Classification and Regression Tree (CART) Analysis for Deriving Variable Importance of Parameters Influencing Average Flexibility of CaMK Kinase Family

Amit Kumar Banerjee, Neelima Arora, U.S.N Murty*

Bioinformatics Group, Biology Division, Indian Institute of Chemical Technology, Uppal Road , Tarnaka, Hyderabad-500607, Andhra Pradesh, India

* Corresponding author. Tel: +91 40 27193134; Fax: +91 40 27193227; E-mail: murty_usn@yahoo.com

Abstract

In this study, data mining approach was used to derive decision rules for predicting average flexibility from the various derived sequence and structural features. 21 parameters were calculated and variable importance was calculated for 101 sequences of CaMK kinase family belonging to mouse and human using Classification and Regression Tree (CART). Coils were found to have maximum influence on average flexibility while the Parallel beta strands were found to exert minimum impact on average flexibility. Understanding the variable importance will prove useful as a simple predictor of flexibility from an amino acid sequence. This will aid in better understanding of phenomenon underlying the average flexibility and thus, will pave a way for rational design of therapeutics and development of proper parametric weight distribution for existing molecular dynamics and protein folding algorithms.

Keywords: Average flexibility, CaMK Kinase, Bioinformatics, Data mining, Classification & Regression tree (CART).

1. Introduction

In this data–rich, information-poor world, extracting meaningful information from the flood of data is a formidable task. Though at its nascent stage, data mining is enabling researchers in demystifying biological processes. The multiplicity of functions of function is attributed to their structure. Given the dynamic nature of proteins, their structure function relationship is being actively investigated. Protein flexibility constitutes a significant linkage between protein structure and function. Conformational changes as and when required in biological processes are facilitated by their inherent flexibility. Proteins are the lead players encompassing a varied range of functions like transport of metabolites [1, 2], catalysis [3, 4] and regulation of protein activity [5, 6] etc, average flexibility holds prime importance in this context. Protein flexibility may influence diminutive changes in conformation to large-scale molecular motions. Various degree of flexibility exhibited by protein molecules often perplexes the researchers. Various studies have been incited after the discovery of role of some highly flexible proteins with implications in pathologies like AIDS (HIV gp41) and scrapie [7].

A comprehensive knowledge of fundamental nature of average flexibility will facilitate the unraveling of structure-function relationship and will also aid in development of novel therapeutics [8]. Thus, a comprehensive understanding of the intricate relationship of factors influencing protein flexibility will aid in the rational design.

The Ca^{2+}/calmodulin-dependent kinase (CaMK) family, which is activated in response to elevation of intracellular Ca^{2+}, includes CaMKI, CaMKII, CaMKIV and CaMK-kinases (CaMKKs). CaMKK/CaMK cascade plays an important role in regulating Ca^{2+} mediated cellular response. There is no dearth of data on flexibility of proteins but most of the studies have focused only on 3-D structure and related parameters. This study is an attempt to investigate the significance of diverse parameters influencing the average flexibility of CaMK kinase family by means of data mining approach.

2. Materials and Methods

2.1 Sequence Collection and Pre-Processing

Protein sequences of the enzymes belonging to CaMK kinases were collected in FASTA format from the NCBI's protein database (http://www.ncbi.nlm.nih.gov) (Supplement). The collected sequences were filtered in order to exclude redundancy. From the available sequences, 101 sequences belonging to *Homo sapiens* (55) and *Mus musculus* (46) were considered for this study.

2.2 Feature Extraction

Sequence features were extracted for these sequences using ProtScale (http://expasy.org/tools/protscale.html). 21 scales like molecular weight, number of codons, bulkiness, polarity [9], refractivity [10], recognition factors [11], hydrophobicity [12], transmembrane tendency [13], % burried residues, % accessible residues, average area buried[14], average flexibility [15], alpha-helix [16], beta-sheet [16], beta-turn [16], coil [17], total beta-strand [18] , antiparallel beta-strand [18], parallel beta-strand [18], amino acid composition [19] and relative mutability [20] were calculated for all the sequences. Being a categorical variable of little importance for further analysis, accession numbers were excluded from the analysis.

2.3 Data mining

CART (Classification And Regression Tree) from Salford Systems Inc, USA is a data-mining tool based on recursive binary partitioning (21). For gaining a comprehensive understanding on influence of different variables on average flexibility, CART was employed to determine variable importance. 20 parameters were considered as predictor (independent) variables and average flexibility was considered as predictive (dependent) variable. As the target variable is continuous variable, regression model using Least Square (LS) method was selected.10 fold cross validation and default options for penalty were kept for the analysis.

3. Results

CART yielded the output of basic statistical analyses performed for all the parameters and the results are represented in Table 1 and frequency distribution for these are presented in Figure 1.

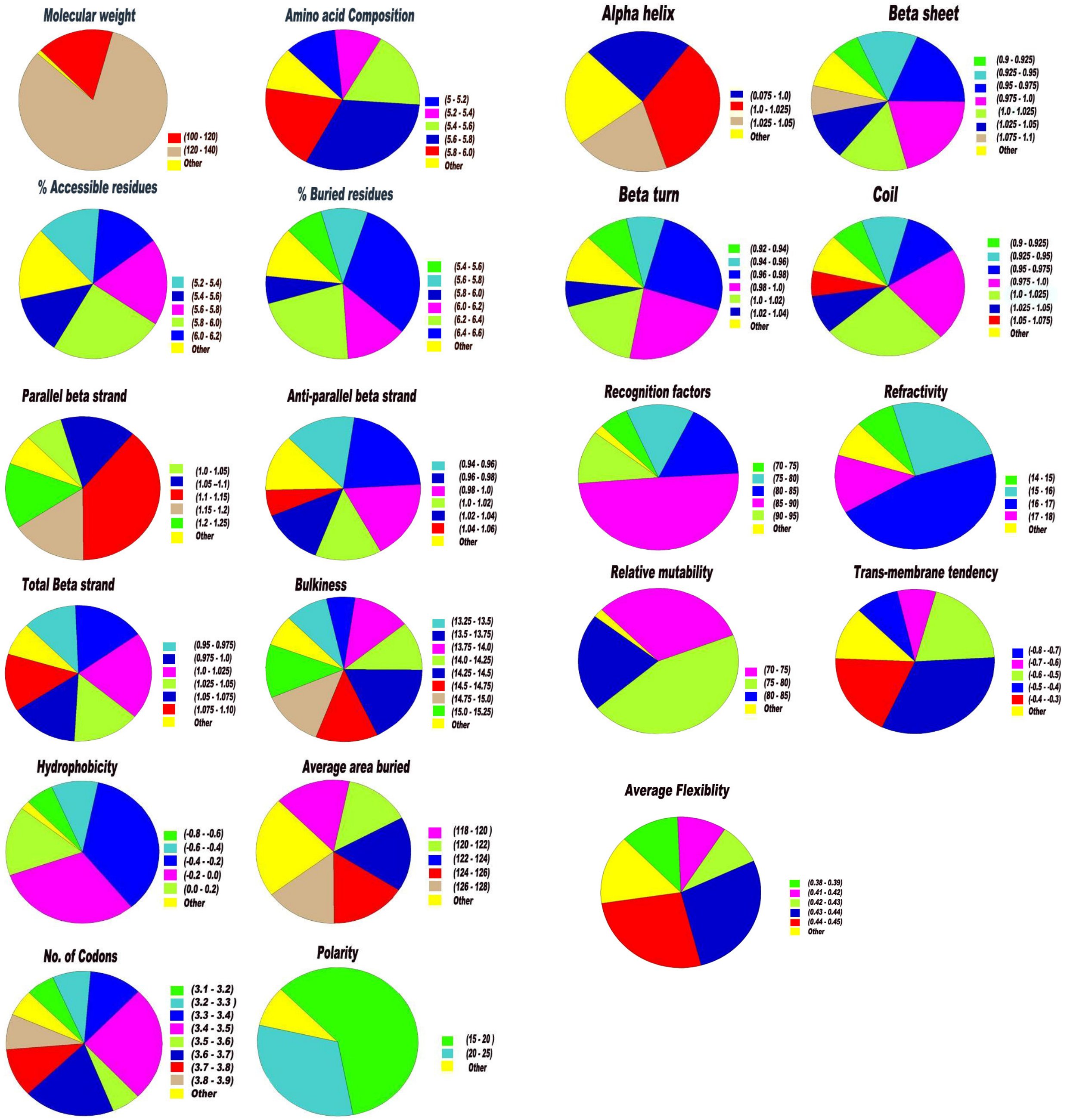

Figure 1. Frequency distribution chart for different parameters generated in CART.

Table 1. Basic statistical features of parameters considered for the study.

Parameters	Mean	Std Deviation	Skewness	Coeff Variation	Cond. Mean	Variance	Kurtosis	Std Error Mean
Accessible residues	5.7018	0.37534	-0.33589	0.065827	5.7018	0.14088	-0.014461	0.037347
Buried Residues	6.0371	0.3741	0.14695	0.061966	6.0371	0.13995	0.28716	0.037224
A.A composition	5.5674	0.32345	-0.87867	0.058097	5.5674	0.10462	0.88994	0.032184
Alpha helix	1.0014	0.042836	-1.3922	0.042775	1.0014	0.0018349	3.0152	0.0042623
Antiparallel Beta strand	0.98968	0.041571	0.21073	0.042004	0.98968	0.0017281	-0.11728	0.0041365
Average area buried	122.66	4.6318	-0.41521	0.037762	122.66	21.453	0.21574	0.46088
Average flexibility	0.42386	0.024286	-0.93356	0.057297	0.42386	0.0005898	-0.43531	0.0024165
Beta sheet	0.99057	0.050085	0.13276	0.050561	0.99057	0.0025085	-0.19328	0.0049836
Beta turn	0.98626	0.035704	0.17487	0.036201	0.98626	0.0012748	0.25416	0.0035527
Bulkiness	14.342	0.60675	-0.52613	0.042305	14.342	0.36814	-0.059278	0.060374
Coil	0.98701	0.050084	-0.39007	0.050744	0.98701	0.0025084	0.14176	0.0049836
Hydrophobicity	-0.21569	0.22909	-0.46668	-1.0621	-0.21569	0.052482	0.6858	0.022795
Molecular weight	127.58	14.506	7.2921	0.11371	127.58	210.43	64.252	1.4434
No. of codons	3.5407	0.21474	-0.079153	0.060648	3.5407	0.046112	-0.19733	0.021367
Parallel Beta strand	1.1224	0.076267	-1.4981	0.067949	1.1224	0.0058167	4.8368	0.0075889
Polarity	19.322	3.7942	4.0298	0.19637	19.322	14.396	29.632	0.37753
Recognition factors	84.57	6.0721	-1.2564	0.0718	84.57	36.871	1.3626	0.6042
Refractivity	16.074	1.0731	-0.78151	0.066759	16.074	1.1515	1.3996	0.10677
Relative Mutability	76.701	4.6985	-2.826	0.061257	76.701	22.076	17.774	0.46752
Total beta strands	1.0238	0.048899	-0.16652	0.047762	1.0238	0.0023912	-0.041638	0.0048657
Trans-membrane tendency	-0.51632	0.16864	-1.0046	-0.32662	-0.51632	0.028439	1.4962	0.01678

While CART will highlight the optimal tree based on the lowest cross-validated relative error, the overall goal was to obtain a tree which can yield maximum number of association rules. For the sake of simplicity, best regression tree should be with least number of nodes while for accuracy, best regression tree should have maximum possible number of nodes. 14 trees with different complexities and error values obtained using CART based on splitting criteria are reflected in Table 2. Out of these trees, tree with 20 terminal nodes (Figure 2) with minimum complexity and re-substitution relative error of 0.03218 and cross validated error of 0.34002 ± 0.09877 generated by Least Square splitting criteria was selected for generating decision rules. Decision rules obtained using CART are summarized in (Suppl. Table 1).

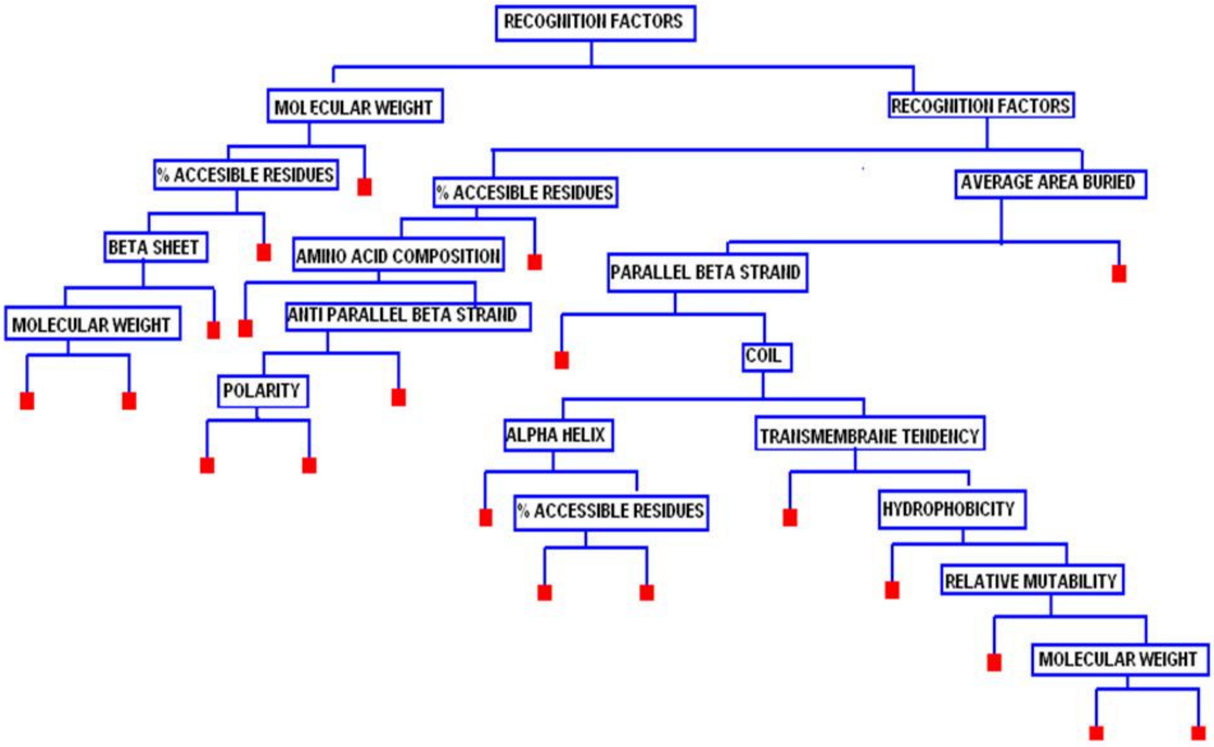

Figure 2. Splitters for the tree generated using CART.

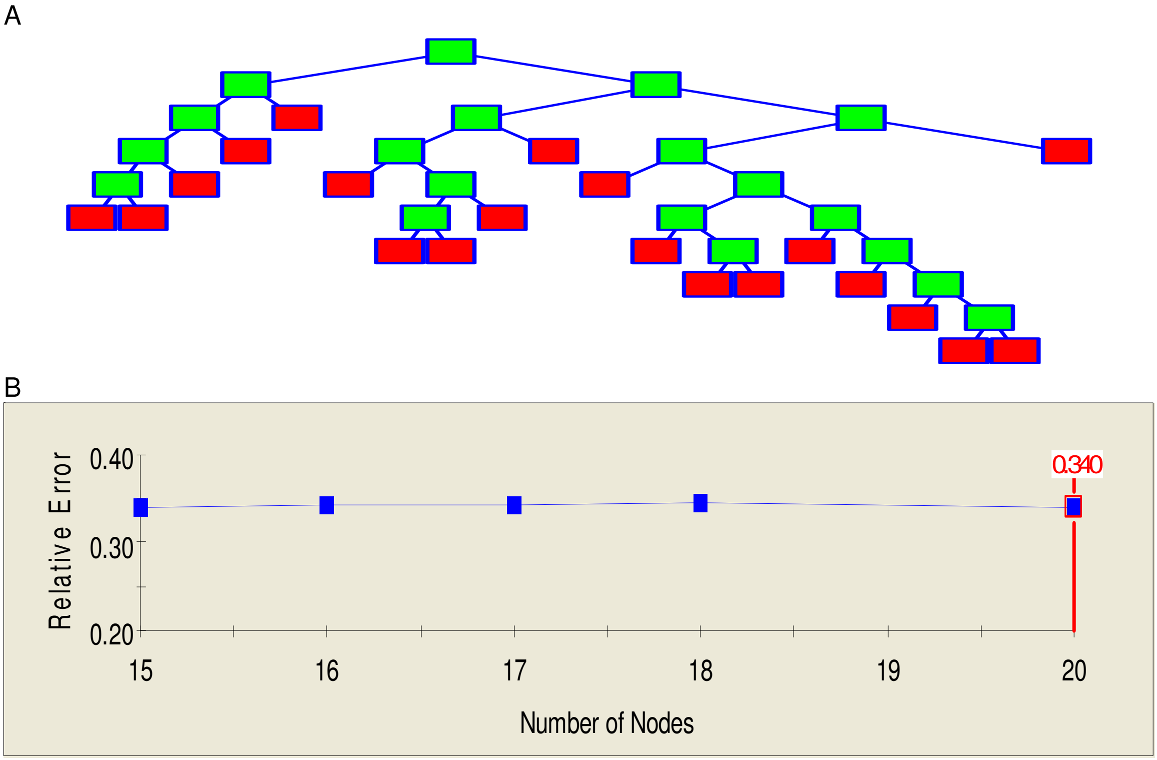

Figure 3. The tree sequence of lowest complexity which yielded 21 terminal nodes (A) with the cross validation error rate (B).

Table 2. Details of trees generated using CART along with relative error and complexities.

Tree No.	Terminal Nodes	Cross-Validated Relative Error	Re-substitution Relative Error	Complexity
1	20	0.34002 ± 0.09877	0.03218	0.00000
2	18	0.34556 ± 0.10138	0.03347	0.00004
3	17	0.34333 ± 0.10030	0.03612	0.00017
4	16	0.34401 ± 0.10030	0.03903	0.00018
5	15	0.33952 ± 0.09889	0.04220	0.00020
6	14	0.33967 ± 0.09835	0.04605	0.00024
7	13	0.32773 ± 0.09772	0.05078	0.00029
8	10	0.32363 ± 0.09534	0.06577	0.00030
9	9	0.31195 ± 0.08421	0.07331	0.00045
10	8	0.30844 ± 0.08397	0.08104	0.00047
11	7	0.29577 ± 0.08380	0.08993	0.00053
12	6	0.30413 ± 0.08550	0.10551	0.00093
13	5	0.29419 ± 0.08711	0.12323	0.00106
14	4	0.23277 ± 0.06945	0.14744	0.00144
15	3	0.32305 ± 0.07822	0.22063	0.00433
16	2	0.35356 ± 0.07602	0.32368	0.00609
17	1	1.00033 ± 0.00083	1.00000	0.03990

The tree selected for deriving decision rules is shown in Figure 3 along with error rate.

To calculate a variable importance score, CART looks at the improvement measure attributable to each variable in its role as a surrogate to the primary split. The values of these improvements are summed over each node and summed, and are scaled relative to the best performing variable. The variable with the highest sum of improvements is scored 100, and all other variables will have lower scores ranging downwards towards zero. Importance of different variables was calculated and summarized in Table 3.

Rules derived from CART can be interpreted in simple context of "If "and "Then" based statement and thus are self-explanatory.

For example: Rule 1 can be interpreted as:
Rule 1: IF "RECOGNITION FACTORS<= 81.5417" & "MOLECULAR WEIGHT<= 114.042" & "% ACCESSIBLE RESIDUES<= 5.497" & "BETA SHEET<= 1" THEN "AVERAGE FLEXIBILITY =0.374"

Rule 14 can be explained as:
Rule 14: IF "RECOGNITION FACTORS> 87.4445" & "% ACCESSIBLE RESIDUES > 5.8055" & "COIL<= 1" & "ALPHA HELIX>1" & "PARALLEL BETA SHEET> 1" &"AVERAGE AREA BURIED <= 129.268" THEN "AVERAGE FLEXIBILITY =0.444".

Table 3. Variable importance of parameters influencing average flexibility.

S. No.	Parameter	Importance
1	Recognition Factors	100.00
2	Amino acid composition	71.73
3	Molecular Weight	70.94
4	% Accessible_Residues	63.68
5	Coil	52.12
6	Antiparallel Beta Strand	43.29
7	Bulkiness	11.21
8	Alpha_Helix	8.86
9	Beta Sheet	5.90
10	% Buried Residues	3.10
11	Hydrophobicity	2.44
12	Refractivity	2.14
13	Polarity	1.61
14	Beta Turn	1.49
15	Total Beta Strand	1.48
16	Transmembrane tendency	1.42
17	Relative Mutability	1.31
18	Parallel Beta Strand	1.21
19	Average area buried	0.95
20	Number of Codons	0.85

4. Discussion

Many biological processes require change in conformations of proteins, thus, are influenced by the flexibility of the particular protein. This very property of proteins allows a spectrum of interactions between Enzyme-substrate/inhibitor in catalysis and hormone-receptor in biological systems. Thus, average flexibility, an inherent property of protein molecules is correlated with functions [22-26]. The discovery that some flexible proteins were found to have implications in pathological conditions has fuelled the studies relating to average flexibility of proteins. The complexity of such studies is often bewildering, given the enormous data available.

Data mining approaches based on decision tree based methods have been successfully exploited in elucidating importance of features affecting important biological processes [27]. Decision tree based methods are effective and simple means for sifting complex biological data for hidden explicit patterns and information. More and more biological studies are harnessing CART methodologies owing to its simplicity and ability to handle missing values. The CART methodology is being increasingly employed in biological studies like in ecological studies [28], diagnosis decision processes [29], epidemiology [30], microbiology [31], histology [32], genetics [33] and biochemical analysis [34].

CaMKK is known to control the activity of both CaMKI and CaMKIV. CaMK kinase, a part of CaMK cascade has been characterized in many organisms. Although various studies have focussed on kinetics of CaMK Kinases but the impact of various factors influencing their average flexibility is yet to be explored. Our analysis revealed that in CaMK kinases, recognition factors, amino acid composition, molecular weight, percent accessible residues, bulkiness, hydrophobicity, refractivity, polarity, transmembrane tendency, relative mutability, average area buried, numbers of codons among the sequence features were found to exert the influence on average flexibility in descending order. Among secondary structures, coil, anti parallel beta strand, alpha helix, beta sheet, beta turn, total beta strand, parallel beta strand were found to influence the average flexibility in decreasing order.

Keeping in mind, the recent enthusiasm for the inclusion of protein flexibility in docking algorithms, it will be interesting to gain an insight on features influencing the flexibility of proteins.

It is anticipated that an extensive knowledge of protein flexibility and the various parameters contributing towards is important for rational drug design. Such an approach will lead to better understanding of underlying biological phenomena and aid in enzyme engineering processes

Acknowledgements

Authors thank Dr. J.S. Yadav, Director, IICT for his continuous support and encouragement. Authors are thankful to anonymous reviewers for their critical suggestions for the improvement of manuscript.

References

[1] Anderson, B.F., Baker H.M., Morris, G.E., et al. (1990) Apolactoferrin structure demonstrates ligand-induced conformational change in transferrins. *Nature*, **344**: 784–787.

[2] Spurlino, J.C., Lu, G.Y., Quiocho, F.A.(1991) The 2.3-Å resolution structure of the maltose- or maltodextrin-binding protein, a primary receptor of bacterial active transport and chemotaxis. *J. Biol. Chem*, **266**: 5202–5219.

[3] Bennett, W.S., Jr, Steitz, T.A. (1978) Glucose-induced conformational change in yeast hexokinase. *Proc. Natl Acad. Sci. USA*, **75**: 4848–4852.

[4] Remington, S., Wiegand, G., Huber, R. (1982) Crystallographic refinement and atomic models of two different forms of citrate synthase at 2.7 and 1.7 Å resolution. *J. Mol. Biol*, **158**: 111–152.

[5] Perutz, M.F., (1970) Stereochemistry of cooperative effects in haemoglobin. *Nature*, 228: 726–739.

[6] Perutz, M.F. (1989) Mechanisms of cooperativity and allosteric regulation in proteins. *Q. Rev. Biophys*, **22**: 139-237.

[7] Chan, D.C, Fass, D. Berger J.M. et al. (1997) Core structure of gp41 from the HIV envelope glycoprotein. *Cell*, **89**: 263-273.

[8] Teague, S. J. (2003) Implications of protein flexibility for drug discovery. *Nat Rev Drug Discov*, **2**: 527-41.

[9] Zimmerman, JM, Naomi Eliezer, Simha R.(1968) The characterization of amino acid sequences in proteins by statistical methods. *Journal of Theoretical Biology*. **21**(2): 170-201.

[10] Jones DD, (1975) Amino acid properties and side-chain orientation in proteins: a cross correlation approach. *J Theor Biol*. **50**(1): 167-83.

[11] Fraga, S. (1982) Theoretical prediction of protein. antigenic determinants from amino acid sequences. *Can. J. Chem*. **60**: 2606-2610.

[12] Kyte, J., Doolittle, RF.(1982) A simple method for displaying the hydrophobic character of a protein, *J. Mol. Biol*. **157**: 105-132.

[13] Zhao, G, London E. (2006) An amino acid "transmembrane tendency" scale that approaches the theoretical limit to accuracy for prediction of transmembrane helices: Relationship to biological hydrophobicity. *Protein Sci*. 15: 1987-2001.

[14] Joël Janin, (1979) Surface and inside volumes in globular proteins. *Nature*. **277**: 491-492.

[15] Bhaskaran, R, Ponnuswamy.PK. (1988) Positional flexibilities of amino. acid residues in globular proteins. *Int. J. Pept. Prot. Res*., **32**: 242-255.

[16] Chou PY, Fasman GD, (1978) Prediction of the secondary structure of proteins from their amino acid sequence. *Adv Enzymol Relat Areas Mol Biol*. **47**: 45-148.

[17] Deléage and Roux,(1987) An algorithm for protein secondary structure prediction based on class prediction. *Protein Engineering, Design and Selection*. **1**: 289-294.

[18] Lifson, S., Sander C., (1979) Antiparallel and parallel-strands differ in amino acid residue preferences. *Nature* **282**: 109-111.

[19] McCaldon, P, Argo, P (1988) Oligopeptide biases in protein sequences and their use in predicting protein coding regions in nucleotide sequences. *Proteins: Structure, Function, and Genetics*. **4**: 99-122.

[20] Dayhoff, MO, Schwartz, RM., Orcutt BC (1978) A model of evolutionary change in protein; in: M.O. Dayhoff (Ed.), *Atlas of Protein Sequence and Structure*, Nat. Biomed. Res. Foundation, Washington, DC. 5, **Suppl. 3**: 345–352

[21] Briemann L., Friedman, J.H., Olshen, R.A., Stone C.J. (1984) *Classification and regression trees*. Chapman & Hall, New York, NY.

[22] Wright, P.E., Dyson, H. J., (1999) Intrinsically Unstructured Proteins: Re-assessing the Protein Structure-Function Paradigm. *J. Mol. Biol*., **293**: 321–331.

[23] Bright, J.N., Woolf, T.B.,.Hoh, J. H, (2001) Predicting properties of intrinsically unstructured proteins. *Prog. Biophys. Mol. Biol*., **76**: 131-173.

[24] Dunker, A.K., Lawson D.J., Brown, C. J. et al. (2001) Intrinsically disordered protein. *J. Mol. Graph. Model*, **19**: 26-59.

[25] Namba, K. (2001) Roles of partially unfolded conformations in macromolecular self-assembly. *Gene Cells*, **6**: 1-12.

[26] Garbers, A., Reifarth, F., Kurreck, J., Renger, G. & Parak, F., (1998) Correlation between protein flexibility and electron transfer from QA to QB in PSII membrane fragments from spinach. Biochemistry, **37**: 11399-11404.

[27] Banerjee, A.K., N. Arora and U. S.N. Murty, (2007) Stability of ITS2 Secondary Structure in Anopheles: What Lies Beneath? *International Journal of Integrative Biology*, **3**:232-238.

[28]De'ath. G., Fabricius, K.E., (2000) Classification and regression trees: a powerful yet simple technique for ecological data analysis, *Ecology*. **81**: 3178-3192.

[29] Guzick, D.S., Overstreet, J.W., Factor-Litvak, P. et al. (2001) Sperm morphology, motility, and concentration in fertile and infertile men. *N. Engl. J. Med*. **345**: 1388-1393.

[30] Lemon, S.C., Roy, J., Clark, M.A., et al. (2003) Classification and Regression Tree Analysis in Public Health: Methodological Review and Comparison with Logistic Regression. *Annals of Behavioral Medicine*. **26**(3): 172-181.

[31] Smolle, J., Kahofer, P. (2001) Automated detection of connective tissue by tissue counter analysis and classification and regression trees. *Anal Cell Pathol*. **23** (3-4): 153-8.

[32] Ambrose, P.G., Dennis M. Grasela, Thaddeus H. Grasela et al. (2001) Pharmacodynamics of Fluoroquinolones against Streptococcus pneumoniae in Patients with Community-Acquired Respiratory Tract Infections. *Antimicrobial Agents and Chemotherapy*. **45**(10): 2793-2797.

[33] Davuluri, V., Suzuki, Y., Sugano, S., et al. (2000) CART classification of human 5′ UTR sequences, Genome Res. **10**: 1807-1816.

[34] Bai, J.P., Utis, A.., Crippen, G. et al.(2003) A genome-wide scan using tree-based association analysis for candidate loci related to fasting plasma glucose levels. *BMC, Genetics* **4**(Suppl. 1), S65.

Suppl. Table 1. Decision rules derived using CART.

Node	Recognition Factors	A.A.	M. weight	% Accessible residues	Coil	Anti parallel Beta strand	Alpha helix	Beta sheet	Hydro phobicity	Polarity	Trans-membrane tendency	Relative mutability	Parallel beta sheet	Average area buried	Average Flexibility
1	<= 81.5417		<= 114.042	<= 5.497				<= 1							0.374
2	<= 81.5417		> 114.042 & <= 124.306	<= 5.497				<= 1							0.384813
3	<= 81.5417		<= 124.306	<= 5.497				> 1							0.3711
4	<= 81.5417		<= 124.306	> 5.497											0.394417
5	<= 81.5417		> 124.306												0.417286
6	> 81.5417 & <= 87.4445	<= 5.33325		<= 6.222											0.409
7	> 81.5417 & <= 87.4445	> 5.33325		<= 6.222		<= 1				<= 20.1165					0.427875
8	> 81.5417 & <= 87.4445	> 5.33325		<= 6.222		<= 1				> 20.1165					0.417667
9	> 81.5417 & <= 87.4445	> 5.33325		<= 6.222		> 1									0.44
10	> 81.5417 & <= 87.4445			> 6.222											0.386
11	> 87.4445												<= 1	<= 129.266	0.433875
12	> 87.4445				<= 1								> 1	<= 129.266	0.4327
13	> 87.4445			<= 5.8055	<= 1		> 1						> 1	<= 129.267	0.4367
14	> 87.4445			> 5.8055	<= 1		> 1						> 1	<= 129.268	0.444
15	> 87.4445				> 1						<= -0.58475	<= 72.0835	> 1	<= 129.269	0.439889
16	> 87.4445				> 1				<= -0.3225		> -0.58475		> 1	<= 129.266	0.4525
17	> 87.4445				> 1				> -0.3225		> -0.58475		> 1	<= 129.266	0.438
18	> 87.4445		<= 130.278		> 1				> -0.3225		> -0.58475	> 72.0835	> 1	<= 129.266	0.4465
19	> 87.4445		> 130.278		> 1				> -0.3225		> -0.58475	> 72.0835	> 1	<= 129.266	0.44325
20	> 87.4445													> 129.266	0.432333