Paramount role of sequencing in Ubiquitous forms of Nucleotide

Dineshnath.G[#1], Nagarajan.E[*2] and Karthick.R[#3]

#Post Graduate, Department of Computer Science and Engineering,
Sathyabama University,
Chennai, India
[1]dinesh31m90@gmail.com
[3]arjunsehar@gmail.com
[*]Assistant Professor, Department of Computer Science and Engineering,
Sathyabama University, Chennai, India
[2]enagaraj99@yahoo.com

***ABSTRACT*-- Protein microarrays allow biologist to measure the characteristics of amino acid sequence on a small chip. It generates voluminous of data and new intelligent methods are needed to analyze them. Sequence alignment is challenging task for protein structure prediction. Also for better structural alignments, various algorithms have been proposed. In this paper familiar pair wise alignments are discussed and their significant merits and demerits Finally, this study concludes with an analysis of challenges to overcome for similarity mismatch of given candidate query is evaluated in every structural organization with conventional techniques that still remain to be resolved with Dynamic approach.**

KEYWORDS— Peptide bond formation, Protein fold Recognition, Structural organization of protein.

I. INTRODUCTION

DNA/RNA undergoes post translation mechanism to transfer genetic code of nucleotide sequence by T-RNAs for synthesis of Protein. Nucleic acids are secreted by codons in the transcription process and form a helical structure. m-RNA and t-RNA plays a vital role in synthesis and production of macro molecules(protein , mitochondria).Those nucleic acids are also called amino acid.

Bio-Chemical reactions generate two types of tautomers L-amino acid (present in body) and S-amino acid (produced by enzymes).

Biochemistry researchers compute helical structure angles(Phi, Psi, omega and torsion angles) and electrochemical properties ,Presence of hydrogen and energy(Electro potential) to interpret covalent or ionic bond nature of chemical compound .Since hydrophobic and hydrophilic property is required for amino acid for peptide bond formation. Hence these properties are taken into account in secondary structure prediction.

Vander Waals force, chemical composition of carbon, nitrogen and other cubical properties are analyzed and model tertiary structure of protein using Swiss-Pdb consortium. Numerous Protein structures observed in various enzymes are re –engineered to form new structural assembly. Whatever may be mechanism handled to model a protein in different hierarchy of structures, for its structural prediction[4] (ordered) or Disorder in secondary structure call for sequence alignments.

The Following Sections addresses keen interest to be taken for sequence alignments is mandatory for any structural representation of protein for modeling new structures.

II. PROTEIN STRUCTURE

Amino acid sequences are the prime constituents of protein. Proteins are macromolecules that are found abundant in living organisms and it is organized into primary. Secondary, tertiary[7] and quaternary structure. Structural proteins and dynamic proteins perform their own assigned function. Example for structural protein is collagen found in bone matrix.

Protein is comprised of 20 natural amino acids are interconnected by innumerable combinations of amide bonds. Their diverse biological functions are not varied not only by the amino acid and its connectivity but also from their folded three dimensional structure. Although overall folded structure are complex to infer meaning, it is indicated in simpler characterization such as α-helices,β-sheets, turns and loops[5].

1. **PRIMARY STRUCTURE**:

It is the amino acid sequence (1940) that “exclusively” determines the 3D structure of a protein. There are essential amino acids exists in naturally and it is of two types L and S-amino acid. Since sequence alignment is performed in its nucleotide sequence [nucleobase pair A->G, T->C in DNA and purine (thymine) in RNA].Individual amino acids form a polypeptide chain which is a component of a hierarchy for describing macromolecular structure chain has its own set of attributes is planar and rigid peptide linkage.

2. **SECONDARY STRUCTURE** :

The chemical nature of the carboxyl and amino groups of all amino acids consent hydrogen bond formation (firmness) and hence defines secondary structures in the protein. The R group has an influence on the possibility of secondary structure formation (proline is an extreme case) This indications to a tendency for amino acids to exist in a particular secondary structure conformation Helices and sheets are the regular secondary structures, but irregular secondary structures exist and can be critical for biological function. For irregular structure sequence alignment is essential.

A dihedral angle is that angle between 2 planes defined by 4 atoms – 123 form one plane; 234 the other. Omega is the rotation around the peptide bond Cn – Nn+1 – it is planar and Phi is the angle around N – C alpha Psi is the angle around C alpha C’ the values of phi and psi are constrained to certain values based on steric clashes of the R group. Thus these values show characteristic patterns as defined by the Ramachandran plot. The typical view of peptide bond formation helical, β-interchain based on chirality of carbon atom and other angular measures in amino acids are shown in Figure 1.

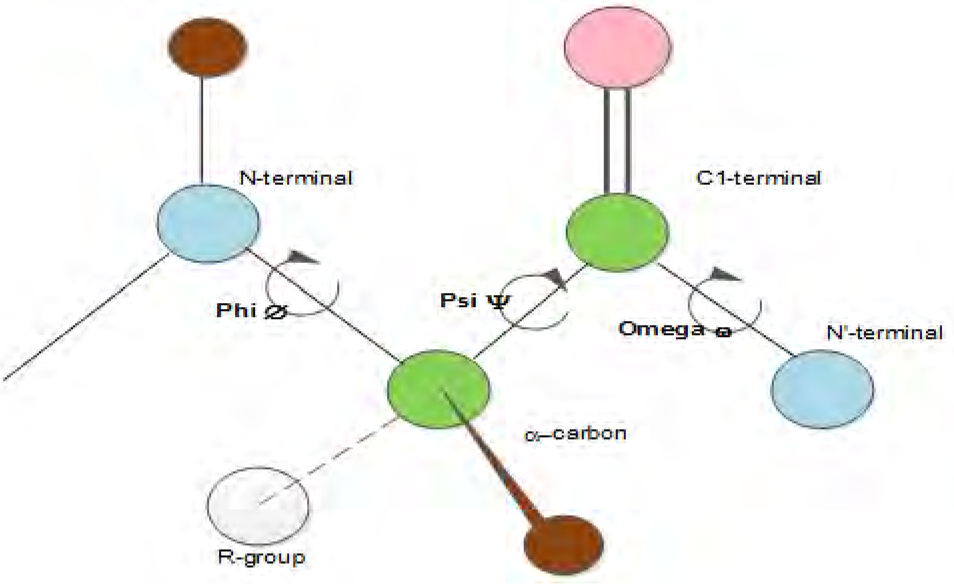


Figure 1:Fundamental Concepts of Learning Protein Residue

Linguistics (Languages) has appropriate letters, words and sentences for understanding and acquiring knowledge. Similarly protein flow interpretation is made from atomic residues, secondary Structure and tertiary structure. Various Crystallographic techniques NMR and mass spectrometry experimental results of protein disorder flavors are gathered in CASP (Critical Assessment of Structural Proteins) conduct research and analyze structural proteins. The training and testing of given primary structure with secondary structure for learning states are shown below and corresponding CASP standard is mentioned.

SAMPLE PROTEIN PRIMARY STRUCTURE SEQUENCE

INPUT: GHWIATRGQLIREAYEDYRHFSSECPFIP

Predict its secondary structure content

(C=Coils H=Alpha Helix E=Beta Strands)

CEEEEECHHHHHHHHHHCCCHHCCCCCC

Eight states from DSSP[6]

□ H: a-helix

□ G: 310 helix

□ I: p-helix

□ E: b-strand

□ B: bridge

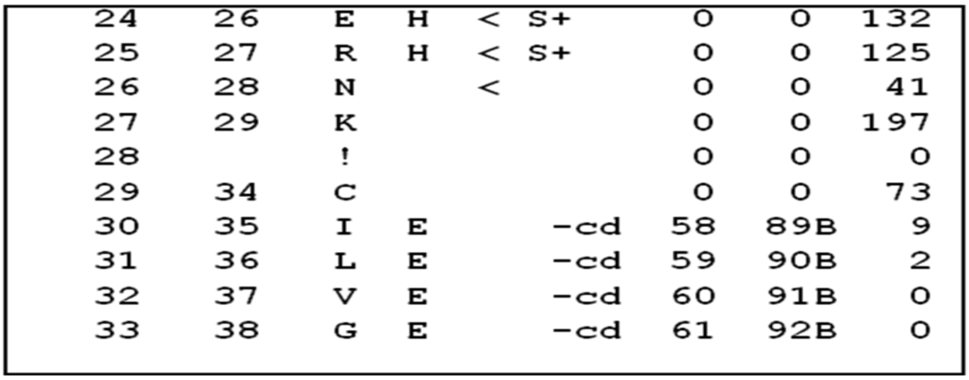
□ T: b-turn

□ Y S: bend

□ C: coil

CASP Standard

□ H = (H, G, I), E = (E, B), C = (C, T, S)



```
  24    26    E   H   <  S+       0    0  132
  25    27    R   H   <  S+       0    0  125
  26    28    N       <           0    0   41
  27    29    K                   0    0  197
  28          !                   0    0    0
  29    34    C                   0    0   73
  30    35    I   E        -cd   58  89B    9
  31    36    L   E        -cd   59  90B    2
  32    37    V   E        -cd   60  91B    0
  33    38    G   E        -cd   61  92B    0
```

Figure 2: ATOMIC RESIDUES with b-strand, Score and other parameters.

3. PDB CONTRIBUTIONS

The Protein Data Bank (PDB)[9] is associate archive of by experimentation determined three-dimensional structures of biological macromolecules that serves a global community of researchers, educators, and students. The data contained in the archive consists of atomic coordinates, crystallographic structure factors and NMR experimental data. Aside from coordinates, each deposition also includes the names of molecules, primary and secondary structure information, sequence database references, where appropriate, and ligand and biological assembly information, details about data collection and structure solution, and bibliographic citations.

NCBI, Swiss Prot, DisProt , Genbank and TrEmBL have their own Web server that perform autonomous activity which serves a Knowledge Base(KB). Every Protein database have common paradigm to design a new model of protein have standard templates which is shown in figure 3.

The detailed view of homology (comparative) model is described in preceding Section-4 which insists sequence alignment is compulsory.

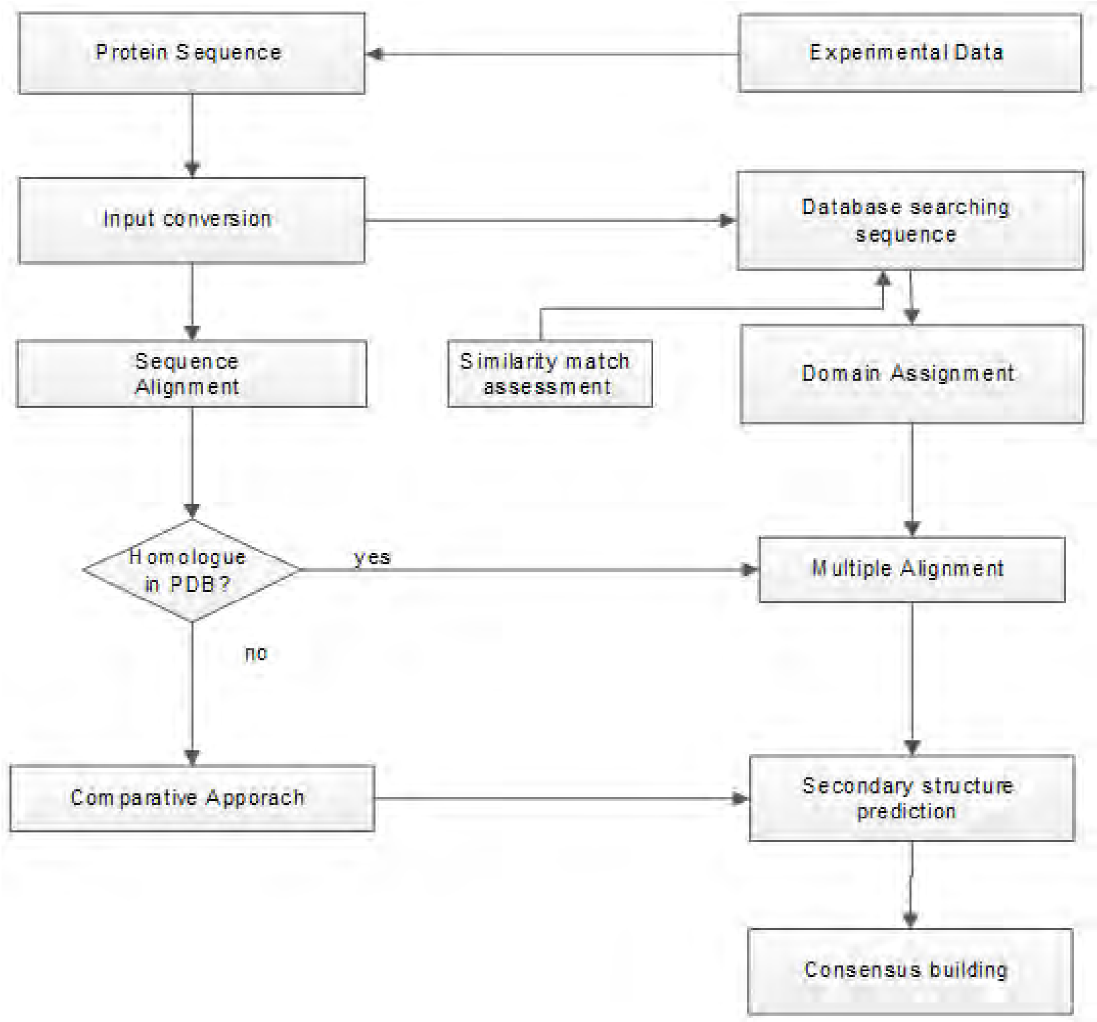


Fig:3. Function flow diagram of protein sequence alignment

4. HOMOLOGY MODEL

The homology [3] protein structure is concerned with three-dimensional (3D) model for a protein of the target sequence correspondence to proteins of identified structure (the templates)

Constructing a 3D model must employ two conditions. One, the resemblance between the target sequence and the pattern structure must be noticeable, on the other hand appropriate alignment between the target sequence and the template structures must be computed. Comparative modeling is possible since milder changes in the protein sequence usually results in minute changes in its 3D structure. Although remarkable progress has been made in AB- initio protein structure prediction, comparative protein structure modeling remains the accurate prediction method. In general , the accuracy of comparative models covers a ample range from minor models with a single exact fold to more accurate models comparable to medium range contacts of Protein residue structures determined by crystallography or nuclear magnetic resonance (NMR) spectroscopy. Even low range models can be useful in biology because some aspects of perform will generally be foreseen solely from the coarse structural options of a model.

Therefore, if relationship between 2 proteins is tracked at the sequence level, structural similarity will routinely be assumed. Moreover, even proteins that accompany sequence affiliation will have similar structures. This theme estimation shows that just about sequences square measure associated with one renowned organic compound structure.

The importance of comparative modeling is systematically progressive in nature from the time once range of distinctive structural folds that proteins depends on adoption remains restricted, because the consequence by experimentation determined new structures raises exponentially.

5. EXISTING SEQUENCE ALIGNMENT ALGORITHM

Pair wise sequence alignment methods are employed in both local (Smith Waterman Algorithm) and global (Needle man Wunch) to assess sequence similarity detection. Dot Plot paradigm is applied in case of global, its alignment is limited and process only queries sequence of unique length. Multiple Sequence alignment procedure is impossible for this case.

Local alignment compares the input query sequence of short with large database sequence as well as identify recently or newly determined sequences. It also works well for single sequence to entire database and partial sequence to the Whole. FASTA handles multiple sequence alignments, Gaps in amino acid sequence are tackled effectively for better matching.

BLAST is the modified scheme of FASTA, it does not require identical words and gaps are also not considered much. BLAST holds significant benefits such as speed, user friendly, statistically rigor and more sensitive. Blast finds only similar words. Both BLAST and FASTA are also collectively referred as K-tuple methods.

6. RELEVANT LITERARY WORKS

The linear encoding algorithms have the ability to attain competitive accuracy with existing structure alignment methods.

However, it is extent to most effective way for resolving the structure alignment problem. Further enhancement is concerned with linear encoding of secondary structure with proper geometrical symmetry of protein structure in linear sequences [1].

A novel manipulation rule for model quality assessment using molecular binding site comparisons to judge the quality of the models. The binding site residues are processed by COFACTOR scans the query 3D structure against the template library, first based on global structure similarity, followed by a local similarity refinement search on selected templates, with the purpose of filtering out template proteins that do not share binding site similarity with the query protein. During the local structure similarity search, template proteins are scored against the query protein using an Innovative structure-sequence similarity measure (BS-score), which is designed to capture both chemical and structural similarity between the query and the template proteins. Template protein, which shared the highest local similarity with the CASP models, was finally used to rank all the models based on their local similarity score (BS-score).

AOBA-server is an automated server for protein structure modeling that assesses the homology behavior of target amino acid compared with template. Pre-requiristics is sequence alignment [1]. Template identification , structure assembly and model selection is made and hard(complex)range residue lacks modeling of its tertiary structure(3D) and accuracy is evaluated by quality of template chosen[2]. Use of Meta- Heuristics[8] dynamic approach solves the complexity of structural protein prediction and disorder modeling.

7. CONCLUSION

The clear distinction of protein structure is varied in levels of organization are thoroughly studied which have significant deficiency in sequence alignment will have adverse effects in experimental research for various enzymes. The modeling of newly determined genes or proteins may have improper structural alignment.

Knowledge base applies various alignment procedures to assess those gene or protein with acknowledged template and update database whenever it forms unique sequence(structure). Likewise genes also need alignment for their nucleotide sequence. Sequence Alignment is fundamental for prediction of accurate protein structural organization in secondary structure and in latter stages. Requirement of meta-heuristic approach and intelligent methods for better machine learning which tends to adapt rapidly provides a alignment solution within time frame and preserves the cost that will persist merely in future.

REFERENCE

[1] Matsuyuki Shirota. “Protein structure modeling guided by homology and hydrophobic residue interactions”. Tohoku University mshirota@hgc.jp.

[2] T. Brunette*, R. Wang*, D.E. Kim*, F. DiMaio, S. Ovchinnikov, K. Jung, H. Kamichetty, Y. Song, F. Khatib, C. Miles, J. Thompson, D. Baker.” Modeling of Protein Structures Using Rosetta in CASP 10”. University of Washington, Seattle, WA dbaker@uw.edu.

[3] Söding,J. (2005) “Protein homology detection by HMM-HMM comparison.Bioinformatics”.21(7):951-60 http://swissmodel.expasy.org/workspace/

[4] Ambrish Roy, Yang Zhang. ”Protein Structure Prediction”. John Willey and sons Ltd 2012.

[5] Hung-Pin Peng and An-Suei Yang ,“Modeling protein loops with knowledge-based prediction of sequence-structure alignment” , Oxford journals 2007.

[6] Wolfgang Kabsch, Christian Sander.” Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features”. Biopolymer Issue published online: 1 FEB 2004 John Willey and sons Inc.

[7] Yang Zhang.” Progress and challenges in protein structure prediction”. Published online 2008 April 22. doi: 10.1016/j.sbi.2008.02.004.

[8] Sayantan Mandal, Nanda Dulal Jana. “Protein Structure Prediction using 2D HP Lattice Model Based on Integer Programming Approach”. 2012 International Congress on Informatics, Energy and Applications-IEEA 2012, IPCSIT vol.38 (2012).

[9] http://www.wwpdb.org/documentation/

[10] http://predictioncenter.org/

[11] http://www.rcsb.org/pdb/home/home.do