IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727
PP 07-10
www.iosrjournals.org

An Overview of Three Dimensional Protein Structure Predictions

[1]Soumya Sasi II MSc (Computer Science), [2]Dr.D. Ramyachitra Assistant Professor,
[3]P. Lakshmi PHD Scholar,
Department of Computer Science, Bharathiar University,

Abstract: *Protein structure is the three dimensional arrangements of atoms (amino acids) in a protein molecule .The three dimensional protein structure prediction is used to easily understand the function and molecular level of a protein. In this paper summarized about some evolutionary algorithms, methods and tools that are used for solving the protein tertiary structure prediction.*
Keywords: *protein structure, three dimensional protein structure prediction, evolutionary algorithms.*

I. Introduction

Proteins are large biomolecules, or macromolecules that are exist in living organisms. This protein consists of one or more long chains of amino acid sequence, which are the important nutrients for the human body to maintain growth. The proteins have 20 amino acids and each one is differ from other only by the variations in the amino acid sequence (Nelson DL, Cox MM,2005). When you consider a protein structure, atoms in a protein molecule are arranged in three dimensional formats. Depending on the structure, proteins can be divided into four different levels(Murray *et al*). They are given below

1. Primary structure
2. Secondary structure
3. Tertiary structure
4. Quaternary structure

Primary structure of proteins consists of linear sequences of twenty natural amino acids joined together by peptide bonds. The secondary structure can be defined by the pattern of hydrogen bonds between the amine hydrogen and carbonyl oxygen atoms in the peptide backbone(Mount DM, 2004). Protein tertiary structure is that the proteins that are represented in a three dimensional shape. These tertiary structures of proteins have a single polypeptide chain with one or more protein secondary structures (Zhang Y, 2008). The quaternary structure is the final dimensional structure formed by all the polypeptide chains making up a protein (Battey JN, Kopp J,etal.(2007)).

II. 3d Protein Structure Prediction Overflow

Predicting the tree dimensional protein structure is very difficult task in structural bio informatics. Imperialist competitive algorithm was used to accelerate the minimization process (Erfan Khajia, etal 2016). Using genomic sequences, the 3D protein structure prediction by some computational methodology can be done that involving the genetic algorithm. These were analyzed by using different tools (Jyotshna Dongardive S, 2013).

From the 3D protein structure prediction of protein sequences differentiate the interacting and non-interacting protein pairs are complex task. For that many methods are developed (Thomas A Hopf, etal, 2014). The 3D structure of proteins prediction some computational strategies are used for obtained structural information that are called Angle Probability List (APL) combined with distributed knowledge based Genetic algorithm (Bruno Borguesan,etal 2016).

For protein structure prediction problem the novel strategies are introduced. Based on the Memetic algorithm the search method is extracted information from protein data bank (Leonardo Corrˆea,etal 2016).Newly improved hybrid optimization algorithm is PGATS algorithm. This is the combination of three algorithms they are Particle Swarm Optimization (PSO), Genetic Algorithm (GA) and Tabu Search (TS) algorithms (Changjun Zhou,etal 2014).

This section discuss about the 3D protein structure prediction using variety of algorithms and, methods tools that are given in different journal papers. For each of these different algorithms in different papers are evaluated by using of different proteins sequences as the input.

III. Algorithms And Methods

Imperialist competitive algorithm: In the imperialist competitive algorithm more colonies ruled by the powerful imperialist country which generate the random population. The imperialist counties are chosen by best country that had lowest cost.

Heuristic algorithm: Heuristic algorithm is used for the energy function, in the specific task heuristic methods are efficient (Chen and Huang, 2005).

Future work is to provide a web-server for the three dimensional protein structure using ICA.(Erfan Khajia, etal 2016)**.**

The drawback of this algorithm is that it can't produce results for many of the computational problems

Genetic algorithm: Genetic algorithm use genomic sequences for its experimental works, the collection of sequences are evolved. These sequences are offer consensus of the sequences, then this consensus generate a primary protein sequence. That will used for predict the 3D structure of protein. (Siby Abraham, 2013).

The drawback of the algorithm is to consider low mutation level and also give accurate value for writing of fitness function.

Weight matrix algorithm: Weight matrix algorithm is commonly used for the representation of patterns in biological sequence (Siby Abraham, 2013).

Future enhancement is the comparative molecular modeling allows expanding the number of protein sequences for which have structural information, such models can be effectively used to design mutagenesis experiments and to support drug design projects. (Siby Abraham, 2013)**.**

Distributed genetic algorithm: Distributed genetic algorithm is a type of search algorithm with a structured population , which can handile sixteen times for 12 hours using APL (Bruno Borguesan, 2016).

Memetic algorithm: This memtic algorithm is used to deal with the PSP problem that uses a structured ternary tree population allied to different kinds of global search operators and also a Simulated Annealing implementation as a local search technique (Mario Inostroza-Ponta, 2016).

PGATS algorithm: PGATS algorithm is an improved hybrid algorithm, which is the combination of particle swarm optimization (PSO), genetic algorithm (GA) and tabu search(TB)algorithms.(Changjun Zhou,etal, 2014).

EVcomplex method: EVcomplex can determine which proteins interact with each other at the same time as specific residue pair couplings across the proteins.

An important technical challenge for future work is to determine models of these complexes one must deconvolute homomultimeric inter-ECs from the intra-protein signal. (Oliver Kohlbacher, 2014).

RMSD method: Root-meen-square deviation is the method used to measure the average distance between the atoms (Mario Inostroza-Ponta, 2016).

Local search method: Local search method is used on hard computational problems that maximize the criteria based on that find the optimal solution. (Mario Inostroza-Ponta, 2016).

Random linear method: Random linear method formed by the operations of crossover and mutation that are in the genetic algorithm. (Changjun Zhou ,etal 2014).

IV. Tools

PSIPRED Tool: PSI-blast based secondary structure prediction is used to search the protein,for the secondary structure.(Erfan Khajia, etal, 2016)**.**

HSEpred Tool: HSEpred, is used for the number of amino acid neighbors within two half spheres of chosen radius around the amino acids. (Erfan Khajia, etal, 2016)**.**

BLASTp: BLASTp(Basic Local Alignment Tool) is the tool for finding the amino acid sequence obtained to detect a closely related homolog(Jyotshna Dongardive, 2013).

V. Datasets

DNA binding protein sequences can be taken from Protein Data Bank (PDB: http//www.rcsb.org/pdb/home/home.do). Sequences are downloaded from the server (http://server.malab.cn/Local-DPP/Datasets.html).

The predicted 3D structures of the sequences in the tested benchmark: 1FC2, 1ENH, 2GB1, 2CRO, 1CTF, and 4ICB respectively. Right pictures: experimentally determined 3D structures of the protein in the tested benchmark obtained from PDbj website (Berman HM,etal 2000, "Protein Data Bank").

DNA sequences of HPV strains obtained from NCBI GenBank as the primary data source.The multiple sequence alignment was performed using Clustal X 2. In this the datasets are training datasets and testing datasets were used as the input (Berman HM,etal 2000, "Protein Data Bank").

The Data Sequence co-evolution gives 3D contacts and structures of protein complexes, Publicly available at Dryad Digital Repository, and is taken from the protein data bank (Hopf T, etal, 2014 "Sequence co-evolution gives 3D contacts and structures of protein complexes").

Blinded prediction of evolutionary coupling between complex subunits with known 3D structure . APL method predict the three dimensional structure of eight protein sequences obtained from the PDB: 1L2Y , 1WQC , 2F4K , 2MR9, 2MTW , 3P7K , 1K43 and 1ACW (H.M. Berman, etal, 2000, "The protein data bank").

As the form of input secondary structures of proteins are taken from protein data bank, and the angle probability list used for reduce the search sequences(K. D. Pruitt, etal, 2002, "Ncbi reference sequence").

In the sequence D, E, F, H, K, N, Q, R, S, T, W and Y are hydrophobic, and I, V, L, P, C, M, A and G are hydrophilic, which was taken from the protein data bank("PDB database").

VI. Perfomance Meatures

The radius and the total energy Q(p) is defined as the root- mean-square deviation (RMSD) of the residue energy contributions Q(p)(Ai),

$$Rg = 2.2n^{0.38} \quad \text{---------- (1)}$$

$$\mathbf{Q(P)}=\sqrt{\sum_{I=1}^{n} \mathbf{Qp(Ai)/(2*n)}} \text{----------(2)}$$

Total power of an empire=power of imperialist+ r (mean (powers of colonies)) ---------- (3)

Each inter-protein pair of sites i and j with pair coupling strength ECinter(i, j), we therefore calculate a raw reliability score defined by

$$Q_{inter}^{row}(i,j) = \frac{EC_{inter}(i,j)}{|\min_{i,j}(EC_{inter}(i,j))|} \text{ ------- (4)}$$

The normalized EVcomplex score is defined as

$$EVcomplex - Score(i,j) = \frac{Q_{inter}^{row}(i,j)}{1+\left[\frac{N_{eff}}{L}\right]^{-\frac{1}{2}}} \text{--(5)}$$

These equations are taken from the previously discussed paper (Thomas A Hopf, etal, 2014).

For each amino acid residue and secondary structure we compute the torsion Angle Probability List APLaa,ss that represents the normalized frequency of each square(Bruno Borguesan,etal 2016).

$$\mathbf{APLaa,ss(i,j)} = \frac{\mathbf{Haa,ss(i,j)}}{\sum(\mathbf{Haa,ss})} \text{ ---------- (6)}$$

To evaluate DNA binding protein predictors LOOCV (Leave one out cross validation) test carried out for fair comparison with existing methods

Four evaluation metrics are: Sensitivity (SE), Specificity (SP), Accuracy (ACC), and Mathew's Correlation Coefficient (MCC) ("PDB database").

SE = (TP/ (TP+TN))*100% ---------- (7)

SP = (TN/ (TN+FP))*100% ---------- (8)

ACC = ((TP+TN) /(TP+TN+FN+FP)) *100% ---------- (9)

$$\mathbf{MCC} = \frac{(\mathbf{TP*TN-FP*FN})}{\sqrt{(\mathbf{FP+FN})(\mathbf{TP+FP})(\mathbf{TN+FP})(\mathbf{TN+FN})}} \text{- (10)}$$

Minimization measure of Root Mean Square Deviation (RMSD) and the Maximum measure of Global distance total score test (GDT_TS) are used to analyze the 3D protein structures.

$$\mathbf{RMSD_{(a,b)}} = \sqrt{(\sum_{1=1}^{n} |\mathbf{rai - rbi}|)/\mathbf{n}} \text{ - (11)}$$

Where $r_{ai \text{ and }} r_{bi}$ are the vectors representing the positions of the same atom I of the two structures a and b respectively.

$$\mathbf{GDT_{TS} = (GDT_{P1}+GDT_{P2}+GDT_{P3}+GDT_{P8})\ /4} \text{ ---------- (12)}$$

Where GDT_{pn} represents the percentage of residues under the cutoff distance <=n ◌̊A (Leonardo Corrˆea, etal 2016).

The energy function for any chains of n monomers is described as

$$E = \sum_{i=1}^{n-2} \frac{1}{4}(1-\cos\theta_i) + \sum_{i=1}^{n-2} \sum_{j=i+2}^{n} 4\left[r_{ij}^{-12} - C(\varepsilon_i,\varepsilon_j) r_{ij}^{-6}\right] \quad \text{---(13)}$$

The variable formula of weight coefficient is

$w = w_{max}$-**time** $\times (w_{max} - w_{min})/MaxDT$-------(**14**)

wmax and wmin are the maximum and minimum of w, respectively.Time is the current iterations, MaxDT is the maximum iterations(Changjun Zhou,etal 2014).

VII. Conclusion

In bioinformatics predicting the native structure of protein from its amino acids is a big problem. The 3d protein structure prediction is very useful for the drug development. This paper illustrate about algorithms, methods and tools for predicting the three dimensional protein structure.

References

[1]. Nelson DL, Cox MM (2005). Lehninger's Principles of Biochemistry (4th ed.). New York, New York: W. H. Freeman and Company.

[2]. Murray *et al*., pp. 30–34.

[3]. Mount DM (2004). Bioinformatics: Sequence and Genome Analysis. **2**. Cold Spring Harbor Laboratory Press. ISBN 0-87969-712-1.

[4]. Zhang Y (2008). "Progress and challenges in protein structure prediction". *Curr Opin Struct Biol*. **18** (3): 342–8. PMC 2680823 . PMID 18436442. doi: 10.1016/j.sbi.2008.02.004

[5]. Battey JN, Kopp J, Bordoli L, Read RJ, Clarke ND, Schwede T; Kopp; Bordoli; Read; Clarke; Schwede (2007). "Automated server predictions in CASP7". *Proteins.* **69** (Suppl 8): 68–82. PMID 17894354. doi:10.1002/prot.21761.

[6]. "3D protein structure prediction using Imperialist Competitive algorithm and half sphere exposure prediction", Erfan Khaji a, Masoumeh Karami b, Zahra Garkani - Nejadc,n, 0022-5193/& 2015 Elsevier Ltd

[7]. "Predicting 3D Structure of Proteins from Genomic Sequences: A Genetic Algorithm Approach", Jyotshna Dongardive S, 978-1-4673-6217-7/13/$31.00_c 2013 IEEE

[8]. "Sequence co-evolution gives 3D contacts and structures of protein complexes", Thomas A Hopf1,2†, Charlotta P I Schärfe1,3,4†, João P G L M Rodrigues5†, Anna G Green1, Oliver Kohlbacher3,4, Chris Sander6*, Alexandre M J J Bonvin5*, Debora S Marks1*, Hopf *et al.* eLife 2014;3:e03430. DOI: 10.7554/ eLife.03430

[9]. "Improving protein tertiary structure prediction with conformational propensities of amino acid residues", Bruno Borguesan, Jonas Bohrer, Mariel Barbachan e Silva, 978-1-5090-0623-6/16/$31.00_c 2016 IEEE

[10]. "A Memetic Algorithm for 3-D Protein Structure Prediction Problem", Leonardo Corrˆea, Bruno Borguesan, Camilo Farf ´an, Mario Inostroza - Ponta, and M´arcio Dorn, IEEE 2016

[11]. "Improved hybrid optimization algorithm for 3D protein structure prediction", Changjun Zhou & Caixia Hou & Xiaopeng Wei &Qiang Zhang ,Springer-Verlag Berlin Heidelberg 2014

[12]. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al.(2000): "The Protein Data Ban", .Nucleic Acids Res. 28:235–42.

[13]. Hopf T, Schärfe C, Rodrigues J, Green A, Sander C, Bonvin A, Marks D , 2014 , Data from: "Sequence co-evolution gives 3D contacts and structures of protein complexes" , http://dx.doi.org/10.5061/ dryad.6t7b8, Publicly available at Dryad Digital Repository.

[14]. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N.Bath, H. Weissig, I.N. Shindyalov, and P.E. Bourne. "The protein data bank". *Nucleic Acids Res.*, 28(1):235–242, 2000

[15]. K. D. Pruitt, T. Tatusova, and D. R. Maglott, "Ncbi reference sequence (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins," Nucleic Acids Res., vol. 33, no. suppl 1, pp. D501–D504, 2005.

[16]. The protein sequences are downloaded from the "PDB database". The URL is http://pdbbeta.rcsb.org/pdb/Welcome