Lectures 22: Multiple Sequence Alignment

Spring 2017
April 27, 2017
Multiple Alignment vs Pairwise Alignment

- Up until now we have only tried to align two sequences.
Multiple Alignment vs Pairwise Alignment

- Up until now we have only tried to align two sequences.
- **What about more than two?**
  And what for?
MSA: why?

- A faint similarity between two sequences becomes significant if present in many.
- Multiple alignments can reveal subtle similarities that pairwise alignments do not reveal.

“Pairwise alignment whispers... multiple alignment shouts out loud” Hubbard, Lesk, Tramontano, Nature Structural Biology 1996.
Identify regions in the protein that have an important role in the structure and function of a group of related proteins: conservation implies importance

Construct *motifs/domains* that describe those regions
MSA: why?

- Identify and represent protein families
- Deduce evolutionary history
Genome multiple alignment?

- Yes, but need special methods for that
Correctness of an alignment

- Homologous residues should be in the same column of an MSA.
Generalizing the Notion of Pairwise Alignment

- Alignment of 2 sequences: a 2-row matrix
- Alignment of 3 sequences: a 3-row matrix

\[
\begin{align*}
\text{ATGC} & \quad \text{GCG} \\
\text{A} & \quad \text{CGT} \\
\text{ATCAC} & \quad \text{A}
\end{align*}
\]

- Score: more conserved columns, better alignment
Alignment = Paths in...

Align 3 sequences: ATGC, AATC, ATGC
Alignment Paths

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>--</td>
<td>T</td>
<td>G</td>
<td>C</td>
<td></td>
</tr>
</tbody>
</table>

x coordinate

| A | A | T | -- | C |

| -- | A | T | G | C |
Alignment Paths

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>--</td>
<td>T</td>
<td>G</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>A</td>
<td>A</td>
<td>T</td>
<td>--</td>
<td>C</td>
<td></td>
</tr>
</tbody>
</table>

x coordinate

y coordinate
Alignment Paths

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>--</td>
<td>T</td>
<td>G</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>A</td>
<td>A</td>
<td>T</td>
<td>--</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>--</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>C</td>
<td></td>
</tr>
</tbody>
</table>

- Resulting path in (x,y,z) space:
  
  (0,0,0) → (1,1,0) → (1,2,1) → (2,3,2) → (3,3,3) → (4,4,4)
Aligning Three Sequences

• Same strategy as aligning two sequences
• Use a 3-D grid, with each axis representing a sequence to align
• For global alignment, go from source to sink
2-D vs 3-D Alignment Grid

2-D edit graph

3-D edit graph
2-D cell versus 3-D Alignment Cell

In 2-D, 3 edges in each unit square

In 3-D, 7 edges in each unit cube
Architecture of 3-D Alignment Cell

(i-1,j-1,k-1)  (i-1,j-1,k)  (i-1,j,k)  (i-1,j,k-1)

(i,j-1,k-1)  (i,j-1,k)  (i,j,k)
Multiple Alignment: Dynamic Programming

\[ s_{i,j,k} = \max \left\{ \begin{array}{l}
  s_{i-1,j-1,k-1} + \delta(v_i, w_j, u_k) \\
  s_{i-1,j-1,k} + \delta(v_i, w_j, _) \\
  s_{i-1,j,k-1} + \delta(v_i, _, u_k) \\
  s_{i,j-1,k-1} + \delta(_, w_j, u_k) \\
  s_{i-1,j,k} + \delta(v_i, _, _) \\
  s_{i,j-1,k} + \delta(_, w_j, _) \\
  s_{i,j,k-1} + \delta(_, _, u_k) \\
\end{array} \right\} \]

\( \delta(x, y, z) \) is an entry in the 3-D scoring matrix

cube diagonal: no indels
face diagonal: one indel
edge diagonal: two indels
Multiple Alignment: Running Time

- For 3 sequences of length $n$, the run time is $7n^3$: $O(n^3)$

- Running time for $k$ sequences?
Multiple Alignment: Running Time

• For 3 sequences of length $n$, the running time is $O(n^3)$

• For $k$ sequences: $k$-dimensional grid. Running time: $(2^k-1)(n^k)$ i.e. $O(2^k n^k)$

• Conclusion: dynamic programming approach is easily extended to $k$ sequences but is impractical.
Additive MSA scores

- General structure of an MSA scoring scheme:

\[ S(m) = G + \sum_i S(m_i) \]

- Where \( m \) is the alignment, \( m_i \) is column \( i \) of the alignment and \( S(m_i) \) is the score of column \( i \), and \( G \) scores the gaps.

- Assumption: columns are independent
Sum of Pairs (SP) score

$$S(m_i) = \sum_{j < k} \delta(m_{ij}, m_{ik})$$

ATG–C–AAT
A–G–CATAT
ATCCCATTT

Score=3

Score = $1 - 2\mu$
Problem

- MSA with SP-score is NP-complete
What next?

- Heuristics: reducing MSA to pairwise alignment.
Reducing MSA to pairwise alignment

• **Progressive alignment** - a succession of pairwise alignments: at each step align sequence to an MSA that was already computed

• The primary heuristic: start with the most similar pairs of sequences since these will produce the most reliable alignments.

• Pro: Fast

• Con: What are we optimizing?
Profile Representation of an MSA

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>G</th>
<th>C</th>
<th>T</th>
<th>A</th>
<th>T</th>
<th>C</th>
<th>A</th>
<th>C</th>
<th>C</th>
<th>T</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>A</td>
<td>G</td>
<td>C</td>
<td>T</td>
<td>A</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>G</td>
</tr>
<tr>
<td>C</td>
<td>A</td>
<td>G</td>
<td>C</td>
<td>T</td>
<td>A</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>G</td>
</tr>
<tr>
<td>C</td>
<td>A</td>
<td>G</td>
<td>C</td>
<td>T</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>C</td>
<td>-</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>C</td>
<td>A</td>
<td>G</td>
<td>C</td>
<td>T</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>G</td>
<td>C</td>
<td>G</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A: 0 1 0 0 0 0 1 0 0 0 .8 0 0 0 0 0
C: .6 0 0 0 1 0 0 .4 1 0 .6 .2 0 0 0 0
G: 0 0 1 .2 0 0 0 0 0 0 .2 0 0 0 .4 1
T: .2 0 0 0 0 1 0 .6 0 0 0 0 .2 0
-: .2 0 0 .8 0 0 0 0 0 0 0 .4 .8 .4 0
Profile alignment

• How to align a sequence against a profile?
• How to align a profile against a profile?
• An alignment of two profiles induces a multiple sequence alignment of the sequences
Overview of ClustalW

```
1  PEEKSAVTALWGKVN--VDEVGG
2  GEEKAAVLALWDKVN--EEEVGG
3  PADKTNVKAAWGKVGAHAGEYGA
4  AADKTNVKAAWSKVGGHAGEYGA
5  EHEWQLVLHVWAKVEADVAGHGQ
```

sequences

→

distance matrix

→

guide tree (neighbor-joining)

→

progressive alignment following guide tree
Progressive alignment - step 1

gctcgatacgcgatagcta
gctcgatacacaagacgatgacagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagctagc-
Progressive alignment - step 2
Progressive alignment - step 3

gctcgatacgcgatacgcgactagcta
gctcgataacaagacgatgac-agctagcta
+
gctcgatacacadacgatgactagctagcta
gctcgataacacgatgacgagcga

gctcgatacgcgatacgcgactagcta
gctcgataacaagacgatgac-agctagcta
gctcgataacacgag---tgactagcta
gctcgataacacgag---tgaegagagega
Progressive alignment - final step

gctcgataacgatacgcgtgactagcta

gctcgatacaagacgatgac-agcta

gctcgatacacacgac---tgactagctga

gctcgatacacacga---tgacgagcga
+
ctcgaacgatacgcgtgactagcta

gctcgcggatacgcgtgactagcta

gctcgatacaagacgatgac-agcta

gctcgcggatacacacga---tgactagctga

gctcgcggatacacacga---tgacgagcga
-ctcga-acgatacgcgtgactagcta-
Progressive alignment

- Many flavors of progressive alignment.
- Algorithms differ in:
  - The order in which sequences are aligned (guide tree/no guide tree/what type of guide tree)
  - Scoring and alignment of a sequence/alignment to an alignment.
Where things can go wrong...

Sequence A GARFIELD THE LAST FAT CAT
Sequence B GARFIELD THE FAST CAT
Sequence C GARFIELD THE VERY FAST CAT
Sequence D THE FAT CAT

Clustal alignment

Sequence A GARFIELD THE LAST FA–T CAT
Sequence B GARFIELD THE FAST CA–T ---
Sequence C GARFIELD THE VERY FAST CAT
Sequence D -------- THE ---- FA–T CAT
Another example

A problem of progressive alignment:
- Initial alignments are “frozen” even when new evidence comes

Example:

<table>
<thead>
<tr>
<th>x:</th>
<th>GAAGTT</th>
<th>Frozen!</th>
</tr>
</thead>
<tbody>
<tr>
<td>y:</td>
<td>GAC-TT</td>
<td></td>
</tr>
<tr>
<td>z:</td>
<td>GAACTG</td>
<td></td>
</tr>
<tr>
<td>w:</td>
<td>GTACTG</td>
<td></td>
</tr>
</tbody>
</table>

Now clear that y = GA- CTT is correct
Iterative Refinement

- Barton-Sternberg method:
  Create an alignment using a progressive method.
  for \( i=1,\ldots,N \)
    Remove sequence \( i \) and realign it to a profile of the other aligned sequences.
  Repeat until convergence
MUSCLE

Combines several ideas:

- Draft progressive alignment
- Recompute a guide tree and recompute a progressive alignment.
- Refinement
MUSCLE

• Build a draft progressive alignment
  • Determine pairwise similarity through k-mer counting (no alignment)
  • Construct tree
  • Construct draft progressive alignment following tree
MUSCLE

- Improve the progressive alignment
  - Compute pairwise similarity using the current MSA
  - Construct new tree with Kimura distance measure
  - Compare new and old trees: if improved, repeat this step, if not improved, then we are done
Multiple Alignment: Timeline

1975 Sankoff
  Formulated multiple alignment problem and gave dynamic programming solution
1990 Feng-Doolittle
  Progressive alignment
1994 Thompson-Higgins-Gibson: ClustalW
  Most popular multiple alignment program
  Using the library of pairwise alignments
2004 MUSCLE
2005 PROBCONS, MAFFT
2012 Alignathon Contests
What can we do with an MSA?

• Represent it as a profile and use it for searching (alternative: HMM): PSI-BLAST
• Extract motifs: can then search for a particular feature of a protein
From MSA to discrete sequence motifs
Problems with the Formulation of MSA

- Multidomain proteins evolve not only through point mutations but also through domain duplications and domain recombinations.
- Often impossible to align all protein sequences throughout their entire length.
- Although MSA is a 30 year old problem, there were no MSA approaches for aligning rearranged sequences (i.e., multi-domain proteins with shuffled domains) prior to 2002.
Alignment as a Graph


Conventional Alignment

B [Graph representation of sequence as a path]

Sequence as a path

C [Graph representation of two paths]

Two paths

D [Graph representation of combined graph (partial order) of both sequences]

Combined graph (partial order) of both sequences
Representing Sequences as Paths in a Graph

Each protein sequence is represented by a path. Dashed edges connect “equivalent” positions; vertices with identical labels are fused.
Partial Order Alignment (POA) Algorithm

Aligns sequences onto a directed acyclic graph (DAG)

Steps:
1. Guide Tree Construction
2. Progressive Alignment Following Guide Tree
3. Dynamic Programming Algorithm to align two PO-MSAs (PO-PO Alignment).
POA Advantages

- POA is more flexible: standard methods force sequences to align linearly
- PO-MSA representation handles gaps more naturally and retains (and uses) all information in the MSA
A-Bruijn Alignment (ABA)

• POA: represents alignment as directed graph; no cycles
• ABA: represents alignment as directed graph that may contain cycles
ABA vs. POA vs. MSA
Protein MSA programs

http://www.ebi.ac.uk/clustalw/
CLUSTALW – most widely used

http://phylogenomics.berkeley.edu/cgi-bin/muscle/input_muscle.py
MUSCLE – very scalable

http://mafft.cbrc.jp/alignment/software/
MAFFT – very scalable

http://probcons.stanford.edu/
PROBCONS – very accurate

http://www.bioinformatics.ucla.edu/poa/
POA - a different approach (very fast!)

http://tcoffee.org
T-Coffee – accurate, can incorporate other information (3D-Coffee)