Multiple sequence alignment 1

Multiple alignment methods

Algorithms in Sequence Analysis 2010

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 ...

... and they are used!
- BLAST 24987 citations (ISI - 09/12/08)
- ClustalW 25996 citations (ISI - 09/12/08)

Think at journal impact factors:
- A = the number of times articles published in 2007-2008 were cited in indexed journals during 2009
- B = the number of articles, reviews, proceedings or notes published in 2007-2008
- impact factor 2009 = A/B
- BLAST paper from 1990, ClustalW paper from 1994
**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 function.
- Sequences can be mutated or rearranged to perform an altered function, which changes in the sequences have caused a change in function.

**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**
  - Examine all possible aligned positions simultaneously.
  - Look 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.

**Heuristic algorithm**

- Does not guarantee that an optimal solution will be found.
- But makes it likely.
- Typically, a heuristic is used to make an algorithm faster (or sometimes makes solving a problem feasible).
- Some problems have an exponential complexity; this relates to so-called NP-complete (or NP-hard) problems in computer science.
- NP stands for Non Polynomial, meaning that the problem grows faster than any fixed exponent.
- So, $a^x$ rather than $x^a$, where $x$ is the problem size.
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

Calculate all pair

Let's consider 3 sequences

Comparison number increases exponentially

2 sequences of length \( n \)

This renders the impractical.

Calculate lower bound score for.

Impractical for even a small number of short sequences

Use the scores to predict a tree

Correct for problems with progressive alignment by

Produce a heuristic multiple align.

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.
5. Based on the tree (quick & dirty)
6. Calculate lower bound score for each sequence pair from this multiple alignment & determine paths with scores \( \geq \) lower bound.
7. Determine spatial positions in multidimensional matrix that must be calculated to obtain the optimal alignment (intersecting areas)

Note: Redundancy caused by highly correlated sequences is avoided

The MSA method *in practice*

Notwithstanding the complexity optimisation by means of bound determination (intersecting areas), the MSA method is only able to align 7-9 sequences of moderate length.

- This renders the method infeasible for biological data sizes

The MSA approach

Lipman et al. 1989

**Key idea**: restrict the computational costs by determining a minimal region within the \( n \)-dimensional space that will contain the optimal path.
The DCA (Divide-and-Conquer) approach

- 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.
- The algorithm allows the alignment of up to 10 sequences in practice.

Progressive alignment strategy

Four steps:
1. All-against-all pairwise alignment
2. Distance matrix
3. Guide tree
4. Progressive build-up of alignment (following tree order)

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

Calculating a guide tree:
- C & B the closest pair
- A & B the next closest pair

But how can we align blocks of sequences?

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 ...

So in effect ...

- Multiple sequence alignment 1
- Algorithms in Sequence analysis 2010
- Progression in Sequence analysis 2010
- Sequence analysis 2010

Multiple alignment methods

- Principal: Optimal alignment by MSA.
- Principle: construct an approximate phylogenetic tree for the sequences to be aligned and then build up the alignment by progressively adding sequences in the order specified by the tree.

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

Four steps:
1. All-against-all pairwise alignment
2. Distance matrix
3. Guide tree
4. Progressive build-up of alignment (following tree order)

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 ...

The DCA (Divide-and-Conquer) approach

- 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.
- The algorithm allows the alignment of up to 10 sequences in practice.

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

Four steps:
1. All-against-all pairwise alignment
2. Distance matrix
3. Guide tree
4. Progressive build-up of alignment (following tree order)

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 ...
Historically, Gribskov et al. 1987 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 ($N =$ 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 gap penalty weight for insertions or deletions at that position (gap-opening and gap-extension are weighted).

### 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.

### Problem: loss of information

- For larger blocks of sequences it particularly “punishes” more distant sequence members.

### Alignment profiles

- **Advantages:** 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)
- **Disadvantage:** still loss of information

### Multiple alignment profiles

- 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 ($N =$ 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 gap penalty weight for insertions or deletions at that position (gap-opening and gap-extension are weighted).
Profile building

Each aa is represented as a frequency, with an additional gap penalty weight.

Profile-sequence alignment

Sequence to profile alignment

Profile-profile alignment

Profile to profile alignment

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 $M(a_{aa}, a_{aj})$ (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

Profile building

Each aa is represented as a frequency, with an additional gap penalty weight.

Profile to profile alignment

- Match score of these two alignment columns using the aa frequencies at the corresponding profile positions:
  \[
  \text{Score} = 0.4 \cdot 0.75 \cdot s(A, A) + 0.2 \cdot 0.75 \cdot s(L, G) + 0.4 \cdot 0.75 \cdot s(V, G) + 0.4 \cdot 0.25 \cdot s(A, A) + 0.2 \cdot 0.25 \cdot s(L, G) + 0.4 \cdot 0.25 \cdot s(V, G)
  \]
- $s(x,y)$ is value in amino acid exchange matrix (e.g. PAM250, Blosum62) for amino acid pair (x,y)
At each position (column) we have different residue frequencies for each amino acid (rows).

Log-average score

- Remember the substitution matrix formula?

\[ S = \sum_{i=1}^{20} \sum_{j=1}^{20} f_{aa_i} \times f_{aa_j} \times \log\frac{p_{aa_{ai}aa_{aj}}}{q_{aa_i}q_{aa_j}} \]

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

\[ S = \log \sum_{i=1}^{20} \sum_{j=1}^{20} f_{aa_i} \times f_{aa_j} \times \frac{p_{aa_{ai}aa_{aj}}}{q_{aa_i}q_{aa_j}} \]

- What is the effect?

Progressive multiple alignment

Scores to distances

Scores

Guide tree

Multiple alignment

Return to 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 (LÖ, Morgenstern 1996)
- PROPR (Ocsten 1996)
- ClustalW (Thompson et al 1994)
- PRALINE (Heringa 1999)
- T Coffee (Notredame 2000)
- POA (Lee 2002)
- MUSCLE (Edgar 2004)
- Procon (Do 2004)

**PRALINE progressive strategy**

(builds a tree “on the fly)

http://ibi.vu.nl/programs/praline/
There are problems ...
Accuracy is very important
- Alignment errors during the construction of the MSA cannot be repaired anymore: errors become propagated during 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.
- The progressive multiple alignment strategy is a greedy algorithm, things that are done (i.e. sequences that have been aligned) are not reconsidered or changed anymore:
  "Once a gap, always a gap"

Feng & Doolittle, 1987

Pair-wise alignment quality versus sequence identity
- Vogt et al., JMB 249, 816-831, 1995

Alignment quality quickly deteriorates with more distant sequences: the guide tree concept (W X, Clustal X alignments: close sequences and more distant alignments later on during progressive alignment) appears sound.

Strategies for progressive alignment optimization
- Heuristics
- Profile pre-processing
- Secondary structure-guided alignment
- Globalised local alignment
- Matrix extension

Objective: try to avoid (early) errors

Clustal, ClustalW, ClustalX
Higgins et al. 1994
- Probably most well-known progressive alignment program
- Neighbour Joining (NJ) algorithm (Saitou and Nei, 1984) to construct guide tree
- NJ does not require that all lineages have diverged by equal amounts
- Sequence blocks are represented by profiles
- Individual sequences are additionally weighted according to the branch lengths in the NJ tree

Clustal, ClustalW, ClustalX
- Further carefully crafted heuristics include:
  - local gap penalties
  - automatic selection of the amino acid substitution matrix
  - automatic gap penalty adjustment
  - mechanism to delay alignment of sequences that appear to be distant at the time they are considered.
- Clustal (WX) does not allow genuine iteration (Hogeweg and Heaper, 1984; Corpet, 1988, Gotoh, 1996; Heringa, 1999, 2002)
- Recently, a corrective iteration step was implemented
  - Each sequence is taken out of the MSA and realigned until the alignment score cannot be improved anymore (Barton and Sternberg, 1987)
  - Still limited by "greedy" nature of progressive algorithm

Remember the greediness of the progressive alignment algorithm?
Flavodoxin fold: helix-beta-helix

- mainly bacterial
- electron-transfer proteins (transport systems)
- two alpha-helical layers sandwiching 5 stranded parallel beta-sheet

Flavodoxin cheY NJ tree

12 Dec 2007
**T-COFFEE**
Notredame, Higgins and Heringa 2000

- Performs progressive alignment as in Clustal
- But there are many differences …
  - Integrating different pair-wise alignment techniques (NW, SW, …)
  - Combining different multiple alignment methods (consensus multiple alignment)
  - Combining sequence alignment methods with structural alignment techniques
  - Plug in user knowledge

---

**Search matrix extension**

---

**So for T-COFFEE …**

- Original pairwise alignments are refined based on the consistency with other “intermediate” sequences
- Distance matrix and guide tree are built based on the refined alignments
The T-COFFEE effect

- Although a direct alignment might choose one path, using information from the other sequences (T-COFFEE) finds an alternate and usually better one
- Minimize errors in the early stages of alignment assembly
- Extra computation time for the calculation of the consistency scores

3D-COFFEE

- Computes structural based alignments
- Structures related to the sequences are retrieved
- More accurate … but for many (many) proteins we do not have the structure!

but ... T-COFFEE(v1.23) Flavodoxin-cheY

Recap

- Simultaneous multiple sequence alignment
  - MSA (multi-dimensional DP with restricted matrix search)
  - DCA (divide-and-conquer)
- Progressive alignment strategy (greedy, "once a gap always a gap", error propagation)
  - Clustal(W) (various tricks to optimise alignment)
  - T-Coffee (local and global alignment, matrix extension by transitivity)

References