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

Catherine S. Grasso
Christopher J. Lee
Overview of Talk

- 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.
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.
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’.

\[
\begin{array}{c}
A: \\
\text{ACATGTCGAT} \quad \text{AGGTG} \\
\text{TGCAC} \quad \text{TCGATACATAAGGTG}
\end{array}
\quad \begin{array}{c}
A’: \\
\text{ACATG} \quad \text{TCGAT} \quad \text{AGGTG} \\
\text{TGCACTCGATACATAAGGTG}
\end{array}
\]

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)*


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

B

C

Fuse aligned, identical letters

*(PO-MSA)*

D
Returning to the previous example...

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

\[ A: \quad \ldots \text{ACATGTCGAT} \ldots \text{AGGTG} \quad TGCAC \ldots TCGATACATAAGGTG \]

\[ A': \quad \ldots \text{ACATGTCGAT} \ldots \text{AGGTG} \quad \text{TCGACTCGATACATAAGGTG} \]

Can be represented as
Real Example: Human SH2 Domain Containing Proteins

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

   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 \( A' \).

\[
\begin{align*}
A: & \quad \ldots ACATGTCGAT \ldots AGGTG & A': & \quad ACATG \ldots TCGAT \ldots AGGTG \\
S: & \quad TGCACTCGATACATAAGGTG
\end{align*}
\]

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

\[
S: \quad TGCACTCGATACATAAGGTG
\]

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

$A'$: TGCAC......TCGATACATAAGGTG

Can be represented as
We can align S to A using Sequence to PO-MSA alignment algorithm.
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:

\[
S(n, m) = \max \left\{ S(p, m - 1) + s(n, m), \quad S(p, m) + \Delta(m), \quad S(n, m - 1) + \Delta(n) \right\}
\]

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:

\[
S(n,m) = \max_{p \rightarrow n, q \rightarrow m} \left\{ S(p,q) + s(n,m), S(p,m) + \Delta(n), S(n,q) + \Delta(m) \right\}
\]

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