Available online at www.sciencedirect.com

ScienceDirect

Pattern Recognition 39 (2006) 2494–2505

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

A coprocessor architecture for fast protein structure prediction

Rahim Khoja, Mehul Marolia, Tinku Acharya, Chaitali Chakrabarti*

Department of Electrical Engineering, Arizona State University, Tempe, AZ, USA

Received 1 September 2005; received in revised form 9 December 2005; accepted 12 December 2005

Abstract

Predicting the protein structure from an amino acid sequence is computationally very intensive. In order to speed up protein sequence matching and processing, we present a novel coprocessor architecture for fast protein structure prediction. The architecture consists of systolic arrays to speed up the data intensive sequence alignment and structure prediction steps, and finite state machines for the control dominated steps. The architecture has been synthesized using Synopsys DC Compiler in 0.18 micron CMOS technology and details of its area and timing performance have been provided. A procedure to develop architectures with area-time trade-offs has also been presented.

Keywords: Protein structure prediction; PSIPRED; PSI BLAST; Neural network; VLSI architecture

1. Introduction

In the last century, biologists have made vast progress in studying the characteristics and complexity of living organisms. In this century, significant advances have been made in understanding the cellular processes, in terms of molecular force interactions. The advances in computational molecular biology (also called *bioinformatics*) is likely to help answer questions such as “Why is one person different from another”, or “Why is one person more susceptible to a particular disease than another”, or “What is the cure of AIDS or cancer”. Today, bioinformatics is an interdisciplinary area of study involving diverse fields such as biology, biochemistry, medicine, genetics, computer science, information technology, mathematics, statistics, physics, etc. that has very high potential [1,2].

A better understanding of the cell structure at the molecular level, and knowledge of the underlying chemical structures and functions has led to a special interest in the study of protein molecules. Proteins are organic polymers built by polypeptide chains of amino acids inside a cell. Many of the hormones and enzymes in our body are proteins, and the functionality and characteristics of the proteins are responsible for many of the cellular functions. Proteins are the most varied group of molecules involved in biochemical processes and understanding them is a core area of research in bioinformatics.

In order to understand the structure of a protein, we start with the DNA (deoxyribonucleic acid) of an organism that is a chain of four types of bases Adenine (A), Cytosine (C), Guanine (G), and Thymine (T). The DNA is transcribed to produce m-RNA (messenger ribonucleic acid), which is then translated to produce protein in our body. Each nonoverlapping triplet (codon) of DNA bases is treated as an amino acid and the translation process maps the triplets of four bases (A, C, G, T) to the set of 20 amino acids which form the basic building block of proteins [2]. The length of a protein, represented in the form of a linear chain of amino acids, can vary from 10 to 1000 s. The molecular interactions cause the linear chain of amino acids to fold and twist, resulting in different functionalities.

The protein structure can be predicted by matching an amino acid query sequence with existing sequences in biological databases. Since a query sequence ranges from 50 to 400 amino acids and the databases have ~500 thousand amino acid sequences, the prediction problem is

* Corresponding author. ASU Main Campus, Ira A Fulton School of Engineering, Department of Electrical Engineering, Tempe, AZ, 85287-5706, USA. Tel.: +1 480 9659516; fax: +1 480 965 8325.

E-mail address: chaitali@asu.edu (C. Chakrabarti).

0031-3203/$30.00 © 2006 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2005.12.006

computationally very intensive, and processing this enormous data quickly is an area of active research.

According to Moore's law of genomics, "The Protein data Bank (PDB) of known structures will continue to double every three years and the number of sequences in want of a predicted structure will continue to double every few months" [3]. Thus, in order to process the large volume of protein data, it is becoming increasingly important to develop faster and simpler techniques for protein structure prediction. Experimental ways of finding the protein structure is very time consuming, expensive and does not meet the current demands. Also, they do not provide insight into how the proteins fold due to different molecular interactions and why they fold into a particular structure only. Methods such as ab initio [4] that require exhaustive molecular force interaction study to develop the protein structure, fails to provide the necessary speed. Hence, the trend has shifted towards developing structure prediction methods and modeling techniques which can predict the corresponding three dimensional protein structure from the amino acid sequences.

There are numerous structure prediction methods [5–7] that differ in the type of approach as well as the degree of accuracy. Most of these methods are based on homologous template matching and use databases of existing predicted structures to develop the three dimensional structure for the amino acid sequence under study. We have selected PSIPRED [8] as the preferred method of prediction because it is simple and demonstrates a high level of accuracy.

The PSIPRED algorithm predicts the secondary structure of a protein by using neural networks to process the position specific scoring matrix (PSSM) generated by position specific iterated—basic local alignment search tool (PSI-BLAST) [9]. The steps involve scanning the database to get a list of highly probable hit locations which are then extended with dynamic programming methods to align with the query sequence. The process of generating the alignment profile is iterated, and the PSSM matrix is used by the neural network to predict the secondary structure.

The computationally intensive nature of the PSIPRED algorithm makes it necessary to develop a special purpose VLSI architecture that would serve as a coprocessor. In this paper, we present a specialized architecture that uses a combination of systolic arrays for sequence alignment and structure prediction and finite state machines for the remaining steps. This is an extension of our earlier architecture presented in Ref. [10]. The coprocessor architecture has been implemented using VHDL and synthesized using Synopsys DC Compiler in 0.18 micron CMOS technology. The synthesized architecture requires 663 940 units of gate area and 346 KB of memory, and can be clocked at 100 MHz. It can predict the secondary structure of a query sequence of length $\sim$ 150 amino acids using a database of 135 million amino acids in less than 10 s. To the best of our knowledge, this is the first VLSI architecture for protein structure prediction proposed in the literature to date.

To better understand the performance of the proposed architecture, we have analyzed it with respect to datapath size, memory and timing requirements. We have suggested a procedure for allotment of area and memory resources to achieve the desired timing. We have also considered the resource requirements to facilitate processing larger query sequences and larger protein databases.

The rest of the paper is organized as follows. We first explain the protein structure prediction algorithm in Section 2. Then we present the details of the proposed architecture in Section 3. We highlight the synthesis results for the architecture, the tradeoffs between area (including memory requirement) and time, and the effects of query sequence lengths and database sizes on the architectural performance in Section 4. We make some concluding remarks in Section 5.

2. Protein structure prediction technique

Protein is a biopolymer of amino acids arranged in a specific sequence that fold into a three-dimensional shape. The shape of a three dimensional protein molecule is determined by the sequential order of the amino acids. The folded structure can be constructed with three types of substructure (secondary structure): helix structure (α), sheet structure (β) or other coil structures (γ).

Of the various secondary structure prediction methods that exist today [4–7], PSIPRED is a method that demonstrated average accuracy of 77.3% at CASP3 (third critical assessment of structure prediction) meeting [8]. We selected this method because of its simplicity and the capability of giving the highest level of accuracy published to date.

The PSIPRED algorithm takes a sequence of amino acids as a query sequence input and predicts the corresponding secondary structure. It is implemented in two steps: (1) PSI-BLAST, for multiple sequence alignment followed by updating a sequence profile iteratively to generate position specific scoring matrix (PSSM) and (2) secondary structure prediction and filtering using neural networks. The key components of the algorithm and their interactions are shown in Fig. 1.

2.1. PSI-BLAST

BLAST (Basic Local Alignment Search Tool) [11] is a heuristic tool based on dynamic programming [12] to find sequence similarities in protein and DNA sequences. The central idea of BLAST is that any statistically significant alignment between query and database sequences must have at least one word pair of length w scoring above a threshold T. Once the location of this word pair is known, it is extended on either side to obtain high scoring pair (HSP) of aligned sequences with a score S. This score is obtained from the BLOSUM-62 [13] substitution matrix by comparing the aligned amino acids in the HSP. PSI-BLAST [9] is an iterated version of BLAST which generates a position