Interdiscip Sci Comput Life Sci (2012) 4: 296–301
DOI: 10.1007/s12539-012-0143-8

Prot-Prop: J-Tool to Predict the Subcellular Location of Proteins Based on Physiochemical Characterization

Brindha SENTHILKUMAR[1], Sangzuala SAILO[1], Gurusubramanian GURUSWAMI[2], Senthilkumar NACHIMUTHU[1*]
[1](Bioinformatics Infrastructure Facility, Department of Biotechnology, Mizoram University, Aizawl 796004, Mizoram, India)
[2](Department of Zoology, Mizoram University, Aizawl 796004, Mizoram, India)

Received 14 March 2012 / Revised 28 April 2012 / Accepted 7 June 2012

Abstract: PROT-PROP is a computational tool to characterize 27 physicochemical properties of a protein along with its subcellular location (intra or extra) in a single-window application. Other significant features of this software include calculation of numerical values for hydrophobicity, hydrophilicity; composition of small and large amino acids; net hydrophobic content in terms of low/high; and Navie's algorithm to calculate theoretical pI. PROT-PROP is an easy-to-install platform independent implementation of JAVA under a user-friendly interface. It is a standalone version as a virtual appliance and source code for platforms supporting Java 1.5.0 and higher versions, and downloadable from the web www.mzu.edu.in/schools/biotechnology.html. PROT-PROP can run under Windows and Macintosh Operating Systems. PROT-PROP is distributed with its source code so that it may be adapted or customized, if desired.
Key words: software, protein analysis, physicochemical properties, sub-cellular location.

1 Introduction

In nature, the amino acids constitute a protein molecule joined by peptide bonds: there are twenty-two different amino acids having different side chains and hence different physicochemical properties (Aftabuddin *et al.*, 2007). Proteins constantly change shape and form to perform their biological roles. Amino acid residues are likely to be evolutionarily conserved in a protein family because they play an important role in stability and function (Ora and Baker, 2003). Residues important for stability are found in the hydrophobic core (Shortle, 1992) and functional residues are close together in protein-protein interfaces (DeLano, 2002). Understanding the amino acids composition and physicochemical features of protein is essential for knowing the structure and function that are evolutionarily conserved (Ora and Baker, 2003).

The existing methods predict cellular localization of protein sequences according to their amino acid frequencies or by using data mining techniques (Cedano *et al.*, 1997; Ahmad and Sarai, 2004; Bhasin *et al.*, 2005; Gasteiger *et al.*, 2005; Wang *et al.*, 2005; Rashid *et al.*, 2007). The objective of the study is to develop an integrated tool to calculate the physicochemical properties of proteins and to predict their localization using amino acid characteristics under a single-window application.

*Corresponding author.
E-mail: nskmzu@gmail.com

2 Materials and methods

2.1 System specification

PROT-PROP is converted to a JAR (Java Archive) file for portability and convenience. It can either be used online from www.mzu.edu.in/schools/biotechnology.html or downloadable. It runs on DOS, Windows and Macintosh operating systems provided with jdk 1.5.0 or above. PROT-PROP is a pioneer tool which has incorporated 27 unique physicochemical properties under a single window application (Fig. 1).

2.2 Physicochemical characterization

Molecular weight was calculated using IUPAC 1997 standards (pH = 7.0). The following physicochemical properties were calculated using standard formulas: theoretical pI (Navie Alogirthm); extinction co-efficient (Gill and von Hippel, 1989); absorbance = extinction co-efficient/molecular weight; aliphatic index (Ikai, 1980); GRAVY (Kyte and Doolittle, 1982); residue volume (Creighton, 1993); half life (Varshavsky, 1997; Gonda *et al.*, 1989); aromaticity score = phe+trp+tyr/sequence length x100; amino acid residue = No. of individual amino acid; amino acid composition = amino acid residue/sequence length x 100; hydrophobic amino acids = total No. of hydropho-

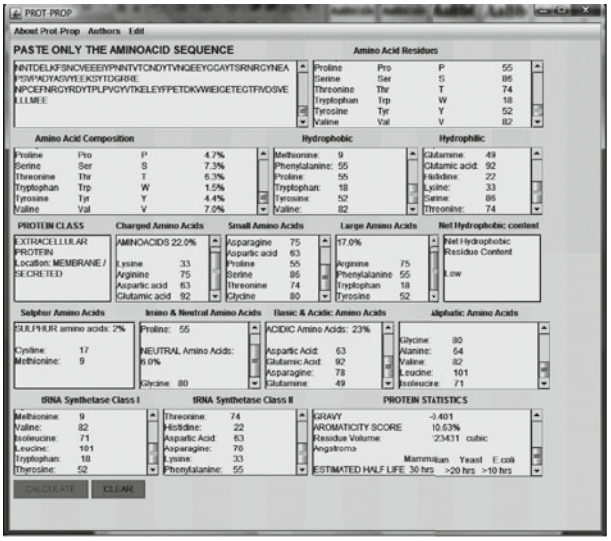

Fig. 1 Snap shot of PROT-PROP showing 27 physicochemical properties.

bic residues/sequence length x 100 (hydrophobic amino acids: Ala, Cys, Ile, Leu, Met, Phe, Pro, Trp, Tyr and Val); hydrophilic amino acids = total No. of hydrophilic residues/sequence length x 100 (hydrophilic amino acids: Arg, Asn, Asp, Gln, Glu, His, Lys, Ser and Thr); net hydrophobic content = hydrophobic – hydrophilic (negative – net hydrophobic content is low and positive – net hydrophobic content is hight); neutral amino acid = total No. of glycine residues/sequence length x 100; charged amino acids = total No. of charged residues/sequence length x 100 (charged amino acids: Lys, Arg, Asp, and Glu); small amino acids = total No. of small amino acid residues/sequence length x 100 (small amino acids: Ala, Asn, Asp, Pro, Ser, Thr and Gly); large amino acids = total No. of large amino acid residues/sequence length x 100 (large amino acids: Arg, Phe, Trp and Tyr); sulfur amino acids = total No. of sulfur amino acid residues/sequence length x 100 (sulfur amino acids: Met and Cys); basic amino acids = total No. of basic amino acid residues/sequence length x 100 (basic amino acids: Lys, Arg, and His); acidic amino acids= total No. of acidic amino acid residues/sequence length x 100 (acidic amino acids: Asp, Glu, Asn and Gln); aliphatic amino acids = total No. of aliphatic (hydroxyl R-groups & R-groups) amino acid residues/sequence length x 100 (aliphatic hydroxyl R-group amino acids: Ser and Thr, aliphatic R-group amino acids: Gly, Ala, Val, Leu and Ile); tRNA synthetase class I = total No. of tRNA synthetase class I amino acid residues/sequence length x 100 (tRNA synthatase class I amino acids: Glu, Gln, Arg, Cys, Met, Val, Ile, Leu, Trp and Tyr); tRNA synthetase class II = total No. of tRNA synthetase class II amino acid residues/sequence length x 100 (tRNA synthetase class II amino acids: Gly, Ala, Pro, Ser, Thr, His, Asp, Asn, Lys and Phe).

2.3 Protein sub-cellular localization

One hundred proteins from membrane integral, membrane anchored, nuclear, extracellular and intracellular proteins were retrieved from Swiss Prot Protein Database (web.expasy.org/groups/swissprot) to find their variations with respective to physicochemical features (Supplementary Table 1).

2.4 Validation and testing

The input data was validated for amino acid sequence and pops-up an error message for invalid sequence (Fig. 2). For a valid amino acid sequence, the extra white spaces were removed to find the appropriate sequence length and 27 physicochemical properties of a given protein were computed (Fig. 1).

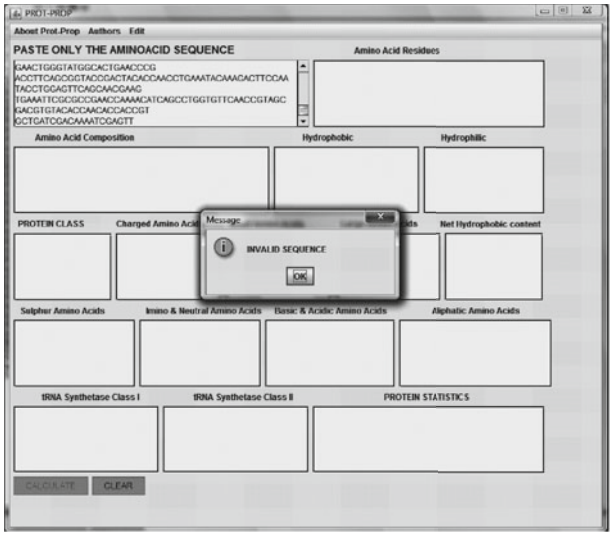

Fig. 2 Snap shot of PROT-PROP showing the validation for amino acid sequence.

3 Results and discussion

3.1 Hierarchical work flow

A hierarchical work flow is shown in Fig. 3 with an objective to calculate 27 physicochemical properties of the input protein sequence and to classify the given protein either as extracellular or intracellular protein.

3.2 Physicochemical characterization

The significance of PROT-PROP is that 27 physiochemical properties are integrated under a single-window system (Fig. 1). The hydrophobic amino acids repel the aqueous environment, reside in the interior of proteins and do not ionize nor form H-bonds. They are involved in the interactions with the receptor protein (Ota *et al.*, 1998) and in transcriptional activation suggesting a relationship between hydrophobicity and activity (Almlof *et al.*, 1997). The hydrophilic amino acids interact with the aqueous environment and are found on the exterior surfaces proteins or in the reac-

Table 1 Algorithm used for protein sub-cellular localization

PROT-PROP Algorithm for predicting sub-cellular location
If (HIGH charged amino acids composition)
If (POOR hydrophobic) AND (HIGH aliphatic index) AND (LOW cystine residues)
"Extracellular protein: Location - Membrane/Secreted"
else
"Intracellular Protein: Location - Nuclear/Cytoplasm
else
if (LOW aliphatic index) AND (HIGH aromatic score) AND (LOW valine residues)
"Intracellular Protein: Location - Nuclear/Cytoplasm
else if (POOR hydrophobic) AND (HIGH aromatic score) AND (HIGH valine residues)
"Intracellular Protein: Location - Nuclear/Cytoplasm
else if (LOW aliphatic index) AND (HIGH valine residues)
"Extracellular protein: Location - Membrane/Secreted"
Else
"Extracellular protein: Location - Membrane/Secreted"

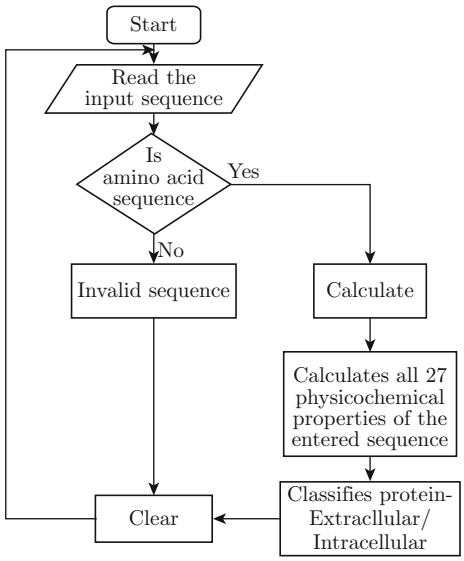

Fig. 3 Hierarchical work flow to find physicochemical characters and sub-cellular location of proteins.

tive centers of enzymes. They are able to make hydrogen bonds to one another, to the peptide backbone, to polar organic molecules and to water (Gregory *et al.*, 2003). Depending on the polarity of the side chain, amino acids vary in their hydrophilic or hydrophobic character. These properties are important in protein structure and protein-protein interactions. The distribution of hydrophilic and hydrophobic amino acids determines the tertiary structure of the protein, and their physical location on the outside structure of the proteins influences their quaternary structure (Meierhenrich, 2008).

Intracellular proteins are relatively poor in cysteine and rich in charged and aliphatic amino acids (Nakashima and Nishikawa, 1984). Nuclear proteins have low hydrophobicity and aromatic residues and rich in charged residues. In the present study, it is significant that both intracellular and nuclear proteins have rich charged residues, high aliphatic index and low aromaticity score, as they both share common properties they are grouped into a single class of intracellular proteins.

The protein hydrophobicity has shown good correlation with the extent to which residues are buried and it could be used to characterize tertiary structures (Manavalan and Ponnusamy, 1978). The hydrophobic residues have an architectural role in protein folding and structure (Betney and McEwan, 2003). A positive hydropathy value indicates increased non-polarity and an increased likelihood that the amino acid would be found inside the hydrophobic core of the protein. The most hydropathic residues are isoleucine, valine, and leucine. The least hydropathic are arginine, lysine, asparagine, and aspartate. The results show that hydropathy is an important amino acid property for disorder.

The higher content of charged residues is related to the polar character of the extracellular medium, favoring its solubility and stability. Charged residues are a significant component in determining the protein subcellular location (Ota *et al.*, 1998). Serine and threonine belong to Aliphatic Hydroxyl Amino Acids group and contain hydroxyl group in their neutral side chains and are polar and hydrophilic.

Nuclear proteins are generally poor in hydrophobicity, especially aromatic amino acid residues (tyrosine, phenylalanine and tryptophan) and rich in charged residues (Cedano *et al.*, 1997). Aromatic residues constitute a special class of amino acids with a partial negative charge and a partial positive charge at their edges. Interactions between aromatic residues have

been shown to contribute substantially to protein stability (Burley and Petsko, 1985). Aromaticity score is the frequency of aromatic amino acids (Phe, Tyr, Trp) in the protein sequence. The hydropathicity and aromaticity protein scores are indices of amino acid usage and are correlated with the variation in the amino acid composition in the *E. coli* (Lobry and Gautier, 1994).

Small amino acids, such as Gly, Ala, Ser, Pro, Val, Thr, Asp and Glu, are relatively stable than large amino acids such as His, Phe, Arg, Tyr, and Trp (Zhang, 2007). Imino amino acid, proline, has a special property of creating kinks in polypeptide chains and disrupting ordered secondary structure. tRNA Synthetase Class I & II Amino Acids are important to know the accuracy of protein translation which inturn depends on the fidelity with which the correct amino acids are esterified to their cognate tRNA molecules by aminoacyl tRNA synthetases (Qiu *et al.*, 1999). The sulfur amino acids (methionine and cysteine) are generally considered to be non-polar and hydrophobic. Methionine is one of the most essential amino acid and is incorporated into the N-terminal position of all proteins in eukaryotes and archaea during translation, although it is usually removed by post-translational modification. Cysteine residues are most frequently buried inside the proteins and in an oxidation reaction yield disulfide bond which plays an important role in the folding and stability of some proteins. Inside the cell, disulfide bridges between cysteine residues within a polypeptide support the protein's secondary structure (Sevier and Kaiser, 2002). In general, sulfur-containing amino acids are essential for a variety of biological activities, including protein synthesis, methylation, polyamine synthesis, coenzyme A production, cysteamine production, taurine production, iron-sulfur cluster (ISC) biosynthesis, and antioxidative stress defense (Ali and Nozaki, 2007; Nozaki *et al.*, 2005). Acidic amino acids (Asp, Glu) are polar and negatively charged and have a second carboxyl group. They play important roles in maintaining the solubility and ionic character of proteins. Basic amino acids are polar, positively charged and are hydrophilic. A critical role for basic amino acid residues in the interaction of numerous proteins with a variety of anionic polymers has been established (DeAngelis and Glabe, 1988).

Extinction coefficients for proteins are generally reported with respect to an absorbance measured at or near a wavelength of 280 nm (Gill and von Hippel, 1989). The half-life is a prediction of the time it takes for half of the amount of protein in a cell to disappear after its synthesis in the cell. PROT-PROP relies on the "N-end rule", which relates the half-life of a protein to the identity of its N-terminal residue; the prediction is given for 3 model organisms (human, yeast and *E. coli*). The N-terminal residue of a protein plays an important role in determining its stability *in vivo* (Gonda *et al.*, 1989). The aliphatic index of a protein is defined as the relative volume occupied by aliphatic side chains (alanine, valine, isoleucine, and leucine). The GRAVY value for a peptide or protein is calculated as the sum of hydropathy values (Kyte and Doolittle, 1982) of all the amino acids, divided by the number of residues in the sequence.

3.3 Protein sub-cellular localization

To frame algorithm for protein classification, 100 proteins each from membrane integral, membrane anchored, nuclear, extracellular and intracellular proteins were used as a training dataset for analysis (Cedano *et al.*, 1997). PROT-PROP classified these five groups under two major classes: extracellular (location: membrane/secreted) and intracellular (location: Nucleus/Cytoplasm) proteins (Table 1). Membrane integral and membrane anchored proteins have the same features as extracellular proteins, whereas nuclear protein has similar characteristics of intracellular proteins. So, in the present study they were combined into two distinct classes. The extracellular proteins have higher cysteine composition resulting in more disulphide bridges (Bradshaw, 1989). Our statistics reveal poor hydrophobicity when charged residues are rich; high hydrophobicity when charged residues are poor. Interestingly, the net hydrophobic content is high/very high and charged residues are poor in most of the membrane proteins especially in membrane anchored. Membrane proteins are rich in hydrophobic amino acids composition corresponding to proteins having several transmembrane stretches of secondary structure and poor charged residues. On contradictory, anchored membrane proteins have one transmembrane stretch (Rost *et al.*, 1985).

3.4 Model validation

The PROT-PROP classified the testing dataset protein with 92% accuracy (Table 2). The 10% loss of accuracy is when an extracellular protein poses the features of intracellular protein or an intracellular protein imitates the features of extracellular protein. This may be due to their function which is irrespective of the protein location. Proteins anchored through a lipid group have a similar composition of an extracellular or intracellular protein. There is a difference between sequence information and the functional characterization of a protein. PROT-PROP includes a tool for analyzing the sequence information via. Physicochemical properties and in turn, predicting the sub-cellular location of the protein. The results are validated with the currently available online software tools and had turned out to be very satisfactory. The time taken to calculate the physicochemical properties in PROT-PROP is quicker than ExPASy-prot param, Gene Infinity – Protein Statistics and Protein Information Resource. Except the input box all other output areas are disabled to avoid accidental modification of predicted results.

Table 2 Protein Testing Dataset used in PROT-PROP

Acc_No.	Location	PROT-PROP Prediction
P02638	Nucleus	IP
O75487	Membrane	EP
Q0657	Anchored	EP
P00270	Cytoplasm	IP
P69986	Multi-pass membrane	EP
P53104	*Cytoplasm*	*EP*
Q92838	*Membrane*	*IP*
P70371	Cytoplasm	IP
P25870	Membrane	EP
Q765A7	Multi-pass membrane	EP
P32303	Anchored	EP
P0CAW7	Cytoplasm	IP
Q57RK6	Membrane	EP
P25714	Multi-pass membrane	EP
P46116	Anchored	EP
Q06118	Cytoplasm	IP
Q9N4M4	Cytoplasm	IP
P35052	Membrane	EP
Q60760	Cytoplasm	IP
Q9JLI3	Membrane + Secreted	EP
Q12355	Membrane + Secreted	EP
P08154	Nucleus	IP
P92934	Anchored	EP
Q9JI78	Cytoplasm	IP
P39935	Cytoplasm	IP

Nucleus/Cytoplasm – Intracellular Protein (IP); Membrane/Secreted - Extracellular Protein (EP).

Twenty-five proteins were taken as testing dataset from the Swiss PROT and tested using the PROT-PROP, the accuracy is about 92%. The proteins (P53104 and Q92838) had been predicted incorrectly because the physiochemical properties of P53104 (rich charged and high aliphatic index with poor aromaticity) are like intracellular protein characteristics. Q92838 is a membrane protein with high hydrophobicity and low charged residues resembling the anchored protein which is an extracellular protein feature as per this program classification.

The uniqueness of PROT-PROP in comparison to the existing software is that it predicts protein localization based on amino acid characteristics and twenty-seven physicochemical properties are calculated in a single-window application.

Electronic Supplementary Material

Supplementary material is available in the online version of this article at http://dx.doi.org/10.1007/s12539-012-0143-8 and is accessible for authorized users.

Acknowledgements The authors thank the Department of Biotechnology (DBT), Government of India, New Delhi for the financial support in the form of Bioinformatics Infrastructure Facility under BTISNet.

References

[1] Aftabuddin, M., Kundu, S. 2007. Hydrophobic, hydrophilic and charged amino acid networks within protein. Biophysic J 93, 225–231.

[2] Ahmad, S., Sarai, A. 2004. Qgrid: Clustering tool for detecting charged and hydrophobic regions in proteins. Nucl Acid Res 32, 104–107.

[3] Ali, V., Nozaki, T. 2007. Current therapeutics, their problems and sulfur-containing-amino acid metabolism as a novel target against infections by amitochondriate protozoan parasites. Clin Microbiol Rev 20, 164–187.

[4] Almlof, T., Gustafsson, J.A., Wright, A.P. 1997. Role of hydrophobic amino acids clusters in the transactivation activity of the human glucocorticoid receptor. Mol Cellular Biol 17, 934–945.

[5] Betney, R., McEwan, I.J. 2003. Role of conserved hydrophobic amino acids in androgen receptors AF-1 function. J Mol Endocrinol 31, 427–439.

[6] Bhasin, M., Garg, A., Raghava, G.P. 2005. PSLpred: Prediction of subcellular localization of bacterial proteins. Bioinformatics 21, 2522–2524.

[7] Bradshaw, R.A. 1989. Protein translocation and turnover in eukaryotic cells. Trends Biochem Sci 14, 276–279.

[8] Burley, S.K., Petsko, G.A. 1985. Aromatic-aromatic interaction: A mechanism of protein structure stabilization. Science 229, 23–28.

[9] Cedano, J., Aloy, P., Perez-Pons, J.A., Querol, E. 1997. Relation between amino acid composition and cellular location of proteins. J Mol Biol 266, 594–600.

[10] Creighton, T.E. 1993. Proteins: Structures and Molecular Properties, 2nd Edition, W.H. Freeman & Company, New York.

[11] DeAngelis, P.L, Glabe, C.G. 1988. Role of basic amino acids in the interaction of binding with sulfated fucans. Biochemistry 27, 8189–8194.

[12] DeLano, W.L. 2002. Unraveling hot spots in binding interfaces: Progress and challenges. Curr Opinion Structural Biol 12, 14–20.

[13] Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M.R., Appel, R.D., Bairoch, A. 2005. Protein identification and analysis tools on the ExPASy server. In: Walker, J.M. (Ed.) The Proteomics Protocols Handbook, Humana Press, New Jersy, 571–607.

[14] Gill, S.C., von Hippel, P.H. 1989. Calculation of protein extinction coefficients from amino acid sequence data. Anal Biochem 182, 319–326.

[15] Gonda, D.K., Bachmair, A., Wunning, I., Tobias, J.W., Lane, W.S., Varshavsky, A.J. 1989. Universality and structure of the N-end rule. The J Biol Chem 264, 16700–16712.

[16] Gregory, A.P., Dagmar, R. 2003. Protein motifs. In: Gregory, A.P., Dagmar, R., Waltham, M.A. (Eds.) Protein Structure and Function, 4th Edition, New Science Press, London, 89–101.

[17] Ikai, A.J. 1980. Thermostability and aliphatic index of globular proteins. The J Biochem 88, 1895–1898.

[18] Kyte, J., Doolittle, R.F. 1982. A simple method for displaying the hydropathic character of a protein. J Mol Biol 157, 105–132.

[19] Lobry, J.R., Gautier, C. 1994. Hydrophobicity, expressivity and aromaticity are the major trends of amino acid usage in 999 *Escherichia coli* chromosome encoded genes. Nucleic Acids Res 22, 3174–3180.

[20] Manavalan, P., Ponnusamy, P.K. 1978. Hydrophobic character of amino acid residues in globular proteins. Nature 275, 673–674.

[21] Meierhenrich, U.J. 2008. Amino acids and the asymmetry of life. In: Advances in Astrobiology and Biogeophysics, Springer-Verlag, Berlin, Heidelberg, New York.

[22] Nakashima, H., Nishikawa, K. 1994. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. J Mol Biol 238, 54–56.

[23] Nozaki, T., Ali, V., Tokoro, M. 2005. Sulfur-containing amino acid metabolism in parasitic protozoa. Adv Parasitol 60, 1–99.

[24] Ora, S.F., Baker, D. 2003. Conserved residue clustering and protein structure prediction. Proteins: Str Func Genet 52, 225–235.

[25] Ota, M., Shimizu, Y., Tonosaki, K., Ariyoshi, Y. 1998. Role of hydrophobic amino acids in gurmarin, a sweetness-suppressing polypeptide. Biopolymers 45, 231–238.

[26] Qiu, X., Janson, C.A., Blackburn, M., Chhohan, I., Hibbs, M., Abdel-Meguid, S. 1999. Cooperative structural dynamics and a novel fidelity mechanism in histidyl-tRNA synthetases. Biochemistry 38, 12296–12304.

[27] Rashid, M., Saha, S., Raghava, G.P. 2007. Support Vector Machine-based method for predicting subcellular localization of mycobacterial proteins using evolutionary information and motifs. BMC Bioinformatics 8, 337–345.

[28] Rost, B., Casadio, R., Fariselli, P., Sander, C. 1995. Transmembrane helices predicted at 95% accuracy. Protein Sci 4, 521-533.

[29] Sevier, C.S., Kaiser, C.A. 2002. Formation and transfer of disulphide bonds in living cells. Nat Rev Mol Cell Biol 3, 836–847.

[30] Shortle, D. 1992. Mutational studies of protein structures and their stabilities. Quarterly Rev Biophysics 25, 205–250.

[31] Varshavsky, A. 1997. The N-end rule pathway of protein degradation. Genes Cells 2, 13–28.

[32] Wang, J., Sung, W.K., Krishnan, A., Li, K.B. 2005. Protein subcellular localization prediction for Gram-negative bacteria using amino acid subalphabets and a combination of multiple support vector machines. BMC Bioinformatics 6, 174–183.

[33] Zhang, H.Y. 2007. Exploring the evolution of standard amino-acid alphabet: When genomics meets thermodynamics. Biochem Biophy Res Comm 359, 403–405.