaAByLLSsain.GSFLVRESSES
grb2  FIPkNYI.....eMKpHP.....WFFGKIprRaKAEEMLSkQRHDGAFLIRESES
crkl  .....mssarfDsSDRSa.....WYMGpVSRQEAtRLQQRH.GMFLVRDSSST
matk  hPGDYVLeVSPGrDViHYRV.lhrDGHLtIdeAVfEeNLMdMVEHYSkdkgaIeTklLvrP
abl  SPGqrSISLRYegRVyHYRINTasDGKLYVssEsRFNTLaELVhHHSTVadgLiTTLhyP
grb2  APGDFLSLVKFGNDVgHFkVlrdgaGK.YFlwvKFNLSLnlELVDYHRS....TSV...
crkl  ePGDYVLSVSenSRVsHYiINSLpNrRfkiGdQe.FDHLpaLLEfYK.IhyldtTTLIEP
matk  ..KRKHgTksaeelaragWllnlqhLTLgaIQeGEPGaVlQGeY..lgqkVAVKNIKc
abl  APKRnkFTVYgVS.PNydkWemertdFTMkhkLgGQYGeVYEGVWkysltVAVKTLKE
grb2  ..sRNO.qIF.Lr..D.....
crkl  APRYpsPpMgsVsaPN.....
matk  DVT.aQaFLdeAVMtKMOHeNLVRLlGVilHqG.LYIVmEhVSkGNLVNfLrtRgRalV
abl  DTmevBeFLkEaAVMkEIKHpNLVQLLGVctREppFYIItBfMTYGNLLDYLReCnRgeV
matk  NtaqLLqFSlHVAegMEYLEsKKLVHRDLAARNiLVsEDLVaKVSDfGLAK...aErkgI
abl  NavvLLyMatQISsaMEYLEkKNFIHRDLAARNcLVgEnHlVkvADfGLSRlmtgDtya
matk  GS.SRLPVKWTAPeALkHgKFTeKSDVWAFGVLLWEVfSGrAPYpkMsLkEvsEaVEKq
abl  hAgAKFPiKWTAPESLaYnKFSiKSDVWAfGVLLWEaTYGMSPYPgIdLsQVYElLEKq
matk  YRMEpPEGCPgpVHvLmsSCWEaePArRppF.....
abl  YRMErPEGCPekVYeLmRACWQwnPsdRPsFaeihqafetmfqessisdevekelgkqv
abl  rgavstllqapelptktrtsrraaehrtdtdvpemphskgggesdpldhepavsp1lprk
abl  ergppegglnederllpkdkktnlfsalikkkkktaptppkrsssfiremdgqpergagae
abl  eegrdisngalafptldtadpakspkpsngagvpngalresggsgfrsphlwkssstlts
abl  srlatgeeeeggssskrflrscsascvphgakdtewsvtlprdlqstgrqfidsstfggh
abl  ksekalprkragenrsdqvtrgtvtppprlvkkneeaadevfkdimesspsppnltp
abl  kplrrqvtvapasglphkeeaekgsalgtpaaapevptptskagsgagpgtskqpaeesrv
abl  rrrkhsesepgrdkglslr1kpapppppaasagkaggkpsqpsqeaageavlgaktkat
abl  slvdavnsdaakpsqpggglkpvlpatpkpqsakpgtppisapvpstlpsasalagd
abl  qpsstafiplistrvslrktrqpperiasgaitkgvldstealc1aisrnseqmasha
abl  vleagknlytfcvsyvdsiqqmrnkfafreainkennlrelqicpatagsgpaatqdfs
abl  kllssvkeisdivqr.....
grb2  .....IeqvpgOptYVQALFDfDPeDgE.LgFRRGDFIhVMDNsDpNWW
crkl  .....LptaedNleYVRTLYDF.PgnDaEdLpFKKGEILvII EKpEeQWW
grb2  kgach.GQTGMFPrnYVtPvNRN.....
crkl  sarnkdGRVGMIPvpYVekLVRs.sphghgnrnsnsygipepahayaqqttt1plpavs
crkl  gspgaa1t1p1stqhgvpfakaiqkrvpcaydktalalevgd1kvtrmningqwegevn
matk  .....k1laeklare1rsagapasvsgdqadgstsprsqep
crkl  grkg1pfthvkifdpqnpdene.....

```

RC-MSA in Text Format



Hand Rendered PO-MSA Showing Domain Structure



POAVIZ Rendered PO-MSA Reflects Domain Structure



What do we really want to do with an MSA?

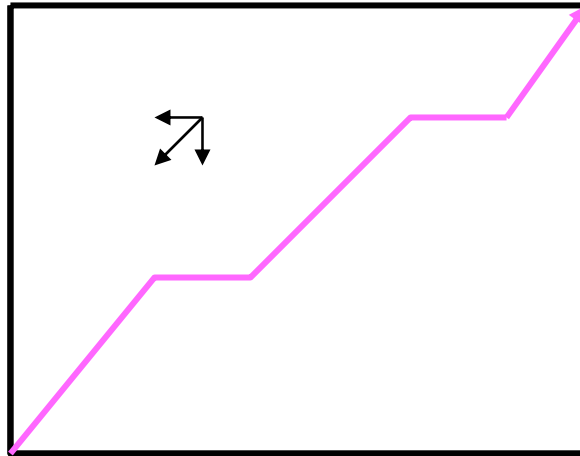
We want to analyze it for biologically interesting features.

We want to use it as an object for building multiple sequence alignments.



# Multiple Sequence Alignment Construction Using PO-MSAs

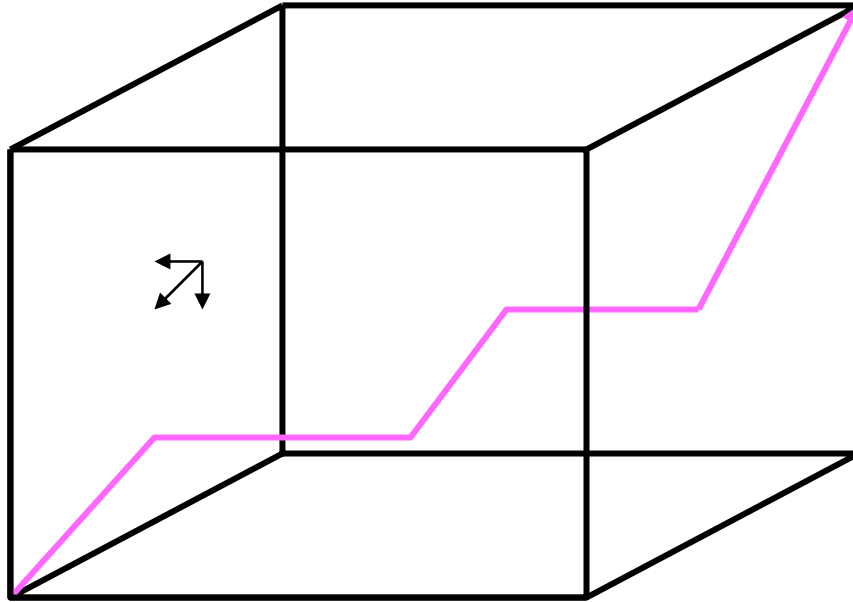
# Pair-wise Sequence Alignment Using Dynamic Programming



Finding a PSA = Finding a path through a 2-Dim matrix.

It's  $O(L^2)$ , where  $L$  is the sequence length.

# Multiple Sequence Alignment Using Dynamic Programming



Finding an MSA = Finding a path through an N-Dim matrix. It's  $O(L^N)$ , where N is the number of sequences and L is the sequence length.

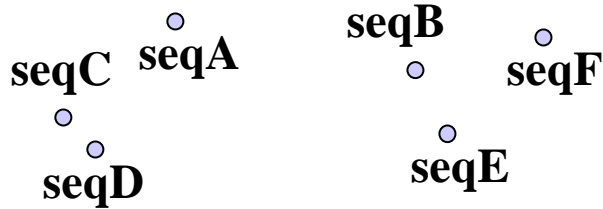
Note: More than 5 sequences takes a prohibitive amount of time.

Heuristic methods, such as those used by CLUSTAL W, are used instead.

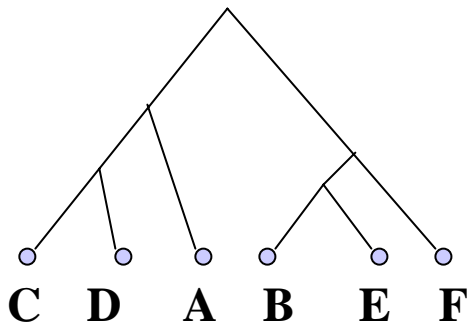


# Progressive Alignment (CLUSTAL W) Approach

1. Compute pairwise distances of all N sequences.

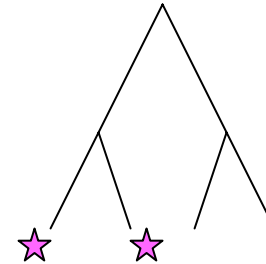


2. Build Guide Tree



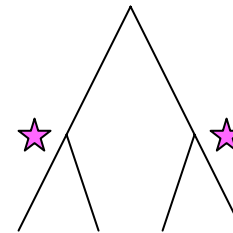
3. Align N sequences using guide tree.

a. Use standard PSA to align leaf sequence.



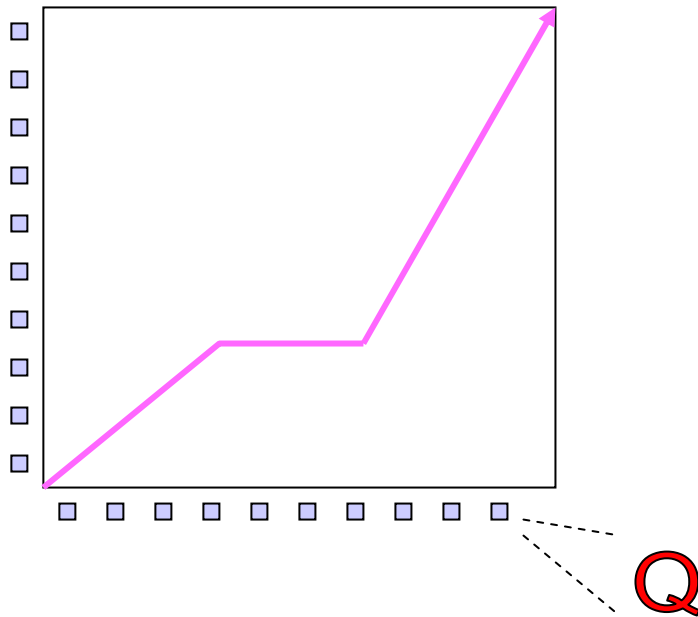
b. Profile multiple sequence alignments at branch nodes.

c. Use standard PSA on profiles.

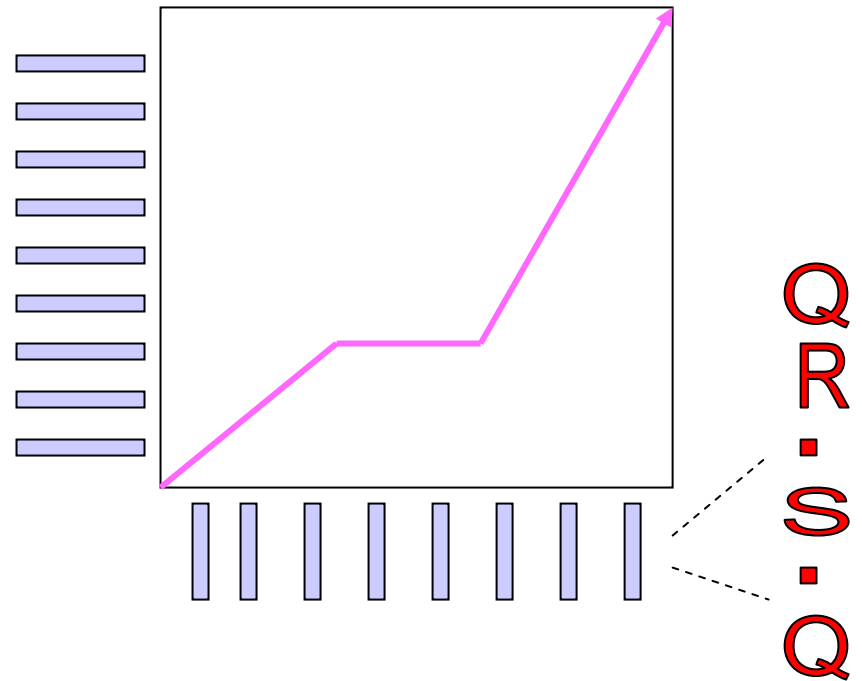


d. Recurse.

# Pair-wise Sequence Alignment of Leaf Nodes V. Branch Nodes



- PSA of sequences at leaf nodes:  
Requires a scoring function which can score a match between residues.



- PSA of profiles at branch nodes:  
Requires a scoring function which can score a match between profiles of columns of residues and gaps.

# Problem with Aligning Profiles: Gap Artifacts!

Alignment  $A$  is biologically equivalent to alignment  $A'$ .

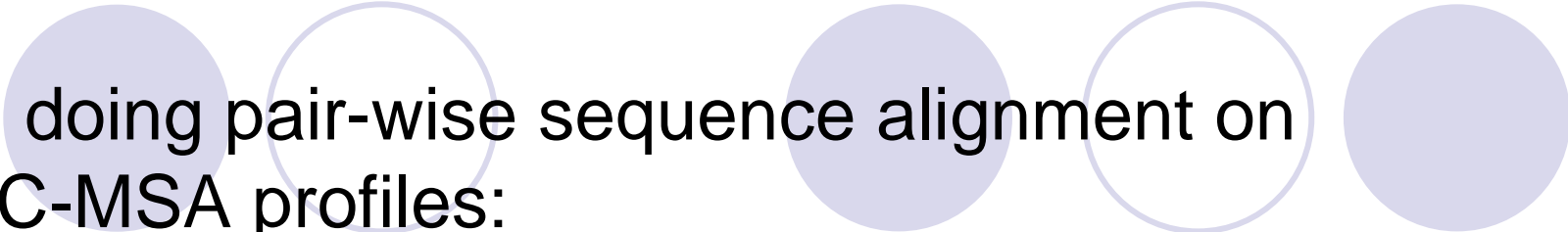
$A$ :   .....ACATGTCGAT.....AGGTG  
TGCAC.....TCGATACATAAGGTG

$A'$ :   ACATG.....TCGAT.....AGGTG  
.....TGCACTCGATACATAAGGTG

If we try to align another sequence which is identical to the second sequence in the alignment...

$S$ :   TGCACTCGATACATAAGGTG

We find that  $\text{Score}(S,A)$  not equal to  $\text{Score}(S,A')$ , but it should be.

The slide features five decorative circles at the top. From left to right: a solid light purple circle, a hollow light purple circle, a solid light purple circle, a hollow light purple circle, and a solid light purple circle. The text 'In doing pair-wise sequence alignment on RC-MSA profiles:' is positioned below the first two circles and above the last three circles.

# In doing pair-wise sequence alignment on RC-MSA 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.

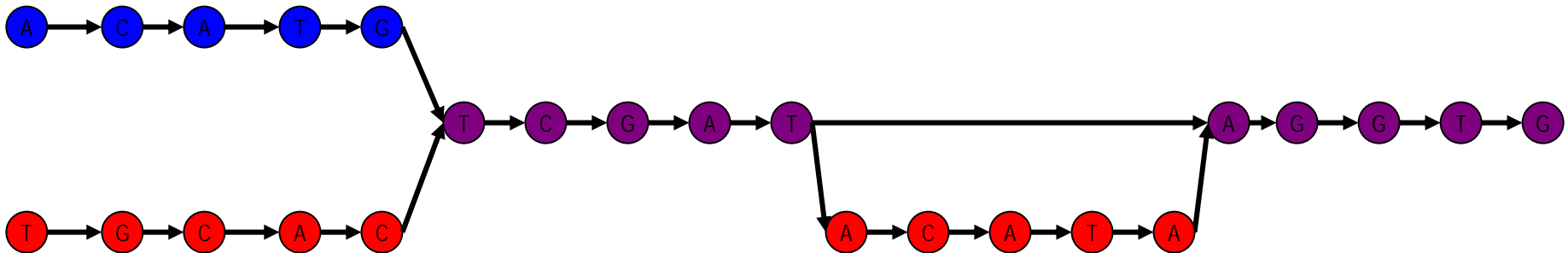
We can instead replace the profile RC-MSA representation with the PO-MSA representation.

In the PO-MSA representation, both  $A$  and  $A'$

$A$ : .....ACATGTCGAT.....AGGTG  
TGCAC.....TCGATACATAAGGTG

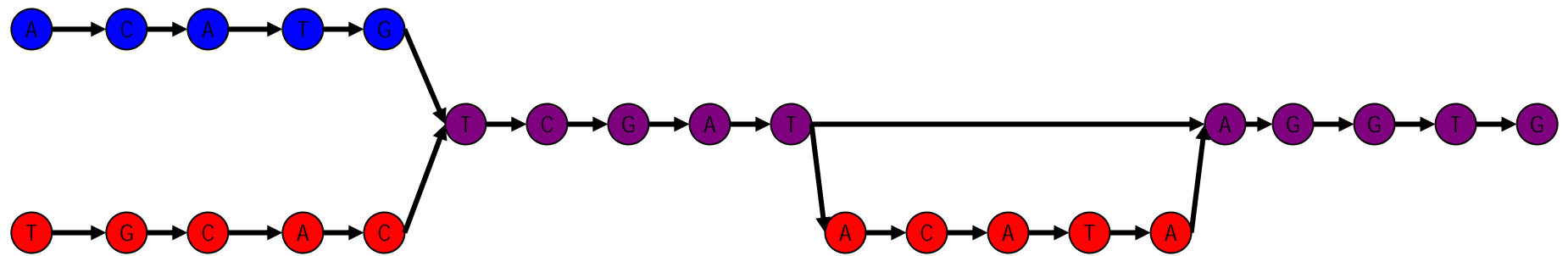
$A'$ : ACATG.....TCGAT.....AGGTG  
.....TGCACTCGATACATAAGGTG

Can be represented as

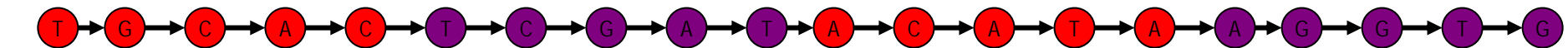


We can align  $S$  to  $A$  using Sequence to PO-MSA alignment algorithm.

**A:**

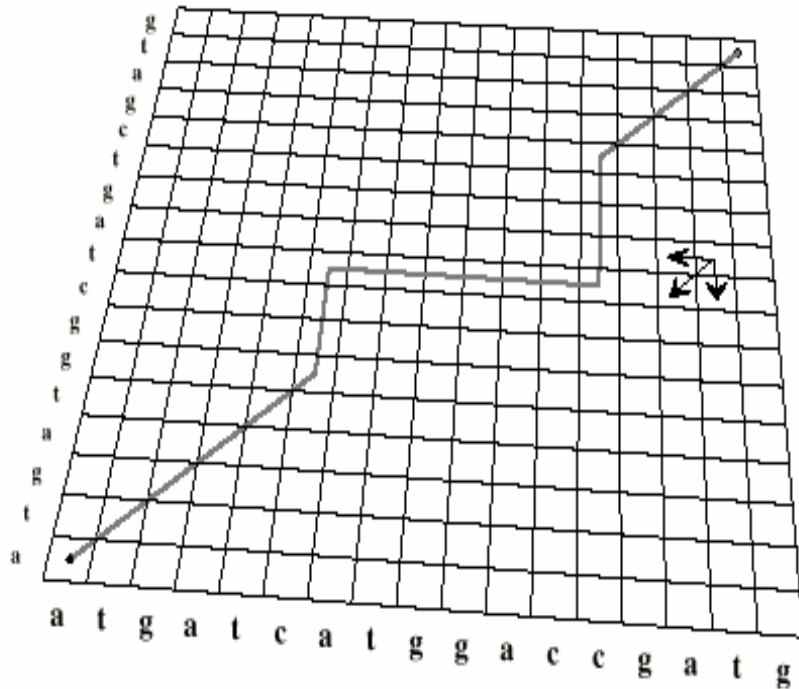


**S:**

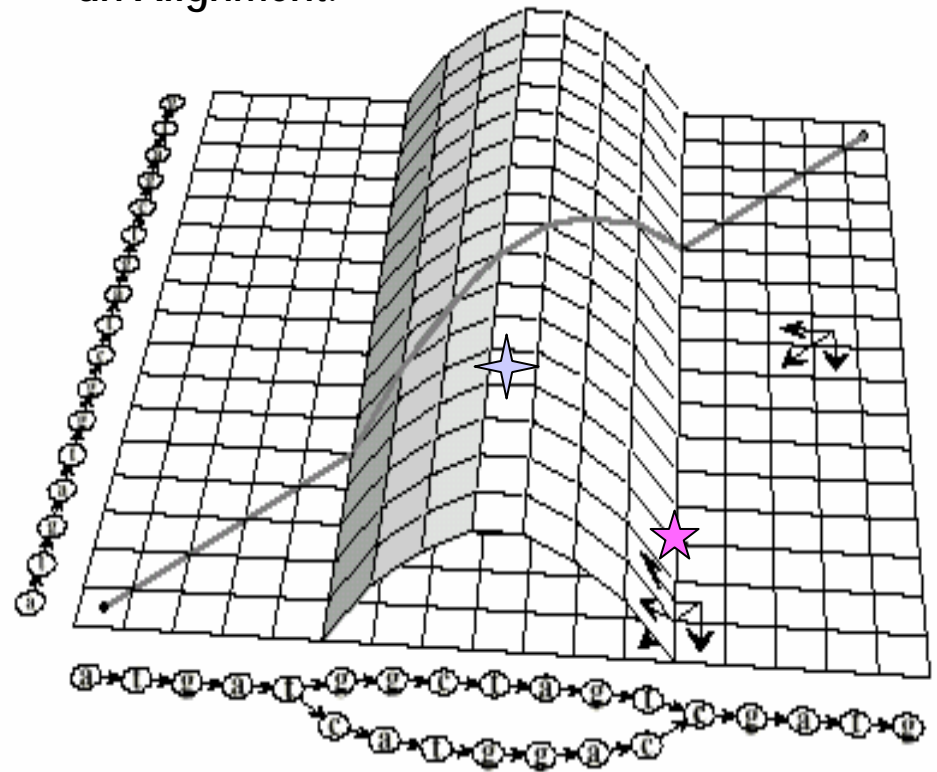


# Sequence to PO-MSA Alignment Algorithm

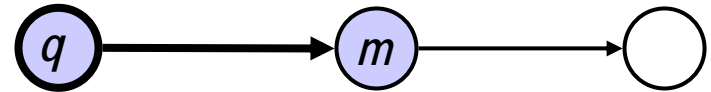
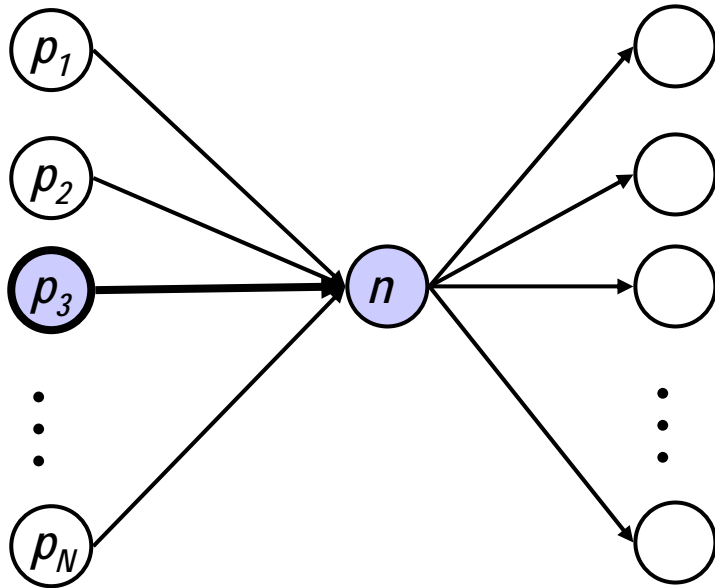
Conventional Alignment of Two Sequences



Partial Order Alignment of a Sequence to an Alignment.



# Sequence to PO-MSA Alignment Algorithm Requires a Simple Extension of Sequence to Sequence Alignment Algorithm



Simply extend dynamic programming move set to include partial order moves: at each position  $(n,m)$  in the matrix, choose best move by:

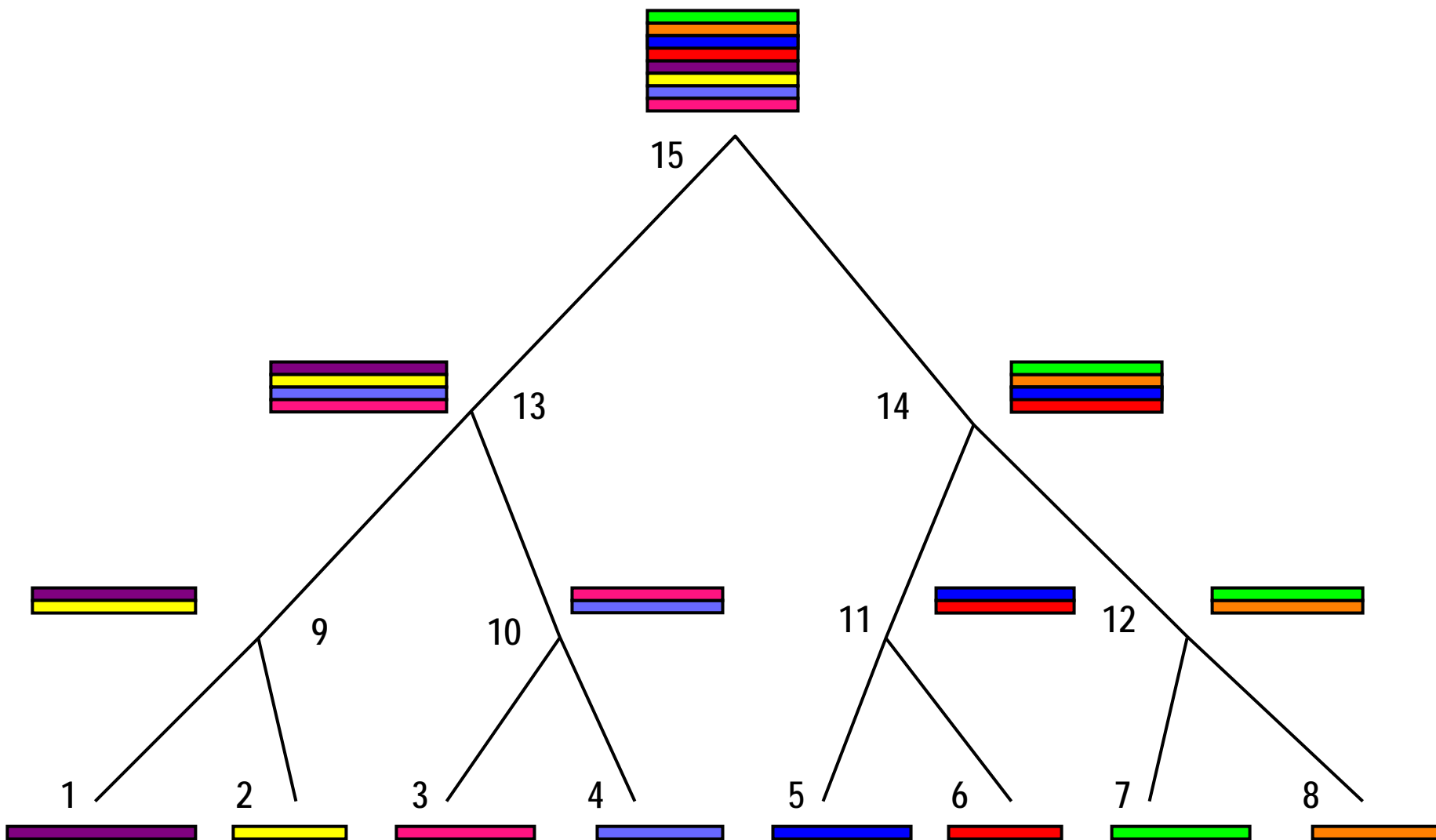
$$S(n, m) = \max \begin{cases} S(p, m - 1) + s(n, m) \\ S(p, m) + \Delta(m) \\ S(n, m - 1) + \Delta(n) \end{cases}$$

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

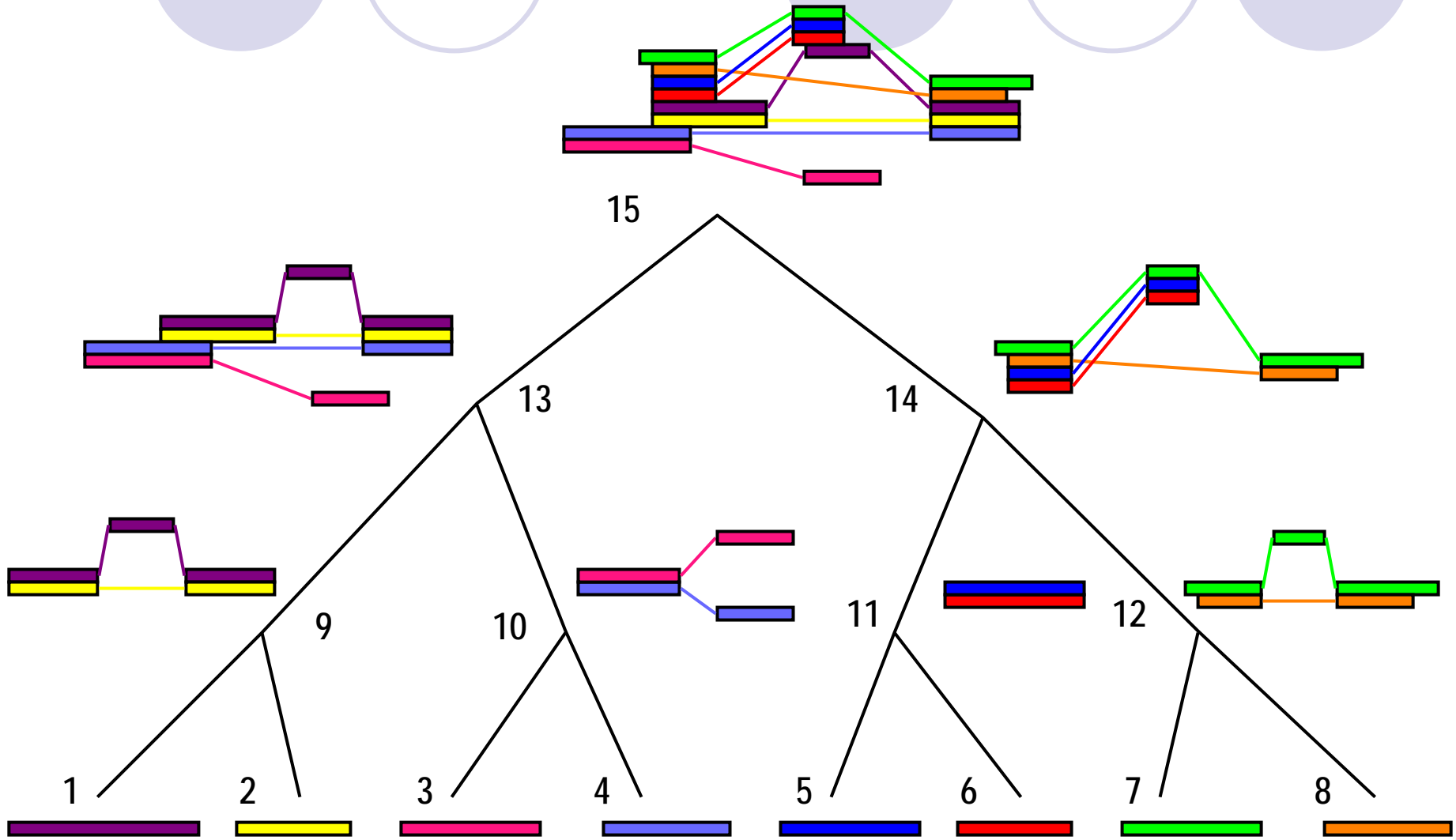
Note: MATCH and INSERT moves may have **more than one** incoming edge  $p$ .



# Recall Progressive Multiple Sequence Alignment with Profile Intermediates

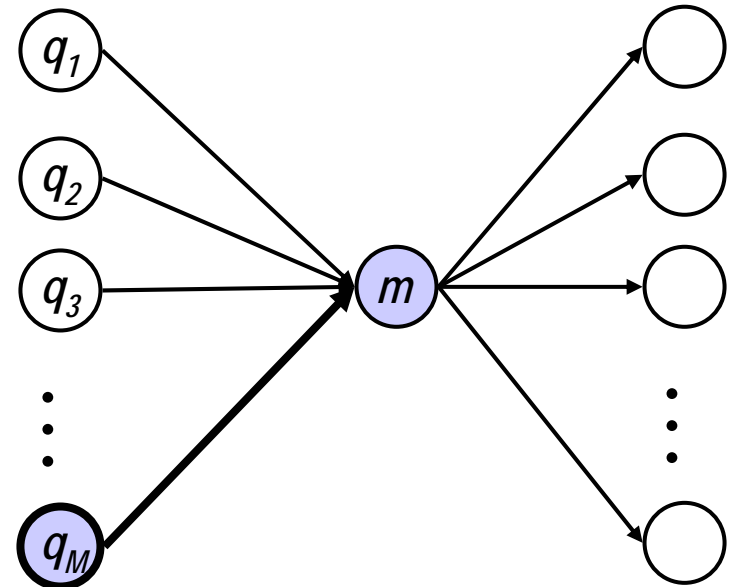
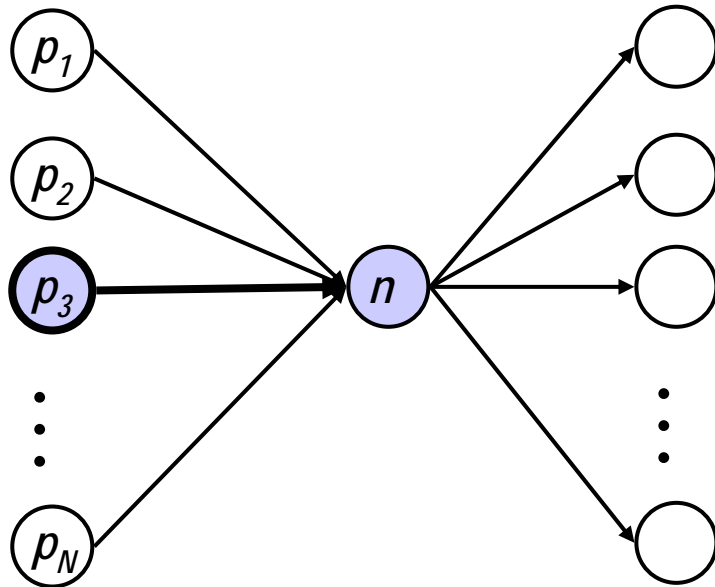


# Progressive Multiple Sequence Alignment with PO-MSA Intermediates



Requires PO-MSA to PO-MSA Alignment Algorithm

# PO-MSA to PO-MSA Alignment Algorithm Requires a Simple Extension of Sequence to PO-MSA Alignment Algorithm



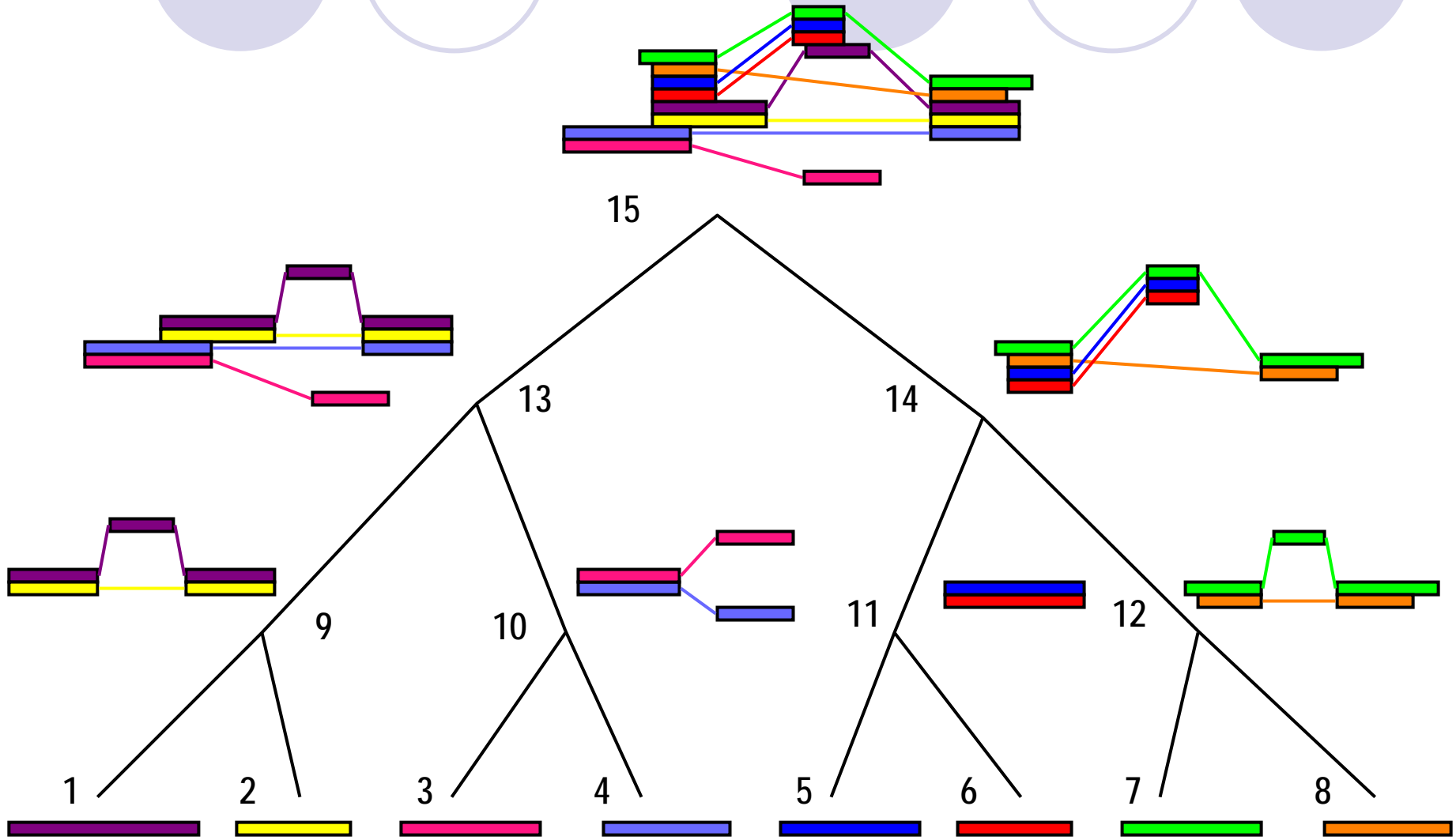
Simply extend dynamic programming move set to include partial order moves: at each position  $(n,m)$  in the matrix, choose best move by:

$$S(n, m) = \max_{\substack{p \rightarrow n \\ q \rightarrow m}} \begin{cases} S(p, q) + s(n, m) \\ S(p, m) + \Delta(n) \\ S(n, q) + \Delta(m) \end{cases}$$

Considering all predecessor nodes that have a directed edge from  $p \rightarrow n$  and  $q \rightarrow m$ .

Note: MATCH and INSERT moves may have **more than one** incoming edge  $p$  or  $q$ .

# PO-MSA to PO-MSA Alignment Algorithm Finds Best Linear Match Between the PO-MSAs



Can be extended heuristically to find best match

# Thesis Work



- Developed partial order alignment visualizer
- Combined partial order alignment and progressive alignment
- Applied POA to detect alternative splicing events in expressed sequence data
- Formalized relationship between PO-MSAs and HMMs

# Acknowledgements



- I'd like to thank:
  - Chris Lee for all of his guidance and support.
  - Michael Quist for all of his help with this project.
  - Barmak Modrek with whom I worked on annotating alternative splicing using PO-MSAs and POA.
  - Everyone in the Lee Lab for hours of helpful discussion.
  - DOE CSGF for supporting the work.

To use or download POA or POAVIZ go to:

<http://www.bioinformatics.ucla.edu/poa>