Research Inventy: International Journal Of Engineering And Science
Issn: 2278-4721, Vol.2, Issue 2 (January 2013), Pp 05-12
Www.Researchinventy.Com

In Silico Attributes Of Cold Resistance Protein 1 (Crp1) From Brassica Oleraceae

[1]nishawar Jan, [2]meenu A. Qureshi, [3]sheikh T. Majeed, [4]khurshid I. Andrabi
[1,2,3,4(]*Department of Bioinformatics and Biotechnology, University of Kashmir, J&K, India,)*

Abstract- *Cold Resistance Protein 1 (CRP1) gene arabidopsis kin family homolog was previously isolated by us from cold tolerant varieties of Brassica oleraceae (AC# GQ461797-800). CRP1 expression was found to relate with cold acclimation and resistance. The mechanistic understanding of potential cold tolerance induced by the protein however remains obscure due to lack of knowledge about its structure and precise biological function. Therefore, bioinformatics analysis and abinitio 3-D modelling of the protein sequence was performed using various computational analysis tools that suggest a helical organization for the protein with four transmembrane α-helices giving rise to a unique 3-D structure. Consensus sequence motifs for myristoylation further indicate that its localization may be membrane oriented.*

Keywords- *Cold resistance protein, in silico, modeling, abinitio*

I. Introduction

Structural analysis is often crucial in elucidating the function of a protein and understanding its biological role. we have previously isolated cDNA and genomic DNA sequences for a low molecular weight cold resistance protein 1 (CRP1) from several varieties of Brassica oleraceae (AC# GQ461797-800), with established potential to induce cold acclimation and tolerance [1]. Accordingly CRP1 has an appeal to be genetic tool for manipulation of cold tolerance in plants. Based on similarity in the amino acid sequence and multiple sequence alignments, a domain, Kin Homology Domain (KHD) has been recognized as common to many cold induced proteins and is conserved right from *Arabidopsis*. This is suggestive of a critical role associated with the CRP1 in general and the conserved domain in particular in the context of cold tolerance.Protein three-dimensional (3D) structure (i.e., the coordinates of all atoms) determines protein function. The hypothesis that structure (also referred to as 'the fold') is uniquely determined by the specificity of the sequence, has been verified for many proteins [2]. While it is now known that particular proteins (chaperones) often play a role in the folding pathway, and in correcting misfolds [3], it is still generally assumed that the final structure is at the free-energy minimum. Thus, all information about the native structure of a protein is coded in the amino acid sequence, plus its native solution environment. In principle, the code could be deciphered from physicochemical principles using, for example, molecular dynamics methods [4]. In practice, however, such approaches are frustrated by two principal obstacles. First, energy differences between native and unfolded proteins are extremely small (order of 1 kcal mol^{-1}). Second, the high complexity (i.e., co-operativity) of protein folding requires several orders of magnitudes more computing time than we anticipate having over the next decades. Thus, the inaccuracy in experimentally determining the basic parameters, and the limited computing resources become fatal for predicting protein structure from first principles [5]. The only successful structure prediction tools are knowledge-based, using a combination of statistical theory and empirical rules.

Currently, databases for protein sequences (e.g., SWISSPROT7) are expanding rapidly, largely because of large-scale genome sequencing projects. The classical methods for structure analysis of proteins are X-ray crystallography and nuclear magnetic resonance (NMR). Unfortunately, these techniques are expensive and can take a long time (sometimes more than a year). On the other hand, the sequencing of proteins is relatively fast, simple, and inexpensive. As a result, there is a large gap between the number of known protein sequences and the number of known three-dimensional protein structures. This gap has grown over the past decade (and is expected to keep growing) as a result of the various genome projects worldwide. Thus, computational methods which may give some indication of structure and/or function of proteins are becoming increasingly important. Unfortunately, since it was discovered that proteins are capable of folding into their unique native state without any additional genetic mechanisms, over 25 years of effort has been expended on the determination of the three-dimensional structure from the sequence alone, without further experimental data. Despite the amount of effort, the protein folding problem remains largely unsolved and is therefore one of the most fundamental unsolved problems in computational molecular biology today.

How can the native state of a protein be predicted? There are three major approaches to this problem: 'comparative modelling', 'threading', and 'ab initio prediction'. Comparative modelling exploits the fact that evolutionarily related proteins with similar sequences, as measured by the percentage of identical residues at each position based on an optimal structural superposition, often have similar structures. For example, two sequences that have just 25% sequence identity usually have the same overall fold. Threading methods compare a target sequence against a library of structural templates, producing a list of scores. The scores are then ranked and the fold with the best score is assumed to be the one adopted by the sequence. Finally, the ab initio prediction methods consist in modelling all the energetics involved in the process of folding, and then in finding the structure with lowest free energy. This approach is based on the 'thermodynamic hypothesis', which states that the native structure of a protein is the one for which the free energy achieves the global minimum. While ab initio prediction is clearly the most difficult, it is arguably the most useful approach and hence this attempt.

II. Methods

2.1 Protein Primary Structure Studies:

Most of the protein primary structure studies including the Amino acid composition, Protein statistics, Hydrophobicity analysis, Protein charge studies were done using ExPASy Proteomics server [http://www.expasy.ch/]. The hydrohobicities of the individual Amino acids were calculated using a software, CLC Main Workbench. This software uses the algorithm of Kyte and Doolittle [6] for the Hydrophobicity predictions. The Kyte-Doolittle scale is widely used for detecting hydrophobic regions in proteins. Regions with a positive value are hydrophobic. This scale can be used for identifying both surface-exposed regions as well as transmembrane regions.

2.2 Secondary Structure Prediction:

Secondary structure predictions were done by PSIPRED [http://www.psipred.net] [7] JNet [http://www.jalview.org/ help/html/webServices/jnet.html] [8,9], sspro [http://download.igb.uci.edu/ sspro4.html] [10] and the consensus sequence was presumed as the final secondary sequence.

2.3 Template Searching and 3-D Modelling:

An attempt was made to find a suitable template protein for the modeling of the target protein. The template protein was searched through mGenTHREADER [http://www.psipred.net] [7,11] which is an online tool for searching similar sequences, based on sequence and structure-wise similarity. No templates structures were found that significantly matched it. Therefore, an automated server, I-TASSER, [zhang.bioinformatics.ku.edu/I-TASSER/; [12-14] was used for predicting the 3-D structure of the protein which furnishes four PDB files representing the probable 3-D structure. I-TASSER is a hierarchical protein structure modeling approach based on the secondary-structure enhanced Profile- Profile threading Alignment (PPA) [13] and the iterative implementation of the Threading ASSEmbly Refinement (TASSER) program [15]. The target sequences are first threaded through a representative PDB structure library (with a pair-wise sequence identity cut-off of 70%) to search for the possible folds by four simple variants of PPA methods, with different combinations of the hidden Markov model [16] and PSIBLAST [17] profiles and the Needleman-Wunsch [18] and Smith-Waterman [19] alignment algorithms. The continuous fragments are then excised from the threading aligned regions which are used to reassemble full-length models while the threading unaligned regions (mainly loops) are built by ab initio modeling [20]. The conformational space is searched by replica-exchange Monte Carlo simulations [21]. The structure trajectories are clustered by SPICKER [22]; SPICKER package) [http://zhang.bioinformatics.ku.edu/ SPICKER] and the cluster centroids are obtained by the averaging the coordinates of all clustered structures.

To rule out the steric clashes on the centroid structures and to refine the models further, the fragment assembly simulation is implemented again, which starts from the cluster centroid of the first round simulation. Spatial restraints are extracted from the centroids and the PDB structures searched by the structure alignment program TM-align [23], which are used to guide the second round simulation. Finally, the structure decoys are clustered and the lowest energy structure in each cluster is selected, which has the Cα atoms and the side chain centers of mass specified.

I-TASSER simulations are run for the full chain as well as the separate domains. The final full-length models are generated by docking the model of domains together. The domain docking is performed by a quick Metropolis Monte Carlo simulation where the energy is defined as the RMSD of domain models to the full-chain model plus the reciprocal of the number of steric clashes between domains. The goal of the docking is to find the domain orientation that is closest to the I-TASSER full-chain model but has the minimum steric clashes. This procedure does not influence the multiple domain proteins which have all domains completely aligned by the PPAs.

2.4 PDB File Visualisation and Free Energy Calculations:

SWISS PDB viewer [24] was used for visualizations of the PDB files and computing the free energy of the predicted 3-D structures using the "COMPUTE ENERGY" tool of the viewer.

2.5 Quality Assessment of the Predicted 3-D Structure:

The quality of the structures modeled, was assessed using the program PROCHECK [25]. It is a suite of programs to check the stereochemical quality of protein structures producing a number of PostScript plots analyzing its overall and residue-by-residue geometry.

2.6 Functional Prediction:

One of the tools of ExPASy called PROSITE [http://www.expasy.ch/cgi-bin/prosite] was used for search for any post-translational modification consensus sites present the protein sequence.

III. Results and Discussion

3.1 Physico-Chemical Parameters of the Protein:

Every protein holds specific and individual features which are unique to that particular protein. Features such as isoelectric point or amino acid composition can reveal important information about a novel protein. Amino acid composition is generally conserved through family-classes in different organisms which can be useful when studying one particular protein or enzymes across species borders. Another interesting observation is that amino acid composition deviates slightly between proteins from different subcellular localizations. This fact has been used in several computational methods, used for prediction of subcellular localization. All 20 amino acids consist of combinations of only five different atoms. The atoms which can be found in these simple structures are; Carbon, Nitrogen, Hydrogen, Sulfur, Oxygen. The atomic composition of a protein can, for example, be used to calculate the precise molecular weight of the entire protein. The count of charged residues can also give a feel about the location of the protein. At neutral pH, the fraction of negatively charged residues implies information about the location of the protein. Intracellular proteins tend to have a higher fraction of negatively charged residues than extracellular proteins. At neutral pH, nuclear proteins have high relative percentage of positively charged amino acids. Nuclear proteins often bind to the negatively charged DNA, which may regulate gene expression or help to fold the DNA. Nuclear proteins often have a low percentage of aromatic residues [26]. The Isoelectric point (pI) of a protein is the pH where the proteins have no net charge. The pI is calculated from the pKa values for 20 different amino acids. At a pH below the pI, the protein carries a positive charge, whereas if the pH is above pI the proteins carry a negative charge. In other words, pI is high for basic proteins and low for acidic proteins. This information can be used in the laboratory when running electrophoretic gels. Here the proteins can be separated, based on their isoelectric point. The aliphatic index of a protein is a measure of the relative volume occupied by aliphatic side chain of the following amino acids; alanine, valine, leucine and isoleucine. An increase in the aliphatic index increases the thermostability of globular proteins.

Amino acid statistics of the protein under study (CRP1) using EXPASY proteomics server (Methods Sec. 2.1) reveals that the protein of total 65 Amino acids "Table 1" has an average Molecular weight of 6.6 kDa and a theoretical PI of 9.16 "Fig.1". The atomic composition of the protein under study is $C_{271}H_{457}N_{85}O_{95}S_3$. The protein is composed of a total of 7 negatively charged residues and 9 positively charged residues "Table 2". The maximum percentage of the protein is neutral amino acids "Fig.2". This cancels out the probability of the protein being a nuclear or a cytoplasmic protein. As there are no Trp, Tyr or Cys in the protein, the protein should not be visible by UV spectrophotometry "Table.3". The instability index is computed to be 25.36. This classifies the protein as reasonably stable. Aliphatic index is calculated as 57.38 that explains the thermostability of the protein. Grand average of hydropathicity (GRAVY) of -0.571 tells that the protein is reasonably hydrophilic "Table 4" and "Fig.3"

.

3.2 Protein Secondary Structure:

The consensus of PSIPRED, sspro and JNet secondary structure prediction servers (Methods Sec. 2.2) discloses that the protein is composed of four helices separated by coils "Fig.4". PSIPRED calculates the secondary structure propensities of the individual amino acids using a simple and accurate secondary structure prediction method, incorporating two feed-forward neural networks which perform an analysis on output obtained from PSI-BLAST (Position Specific Iterated - BLAST) (www.nbi.nlm.nih. gov/blast). JNet secondary structure prediction methods attempts to infer the likely secondary structure for a protein based on its amino acid composition and similarity to sequences with known secondary structure. The JNet method uses several different neural networks and decides on the most likely prediction via a jury network.

3.3 Protein Tertiary Structure:

I-TASSER server furnished four PDB files (See supplementary Data) representing the probable tertiary structures of the protein under study with the C-Scores as -2.36 "Fig.5", -4.40, -2.63 (Methods Sec. 2.3). C-score is a confidence score for estimating the quality of predicted models by I-TASSER. It is calculated based on the significance of threading template alignments and the convergence parameters of the structure assembly simulations. C-score is typically in the range of [-5, 2], where a C-score of higher value signifies a model with a high confidence and vice-versa. Analysis of the four PDB files predicted by I-TASSER using PDB Viewer (Methods Sec. 2.3) confirms the first structure with C- score -2.36 as the most energetically favourable one. The total energy of the respective structures calculated by the PDB Viewer (Methods Sec. 2.4) comes out to be +3833KJ/mol, +78921304 KJ/mol, +3885KJ/mol. Analysis of the structures through PROCHECK (Methods Sec. 2.4) also confirms the first structure with C- score -2.36 as the most stable structure among the four "Fig 6".

3.4 Protein Localization:

An appreciable number of eukaryotic proteins are acylated by the covalent addition of myristate (a C14-saturated fatty acid) to their N-terminal residue via an amide linkage [27-28]. The sequence specificity of the enzyme responsible for this modification, myristoyl CoA: protein N-myristoyl transferase (NMT) has been derived from the sequence of known N-myristoylated proteins and from studies using synthetic peptides. It seems to be the following:

- The N-terminal residue must be glycine.
- In position 2, uncharged residues are allowed. Charged residues, Proline and large hydrophobic residues are not allowed.
- In positions 3 and 4, most, if not all, residues are allowed.
- In position 5, small uncharged residues are allowed (Ala, Ser, Thr, Cys, Asn and Gly). Serine is favored.
- In position 6, Proline is not allowed.

PROSITE (Methods Sec. 2.6) finds four probable N-myristoylation sites in CRP1 sequence shown in Table 5. The presence of the myristoylation sites in the protein suggests that the protein perhaps, might be attached to some sort of membrane.

1. Figures And Tables:

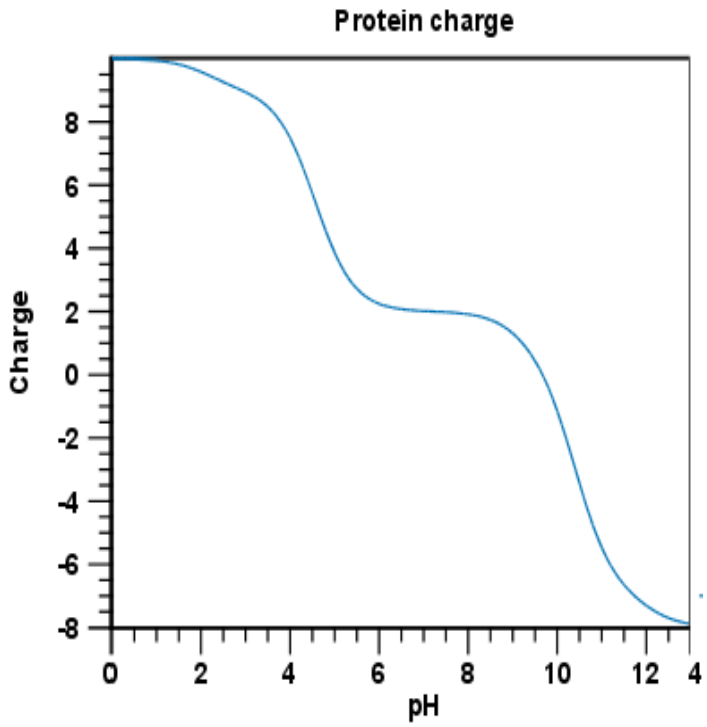

Fig 1: Graph showing charge carried by CRP1 at different pH values. The graph shows that the charge on the protein is neutralized between pH 6-8.

Table 1: Distribution of Amino acids in CRP1

Amino acid	Count	Frequency
Alanine (A)	14	0.215
Cysteine (C)	0	0.000
Aspartic Acid (D)	3	0.046
Glutamic Acid (E)	4	0.062
Phenylalanine (F)	1	0.015
Glycine (G)	9	0.138
Histidine (H)	0	0.000
Isoleucine (I)	1	0.015
Lysine (K)	8	0.123
Leucine (L)	2	0.031
Methionine (M)	3	0.046
Asparagine (N)	4	0.062
Proline (P)	0	0.000
Glutamine (Q)	5	0.077
Arginine (R)	1	0.015
Serine (S)	2	0.031
Threonine (T)	4	0.062
Valine (V)	4	0.062
Tryptophan (W)	0	0.000
Tyrosine (Y)	0	0.000

Table 2: Count of Charged Residues in CRP1

Charge Type	Count	Frequency
Negatively charged(D & E)	7	0.108
Positively Charged	9	0.138
Other	49	0.754

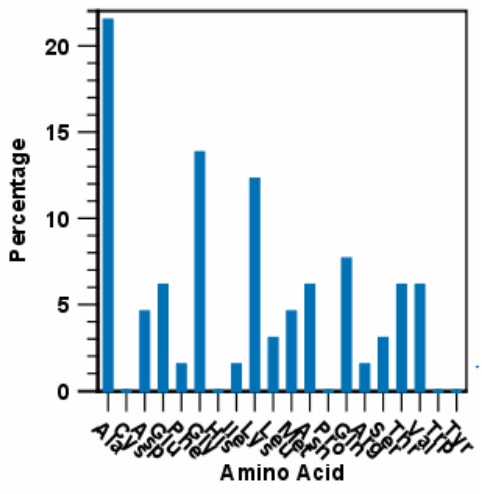

Fig 2: Histogram representation of the percentage of various Amino acids in CRP1.

Conditions	Extinction coefficient at 280nm	Absorption at 280nm 0.1(=1g/l)
Non reduced Cystines	No Trp, Tyr or Cys in protein	Not visible by UV spectrophotometry
Reduced Cysteines	No Trp, Tyr or Cys in protein	Not visible by UV spectrophotometry

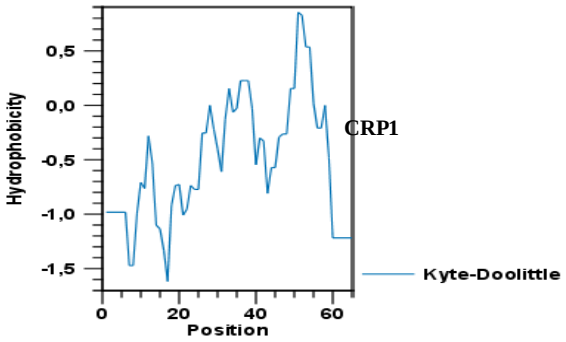

Fig .3: Hydrophobicity plot of CRP1 using Kyte-Doolittle algorithm
Hydrophobicity plot CRP1

Table 4: Count of Hydrophobic and Hydrophillic Residues in CRP1

Hydrophobicity	Count	Frequency
Hydrophobic (A, F, G, I, L, M, P, V, W)	34	0.523
Hydrophillic (C, N, Q, S, T, Y)	15	0.231
Other	16	0.248

Fig 4: Secondary structure prediction of CRP1.Amino acid sequence of CRP1 is included as Query sequence. The secondary predictions made by various servers including PSIPRED, sspro and JNet are shown against the respective sever name. The final secondary structure is considered as the consensus of the three.

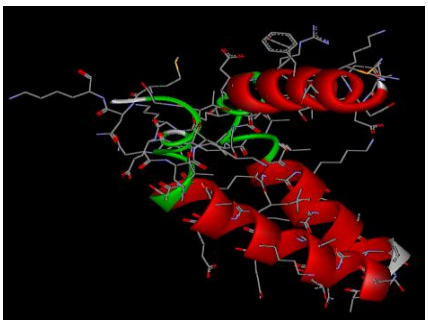

Fig 5: Diagrammatic Representation of most stable Predicted 3-D conformation of CRP1.

Analysis of the four PDB files predicted by I-TASSER using PDB Viewer (Methods Sec. 2.3) confirms the first structure with C- score -2.36 as the most energetically favourable one. The total energy of the respective structures calculated by the PDB Viewer (Methods Sec. 2.4) comes out to be +3833KJ/mol, +78921304 KJ/mol, +3885KJ/mol. Analysis of the structures through PROCHECK (Methods Sec. 2.4) also confirms the first structure with C- score -2.36 as the most stable structure among the four (Fig 6).

Table 5: Location of probable N-myristoylation sites in the CRP1 sequence

Amino acid position	*Sequence*
11 - 16:	*GQaaGR*
32 - 37:	*GTaaGA*
43 - 48:	*GQkiTE*
51 - 56:	*GGavNL*

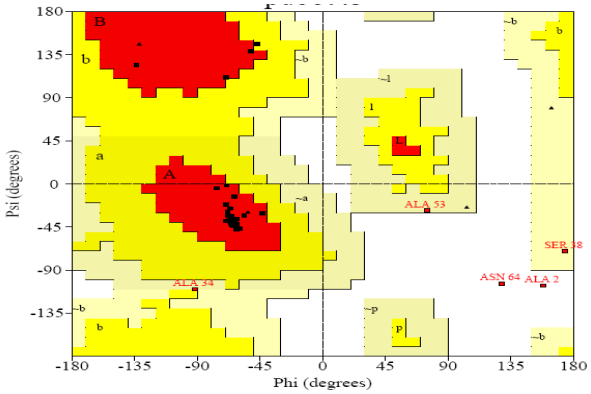

Fig 6: Ramachandran plot showing the phi-psi torsion angles for all residues in most stable Predicted 3-D conformation of CRP1 (except those at the chain termini). Glycine residues are separately identified by triangles, as these are not restricted to the regions of the plot appropriate to the other sidechain types. The colouring/shading on the plot represents the different regions described in Morris et al. (1992): the darkest areas (here shown in red) correspond to the "core" regions representing the most favourable combinations of phi-psi values. The different regions on the Ramachandran plot are as described in Morris et al. (1992). The regions are labelled as: A - Core alpha, L- Core left-handed alpha, a- Allowed alpha, l- Allowed left-handed alpha, ~a- Generous alpha, ~l- Generous left-handed alpha, B- Core beta, p - Allowed epsilon, b- Allowed beta, ~p- Generous epsilon, ~b- Generous beta.

Plot statistics		
Residues in most favoured regions [A,B,L]	**48**	**88.9%**
Residues in additional allowed regions [a,b,l,p]	**1**	**1.9%**
Residues in generously allowed regions [~a,~b,~l,~p]	**3**	**5.6%**
Residues in disallowed regions	**2**	**3.7%**
----------	----	------
Number of non-glycine and non-proline residues	**54**	**100.0%**
Number of end-residues (excl. Gly and Pro)	**2**	
Number of glycine residues (shown as triangles)	**9**	
Number of proline residues	**0**	
----------	----	
Total number of residues	**65**	

IV. Conclusion:

This study presents a comprehensive in silico assessment of physical properties associated with Cold resistance protein 1 (CRP1) from Brassica *oleraceae* and provides indications about its possible 3D structure and cellular localization.

V. Acknowledgements:

This work was partly supported by Jammu and Kashmir, Science and Technology council. Fellowships in favor of NJ from University Grants Commission, India is duly acknowledged.

References

[1]. N.Jan, Cloning, sequence analysis and characterization of cold resistance genes from Brassica Species, doctoral diss., University of Kashmir, J&K, India ,2009.
[2]. C. B.Anfinsen, Principles that govern the folding of protein chains. *Science,181,* 1973, 223-230
[3]. F.J.Corrales and A.R.Fersht, Kinetic significance of GroEL14. (GroES7)2 complexes in molecular chaperone activity. *Folding & Design, 1,* 1996, 265-273.
[4]. M.Levitt and A.Warshel, Computer Simulation of Protein Folding. *Nature, 253,* 1975, 694-698.
[5]. W.F.Van Gunsteren, Molecular dynamics studies of proteins. *Current Opinion in Structural Biology, 3,* 1993,167-174.
[6]. J.Kyte and R.F.Doolittle, A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology, 157,*1982,105–132.
[7]. D.T.Jones, Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology,292,*1999,195-202.
[8]. C.Cole,J.D Barber and G.J.Barton, The Jpred 3 secondary structure prediction server *Nucleic Acids Research,36,* 2008, W197-W201.
[9]. J.A.Cuff and G.J.Barton, Application of enhanced multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins,40,* 1999,502-511.

[10]. J.Cheng, A.Randall, M.Sweredoski and P.Baldi, SCRATCH: a Protein Structure and Structural Feature Prediction Server, *Nucleic Acids Research,33*, 2005,72-76.
[11]. L.J.McGuffin, and D.T.Jones, Improvement of the GenTHREADER method for genomic fold recognition. *Bioinformatics,19*,2003, 874-881.
[12]. Y.Zhang, Template-based modeling and free modeling by I-TASSER in CASP7. *Proteins,8*, 2007,108-117.
[13]. S.Wu, J.Skolnick and Y. Zhang, Ab initio modeling of small proteins by iterative TASSER simulations. *BMC Biology,5*,2007,5-17.
[14]. Y.Zhang, I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics,9*, 2008, 40
[15]. Y.Zhang, and J.Skolnick, Automated structure prediction of weakly homologous proteins on a genomic scale. *Proceedings of the National Academy of Sciences (USA),101*, 2004, 7594-7599.
[16]. K. Karplus, C. Barrett, and R. Hughey, Hidden Markov models for detecting remote protein homologies. *Bioinformatics 14*,1998, 846-856.
[17]. S.F.Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, & and D.J, Lipman,. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research, 25*,1997, 3389-3402.
[18]. S.B. Needleman and C.D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology, 48*,1970, 443-453.
[19]. T.F. Smith and M.S. Waterman, Identification of common molecular subsequences. *Journal of Molecular Biology*,147, 1981, 195-197.
[20]. Y. Zhang, A. Kolinski and J. Skolnick, TOUCHSTONE II: A new approach to ab initio protein structure prediction. *Biophysical journal*, 85, 2003, 1145-1164.
[21]. Y. Zhang, D. Kihara and J. Skolnick, Local energy landscape flattening: Parallel hyperbolic Monte Carlo sampling of protein folding. *Proteins*, 48, 2002, 192-201.
[22]. Y. Zhang and J. Skolnick, Automated structure prediction of weakly homologous proteins on a genomic scale. *Proceedings of the National Academy of Sciences of the United States of America*, 101, 2004a,. 7594-7599.
[23]. Z Y. Zhang and J. Skolnick, TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic acids research, 33*, 2005, 2302-2309.
[24]. N. Guex, and M.C. Peitsch, SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative proteinmodeling. *Electrophoresis, 18*, 1997, 2714-2723.
[25]. R. A. Laskowski, M.W.MacArthur, D. S. Moss, and J. M. Thornton, PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography, 26*, 1993, 283-291.
[26]. M. A. Andrade, S. I. O'Donoghue, and B. Rost, Adaptation of protein surfaces to subcellular location. *Journal of Molecular Biology, 276*, 1998, 517-525.
[27]. D.A. Towler, J.I. Gordon, S.P. Adams, and L. Glaser, The biology and enzymology of eukaryotic protein acylation. *Annual Reviews in Biochemistry, 57*, 1988, 69-99.
[28.] R.J.A. Grand, Acylation of viral and eukaryotic proteins. *Biochemical Journal, 258*, 1989, 625-638.