Volume 6, Issue 5, May 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Review on FPGA Accelerator for Protein Secondary Structure Prediction Based on the GOR Algorithm

Sunnit Kaur
Mtech (CSE) ACET,
Manawala Asr, India

Er. Rajbir Singh
Associate Professor, ACET,
Manawala Asr, India

Abstract— Protein is the basic building block. Protein is a substance that has amino acid,compound found in many foods.The most important and succesful used methods is GOR algorithm . This method has been widely used as an efficient analysis tool to predict secondary structure from protein sequence.GOR algorithm is based on probability parameters derived from empirical studies of know protein tertiary structures solved by X-ray crystallography.Improvs upon the Chu-Fasman method however the execution time is till intolerable with the growth.But .nowdays one promising application accelerator to accelerate bioinformatics algorithm by exploiting fine-grained custom design is FPGA chips.

Keywords— Protein, GOR, FPGA, Task Pipeline

I. INTRODUCTION

Protein is an important molecule that peforms a wide range of function.Protein is a substance that has amino acid ,compound found in many foods.Generally from tertiary structure function of protein molecue can be derived.By protein nuclear magnetic resonance and X- ray diffraction the most structures are determine, but these methods are time consuming and very expensive. The growth of protein sequence databae,such as the EMB-EB1 protein database which has doubled in size every 1824 months for the decade and now it contain 10867,798 sequence entries, comprising 3502326038 amino acid. In the protein data bae (PDB) , the number of know structures is just more than 65,000 at present.In PDB and the number of know protein sequence the gap between them grows conytinously at an incredible rate .By computational metod, the prediction of protein structure and function from amino acid becomes a most important problem in modern biology and bioinformatics.The ultimate goals of protein science is the prediction of tertiary structure .Various methods ,such as [3] ab initio modelling or [4] protein fold recognition have been presented for protein 3-Dimensional structure prediction from amino acid sequence , but we don't use these methods are too complex and even not feasible in some conditions, Its easier to predicit fundamental elemtents .The elements of secondary structure of proten :alpha-helics ,beta-sheet ,coil insead of predicting the full 3D structure directly .In protein 3D structure and 2D structure the knowledge of elements can be easily observed as an input for protein tertiary structure.

To predict 2 Dstructure from amino acid sequence several approaches have been developed .

1. The Choufasman 2. GOR
2. HP 4. ANN
5 Machine Learning 6. NNM

But from all theses methods one of these methods one of the earliest and most successful method for secondary structure prediction is the GOR methd.

At present ,according to our knowledge ,there is no parallel implementation of the GOR algorithm running on general purpose CPU for protein 2D structure prediction .Usualy , to accelerate bioinformatics application such as pairwise / multiple sequence alignment , database searching and RNA secondary structure prediction high performance parallel computer are widely used.

A parallel approach to accelerate clustalW using MPI on traidtonal parallel computer in 2002 done ny Kuo-Bin li.This implementation achieves a speed up of 4.3x using 16 processors.

Christopher Dwan et.al presented a parallel

Implementation of the clustalW package on the SGI Alix XE cluster with 32 CPUs .For Dna and protein alignment they reported the speed ups of 7.2 x and 17.1 x.

Yu-Lun Ku et.al implemented parallel mpiBLAST for database searching application on a symmetric multiprocessors cluster which consist of 16 CPUs.

The BLASTp algorithm also introduces in this paper which is a parallel implementation for protein database searching and they reported a 23.6 x speedup over the NCBI. BLAST version for searching 1000 proetin sequence against the 'NR' dataset from NCBI, which consists of approximately 2 GB of amino acid sequences.

Large scale supercomputers like the IBM Blue Gene also used to accelerate the BLAST algorithm.

A complete fine-grained parallel hardware implementation on FPGA to accelerate the GOR IV package for 2 D protein structure prediction has been proposed in this paper.

Partition the parameter table into small section and access them in parallel to improve the computing efficiency . Data reuse schemes are applied to minimize the need for loading data from external memory .In order to overlap sequence load ,computing ,back writing the whole computation structure is carefully pipeline.The GOR algorithm accelerator is implemented on single FPGA chip.Power consumption is only about 30% of that of te current general purpose CPU.

II. OVERVIEW OF GOR ALGORITHM

The first major methods proposed for protein secondary structure prediction from sequence is GOR program.The original version (GOR- IV) was released in 1978 by Garnier,Osguthorpe and Robson .The GOR method is the use of the information theory and Bayesian statics method to relate the amino acid sequence to the potein secondary structure . Unlike choufasman ethod, the GOR method takes the conditional probability of amino acid to form a secondary structure but also it takes into account the prospensities of individual amino acid.By adopting larger structure databases and more exact statical model for computing information function from the past twenty years,the GOR method has been improved GOR- IV analyzes sequences to predict wheather it is alpha heix,beta sheet ,coil.The prediction of protein secondary structure can be done at each postion based on 17 amino acid sequence windows to consider the information of local segement .The most crucial change between the GOR IV and GOR V, GOR- V is online at the web based protein secondary structure internet prediction server.

The major change was the inclusion of evolutionary information using PSI-blast to increase the information content for improved discrimination among secondary structure ,which combines information theory,Bayesian statistics and evolutionary information.It reaches an accuracy of prediction Q3 of 73.5%.Careful surveys has done between various methods,we choose the GOR- IV as the candidate for fine-grained parallel implementation .The GOR runs with a single protein sequence as input.

The kernel algorithm executes in three steps which are defined as:-

1. It predicts the 2D structure of the input protein sequence based on the Information theory combined with the Bayesian statistics .For each input amino acid ,it computes three probability values and select the largest one for judging the current amino acid.
2. The later two stages perform a scanning procedure to correct the secondary structure generated by first stage.
3. The GOR IV gives the output consists of protein sequence and the predicted secondary structure in rows ,H-helix,E- ectended or C-coil with probability values.

III. LITERATURE SURVEY

Ibrahim Darwish1, Amr Radi2, Salah El-Bakry3 and El-Sayed M. El-Saye (2015) [1]disscussed that Precise prediction of protein secondary structures from the associated amino acids sequence is of great importance in bioinformatics and yet a challenging task for machine learning algorithms. .

Hanan Hendy, et. al(2015)- [**2**]considered that Protein secondary structure prediction has been and will continue to be a rich research field. This paper presents a technical study on recent methods used for secondary structure prediction using amino acid sequence. The paper shows different approaches for predicting the protein structures that showed different accuracies that ranged from 50% to over than 90%. The most commonly used technique is Neural Networks. However, Case Based Reasoning and Mixed Integer Linear Optimization showed the best accuracy among the machine learning techniques and provided accuracy of approximately 83%.

Ravdeep Singh ,**Prof Rajbir Singh, et.al.(2015)** –[**3**]has proposed the Machine learning technique is introduced as a method for the classification of proteins into functionally distinguished classes. Protein function classification is one of the most important problems in modern computational biology. Studies are conducted on a number of protein classes including RNA-binding proteins; protein homodimers, proteins responsible for drug absorption, proteins involved in drug distribution and excretion, and drug metabolizing enzymes.

Shivani Agarwal, **et.al. (2014)**-[**4**] considered that the tertiary structure of protein is difficult to predict accurately directly from a protein sequence. The intermediate step is required to predict the structure which project the one dimensional structure into the three dimensional structure.

Anureet Kaur Johal, Prof. Rajbir Singh (2014)-[5] explained that how To solve the Protein folding problem is one of the most important task in computational biology. Protein secondary structure prediction is key step in prediction of protein tertiary structure. There have emerged many methods such as meta predictor based, neighbor based and model based methods to predict protein structure.

IV. METHODS

System architecture

The prediction of protein 2D structures platforms consists of a reconfigurable algorithm accelerates and a PC.The accelerator receives input protein data stream of length N with 3 bit binary encoding and datavase of 267 sequences with know secondary structure,then executes the 2D structure prediction ans return to host for display .The core of GOR algorithm accelerator is composed of

1. GOR Control Module:- which is responsible for initializing the computing Pipeline.
2. Data Back Writing :- controller is responsible for buffering protein .
3. Computing Pipline :-assigning protein sequence dynamically to the computing pieline.

2D structure prediction of protein which is perform by GOR computing consists of 3 submodules

1. To predict preliminary secondary structure for each amino acid and the latter two stages for correction.
2. Adjoining stages between data buffers are used for delivering middle results.
3. The complete GOR algorthim procedure in data driven can be perform by three computation modules.

The preliminary secondary structure of sequence with TD#1 is generated by the prediction module,it will be delivered to next stage

FPGA Implementation

The conformation for each amino acid involves looking up parameters tables to get conditional probability values. Three probability values of fundamental structure of current amino acid can be calculated by –

First converse the address according to the residue type and relative position in computing window ,then lookup the table to obtain the probability values.

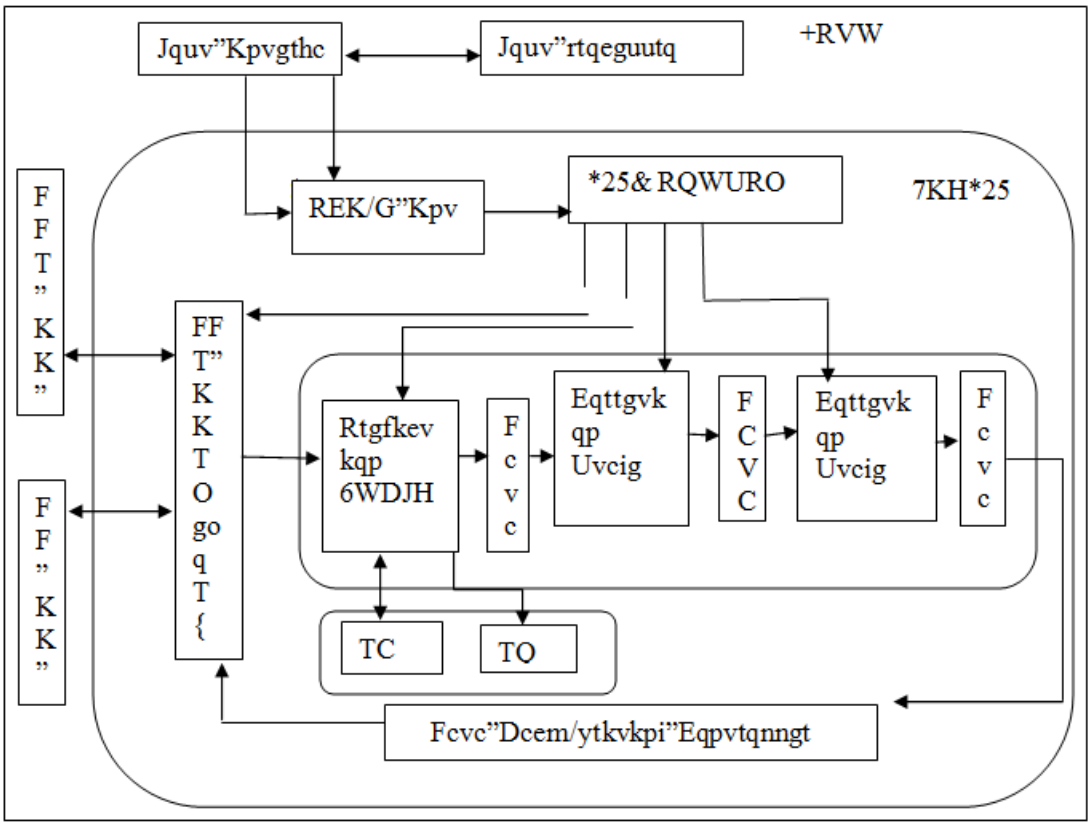

Figure 1. Structure of the GOR algorithm.

The number of query operations for predicting 2D structure of protein with length N is 272XN.

Additionaly , the address interval between adjacent two query operations equal operation equal 21*21*138*8 byte in infor_pair table .The serial look-up order implementation in GOR-4 software results in a great deal of small granularity discotinous memory

The task pipeline

The task pipeline is similar to software version. In the GOR-IV algorithm accelerator also comprises a prediction stage and two correction stages as shown in Figure 2. Implementation of a competed computing pipeline to accelerate the critical part, preliminary prediction, in the GOR program since most of the execution time, more than 99%, is spent in this stage.

Decomposition of the whole preliminary prediction module is done by 109 pipeline levels and then Data Buffers are used to store the prediction results .Our experimental results show that the execution time of the three stages in the GOR algorithm accelerator is basically balanced generally. The first stage has powerful computing capability as the parallel table look-up strategy is adopted and there is no obvious performance bottleneck in our pipeline architecture because the balance of pipeline levels was carefully considered. Moreover, the overhead of the prediction stage for an input protein sequence with length N is fixed. However, because of the uncertain execution time of back-end processing for correction, which is closely related to the fundamental conformations in preliminary prediction results, and the serial correction procedure. back-end process capability can'tcatch up with the throughput of the prediction unit.

After computing the various stages we can calculate the execution time .We considers the various foldind of protein structure. The power consumption is only about 30%

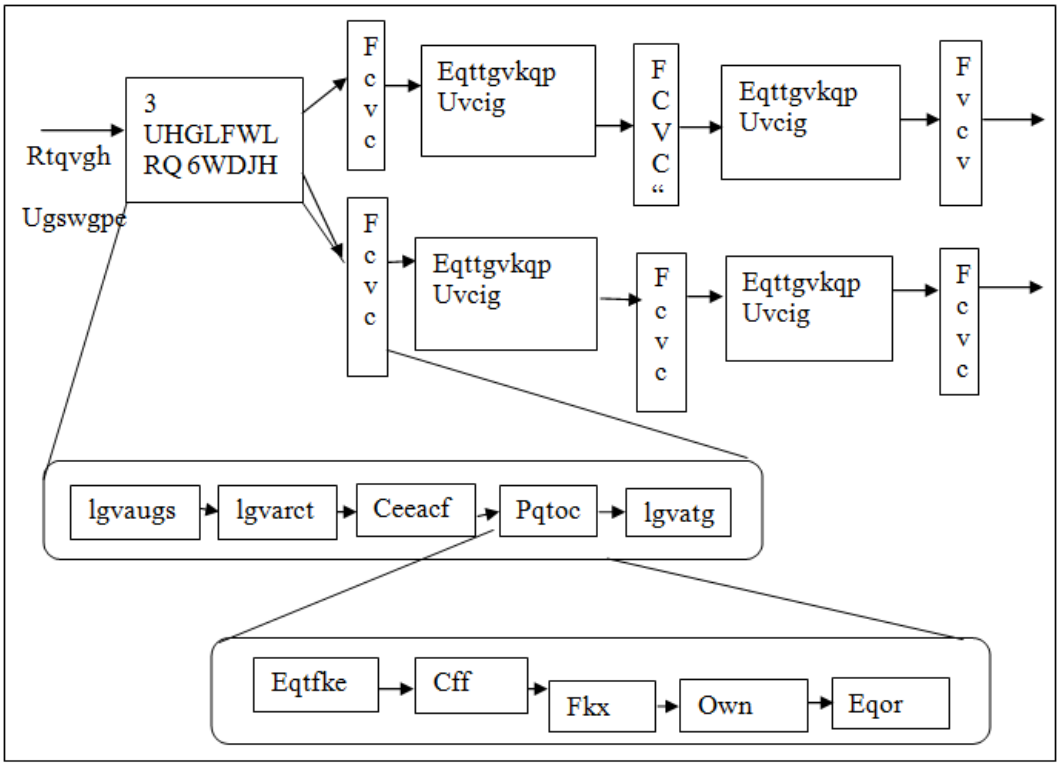

Fig 2. The task pipeline in GOR algorithm

V. CONCLUSIONS

Protein is the basic building block.The main problem in proteins are Protein Folding .This problem has become a central reseach area .The ultimate goal is the prediction of 3 D structure.The feasible intermediate step in th direction is the prediction of 2D structure from protein sequence.The GOR approach is the most efficient approach it nt nly take the conditional probability alsp the propemsities.It is the most successful approach so to predict the 2D strtucture is efficient in classical sense.Acomplete fine grained parallel hardware implementation is proposed on FPGA o acceltrate the GOR packge for 2D protein structure prediction.To increase the performance we partition the parameter table into small sectionsso that paraalely access can take place..Carefully pipelined of whole computational structure is taken in order to overlap sequence load,back-writing and computing.Complete GOR desktop is implemented on a single FPGA chip.The power consumption is only about 30%Our design is the first FPGA implementation for accelerating the protein folding based on GOR algorithm .FPGA algorithm accelerator not only suggest a huge potential performance for parallelizing 3D structure prediction of protein but also can be applied to a desktop platform to resolve large scale bioinformatics.

REFERENCES

[1] Ibrahim Darwish1, Amr Radi2, Salah El-Bakry3 and El-Sayed M. El-Sayed4 (2015) **"***Protein Secondary Structure Prediction Using Artificial Neural Network Implemented on FPGA",* International Journal of Bio-Medical Informatics and e-Health Volume 3, No.1, January - February 2015.

[2] Hanan Hendy, Wael Khalifa, Mohamed Roushdy, (2015) "*A Study Of Intelligent Techniqous For Protein Secondary Structure Prediction"* International Journal "Information Models and Analyses" Volume 4, Number 1, 2015.

[3] Ravdeep Singh ,Prof Rajbir Singh, et.al.(2015) "*Improved Protein Function Classification Using Support Vector Machine"* International Journal of Computer Science and Information Technologies, Vol. 6 (2) , 2015, 964-968.

[4] Shivani Agarwal, et.al. (2014) " *Prediction of Secondary Structure of Protein using Support Vector Machine"* International Journal of Computer Applications® (IJCA) (0975 – 8887)

[5] Anureet Kaur Johal, Prof. Rajbir Singh (2014)" *Protein Secondary Structure Prediction Using Improved Support Vector Machine And Neural Networks"* International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume 3 Issue 1, January 2014 Page No. 3593-3597

