Summary and Conclusions

It has been proved in 1973 that the amino acid sequence is responsible for its tertiary structure. The recent release of Protein Databank have more than 3000 proteins whose three-dimensional structure is known experimentally. We know the amino acid sequence of these proteins as well as tertiary structure. Prediction of tertiary structure from primary sequence remains a major goal of researchers Over the years our understanding in protein tertiary structure has increase, its still far from satisfactory. Protein tertiary structure prediction is one of the major unsolved problem of molecular biology as well as one of the major problem whose solution would make revolution in life sciences. The protein tertiary structure prediction methods can be divided in four categories i) Free energy minimization based methods ii) Homology based methods; iii) Threading protein sequence from structure motifs; iv) Semiempirical methods in which the structure is predicted in hierarchical manner.

Energy minimization methods offer the promise of rigor treatment of the inter- and intramolecular forces in protein structures and therefore, in principle one can predict the detailed atomic coordinates of protein. However, several practical detail of these methods remain unsolved, the available potential functions do not provide an accurate representation of energy states and optimization algorithms can not sample the entirety of conformational space. In practice due to notorious local minima problem, its very difficult to find the real global minimum. Therefore, in most of the cases energy minimization methods were used merely to refine a protein structure determine by X-ray crystallographic and NMR techniques. Homology based methods provide the insight of the protein if homologous protein of known structure is available. The main limitation of homology based methods is that it can not be used to predict tertiary structure of protein which has remote homology with known protein structure available. The threading based methods are successful in a number of cases but unfortunately these methods can not be used to predict a structure for which no example is available. In semiemperical approach the structure is predicted in hierarchical manner. Instead of predicting the tertiary structure of protein directly from amino acid, first secondary structure is predicted from amino acid sequence then secondary structure are used to predict the tertiary structures. The main limitation of these methods are i) the accuracy of secondary structure prediction of available methods is low; and ii) the secondary structure information alone is not adequate to built the tertiary structure of protein.

In this study, semiempirical approach of protein structure prediction is adopted to predict the tertiary structure of protein from its amino acid sequence. First, we studied the available protein secondary structure prediction methods in detail to understand merits and demerits of these methods in depth. An improved method have been developed to predict the protein secondary structure from their amino acid sequence. The secondary structure prediction method developed during this study, is most accurate method as yet reported, which is suitable for proteins who has no homology with known protein structures and sequences.As the secondary structure prediction is intermediate step towards the prediction of tertiary structure. Therefore, we study whether the secondary structure information alone predicted by our method is sufficient to predict the tertiary structure of proteins. We observed two problems in predicting tertiary structure from secondary structure information alone i) the accuracy of prediction of our method is better than previous methods but still quite low (69 %); ii) to predict the 3D structure one needs to translate the secondary structure residue in Cartesian coordinate or more convenient into internal coordinates (dihedral angles),.however its impossible to obtain the dihedral angle for residues predicted in coil. There are more than 50 % residues whose predicted in state coil. The above observations show that the secondary structure predicted using presently available methods including our method, alone is not adequate to predict the tertiary structure.

To overcome this problem, we develop a new method for predicting the sixty-four conformational states of protein backbone, which allow to predict the dihedral angle (ϕ & ψ) corresponding to each residue. Using this approach the dihedral angles of protein backbone were determined, which were further modified by using secondary structure information predicted by our method. Finally, the crude model of protein were constructed from dihedral angles information of backbone, which was refined using free energy minimization based method. The procedure followed in this study in brief can be describe as below.

A computer program called CPSSD has been developed for extracting protein secondary structure information from DSSP files to create protein secondary structure database in compact and user friendly format. CPSSD extracts protein secondary structure from DSSP files and generates the PSSD files. The program provides the option to group the protein secondary structure states into classes, as well as allows to extract any or all chains of protein. Protein secondary structure database in compact form has been derived from DSSP database using program CPSSD of the proteins available in Protein Databank, whose structure is known at resolution 3.2 Å or higher.

To evaluate the quality of protein secondary structure prediction and for appraising the secondary structure prediction methods, a computer program called ASSP has been developed. The program compares the predicted and observed secondary structure of the protein and compute the parameters required for evaluating the quality of secondary structure prediction. The ASSP program compute the number of parameters, like i) accuracy of secondary structure prediction (e.g. percentage of correctly predicted residues in each secondary structure state, percentage of total correctly predicted residues in all the secondary structure states); ii) Probability index (probability of correct prediction of a residue in given secondary structure state), iii) Matthew's correlation coefficients and iv) entropy information. The quality of secondary structure prediction of different methods can be accessed using the above parameters.

In this report, the memory based reasoning method and parallel example-based learning system (PEBLS) method (Salzberg and cost, 1992) have been examined to understand the merits and pitfalls of these methods in protein secondary structure prediction. The strong points of both were utilized to develop a powerful method, for predicting the protein secondary structure, by optimizing the parameters, which we referred as optimized nearest neighbor method (ONNM). It also estimates the probability distribution for the three states, and the probability of correct prediction for each residue in a test protein. ONNM was combined with neural network, which further improves the accuracy of secondary structure prediction. ONNM was applied on 113 protein chains (Zhang et al. 1992) to predict the secondary structure of proteins having less than 50 % homology in training and test data set. It resulted in a prediction accuracy of 67.6 % for three states was achieved.

In the folding of protein certain patterns of secondary structure are more responsible for a given fold rather than single secondary structural state of a local amino acid. This implies that there is considerable correlation between secondary structure state of a residue to the secondary structure pattern in a given protein. Here, this idea is applied to take their correlation into account, at least in part by repredicting secondary structure of protein from predicted secondary structure information. Based on above approach a structure to structure prediction (SSP) method has been developed. SSP method is similar to MBR method, except that the input in case of MBR consists amino acids where as input in case of SSP is predicted secondary structure (output of ONNM).

Finally, a PSSPR (protein secondary structure prediction from residues) computer program has been developed, which first, predicts the secondary structure using ONNM and neural network (combined) and then repredicts the secondary structure using SSP method. Like ONNM, PSSPR was applied on 113 protein chains (Zhang et al. 1992) to predict the secondary structure of proteins which have less than 50 % homology in training and test data set which resulted in a prediction accuracy of 69 % for three states, which is the maximum accuracy reported as yet from sequence information alone. The main advantages of the PSSPR described in this report are i) it predicts the protein secondary structure at fairly high accuracy (69.0 %) using only sequence information; ii) prediction accuracy is independent of the availability of homologous proteins in SwissProt database (unlike, the methods based on multiple sequence alignment approaches (Rost and Sander, 1993; Levin et al. 1993); iii) it computes the probability of correct prediction of each residues and the probability distribution over the three states, which serve as a reliability index of prediction; iv) PSSPR also predicts with a high accuracy, the secondary structure of a protein whose homologous proteins of known structure are available and, v) it is a simple, flexible and a straight forward.

In chapter 5 a procedure is described for predicting the dihedral angles of the backbone (ϕ and ψ angles) of a protein molecule from the data base of known structures. The procedure is basically an adaption of a published secondary structure prediction scheme, applied to predict the conformational states rather than to the secondary structural types. For our study we divided the $\phi - \psi$ angle map in sixty four states referred as conformational states. Then the conformational state prediction method has been developed to predict conformational state corresponding to each residue. The sixty-four conformational state prediction method developed here is based on strategy adopted by secondary structure prediction methods. Here, each conformation state is represented by single unique value of dihedral angles. This way of dihedral angle of protein backbone can be predicted which include all effects of local sequence and are "context sensitive". These conformation state prediction can be used to predict the three-dimensional structure of short polypeptides.

The protein backbone were constructed in two steps, in first phase the conformational state were predicted using PPBCS of protein in protein data set as describe above. Then the dihedral angle of backbone is determined by substituting the conformational states by its representative value of dihedral angle. In second phase the predicted secondary structure information of protein is used to rectify the backbone dihedral angles predicted in first phase. Here we only utilize helix information of secondary structure predicted using PSSPR method. As describe in chapter 3, our method predict the α-helix better than 60 %. The dihedral angles of residues whose predicted in helix by PSSPR in α-helix were replaced by new dihedral angles ϕ = - 60.0 and ψ = - 50.0. This way the predicted secondary structure information is used to improve the quality of prediction of dihedral angles.

The protein backbone 3D coordinates were determined from dihedral angle of protein backbone. In our case we kept the angle between peptides (ω) constant ($\omega = 180$), the bond length were chosen between atoms of amino acids as constant (standard lengths). This way the 3D coordinate of protein were determined in PDB format. These proteins can be displayed by number of programs including Rasmol, MidasPlus . We utilize Rasmol and MidasPlus to present the 3D-structure of protein backbone on screen as well as on lasers printer. This protein model is called crude model of protein. The crude model of protein obtained using the above method was refined using the free energy minimization techniques. We used AMBER program for minimizing free energy of protein crude model. The refined model were than compared with observed structure of protein.

In conclusion, we have attempted to build the protein model from amino acid sequence of protein which does not have the homologous protein of known structure in PDB. The semiempirical approach have been adopted for predicting the tertiary structure of protein. The main achievements of this study may be classify as i) a protein secondary structure prediction method is developed and was applied to predict the secondary structure of proteins. It achieves on an average 69 % accuracy from amino acid sequence information, which is the maximum accuracy of any method reported as yet, ii) A new sixty four state conformational prediction method has been developed for predicting the dihedral angle of protein backbone. This method was applied on 69 highly resolved proteins (resolution better than 2.5 Å) and achieved 32.0 % accuracy which is quite high, iii) the tertiary structure of proteins were obtain from dihedral angle information predicted using PPBCS method, which was further improve by using secondary structure information and iv) tertiary structure was refined using energy minimization approach. To automate the procedure of protein structure prediction, number of computer programs have been developed during this study, like i) CPSSD for extracting secondary structure information from DSSP files, ii) ASSP for calculating the quality of secondary structure prediction, ranking the secondary structure prediction methods, iii) ONNM, SSP, PSSPR for predicting the secondary structure of proteins from its amino acid sequence, probability distribution over the three state and, iv) PPBCS for predicting the sixty four conformational states and prediction of dihedral angle of a protein backbone. The development of our computer programs would be of great aid to protein engineers who are engaged in understanding the relationship between structure and function of proteins, especially those for which detailed structure at atomic level is still not avilable.