NON-SEQUENTIAL AND FLEXIBLE PROTEIN STRUCTURE ALIGNMENT

By

Saeed Salem

An Abstract of a Thesis Submitted to the Graduate

Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the

Requirements for the Degree of

DOCTOR OF PHILOSOPHY

Major Subject: COMPUTER SCIENCE

The original of the complete thesis is on file
in the Rensselaer Polytechnic Institute Library

Examining Committee:

Mohammed J. Zaki, Thesis Adviser
Chris Bystroff, Member
Sanmay Das, Member
Badrinath Roysam, Member

Rensselaer Polytechnic Institute
Troy, New York

July 2009
(For Graduation August 2009)

ABSTRACT

Proteins are macromolecular organic compounds. They are involved either directly or indirectly in all the biological processes in living organisms. Among the major biochemical functions of proteins are binding, catalysis where protein enzymes catalyze chemical reactions, molecular switching to control cellular processes, and serving as structural elements of living systems.

Structural similarity between proteins provides us insights into their evolutionary relationships when there is low sequence similarity. Of particularly interest is the structural similarity between proteins which are remotely homologous where sequence similarity is not strong enough to indicate functional relationship. Moreover, non-sequential alignments highlight the relationship between proteins that are related through circular permutation [1], or proteins that evolved from different ancestors owing to convergent evolution [2].

Most of the algorithms for structural alignment are inherently limited to sequential alignments and therefore cannot capture non-sequential alignments. Another limitation of the existing structural alignment methods is that they only report rigid alignments and thus cannot capture flexible alignments where one protein goes through a conformation change to become similar to the other protein. The ability of finding non-sequential and flexible structure similarity between proteins has important implications for enhancing our understanding of protein structure and the protein folding process.

In this thesis, we present two approaches for addressing the problems of non-sequential protein structural alignment and flexible alignment; more specifically, we introduce two algorithms, namely SNAP and FlexSnap. The SNAP algorithm is an iterative superposition-based algorithm that reports both sequential and non-sequential rigid alignments, from an initial superposition. The initial superpositions are essentially similar well-aligned small substructure pairs, called Aligned Fragment Pairs (AFPs). Each AFP defines a superposition which is used to align the two proteins. A binary similarity scoring matrix is computed from the spatial dis-

tances between all pairs of residues, one residue from each protein. For sequential alignment, we use a sparse dynamic-programming algorithm for finding the optimal sequential chaining of well-aligned segments in the similarity matrix. The problem of optimal non-sequential chaining is computational expensive and thus we propose two greedy approaches. We assessed the quality of SNAP alignments by measuring their agreements with the manually curated reference alignments in two challenging datasets, namely, SISY and RIPC [3]. The SNAP alignments had the highest average agreement as compared to the alignments reported by algorithms such as DALI, CE, STRUCTAL, SARF, and SCALI. Moreover, when used as a topology level classifier on a dataset of 4410 protein pairs selected from the CATH database [4], its classification was both highly sensitive and highly selective; moreover, the SNAP algorithm was competitive to several state-of-the-art alignment methods.

Our second contribution is the FlexSnap algorithm which reports rigid and flexible alignments, both sequential and non-sequential. The FlexSnap algorithm assembles the alignment from small well-aligned fragments (AFPs) and introduces hinges when there is a significant gain in the alignment score. We demonstrate the efficiency and effectiveness of FlexSnap by comparing its alignments to those reported by FlexProt and FATCAT, the two most widely used algorithms for flexible alignments. Moreover, FlexSnap can report rigid alignments. The rigid alignments of FlexSnap are competitive to the state-of-the-art algorithms (SARF, SCALI, MultiProt) for non-sequential alignments, as evident by its high agreements with the manually curated non-sequential alignments in the RIPC dataset. A unique feature of the FlexSnap algorithm is that it is the only alignment algorithm which reports flexible non-sequential alignments.

For our future work, we propose a comprehensive analysis of non-sequentiality and flexibility in the protein database. We are interested in studying how some proteins memberships in the CATH and SCOP classifications would be altered when we consider non-sequential and flexible structural alignments in inferring the relationship between proteins.