Tutorial proposal ISMB2001: Protein Structure Comparison and Structure Patterns, an Algorithmic Approach

Ingvar Eidhammer and Inge Jonassen

February 7, 2001

The tutors

The proposed tutorial will be given by Assoc. professors Ingvar Eidhammer and Inge Jonassen

Ingvar Eidhammer has been Associate professor in informatics since 1980, and has made course lectures and given lectures in bioinformatics since 1995.

Inge Jonassen has a PhD in bioinformatics from 1996, and has been associate professor in (bio-)informatics since 1999. He has given several lectures and talks at different seminars and conferences about finding patterns in sequences and structures, and also given the tutorial "Sequence Pattern Discovery Methods" at ISMB 1998.

Both have worked with methods for the discovery of patterns in biological sequences and in protein structures. The work has resulted in survey papers on both subjects (Brazma et al 1998, J Comp Biol 5, 279-305; Eidhammer et al 2000, J Comp Biol 7, 685-716) and the methods Pratt and SPratt for the discovery of patterns in respectively sequences and protein structures (Jonassen et al 1995, Prot. Sci. 4, 1587-1595; Jonassen 1997, CABIOS 13, 509-522; Jonassen et al 1999, Proteins 34, 206-219). The survey will be based on the survey paper (Eidhammer et al 2000) for the general parts and will use the SPratt method (Jonassen et al 1999) as en example method.

The presentation

Motivation

As we enter a post-genomic age, considerable effort is being expended on structural-genomics programmes with the aim of determining the structure of as many gene-products as possible. These data, combined with the, not inconsiderable volume of current structural data, will lead to the emergence of protein structure comparison as a critical technique — to aid not only in the understanding of the relationships between proteins in detail but also in the classification of the variety of structures into meaningful catagories.

Goals

The goals are to give the audience an understanding of the complexity of structure comparison, present a framework into which most methods can be placed, and present an overview of the various methods that are used.

Contents

We propose a framework for describing structure comparison and pattern discovery methods. For comparison of pairs of structures, the framework consists of four components. The first component is how the methods describe and represent the protein structures, e.g., at residue (atom) and secondary structure level. The second component refers to the solution space of the algorithm. It is common to search for equivalences where elements from each structure are paired up with elements from the other. Different constraints on the allowed equivalences can be imposed. For example, requiring co-linearity of the elements means that the equivalence becomes an alignment. The third component describes how different alternative equivalences are scored. Finally, the fourth component captures how the solution space of possible equivalences is searched for the equivalence with a high score. For example, scoring methods can be based on root-mean-square deviation and algorithmic approaches include dynamic programming, geometric hashing, and clustering. It is discussed how the different methods can be extended to the problem of comparing more than two structures. A method SPratt for the discovery of structure patterns is explained in some detail and it is emphasised how the use of a model/pattern in SPratt helps to avoid the laborious and indirect all-against-all pairwise comparisons.

Intended audience

The audience is expected to have basic knowledge of protein sequence comparison (such as dynamic programming) and sequence patterns such as regular expressions and profiles. Basic knowledge of protein structure is also expected, including coordinate and angular representation, and secondary structures.

Length/duration

Half day

Detailed outline

Introduction

In the introduction the structure comparison and motif discovery problems are explained in a wider context, including databases and searching these, structure classification and structure prediction. The framework for pairwise structure comparison methods is presented.

Description and representation

Which structure features are extracted and included in the structure descriptions depend on the comparison problem at hand. For example, if folds are to be classified into classes, features based on secondary structure element composition may be appropriate. As another example, if one wants to characterise active sites at a fine level, residue or atom level descriptions should be used. In general, a structure description consists of geometry (architecture), topology and properties. Alternative representations are discussed, including coordinates, angles, distance

matrices (for atomic representation), vectors (sticks) for secondary structures or backbone fragments. Formal requirements for representations are discussed.

Superposition

A superposition can be done when the correspondence between elements in two or more structures has been found, and it is often used to quantify the geometric similarity of the structures. The concept of superposition will be explained, and definitions will be given of coordinate and distance based RMSD.

Alternating superposition and alignment

Early methods for structure comparison includes methods that alternated between superposition and alignment operations. Given one alignment, a superposition can be found. Given this superposition, the best alignment is calculated, and if it differs from the preceding alignment, another cycle of superposition and alignment is performed unless a maximum number of cycles has been reached.

Double dynamic programming

An approach based on the dynamic programming algorithm used for sequence alignment, is double dynamic programming (DDP). DDP is related to the approach of alternating between superposition and alignment, but it allows the algorithm to consider several different superpositionings simultaneously (at a lower level) and let the alignment procedure use information from all of these (at a high level). Several variants of DDP will be explained.

Geometric hashing

Geometric hashing is a technique from computer vision that has been used as a basis for several methods. We will explain the basic ideas behind geometric hashing and how it has been applied to protein structure comparison.

Comparison by clustering

Many of the pairwise comparison methods first find seeds of compatible parts from the two structures, and then try to cluster seeds to discover equal substructures. Different clustering techniques are described, and two techniques for measuring the possibility for joining clusters are explained (use of relation or transformation).

Multiple structure comparison

It is shown which of and how the pairwise structure methods can be extended to the multiple case. A common way is to devise a strategy for performing several pairwise alignments, often so that the alignments can be between not only single structures but also between results from earlier alignments.

Local structure patterns

Local structure patterns are defined, and a method (SPratt) for discovering such patterns in a set of structures is explained. The method is not based on pairwise alignments. Instead an external model (pattern) is used to which each structure is compared and a method is explained for exploring the space of possible patterns in order to identify the most interesting patterns shared by sufficiently many structures. This approach avoids the laborious and indirect pairwise comparisons and it allows to discover patterns that may not match the full set of structures. We also discuss how pattern discovery methods can be combined with alignment methods.