International Journal of scientific research and management (IJSRM)
||Volume||2||Issue||10||Pages||1535-1541||2014||
Website: www.ijsrm.in ISSN (e): 2321-3418

Computational Investigation for Secondary Structure Prediction

M.Rithvik[1], G.Nageswara Rao[2]

[1]aditya Institute Of Technology And Management,
Tekkali
Rithvikmadugula@Email.Com

[2] Aditya Institute Of Technology And Management,
Tekkali
Gnraoaitam@Email.Com

Abstract: Protein structure prediction play a key role in every living organisms All living organisms are made up of cells and each cell in its turn consists of certain protein consequences which exercise an important role in catalyzing the chemical reactions. So, a study of a protein structure becomes a search lamp in the diagnosis of a disease .when the percent identity between two protein sequences falls below 33%,it necessities to carry out the analysis of protein secondary structure. Of the several methodologies developed to analyze the protein secondary structure, two methods proved to be sound-DSSP(Dictionary of Secondary Structure of Proteins) and GOR(Garnier,Osguthrope and Robson),Even though the prediction accuracy of GORV is 73.5% due to hazards in its implementation ,GORIV is generally used in spite of its accuracy being only to 64.4%

Keywords:Protein,Residue,DSSP,GOR

1. Introduction

Analysis of secondary structures of proteins is carried out when the percent identity between two sequences fall below 33%. The world wide web hosts many programs that are able to recognize the number of residues that appear in secondary structural elements such as alpha helices, beta strands etc, but the output of such programs is enormous. However, none of the programs are able to reproduce experimental values and are accurate up to a maximum of 70-80%. Hence in this scenario, a java based program was written that takes the outputs of DSSP (Dictionary of secondary structure of proteins) and GOR (Garnier, Osguthorpe and Robson) programs and returns the number of states and conditions when the state was true. The program runs fast and reliable when the outputs of such programs are lengthy and unable to read.

In biochemistry and structural biology, secondary structure (1) is the general three-dimensional form of *local segments* of biopolymers such as proteins and nucleic acids (DNA/RNA). It does not, however, describe specific atomic positions in three-dimensional space, which are considered to be tertiary structure.Secondary structure is formally defined by the hydrogen bonds of the biopolymer, as observed in an atomic-resolution structure. In proteins, the secondary structure is defined by patterns of hydrogen bonds between backbone amide and carboxyl groups (sidechain-mainchain and sidechain-sidechain hydrogen bonds are irrelevant), where the DSSP definition of a hydrogen bond is used. In nucleic acids, the secondary structure is defined by the hydrogen bonding between the nitrogenous bases.

The hydrogen bonding is correlated with other structural features, however, which has given rise to less formal definitions of secondary structure. For example, residues in protein helices generally adopt

backbone dihedral angles in a particular region of the Ramachandran plot; thus, a segment of residues with such dihedral angles is often called a "helix", regardless of whether it has the correct hydrogen bonds. Many other less formal definitions have been proposed, often applying concepts from the differential geometry of curves, such as curvature and torsion. Least formally, structural biologists solving a new atomic-resolution structure will sometimes assign its secondary structure "by eye" and record their assignments in the corresponding PDB file.

The rough secondary-structure content of a biopolymer (e.g., "this protein is 40% α-helix and 20% β-sheet.") can often be estimated spectroscopically. For proteins, a common method is far-ultraviolet (far-UV, 170-250 nm) circular dichroism. A pronounced double minimum at 208 and 222 nm indicate α-helical structure, whereas a single minimum at 204 nm or 217 nm reflects random-coil or β-sheet structure, respectively. A less common method is infrared spectroscopy, which detects differences in the bond oscillations of amide groups due to hydrogen-bonding. Finally, secondary-structure contents may be estimated accurately using the chemical shifts of an unassigned NMR spectrum.Every human body is consistuted with a certain amount of cells.The functioning of the human body depends upon the functioning of cells.Every cell consists of certain protein sequences.Proteins play a vital role in catalyzing the chemical reactions in all the living organisims. Proteins are formed by the combination of several amino acids.the information flow from a DNA sequence to the protein structure is as follows:The information flow from DNA to RNA first.From RNA we acquire a protein sequence .This protein sequence helps us in predicting a protein structure.So,protein sequence plays a key role in predicting the structure of a protein.By knowing the structure we can find the function of the protein.

1.1Necessity to predict the protein structure

Predicted structures can be used in structure based drug design.It can help us understand the effects of mutations on structure or function.Structural knowledge brings understanding of function and mechanism of an action.It Can help in prediction of function

Secondary structure in proteins consists of local inter-residue interactions mediated by hydrogen bonds, or not. The most common secondary structures are alpha helices and beta sheets. Other helices, such as the 310 helix and π helix, are calculated to have energetically favorable hydrogen-bonding patterns but are rarely if ever observed in natural proteins except at the ends of α helices due to unfavorable backbone packing in the center of the helix. Other extended structures such as the polyproline helix and alpha sheet are rare in native state proteins but are often hypothesized as important protein folding intermediates. Tight turns and loose, flexible loops link the more "regular" secondary structure elements. The random coil is not a true secondary structure, but is the class of conformations that indicate an absence of regular secondary structure.

Levels of protein structure

Primary structure - the amino acid sequence of the peptide chains.

Secondary structure - highly regular sub-structures(6) (*alpha helix* and *strands of beta sheet*) which are locally defined, meaning that there can be many different secondary motifs present in one single protein molecule.

Tertiary structure - three-dimensional structure of a single protein molecule; a spatial arrangement of the secondary structures. It also describes the completely folded and compacted polypeptide chain.

Quaternary structure - complex of several protein molecules or polypeptide chains, usually called protein subunits in this context, which function as part of the larger assembly or protein complex.

In addition to these levels of structure, a protein may shift between several similar structures (7)in performing its biological function. In the context of these functional rearrangements, these tertiary or quaternary structures are usually referred to as chemical conformation, and transitions between them are called conformational changes.

The primary structure is held together by covalent or peptide bonds, which are made during the process of protein biosynthesis or translation. These peptide bonds provide rigidity to the protein. The two ends of the amino acid chain are referred to as the C-terminal end or carboxyl terminus (C-terminus) and the N-terminal end or amino terminus (N-terminus) based on the nature of the free group on each extremity.

The various types of secondary structure are defined by their patterns of hydrogen bonds between the main-chain peptide groups. However, these hydrogen bonds are generally not stable by themselves, since the water-amide hydrogen bond is generally more favorable than the amide-amide hydrogen bond. Thus, secondary structure is stable only when the local concentration of water is sufficiently low, e.g., in the molten globule or fully folded states.

The peptide bond

Two amino acids can be combined in a condensation reaction. By repeating this reaction, long chains of residues (amino acids in a peptide bond) can be generated. This reaction is catalysed by the ribosome in a process known as translation. The peptide bond is in fact planar due to the delocalization of the electrons from the double bond. The rigid peptide dihedral angle, ω (the bond between C_1 and N) is always close to 180 degrees. The dihedral angles φ (the bond between N and Cα) and psi ψ (the bond between Cα and C_1) can have a certain range of possible values.

2.Methodology

The secondary structure of protein mainly constitutes of two elements:i)Alpha helix ii)Beta sheet.The regular patterns of the H bonds are formed between neighboring amino acids and the amino acids have similar φ and ψ bondsThis secondary structure has three regular forms 1.Alpha helix 2.Beta sheet 3.loops.We give the input as the amino acid sequences and the results of output will be the predicted structure.A protein consists of about 32% alpha helices 21% beta sheets and 47% loops or non regular structure that is it may be a motif or a fold.In the above figure the rotation angle around the N-c□

is called ¢ and the rotation angle between c-cα is called psy.The proteins that are evolved from a common ancestor are called homologous proteins.Generally we used to predict the protein structure for a non homologous proteins from the protein data bank

The DSSP

DSSP stands for Dictionary of secondary structures.We generally take a non homologous protein structure to calculate the protein structure.Dssp is a program that is used to convert or transfer the number of states that is it is used to decrease the number of states of residues from eight to three[3].The reason behind the decrease of the number of states is to predict the secondary structure that consist of α helices, β sheet, coil and fold.It was first designed by Wolfgang Kabsch and Chris Sander to standardize the secondary structure assignment.DSSP is a database of secondary structure assignments for all protein entries in the Protein Data Bank(PDB)

Brief history: The original DSSP application was written between 1983 and 1988 in the programming language PASCALThis pascal code is automatically converted but maintaining this proved to be a lot of trouble-taking

Working: The Dssp program works by calculating the most likely secondary structure assignment that was given in the 3D structure of a protein[7]It does this by reading the position of the atoms in a protein followed by calculation of the H bond energy between all atoms.The best two H-bonds for each atom are then used to determine the most likely class of secondary structure for each residue in the protein

Structure: The structure of the DSSP is as follows:

G = 3-turn helix (3_{10} helix). Min length 3 residues.
H = 4-turn helix (α helix). Min length 4 residues.

I = 5-turn helix (π helix). Min length 5 residues.

T = hydrogen bonded turn (3, 4 or 5 turn)

E = extended strand in parallel and/or anti-parallel β-sheet conformation. Min length 2 residues.

B = residue in isolated β-bridge (single pair β-sheet hydrogen bond formation)

S = bend (the only non-hydrogen-bond based assignment)

DSSP obtained from non redundant PDB_select dataset,secondary structure assigned by DSSP into eight conformational states to three states H=HGI,E=EB,C=STC

GOR:

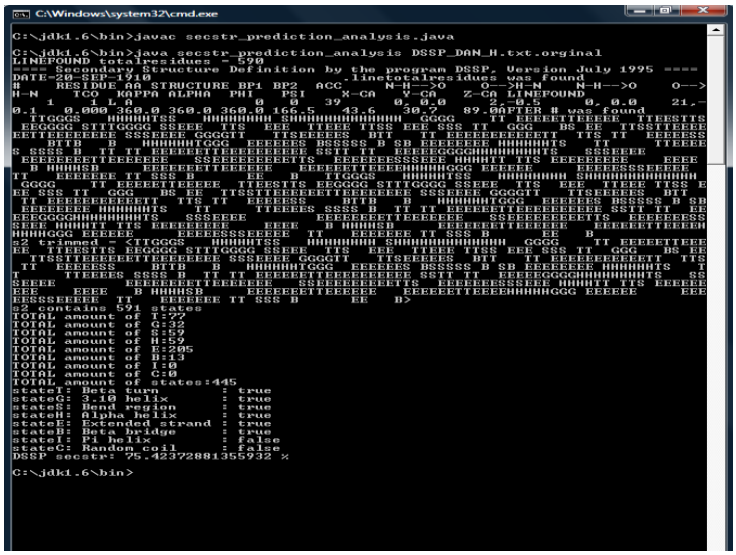
The GOR method was first developed by Garnier,Osguthrope and Robson[5].It is an information theory based method that uses a more powerful probabilistic techniques of Bayseian inference.This method not only takes into account the probability of each amino acid having a particular secondary structure but also the conditional probability of the amino acid assuming each structure given the contributions of its neighbors.This method assumes that amino acids upto eight residues on each side that influence the secondary structure of the central residue.This program has many versions but here we are going to write the program in the fourth version.The algorithm makes use a sliding window of 17 amino acids.All possible pairs of amino acids in this

window are checked for their information content as to predict the central amino acid by comparing the set of 266 other proteins of known structure.In this version we are going to find the certain pairwise combinations of amino acids that are present in the region called as Flanking Region.If a particular amino acid is surrounded by residues that prefer to be a helics it is likely to be a helix.Even if its individual helical preference is low.This method helps in considering a propenties of a single residue position dependent propensities for helics,sheet and turn has been calculated for all types of residues

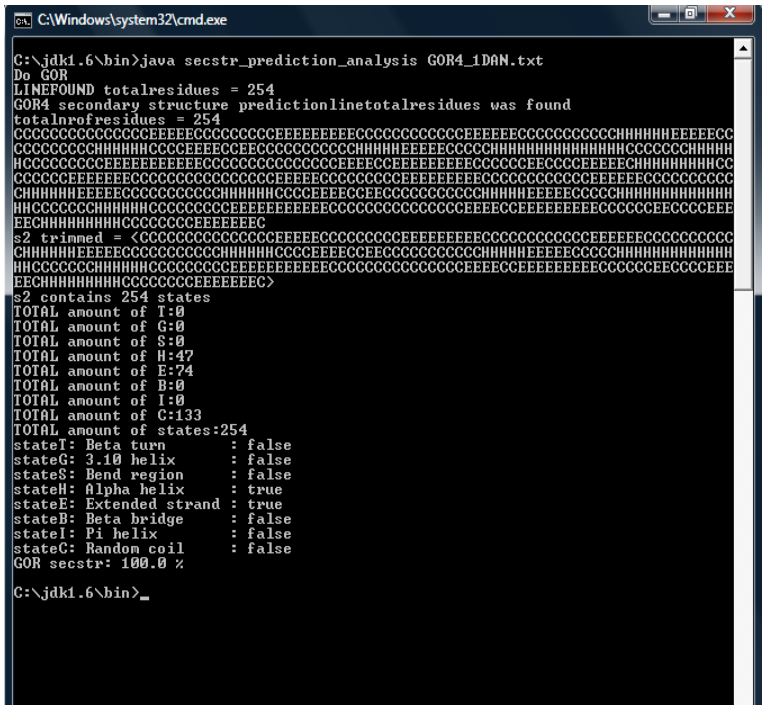
GORV::This method is a bit improved method when compared to GOR IV method[8].The improvement was made in the algorithm.It was inclusion of evolutionary information using PSI-BLASTMultiple alignments are gennerated using PSI-BLAST after five iteratioms based on the non redundant database.The GOR V method prediction takes a longer time for the multiple alignments that is this kind of alignments takes much more time for hits

Accuracy of the two methods: The prediction accuracy of the GOR IV method is 64.4% and the prediction accuracy of GORV method is73.5%.Eventhough GOR V method has more prediction among accuracy people generally use the GOR IV methodThis means that GORV method is difficult to implement

Output of DSSP

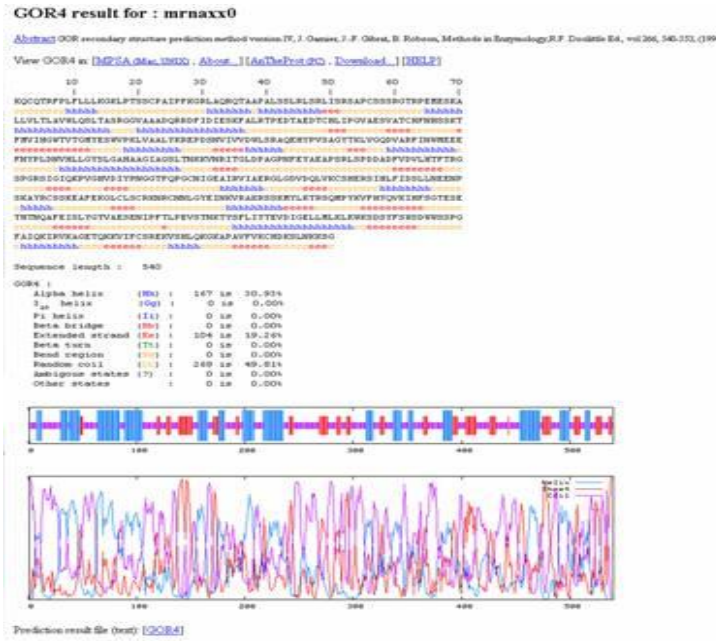


Output of GOR



Results and Discussion

Here we are going to display the results of DSSP and GOR



Output of DSSP

These results show that here we are going to predict the secondary structure of a protein by using these two methods The protein structure can be visualized by using the various tools

SWISS-PDB VIEWER: This is an apllicarion that provides a uer friendly interface allowing to analyze several proteins at the same time.This viewer helps us to visualize a protein secondary structure that was out when performing the DSSP and GOR programs

Accuracy of Prediction of Protein Structure: The accuracy of protein structure prediction depends upon the percentage of correctly predicted residues in sequences of known structure called Q3

A method to calculate the accuracy is to calculate a correlation coefficient for each type of predicted secondary structure. The coefficient indicating the success of predicting residues in the α helical configuration cα is given by:

Cα=(pαnα-uαoα)/squarerootof
[nα+uα][nα+oα][pα+uα][pα+oα])

Where pα is the correct prediction

nα is the negative prediction

oα is the overpredicted positive prediction

uα is the number of unpredicted residues

Conclusion

Here we are writing a program using DSSP and GOR methods to predict the protein secondary structure.Apart from the programs we are going to visualize the protein secondary structure by using the SWISS-PDB viewer.The performance of these two programs gives us an accuracy of 60-70% when compared to the other programs..Eventhough these two programs gives us less accuracy they are easy to write and implement in java.Eventhough new versions of the programs are giving more accuracy people generally prefer these programs .Because these programs are easy in their implementation and robust in nature

Future scope

The further development of these programs helps us in predicting the more accurate protein secondary structures.The programs that will be developed in future must consider these programs in order to predict the secondary structure of the protein.The further applications of these programs must consider the Ramchandran plot in mind for their prediction accuracy

References

[1] Frishman D., Argos P. (1995). "Knowledge-based protein secondary structure assignment". Proteins23(4):566–579doi:10.1002/prot.340230412. PMID 8749853.

[2] Richards F. M., Kundrot C. E. (1988). "Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure". Proteins 3 (2): 71–84. doi:10.1002/prot.340030202. PMID 3399495.

[3] Kabsch W., Sander C. (1983). "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features". Biopolymers 22 (12): 2577–2637. doi:10.1002/bip.360221211. PMID 6667333.

[4]" Prediction of the Number of Residue Contacts in Proteins Piero Fariselli and Rita Casadio CIRB Biocomputing Unit and Lab. of Biophysics," Dept. of Biology University of Bologna via Irnerio 42, 40126 Bologna Italy

[5]. "A Java Based Protein Secondary Structure Prediction" ProgramG. Nageswara Rao, Allam Appa Rao

[6]. "literature survey of prediction of protein"

[7]. "DSSPcont: continuous secondary structure assignments for proteins" Phil Carter1,2,4,*, Claus A. F. Andersen1,3 and Burkhard Rost1,2,5 Nucleic Acids Research, 2003, Vol. 31, No. 13 3293–3295 DOI: 10.1093/nar/gkg626