W232–W237 *Nucleic Acids Research, 2012, Vol. 40, Web Server issue* *Published online 11 June 2012*
doi:10.1093/nar/gks529

CPred: a web server for predicting viable circular permutations in proteins

Wei-Cheng Lo[1,2], Li-Fen Wang[2], Yen-Yi Liu[2], Tian Dai[1,3], Jenn-Kang Hwang[2,*] and Ping-Chiang Lyu[1,4,*]

[1]Institute of Bioinformatics and Structural Biology, National Tsing Hua University, Hsinchu 30013, [2]Department of Biological Science and Technology, National Chiao Tung University, Hsinchu 30068, Taiwan, [3]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA, USA and [4]Department of Life Sciences, National Tsing Hua University, Hsinchu 30013, Taiwan

Received March 11, 2012; Revised May 9, 2012; Accepted May 11, 2012

ABSTRACT

Circular permutation (CP) is a protein structural rearrangement phenomenon, through which nature allows structural homologs to have different locations of termini and thus varied activities, stabilities and functional properties. It can be applied in many fields of protein research and bioengineering. The limitation of applying CP lies in its technical complexity, high cost and uncertainty of the viability of the resulting protein variants. Not every position in a protein can be used to create a viable circular permutant, but there is still a lack of practical computational tools for evaluating the positional feasibility of CP before costly experiments are carried out. We have previously designed a comprehensive method for predicting viable CP cleavage sites in proteins. In this work, we implement that method into an efficient and user-friendly web server named CPred (CP site predictor), which is supposed to be helpful to promote fundamental researches and biotechnological applications of CP. The CPred is accessible at http://sarst.life.nthu.edu.tw/CPred.

INTRODUCTION

The protein structural rearrangement phenomenon termed circular permutation (CP) can be viewed as if the amino- and carboxyl-termini of a protein were relocated along the circularized sequence of the protein. Although the mechanisms underlying natural CP cases are not fully understood (1–5), many CPs have been observed in well-known protein families [see (6) for summaries of proposed mechanisms for CP and natural CP cases]. To study CP, many artificial circular permutants have been generated. The outcomes of these previous studies have indicated that as long as the CP site, i.e. the position for creating new termini, is not at a residue essential for protein folding or function, circular permutants usually retain native structures and biological functions (1,3,7–9), although the structural stabilities, folding mechanisms and enzymatic activities might be changed (10–15). These discoveries have made CP a new mutagenesis method for studying protein structure and function (16–18) and a bioengineering technique to modify the stability, solubility and activities of proteins (13,19–21). Moreover, the CP technique allows two proteins to be covalently linked at positions other than their native termini, facilitating the creation of several useful protein switches, molecular biosensors and fusion proteins (22–24). Despite these interesting applications, the implementation of CP is much more difficult, expensive and time-consuming compared with traditional mutagenesis. Most importantly, not all positions in a protein structure are permissible for CP. However, since practical software for predicting viable CP sites (i.e. positions leading to correctly folded and stable permutants) is still unavailable, researchers interested in CP-based protein engineering may rely on uneconomic trial-and-error for finding appropriate CP sites. To facilitate fundamental researches based on CP and biotech applications of the CP-based mutagenesis, we aim to develop an effective web-based tool for predicting viable CP site in this work.

CPs tend to occur at positions with high solvent accessibility (25), low sequence conservation and low 'closeness' (26), a structure-derived residue feature describing the amount of residues with which a given residue may interact directly or indirectly (27). However, predicting viable CP sites based on these properties yielded only marginal performance; the area under the receiver operating characteristic curve (AUC) values were all

*To whom correspondence should be addressed. Tel: +886 3 5742765; Fax: +886 3 5715934; Email: lslpc@life.nthu.edu.tw
Correspondence may also be addressed to Jenn-Kang Hwang. Tel: +886 3 5131337; Fax: +886 3 5729288; E-mail: jkhwang@cc.nctu.edu.tw

≤0.7 (26). The major difficulty in developing CP viability predictors was the insufficiency of data, particularly the data of inviable CP sites (i.e. negative cases). In fact, the aforementioned predictors were developed and assessed with a data set composed of only one protein—dihydrofolate reductase (DHFR)—the entire 159-residued polypeptide of which had been subjected to systematic CPs (25). Although large data sets of CP-related protein structural homologs, such as the CP Database (CPDB) (28) and the database of GANGSTA+ Internet Services (GIS) (29), have been available since 2009, there is still a lack of negative cases. This is because most wet-lab researches only reported viable CP sites and bioinformatics CP-detecting methods could only identify existent (meaning viable) circular permutants. The DHFR data set contained only 73 negative cases, far from enough for developing reliable predictors.

Recently, we have established several highly different data sets for developing viable CP site prediction methods (30). Among them, nrCPDB-40 and nrGIS-40 contained thousands of proteins with machine-identified viable CP sites, whereas Data set T consisted of six proteins other than DHFR with both experimentally verified viable and inviable CP sites, expanding the number of negative cases by 2.4-fold (30). Based on these data sets, the sequence and structural preferences of CP were extensively examined (30). The identified preferences were quantified into numerous features to develop a CP viability prediction method that combined four machine learning algorithms: artificial neural networks, the support vector machine, a random forest and a hierarchical feature integration procedure (30). As trained with Data set T, this method achieved an AUC of 0.91 for the DHFR data set and a large-scale prediction sensitivity of ≥0.72 for either nrCPDB-40 or nrGIS-40. However, this effective CP site prediction method is not efficient. Due to heavy computational loads caused by several structural features and the time-consuming data flow through numerous prediction models, it took minutes to deal with one protein.

In our present work, we have implemented the developed CP viability prediction method into a user-friendly and quick response web server named CPred. Distributed computation techniques are used to accelerate the whole procedure, which now takes only seconds to make predictions. CPred is currently the most accurate method and is the only web server for predicting viable CP sites. We hope that it can be a good assistant for scientists and bioengineers to study and apply CP.

MATERIALS AND METHODS

The flowchart of CPred is illustrated in Figure 1a. After receiving the query protein structure from the input module, the main program distributes to several processors the computation tasks of feature values, which are collected again by the main program. The main program then creates four threads running different machine learning predictors, the results of which are integrated and processed by the main program and are finally delivered to the output interface. If the protein structure is input by specifying a PDB [Protein Data Bank (31)] or a Structural Classification of Proteins (32) entry identifier, the calculated feature values and final results will be cached to ensure a quick response once the same protein is queried again in the future.

Experimental data sets

Literature-derived data sets: Data set T and the DHFR data set

Information of inviable CP sites is rare, and it is extremely difficult to find a protein with both experimentally verified viable and inviable CP sites. Before our previous work (30), DHFR was the only data set for training and evaluating CP site predictors. By screening the literature, we had additionally collected six such proteins and established Data set T. Collectively, Data set T (76 viable and 100 inviable CP sites) and the DHFR data set (86 viable and 73 inviable CP sites) are the largest CP site data set currently available with both viable and inviable sites. Data set T and the DHFR data set shared very low sequence identities (<9%); the former was used to train and test our prediction system and the latter was used as an independent evaluation data set. These data sets are available in (30).

Database-extracted data sets: nrCPDB-40 and nrGIS-40

CPDB (28) and GIS (29) are the largest protein structural databases providing information about CP-related structural homologs. Previously, we had reduced these databases to 40% sequence identity non-redundant subsets, nrCPDB-40 and nrGIS-40 (30). Any protein in these two data sets that shared sequence identities >40% with any protein in the Data set T or DHFR were further eliminated. Finally, the nrCPDB-40 and nrGIS-40 data sets contained 1059 and 2814 proteins, respectively, and any two data sets among nrCPDB-40, nrGIS-40, Data set T and the DHFR data set shared <40% sequence identities [see (30) for details].

Non-redundant data set of CP site: nrCPsite$_{cpdb}$-40

All CP sites of the proteins in nrCPDB-40 had been extracted (30). Each CP site was represented by a 20-residued segment. These 20-residued CP site representative segments were reduced to a 40% sequence identity non-redundant subset named nrCPsite$_{cpdb}$-40 (1087 CP sites). Note that a protein may possess more than one viable CP site, and thus the number of non-redundant CP sites (1087) in the nrCPsite$_{cpdb}$-40 data set is larger than the number of proteins (1059) in the nrCPDB-40 data set. The aforementioned data sets had been released as a part of the supporting data of (30).

Computation of feature values

The CPred system extracts 46 features from an input PDB file (see Supplementary Table S1). Based on a statistical significance test known as the permutation test (33), we had previously examined the sequence and secondary structural propensities of CP by comparing the compositions of single-, oligo- and coupled-residue patterns of

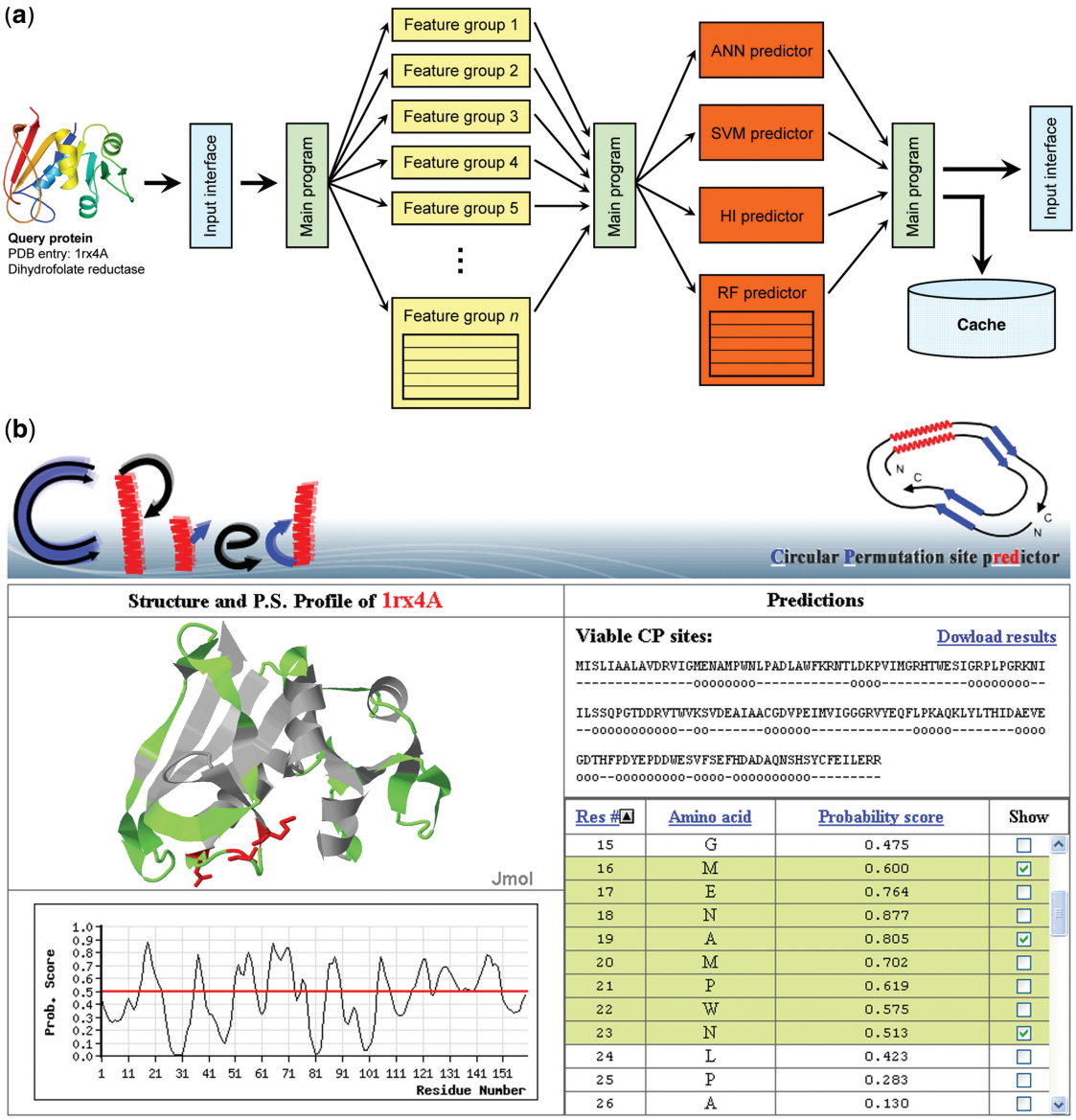

Figure 1. The flowchart and output of CPred. (**a**) CPred is a viable circular permutation cleavage site prediction web server, which is working based on distributed computation techniques. After receiving the query protein data, the main program of CPred will extract feature values, execute machine learning subroutines, integrate the prediction results and deliver the final results to the output interface. The computation loads of many steps are distributed to several processors, as indicated by the radical arrow lines. Some structural features and machine learning methods require much more computation power than others; subroutines responsible for them, as represented by multicelled boxes, are designed by applying distributed techniques as well. (**b**) The output interface of CPred provides a list (lower right) and a graphic profile (lower left) of the probability scores of all residues in the input protein. The structure, along with predicted viable CP sites, is presented by an interactive Jmol (33) object (upper left), which allows the user to change the display mode (cartoon, spacefilled, etc.) and to rotate, resize and dissect the structure. A downloadable text version of the CPred results is provided as well (upper right). The structures shown in panel (a) and (b) were respectively rendered using PyMol (45) and Jmol (33).

amino acid sequence and several secondary structural strings between the CP site segments of nrCPsite$_{cpdb}$-40 and the whole protein sequences of nrCPDB-40 (30). A secondary structural string, for instance, a Ramachandran structural string (34), is a text description of the secondary structure or backbone conformation of a protein. In CPred, these propensities are quantified by using the propensity scoring algorithm proposed in (30). Before CPred extracts tertiary structural features, the reduce program (35) is used to restore hydrogen atoms to the PDB file. Structure-derived residue measures and properties, e.g. the closeness (26), relative solvent accessibility, centroid distance measure (36), weighted contact number (37), farness (30) and the Gaussian Network Model-derived mean-square fluctuation (38–40), are then computed. All the obtained propensity scores and residue measures are standardized based on the conventional Z-score method (30) into real number features suitable for applying machine learning methods.

Application of machine learning methods

Four machine learning methods are applied in the CPred system: (i) a three-layered and back propagation-based (41,42) artificial neural network; (ii) a support vector machine established with the LIBSVM (43); (iii) a random forest of 500 decision trees generated by the C4.5 package (44); and (iv) a hierarchical feature integration procedure, in which features are hierarchically classified into a rooted tree that directs how the feature values are integrated into a single value (30). To efficiently integrate the prediction results from these methods, the output of every method for each residue has been designed to be a real number score between 0 and 1 [see (30) for algorithms]. Because of the range of value of these scores and their being conceptually in direct proportion to the chance that a case is a positive case, we have termed them 'probability scores' for convenience (30). Since these scores have the same range of values, to integrate the prediction power of various methods, we simply average their probability scores into an integrated score, based on which predictions are made. The feasibilities of these individual and integrated methods for predicting viable CP sites had been well established in (30), where the detailed algorithms, parameter settings and performance data are available. In the current work, a major problem in applying these methods lies in the fact that the data flow through the 500 decision trees of the random forest is very time-consuming. To solve this problem, as illustrated in Figure 1a, distributed computation techniques are used.

PERFORMANCE

Evaluations of the prediction system with cross-validation techniques and independent data sets

Before our work, the best viable CP site prediction method was developed based on the closeness measure, which achieved an AUC of 0.7 on the DHFR data set (26) and sensitivity values of 0.62 and 0.61 on the nrCPDB-40 and nrGIS-40 data sets, respectively (30). In our previous study (30), in which the core method of the current CPred system was developed, Data set T was used to establish the prediction model and the 10-fold cross-validated AUC, sensitivity, specificity and Matthews correlation coefficient values for this training data set were 0.91, 0.86, 0.79 and 0.63, respectively. Evaluating the established model with the independent data set DHFR, the aforementioned four performance measures were 0.91, 0.71, 0.92 and 0.64, respectively. A large-scale prediction test on this system registered sensitivity values 0.75 and 0.72 for nrCPDB-40 and nrGIS-40, respectively [refer to (30) for details]. These data indicated that the core method of CPred outperformed previous methods with little data set dependence or overfitting. Since the CPred server is running the same core programs, its performance measures assessed with these data sets are the same with the values listed earlier in the text. In the actual CPred web server, the prediction model is trained with a combined data set of Data set T and DHFR. Evaluations made based on this combined data set with 10-fold cross-validation and based on independent data sets nrCPDB-40 and nrGIS-40 show that accuracy of the actual server is improved as the amount of training data has increased (Table 1).

Evaluations of the developed probability score with information retrieval techniques

To help users interpret the results obtained with CPred, here we examine the average precisions of predictions at various decision thresholds of the probability score by performing 10-fold cross-validated information retrieval experiments. Table 2 demonstrates that a high threshold of probability score would retrieve fewer residues (i.e. a lower recall rate) but obtain a higher proportion of correct positive predictions (i.e. a higher precision) than a low threshold would. In the combined data set of Data set T and DHFR, any residue with a probability score $\geq$0.85 was an actual CP site (precision = 1). Since 82% of the residues predicted as viable CP sites (i.e. probability scores $\geq$0.5) were actual CP sites, this system is quite reliable. Experimenters expecting a high certainty about the viability of the created permutants may choose residues with probability scores $\geq$0.85 to apply CP; at this threshold, only 16% of all residues in a protein will be predicted as viable CP sites (i.e. the predicted positive fraction = 0.16).

Table 1. Performance of CP viability prediction of CPred

Data set	Performance measure	Closeness	CPred
Training set (Data set T + DHFR data set)[a]	AUC	0.753	0.940
	Sensitivity	0.741	0.889
	Specificity	0.687	0.898
	Matthews correlation coefficient	0.428	0.787
nrCPDB-40	Sensitivity	0.622	0.746
nrGIS-40	Sensitivity	0.614	0.719

[a]These results were obtained with 10-fold cross-validation.

Table 2. Performance of CPred at various decision thresholds of the probability score

Probability score	PPF[a]	Recall	Precision
$\geq$0.90	0.06	0.13	1.00
$\geq$0.85	0.16	0.33	1.00
$\geq$0.80	0.26	0.52	0.99
$\geq$0.75	0.33	0.66	0.96
$\geq$0.70	0.39	0.74	0.92
$\geq$0.65	0.43	0.81	0.92
$\geq$0.60	0.49	0.88	0.87
$\geq$0.50	0.54	0.92	0.82
$\geq$0.40	0.61	0.97	0.77
$\geq$0.30	0.69	1.00	0.70
$\geq$0.20	0.78	1.00	0.62
$\geq$0.10	0.90	1.00	0.54
$\geq$0.00	1.00	1.00	0.48

[a]PPF: predicted positive fraction, meaning the proportion of residues predicted as viable CP sites among all residues in the data set.

Speed evaluations

Since some structural features used by CPred such as the closeness (26) and the weighted contact number (37) have a high time complexity in computation and the random forest of CPred possesses many subpredictors, reducing the running time is an important task for developing a quick response server. By applying distributed computation techniques, the computation loads of several time-consuming steps are shared by many processors, greatly enhancing the efficiency of the server. As listed in Supplementary Figure S2, CPred takes only ~3.4 s to make predictions for a protein with 150–200 residues; even for proteins as large as 600 residues, the average running time is <22 s. Without distributed computation, the running time for proteins with 150–200 and approximately 600 residues is, respectively, around 48.8 and 513.6 s. These assessments were performed on the actual CPred server machine, which is a Linux computer with two 3.33 GHz octa-core Intel Xeon CPUs and 128 GB RAM.

WEB SERVER DESCRIPTION

The query interface of CPred accepts three types of input, inclusive of a PDB entry, a Structural Classification of Proteins entry or a PDB file. After the user submits the query data, a notification page will appear to show the status of computation and provide an URL through which the results can be retrieved at a later time if the user decides not to wait. The outputs of CPred include a list of probability scores for all residues of the input protein and an interactive Jmol (33) graphical display of the protein structure that demonstrates the predicted CP sites (Figure 1b). The list of results can be reordered according to the residue number, amino acid type or the probability score.

APPLICATIONS AND FUTURE WORKS

CPred is a user-friendly web server for predicting possible cleavage sites for creating correctly folded and stable circular permutants of proteins. It provides a convenient probability score to help the user select suitable CP cutting sites. An interesting application of CP is to create fusion proteins with tethered sites different from the native termini (22–24). To our knowledge, for every CP-involved fusion protein that has been created, CP was introduced into just one of the fused polypeptides. This is perhaps because of the difficulty in generating two viable circular permutants at the same time. The convenient probability score of CPred may potentiate the production of fusion circular permutants. CP has long been applied to study protein folding. The ability of CPred to predict CP sites implies that it can be used in reverse to predict residues important to folding. Improving protein function is also a useful application of CP. Since residues with low probability scores are unlikely to form viable—not to mention functionally improved—permutants, CPred holds promise for bioengineering by screening out improbable cases. To improve CPred, additional data will be continually collected for training the prediction model. The current CPred server requires protein structures for making predictions. However, there are so many proteins without determined structures. We supposed that a sequence-based viable CP site predictor will further facilitate the application of CP.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Table 1, Supplementary Figure 2 and Supplementary References [34,46–50].

ACKNOWLEDGEMENTS

The authors thank Prof. Chen-Hsiang Yeang at Academia Sinica, Taiwan, for insightful suggestions, and Linwei Technology Ltd. for securing the server machines. The authors of all algorithms and software used in the developed server are sincerely acknowledged.

FUNDING

Funding for open access charge: The National Science Council, Taiwan [100-2745-B-009-001-ASP, Academic Summit Program of National Science Council to J.-K.H.]. National Science Council, Taiwan [100-2627-B-007-005 and 100-2319-B-400-001 to P.-C.L.]; The 'Center for Bioinformatics Research of Aiming for the Top University Program' of National Chiao Tung University and the Ministry of Education, Taiwan, R.O.C.

Conflict of interest statement. None declared.

REFERENCES

1. Cunningham,B.A., Hemperly,J.J., Hopp,T.P. and Edelman,G.M. (1979) Favin versus concanavalin A: circularly permuted amino acid sequences. *Proc. Natl Acad. Sci. USA*, **76**, 3218–3222.
2. Ponting,C.P. and Russell,R.B. (1995) Swaposins: circular permutations within genes encoding saposin homologues. *Trends. Biochem. Sci.*, **20**, 179–180.
3. Lindqvist,Y. and Schneider,G. (1997) Circular permutations of natural protein sequences: structural evidence. *Curr. Opin. Struct. Biol.*, **7**, 422–427.
4. Uliel,S., Fliess,A. and Unger,R. (2001) Naturally occurring circular permutations in proteins. *Protein. Eng.*, **14**, 533–542.
5. Weiner,J. 3rd, Thomas,G. and Bornberg-Bauer,E. (2005) Rapid motif-based prediction of circular permutations in multi-domain proteins. *Bioinformatics*, **21**, 932–937.
6. Lo,W.C. and Lyu,P.C. (2008) CPSARST: an efficient circular permutation search tool applied to the detection of novel protein structural relationships. *Genome. Biol.*, **9**, R11.
7. Ribeiro,E.A. Jr and Ramos,C.H. (2005) Circular permutation and deletion studies of myoglobin indicate that the correct position of its N-terminus is required for native stability and solubility but not for native-like heme binding and folding. *Biochemistry*, **44**, 4699–4709.
8. Tsai,L.C., Shyur,L.F., Lee,S.H., Lin,S.S. and Yuan,H.S. (2003) Crystal structure of a natural circularly permuted jellyroll protein: 1,3-1,4-beta-D-glucanase from Fibrobacter succinogenes. *J. Mol. Biol.*, **330**, 607–620.

9. Vogel,C. and Morea,V. (2006) Duplication, divergence and formation of novel protein topologies. *Bioessays*, **28**, 973–978.
10. Li,L. and Shakhnovich,E.I. (2001) Different circular permutations produced different folding nuclei in proteins: a computational study. *J. Mol. Biol.*, **306**, 121–132.
11. Chen,J., Wang,J. and Wang,W. (2004) Transition states for folding of circular-permuted proteins. *Proteins*, **57**, 153–171.
12. Bulaj,G., Koehn,R.E. and Goldenberg,D.P. (2004) Alteration of the disulfide-coupled folding pathway of BPTI by circular permutation. *Protein Sci.*, **13**, 1182–1196.
13. Qian,Z. and Lutz,S. (2005) Improving the catalytic activity of Candida antarctica lipase B by circular permutation. *J. Am. Chem. Soc.*, **127**, 13466–13467.
14. Anantharaman,V., Koonin,E.V. and Aravind,L. (2001) Regulatory potential, phyletic distribution and evolution of ancient, intracellular small-molecule-binding domains. *J. Mol. Biol.*, **307**, 1271–1292.
15. Todd,A.E., Orengo,C.A. and Thornton,J.M. (2002) Plasticity of enzyme active sites. *Trends. Biochem. Sci.*, **27**, 419–426.
16. Anand,B., Verma,S.K. and Prakash,B. (2006) Structural stabilization of GTP-binding domains in circularly permuted GTPases: implications for RNA binding. *Nucleic. Acids. Res.*, **34**, 2196–2205.
17. Gebhard,L.G., Risso,V.A., Santos,J., Ferreyra,R.G., Noguera,M.E. and Ermacora,M.R. (2006) Mapping the distribution of conformational information throughout a protein sequence. *J. Mol. Biol.*, **358**, 280–288.
18. Nakamura,T. and Iwakura,M. (1999) Circular permutation analysis as a method for distinction of functional elements in the M20 loop of Escherichia coli dihydrofolate reductase. *J. Biol. Chem.*, **274**, 19041–19047.
19. Schwartz,T.U., Walczak,R. and Blobel,G. (2004) Circular permutation as a tool to reduce surface entropy triggers crystallization of the signal recognition particle receptor beta subunit. *Protein Sci.*, **13**, 2814–2818.
20. Yu,Y. and Lutz,S. (2011) Circular permutation: a different way to engineer enzyme structure and function. *Trends Biotechnol.*, **29**, 18–25.
21. Arnold,F.H. (2006) Fancy footwork in the sequence space shuffle. *Nat. Biotechnol.*, **24**, 328–330.
22. Kojima,M., Ayabe,K. and Ueda,H. (2005) Importance of terminal residues on circularly permutated Escherichia coli alkaline phosphatase with high specific activity. *J. Biosci. Bioeng.*, **100**, 197–202.
23. Ostermeier,M. (2005) Engineering allosteric protein switches by domain insertion. *Protein. Eng. Des. Sel.*, **18**, 359–364.
24. Baird,G.S., Zacharias,D.A. and Tsien,R.Y. (1999) Circular permutation and receptor insertion within green fluorescent proteins. *Proc. Natl Acad. Sci. USA*, **96**, 11241–11246.
25. Iwakura,M., Nakamura,T., Yamane,C. and Maki,K. (2000) Systematic circular permutation of an entire protein reveals essential folding elements. *Nat. Struct. Biol.*, **7**, 580–585.
26. Paszkiewicz,K.H., Sternberg,M.J. and Lappe,M. (2006) Prediction of viable circular permutants using a graph theoretic approach. *Bioinformatics*, **22**, 1353–1358.
27. Amitai,G., Shemesh,A., Sitbon,E., Shklar,M., Netanely,D., Venger,I. and Pietrokovski,S. (2004) Network analysis of protein structures identifies functional residues. *J. Mol. Biol.*, **344**, 1135–1146.
28. Lo,W.C., Lee,C.C., Lee,C.Y. and Lyu,P.C. (2009) CPDB: a database of circular permutation in proteins. *Nucleic Acids Res.*, **37**, D328–D332.
29. Guerler,A. and Knapp,E.W. (2010) GIS: a comprehensive source for protein structure similarities. *Nucleic Acids Res.*, **38**, W46–W52.
30. Lo,W.C., Dai,T., Liu,Y.Y., Wang,L.F., Hwang,J.K. and Lyu,P.C. (2012) Deciphering the preference and predicting the viability of circular permutations in proteins. *PLoS One*, **7**, e31791.
31. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic. Acids. Res.*, **28**, 235–242.
32. Chandonia,J.M., Hon,G., Walker,N.S., Lo Conte,L., Koehl,P., Levitt,M. and Brenner,S.E. (2004) The ASTRAL Compendium in 2004. *Nucleic Acids Res.*, **32**, D189–D192.
33. Hesterberg,T., Moore,D.S., Monaghan,S., Clipson,A. and Epstein,R. (2005) *Introduction to the Practice of Statistics*, 5th edn. W.H. Freeman and Company, New York, NY, pp. 14.11–14.70.
34. Lo,W.C., Huang,P.J., Chang,C.H. and Lyu,P.C. (2007) Protein structural similarity search by Ramachandran codes. *BMC Bioinformatics*, **8**, 307.
35. Word,J.M., Lovell,S.C., Richardson,J.S. and Richardson,D.C. (1999) Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.*, **285**, 1735–1747.
36. Shih,C.H., Huang,S.W., Yen,S.C., Lai,Y.L., Yu,S.H. and Hwang,J.K. (2007) A simple way to compute protein dynamics without a mechanical model. *Proteins*, **68**, 34–38.
37. Lin,C.P., Huang,S.W., Lai,Y.L., Yen,S.C., Shih,C.H., Lu,C.H., Huang,C.C. and Hwang,J.K. (2008) Deriving protein dynamical properties from weighted protein contact number. *Proteins*, **72**, 929–935.
38. Bahar,I., Atilgan,A.R. and Erman,B. (1997) Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold Des.*, **2**, 173–181.
39. Haliloglu,T., Bahar,I. and Erman,B. (1997) Gaussian dynamics of folded proteins. *Phys. Rev. Lett.*, **79**, 3090–3093.
40. Zheng,W. (2008) A unification of the elastic network model and the Gaussian network model for optimal description of protein conformational motions and fluctuations. *Biophys. J.*, **94**, 3853–3857.
41. Elarabaty,M. (1989) Aiaa Computers in Aerospace VII Conference. *New Approach for the Solution of Modern Aerospace Systems Using the Artificial-Intelligence*. Monterey, CA, Pts 1 and 2, pp. 300–310.
42. Werbos,P.J. (1994) *The Roots of Backpropagation: from Ordered Derivatives to Neural Networks and Political Forecasting*. Wiley-Interscience, New York, NY.
43. Lin,C.H., Liu,J.C. and Ho,C.H. (2008) Anomaly detection using LibSVM training tools. *Proceedings of the Second International Conference on Information Security and Assurance*. IEEE Computer Society, Washington, DC, USA, pp. 166–171.
44. Quinlan,J.R. (1996) Bagging, boosting, and C4.5. *Proceedings of the Thirteenth National Conference on Artificial Intelligence and the Eighth Innovative Applications of Artificial Intelligence Conference*, Vol. 1 and 2. AAAI Press / The MIT Press, Portland, Oregon, pp. 725–730.
45. DeLano,W.L. (2002) *DeLano Scientific*. San Carlos, CA, USA.
46. Pommie,C., Levadoux,S., Sabatier,R., Lefranc,G. and Lefranc,M.P. (2004) IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. *J. Mol. Recognit.*, **17**, 17–32.
47. Timberlake,K.C. (1992) *Chemistry*, 5th edn. Haper-Collins Publishers Inc., New York, NY.
48. Nelson,D.L. and Cox,M.M. (2004) *Lehninger Principles of Biochemistry*, 4th edn. W.H. Freeman and Company, New York, NY, pp. 75–115.
49. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
50. Yang,J.M. and Tung,C.H. (2006) Protein structure database search and evolutionary classification. *Nucleic Acids Res.*, **34**, 3646–3659.