e-PROPAINOR: A Web-Server for Fast Prediction of C_α Structure & Likely Functional Sites of a Protein Sequence

Rajani R. Joshi* and N.T. Jyothish

Department of Mathematics, Indian Institute of Technology Bombay, India

Abstract: *e-PROPAINOR* (www.math.iitb.ac.in/epropainor/) is a web-server based on extension of *PROPAINOR* for prediction and computational function elucidation of 3-D structure of proteins. It predicts the C_α structure of a given protein sequence. Computational efficiency and reliability are key features of its software. Moreover, it also gives an estimate of the RMSD of the predicted structure. For the structures predicted with estimated RMSD of the order ≤ 5Å, it predicts likely sites of five major types of protein functions.

Keywords: Protein Structure prediction, Nonparametric regression, Nonparametric Discriminant analysis, Distance geometry.

1. INTRODUCTION

Determination of protein structure and function is important in biomedical sciences, and biotechnology. With the advancement of experimental and computational research this has motivated the development of several prominent databanks, web-servers and related Bioinformatics utilities since past few decades.

Many important features of proteins are hidden in their complicated sequences. Therefore, sequence-based prediction methods, such as protein structural class prediction [1,2], tight turn prediction [3, 4], protein quaternary attribute prediction [5, 6], protein folding rate prediction [7, 8] are highly desired because they can timely provide very useful information for both basic and applied research. Towards applied research — especially or relevance in computer aided drug development, sequence based approaches have been successfully deployed to — pKa value prediction in protein [9], HIV protease cleavage site prediction [10-12], signal peptide prediction [13], protein subcellular location prediction [14, 15], identification of enzymes and their functional classes [16], identification of GPCR and their types [17-19], identification of proteases and their types [20], and protein 3D structure prediction based on sequence similarity [21], as well as a series of user-friendly web-servers for predicting various attributes of proteins as recently summarized in Table **3** of [22], and drug development.

In this study, we report a user-friendly web-server developed in our lab for predicting the C_α structure of a protein and its likely functional sites according to its sequence information in hopes that it may become a useful tool for drug design and protein science research.

Large numbers of computational methods of prediction of secondary and tertiary structure of proteins are based on and homology modeling using sequence alignment and or molecular dynamics simulation. The *ab-initio* approaches attempt this without homology modeling. Promising potentials of research in genomics and proteiomics have boosted newer interest protein structural genomics and hence enhanced the significance of *ab-initio* prediction of protein tertiary structure [23-30]. Our recently developed algorithm *PROPAINOR*: PROtein structure Prediction using AI and NOnparametric Regression also contributes in this regard [31-37].

A comprehensive comparative review of different algorithms and servers for *ab-initio* prediction of protein tertiary structures developed since nearly a decade is presented in [38]. Distinct features of *PROPAINOR* are also highlighted there along with the methods of ROSETTA [28] and I-TASSER [30] that are known as best servers so far. Good accuracy — comparable with the best methods and significantly fast computations of *PROPAINOR* made a good case for its web-implementation.

The *PROPAINOR* algorithm makes use of Knowledge-based *Nonparametric Regression modeling* (NPR), *Multivariate Analysis of Variance* (MANOVA) and *Nonparametric Discriminant Analysis*. It solves the computational problem of protein 3D- structure prediction as a probabilistic programming problem based on estimators of inter-residue distances at C_α positions [31, 32].

For short and medium sized proteins (sequence length 70-150 amino acids) this algorithm is found equivalent or better in terms of prediction accuracy as compared to existing best *ab-initio* computational methods. Apart from the non-requirement of sequence-homology, the modularity and computational efficiency of its algorithm, and estimation of *reliability index* of the predicted structure are some significant features of *PROPAINOR* [33].

Successful use of *PROPAINOR* on new biotechnologically and pharmaceutically important proteins like *Human Seminal Plasma Prosthetic Inhibin* [34, 35] and a two domain EF-Hand Calcium Binding Protein from *Entamoeba Histolytica* [36, 37] has motivated us to extend it for longer proteins and provide the utility on the Internet for wider research applications.

*Address correspondence to these authors at the Department of Mathematics, IIT Bombay, Powai, Mumbai: 400076 India; Tel: +91-22-25767485; E-mail: rrj@math.iitb.ac.in

1875-0362/10

As a first step towards its extension for multi-domain proteins we have developed Bayesian methods for prediction of domain boundary points [39, 40]. We have also incorporated some modifications in the method of inter-residue distances for protein sequences of length > 150 amino acids. (Throughout this paper the word residue would imply C_α atom of an amino acid of the protein sequence under consideration).

e-PROPAINOR (www.math.iitb.ac.in/epropainor/) is a web-server based on extension of the above approach. It also incorporates our novel contribution towards prediction of functional sites of a protein structure. In this paper we highlight its methodology and salient features – including performance evaluation and discuss its scope and importance.

2. MATERIALS AND METHODS

The root-software for *e-PROPAINOR* incorporates integration of five major sets of modules consisting of interlinked computer programs written in C^{++}, Perl and Shell scripts. The broad architect of its core software has six interconnected layers of modules; execution of modules in one layer triggers the execution of modules in next layer and so on (see Fig. **1** for illustration). The first level (Layer1) deals with reading the input sequence (in fastA format), predicting its secondary structure using standalone version of PSIPRED [41] and if necessary, identifying the likely structural domain boundary points (*dbp*). It may be noted here that the PSIPRED predicted secondary information is used only if its reliability measure is ≥ 6 for the consecutive segment (in the sliding window) of five amino acids in the sequence. Moreover this is used only for *dbp* prediction [39, 40] and/or at Layer 4 for some heuristics/estimation of certain medium-range pair-wise distances that could not be estimated by the statistical procedure with confidence level above the average of other distances of similar category.

Layer 2 is most crucial as it pertains to estimation of expected inter-C_α residue distances of all pairs of amino-acids on which a probabilistic version of distance-geometry approach is applied to get the C_α coordinates. This layer has ten modules for estimation of sequence features, estimation of distances, estimation of likelihood of contacts, local folds, etc.

Statistical modeling and data mining used here emanates from our idea of considering the inter-residue distance (D_{ij}) between C_α atoms of amino acids at positions i and j as a random variable; $i \neq j$; $i, j = 1, \ldots, L$; L = length (total no. of amino acids) of the protein sequence. In view of Nature's random effects and theoretical possibility that a protein sequence can fold into any 3-D form, this consideration amounts to, as also remarked in [30], most general modeling in the landscape of inter-residue distances. We estimate expected lower and upper bounds on these unknown distances using its nonparametric regression on statistically significant features of the protein sequence.

Additive model of nonparametric regression [42] is used for this purpose. Nineteen features of the protein sequence are found important in the case of long-range distances (D_{ij} for the pairs with $| i - j | \geq 20$). The features include both the sequential and individual amino acid properties [43-45] including — length of the sequence, relative frequencies in the similarity clusters of amino-acids in the patch i+5 to j-5, proportion of sliding window segments of size 5 that have average alpha propensity > 60%, percentage concentration of

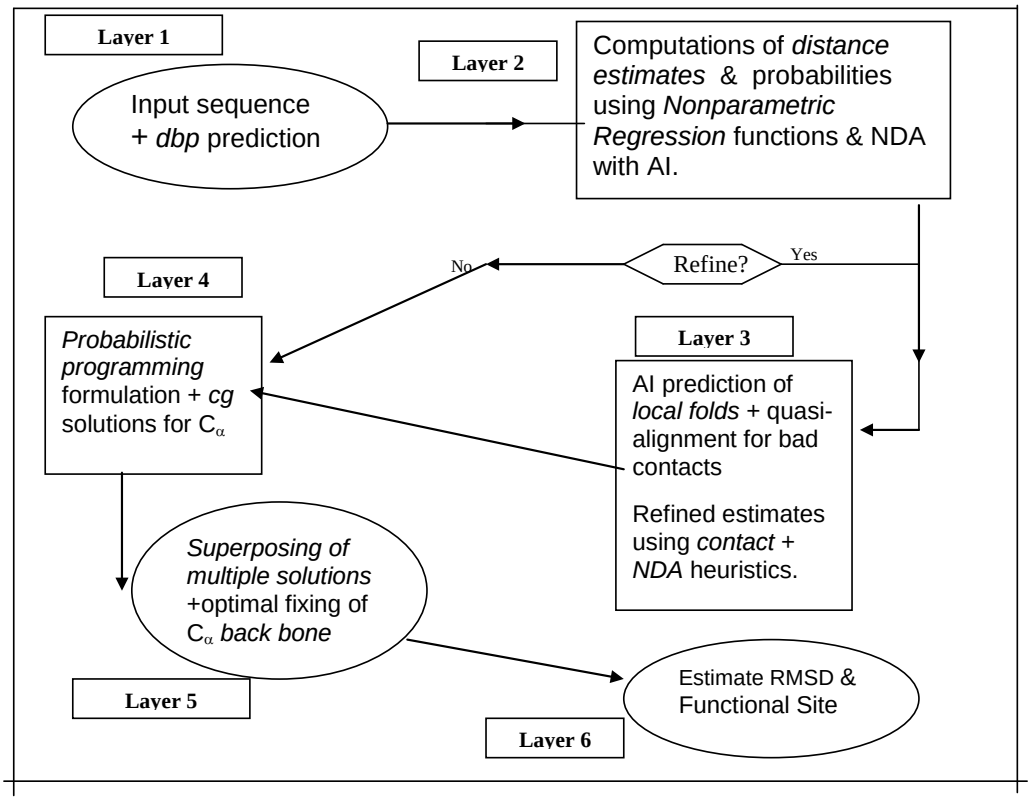

Fig. (1). Schema of Software-Architect of *e-PROPAINOR*.

hydrophobic residues in the specific segments, etc. Only seven of the nineteen features are found sufficient in the case of short and medium-range distance estimates (i. e., D_{ij} for the pairs with $|i-j| \leq 3$ and $4 \leq |i-j| \leq 7$).

Multivariate Nonparametric Discriminant Analysis (NDA) of the pairs in short-, medium- and long-range distances is then carried out to predict the categories (e.g. short contacts, *bump*, hydrophobic core, long-range contact, etc) of the pair of residues and hence of the distance-types. Accordingly the estimates of lower and upper bounds on pair-wise distances are updated [33].

The quantities used in the objective function f (given by equation (1) below) of the probabilistic programming problem, e.g. the bump-distance l_{bump}, the *weights*, ω_t for t^{th} category of distance-types, the probabilities p_{ij}^{*} (= $\Pr\{l_{\text{ij}} \leq D_{\text{ij}} \leq u_{\text{ij}}\}$), etc are also computed by the modules at this layer using the training sample estimates and geometric and probabilistic modeling based heuristics [33]. Unless further refinement is required the Layer 4 is not activated and these estimates are supplied directly to Layer 5.

For the distances that could not be estimated with above-average confidence level due to bad fitting of the *Nonparametric Regression* (NPR) or discrepancies in NDA classification, or predicated type of secondary fold, etc, refined estimates are computed at Layer 4. Heuristics based on PSIPRED-predicted secondary folds [33] and *quasi*-alignment of the segments containing the concerned residues — using stand alone version of BLAST search [46] — are applied for this purpose. Only the prediction made with reliability level ≥ 6 in the output of PSIPRED are used. The term *quasi*-alignment implies approximate and selective alignment: the entire sequence is aligned but the contact information of the aligned portion of the template sequence is used only for segments of interest. If no alignment or no substantial inter-residue contact information is available for some pairs of amino acids, the corresponding distance estimates are not supplied to Layer 5. The parameters (*weights* etc) are recomputed for the refined estimates and all are sent to the next layer. If the total number of inter-residue distance estimates acceptable from Layer 3 and Layer 4 is less than 70% of the desired $\binom{N}{2}$ distances then the system also sends a warning signal to the next layer. This remark on non-reliability of the predicted structure is displayed along with the output, if any.

Layer 5 uses all the refined estimates received from Layer 4 and the acceptable ones from Layer 3 in the following objective function.

$$f = \sum_t \omega_t \sum_{(i,j)\varepsilon\Gamma_t} \Big(p_{ij} * \min^2\{\frac{\|\underline{x}_i - \underline{x}_j\|^2 - l_{ij}^{\,2}}{l_{ij}^{\,2}}, 0\} + p_{ij} * \max^2\{\frac{\|\underline{x}_i - \underline{x}_j\|^2 - u_{ij}^{\,2}}{u_{ij}^{\,2}}, 0\}\Big) + \omega_{bump} * l * \sum_{i=1}^{N-4}\sum_{j=i+1}^{N} \min^2\{\frac{\|\underline{x}_i - \underline{x}_j\|^2 - l_{bump}^{\,2}}{l_{bump}^{\,2}}, 0\} \quad (1)$$

This function is minimized with respect to $\underline{X}_i$'s subject to the triangular inequality constraints using conjugate gradient (CG) method:

$\|\underline{X}_i - \underline{X}_j\| \leq \|\underline{X}_i - \underline{X}_m\| + \|\underline{X}_m - \underline{X}_j\| \;\forall\; i \neq j \neq m$; $i, j, m \in \{1, 2, \ldots, L\}$;

If no optimal solution is obtained, the system does not predict any solution. Else, from among the optimal solutions, the redundant ones are filtered out using our superimposition program at Layer 6. In case there are more than 5 distinct solutions, the best 5 are chosen in terms of the optimal value of f.

Finally, expected range of RMSD is predicted using PERT/CPM approach [33] for each predicted solution, if any.

2.1. Functional Site Predication

If the expected RMSD is ≤5Å then likely functional sites are also predicted at this Layer. The module for functional site prediction deploys our recent algorithm based on logistic regression modeling [47, 48].

A logistic regression model is a statistical model, which estimates the probability of a categorical response variable (say Y) for a given vector (say $\underline{X}$) of regressor variables using a logit function of the latter. We have fitted 5 logistic regression models — one each for major classes of protein functions, namely – Translation Regulation Activity, Transporter Activity, Antioxident Activity, Transcription Regulation Activity, and Enzyme Regulation Activity.

In each model Y has two categories: $Y = 1$ and $Y = 0$ implying respectively "Functional site" and "Not a functional site". A site is predicted as a likely functional site if the estimated Pr (Y=1) is greater than a threshold.

The regressor variables in each model include — structural properties like closeness and relative surface area obtained from the SARIG web-server [49], and some biophysical properties of amino acids.

Necessary details of the models and estimated parameters with RoC — curves of performance evaluation are reported in a separate paper [48].

2.2. User Interaction

The user has to first register on the server site (www.math.iitb.ac.in/epropainor/). This procedure is simple and interactive. Upon successful registration he/she may upload the input sequence online at the server site. Necessary user-guideline is also available on the site (click the link "Help" on the top bar after clicking "run ePropainor). The server (a Fedora6 station on a Pentium-IV dual core PC) automatically picks up the job on first come-first served basis. The text files containing predicted solutions in PDB format and the predicted RMSD information are sent to the user *via* email soon after completion of the job. The output files containing results, if any, on predicted functional sites are also supplied with these. In the mean time, the user may check the job status – e.g. position in the queue – on the server site.

3. PERFORMANCE EVALUATION

For test runs we extracted a set of non-redundant (i.e. belonging to different structural and functional families) proteins of length greater than 40 and less than or equal to 500 amino acids from the Protein Data Bank

(www.rcsb.org/pdb). From these we chose the ones having pair-wise sequence homology ≤ 30%. The final set had nearly 4800 protein sequences, the crystallographic or NMR structures of which are also available in the PDB.

In statistical prediction, the following three cross-validation methods are often used to examine a predictor for its effectiveness in practical application: independent dataset test, subsampling test, and jackknife test [50]. However, as elucidated in [14] and demonstrated by Eq.50 of [15], among the three cross-validation methods, the jackknife test is deemed the most objective that can always yield a unique result for a given benchmark dataset, and hence has been increasingly used by investigators to examine the accuracy of various predictors (see, e.g., (see, e.g., [51-60])

As part of Jackknife approach of cross-validation (e.g. in [61-64]), we have used random subsets of about 1000 of these as training and remaining as validation samples. On an average for 70% of those in validation sample the actual RMSD (between the predicted and the PDB C_α structure) is found to lie between 4.6Å ± 3.5 Å; only for about 12% of the validation candidates the actual RMSD is found to lie between 11Å to 20Å on an average. The performance for the 7.4% in the remaining is in-between the above ranges and no feasible/optimal solution is found for the rest.

In most cases the interval of predicted RMSD is found to contain the actual RMSD. Testing on CASP8 benchmark entries of length between 41 and 500 has shown the predictions close to the top-ranked solution in 4.5% cases; no feasible/optimal solution in 6.1% cases. For most of the others it has shown above-average performance in terms of RMSD as compared to the top ranking solutions.

The prediction of functional sites is so far tested on about 35 proteins from each functional class under consideration. The average *sensitivity* and *specificity* of this prediction are about 68 to 92% and 61.8 to 80.1% respectively.

In all test runs the average CPU time taken per solution (of C_α coordinates) for proteins of length <150 is about 3-4 minutes. That for proteins of length 300 to 500 is about 15-28 minutes. For other proteins, the average computing time is found to lie in-between 5 to12 minutes.

4. DISCUSSION

Computational methods for determining/predicting protein tertiary structure are crucial in Proteiomics research-and-development in the absence of confirmed experimental details. These are also frequently used to complement or refine the experimental findings and to test the flexibility and sensitivity of different structural parameters.

Data-driven (probability distribution-free) statistical mining approaches are of special interest in this context. These also have potential to complement the homology based methods and supervised machine-learning techniques for better understanding of structural genomics and greater applications of Bioinformatics and Computational Biology in full exploration and exploitation of the available databanks. Our approach in *PROPAINOR* contributes in this regard with promising scope. Its extension and web-implementation (*e-PROPAINOR*) offers wider utilization and possibilities in Bioinformatics applications.

Fast prediction with fairly good accuracy is most significant feature of this web- server. Performance evaluation of the core algorithm *PROPAINOR* as reported in our earlier papers [31, 33] shows its superiority in computation time — including the time of atomic structure prediction by Max-Sprout [65] — over other comparable *ab-initio* threading methods. In terms of *RMSD* of predicted structure too, while the average over different validation samples is comparable with the best-known methods, its shows greater consistency of performance as the standard deviations of *RMSD* in these samples are lowest in case of *PROPAINOR*. These strengths are carried in the performance of *e-PROPAINOR* as well.

Because of high diversity in the protein structures with respect to long-range contacts, the statistical estimates associated with these are often predicted with lower confidence level. Specific heuristics of globular geometry (e.g. compact folding of hydrophobic core [33]), beta strand distances, etc, are therefore used in different discriminant classes. The methods based on multiple sequence alignment do not have such constraints of approximation. However, substantial alignment in most parts of the sequence is essential under these methods. Though, indirectly, these methods too use some heuristics such as — homology of structure implied by homology of sequences. Whenever this assumption is not satisfied these methods make drastic errors in structure prediction. Methods (as in servers like I-TASSER [30]) that use structural motif libraries and incorporate several options of sub-structures are generally found better. We are currently working on estimation of probability distribution of structural motifs to modify the heuristics, wherever relevant in the core-method of *e-PROPAINOR*.

A unique feature of *e-PROPAINOR*, which is most desired in the case of prediction of the structure of a newly determined protein sequence, and which should be incorporated in other structure prediction servers too, is that an estimated interval of likely RMSD is provided with each predicted solution. It is also a distinct facet of this server that rather than predicting a wrong or totally random solution without any warning, it either does not predict any or puts a remark on non-reliability of the predicted structure as the case may be.

The solutions are provided in standard PDB format, which could be used for all kinds of further refinement and/or analysis of the properties/functions of the corresponding protein/its translating gene. (The side-chains and other atomic co-ordinates could be attached to the predicted C_α backbone using high-reliability methods/programs like MaxSprout [65].)

The core software of *e-PROPAINOR* is modular for compatible linkage options with gene databanks, NMR (NoE distances) data and chemical activity etc and related softwares. At present we use the link only with SARIG program (http://bioinfo2.weizmann.ac.il/~pietro/SARIG/V3/index.html) for computation of closeness and relative surface accessibility of the residues in predicted structure. Using these features and biophysical properties of amino acids the logistic regression module in the last layer of *e-PROPAINOR* software predicts possible functional sites on the structures that are predicted with substantial accuracy. Currently only five major classes of protein functions are considered. Extension and improvement in this utility is under progress.

Other utilities for detection of specific function class [63, 64] and activity pockets along with possible genome characterization will be added successively with feasible linkage with relevant databanks and web-software.

ACKNOWLEDGEMENTS

This work is part of a research project undertaken by the first author. The author is thankful to the Department of Biotechnology, Govt. of India for the financial grant of this project. The authors would also like to thank Ms. Archana Prabahar, a programmer under this project for her help in data extraction and assistance in testing of several modules.

REFERENCES

[1] K.C. Chou, and C.T. Zhang, "Predicting protein folding types by distance functions that make allowances for amino acid interactions", *J. Biol. Chem*., vol. 269, pp. 22014-22020, 1994.

[2] K.C. Chou, "A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space", *Proteins Struct. Funct. Genet.*, vol. 21, pp. 319-344, 1995.

[3] K.C. Chou, "Prediction and classification of alpha-turntypes", *Biopolymers*, vol. 42, pp. 837-853, 1997.

[4] K.C. Chou, "Review: Prediction of tight turns and their types in proteins", *Anal. Biochem*., vol. 286, pp. 1-16, 2000.

[5] H.B. Shen, and K.C. Chou, "QuatIdent: A web server for identifying protein quaternary structural attribute by fusing functional domain and sequential evolution information", *J. Proteome Res.*, vol. 8, pp. 1577; V1584, 2009.

[6] X. Xiao, P. Wang, and K.C. Chou, "Predicting protein quaternary structural attribute by hybridizing functional domain composition and pseudo amino acid composition", *J. Appl. Crystallogr.*, vol. 42, pp. 169-173, 2009.

[7] K.C. Chou, and H.B. Shen, "FoldRate: A web-server for predicting protein folding rates from primary sequence", *Open Bioinform. J.*, vol. 3, pp. 31-50 (openly accessible at http://www.bentham.org/open/tobioij/), 2009.

[8] H.B. Shen, J.N. Song, and K.C. Chou, "Prediction of protein folding rates from primary sequence by fusing multiple sequential features", Journal of Biomedical Science and Engineering (JBiSE), Vol. 2, pp. 136-143 (openly accessible at http://www.srpublishing.org/journal/jbise/), 2009.

[9] R.B. Huang, Q.S. Du, C.H. Wang, S.M. Liao, and K.C. Chou, "A fast and accurate method for predicting pKa of residues in proteins", *Protein Eng. Des. Sel.*, vol. 23, pp. 35-42, 2010.

[10] K.C. Chou, "A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins", *J. Biol. Chem.*, vol. 268, pp. 16938-16948, 1993.

[11] K.C. Chou, "Review: Prediction of HIV protease cleavage sites in proteins", *Anal. Biochem.*, vol. 233, pp. 1-14, 1996.

[12] H.B. Shen, and K.C. Chou, "HIVcleave: a web-server for predicting HIV protease cleavage sites in proteins", *Anal. Biochem*., vol. 375, pp. 388-390, 2008.

[13] K.C. Chou, and H.B. Shen, "Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides", *Biochem. Biophys. Res. Comm.*, vol. 357, pp. 633-640, 2007.

[14] K.C. Chou, and H.B. Shen, "Cell-PLoc: A package of web-servers for predicting subcellular localization of proteins in various organisms", *Nat Protoc*., vol. 3, pp. 153-162, 2008.

[15] K.C. Chou, and H.B. Shen, "Review: Recent progresses in protein subcellular location prediction", *Anal. Biochem*., vol. 370, pp. 1-16, 2007.

[16] H.B. Shen, and K.C. Chou, "EzyPred: A top-down approach for predicting enzyme functional classes and subclasses", *Biochem. Biophys. Res. Commun*., vol. 364, pp. 53-59, 2007.

[17] K.C. Chou, "Prediction of G-protein-coupled receptor classes", *J. Proteome. Res.*, vol. 4, pp. 1413-1418, 2005.

[18] X. Xiao, P. Wang, and K.C. Chou, "GPCR-CA: A cellular automaton image approach for predicting G-protein-coupled receptor functional classes", *J. Comput. Chem*., vol. 30, pp. 1414-1423, 2009.

[19] W.Z. Lin, X. Xiao, and K.C. Chou, "GPCR-GIA: a web-server for identifying G-protein coupled receptors and their families with grey incidence analysis", *Protein. Eng. Des. Sel.*, vol. 22, pp. 699-705, 2009.

[20] K.C. Chou, and H.B. Shen, "ProtIdent: A web server for identifying proteases and their types by fusing functional domain and sequential evolution information", *Biochem. Biophys. Res. Comm.*, vol. 376, pp. 321-325, 2008.

[21] K.C. Chou, "Review: Structural bioinformatics and its impact to biomedical science", *Curr. Med. Chem.*, vol. 11, pp. 2105-2134, 2004.

[22] K.C. Chou, and H.B. Shen, "Review: recent advances in developing web-servers for predicting protein attributes", Natural Science, vol. 2, pp. 63-92 (openly accessible at http://www.scirp.org/journal/NS/), 2009.

[23] A. Aszodi, M.J. Gradwell, and W.R. Taylor, "Global fold determination from a small number of distance restraints", *J. Mol. Biol.*, vol. 251, pp. 308-326, 1995.

[24] O. A. Kolinski, and J. Skolnick, "Fold assembly of small proteins using Monte-Carlo simulations driven by restraints derived from multiple sequence alignments", *J. Mol. Biol*., vol. 277, pp. 419-448, 1998.

[25] E. Huang, R. Samudrala, and J.W. Ponder, "*Ab-initio* fold prediction of small helical proteins using distance geometry and knowledge-based scoring functions", *J. Mol. Biol.*, vol. 290, pp. 267-281, 1999.

[26] H. E.S. Xia, M. and Levitt, R. Samudrala, "*Ab-initio* construction of protein tertiary structures using a hierarchical approach. *J. Mol. Biol., vol.* 300, pp. 171-185, 2000.

[27] P. Bradley, K.M.S. Misura, and D. Baker, "Towards High-Resolution *de Novo* Structure Prediction for Small Proteins", *Science*, vol. 309, pp. 1868-1871, 2005.

[28] C. Rohl, C.E.M. Strauss, K.M.S. Misura, and D. Baker, "Protein Structure Prediction using ROSETTA", *Meth. Enzymol.*, vol. 383, pp. 66-93, 2004.

[29] Yang, J.S.; Chen, W.W.; Skolnick, J.; Shankhnovich, and E.I. All-atom, *Ab-initio* Folding of a Diverse Set of Proteins. *Structure*, 2006, 15, pp. 53-63.

[30] S. Wu, J. Skolnick, and Y. Zhang, "Ab-initio Modeling of Small Proteins by Iterative TASSER Simulations", *BMC Biol.*, 5:17doi:10.1186/1741-7007-5-17, 2007.

[31] S. Jyothi, and R.R. Joshi, "Protein Structure determination by non-parametric regression and knowledge-based constraints", *Comput. Chem*., 2001, 25, pp. 283-299.

[32] S. Jyothi, and R.R. Joshi, "Ab-initio Computation of the 3d-Structure of Proteins by Nonparametric Statistical Methods - Medium and Short Range Distance Estimates". Sankhyā, vol. 65(3), pp. 593-611, 2003.

[33] R.R. Joshi, and S. Jyothi, "*Ab-initio* Prediction and Reliability of Protein Structural Genomics by PROPAINOR". *Comput. Biol. Chem*., vol. 27(3), pp. 241-252, 2003.

[34] S. Jyothi, and R.R. Joshi, "3D-Structure of Human seminal plasma prostatic inhibin by nonparametric regression". *Protein Pept. Lett.*, vol. 7(3), pp. 167- 174, 2000.

[35] R.R. Joshi, and S. Jyothi, "*Ab-initio* structure of human seminal plasma prostatic inhibin gives significant insight into its biological functions", *J. Mol. Model.*, vol. 8(2), pp. 50-57, 2002.

[36] R.R. Joshi, and S. Jyothi, "Statistical Computational Approach to Function-Elucidation of the Predicted Structure of an EF-Hand $Ca2^+$ Binding Protein", in 7th Ind. Biophysical Soc. Conference, Pune Univ., 2005.

[37] S. Jyothi, S.M. Mustafi, K.V.R. Chary, and R.R. Joshi, "Structure prediction of a multi-domain EF-hand Ca^{2+} binding protein by *propainor*". *J. Mol. Model*., vol. 14, pp. 481-488, 2005.

[38] H. Glennie, "A comparative study of the reported performance of Ab initio protein structure prediction algorithms", *J. Royal Soc.* Interface. vol. 5(21), pp. 387-396, 2008. R. R. Joshi, and V. V. Samant, V.V. "Fast prediction of protein domain boundaries using conserved local patterns", *J. Mol. Model*., vol. 12(6), pp. 943-952, 2006.

[40] R. R. Joshi, and V.V. Samant, "Bayesian data mining of protein domains gives efficient predictive algorithm and new insight", *J. Mol. Mod*., vol. 13 (1), pp. 275-282, 2007.

[41] L. J. Guffin, K. Bryson, and D.T. Jones, "The PSIPRED protein structure prediction server", *Bioinformatics*, vol. 16, pp. 404-405, 2000.

[42] W. Härdle, Applied Nonparametric Regression. (6th Print), Cambridge: Cambridge, University Press, 2002.

[43] L. E. Stanfel, "A new approach for clustering the amino acids", *J. Theor. Biol.*, vol. 183, pp. 195-205, 1996.

[44] K.C. Chou, "A key driving force in determination of protein structure classes", *Biochem. Biophys. Res. Commun.*, vol. 264, pp. 216-224, 1999.

[45] Q. S. Du, Z. Q. Jiang, W. Z. He, D. P. Li, and K. C. Chou, "Amino acid principal component analysis and its applications in protein structural class prediction", *J. Biomol. Struct. Dyn.* vol. 23, pp. 635-640, 2006.

[46] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman, "Gapped BLAST and PSI-BLAST - a new generation of protein database search programs". *Nucl. Acids. Res*., vol. 25, pp. 3389-3402, 1997.

[47] K. Satone, Statistical Exploration of Activity Pockets in Proteins. M. Sc. Project Dissertation, Bioinformatics Centre, Univ. of Poona. (Guide: Prof. Joshi RR, IIT Bombay), 2009.

[48] R.R. Joshi, K. Satone, R. Patil, and N.T. Jyothish, "Prediction of Functional Sites on Proteins using Logistic Regression", (under communication process), 2010.

[49] G. Amitai, A. Shemesh, E. Sitbon, M. Shklar, D. Netanely, I. Venger, and S. Pietrokovski, "Network Analysis of Protein Structures Identifies Functional Residues", *J. Mol. Biol.*, vol. 218, pp. 183-194, 2004.

[50] K.C. Chou, and C.T. Zhang, "Review: Prediction of protein structural classes", *Crit. Rev. Biochem. Mol. Biol*., vol. 30, pp. 275-349, 1995.

[51] H. Lin, The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition", *J. Theor. Biol.*, vol. 252, pp. 350-356, 2008.

[52] K.C. Chou, and D.W. Elrod, "Protein subcellular location prediction", *Protein Eng.*, vol. 12, pp. 107-118, 1999.

[53] Y.H. Zeng, Y.Z. Guo, R.Q. Xiao, L. Yang, L.Z. Yu, and M.L. Li, "Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach", *J. Theor. Biol.*, vol. 259, pp. 366~V372, 2009.

[54] K.C. Chou, and H.B. Shen, "Hum-PLoc: A novel ensemble classifier for predicting human protein subcellular localization", *Biochem. Biophys. Res. Commun.*, vol. 347, pp. 150-157, 2006.

[55] X.B. Zhou, C. Chen, Z.C. Li, and X.Y. Zou, "Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes", *J. Theor. Biol.*, vol. 248, pp. 546; V551, 2007.

[56] C. Chen, L. Chen, X. Zou, and P. Cai, "Prediction of protein secondary structure content by using the concept of Chou's pseudo amino acid composition and support vector machine", *Protein Pept. Lett.*, vol. 16, pp. 27-31, 2009.

[57] H. Ding, L. Luo, and H. Lin, "Prediction of cell wall lytic enzymes using Chou's amphiphilic pseudo amino acid composition", *Protein Pept. Lett.*, vol. 16, pp. 351-355, 2009.

[58] X. Jiang, R. Wei, T.L. Zhang, and Q. Gu, "Using the concept of Chou's pseudo amino acid composition to predict apoptosis proteins subcellular location: an approach by approximate entropy", *Protein Pept. Lett.*, vol. 15, pp. 392-396, 2008.

[59] F.M. Li, and Q.Z. Li, "Predicting protein subcellular location using Chou's pseudo amino acid composition and improved hybrid approach", *Protein Pept. Lett.*, vol. 15, pp. 612-616, 2008.

[60] H. Lin, H. Ding, F.B. Feng-Biao Guo, A.Y. Zhang, and J. Huang, "Predicting subcellular localization of mycobacterial proteins by using Chou's pseudo amino acid composition", *Protein Pept. Lett.*, vol. 15, pp. 739-744, 2008.

[61] Y. S. Ding, T. L. Zhang, and K. C. Chou, "Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network", *Protein Pept. Lett.*, vol. 14, pp. 811-815, 2007.

[62] T. Wang, J. Yang, H.B. Shen, and K. C. Chou, "Predicting membrane protein types by the LLDA algorithm", *Protein Pept. Lett.*, vol. 15, pp. 915-921, 2008.

[63] K. C. Chou, and H. B. Shen, "ProtIdent: A web server for identifying proteases and their types by fusing functional domain and sequential evolution information", *Biochem. Biophys. Res. Comm.*, vol. 376, pp. 321-325, 2008.

[64] G. Y. Zhang, H. C. Li, and B. S. Fang, "Predicting lipase types by improved Chou's pseudo-amino acid composition", *Protein Pept. Lett.*, vol. 15, pp. 1132-1137, 2008.

[65] L. Holm, and C. Sander, "Database algorithm for generating protein backbone and side-chain coordinates from C_{α}-trace", *J. Mol. Biol.*, vol. 218, pp. 183-194, 1991. (www.ebi.ac.uk/MaxSprout)

Received: February 02, 2010 Revised: March 01, 2010 Accepted: March 12, 2010