Frederick Lo
Biochemistry 218
Final Project Paper

Three Recently Developed Algorithms for Aligning Distantly Related Proteins

Introduction

The rate at which new protein sequences are discovered has long outpaced the rate at which those proteins are experimentally assigned functions. To speed the process of function assignment, protein sequences with unknown functions can be compared to sequences with known functions. If two protein sequences are shown to be highly related, then it follows that the protein functions may be closely related as well (Domingues et. al 2000).

Early sequence alignment algorithms such as Smith-Waterman and BLAST focused on comparing the residue sequences only in making alignments. However, function may be conserved even in sequences that appear to have diverged. Much recent effort has been devoted to developing algorithms that align distantly related proteins, also known as remote homologues. The three algorithms discussed in this paper consider three separate approaches to this alignment problem. The three algorithms are the Structure-Dependent Sequence Alignment, a hybrid Iterative-Parametric approach to suboptimal alignment, and the amino acid property-based Proximity Correlation Matrix alignment.

Structure-Dependent Sequence Alignment

Previous studies have shown that structural motifs in related proteins often remain highly conserved even in the presence of significant divergences in sequence (Jaroszewski et. al 2000). Therefore, integrating structural data generally leads to more accurate protein sequence alignments than using sequence data alone, especially for distantly related proteins. Structural alignment is frequently cited as the "gold standard" for sequence alignment (Sunyaev et. al 2004). Since protein structure and function are closely related, it is no surprise that a protein's amino acid sequence is much easier to

determine than the protein's structure, such that the number of known protein sequences far outstrips the number of known protein structures (Jaroszewski et. al 2000). Unless the structures of both the query and template proteins are known, structural alignments cannot be conducted. The Structure-Dependent Sequence Alignment, or SDSA, algorithm attempts to bridge this gap (Yang 2002).

SDSA is a sequence-structure alignment algorithm that is adapted from the Needleman-Wunsch global sequence alignment algorithm. The Needleman-Wunsch algorithm can be expressed with the following equation:

$$H_{ij} = D_{ij} + \max[H_{i+1,j+1},\ \max_{k=i+2,Na}(H_{k,j+1} - g((i,j),(k,j+1))),\ \max_{k=j+2,Nb}(H_{i+1,k} - g((i,j),(i+1,k)))] \qquad \text{(Eq. 1)}$$

D is the sequence-sequence amino acid substitution matrix, where D_{ij} is the score for substituting the template residue at position j with the query residue at position i. H is the scoring matrix, where H_{ij} is the maximum score for all sequences rooted at (i, j). Na and Nb are the number of residues found in the query and template sequences, respectively. The function g((i, j)(k, j+1)) is the penalty for inserting a gap in the query sequence, whereas g((i, j)(i+1, k)) is the penalty for deleting a subsequence in the template. The formulae for determining g in SDSA will be discussed later. The maximal value in H indicates the best alignment, and the entire alignment can be obtained by traversing H diagonally downward from left to right, where position (1, 1) represents the top left corner of H (Yang 2002).

At the core of SDSA is the construction of a structure-based amino acid substitution matrix D. Since the structure of the template sequence is known, residue j in the template sequence can be identified as belonging to an α-helix, β-strand, or coil region. Thus, the following conditional equations were developed for D:

$$D_{ij} = q_j\left[A_{ij} + fP_i(\alpha)\right] + w_j \quad \text{when residue j belongs to an α-helix} \qquad \text{(Eq. 2a)}$$

$$\text{or} \quad D_{ij} = q_j\left[B_{ij} + fP_i(\beta)\right] + w_j \quad \text{when residue j belongs to a β-strand} \qquad \text{(Eq. 2b)}$$

$$\text{or} \quad D_{ij} = C_{ij} + fP_i(c) \quad \text{when residue j belongs to a coil region.} \qquad \text{(Eq. 2c)}$$

A, *B*, and *C* are amino acid substitution matrices for residues in α-helices, β-strands, and coil regions, respectively. Yang used a database of protein structure fragments to obtain

170,673 pairs of local structure alignments, which consisted of 503,466, 1,130,201, and 516,046 pairs of aligned α-helices, β-strands, and coil regions, respectively. By processing these pairs with the Protein Informatics System for Modeling (PrISM.1) structural alignment procedure also developed by Yang and utilizing formulas used by Henikoff and Henikoff (1992) to derive the BLOSUM substitution matrices, the *A*, *B*, and *C* matrices were constructed (Yang 2002).

The functions $P_i(x)$ in Equation 2, where x is α, β, or c, represent the log-odds probability that amino acid i will be found in an α-helix, β-strand, or coil region, respectively. These probabilities were derived from the pairings generated for the *A*, *B*, and *C* substitution matrices above. The general formula for $P_i(x)$ is:

$$P_i(x) = \ln\left(\frac{n_i(x)/n_i}{n(x)/n}\right)$$

where $n_i(x)$ is the number of amino acid type i found the sequences of structure type x, n_i is the number of amino acid type i found in the total set of residues from all pairings, n(x) is the number of residues found in all pairings of type x, and n is the total number of residues found in all pairings. The numerator $n_i(x)/ n_i$ can be interpreted as the probability that a given amino acid type i will be found in structure type x, while the denominator n(x)/n can be interpreted as the probability that a random amino acid will be found in structure type x (Yang 2002).

Parameters q, f, and w_j in Equation 2 are user specified. Parameter q is a measure of whether residue i is exposed to solvent or buried. The residue is considered buried if less than 20% of its surface area is exposed to solvent. The weighting parameter f for the log-odds probability should ideally have a value of 1 but is present to allow flexibility to improve alignment accuracy. Parameter f will be determined by using training sets. Parameter w_j provides extra weight for α-helix and β-strand residues, to differentiate them from coil regions. Although Yang claims that w_j can be varied given information about the template structure, no explanation was given on how this would be done. Parameter w_j is generally set to 1 (Yang 2002).

With the amino acid substitution matrix D now constructed using Equation 2, it can be placed in Equation 1. However, the gap insertion and deletion components of