Chapter 11
Declarative and Efficient Querying on Protein Secondary Structures

Jignesh M. Patel, Donald P. Huddler, and Laurie Hammel

Summary

In spite of the many decades of progress in database research, surprisingly scientists in the life sciences community still struggle with inefficient and awkward tools for querying biological datasets. This work highlights a specific problem involving searching large volumes of protein datasets based on their secondary structure. In this chapter we define an intuitive query language that can be used to express queries on secondary structure and develop several algorithms for evaluating these queries. We have implemented these algorithms in Periscope, which is a native database management system that we are building for declarative querying on biological datasets. Experiments based on our implementation show that the choice of algorithms can have a significant impact on query performance. As part of the Periscope implementation, we have also developed a framework for optimizing these queries and for accurately estimating the costs of the various query evaluation plans. Our performance studies show that the proposed techniques are very efficient and can provide scientists with interactive secondary structure querying options even on large protein datasets.

11.1 Introduction

The recent conclusion of the Human Genome Project has served to fuel an already explosive area of research in bioinformatics that is involved in deriving meaningful knowledge from proteins and DNA sequences. Even with the full human genome sequence now in hand, scientists still face the challenges of determining exact gene locations and functions, observing interactions

between proteins in complex molecular machines, and learning the structure and function of proteins through protein conservation, just to name a few. The progress of this scientific research in the increasingly vital fields of functional genomics and proteomics is closely connected to the research in the database community; analyzing large volumes of genetic and biological datasets involves being able to maintain and query large genetic and protein databases. If efficient methods are not available for retrieving these biological datasets, then unfortunately the progress of scientific analysis is encumbered by the limitations of the database system.

This chapter looks at a specific problem of this nature that involves methods for searching protein databases based on secondary structure properties. This work is a part of the Periscope project at the University of Michigan, in which we are investigating methods for declarative querying on biological datasets. In this chapter, we define a problem that the scientific community faces regarding searching on protein secondary structure, and we develop a query language and query-processing techniques to efficiently answer these queries. We have built a secondary structure querying component, called Periscope/PS^2, based on the work described in this chapter, and we also describe a few experimental and actual user experiences with this component of Periscope.

11.1.1 Biological Background

Proteins have four different levels of structural organization, primary, secondary, tertiary, and quaternary; the latter two are not considered in this chapter. The primary structure is the linear sequence of amino acids that makes up the protein; this is the structure most commonly associated with protein identification [321]. The secondary structure describes how the linear sequence of amino acids folds into a series of three-dimensional structures. There are three basic types of folds: alpha-helices (h), beta-sheets (e), and turns or loops (l). Because these three-dimensional structures determine a protein's function, knowledge of their patterns and alignments can provide important insights into evolutionary relationships that may not be recognizable through primary structure comparisons [307]. Therefore, examining the types, lengths, and start positions of its secondary structure folds can aid scientists in determining a protein's function [10].

11.1.2 Scientific Motivation

The discovery of new proteins or new behaviors of existing proteins necessitates complex analysis in order to determine their function and classification. The main technique that scientists use in determining this information has two phases. The first phase involves searching known protein databases for proteins that "match" the unknown protein. The second phase