Multiple sequence alignment 1

Multiple alignment methods

Sequence analysis 2007

Lecture 5 - 12/12/07
Multiple alignment idea

- Take three or more related sequences and align them so that the greatest number of similar characters are aligned in the same column of the alignment.
Biological definitions for related sequences

- **Homologues** are similar sequences in two different organisms that have been derived from a common ancestor sequence. Homologues can be described as either orthologues or paralogues.

- **Orthologues** are similar sequences in two different organisms that have arisen due to a speciation event. Orthologues typically retain identical or similar functionality throughout evolution.

- **Paralogues** are similar sequences within a single organism that have arisen due to a gene duplication event.

- **Xenologues** are similar sequences that do not share the same evolutionary origin, but rather have arisen out of horizontal transfer events through symbiosis, viruses, etc.
So this means ...

Alignments are useful ...

Conserved patterns
Evolutionary analysis
Structure prediction
Motifs
Function prediction

... and they are used!

- BLAST 22776 citations
- ClustalW 22638 citations
Information content of a multiple alignment

• Sequences can be conserved across species and perform similar or identical functions
  • hold information about which regions have high mutation rates over evolutionary time and which are evolutionarily conserved
  • identification of regions or domains that are critical to functionality

• Sequences can be mutated or rearranged to perform an altered function
  • which changes in the sequences have caused a change in the functionality
Information content of a multiple alignment

What to ask yourself

• How do we get a multiple alignment? (three or more sequences)

• What is our aim?
  • Do we go for max accuracy?
  • Least computational time?
  • Or the best compromise?

• What do we want to achieve each time?
Sequence-sequence alignment
Exhaustive & Heuristic algorithms

- Exhaustive approaches
  - Examines all possible aligned positions simultaneously
  - Looks for the optimal solution by DP
  - Very (very) slow

- Heuristic approaches
  - Strategy to find a near-optimal solution (by using rules of thumb)
  - Shortcuts are taken by reducing the search space according to certain criteria
  - Much faster
Multiple alignment methods

- **Multi-dimensional dynamic programming**
  - extension of pairwise sequence alignment.

- **Progressive alignment**
  - incorporates phylogenetic information to guide the alignment process

- **Iterative alignment**
  - correct for problems with progressive alignment by repeatedly realigning subgroups of sequence
Simultaneous multiple alignment
Multi-dimensional dynamic programming

The combinatorial explosion

• 2 sequences of length n
  • n^2 comparisons

• Comparison number increases exponentially
  • i.e. n^N where n is the length of the sequences, and N is the number of sequences

• Impractical for even a small number of short sequences
Multi-dimensional dynamic programming
Murata et al. 1985
The MSA approach
Lipman et al. 1989

- **Key idea**: restrict the computational costs by determining a minimal region within the n-dimensional space that contains the optimal path
The MSA method *in detail*

1. Let’s consider 3 sequences
2. Calculate all pair-wise alignment scores by Dynamic programming
3. Use the scores to predict a tree
4. Produce a heuristic multiple align. based on the tree (quick & dirty)
5. Calculate maximum cost for each sequence pair from multiple alignment (upper bound) & determine paths with < costs.
6. Determine spatial positions that must be calculated to obtain the optimal alignment (intersecting areas)

*Note* Redundancy caused by highly correlated sequences is avoided
The DCA (Divide-and-Conquer) approach

Stoye et al. 1997

- Each sequence is cut in two behind a suitable cut position somewhere close to its midpoint.
- This way, the problem of aligning one family of (long) sequences is divided into the two problems of aligning two families of (shorter) sequences.
- This procedure is re-iterated until the sequences are sufficiently short.
- Optimal alignment by MSA.
- Finally, the resulting short alignments are concatenated.
So in effect ...
Multiple alignment methods

- Multi-dimensional dynamic programming
  - extension of pairwise sequence alignment.

- Progressive alignment
  - incorporates phylogenetic information to guide the alignment process

- Iterative alignment
  - correct for problems with progressive alignment by repeatedly realigning subgroups of sequence
The progressive alignment method
Stepwise assembly of multiple alignment.

- **Underlying idea**: usually we are interested in aligning families of sequences that are evolutionary related.

- **Principle**: construct an approximate phylogenetic tree (guide tree) for the sequences to be aligned and then build up the alignment by progressively adding sequences in the order specified by the tree.
Progressive alignment strategy

All individual pairwise alignment and construction of distance matrix

Calculating a guide tree; C & D the closest pair; A & B the next closest pair

Aligning C/D and A/B separately using dynamic programming

Figure adapted from Xiong, J. “Essential Bioinformatics”
But how can we align blocks of sequences?

- The dynamic programming algorithm performs well for pairwise alignment (two axes).
- So we should try to treat the blocks as a “single” sequence …
How to represent a block of sequences?

- **Historically**: consensus sequence
  single sequence that best represents the amino acids observed at each alignment position.

- **Modern methods**: alignment profile
  representation that retains the information about frequencies of amino acids observed at each alignment position.
Consensus sequence

<table>
<thead>
<tr>
<th></th>
<th>Sequence 1</th>
<th>Sequence 2</th>
<th>Consensus</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FATNMGSTDSPPTHTRTLRLKLVSQ</td>
<td>FVTNMNNNSDGPTHTKLRLKLVSST</td>
<td>F<em>TNM</em>SD<em>PTHT</em>LRKLVS*</td>
</tr>
</tbody>
</table>

- **Problem**: loss of information

- For larger blocks of sequences it “punishes” especially more distant members within a
Alignment profiles

• **Advantage:** full representation of the sequence alignment (more information retained)

• Not only used in alignment methods, but also in sequence-database searching (to detect distant homologues)

• Also called PSSM (*Position-specific scoring matrix*)
Multiple alignment profiles
Gribskov et al. 1987

- **Gribskov** created a probe: group of typical sequences of functionally related proteins that have been aligned by similarity in sequence or three-dimensional structure (in his case: globins & immunoglobulins).

- Then he constructed a profile, which consists of a sequence position-specific scoring matrix \( M(p,a) \) composed of 21 rows and \( N \) columns \( (N = \text{length of probe}) \).

- The first 20 rows of each column specify the score for finding, at that position in the target, each of the 20 amino acid residues. An additional row contains a penalty for insertions or deletions at that position (gap-opening and gap-extension).
**Multiple alignment profiles**

Position dependent gap penalties
Profile building
Each aa is represented as a frequency, penalties as weights

Position dependent gap penalties

<table>
<thead>
<tr>
<th>Gap penalties</th>
<th>A</th>
<th>C</th>
<th>W</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>0.5</td>
<td>0.3</td>
<td>0.1</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>0</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Profile-sequence alignment
Sequence to profile alignment

Score of amino acid L in sequence that is aligned against this profile position:

\[ \text{Score} = 0.4 \times s(L, A) + 0.2 \times s(L, L) + 0.4 \times s(L, V) \]
Profile-profile alignment
Profile to profile alignment

Match score of these two alignment columns using the a.a frequencies at the corresponding profile positions:

Score = 0.4*0.75*s(A,G) + 0.2*0.75*s(L,G) + 0.4*0.75*s(V,G) +
       + 0.4*0.25*s(A,S) + 0.2*0.25*s(L,S) + 0.4*0.25*s(V,S)

s(x,y) is value in amino acid exchange matrix (e.g. PAM250, Blosum62) for amino acid pair (x,y)
So for scoring profiles ...

- Think of sequence-sequence alignment.
- Same principles but more information for each position.

Reminder

- The sequence pair alignment score $S$ comes from the sum of the positional scores $M(aa_i, aa_j)$ (i.e. the substitution matrix values at each alignment position minus penalties if applicable)
- Profile alignment scores are exactly the same, but the positional scores are more complex
**Scoring a profile position**

At each position (column) we have different residue frequencies for each amino acid (rows)

SO:

- Instead of saying $S = M(aa_1, aa_2)$ (one residue pair)
- *For frequency $f > 0$ (amino acid is actually there) we take:*

$$S = \sum_{i=1}^{20} \sum_{j=1}^{20} faa_i \times faa_j \times M(aa_i, aa_j)$$
Log-average score

• Remember the substitution matrix formula?

\[ S = \sum_{i}^{20} \sum_{j}^{20} faa_i \times faa_j \times \log \frac{p_{aa_i,aa_j}}{q_{aa_i} q_{aa_j}} \]

• In log-average scoring (von Ohsen et al, 2003)

\[ S = \log \sum_{i}^{20} \sum_{j}^{20} faa_i \times faa_j \times \frac{p_{aa_i,aa_j}}{q_{aa_i} q_{aa_j}} \]

• What is the effect?
Return to the **progressive alignment**

1. Perform pair-wise alignments of all of the sequences;
2. Use the alignment scores to produce a dendrogram using neighbour-joining methods (guide-tree);
3. Align the sequences sequentially, guided by the relationships indicated by the tree.

- Biopat (first method ever)
- MULTAL (Taylor 1987)
- DIALIGN (1&2, Morgenstern 1996)
- PRRP (Gotoh 1996)
- ClustalW (Thompson et al 1994)
- PRALINE (Heringa 1999)
- T Coffee (Notredame 2000)
- POA (Lee 2002)
- MUSCLE (Edgar 2004)
- Probcons (Do 2004)
Progressive multiple alignment

1
2
3
4
5

Score 1-2
Score 1-3
Score 4-5

5×5

Scores to distances

Scores

Similarity matrix

Guide tree

Multiple alignment

Iteration possibilities
General progressive multiple alignment technique (followed generated tree)
PRALINE progressive strategy

There are problems ...

Accuracy is very important

- Alignment errors during the construction of the MSA cannot be repaired anymore: propagated into the progressive steps.

- The comparisons of sequences at early steps during progressive alignments cannot make use of information from other sequences.

- It is only later during the alignment progression that more information from other sequences (e.g. through profile representation) becomes employed in the alignment steps.

“Once a gap, always a gap”

Feng & Doolittle, 1987