COMPUTER SCIENCE DEPARTMENT
FACULTY OF COMPUTERS AND INFORMATION SCIENCES
AIN SHAMS UNIVERSITY

Intelligent Technique for Indexed Search of Protein Structures

A Thesis Submitted to Computer Science Department,
Faculty of Computer and Information Science,
Ain Shams University, Cairo, Egypt.

In partial fulfillment of the requirements for
the degree of Master in Computer Science

By
Ahmed Salah Mohammed
B.Sc. in Computer Science,
Faculty of Computer and Informatics, Zagazig University.
Demonstrator at Computer Science Department, Faculty of Computer and Informatics, Zagazig University, Zagazig, Egypt.

Under Supervision of
Professor Adbel-badeeh M.Salem
Professor of Computer Science
Faculty of Computer and Information Sciences,
Ain Shams University.

Professor I. M. El Henawy
Professor of Computer Science
Faculty of Computers and Informatics,
Zagazig University.

Dr. Tarek Fouad Gharib
Associate Professor of Information Systems
Faculty of Computer and Information Sciences,
Ain Shams University.

2009

Acknowledgement

In the first place, I would like to record my gratitude to ALLAH, the most gracious and the most merciful, that helps me in completing this work. I hope that may ALLAH accept that work as a good work.

I would like to express my deep and sincere gratitude to Professor Adbel-Badeeh M. Salem. His wide knowledge and his logical way of thinking have been of great value for me. His understanding, encouraging and personal guidance have provided a good basis for the present thesis. Seeing him is the best motivation for me to just achieve a tiny part of his achievements in science.

I am strongly grateful to Professor Ibrahim M. El-Henawy for his continuous encouragement and support during the thesis period. His advices have lightened my way and without him I was not able to pass the tough obstacles I faced.

I wish to express my warm and sincere thanks to Assoc. Prof. Dr. Tarek Gharib for the immeasurably valuable guidance and support I have received from him. The effort he puts into every ounce of the heavy task of advising me is astounding. I appreciate his continuous guidance, expert suggestions and beneficial discussions. He was always with me to improve the precision and clarity of my writing and speech. It would have been practically impossible to realize this work without his support. I am grateful to have had the opportunity to work with him.

I am greatly indebted to my parents, they were more than my supervisor, and they were a kind people that gave me the hand starting by the work in this thesis and ending by the completion of it. Their continuous and strong support and encouragement had very profound effects on me.

Special thanks devoted to my friends Hany Abd El-Maojod, Abd El-Wahed Khames, Marco and Ahmed Enany for their support and just in time help. My special gratitude is due to my brothers Mohamed and Hussein, and my sister Mona for their continuous support and prayers.

Publications

1. **Tarek F. Gharib, A. Salah, I. M. El Henawy and Abdel-Badeeh M. Salem "Protein Structure Searching using Suffix Arrays" In The International Conference on Bioinformatics & Computational Biology (BIOCOMP), pp. 688-691, 2008.**

2. **Tarek F. Gharib, A. Salah and Abdel-Badeeh M. Salem "PSISA: an Algorithm for Indexing and Searching Protein Structure using Suffix Arrays" In The WSEAS International Conference on COMPUTERS, pp. 775-780, 2008.**

Abstract

Searching for structural similarities of proteins has a central role in bioinformatics field. Most tasks of bioinformatics depend on investigating the homologous protein's sequence or structure. In turn searching for structural similarities has a critical role in many applications like prediction of protein's structure and functions, classification of proteins and drug design and discovery. Proteins with homologous sequence or structure can be concluded to have a common ancestor which is helpful for better understanding of life tree.

Protein Structure Indexing Using Suffix Array (PSISA) is a new technique that provides the ability to retrieve similarities of proteins based on their structures. Indexing the protein structure is one approach of searching for protein similarities. In this thesis we present a new technique for indexing and searching protein structure using suffix arrays. We start by converting protein structure into a sequence by extracting local feature vectors; normalization is applied to these vectors components. These normalized vectors are converted into a sequence. Sequence is indexed using the suffix array structure, which is used effectively in the searching process to retrieve proteins with similar structure. Proteins with high structural similarities are ranked according to their alignment score against the query protein.

The experimental results, which are based on the structural classification of proteins (SCOP) dataset, show that our method outperforms existing similar methods in memory utilization. Our results show an enhancement in the memory usage with factor 35%.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Cα	Central Carbon Atom or C Alpha
GSA	Generalized Suffix Array
GST	Generalized Suffix Tree
NCBI	National Center for Biotechnology Information
NIH	National Institute of Health
PDB	Protein Databank
PIR	Protein Information Resource
ProGress	Protein Grep by Sequence and Structure
PSI	Protein Structure Indexing
PSISA	Protein Structure Indexing using Suffix Array
PSIST	Protein Structure Using Suffix Tree
RCSB	Research Collaboratory for Structural Bioinformatics
RMSD	Root Mean Square Deviation
SCOP	Structural Classification Of Protein
SSEs	Secondary Structure Elements

Chapter 1

Introduction

1.1 Motivation

The rapid growth of the Protein Databank (PDB) current holdings, > 50000 proteins at the first quarter of 2009, raises the need for new tools that perform proteins similarity searching to clarify the similarities in the three dimensional structures between related or similar proteins. Most of these tools search the protein structure rather than protein sequence. Proteins that have similar functionality have similar structure besides it might have not the same primary structure [3]. But if a set of proteins have the same primary structure then they will have the same functionality. So the importance of the protein's 3-D shape comes from the fact that the function of protein depends on its shape rather than its sequence (primary structure). Proteins Primary structure is a sequence of letters that states the amino acid in this protein. Protein Secondary structure is a 3D description of the proteins as a sequence of local segment of proteins.

Several databases that hold information about the protein exist nowadays; some of them provide information about the protein sequence only, while others provide information about the secondary and tertiary structure. The widely used databases are for the secondary structure since the functions of protein can be predicted from its structure better than its sequence. Genbank, the National Institute of Health (NIH), is built by National Center for Biotechnology Information (NCBI), SWISS-PORT, Protein Information Resource (PIR), Protein Data (PDB), and

Structural Classification Of Protein (SCOP) are examples for the most famous proteins databases.

Searching the protein structure has another problem, besides the rapidly growing rate of proteins in PDB, which is the complexity. The protein structure alignment is a NP-hard problem. Many methods were proposed to solve this problem.

Searching for similarities in proteins database is a problem approached by several ways. Firstly, it was approached by sequence alignment [9]. Secondly, and because of the link between protein structure and its functionality, the problem is approached by structural alignment. This means that we can search for partial structure similarities between proteins.

Several approaches were proposed to solve structural alignment problem. Pair-wise structural alignment algorithms can perform the alignment at the secondary structure elements (SSEs) level or intra and inter-molecular atomic level [4, 11]. Another class of algorithms performs the alignment at residue level [12]. The last class is based on geometric hashing, which can be applied at both SSE and residue level. New algorithms are based on multiple structure alignment [13] to enhance the performance. However, pair-wise alignment is not feasible for large databases with more than few thousands of proteins. PSI is an algorithm used to reduce the searching space [6]. Database searching using information retrieval techniques [7] is another approach to be used instead of pair-wise alignment. A final approach to solve this problem is to index the proteins database, indexing the Protein Structure Indexing using Suffix Tree (PSIST) [5], CTSS [14] and Progress [15] are examples of such approach.

Protein structure index (PSI) method prunes unpromising protein for the given protein query. It is based on extracting feature vector for each protein in database then indexing it using the R* tree. R* trees are used to prune the search space to be used by VAST structural alignment algorithm, this reduction in search space results in reducing the searching time [6].

Protein Structure Indexing using Suffix Trees (PSIST) convert the 3D structure to a sequence by extracting feature vector for each protein in the database. The feature vector includes the distance between each two residues and the angle between their planes. Each protein is described as a list of vectors. Each vector is converted to a unique symbol, that maps the list of vectors to a sequence (String) that can be fit in a suffix tree which is an indexing structure that speedup the searching process [5].

1.2 Objectives

We can summarize the objectives of the thesis in the following points:

1. Designing a proposed algorithm for searching the structural similarities between proteins.
2. Investigating the usage of suffix arrays as indexing structure to index the protein structure.
3. Comparing the proposed algorithm performance with previous best known algorithms which use indexes to search proteins structural similarities.
4. Building a tool for searching protein structural similarities using the proposed algorithm. The tool is given a query protein, PDB file, then search all known protein, proteins database, to find all proteins that have similar structure to the query protein.

1.3 Contributions

Our research shows up four contributions. They are as following:

1. We proposed an algorithm that can search for structural similarities between proteins which is called Protein Structure Indexing using Suffix Array (PSISA).
2. Our study gives the insights on how to perform proteins structural searching with different approaches.
3. The experimental results of the proposed algorithm show that we presented a basis for building memory optimized protein structural similarities searching tool for whom gives the memory consumption factor the highest priority between other computing resources.

1.4 Organization of the Thesis

Chapter 1 gives a brief description about the problem this thesis supposed to solve. Also it lists the main contributions of this thesis.

Chapter 2 presents a biological background related to the point of research. The chapter discusses some biological concepts and facts like proteins, amino acids, different protein structures, and proteins databases.

Chapter 3 presents the suffix array data structure and its role in bioinformatics fields. Bioinformatics field is introduced first in terms of its roles and how it is related to computer science field. Finally, it elucidates the suffix array data structure.

Chapter 4 provides an overview of the previous work. Firstly, it lists the different approached which were proposed to solve the problem of searching for structural similarities between proteins and explains each approach briefly.

Chapter 5 provides detailed explanation of our proposed algorithm to solve the problem. Each section in the chapter explains one step of the proposed algorithm. Algorithms are provided also with sections.

Chapter 6 explains experiments which are used to test our proposed algorithm. This chapter provides the details for preparing the dataset and measurements used to evaluate the proposed algorithm.

Chapter 7 concludes the thesis with a summary of the main results and suggests some directions for future work.