Chapter 2.15
Algorithmic Aspects of Protein Threading

Tatsuya Akutsu
Kyoto University, Japan

ABSTRACT

This chapter provides an overview of computational problems and techniques for protein threading. Protein threading is one of the most powerful approaches to protein structure prediction, where protein structure prediction is to infer three-dimensional (3-D) protein structure for a given protein sequence. Protein threading can be modeled as an optimization problem. Optimal solutions can be obtained in polynomial time using simple dynamic programming algorithms if profile type score functions are employed. However, this problem is computationally hard (NP-hard) if score functions include pairwise interaction preferences between amino acid residues. Therefore, various algorithms have been developed for finding optimal or near-optimal solutions. This chapter explains the ideas employed in these algorithms. This chapter also gives brief explanations of related problems: protein threading with constraints, comparison of RNA secondary structures and protein structure alignment.

INTRODUCTION

Inference and mining of functions of genes is one of the main topics in bioinformatics. Protein structure prediction provides useful information for that purpose because it is known that there exists close relationship between structure and function of a protein, where protein structure prediction is a problem of inferring three-dimensional structure of a given protein sequence. Computational inference of protein structure is important since determination of three-dimensional structure of a protein is much harder than determination of its sequence.

There exist various kinds of approaches for protein structure prediction (Clote & Backofen, 2000; Lattman, 2001; Lattman, 2003). Ab initio approach tries to infer structure of a protein based on the basic principles (e.g., energy minimization) in physics. In this approach, such techniques as molecular dynamics and Monte Carlo simulations have been employed. Homology modeling approach tries to infer structure of a protein using

structure of a homologous protein (i.e., a protein whose structure is already known and whose sequence is very similar to the target protein sequence). In this approach, backbone structure of a protein is first computed from structure of a homologous protein and then the whole structure is computed by using molecular dynamics and/or some optimization methods. Secondary structure prediction approach does not aim to infer three-dimensional structure. Instead, it tries to infer which structural class (α, β, others) each residue belongs to. Such information is believed to be useful for inference of three-dimensional structure and/or function of a protein. In secondary structure prediction approach, various machine learning methods have been employed, which include neural networks and support vector machines.

Protein threading is another major approach for protein structure prediction. In this approach, given an amino acid sequence and a set of protein structures (structural templates), a structure into which the given sequence is most likely to fold is computed. In order to test whether or not a sequence is likely to fold into a structure, an alignment (i.e., correspondence) between spatial positions of a 3-D structure and amino acids of a sequence is computed using a suitable score function. That is, an alignment which minimizes the total score (corresponding to the potential energy) is computed. This minimization problem is called the protein threading problem. Though there exists some similarity between protein threading and homology modeling, these are usually considered to be different: alignment between a target sequence and a template structure is computed in protein threading whereas alignment between two sequences is computed in homology modeling.

Many studies have been done on protein threading. Most of them focus on improvement of practical performances using heuristic and/or statistical techniques, and few attentions had been paid to algorithmic aspects of protein threading. Since protein threading is NP-hard in general, heuristic algorithms have been widely employed, which do not necessarily guarantee optimal solutions. However, recent studies (Xu, Xu, Crawford, & Einstein, 2000; Xu, Li, Kim, & Xu, 2003) suggested that it is possible to compute optimal solutions in reasonable CPU time for most proteins and computation of optimal threadings is useful for improving practical performances of protein threading. Furthermore, there exist several important problems in bioinformatics, which are closely related to protein threading. Therefore, in this chapter, we overview algorithmic aspects of protein threading and related problems.

This chapter is organized as follows. First, we formally define the protein threading problem and show NP-hardness of protein threading. Then, after briefly reviewing heuristic methods, we describe three exact approaches for computing optimal threading: branch-and-bound approach (Lathrop & Smith, 1996), divide-and-conquer approach (Xu, Xu, & Uberbacher, 1998) and linear programming approach (Xu et al., 2003). Next, we briefly explain a variant of protein threading (protein threading with constraints) and related problems (comparison of RNA secondary structures and protein structure alignment). Finally, we conclude with future directions.

PROTEIN THREADING ALIGNMENT BETWEEN SEQUENCE AND STRUCTURE

As mentioned in the introduction, protein threading is a problem of computing an alignment between a target sequence and a template structure. First we define a usual alignment for two sequences. Let Σ be the set of amino acids (i.e., $|\Sigma|=20$). Let $s=s_1 \ldots s_m$ and $t=t_1 \ldots t_n$ be strings over Σ. An alignment between s and t is obtained by inserting gap symbols ('-') into or at either end of s and t such that the resulting sequences s' and t'