Multiple Sequence Alignments

Definition

Given N sequences $x_1, x_2, \ldots, x_N$:
- Insert gaps (-) in each sequence $x_i$, such that
  - All sequences have the same length $L$
  - Score of the global map is maximum

The sum-of-pairs score of an alignment is the sum of the scores of all induced pairwise alignments

$$S(m) = \sum_{k<l} s(m_k, m_l)$$

$s(m_k, m_l)$: score of induced alignment $(k,l)$

Consensus

• Find optimal consensus string $m^*$ to maximize

$$S(m) = \sum_s s(m, m)$$

$s(m, m)$: score of pairwise alignment $(k,l)$

Multidimensional Dynamic Programming

- Example: in 3D (three sequences):
  - 7 neighbors/cell

$$F(i,j,k) = \max \{ F(i-1,j-1,k-1)+S(x_i, x_j, x_k),$$
$$F(i-1,j-1,k)+S(x_i, x_j, -),$$
$$F(i-1,j,k-1)+S(x_i, -, x_k),$$
$$F(i-1,j,k)+S( -, x_j, x_k),$$
$$F(i,j-1,k-1)+S( -, -, x_k),$$
$$F(i,j-1,k)+S( -, - , x_k) \}$$
Progressive Alignment

- Multiple Alignment is NP-complete
- Most used heuristic: Progressive Alignment

**Algorithm:**
1. Align two of the sequences $x_i$, $x_j$
2. Fix that alignment
3. Align a third sequence $x_k$ to the alignment $x_i$, $x_j$
4. Repeat until all sequences are aligned

Running Time: $O(N L^2)$

Progressive Alignment: CLUSTALW

CLUSTALW: most popular multiple protein alignment

**Algorithm:**
1. Find all $d_{ij}$ alignment dist $(x_i, x_j)$
2. Construct a tree
   (Neighbor-joining hierarchical clustering)
3. Align nodes in order of decreasing similarity
4. A large number of heuristics

CLUSTALW & the CINEMA viewer

Iterative Refinement

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

**Example:**

- $x$: GAAGTT
- $y$: GAC-TT
- $z$: GAACTG
- $w$: GTACTG

Frozen!

New correct $y = GA-CTT$

Iterative Refinement

**Algorithm (Barton-Stenberg):**
1. Align most similar $x_i$, $x_i$
2. Align $x_i$ most similar to $(x_i, x_k)$
3. Repeat 2 until $(x_i, x_k)$ are aligned
4. For $j = 1$ to $N$
   - Remove $x_i$ and realign to $x_i, x_i, x_k$
5. Repeat 4 until convergence

**Note:** Guaranteed to converge
2. Iterative Refinement (cont’d)

For each sequence y
1. Remove y
2. Realign y
(while rest fixed)

Iterative Refinement

Example: align (x,y), (z,w), (xy,zw):

x: GAAGTTA
y: GAC-TTA
z: GAACTGA
w: GTAAGTA

After realigning y:

x: GAAGTTA
y: G-ACTTA
+ 3 matches
z: GAACTGA
w: GTAAGTA

Restricted MDP

• Here is a final way to improve a multiple alignment:

1. Construct progressive multiple alignment m
2. Run MDP, restricted to radius R from m

Running Time: O(2^N R^{N-1} L)

1. Restricted MDP

Run MDP, restricted to radius R from m

Running Time: O(2^N R^{N-1} L)

Restricted MDP (2)

x: GAAGTTA
y1: GAC-TTA
y2: GAC-TTA
y3: GAC-TTA
z: GAACTGA
w: GTAAGTA

• Within radius 1 of the optimal
⇒ Restricted MDP will fix it
MLAGAN: Multiple Alignment

1. Multi-anchoring
   - To anchor the (X/Y) and (Z) alignments:

2. Progressive Alignment
   - Given N sequences, phylogenetic tree
   - Align pairwise, in order of the tree (LAGAN)

3. Iterative Refinement
   - For each sequence y:
     1. Remove y
     2. Anchor "good" spots
     3. Realign y using LAGAN

Cystic Fibrosis (CFTR), 12 species
The "zoo" project

- Human sequence length: 1.8 Mb
- Total genomic sequence: 13 Mb

Performance in the CFTR region

<table>
<thead>
<tr>
<th>Method</th>
<th>Exons Perfect</th>
<th>Exons &gt;= 50%</th>
<th>TIME (sec)</th>
<th>MAX MEMORY (Mb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MUMmer</td>
<td>94%</td>
<td>40%</td>
<td>45</td>
<td>45</td>
</tr>
<tr>
<td>AVID</td>
<td>94%</td>
<td>40%</td>
<td>7</td>
<td>28</td>
</tr>
<tr>
<td>LAGAN</td>
<td>94%</td>
<td>40%</td>
<td>22</td>
<td>31</td>
</tr>
<tr>
<td>DIRECT</td>
<td>94%</td>
<td>40%</td>
<td>25</td>
<td>56</td>
</tr>
<tr>
<td>MLAGAN</td>
<td>94%</td>
<td>40%</td>
<td>45</td>
<td>670</td>
</tr>
</tbody>
</table>
Alignment & Rearrangements

Evolution at the DNA level

- Mutation
- Deletion

SEQUENCE EDITS

REARRANGEMENTS
- Inversion
- Translocation
- Duplication

Local & Global Alignment

Glocal Alignment Problem

Find least cost transformation of one sequence into another using new operations:

- Sequence edits
- Inversions
- Translocations
- Duplications
- Combinations of above

Shuffle-LAGAN

A glocal aligner for long DNA sequences

S-LAGAN: Find Local Alignments

1. Find Local Alignments
2. Build Rough Homology Map
3. Globally Align Consistent Parts
S-LAGAN: Build Homology Map

1. Find Local Alignments
2. Build Rough Homology Map
3. Globally Align Consistent Parts

Building the Homology Map

Chain (using Eppstein Galil); each alignment gets a score which is MAX over 4 possible chains. Penalties are affine (event and distance components)

Penalties:
a) regular  
b) translocation  
c) inversion  
d) inverted translocation

S-LAGAN: Global Alignment

1. Find Local Alignments
2. Build Rough Homology Map
3. Globally Align Consistent Fragments

S-LAGAN alignments
**S-LAGAN alignments (Chr 20)**

- Human Chr 20 v. homologous Mouse Chr 2.
- 270 Segments of conserved synteny
- 70 Inversions

**Some more examples**

- Hum/Mus
- Hum/Rat

**Some more examples**

- Hum/Mus
- Hum/Rat

**Some more examples**

- Hum/Mus
- Hum/Rat

**Some more examples**

- Hum/Mus
- Hum/Rat
Some more examples