AIS-PSMACA: Towards Proposing an Artificial Immune System for Strengthening PSMACA: An Automated Protein Structure Prediction using Multiple Attractor Cellular Automata

P.Kiran Sree[1] and Dr. Inampudi Ramesh Babu[2]

[1] BVCEC

Received: 7 December 2012 Accepted: 31 December 2012 Published: 15 January 2013

Abstract

Predicting the structure of proteins from their amino acid sequences has gained a remarkable attention in recent years. Even though there are some prediction techniques addressing this problem, the approximate accuracy in predicting the protein structure is closely 75

Index terms— protein structure, cellular automata, MACA.

1 Introduction

roteins are molecules with macro structure that are responsible for a wide range of vital biochemical functions, which includes acting as oxygen, cell signaling, antibody production, nutrient transport and building up muscle fibers. Specifically, the proteins are chains of amino acids, of which there are 20 different types, coupled by peptide bonds [2]. The three-tiered structural hierarchy possessed by proteins is typically referred to as primary and tertiary structure. This is because the higher-level and secondary level [1], [2] structures determine the function of the proteins and consequently, the insight into its function can be inferred from that.

As genome sequencing projects are increasing tremendously. The SWISS-PORT databases [3], [4] of primary protein structures are expanding tremendously. Protein Data Banks are not growing at a faster rate due to innate difficulties in finding the levels of the structures. Structure determination [5], [6] procedure experimental setups will be very expensive, time consuming, require more labor and may not applicable to all the proteins. Keeping in view of shortcomings of laboratory procedures in predicting the structure of protein major research have been dedicated to protein prediction of high level structures using computational techniques. Anfinsen did a pioneering work predicting the protein structure from amino acid sequences [6], [7]. This is usually called as protein folding problem which is the greatest challenge in bioinformatics. This is the ability to predict the higher level structures from the amino acid sequence.

By predicting the structure of protein the topology of the chain can be described. The tree dimensional arrangement of amino acid sequences can be described by tertiary structure. They can be predicted independent of each other. Functionality of the protein can be affected by the tertiary structure, topology and the tertiary structure. Structure aids in the identification of membrane proteins, location of binding sites and identification of homologous proteins [9], [10], [11] to list a few of the benefits, and thus highlighting the importance, of knowing this level of structure This is the reason why considerable efforts have been devoted in predicting the structure only. Knowing the structure of a protein is extremely important and can also greatly enhance the accuracy of tertiary structure prediction. Furthermore, proteins can be classified according to their structural elements, specifically their alpha helix and beta sheet content.

2 Related Works in Structure Prediction

The Objective of structure prediction is to identify whether the amino acid residue of protein is in helix, strand or any other shape. In 1960 as a initiative step of structure prediction the probability of respective structure element is calculated for each amino acid by taking single amino acid properties consideration [1], [3], [6]. This method of structure prediction is said to be first generation technique. Later this work extended by considering the local environment of amino acid said to be second generation technique. In case of particular amino acid structure prediction adjacent residues information also needed, it considers the local environment of amino acid it gives 65% structure information. So that extension work gives 60% accuracy. The third generation technique includes machine learning, knowledge about proteins, several algorithms which gives 70% accuracy. Neural networks [10], [11] are also useful in implementing structure prediction programs like PHD, SAM-T99.

The evolution process is directed by the popular Genetic Algorithm (GA) with the underlying philosophy of survival of the fittest gene. This GA framework can be adopted to arrive at the desired CA rule structure appropriate to model a physical system. The goals of GA formulation are to enhance the understanding of the ways CA performs computations and to learn how CA may be evolved to perform a specific computational task and to understand how evolution creates complex global behavior in a locally interconnected system of simple cells. Artificial immune systems are motivated by the theory of immunology. The biological immune system functions to protect the body against pathogens or antigens that could potentially cause harm. It works by producing antibodies that identify, bind to, and finally eliminate the pathogens. Even though the number of antigens is far larger than the number of antibodies, the biological immune system has evolved to allow it to deal with the antigens. The immune system will learn the criteria of the antigens so that in future it can react both to those antigens it has encountered before as well as to entirely new ones. In 2002, de Castro and Timmis [17], suggested that "for a system to be characterized as an artificial immune system, it has to embody at least a basic model of an immune component (e.g. cell, molecule, organ), it has to have been designed using the ideas from theoretical and/or experimental immunology.

IV. Step 1: Generate a AIS-PSMACA with k number of attractor basins.

3 Design of MACA based Pattern Classifier with Artificial Immune System

Step 2: Distribute S into k attractor basins (nodes).

Step 3: Evaluate the distribution of examples in each attractor basin

Step 4: If all the examples (S") of an attractor basin (node) belong to only one class, then label the attractor basin (leaf node) for that class.

Step 5: If examples (S") of an attractor basin belong to K" number of classes, then Partition (S", K").

Step 6: Stop.

A special class of non-linear CA, termed as Multiple Attractor CA (SPECIAL MACA), has been proposed to develop the model. Theoretical analysis, reported in this chapter, provides an estimate of the noise accommodating capability of the proposed SPECIAL MACA based associative memory model. Characterization of the basins of attraction of the proposed model establishes the sparse network of nonlinear CA (SPECIAL MACA) as a powerful pattern recognizer for memorizing unbiased patterns. It provides an efficient and cost-effective alternative to the dense network of neural net for pattern recognition. Detailed analysis of the SPECIAL MACA rule space establishes the fact that the rule subspace of the pattern recognizing/classifying CA lies at the edge of chaos. Such a CA, as projected in [20], is capable of executing complex computation. The analysis and experimental results reported in the current and next chapters confirm this viewpoint. A SPECIAL MACA employing the CA rules at the edge of chaos is capable of performing complex computation associated with pattern recognition.

4 c) Algorithm Single Point Crossover

Input : Two randomly selected rule vectors (Parent 1 and 2). Output : Resultant rule vectors (Offspring 1 and 2).

Step 1: Randomly generate a number "q" in between 1 and n.

Step 2: Take the first q rules (symbols) from first rule vector (Parent 1) and the (n-q) rules of Parent 2. Form a new rule vector (Offspring 1) concatenating these rules.

Step 3: Form Offspring 2 by concatenating the first q rules of Parent 2 and the last (n-q) rules of Parent 1.

Step 4: Stop.

5 d) Random Generation of Initial Population

To form the initial population, it must be ensured that each solution randomly generated is a combination of an n-bit DS with 2m number of attractor basins (Classifier #1) and an m-bit DV (Classifier #2). The chromosomes are randomly synthesized according to the following steps. V.

6 Experimental Step

? Select the target CA protein (amino acid sequence) T, whose structure is to be predicted.

? Perform a AIS-PSMACA search, using the primary amino acid sequence Tp of the target CA protein T.

The objective is being to locate a set of CA proteins, S = {S1, S2?} of similar sequence

? Select from S the primary structure Bp of a base CA protein, with a significant match to the target CA protein. A AIS-PSMACA [16],[18] search produces a measure of similarity between each CA protein in S and the target CA protein T. Therefore, Bp can be chosen as the CA protein with the highest such value

? Obtain the base CA protein"s structure, Bs, from the PDB

? Using Bp, create an input sequences Ib (corresponding to the base CA protein) by replacing each amino acid in the primary structure with its hydrophobia city value. The output sequences Ob is created by replacing the structural elements in Bs with the values, 200, 600, 800 for helix C, strand and coil respectively

? Solve the system identification problem, by performing CA de convolution with the output sequences Ob and the input sequence Ib to obtain the CA response, or the sought after running the algorithm.

?

7 Experimental Results

In the experiments conducted, the base proteins are assigned the values 300,700,900 for helix C, strand and coil respectively. We have found an structure numbering scheme that is build on Boolean characters of CA which predicts the coils, stands and helices separately. The MACA based prediction procedure as described in the previous section is then executed, and each occurrence of each sequences in the resulting output, is predicted. The query sequence analyzer was designed and identification of the green terminals of the protein is simulated in the figure **??**. The analysis of the sequence and the place of joining of the proteins are also pointed out in the figure **??**. Experimental results Figure **??**, 8 which include the similarity and accuracy graph with each of the components are separately plotted.

8 Conclusion

Existing structure-prediction methods can predict the structure of protein with 75% accuracy. To provide a more thorough analysis of the viability of our proposed technique more experiments will be conducted .Our results indicate that such a level of accuracy is attainable, and can be potentially surpassed with our method. AIS-AIS-PSMACA provides the best overall accuracy that ranges between 80% and 89.8% depending on the dataset.

9 Global Journal of Computer Science and Technology

Volume XIII Issue IV Version I [1]

Figure 1:

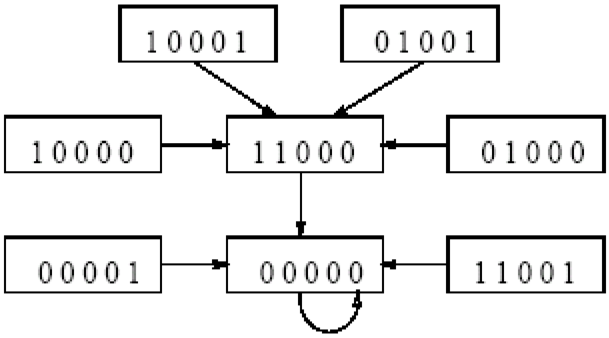


Figure 2:

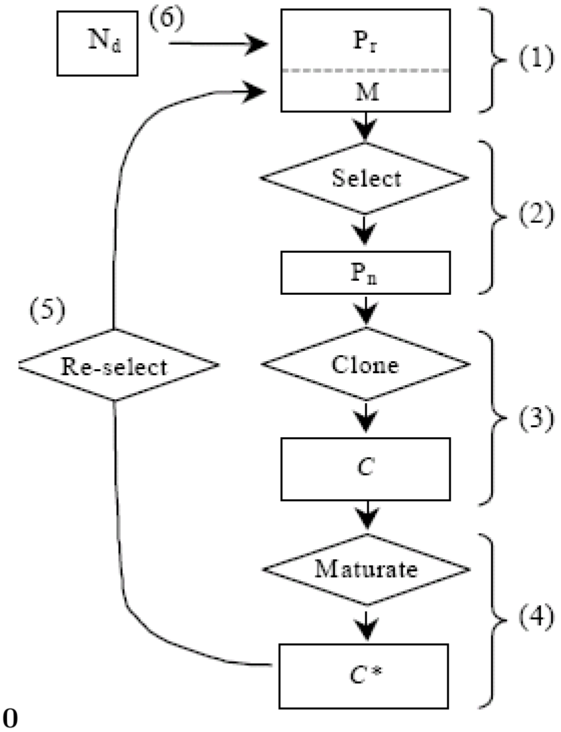
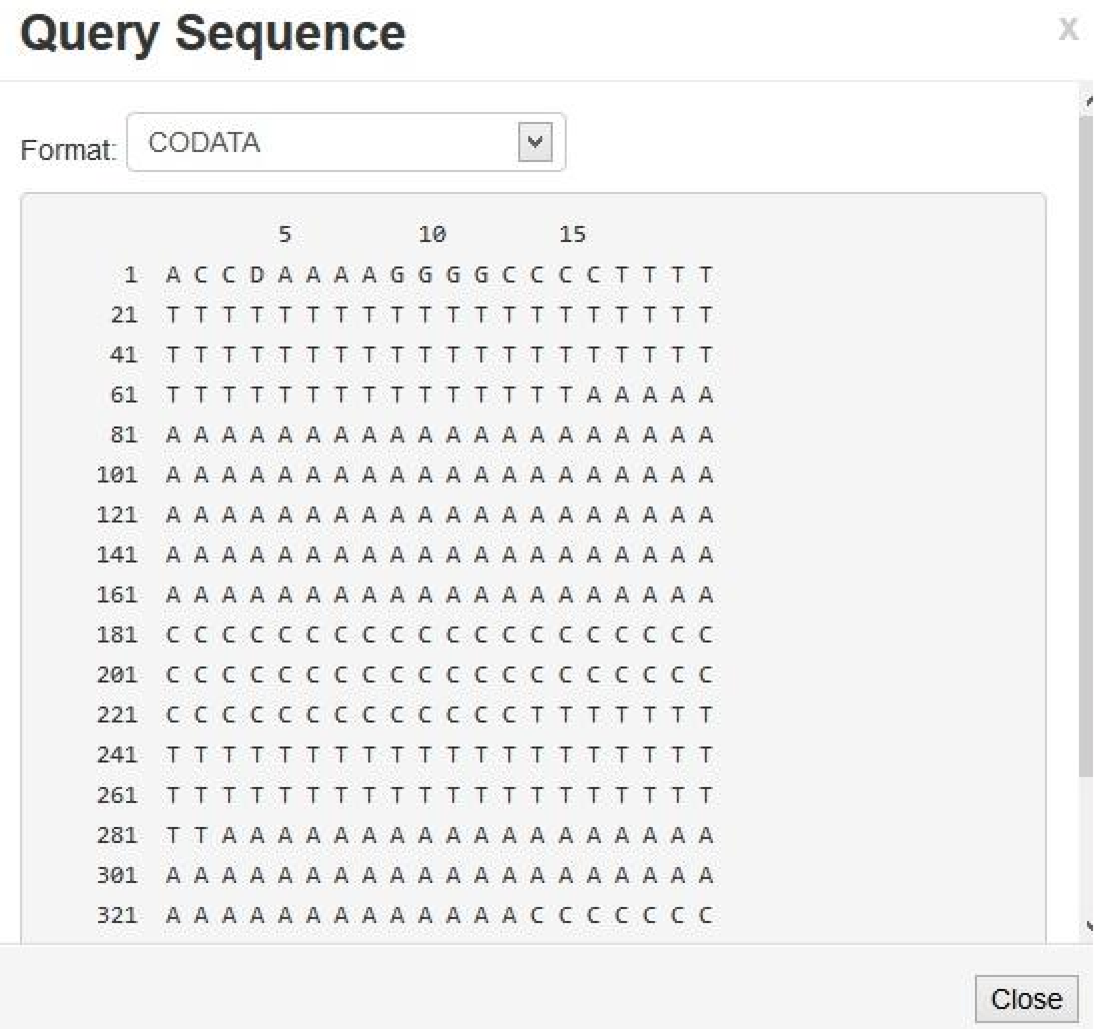


Figure 3: Figure 1 : 10 G

1



Query Sequence

Format: CODATA

```
               5         10         15
  1  A C C D A A A A G G G G C C C C T T T T
 21  T T T T T T T T T T T T T T T T T T T T
 41  T T T T T T T T T T T T T T T T T T T T
 61  T T T T T T T T T T T T T T T A A A A A
 81  A A A A A A A A A A A A A A A A A A A A
101  A A A A A A A A A A A A A A A A A A A A
121  A A A A A A A A A A A A A A A A A A A A
141  A A A A A A A A A A A A A A A A A A A A
161  A A A A A A A A A A A A A A A A A A A A
181  C C C C C C C C C C C C C C C C C C C C
201  C C C C C C C C C C C C C C C C C C C C
221  C C C C C C C C C C C C C T T T T T T T
241  T T T T T T T T T T T T T T T T T T T T
261  T T T T T T T T T T T T T T T T T T T T
281  T T A A A A A A A A A A A A A A A A A A
301  A A A A A A A A A A A A A A A A A A A A
321  A A A A A A A A A A A A C C C C C C C C
```

Figure 4: 1 .

2

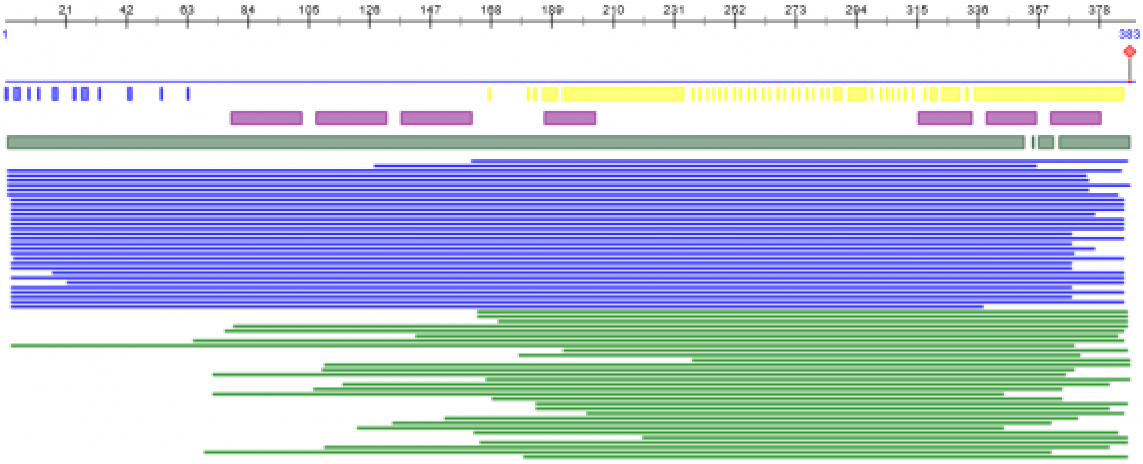
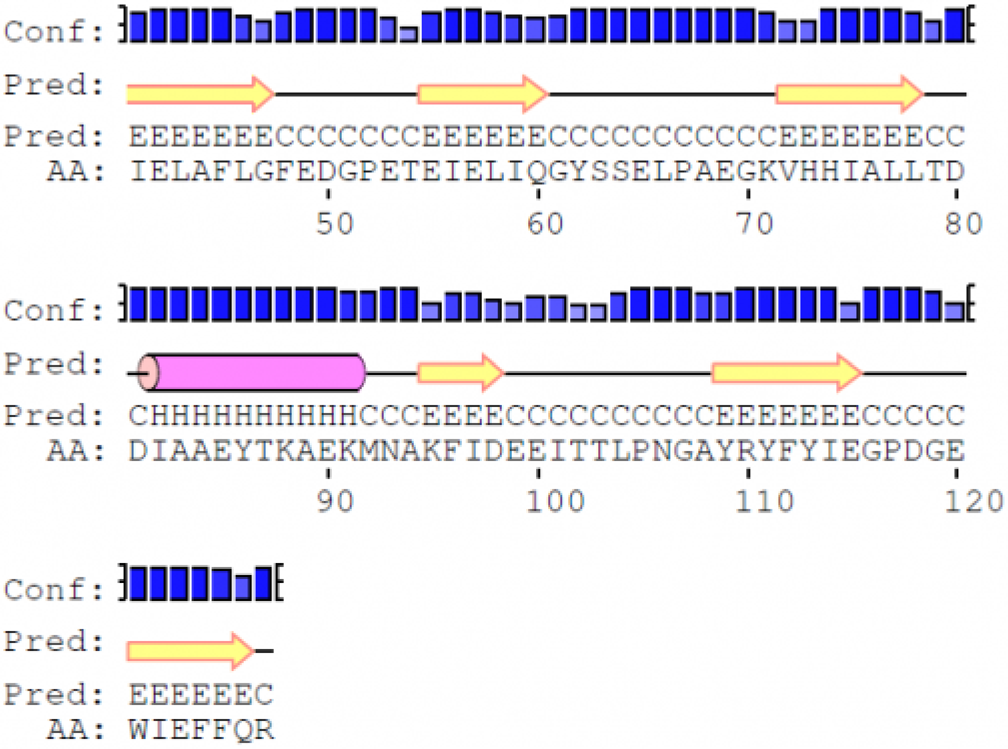


Figure 5: 2 .



```
Conf:
Pred:
Pred: EEEEEEECCCCCCCEEEEEECCCCCCCCCCCEEEEEEECC
  AA: IELAFLGFEDGPETEIELIQGYSSELPAEGKVHHIALLTD
                     50         60         70         80

Conf:
Pred:
Pred: CHHHHHHHHHCCCEEEECCCCCCCCCCEEEEEEEECCCCC
  AA: DIAAEYTKAEKMNAKFIDEEITTLPNGAYRYFYIEGPDGE
                     90         100        110        120

Conf:
Pred:
Pred: EEEEEEC
  AA: WIEFFQR
```

12

Figure 6: 12 G

14

Target : 1PFC	Prediction Accuracy	Target: 1PP2	Prediction Accuracy	Target: 1QL8
Exp 1 65%		Exp 5	85%	Exp 9
Exp 2 65%		Exp 6	90%	Exp 10

Exp 3 69% Exp 4 71% Prediction Method Prediction Accuracy for 1PFC Prediction Accuracy for Exp 7 83

SS Pro AIS-PSMACA AIS-AIS-PSMACA		70% 90% 92%	73% 85% 83%

Figure 7: 14 G

[1] G© 2013 Global Journals Inc. (US)

[Bienkowsk and Lathrop] , Jadwiga Bienkowsk , Rick Lathrop . *THREADING ALGORITHMS*

[Bioinformatics] , *Bioinformatics* 14 (10) p. .

[Lippmann ()] 'An introduction to computing with neural nets'. R Lippmann . *IEEE ASSP Mag* 2004. 4 (22) p. .

[Baldi ()] 'Bidirectional Dynamics for Protein Secondary Structure Prediction'. [Baldi . *Sequence Learning: Paradigms, Algorithms and Applications*, 2000. 2000. Springer. p. .

[Bourne and Weissig ()] Philip E Bourne , Helge Weissig . *Structural Bioinformatics*, 2003. 2003. John Wiley & Sons.

[Dunbrack ()] 'Comparative Modeling of CASP3 Targets Using PSI-BLAST and SCWRL'. R Dunbrack . *PROTEINS: Structure, Function and Genetics* 1999. 1999. 3 p. .

[Skolnick and Kolinski ()] 'Computational Studies of Protein Folding'. J Skolnick , A Kolinski . *Computing in Science and Engineering* 2001. 2001. (Skolnick and Kolinski)

[Mitra and Smith ()] 'Digital Signal Processing in Protein Secondary Structure Prediction'. Debasis Mitra , Michael Smith . *Innovations in Applied Artificial Intelligence Lecture Notes in Computer Science* 2004. 3029 p. .

[Alexandrov and Solovyev ()] 'Effect of secondary structure prediction on protein fold recognition and database search'. N Alexandrov , V Solovyev . *Genome Informatics* 1996. 1996. 7 p. . (Alexandrov and Solovyev)

[Irback ()] 'Evidence for nonrandom hydrophobicity structures in protein chains'. Irback . *Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features*, 1996. 1996. September. 1983. 93 p. . (Biopolymers)

[Sree and Babu ()] 'Face Detection from still and Video Images using Unsupervised Cellular Automata with K means clustering algorithm'. P Sree , I Ramesh Babu . *Vision and Image Processing*, 2008. 8 p. . (Issue II)

[Flocchini et al. ()] P Flocchini , F Geurts , A Mingarelli , Santoro . *Convergence and Aperiodicity in Fuzzy Cellular Automata: Revisiting Rule 90*, 2000.

[Maji and Chaudhuri ()] 'FMACA: A Fuzzy Cellular Automata Based Pattern Classifier'. P Maji , P P Chaudhuri . *Proceedings of 9th International Conference on Database Systems*, (9th International Conference on Database SystemsKorea) 2004. 2004. p. .

[Maji and Chaudhuri ()] 'FMACA: A Fuzzy Cellular Automata Based Pattern Classifier'. P Maji , P P Chaudhuri . *Proceedings of 9th International Conference on Database Systems*, (9th International Conference on Database Systems) 2004. 2004. p. .

[Maji and Chaudhuri ()] 'Fuzzy Cellular Automata for Modeling Pattern Classifier'. P Maji , P P Chaudhuri . *IEICE* 2004.

[Global Journal of Computer Science and Technology Volume XIII Issue IV Version I 2] *Global Journal of Computer Science and Technology Volume XIII Issue IV Version I 2*,

[Karplus ()] *Hidden Markov Models for Detecting Remote Protein Homologies*, Karplus . 1998. 1998.

[Abagyan ()] 'Homology Modeling With Internal Coordinate Mechanics: Deformation Zone Mapping and Improvements of Models via Conformational Search'. Abagyan . *PROTEINS: Structure, Function and Genetics* 1997. 1997. 1 p. .

[Snyder and Stormo (1993)] 'Identification of coding regions in genomic DNA sequences: an application of dynamic programming and neural networks'. E E Snyder , G D Stormo . *Nucleic Acids Res* 1993 February 11. 21 (3) p. .

[Snyder and Stormo ()] 'Identification of Protein Coding Regions in Genomic DNA'. Eric E Snyder , Gary D Stormo . *ICCS Transactions* 2002.

[Sree and Babu ()] 'Identification of Protein Coding Regions in Genomic DNA Using Unsupervised FMACA Based Pattern Classifier'. P Sree , I Ramesh Babu . *International Journal of Computer Science & Network Security* 1738-7906. 2008. Number (1) .

[Sree et al. ()] 'Improving Quality of Clustering using Cellular Automata for Information retrieval'. P Sree , G V S Raju , I Ramesh Babu , S Viswanadha , Raju . *International Journal of Computer Science* 1549-3636. 2008. Science Publications-USA. 4 (2) p. .

[Brandon and Tooze ()] *Introduction to Protein Structure*, C Brandon , J Tooze . 1999. 1999. Garland Publishing. (Brandon and Tooze)

[Veljkovic ()] 'Is It Possible To Analyze DNA and Protein Sequences by the Methods of Digital Signal Processing'. Veljkovic . *IEEE Transactions on Biomedical Engineering* 1985. 1985. 1985. 32 (5) p. .

[Chandonia and Karplus ()] 'New Methods for Accurate Prediction of Protein Secondary Structure'. J Chandonia , M Karplus . *PROTEINS: Structure, Function and Genetics* 1999. 1999. 35 p. .

[Irback and Sandelin ()] 'On Hydrophobicity Correlations in Protein Chains'. A Irback , E Sandelin . *Biophysical Journal* 2000. 2000. 79 p. . (Irback and Sandelin)

[Hirakawa and Kuhara ()] 'Prediction of Hydrophobic Cores of Proteins Using Wavelet Analysis'. H Hirakawa , S Kuhara . *Genome Informatics* 1997. 1997. 8 p. .

[Chou and Fasman ()] 'Prediction of the secondary structure of proteins from their amino acid sequence'. P Chou , G Fasman . *Advanced Enzymology* 1978. 1978. 47 p. .

[Anfinsen ()] 'Principles that govern the folding of protein chains'. C B Anfinsen . *Science* 1973. 181 p. .

[Thiele ()] 'Protein Threading by Recursive Dynamic Programming'. Thiele . *Journal of Molecular Biology* 1999. 290 p. .

[Kiran Sree Dr Inampudi and Babu] 'PSMACA: An Automated Protein Structure Prediction using MACA (Multiple Attractor Cellular Automata)'. P Kiran Sree & Dr Inampudi , Ramesh Babu . *Journal of Bioinformatics and Intelligent Control* American Scientific Publications. 2 (3) . (JBIC) in)

[Bonneau ()] 'Rosetta in CASP4: Progress in Ab Initio Protein Structure Prediction'. [Bonneau . *PROTEINS: Structure, Function and Genetics* 2001. 2001. 5 p. .

[Boeckmann ()] 'The SWISS-PROT protein knowledgebase and its supplement TrEMBL in'. Boeckmann . *Nucleic Acids Res* 2003. 2003. 2003. 31 p. .