Indian Journal of Biochemistry & Biophysics
Vol. 47, December 2010, pp. 370-377

Assessing the relationship among physicochemical properties of proteins with respect to hydrophobicity: A case study on AGC kinase superfamily

Amit Kumar Banerjee, B Poorna Manasa and Upadhyayula Suryanarayana Murty*

Bioinformatics Group, Biology Division, Indian Institute of Chemical Technology (C.S.I.R), Hyderabad 500 007, Andhra Pradesh, India

Received 11 November 2009; revised 10 July 2010

Understanding the protein structures is crucial, as it is involved in every cellular activity. Several experimental techniques, such as X-Ray crystallography, nuclear magnetic resonance and electron microscopy are available to gain insight about the structure and function of a protein molecule. Gigantic data on protein structural and sequential information is deposited in various repositories regularly which provide us the scope for more theoretical studies. Hydrophobicity always plays a vital role in tertiary structure formation and behavior of a protein molecule. This study focuses on elucidating influence of several physicochemical properties on hydrophobicity of AGC kinase proteins. AGC kinase superfamily is selected due to its tremendous structural and functional variability and sequence data availability. A combined data mining and stochastic approach confirmed that out of 47 parameters, transmembrane tendency influences the target variable most, followed by percent buried residues, GRAVY (Grand Average Hydropathicity) and aliphatic index. Calculating the influence of different physicochemical parameters and their interrelation will aid tremendously in the future of protein science.

Keywords: CART, AGC Kinase, Hydrophobicity**,** Data Mining**,** *In Silico*, Stochastic method.

Protein folding continues to lure the researchers across the globe for its complex multidimensional nature. How a protein attains its particular native conformation as a product of the regular expression of a gene remains a paradox[1]. Numerous factors support a protein molecule to achieve its native and stable conformation successfully on gene expression invariably. The "hydrophobic effect" is considered as one of the major driving force among all sundry factors in protein folding[2].

The upsurge in structural and sequence data submission in public domain repository and the burgeoning rise in modern data analysis methodologies have provided a major thrust to studies aiming at seeking molecular insights into behavior and interaction of proteins. It is suggested that surface hydrophobicity contributes to protein-protein recognition[3]. Hydrophobicity plays an important role in determining protein disorder[4]. This key factor has help directly or indirectly in developing several computing tools for protein property calculation[5]. Hydrophobic residues tend to be embedded in the core of a protein molecule, whereas hydrophilic residues prefer the surface region and interact with the solvent molecule[4-7]. These hydrophobic residues residing in the core exhibit greater conservation.

In order to appreciate and comprehend the complexity of protein-protein interaction, inter-relationship of affecting physicochemical factors needs to be analyzed intensely. Real time calculations involve dynamics simulations to mimic the cellular environment and understand the behavior of the molecules in aqueous solution. Molecular dynamics calculations provide the answers to such problem[8]. Dynamics studies increase the complexity in computational calculations and warrant the need for alternative approaches for simplification of the equations involved[9]. This can be made possible by considering a quantified view of interrelationship among properties and can aid in achieving more realistic outputs.

With time, a tremendous improvement has been observed in the data analyses methodologies and statistical analyses. Several algorithms and computer programs are now available which can handle ample amount of data and perform efficient analysis. Different algorithms are available which can mimic

*Author for correspondence:
Phone: 91-40-27193134,
Fax: +91-040-27193227
E-mail: murty_usn@yahoo.com

the natural process (popularly known as "evolutionary algorithms"), such as neural networks, genetic algorithm and ant colony optimization are extensively used for protein data analysis across the globe[10-12]. Besides these methodologies, decision based trees also occupy a prominent place in data classification and understanding the behavior of complex biological data set[13]. Recent developments in bioinformatics and data mining technology[14-16] may aid in approaching this multidimensional problem with a new perspective by unveiling some unique evidences or clues on the pattern or the interrelationship of the complex factors.

AGC kinase superfamily displays functional and structural divergence[17]. Data mining techniques have been used in past to facilitate generation of knowledge on interplay of physicochemical parameters in protein families[18,19]. Most of the AGC kinases play a vital role in cell signaling which is manifested by their existence as membrane proteins. The hydrophobicity of these kinases aids in achieving proper conformation and function of these protein kinases. Therefore, understanding the influence of different parameters on hydrophobicity is essential. To be able to perform their biological function, proteins fold into one or more specific spatial conformations, driven by a number of noncovalent interactions such as hydrophobicity.

In this investigation, we have assessed the relationship and interdependence of various parameters that help the protein to attain their structural and functional specificity of various protein kinases belonging to AGC kinase superfamily, employing a simplistic stochastic approach. The objective of this study is to dissect the impact of several physicochemical parameters on hydrophobicity in AGC kinase family. The sequences have been collected from a wide spectrum of organisms and thus, chance of obtaining biased result is eliminated. Though the apparent variation prevails in the primary data sequences, yet there remains a strong opportunity of generating association rules using data mining technology on account of diverse source organisms.

Methodology

AGC kinase protein family was selected as the target protein family for this study and all the sequence data was extracted from NCBI (http://www.ncbi.nlm.nih.gov)[20] protein sequence database. The collected raw data sequences were filtered to ensure elimination of partial sequences and minimizing redundancy present in the initial dataset. The overall workflow is represented in Fig. 1.

Sequence collection

Total 1247 sequences belonging to different source organisms ranging from prokaryotes to eukaryotes were collected. A strict filtering process was adopted for removing any kind of redundancy in the dataset. After initial filtering, 656 sequences were obtained. The variation in number of amino acid residues in considered proteins is shown in Fig. 2.

Extraction of features

Physicochemical properties of a protein provide proper insight of the structural and functional behavior of the molecule. To obtain data regarding certain physicochemical features, standard EXPASY tools, Protscale (http://expasy org/tools/protscale.html) and Protparam (http://expasy org/tools/protparam.html)[21] were employed. Total 47 parameters that are known to influence structural and functional behavior of a protein were calculated by these servers and considered for further analyses.

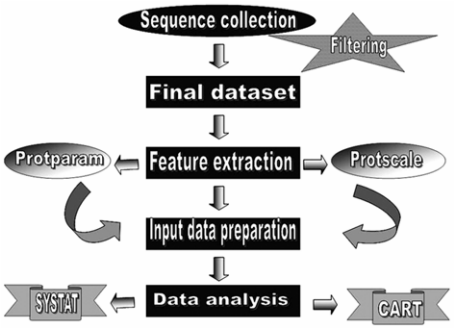

Fig. 1—Strategic workflow of the study

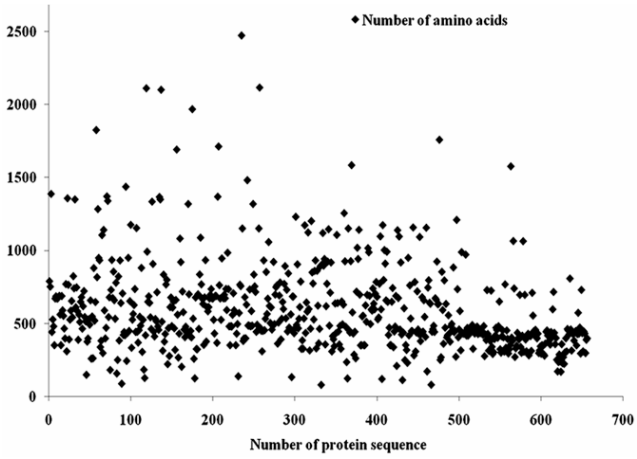

Fig. 2—Scatter plot of the number of amino acids present in the sequences considered in this study

Protparam calculated parameters were considered directly, whereas computed average scales were considered from the Protscale server for the study. All the parameters considered for this study are listed in Table 1.

Data analysis

Classification and regression trees are ideally suited for the analysis of complex data. For such data, we require flexible and robust analytical methods, which can deal with non-linear relationships, high-order interactions, and missing values. CART is a robust decision–tree based tool for data mining and predictive modeling. It was applied for analyzing the obtained data.

Selection of target and predictor variables

Out of 47 parameters, hydrophobicity was considered as the target variable and all other parameters were considered as predictor variables. Since target variable hydrophobicity is continuous in nature, regression tree was selected for analysis. During the parameter adjustment of the tool, no priors were selected and default parameters were used for the penalty, as there were no missing values and bias in the data. Least square splitting method was applied for this study. V-fold cross-validation was selected for testing the obtained trees. This method is highly accurate and has advantage of not requiring a separate, independent dataset for assessing the accuracy and the size of the tree. V-fold cross validation performs partitioning of data into equal-sized segments and holds out one segment at a time for test purposes.

Table 1—Calculated parameters with respective tools applied

S.No	Parameters obtained from Protscale calculation
1	Molecular weight
2	Number of codon(s)
3	Bulkiness
4	Polarity (Zimmerman)
5	Refractivity
6	Hydrophobicity (Kyte & Doolittle)
7	Transmembrane tendency
8	% Buried residues
9	% Accessible residues
10	Average area buried
11	Average flexibility
12	α helix(Chou & Fasman)
13	β sheet (Chou & Fasman)
14	β turn (Chou & Fasman)
15	Coil (Deleage & Roux)
16	Total β strand
17	Antiparallel β strand
18	Parallel β strand
19	Amino acid composition
20	Relative mutability
	Parameters obtained from Protparam calculation
1	Theoretical pI (Isoelectric point)
2	Individual amino acid composition
3	Total number of negatively charged residues
4	Total number of positively charged residues
5	Extinction coefficients
6	Instability index
7	Aliphatic index
8	Grand Average of Hydropathicity (GRAVY)

Statistical calculation

Statistical analysis was carried out by SYSTAT software (http://www.systat.com/Default.aspx)[22]. The analysis was performed by selecting the correlation option, where Pearson's correlation coefficient (r_{xy}) was calculated for the variables using the following formula.

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(n-1)S_x S_y}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of X and Y, S_x and S_y are the sample standard deviations of X and Y and the sum is from i = 1 to n. The correlation coefficients were calculated for each parameter with other parameters.

Results and Discussion

Data mining techniques have proved their efficiency as meaningful technology long back. Enormous literature supports their ability to solve complex non-linear classification and clustering problems.

Classification and regression analysis

Decision tree learning is a common method used in data mining. A decision tree is a flow chart of diagram representing a classification system or predictive model. The tree is structured as a sequence or simple questions and the answers to these questions trace a path down the tree. A decision tree can be described also as the combination of mathematical and computing techniques to aid the description, categorization and generalization of a given set of data. CART[23] is known to be one of the best data mining tool in recent times and finds enormous application in complex biological data analysis, especially in molecular biology[24], microbiology[25,26], medical sciences[27,28], genomics[29], proteomics[30] and other important areas.

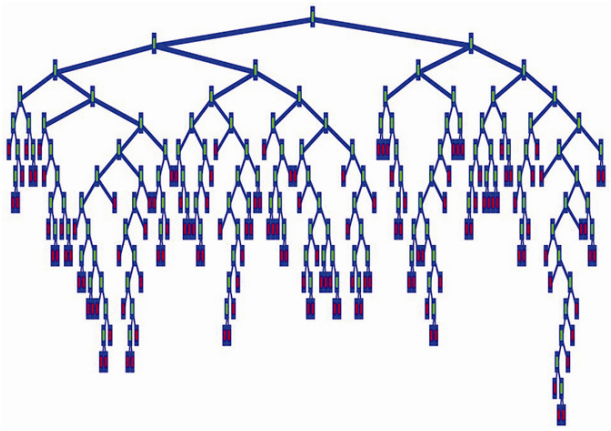

Fig. 3—Obtained tree topology for hydrophobicity (Kyte & Doolittle) target against all other parameters as predictor

Optimal tree with 29 (Tree no. 102 in Annexure I) nodes was obtained with cross-validated error of 0.32673 ± 0.02336, resubstitution error of 0.15574 and complexity of 0.28214. The full grown tree topology is represented in Fig. 3. When tree with 29 nodes was considered, parameters like grand average hydropathicity (GRAVY), polarity, transmembrane tendency, relative mutability, aliphatic index, antiparallel β strand, percent buried residues appeared into the decision tree development. Using this decision tree, the impact of various parameters on hydrophobicity was assessed, estimated and deduced to numerical representation.

Figure 4A depicts the obtained error curve along with the relative error observed for the optimal tree (0.327) with 29 nodes. The full-grown tree represents 136 nodes with a relative error of 0.398. The terminal nodes sorted by hydrophobicity (Kyte & Doolittle) are represented in Fig. 4B.

Variable importance

In this rigorous regression analysis, the variable importance determined by the algorithm clearly indicates the impact of several variables on the target variable. Variable importance is determined by looking at every node in which a variable appears and taking into account how good a splitter it is. Variable importance ranking is a summary of a variable's contribution to the overall tree, when all nodes are examined. Variables earn credit towards their importance in a CART tree in two ways, as primary splitters that actually split a node, and as surrogate splitters (backup splitters to be used, when the primary splitter is missing). Variable importance of various parameters on hydrophobicity was obtained using CART (Table 2).

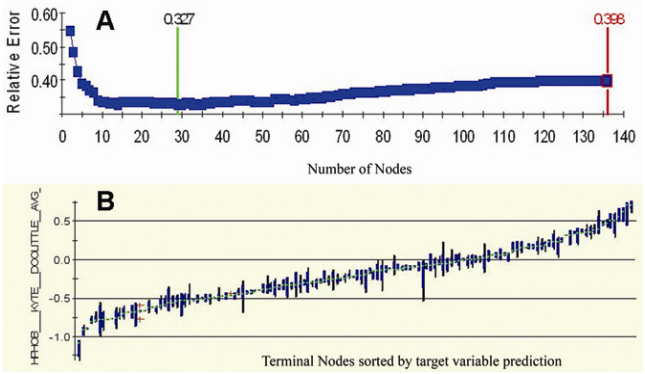

Fig. 4—(A): Obtained error curve where relative error (X axis) versus number of nodes (Y-axis) is represented; and (B): Terminal nodes sorted by target variable prediction. These terminal nodes are linked to association rules, which are used to define the interrelationship among properties

Out of all, predictor variables transmembrane tendency, percent buried residues, GRAVY and aliphatic index were found to influence hydrophobicity in descending order.

Association rules generation

Association rules were generated using a 136-node tree (maximal grown tree). Tree with least complexity and minimum relative resubstitution error was considered for obtaining association rules. For example:

Rule (136^{th} node of 136 nodes tree), IF "transmembrane tendency > -0.26975 and parallel beta strand > 1.316 and phenylalanine > 3.3" THEN "hydrophobicity is 0.712643" (maximum in this case) (In Annexure).

Rule (1^{st} node of 136 nodes tree), IF "grand average of hydropathicity < = -0.351 and transmembrane tendency <= -0.7605 and % buried residues < = 5.58325, and beta sheet < = 0.93025" THEN "hydrophobicity is -1.13063" (minimum in this case) (In Annexure).

Rule (70^{th} node of 136 nodes tree), IF "grand average of hydropathicity > -0.351 and transmembrane tendency > -0.560725 & <= -0.41575 and % buried residues > 6.064 and parallel beta strand <= 1.14375 and phenylalanine <= 4.4 and threonine > 5 and tryptophan > 0.25 and proline <= 4.6 and instability index > 24.15 and coil (Deleage & Roux) > 0.98" THEN "hydrophobicity is -0.216667" (intermediate in this case) (In Annexure).

Statistical analysis

SYSTAT package was used to validate the CART analysis. Hydrophobicity was selected as the target variable and rest of the parameters was selected

Table 2—Importance of Variables as elucidated using CART

S.No.	Variable	Score	S.No.	Variable	Score
1	Transmembrane tendency	100	24	Proline	2.55
2	Percent Buried Residues	89.78	25	Average flexibility	2.49
3	Grand average of hydropathicity	83.1	26	Instability index	2.47
4	Aliphatic index	69.74	27	Coil (Deleage & Roux)	2.21
5	Total β strand	56.58	28	Relative mutability	2.16
6	β sheet (Chou& Fasman)	45.89	29	Total number of positive charged residues	2.16
7	Parallel β strand	31.63	30	α helix(Chou & Fasman)	2.14
8	Alanine	11.2	31	Antiparallel beta strand	2.12
9	Lysine	9.41	32	Isoleucine	2.12
10	Polarity (Zimmerman)	7.35	33	Theoretical pI	2.11
11	Valine	6.33	34	Phenylalanine	1.77
12	Amino acid composition	5.23	35	Tyrosine	1.74
13	Refractivity	5.21	36	Cysteine	1.71
14	Histidine	4.91	37	β turn (Chou & Fasman)	1.7
15	Average area buried	4.76	38	Glycine	1.69
16	Leucine	4.49	39	Arginine	1.67
17	Molecular weight	3.61	40	Serine	1.34
18	Aspartic acid	3.56	41	Number of codon (s)	1.31
19	Bulkiness	3.27	42	Extinction coefficient	1.26
20	Glutamine	3.16	43	Metheonine	1.19
21	Total number of negatively charged residues	3.14	44	Threonine	1.14
22	Asparagine	2.73	45	Glutamic acid	1.07
23	Percent Accessible residues	2.71	46	Tryptophan	0.51

as independent variables. The correlation between hydrophobicity and transmembrane tendency was 0.774 (maximum) (Annexure 2) in this case, followed by percent buried residues, GRAVY, aliphatic index, total β strand which exerted a positive influence on it. Other parameters β turn, lysine, total number of negatively charged residues, total number of positively charged residues and polarity showed negative correlation in the statistical calculations.

Both the analyses performed by CART and SYSTAT suggested the importance of tansmembrane tendency, percent buried residues, GRAVY and aliphatic index as the most influential factors for the hydrophobicity of the protein molecules in AGC protein kinase family.

Hydrophobicity is known to be an important determinant of transmembrane tendency. This study attempts at gaining an insight on the influence of various sequence and structural features on an inherent property of proteins viz. hydrophobicity. CART yielded association rules, which can assist in decision making process by serving as "rule of thumb", when applied to a huge dataset in crucial experimental procedures, where knowledge of parametric influence on transmembrane tendency and hydrophobicity is vital.

Conclusion

Numerous statistical analyses in the past have reinstated the complexity of protein folding problem. In essence, protein structure depends on its physiochemical properties. This study paves a way for understanding several physicochemical parameters in detail and their influence on a particular property stochastically. A case on AGC kinase protein superfamily was taken into account owing to their functional and structural diversity. Similar kind of approach can be implemented in more complex and larger dataset.

Understanding how physiochemical properties influence protein folding is a major challenge. An interdisciplinary approach comprising of both experimental and computational methodologies is needed to solve this paradox[31]. Numerous efforts have been made to understand the impact of hydrophobicity on protein folding[32]. We have investigated the impact of different physicochemical parameters on hydrophobicity and generated simplified association rules based on the quantification of their effects. This will help in understanding and quantifying the contribution of various physicochemical parameters involved in protein folding.

This kind of statistical insight will enable us to understand the different structural and functional properties of a protein molecule and quantify or rank them according to their influence on a parameter, thus enlightening the future path for *in silico* elucidation of roles of diverse factors in determining the crucial properties and how such knowledge can be exploited in several aspects of protein folding and enzyme engineering.

Acknowledgments

The authors are thankful to the Director, Indian Institute of Chemical Technology, Hyderabad for his constant support and encouragement through out the study. AKB thanks Council of Scientific and Industrial Research (C.S.I.R) for Senior Research Fellowship (SRF).

References

1 Dill K A, Ozkan S B, Weikl T R, Chodera J D & Voelz V A (2007) *Curr Opin Struct Biol* 17, 342-346
2 Dill K A (1990) *Science* 250, 297-298
3 Young L, Jernigan R L & Covell D G (1994) *Protein Sci* 3, 717-729
4 Keskin O, Gursoy A, Ma B & Nussinov R (2008) *Chem Rev* 108, 1225-1244
5 Higa R H, Togawa R C, Palandrani A J M J C F, Okimoto I K S, Kuser P R, Yamagishi M E B, Mancini A L & Neshich G (2004) *BMC Bioinform* 25, 107
6 Tsai C J, Lin S L, Wolfson H J & Nussinov R (1997) *Protein Sci* 6, 53-64
7 Tsai CJ & Nussinov R (1997) *Protein Sci* 6, 24-42
8 Straub J E, Guevara J, Huo S & Lee J P (2002) *Acc Chem Res* 35, 473–481
9 Komatsu H, Yamasaki T & Ichikawa S (2008) *Fujitsu Sci Tech J* 44, 449-457
10 Yang Z R, Thomson R, McNeil P & Esnouf R M (2005) *Bioinformatics* 21, 3369-3376
11 Del Carpio C A (1996) *J Chem Inf Comput Sci* 36, 258–269
12 Shmygelska A & Hoos H H (2005) *BMC Bioinform* 6, 1-22
13 Vlahou A, Schorge J O, Gregory B W & Coleman R L (2003) *J Biomed Biotech* 5, 308–314
14 Valera V A, Walter B A, Yokoyama N, Koyama Y, Iiai T, Okamoto H & Hatakeyama K (2007) *Ann Surg Oncol* 14, 34-40
15 Valkonen V P, Kolehmainen M, Lakka H M & Salonen J T (2002) *Int J Epidemiol* 31 864-871
16 Huang K & Murphy R F (2004) *BMC Bioinform* 5, 78, 1-19
17 Hanks S K (2003) *Genome Biol* 4, 111, 1-7
18 Murty U S N, Banerjee A K & Arora N (2009) *J Proteomics Bioinform* 2, 97-107
19 Banerjee A K, Arora N & Murty U S N (2008) *Electr J Biol* 4, 27-33
20 Sayers E W, Barrett T, Benson D A, Bryant S H, Canese K, Chetvernin V, Church D M, DiCuccio M, Edgar R, Federhen S, Feolo M, Geer L Y, Helmberg W, Kapustin Y, Landsman D, Lipman D J, Madden T L, Maglott D R, Miller V, Mizrachi I, Ostell J, Pruitt K D, Schuler G D, Sequeira E, Sherry S T, Shumway M, Sirotkin K, Souvorov A, Starchenko G, Tatusova T A, Wagner L, Yaschenko E & Ye J (2009) *Nucleic Acids Res* 37, D5–D15
21 Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins M R, Appel R D & Bairoch A (2005) In: *The Proteomics Protocols Handbook* (John M. Walker, ed), pp 571-607, Humana Press
22 http://www.systat.com/Default.aspx
23 Breiman L, Friedman J H, Olshen R A & Stone C J (1984) Wadsworth International Group, Belmont, California, p 358. Steinberg *et al.*, CART® 6.0 *User's Guide*, Salford Systems
24 Cummings M P & Myers D S (2004) *BMC Bioinform* 16:132, 1-7
25 Cummings M P & Segal M R (2004) *BMC Bioinform* 28, 137-143.
26 Kashuba A D M, Nafziger A N, Drusano G & Bertino J S Jr. (1999) *Antimic Agents Chemother* 43, 623–629
27 Vladimar A V, Beatriz A W, Naoyuki Y, Koyama Y, Tsuneo I, Haruhiko O & Katsuyoshi H (2007) *Ann Surg Oncol* 14, 34-40
28 Takahashi O, Cook E F, Nakamura T, Saito J, Ikawa F & Fukui T (2006) *QJM* 99, 743-750
29 Jin V X, Leu Y W, Liyanarachchi S, Sun H, Fan M, Nephew K P, Huang T H M & Davuluri R V (2004) *Nucleic Acids Res* 32, 6627-6635
30 Markey M K, Tourassi G D & Floyd Jr. C E (2003) *Proteomics* 3, 1678–1679
31 Oleg B P (1987) *J Protein Chem* 6, 273-293
32 Laurence L & Brasseur R (1995) *FASEB J* 9, 535-540

ANNEXURE 1—Number of trees along with cross-validation error, resubstitution error and complexity

Tree No	Terminal nodes	Cross-validated relative error	Resubstitution relative error	Complexity	Tree No	Terminal nodes	Cross-validated relative error	Resubstitution relative error	Complexity
1	136	0.39773 ± 0.02914	0.03347	0.00000	58	78	0.36387 ± 0.02714	0.06378	0.09102
2	135	0.39800 ± 0.02915	0.03350	0.00235	59	77	0.36270 ± 0.02707	0.06478	0.09209
3	134	0.39782 ± 0.02914	0.03354	0.00405	60	76	0.36267 ± 0.02705	0.06579	0.09362
4	133	0.39763 ± 0.02912	0.03361	0.00651	61	75	0.36328 ± 0.02706	0.06681	0.09382
5	132	0.39730 ± 0.02911	0.03374	0.01227	62	74	0.36234 ± 0.02702	0.06784	0.09562
6	131	0.39769 ± 0.02913	0.03390	0.01441	63	73	0.36119 ± 0.02687	0.06889	0.09735
7	130	0.39726 ± 0.02911	0.03408	0.01694	64	72	0.35839 ± 0.02670	0.06997	0.09905
8	129	0.39766 ± 0.02913	0.03427	0.01745	65	71	0.35770 ± 0.02666	0.07107	0.10222
9	128	0.39780 ± 0.02913	0.03449	0.01970	66	70	0.35745 ± 0.02665	0.07221	0.10539
10	127	0.39646 ± 0.02904	0.03473	0.02234	67	69	0.35364 ± 0.02646	0.07337	0.10652
11	126	0.39626 ± 0.02905	0.03499	0.02467	68	67	0.35106 ± 0.02644	0.07572	0.10874
12	125	0.39590 ± 0.02904	0.03528	0.02633	69	66	0.35062 ± 0.02643	0.07694	0.11299
13	124	0.39570 ± 0.02903	0.03558	0.02800	70	65	0.34674 ± 0.02595	0.07817	0.11331
14	123	0.39572 ± 0.02903	0.03589	0.02878	71	64	0.34700 ± 0.02601	0.07950	0.12313
15	122	0.39572 ± 0.02902	0.03621	0.02898	72	63	0.34705 ± 0.02602	0.08087	0.12681
16	121	0.39572 ± 0.02902	0.03652	0.02921	73	62	0.34672 ± 0.02609	0.08226	0.12773
17	120	0.39572 ± 0.02902	0.03684	0.02934	74	61	0.34138 ± 0.02573	0.08377	0.13950
18	119	0.39561 ± 0.02900	0.03716	0.02988	75	60	0.34223 ± 0.02572	0.08531	0.14233
19	118	0.39508 ± 0.02898	0.03751	0.03169	76	59	0.34223 ± 0.02572	0.08686	0.14367
20	117	0.39517 ± 0.02895	0.03787	0.03339	77	58	0.34086 ± 0.02557	0.08843	0.14512
21	116	0.39481 ± 0.02898	0.03824	0.03464	78	56	0.34372 ± 0.02553	0.09190	0.16034
22	115	0.39481 ± 0.02898	0.03863	0.03600	79	55	0.34465 ± 0.02558	0.09366	0.16255
23	114	0.39495 ± 0.02899	0.03902	0.03602	80	53	0.34454 ± 0.02557	0.09722	0.16425
24	113	0.39468 ± 0.02899	0.03942	0.03680	81	52	0.33679 ± 0.02507	0.09900	0.16462
25	112	0.39392 ± 0.02898	0.03983	0.03830	82	51	0.33687 ± 0.02522	0.10092	0.17706
26	111	0.39282 ± 0.02889	0.04029	0.04159	83	49	0.33647 ± 0.02528	0.10487	0.18250
27	110	0.39326 ± 0.02889	0.04074	0.04197	84	48	0.33633 ± 0.02520	0.10700	0.19669
28	109	0.39243 ± 0.02885	0.04120	0.04270	85	47	0.34084 ± 0.02573	0.10920	0.20329
29	108	0.39175 ± 0.02882	0.04169	0.04531	86	46	0.34023 ± 0.02572	0.11140	0.20364
30	107	0.39083 ± 0.02890	0.04220	0.04670	87	44	0.33933 ± 0.02569	0.11587	0.20663
31	106	0.38786 ± 0.02858	0.04272	0.04818	88	43	0.33839 ± 0.02571	0.11813	0.20838
32	105	0.38592 ± 0.02860	0.04330	0.05391	89	42	0.33553 ± 0.02561	0.12046	0.21546
33	104	0.38328 ± 0.02849	0.04390	0.05495	90	41	0.33465 ± 0.02556	0.12286	0.22132
34	103	0.38255 ± 0.02850	0.04452	0.05739	91	40	0.33508 ± 0.02564	0.12528	0.22390
35	102	0.38255 ± 0.02850	0.04514	0.05742	92	39	0.33502 ± 0.02559	0.12775	0.22829
36	101	0.38110 ± 0.02840	0.04576	0.05769	93	38	0.33434 ± 0.02558	0.13028	0.23338
37	100	0.38069 ± 0.02839	0.04642	0.06078	94	37	0.32978 ± 0.02530	0.13281	0.23358
38	99	0.38031 ± 0.02842	0.04708	0.06082	95	36	0.33069 ± 0.02517	0.13541	0.24030
39	97	0.37929 ± 0.02829	0.04842	0.06210	96	35	0.32862 ± 0.02505	0.13807	0.24594
40	96	0.37801 ± 0.02817	0.04911	0.06332	97	34	0.32830 ± 0.02483	0.14088	0.25935
41	95	0.37840 ± 0.02819	0.04980	0.06398	98	33	0.32896 ± 0.02494	0.14379	0.26896
42	94	0.37843 ± 0.02819	0.05049	0.06407	99	32	0.32927 ± 0.02494	0.14674	0.27312
43	93	0.37754 ± 0.02816	0.05120	0.06475	100	31	0.32927 ± 0.02494	0.14971	0.27393
44	92	0.37539 ± 0.02807	0.05190	0.06536	101	30	0.32919 ± 0.02481	0.15269	0.27506
45	91	0.37528 ± 0.02801	0.05262	0.06587	102**	29	0.32673 ± 0.02336	0.15574	0.28214
46	90	0.37481 ± 0.02794	0.05338	0.07029	103	28	0.33146 ± 0.02279	0.15926	0.32492
47	89	0.37417 ± 0.02793	0.05414	0.07058	104	27	0.32957 ± 0.02265	0.16304	0.34955
48	88	0.37344 ± 0.02789	0.05492	0.07160	105	26	0.32958 ± 0.02261	0.16692	0.35864
49	87	0.37175 ± 0.02779	0.05571	0.07319	106	25	0.33034 ± 0.02255	0.17088	0.36523
50	86	0.37222 ± 0.02781	0.05650	0.07362	107	24	0.33034 ± 0.02255	0.17494	0.37539
51	85	0.36982 ± 0.02739	0.05732	0.07492	108	23	0.33213 ± 0.02246	0.17907	0.38116
52	84	0.36860 ± 0.02730	0.05817	0.07887	109	22	0.33344 ± 0.02218	0.18359	0.41775
53	83	0.36873 ± 0.02736	0.05906	0.08238	110	21	0.33604 ± 0.02225	0.18813	0.42012
54	82	0.36917 ± 0.02744	0.05998	0.08478	111	20	0.33584 ± 0.02200	0.19302	0.45092
55	81	0.36775 ± 0.02735	0.06090	0.08480	112	19	0.33625 ± 0.02202	0.19793	0.45379
56	80	0.36812 ± 0.02737	0.06184	0.08727	113	18	0.33625 ± 0.02202	0.20287	0.45666
57	79	0.36631 ± 0.02725	0.06279	0.08816	114	17	0.33618 ± 0.02200	0.20782	0.45761

(*Contd.*)

ANNEXURE 1—Number of trees along with cross-validation error, resubstitution error and complexity

Tree No	Terminal nodes	Cross-validated relative error	Resubstitution relative error	Complexity	Tree No	Terminal nodes	Cross-validated relative error	Resubstitution relative error	Complexity
115	16	0.33419 ± 0.02195	0.21298	0.47625	123	8	0.36106 ± 0.02228	0.27883	1.56528
116	15	0.33407 ± 0.02188	0.21823	0.48506	124	7	0.37110 ± 0.02250	0.29782	1.75451
117	14	0.32737 ± 0.02144	0.22386	0.51996	125	6	0.38027 ± 0.02274	0.31982	2.03266
118	13	0.33052 ± 0.02116	0.23024	0.58941	126	5	0.38994 ± 0.02300	0.34695	2.50587
119	12	0.33288 ± 0.02139	0.23704	0.62803	127	4	0.42414 ± 0.02377	0.37493	2.58519
120	11	0.33556 ± 0.02157	0.24396	0.63938	128	3	0.48132 ± 0.02559	0.44395	6.37653
121	10	0.33446 ± 0.02109	0.25232	0.77276	129	2	0.54566 ± 0.02579	0.53349	8.27122
122	9	0.33746 ± 0.02131	0.26189	0.88379	130	1	1.00001 ± 0.00002	1.00000	43.09816

ANNEXURE 2—Correlation between dependent and independent variables

Parameters	Hydrophobicity (Kyte & Doolittle)
Molecular weight	-0.274
Number of codon(s)	0.180
Bulkiness	0.159
Polarity (Zimmerman)	-0.376
Refractivity	-0.200
Hydrophobicity (Kyte & Doolittle)	1.000
Transmembrane tendency	0.774
% Buried residues	0.691
% Accessible residues	-0.124
Average area buried	0.043
Average flexibility	-0.022
α helix (Chou & Fasman)	0.066
β sheet (Chou & Fasman)	0.229
β turn (Chou & Fasman)	-0.462
Coil (Deleage & Roux)	-0.140
Total β strand	0.500
Antiparallel β strand	0.247
Parallel β strand	0.678
Amino acid composition	0.312
Relative mutability	-0.085
Theoretical pI	-0.007
Alanine	0.461
Arginine	0.218
Aspargine	-0.250
Aspartic acid	-0.204
Cysteine	-0.262
Glutamine	-0.026
Glutamic acid	-0.293
Glycine	0.197
Histidine	-0.149
Isoleucine	0.027
Leucine	0.475
Lysine	-0.501
Metheonine	-0.092
Phenylalanine	-0.273
Proline	-0.060
Serine	-0.092
Threonine	0.050
Tryptophan	-0.036
Tyrosine	-0.314
Valine	0.179
Total number of negatively charged residues	-0.135
Total number of positively charged residues	-0.160
Extinction coefficients	-0.088
Instability index	-0.207
Aliphatic index	0.646
Grand average of hydropathicity (GRAVY)	0.662