Protein Structure Prediction

Jayanthi Sourirajan

Final Project

Computational Molecular Biology BIOC218

June 4, 2004

Protein Structure Prediction

Proteins are building blocks of life. Proteins exhibit more sequence and chemical complexity than DNA or RNA. A protein sequence is a linear hetero polymer made up of one of the 20 different amino acids. They perform a wide variety of functions in the living organism, playing various catalytic, structural, regulatory and signaling roles required for the cellular development, differentiation, replication and survival. The key to the wide variety of functions exhibited by the individual proteins is not its linear sequence but its three dimensional structure. The knowledge of the 3D structure is useful for rational drug design, protein engineering, detailed study of protein –bio-molecular interactions, study of evolutionary relationship between proteins or protein families etc.

The 3D structure of proteins can be solved by 1) Experimental methods, or 2) Structure prediction. Solving structures experimentally is very hard. Solving through X-ray crystallography produces very good results but we need to have a very pure protein sample which must form crystals that are relatively flawless. Solving through NMR is limited to small soluble proteins. In addition, large scale sequencing projects like the human genome project produce protein sequences at a very fast rate. Thus there is a huge gap between the number of known protein sequences and the number of solved structures. Protein structure prediction aims at reducing this gap.

Protein structure prediction is not as easy as it sounds. There are a number of facts that exist that make structure prediction a difficult task.

- There are a number of ways a protein could fold to attain the native state.
- The physical basis of protein structural stability is not fully understood.
- The primary sequence may not fully specify the tertiary structure. There are proteins called chaperones that induce the protein to fold in specific ways. E.g.: the chaperons help in the folding of the heat shock proteins.

Steps Involved in Prediction

Although there are many methods and algorithms to predict the structure, the general steps involved can be summarized as follows

Prediction in 1D involves

- Prediction of secondary structure(SSP)
- Prediction of solvent accessibility
- Prediction of trans-membrane helices

Prediction in 2D involves

- Prediction of the inter-residue and strand contacts

Prediction in 3D involves

Searching the database to find a suitable template for modeling.

If the sequence identity is $\geq$ 25% then modeling is carried out through homology modeling. If the sequence identity is <25 then the model is obtained through fold recognition or threading. If no suitable template is found then structure is predicted using Ab-initio prediction.

Secondary structure prediction

The secondary structure of a protein has three regular forms, an alpha helix, a beta sheet and loop or turns. SSP involves predicting the secondary structure state for each amino acid residue. The most widely used accuracy index for SSP is the *three state accuracy* which gives the percentage of the correctly predicted residues in any of the three states.
$Q=(P_\alpha + P_\beta + P_{loop})/T \ \times 100$
Where T is the total number of residues, P_α is the number of residues predicted correctly to be in alpha helix, P_β is the number of residues predicted correctly to be in beta sheet and P_{loop} is the number of residues predicted correctly to be in loops or turns. The quality of the prediction is assessed by the number of segments in a protein, the average segment length and the distribution of the number of segments with the length.

First generation SSP

Most of the methods in this generation were based on single residue statistics. In the Chou-Fasman method developed in 1974, the residues were aligned according to their ability to form or break a secondary structure. They were classified into strong formers, weak formers, formers, indifferent formers, strong breakers and breakers. He identified an alpha helix by locating a cluster of 4 formers or strong formers within 6 residues which was extended in both directions until terminated by a tetra-peptide with an average alpha propensity of less than 1. For beta sheets, he looked in a cluster that had 3 out of 5 formers and strong formers and then extended in both directions. Turns were predicted in a window of 4 residues first with an overall score that is significantly greater than that for helix or strand and then by a position specific score for each of the 4 residues in reverse turn.

The GOR algorithm (Garnier, Osuguthorpe, and Robson) was developed in 1978 to improve upon the Chou-Fasman method. The GOR method not only took the relative occurrence of residue in a particular element of the structure but also took the accuracy of the data into consideration. The method first analyzed the protein of a known structure based on the query. It then considered the effect that a residue has on the secondary structure of another residue say 'n' residues from it. This gave the likelihood of a residue and its neighbors being in particular secondary structure. This information is used to scan the protein using a sliding window of 17 residues and assigning a value to each residue which expresses the likelihood of it being in a particular secondary structure.

Second generation SSP

These methods depended on sequence structure relationship and modeled using algorithms based on statistical information, physio-chemical properties, sequence patterns, multilayered neural networks, graph theory, multivariate statistics and nearest neighbor algorithm. The neural network based algorithm by Qian and Sejnowksi predicted the alpha helix and beta sheet of 15 test proteins. The neural network had 3 layers with 40 hidden layers and 13 input residues. The output of the first network was fed into a second neural network.

Although the first generation method gave an accuracy of 50-60% and the second generation methods gave an improved accuracy of about 70%, these method had some drawbacks. This may be due to 2 reasons namely the secondary structures differ even between crystals of the same protein. Moreover the long range interaction plays a role in secondary structure formation.

Third generation SSP

These methods were superior in terms of accuracy and also dealt with the drawbacks of the other two generation. Their accuracy is about 76%. In PREDATOR the secondary structure propensities is based on both local and long range effects, utilizing the similar sequence information in the form of pair wise alignment fragments and relying on a large collection of known proteins. PHD developed by Rost and Sander in 1993 is composed of several cascading neural networks. In the neural network, aligned homologous sequences of known structures are used to "train" the network, which then can be used to predict the secondary structure of the aligned sequences of the unknown protein. The homologous sequences are determined by BLAST and are aligned using MaxHom. In the first step, the occurrence of various residues in a window of 13 amino acids is correlated with the secondary structure of the central residue. In the second step (structure-structure layer), the output from the first layer in a window of 17 residues is used to predict the secondary structure of the central residue. In this case, the network will be trained not to predict unreasonably short segments of secondary structure. Another step consists of averaging the output from independently trained network. Others that predict via neural networks and PSSM are PHDhtm, TMAP, and TMpred etc. Some of the best secondary structure prediction programs are PHD with an approximate 72% accuracy, Jpred with about 73-75% accuracy, PREDATOR with about 75% accuracy, Sam T99 with about 74% accuracy.

One of the difficulties in predicting secondary structures at high accuracy is the presence of non-local contacts in protein folding. This is because amino acids which are quite distant in the primary sequence may be close to each other in the 3D structure as the protein folds. Bayesian network which is based on parameterization of the sequence structure relationship in terms of structural segments can be used for predicting secondary structures.

Prediction of solvent accessibility

Usually non-polar amino acids tend to be buried inside the protein and the polar amino acids are in contact with the solvent. The solvents in a cell are usually vehicle for transporting metabolites to protein active sites. Hence determination of solvent accessibility is an important to find out how much of a particular amino acid is in contact with the solvent and how frequently does the amino acid of that type occur in a site with that degree of accessibility. The accuracy used is a *two state per residue* accuracy depending on whether a residue is exposed (relative solvent accessibility >16%) or buried (relative solvent accessibility < 16%). Although residue solvent accessibility is not as well conserved within a structural family as secondary structure, prediction can be improved by including evolutionary information. A neural network prediction of accessibility has been shown to be superior to simple hydrophobicity analyses. Prediction of solvent accessibility has been used successfully in prediction based threading as well. The average accuracy of predicting the solvent accessibility is around 70-75%

Prediction of Trans-membrane helices

There are two main classes of membrane protein

- Protein with long about 17 to 27 residue forming transmembrane helices that spans the membrane
- Porins which form a 16 strand beta barrel fold that forms a pore through the membrane.

Predicting the location of the trans-membrane helix is a task comparable to the secondary structure prediction. Accuracy has been improved by combing hydrophobicity analyses, statistical information and multiple sequence information. The hydropathic profiles of a protein are calculated by assigning each amino acid a "hydropathy index" and then averaging the values along the peptide chain. The values assigned can be either Hoop-wood values or the Kyle-Doolittle values. Another way is by calculating the hydrophobic moment which is a measure of the amphilicity or asymmetry or hydrophobicity of the polypeptide chain. An alpha helix has a periodicity of 3.6 , hence the residues at position i,i+3,i+4,i+7 etc will lie on the exposed face of the helix . Similarly the beta strands that are half buried in the protein core will tend to have hydrophobic residues at i,i+2,i+4, and polar residues at i+1,i+3 etc. Some of the programs that predict the topology of membrane proteins are TMHMM, PHDhtm ,TOPPred2 etc.

Prediction of inter-residue and strand contacts

The NMR spectroscopy produces experimental data of distances between the protons. Using these distances, the 3D structure can be reconstructed using distance geometry or molecular dynamics. Hence if the secondary structure can be predicted successfully, some fraction (helices and strands which can be assigned based on hydrogen bonding pattern) of the contacts is known and its 3D structure can be determined by distance geometry. But the contacts predicted by secondary structure are short range contacts. For application of distance geometry, contacts between residues far apart in sequence should also be considered. One of the methods to predict such long range inter-residue contacts is by analyzing correlated mutations. Other methods use statistics, mean-force potentials, or neural networks.

One way to simply the problem to predict inter-residue contacts is by predicting the contacts between residues in adjacent strands. This is because such interactions are more specific than the long range contacts and thus easier to predict. One method to predict the inter-strand contact is by mean force potentials which can be improved by using multiple sequence alignment information.

Prediction in 3D

The tertiary structure of proteins involves the folding of the secondary structural elements. The physical properties that determine fold are the backbone rigidity, interaction between the amino acids which include the electrostatic interaction, the vander-waals interaction, hydrogen and disulphide bonds and interaction with water. There are three methods for protein structure prediction namely 1) homology modeling 2) Fold recognition or threading and 3) Ab-initio method. All these methods involve searching the database for a homologue to the target protein. If the sequence similarity between the template and the target is ≥ 25% then comparative or homology modeling is carried out. If the sequence similarity is < 25% the prediction is done thorough fold recognition or threading. If no suitable homologue it found in the database then the 3D structure is predicted through ab-initio predictions. Template selection and alignment accuracy have a large impact on the model accuracy. If there is a 90% or more sequence similarity between the template and target then the errors of the final model are as low as that obtained from X-ray crystallography except for some side chain errors. For template sequence identity between 30-50%, 90% of the main chain can be modeled with 1.5 A^o RMSD. The quality of the model is limited by side chain packing, core distortion and loop modeling. For template sequence identity <25% accuracy of the alignment is the main limiting factor for errors.

Critical Assessment of Structural Prediction (CASP)

The idea to test the different prediction methods in a blind manner which enables a direct comparison of a protein model to its real structure was the basis of CASP experiments initiated by John Moult in 1994. CASP is held every two years and the next one CASP6 is to be held in December 2004. In CASP a few dozen proteins of known sequence but unknown structure are used as prediction targets. Contestants are to predict the structure of protein using different algorithms and different methods- homology, fold recognition and ab-initio. Subsequently once the 3D structure is released an assessment of the accuracy of the predictions is carried out. CASP concludes with a meeting in Asilomar to discuss the results.

Servers

There are various servers available for performing these modeling. The models generated from each one of them may be different. This actually depends on the templates chosen, the alignment of the query protein with the template and also the algorithm or the method used in their prediction. The Live Bench Project is a continuous benchmarking program. Every week it evaluates the sensitivity and specificity of the different servers. Some of the servers that did well in both comparative and fold recognition modeling in CASP4 and CASP5 were 3DPSSM,FFAS,mgenthreader, inbgu and samT99 .

Meta servers

Metaservers are servers that make their predictions based on the results of two or more different methods. Some simply look for consensus prediction between several methods while others calculate their own scores based on the results that they get back from the server. Some of the metaservers are shotgun on 3 , shotgun on 5, Pcons, Pmodeller etc Metaservers helped many groups to win in CASP5.

Homology modeling

Homology modeling is based on the fact that if two sequences have a high sequence similarity then they have similar 3D structure. But this is always not the case. Two sequences which don’t share much sequence similarity do have similar folds. This zone that defines low sequence similarity (15-30%) between target and template was termed by Doolittle (1986) as the *Twilight zone*. The steps in homology based predictions are

- *Database search*: Use the query sequence to search the database for known protein structures. This can be done by BLAST which does a pair-wise comparison .PSIBLAST and HMM which is profile based are better as they will be able to detect remote homologues as well
- *MSA*: Multiple sequence alignment of the query protein to the templates and identify structurally conserved region (SCR), active site residues, disulphide bridges, salt bridges. Most of the search methods are tuned to detect remote homologous and not for optimal sequence alignment. The alignment is a simple sequence-sequence alignment for sequence identity >40%, but for sequence identity <40% the alignment has gaps. In such cases, manual intervention through the knowledge of structural information can give better alignments. It should be seen that there are no gaps in the SCR or in secondary structural regions. The insertions should nor be buried in the secondary structure region but rather be at the ends or in loops. Within the homologous proteins secondary structures can move relative to each other or even disappear but neither the order not the orientation should differ- an alpha helix cannot be a beta sheet.
- *Main chain modeling*: This involves exchanging the residues in the template to that of the target in the SCR. When more than one template is used for modeling, the relative contribution or the weight of each structure is determined by its local degree of sequence identity with the target sequence.
- *Loop modeling*: If the template protein has a similar loop then it can be copied. The database approach of loop modeling involves searching the database for a segment of the main chain that fits the two stem region of a loop. The segment may be from homologous or non-homologues protein. These are sorted according to geometric criteria or sequence similarity and then superimposed and annealed on the stem region. Refinement is then carried out by energy optimization. The de-novo method is based on searching for a conformation in a given environment.
- *Side chain modeling:* Accurate prediction of the side chain conformation is an important step as the side chains mostly determine the interactions of the proteins with their ligands. One way of side chain modeling is to look for closely related sequences having similar conformation. If the side chains are very different, then we use the most common conformation found in that particular secondary structure and evaluate its energy. Although the conformation of one amino acid depends on the position of its neighbors, side chains in proteins tend to cluster independently into groups. SCWRL is an algorithm that solves the conformation of each cluster by braking up the cluster into groups connected by single amino acid. It is based on graph theory. SCWRL3.0 combines the dead–end algorithm and branch bound algorithm to give a powerful method for side chain prediction.
- *Energy refinement:* Energy minimization does not normally lead to big changes unless the structure was very bad to start with it. Energy minimization will bring the conformation to the nearest local minimum. But there are many local minima. To overcome this problem, molecular dynamics or Monte Carlo method is applied.
- *Model evaluation:* Usually if the sequence similarity between the target and the template is the model obtained will be good. There are two types of evaluation- the internal evaluation check whether or not a model satisfies the restraints used

to calculate it. It involves assessment of the bond angles, bond lengths, dihedral angles etc. Some of the programs that evaluates are PROCHECK, WHATCHECK etc. The external evaluation checks whether the template chosen is the best one, by comparing the Z scores. It is best to select a template that has a high sequence similarity, is in similar environment, and has a high X ray/NMR accuracy.

Some of the most commonly used servers for homology modeling are the Swiss-model, Modeler, 3D Jigsaw, 3DPSSM,SAMT-99,fugue-cam etc.

Fold Recognition or Threading

Threading or remote homologue design is a protein structure prediction technique carried when there is not enough sequence similarity between the target and template. The recognition of the template is a problem by itself and hence it is also called “fold recognition”. Threading involves steps similar to comparative modeling. It differs in the fold identification and fold fitting step. There are many approaches but the main theme is to try to find “folds’ from a library of folds of known protein structures. Fold recognition is carried based on Chothia’s 1000 fold hypothesis which states that there only a finite number of new folds. Instead of predicting how the sequence folds, it predicts how well a fold will fit the sequence and hence also called inverse folding.

The structural properties which are used to evaluate the fit include the local secondary structure, the environment and the pair-wise interaction of side chain of close amino acids. The 3D profile method involves the alignment of the sequence to a string of descriptors that describe the 3D environment of the target structure namely whether it is polar or nonpolar, whether it is in an alpha helix, beta sheet or a turn and whether it is fully buried, exposed or partially buried. Most of the threading programs use the contact potential method which models interaction in a protein structure as a sum over pair wise interactions. Here each known fold is represented as a 2D matrix of interresidue distance. An energy potential describes these distance dependent pair wise interactions of all combinations of amino acids. These are then compared with the 2D distance matrix of known protein structures and adjusted .The trained contact potentials are then used to align an unknown protein sequence against a group of folds. The top scoring alignments are considered as possible templates.

There is only 70% chance that the top 10 predictions will contain the correct fold. To increase the accuracy and eliminate the decoy folds we must consider more information structural or functional information, motifs, domain etc. Structures modeled thorough fold recognition has about 3-6A^{o} RMSD from the actual structure.

Some of programs that predict through fold recognition are Threader, 123D, 3DPSSM, PROSPECT etc. Threader, threads the protein through a library of folds derived from CATH. It aligns by optimizing the interaction partners fro each pair of residues in contact with the structure. It can use secondary structure predictions to constrain the threading. 123D uses contact potentials. 3DPSSM is uses position specific scoring matrix to align the target with the template.3DPSSm is a hybrid method that combines optimization of alignment to a family with optimization of threading energy.

AB-initio prediction

Ab-initio prediction is carried out when there is no suitable homologue found in the database. Prediction is done completely from the sequence It is based on Anfinsen's hypothesis that the native state of the protein represents the global free energy minimum. Ab-initio method tries to find these global minima of the protein. Finding the correct native like protein conformation requires

- An efficient search method for exploring the conformational space to find the energy minima.
- An accurate potential function that calculates the free energy of a given structure

To simply the computation, models are used to reduce the search space. There are 3 kinds of model.1) Lattice model: This represents peptide chain as lattices. But fails to represent subtle geometric consideration like strand twist and its backbone prediction is not all that accurate. 2)Discrete state off-lattice model: It improves upon the lattice model by applying restraints like allowing only certain side chain structure and limiting the peptide bond rotation.3)Using local structure prediction: In order to reduce the complexity, local structure biases are used. But the strength and multiplicity of the local structure prediction is highly sequence dependent. There are two type of scoring functions namely knowledge based scoring function and force field based function which are used. Currently there does not exist a reliable scoring function or search method.

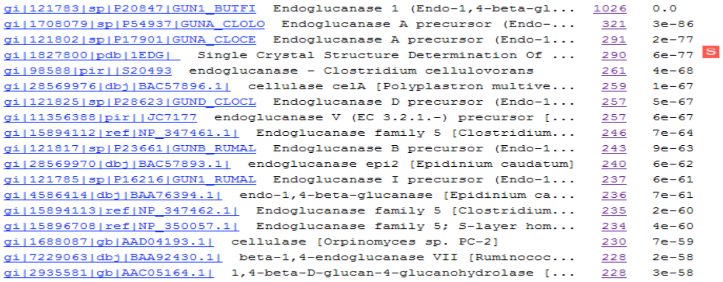
Some of the methods that did well in CASP4 and CASP5 were the segment insertion Monte-Carlo method in Rosetta, threading and Monte Carlo method by Friesner, the lattice Monte Carlo method by Jeff Skolnick and Andrew Kolinski where side chains were used for the lattice model etc.

There is a new fully automated ab-intio prediction method in which the Monte Carlo fragment insertion method (ROSETTA) of Baker and others has been merged with the I-site (library of sequences structure motifs) and the HMMSTR model for local structure in proteins. Here the input sequence after filtering out the low complexity region is submitted to Psi-Blast and then converted to sequence profiles. The sequence profile is compared in a sliding window with each of the I- site library scoring matrices. The highest confidence fragments retuned are the "I –site predictions". For the server1, I-sites fragment list is converted to Rosetta move set. The move set is fragment libraries of length 3 to 9 peptides which are used for Monte Carlo insertion. For server 2, the profile was submitted to each of the HMMSTR models namely HMMSTR-r for prediction of backbone angles, HMMSTR-d for the prediction of secondary structure and lastly HMMSTR-c for the prediction of super secondary structure. Rosetta searches the conformational protein space using fragment insertion moves and by applying the Monte-Carlo acceptance criteria. The point where the fragment is inserted is selected at random and then a fragment of length 3 or 9 residues is selected at random from the move set. The backbone angles are changed to that of the fragment and the co-ordiantes are then calculated. The move is accepted or rejected based on Monte Carlo criteria. The energy function is a structure based Bayesian conditional probability drawn form same PDB select database.

Case Study:

To illustrate the prediction of protein structure, I have chosen the entry P20847 from Swiss –prot and will try to predict its structure

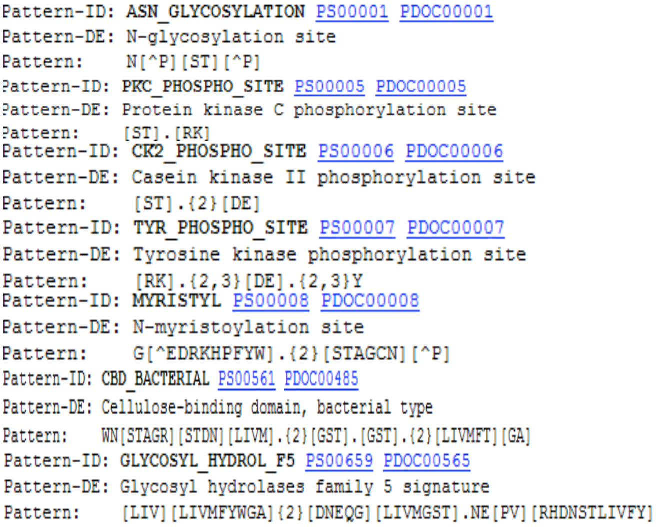
Step 1: I searched the database for possible homologues using Psi-Blast. The results of scores of the first 18 hits are given below,



```
gi|121783|sp|P20847|GUN1_BUTFI  Endoglucanase 1 (Endo-1,4-beta-gl...  1026   0.0
gi|1708079|sp|P54937|GUNA_CLOLO  Endoglucanase A precursor (Endo-...   321   3e-86
gi|121802|sp|P17901|GUNA_CLOCE  Endoglucanase A precursor (Endo-1...   291   2e-77
gi|1827800|pdb|1EDG|   Single Crystal Structure Determination Of ...   290   6e-77 S
gi|98588|pir||S20493  endoglucanase - Clostridium cellulovorans        261   4e-68
gi|28569976|dbj|BAC57896.1|  cellulase celA [Polyplastron multive...   259   1e-67
gi|121825|sp|P28623|GUND_CLOCL  Endoglucanase D precursor (Endo-1...   257   5e-67
gi|11356388|pir||JC7177  endoglucanase V (EC 3.2.1.-) precursor [...   257   6e-67
gi|15894112|ref|NP_347461.1|  Endoglucanase family 5 [Clostridium...   246   7e-64
gi|121817|sp|P23661|GUNB_RUMAL  Endoglucanase B precursor (Endo-1...   243   9e-63
gi|28569970|dbj|BAC57893.1|  endoglucanase epi2 [Epidinium caudatum]   240   6e-62
gi|121785|sp|P16216|GUN1_RUMAL  Endoglucanase I precursor (Endo-1...   237   6e-61
gi|4586414|dbj|BAA76394.1|  endo-1,4-beta-glucanase [Epidinium ca...   236   7e-61
gi|15894113|ref|NP_347462.1|  Endoglucanase family 5 [Clostridium...   235   2e-60
gi|15896708|ref|NP_350057.1|  Endoglucanase family 5; S-layer hom...   234   4e-60
gi|1688087|gb|AAD04193.1|  cellulase [Orpinomyces sp. PC-2]            230   7e-59
gi|7229063|dbj|BAA92430.1|  beta-1,4-endoglucanase VII [Ruminococ...   228   2e-58
gi|2935581|gb|AAC05164.1|  1,4-beta-D-glucan-4-glucanohydrolase [...   228   3e-58
```

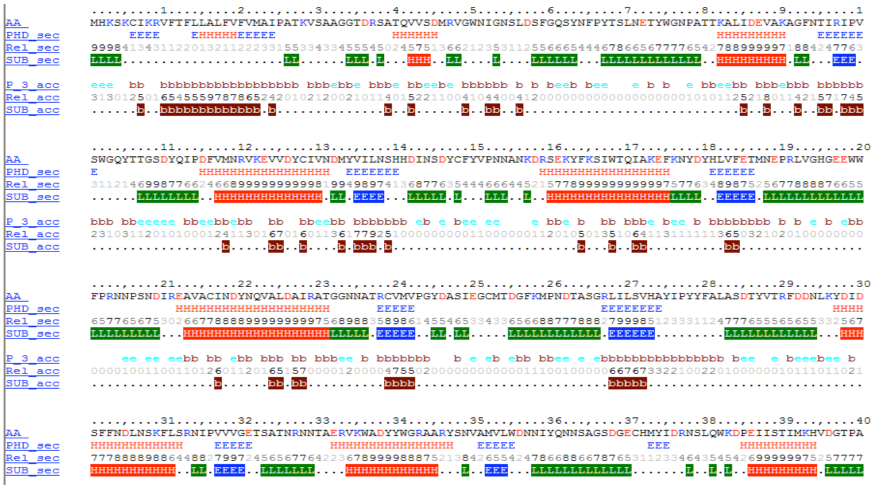
Step 2: To predict the secondary structure of the query protein, I submitted the protein to the PHD server. The results returned by the server is summarized as follows,

- ScanProsite has identified the following functional motifs that are annotated in the Prosite database

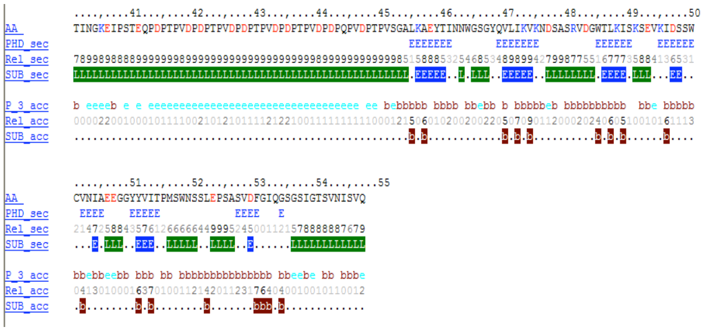


```
Pattern-ID: ASN_GLYCOSYLATION PS00001 PDOC00001
Pattern-DE: N-glycosylation site
Pattern:    N[^P][ST][^P]
Pattern-ID: PKC_PHOSPHO_SITE PS00005 PDOC00005
Pattern-DE: Protein kinase C phosphorylation site
Pattern:    [ST].[RK]
Pattern-ID: CK2_PHOSPHO_SITE PS00006 PDOC00006
Pattern-DE: Casein kinase II phosphorylation site
Pattern:    [ST].{2}[DE]
Pattern-ID: TYR_PHOSPHO_SITE PS00007 PDOC00007
Pattern-DE: Tyrosine kinase phosphorylation site
Pattern:    [RK].{2,3}[DE].{2,3}Y
Pattern-ID: MYRISTYL PS00008 PDOC00008
Pattern-DE: N-myristoylation site
Pattern:    G[^EDRKHPFYW].{2}[STAGCN][^P]
Pattern-ID: CBD_BACTERIAL PS00561 PDOC00485
Pattern-DE: Cellulose-binding domain, bacterial type
Pattern:    WN[STAGR][STDN][LIVM].{2}[GST].[GST].{2}[LIVMFT][GA]
Pattern-ID: GLYCOSYL_HYDROL_F5 PS00659 PDOC00565
Pattern-DE: Glycosyl hydrolases family 5 signature
Pattern:    [LIV][LIVMFYWGA]{2}[DNEQG][LIVMGST].NE[PV][RHDNSTLIVFY]
```

- SEG has identified the low complexity regions to be around residues 10-21 and from residue 413 to 451 (which has a PDPVPDT repeat)
- Multiple sequence alignment was done by MaxHom
- CYSPRED has not predicted the protein to be having disulphide bridges
- GLOBE has predicted the protein to appear as compact as a globular protein
- Ambivalent sequence predictor has identified the location of conformational switches at between residues 9-19, 23-26, 30-32, 49-57,101-103,276-280.This does not predict whether the sequence has any switch or not.
- PHDsec has predicted the protein to be composed of 21.21 % alpha helix, 17.55% beta sheets and the rest 61.24 to be loops. PHDacc has predicted the 51.37 % to be buried and 48.63% to be exposed with more than 16% of their surface.
- Secondary structure is predicted by PHDsec by a system of neural networks. From the data given below, it is seen that the protein has both alpha helices and beta sheets.

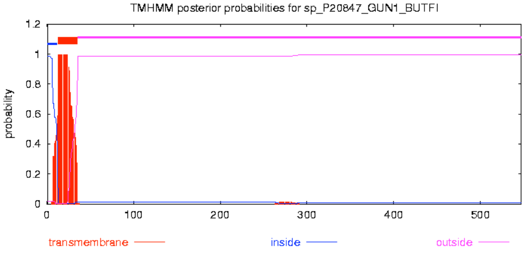


```
          ....,....1....,....2....,....3....,....4....,....5....,....6....,....7....,....8....,....9....,....1
AA        MHKSKCIKRVFTFLLALFVFVMAIPATKVSAAGGTDRSATQVVSDMRVGWNIGNSLDSFGQSYNFPYTSLNETYWGNPATTKALIDEVAKAGFNTIRIPV
PHD_sec        EEEE    EHHHHHEEEEE               HHHHHH                                    HHHHHHHHH    EEEEEE
Rel_sec   9998413431122013211222331553343345554502457513662123531125566654446786656777765427889999971884247763
SUB_sec   LLLL.....................LL......LLL.L...HHH..LL....L....LLLLLL...LLLLLLLLLLLLL..HHHHHHHHH.LL...EEE.

P_3_acc   eee  bb  bbbbbbbbbbbbbbbbbb bbbebbe bbbe bbeebe bbbbbb b b beeb bee   e b b  e bbeebb bbbebbb bbbbbb
Rel_acc   3130125016545559787865242010212002101140152211004104400412000000000000000000010101125218011421571745
SUB_acc   ......b..bbbbbbbbbbbbb.b..............b..b......b..bb..b............................b..b...b..bb.bbb

          ....,....11...,....12...,....13...,....14...,....15...,....16...,....17...,....18...,....19...,....20
AA        SWGQYTTGSDYQIPDFVMNRVKEVVDYCIVNDMYVILNSHHDINSDYCFYVPNNANKDRSEKYFKSIWTQIAKEFKNYDYHLVFETMNEPRLVGHGEEWW
PHD_sec   E             HHHHHHHHHHHHHHHHH  EEEEEEE                  HHHHHHHHHHHHHHHHH     EEEEEE
Rel_sec   3112146998776624668999999999981994989741368776354446664452157789999999999975776348987525677888876655
SUB_sec   ......LLLLLLLL..HHHHHHHHHHHHHH.LL.EEEE...LLLLL.L...LLL..L..HHHHHHHHHHHHHHHHLLLL..EEEEE.LLLLLLLLLLLLL

P_3_acc   bbb bbeeeee bbeebbebb  bb  bbeebb bbbbbbb eb e bee ee    e bbe b  bb bbbe bee b bbbbbbbb b b e b ebb
Rel_acc   2310311201010001241130167016011361779251000000000110000001120105013510641131111113650321020100000000
SUB_acc   .................b.....bb..b....b.bbb.b........................b...b..bb..........bb................

          ....,....21...,....22...,....23...,....24...,....25...,....26...,....27...,....28...,....29...,....30
AA        FPRNNPSNDIREAVACINDYNQVALDAIRATGGNNATRCVMVPGYDASIEGCMTDGFKMPNDTASGRLILSVHAYIPYYFALASDTYVTRFDDNLKYDID
PHD_sec              HHHHHHHHHHHHHHHHHHHH      EEEEE                        EEEEEEEE                      HHHH
Rel_sec   6577656753026677888899999999975689883589861455465334336566887778882799985123331124777655565655332567
SUB_sec   LLLLLLLLL...HHHHHHHHHHHHHHHHHHHLLLLL.EEEEE..LL.LL.....LLLLLLLLLLLL.EEEEEE.........LLLLLLLLLLLL...HHH

P_3_acc       ee ee eebb bb ebb bbb bb bbbee b bbbbbbb   b e eb ebb bbee e ebbbbbbbbbbbbbbbb bee  e beeebee b
Rel_acc   0000100110011012601120165157000012000047550200000000000011100100000667673322100220100000010111011021
SUB_acc   ................b......bb.bb..........bbbb.........................bbbbb............................

          ....,....31...,....32...,....33...,....34...,....35...,....36...,....37...,....38...,....39...,....40
AA        SFFNDLNSKFLSRNIPVVVGETSATNRNNTAERVKWADYYWGRAARYSNVAMVLWDNNIYQNNSAGSDGECHMYIDRNSLQWKDPEIISTIMKHVDGTPA
PHD_sec   HHHHHHHHHHHH    EEEEE          HHHHHHHHHHHHHHHH   EEEEE                 EEE         HHHHHHHHHH
Rel_sec   7778888988644882799724565677642236789999888752138426554247866886678765311233464354542699999975257777
SUB_sec   HHHHHHHHHHH..LL.EEEE..LLLLLLL....HHHHHHHHHHHH...L..EEE...LLLLLLLLLLLLL.......L..L.L..HHHHHHHHH.LLLLL
```



```
          ....,....41...,....42...,....43...,....44...,....45...,....46...,....47...,....48...,....49...,....50
AA        TINGKEIPSTEQPDPTPVDPDPTPVDPDPTPVDPDPTPVDPDPQPVDPTPVSGALKAEYTINNWGSGYQVLIKVKNDSASRVDGWTLKISKSEVKIDSSW
PHD_sec                                                         EEEEEEE       EEEEEEE         EEEEEE    EEEEEE
Rel_sec   7899898888999999998999999999999999999999998999999999851588853254685348989942799877551677735884136531
SUB_sec   LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.EEEEE..L.LLL..EEEEE..LLLLLLLL.EEEE.LLL...EE..

P_3_acc   b eeeeb e e eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee ee bebbbbb bbbb bbebb b bbbbbeb bbbbbbbbbb  bbe bbbbb
Rel_acc   0000220010001011110021012101111212210011111111110001215060102002002205070901120002024060510010161113
SUB_acc   ......................................................b.b............b.b.b..........b.b.b......b....

          ....,....51...,....52...,....53...,....54...,....55
AA        CVNIAEEGGYYVITPMSWNSSLEPSASVDFGIQGSGSIGTSVNISVQ
PHD_sec    EEEE    EEEEE            EEEE   E
Rel_sec   21472588435761266666449995245001121578888887679
SUB_sec   ...E.LLL..EEE..LLLLL..LLLL..E......LLLLLLLLLLLL

P_3_acc   bbebbeebb bbb bb bbbbbbbbbbbbbbb bbeebe bb bbbe
Rel_acc   04130100016370100112142011231764040010010110012
SUB_acc   .b........b.b........b.......bbb.b.............
```

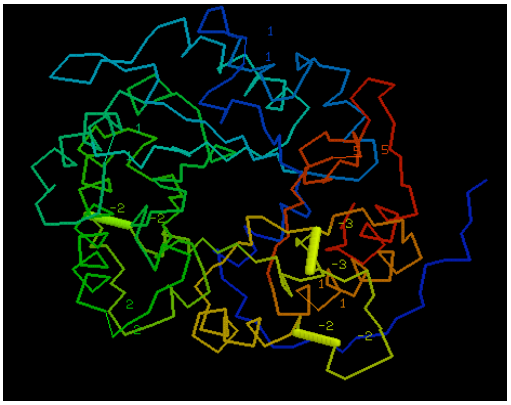
- TMHMM: prediction of the trans-membrane segments was done by submitting the protein to the TMHMM server. From the results obtained, it is seen that there is one TMhelix between residues 13 and 35. Since the expected number of amino acids in trans-membrane helices in the first 60 amino acids of the protein is around 21.34 the predicted trans-membrane helix in the N-term could be a signal peptide



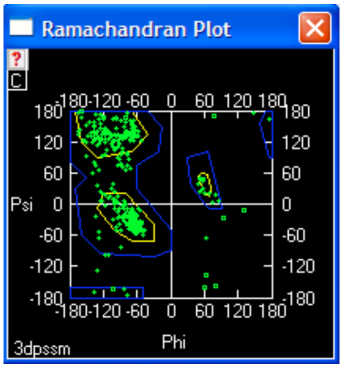
Step 3: Model Building: The sequence identity for the first few hits obtained form Psi-Blast was an average value below 40%. Hence I submitted the sequence to Swiss-model for comparative modeling. Although it found 1edg.pdb, 1exg.pdb and 1exh.pdb as templates, ProMoII failed to build a suitable model. Hence I submitted the protein to the fold recognition server 3DPSSM.

A Pseudo multiple sequence alignment is generated through Psi-Blast. Psipred is used to predict the secondary structural elements. It is seen from the pattern of distribution of the alpha helices and beta sheets that it must be an alpha and beta class protein. The secondary structure predicted by Psi-pred and PHD are similar in the alpha helical regions between 39-45,82-90,115-130,160-175,210-231,295-312,329-347,385-395. The beta regions that are similar in both the predictions are between 95-100,134-140,180-185,238-242,267-272,316-322,352-355,453-461,467-476,485-490,495-501,510-515,527-534. Both have shown the query protein to be composed of many loop regions. The query sequence, its profile and the secondary structure predictions are scanned against a fold library and the top 20 hits are retrieved. From the hits obtained, it was seen that the template 1edg had a higher sequence identity of 42% over a length of 380 compared to the other templates. Hence this was chosen for further model building. The model generated by 3DPSSM is simple mappings from the co-ordinates of the template structure and the query residues aligned to them. Side chains are modeled using SCWRL.

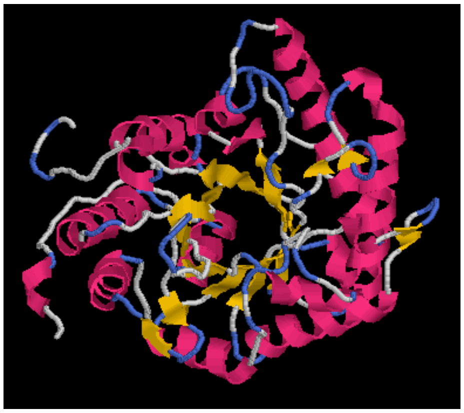
3DPSSM has identified the query protein to be belonging to the class of alpha and beta proteins having a TIM beta/alpha barrel fold and in the super family of glycosy ltransferases and family of beta glycanases. It has identified the protein to be an endo-glucanase celA protein. The 3D model obtained using 1edg as the template is shown below. It is seen that the model does have many insertions and deletions. But most of these are in the coil regions except for a single residue insertion in the helical region and a 2 residue deletion in the helical region.

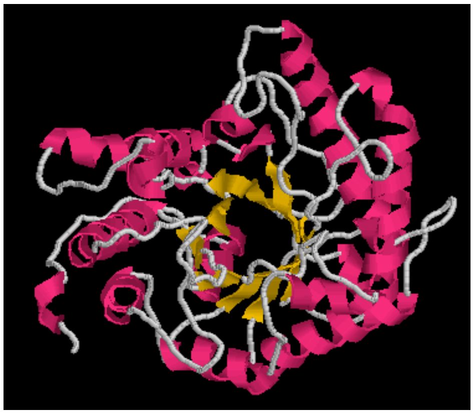


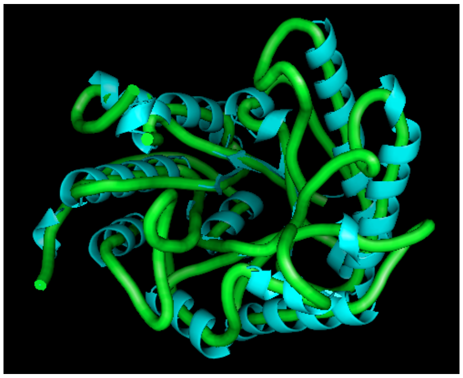
Step 4: Model evaluation and comparison: The model obtained was evaluated using PROVE, PROCHECK and whatif checks. There are a number of unusual bond angle and bond length reported although the there is a normal bond angle and bond length variability. There are number of buried unsatisfied hydrogen bond donors and acceptors as checked by bpocheck. Gly 17 has an unusual back bone oxygen position and prolines at 48, 60 and 372 have improper dihedral angles. Proline at 171 has an unusual puckering. A stretch of three residues from Tyr 360 to asn 362 have an abnormal packing environment. This may be because these residues are far apart in the loop or might be an indication of misthreading. As seen from the Ramachandran plot there are residues in the forbidden regions. The Ramchandran Z-score expressing how well the backbone conformations of all residues are corresponding to the known allowed areas in the Ramachandran plot is within expected ranges. Omega angles are too tightly restrained.



The structure of the query P20847 is shown in left and that of the template 1edg is shown in right and the superposition of the query (in green) on the template (cyan) is shown below them.







The query having 547 is an alpha and beta class endoglucanase cel A protein with a TIM beta/alpha barrel fold. It belongs to the family of glycosyl transferase. Glycosyl hydrolases (EC 3.2.1-) are a widespread group of enzymes that hydrolyse the glycosidic bond between two or more carbohydrates or between a carbohydrate and a non carbohydrate moiety. Glycoside hydrolase family 5 contains enzymes with several known activities- endoglucanase ,beta mannose, exo 1,3 glucanase, endo 1,6 glucanase, Xylanase and endoglycoceramidase. The template protein is an endoglucanase. By similarity with the template the catalytic site must be at the C terminal ends of the beta strands.

Conclusion

Predicting the structure of a protein is a difficult task. Different approaches to predict the structure take into account different chemical and physical properties. This has given rise to a number of tools and techniques, some of which being specialized to work on either some aspects of predictions or some categories of proteins. Nevertheless these are not significantly accurate or reliable enough to predict all kinds of proteins. One way that the quality of the model can be improved is by jointly applying different prediction technique and combines their results. But this has to be done delicately as different techniques have different ways or formats of representing their input and output. As each one uses a different approach, comparison and interpretation of the result can be hard. Sometimes integration may not always give the best results. But generally, good detection of homologues sequence and a good multiple sequence alignment can improve the results. The high quality crystal structure on which the model is built also plays a part in giving good models. Metaservers when used can also give good models.

Method	Sequence identity	Template coverage	Accuracy	Difficulty
Homology	>30%	>90%	1-3A^{o}	Trivial
Fold recognition/homology	20-30%	>75%	2-5A^{o}	Easy
Fold recognition	<20%	>50%	3-10A^{o}	Moderate
Ab-initio	<5	0	5-20A^{o}	Hard

High and medium accuracy model is helpful in refining of functional predictions that have been based on a sequence match alone. This is because the binding of a ligand to a substrate is determined by the structure of the binding site. Ab-initio prediction methods even with their very low accuracy can give a reliable functional annotation.

References

Baker D., Sali A. (2001). Protein structure prediction and structural genomics. *Science.* 294, 93.

Blundell, T. L.; Sibanda, B. L.; Sternberg, M. J.; Thornton, J. M. Knowledge-Based Prediction of Protein Structures and the Design of Novel Molecules. Nature 1987, 326, 347-352.

Bonneau R. Baker, D. (2001) Ab Initio protein structure prediction: progress and prospects. *Annul. Rev. Biophys. Biomol. Struct*. 30**,** 173.

Bonneau R., Tsai Jerry, Ruczinski I., Chivian D., Rohl C., Strauss C. E. M. and Baker D. (2001) ROSETTA in CASP4: Progress in Ab Initio protein structure prediction. *Proteins: Structure, Function, and Genetics* Suppl 5, 119

Bonneau R., Ruczinski I., Tsai J., and Baker D. (2002). Contact order and Ab Initio protein structure prediction. *Protein Science*. 11, 1937.

Bowie J. U., Luthy R. & Eisenberg, D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**, 164

Brenner S. E., Chothia C. & Hubbard, T. J. P. (1998). Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci*. USA, 95, 6073.

Chothia, C. (1992). Proteins. One thousand families for the molecular biologist. *Nature*, 357, 543.

Chung, S., Subbiah, S. (1996). A structural explanation for the twilight zone of protein sequence homology. *Structure*, 4, 1123

Dunbrack RL Jr, Dunker K, Godzik A. Protein structure prediction in biology and medicine Pac Symp Biocomput. 2000;(12):93-4.

Huang ES, Samudrala R, Ponder JW. Ab initio fold prediction of small helical proteins using distance geometry and knowledge-based scoring functions. *JMB* 290:267-281, 1999.

Jeffrey Skolnick, Andrzej Kolinski, Daisuke Kihara, Marcos Betancourt, Piotr Rotkiewicz,2 and Michal Boniecki2 :Ab Initio Protein Structure Prediction via a Combination of Threading, Lattice Folding, Clustering, and Structure Refinement Proteins: Structure, Function, and Genetics Suppl 5:149–156 (2001)

Jones D. T., Hadley C. (2001) Threading methods for protein structure prediction. In *Bioinformatics: Sequence, Structure and Databanks: A Practical Approach* (ed. Higgins D., and Taylor W.). Oxford University Press, Oxford.

Karplus K., Barrett C., Hughey R. (1998). Hidden Markov models for detecting remote protein homologies. *Bioinformatics*. **14**, 846.

Lawrence A. Kelley, Robert M. MacCallum and Michael J. E. Sternberg Enhanced Genome Annotation Using Structural Profiles in the Program 3D-PSSM J. Mol. Biol. 2000, 299, 499-520.

Leonid A. Mirny and Eugene I. Shakhnovich Protein Structure Prediction by Threading. Why it Works and Why it Does Not J. Mol. Biol. (1998) 283, 507±526

Rost B,Protein secondary structure prediction continues to rise Journal of Structural Biology, 134, 204-218, 2001

Rost B. (1999) Twilight zone of protein sequence alignment. *Protein Eng.* **12**, 85

Samudrala R, Xia Y, Huang E, Levitt M
Ab initio protein structure prediction using a combined hierarchical approach. Proteins. 1999;Suppl 3:194-8

Samudrala R, Xia Y, Levitt M, Huang ES. A combined approach for ab initio construction of low resolution protein tertiary structures from sequence. In Altman R Dunker K, Hunter L, Klein T, Lauderdale K, eds. *Proceedings of the Pacific Symposium*

Sathya nanda,Saraai Deris , Rosli Md Illias Prediction of protein secondary structure Jurnal Teknologi 35(C), Pages: 81-90, December 2001

Schmidler SC, Liu JS, Brutlag DL.Bayesian segmentation of protein secondary structure. J Comput Biol. 2000 Feb-Apr;7(1-2):233-48

Shao Y, Bystroff C. Predicting interresidue contacts using templates and pathways. Proteins. 2003;53 Suppl 6:497-502.1)

Shortle D. Protein fold recognition. *Nature Struct Biol* 1995; 2: 91–93.