Table of Contents

Table of Contents 1
Multiple Sequence Alignment 1
  What a multiple alignment means 2
Scoring a multiple alignment 2

Multiple Sequence Alignment

Multiple sequence alignment techniques are most commonly applied to protein sequences; ideally they are a statement of both evolutionary and structural similarity among the proteins encoded by each sequence in the alignment.

Multiple alignments must usually be inferred from primary sequences alone. Biologists produce high quality multiple sequence alignments by hand using expert knowledge of protein sequence evolution. This knowledge comes from experience. Important factors include:

- specific sorts of columns in alignments, such as highly conserved residues or buried hydrophobic residues
- the influence of secondary or tertiary structure, such as the alteration of hydrophobic and hydrophilic columns in exposed beta sheet
- expected patterns of insertions and deletions, that tend to alternate with blocks of conserved sequence

The phylogenetic relationships between sequences dictate constraints on the changes that occur in columns and in the patterns of gaps.

Manual alignment is tedious. To automate the process, it is hard to define exactly what an optimal multiple sequence alignment is, and impossible to set a standard for a single correct multiple alignment. In theory, there is one underlying evolutionary process and one evolutionarily correct alignment generated from any group of sequences. However, the differences between sequences can be so great in parts of an alignment that there isn't an apparent, unique solution to be found by an alignment algorithm. Those same divergent regions are often structurally unalignable as well. Most of the insight that we derive from multiple alignments comes from analyzing the regions of similarity, not from attempting to align highly diverged regions.

In general, an automatic method must have a way to assign a score so that better multiple alignments get better scores. We should carefully distinguish the problem of scoring a multiple alignment from the problem of searching over possible multiple alignments to find the best one. Descriptions of multiple alignment programs tend to emphasize the alignment algorithm rather than the scoring function. However, the scoring function is our primary concern in probabilistic modeling. We wish to incorporate an expert's evaluation criteria into our scoring procedure.

To automate multiple alignment, we need to do the following:

- looking at what we need to do for automatic multiple alignment structurally and evolutionarily
- considering how to turn the biological criteria into a numerical scoring scheme, so that a program will recognize a good multiple alignment.
- Examine various approaches by different multiple alignment programs
- Describing a full probabilistic multiple alignment approaches based on profile HMM

What a multiple alignment means

In a multiple sequence alignment, homologous residues among a set of sequences are aligned together in columns. 'Homologous' is meant for both structural and evolutionary sense. Ideally, a column of aligned residues occupy similar 3D structural positions and all diverge from a common ancestral residue.

Except for trivial cases of highly identical sequences, it is not possible to unambiguously identify structurally or evolutionarily homologous positions and create a single 'correct' multiple alignment. Since protein structures also evolve, we do not expect 2 protein structures with different sequences to be entirely superposable. Even the definition of 'structurally superposable' is subjective and can be expected to vary among experts.

In principle, there is always an unambiguously correct alignment even if the structures diverge. In practice, however, an evolutionarily correct alignment can be even more difficult to infer than a structural alignment. Structural alignment has an independent point of reference, superposition of NMR structures. The evolutionary history of the residues of a sequence family cannot be independently known from any source. It must be inferred from sequence alignment.

The program should not be asked to produce exactly the same alignment. Instead, it should be focused on the subset of columns corresponding to key residues and core structural elements that can be aligned with confidence.

Scoring a multiple alignment

The scoring system should take 2 important features into account:

1. some positions are more conserved than others
2. sequences are not independent, but instead are related by a phylogenetic tree

An idealized way to score a multiple alignment would be to specify a complete probabilistic model of molecular sequence evolution. Given the correct phylogenetic tree for the sequences, the probability of a multiple alignment is the product of all the evolutionary events necessary to produce that alignment via ancestral intermediate sequences times the prior probability of the root ancestral sequence. This evolutionary model would be very complex. The probabilities of evolutionary change would depend on the evolutionary times along each branch of the tree, as well as position specific structural and functional constraints imposed by natural selection. This way key residues

and structural elements would be conserved. High probability alignments would then be good structural and evolutionary alignments under this model. Unfortunately, we don't have enough data to parameterize this model. Therefore assumptions must be made.

Almost all alignment methods assume that the individual columns of an alignment are statistically independent. Such a scoring function is written as follows:

$$S(m) = G + \sum_i S(m_i) \qquad (1)$$

Where mi is the column i of the multiple alignment m, S(mi) is the score for the column i, and G is a function for scoring the gaps that occur in the alignment. Most multiple alignment methods use affine scoring functions.

Minimum entropy

If the phylogenetic tree for sequences has many intermediate ancestors, then the statistical dependence between sequences is complex. The scoring problem is greatly simplified if we assume that sequences have all been generated independently. If we assume that residues within the column are independent, as well as being independent between columns, then the probability of a column mi is

$$P(m_i) = \prod_a Pia^{Cia} \qquad (2)$$

Where Pia is the probability of residue a in column i. We define a column score as the negative logarithm of this probability

$$S(m_i) = -\sum_a cia \log pia$$

This is an entropy measure directly related to the equation for Shannon entropy in information theory. It is a convenient measure of the variability observed in an aligned column of residues. The more variable the column, the higher the entropy. A completely conserved column would score 0. A good alignment is one which minimizes the total entropy of the alignment.

Thus, in return for giving up evolutionary tree and assuming independence between sequences, we gain the ability to straightforwardly estimate a position-specific model of both residue probabilities in columns and insertions and deletions. This assumption, however, can only be reasonable if representative sequences of a sequence family are chosen carefully. In practice, sample sequences are often biased with under or over representations of sub families. Several tree-based weighting schemes have been devised to deal with this.

Sum of pairs: SP scores

The standard method of scoring multiple alignments is not the HMM formulation, but is similar in that it does not use a phylogenetic tree and it assumes statistical independence for the columns. Columns are scored by an SP function using substitution scoring matrix. The SP score for a column is defined as:

$$S(m_i) = \sum_{k<l} s(m_i^k, m_i^l)$$

Where scores s(a,b) come from a substitution matrix such as PAM or BLOSUM.

Multidimensional dynamic programming

The dynamic programming algorithms used for pairwise sequences alignment can theoretically be extended to any number of sequences. However, the time and memory requirements of this algorithm increase exponentially with the number of sequences.

The only assumption necessary to make multidimensional dynamic programming to work is that column scores are independent.

A common approach to multiple sequence alignment is to progressively align pairs of sequences. The general strategy is:

1. A starting pair of sequences is selected and aligned
2. Each subsequent sequence is aligned to the previous alignment

This is a greedy heuristic algorithm. A greedy algorithm decomposes a problem into pieces, and then choose the best solution to each piece without paying attention to the problem as a whole. Since it is a heuristic algorithm, progressive alignment is not guaranteed to find the best solution. In practice, however, progressive alignment methods are efficient and produce biologically meaningful results.