D328–D332 *Nucleic Acids Research, 2009, Vol. 37, Database issue* *Published online 8 October 2008*
doi:10.1093/nar/gkn679

CPDB: a database of circular permutation in proteins

Wei-Cheng Lo, Chi-Ching Lee, Che-Yu Lee and Ping-Chiang Lyu*

Institute of Bioinformatics and Structural Biology, National Tsing Hua University, Hsinchu 30013, Taiwan

Received August 13, 2008; Revised September 22, 2008; Accepted September 23, 2008

ABSTRACT

Circular permutation (CP) in a protein can be considered as if its sequence were circularized followed by a creation of termini at a new location. Since the first observation of CP in 1979, a substantial number of studies have concluded that circular permutants (CPs) usually retain native structures and functions, sometimes with increased stability or functional diversity. Although this interesting property has made CP useful in many protein engineering and folding researches, large-scale collections of CP-related information were not available until this study. Here we describe CPDB, the first CP DataBase. The organizational principle of CPDB is a hierarchical categorization in which pairs of circular permutants are grouped into CP clusters, which are further grouped into folds and in turn classes. Additions to CPDB include a useful set of tools and resources for the identification, characterization, comparison and visualization of CP. Besides, several viable CP site prediction methods are implemented and assessed in CPDB. This database can be useful in protein folding and evolution studies, the discovery of novel protein structural and functional relationships, and facilitating the production of new CPs with unique biotechnical or industrial interests. The CPDB database can be accessed at http://sarst.life.nthu.edu.tw/cpdb

INTRODUCTION

Circular permutation (CP) in the protein structure is a rearrangement of the amino acid sequence, such that the original amino- and carboxyl-termini of the polypeptide seem to be linked and new ones created elsewhere (1–4). This phenomenon was first observed in plant lectins 30 years ago (5). Since then, many natural cases have been discovered, including some carbohydrate-related enzymes and binding proteins, swaposins, transaldolases, FMN-binding proteins, glutathione synthetases, methyltransferases, ferredoxins, protease inhibitors, etc. (6). To reveal the effects of CP, many artificial circular permutants (CPs) have been generated, inclusive of the anthranilate isomerase, dihydrofolate reductase, T4 lysozyme, ribonucleases, aspartate transcarbamoylase, SH3 domain, ribosomal protein S6 and so on (7,8). The outcomes of these previous studies have indicated that CPs usually retain native structures and biological functions (3–5,9,10), although the stabilities and folding mechanisms might be altered (7,11,12). Since CP may sometimes increase the stability (13), activity or functional diversity (14–16) of proteins, it has been applied to trigger crystallization (13), improve enzyme activities (14), determine critical elements (17,18) and create novel fusion proteins (19–22).

In spite of these interesting properties and applications, there is still much uncertainty about the evolutionary mechanism, importance and natural prevalence of CP (7,9,23,24). Besides, even if there have been a few methods developed for the prediction of viable CPs, their performances were not well-assessed. The major cause of these uncertainties may be the lack of comprehensive resources of CP that can serve as a good base for studying it. This lack was basically because of the complicated rearrangement nature of circular permutation.

Conventional sequence and structural comparison methods employ collinear alignments and are inefficient to identify CP (9,25,26). To detect CP, several brilliant approaches have been developed, such as the sequence-based algorithms by Uliel *et al.* (27) and Weiner *et al.* (2), and the structure-based SHEBA (23), SAMO (26) and FASE (28). Sequence-based methods are fast, but they may miss many far-related CPs with low sequence similarities that can only be identified by structure-based methods (23), which are very time-consuming (6). We have developed an efficient CP-detecting procedure called CPSARST (Circular Permutation Search Aided by Ramachandran Sequential Transformation). The linear encoding methodology (29) and 'double filter-and-refine' strategy of CPSARST not only make it inherit the speed advantages of sequence-based methods but also retain the sensitivity to detect far-related CPs (6).

Here we present CPDB to be the first CP database. The primary data were screened from the Protein Data Bank

*To whom correspondence should be addressed. Tel: +886 3 5742762; Fax: +886 3 5715934; Email: lslpc@life.nthu.edu.tw

(PDB) (30) by using CPSARST and then refined manually. There are currently 4169 nonredundant pairs of circular permutants recorded in the CPDB. CP pairs were grouped into CP clusters according to their direct and indirect CP relationships. Clusters were further grouped into folds and then classes based on their structural similarities. In addition, CPDB hosts a variety of tools and resources for studying CP, such as CP-based structural similarity search services, circularly permuted sequence/structure alignment and visualization tools, network representations of CP relationships, basic statistics of the properties of CPs and CP sites, and a well-organized list of CP-related literatures. Prediction methods for viable CPs described by Paszkiewicz *et al.* (31) are also implemented in the CPDB with some improvements. After an assessment, a measure known as 'closeness' (32) has been found successfully hitting 66.5% of the nonredundant CP sites in CPDB.

CP has long been used to study the folding mechanism of proteins. The evolutionary mechanism of CP itself is also interesting and has drawn many attentions (6). The information compiled in the CPDB is supposed to be helpful to move these research areas forward. Furthermore, most of the bioengineering and biotechnological applications of CP depend on a proper choice of position to create CP. The CP site information and viable CP site prediction methods provided by CPDB shall be advantageous to these fields.

CONTENTS AND METHODS

Identification of CP

Candidate pairs of circular permutants were first retrieved from a nonredundant PDB data set (26 349 polypeptides; see Supplementary List S1) by performing all-against-all searches with CPSARST (6) and then examined by visual inspections. After false cases were eliminated, the determined permutation sites of each pair were refined by the theoretically most accurate approach to identify CP (2,27), that is, generating all possible circularly permuted alignments to find the best way of aligning a pair of proteins. FAST (33) was applied as the structural alignment engine in this step. Finally, 4169 CP pairs consisting of 2238 proteins were identified. Among these cases, some bear multi-domain architectures with intact domain sequences, such as those reported in (34), but most of them are multi-domain proteins with one domain disrupted by CP or single-domain proteins.

There are two major categories of genetic mechanisms proposed to be responsible for CP (1). Duplication/deletion (9,35) and duplication-by-permutation models (1,36) both rely on independent events of gene duplication and partial deletion of terminal regions, while the latter one also emphasizes that an in-frame fusion had occurred along with the duplication. (2) Fusion/fission models (2,24,34) indicate that a pair of circular permutants were created by independent fusions of two smaller components, or, after a protein undergone fission, the resulting two distinct genes subsequently reassembled in a different order. Although it was reported by using sequence-based analyses that, for multi-domain proteins, fusion/fission mechanisms seem more dominant (34), whether this is also true for those permutations within single-domain proteins, however, remains uncertain. A large amount of new structural data has now been retrieved by CPSARST, including those of many functionally and/or structurally similar circular permutants with extremely low sequence identities. We hope that these data provided by CPDB can be helpful to elucidate more clearly the evolutionary mechanism of CP.

Categorization of circular permutants

Circular permutants in the CPDB were categorized in a hierarchical way. First, proteins with direct or indirect CP relationships were grouped into a 'cluster'. For instance, if proteins A and B is a CP pair (designated as A↔B), B↔C is another CP pair and there is no significant CP relationship detected between proteins A and C, then A↔B and B↔C will be considered to have direct while A and C have indirect CP relationships. In this simple cluster (A↔B↔C), A and C may still be related by an unobvious CP, such as a very small permutation size, or they are just linear structural homologs. Next, structural similarities among representative proteins of each cluster, i.e. the most highly connected proteins, were calculated by FAST (33) and then a nearest-neighbor clustering algorithm (37) followed by manual adjustments were performed to group structurally similar clusters into the same 'fold'. Finally, folds were classified into three classes, i.e. mainly-alpha, mainly-beta and alpha–beta mixed proteins according to their secondary structure elemental contents (Supplementary Data S2). The titles and descriptions of each level of categories were given based on the structural and functional information provided by the SCOP (38), PDB (30) and GO (39) databases.

Circularly permuted alignments and the visualization of CP relationships

Circularly permuted structural alignments can be performed by FAST with suitable manipulations to the PDB file, as described in (6). We have implemented this strategy with a user-friendly way of visualization in the CPDB. As Figure 1a illustrates, the different locations of the termini and the position of CP sites can be easily recognized. The structure-based sequence alignment is shown in two different ways. The first is a plain text format in which unaligned regions are represented as gaps (-). The second is a graph with circularized text in which unaligned regions are represented as budding loops. Fewer loops or a smaller size of the loops stand for a larger number of residues that can be well aligned. If a pair of proteins is better aligned with a CP than without it, a CP relationship can be identified (2). If they can be well aligned both with and without a CP, they may be symmetric CPs (23). This circularized sequence alignment is especially helpful when the protein structures are too complicated for the user to trace their details.

CPDB provides two methods to visualize the CP relationships among a group of proteins. For each CP cluster, a graphic 'CP network' was drawn by Osprey (40)

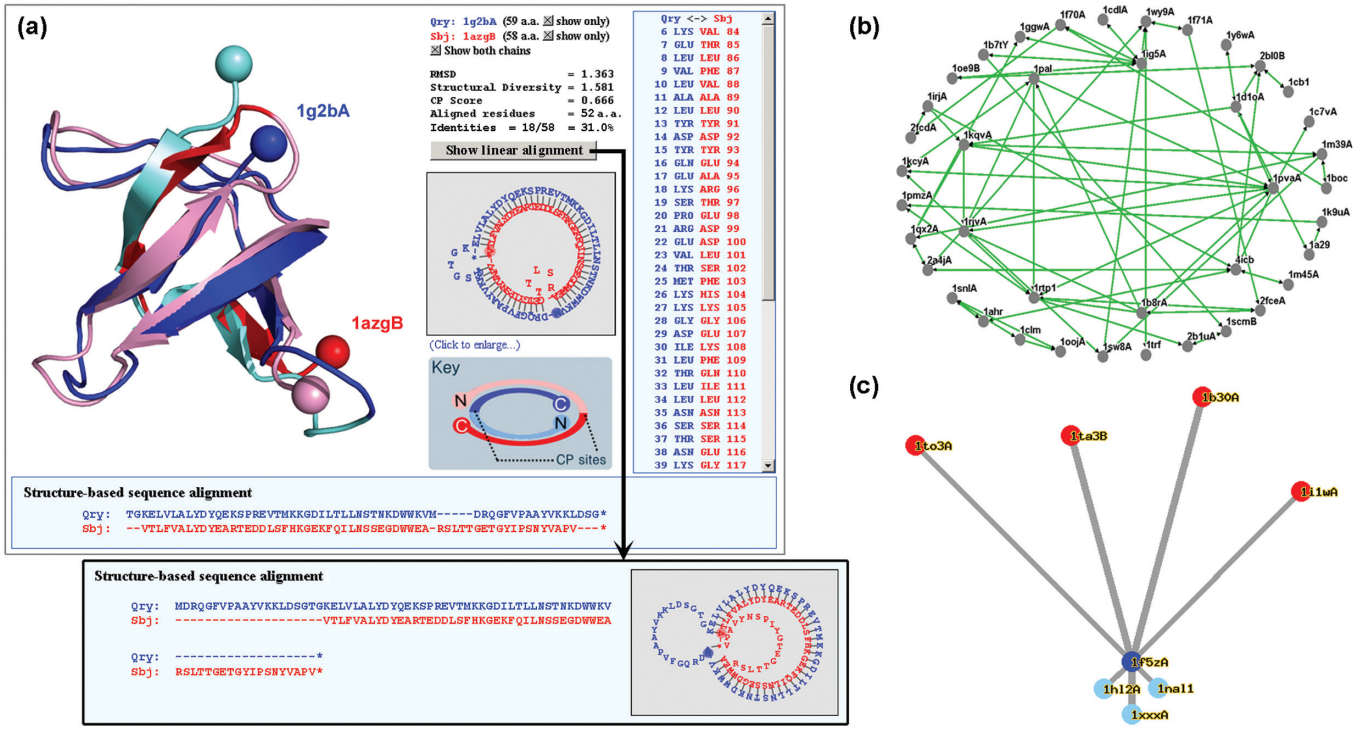

Figure 1. Various methods provided by CPDB for visualizing CP relationships among proteins. (**a**) Circularly permutated structure and sequence alignments. Cα atoms of terminal residues of the superimposed structures are shown as balls so that the different locations of termini, which are a property of CP, can be easily recognized. Two proteins are colored very differently. The boundaries between the lighter and darker colors are the positions of CP site. (**b**) Network view of a CP cluster. A CP cluster usually contains several CP pairs with direct or indirect linkages. Proteins with more complicated CP relationships are placed closer to the center of this network. (**c**) Star-like map of structural homologs. Query protein is at the center (the blue circle) with its circular permutants (red circles) radiating upwards and linear structural homologs (light blue circles) radiating downwards. Connecting lines are drawn in a way that their lengths are in proportion to the structural diversities (41) between proteins.

(Figure 1b). For every protein, a star-like map was generated to show the structural diversities (41) from its circular permutants and linear homologs (Figure 1c).

Prediction of viable circular permutants

A measure known as residue closeness is useful for the identification of active site residues (32). Paszkiewicz *et al.* (31) have proven it also applicable to predict viable CP sites in protein structures and the accuracy is higher than that of relative side-chain area (RSA) or sequence conservation. We have re-implemented their methods of closeness and RSA. The results showed that 62.9% of the nonredundant CP sites in the CPDB could be successfully hit by using closeness and the successful rate of RSA is 60.4%. If we first added hydrogen atoms to PDB structures using the LEaP program of the Amber 6 package (42), the successful rate of closeness and RSA could be raised to 66.5 and 60.9%, respectively.

WEB INTERFACE

CPDB is implemented with MySQL 4 on a HP ProLiant ML570 machine with Linux operating system. A user-friendly web interface was developed by using PHP 5 scripting language, GD graphic library, JavaScript and Chime scripts for easy viewing and retrieval of the data. Figure 2 shows the navigation of the web pages:

- Home page gives the background of CP and some basic statistics of the circular permutants recorded in CPDB.
- Hierarchy browsing, batch browsing and the keyword search pages offer various methods for the users to obtain the information in which they are interested.
- Protein page provides a variety of information including the functions, related references, protein and gene sequences, determined CP sites and CP site predictions. This page is cross-linked with many other pages of CPDB.
- Alignment page offers novel visualization tools to examine circularly permuted sequences and structures.
- CPSARST (6) and SARST (29) are provided to perform rapid structural similarity searches.
- Literature list page offers greatly useful information about CP. Previous reports are well organized according to their purposes and methods. Both wet-lab experimental procedures and computational resources can be found through this page.

FUTURE WORKS

Since the source of protein structures for the current release of CPDB is PDB, according to (6), the type of CP recorded in this database is basically the global CP

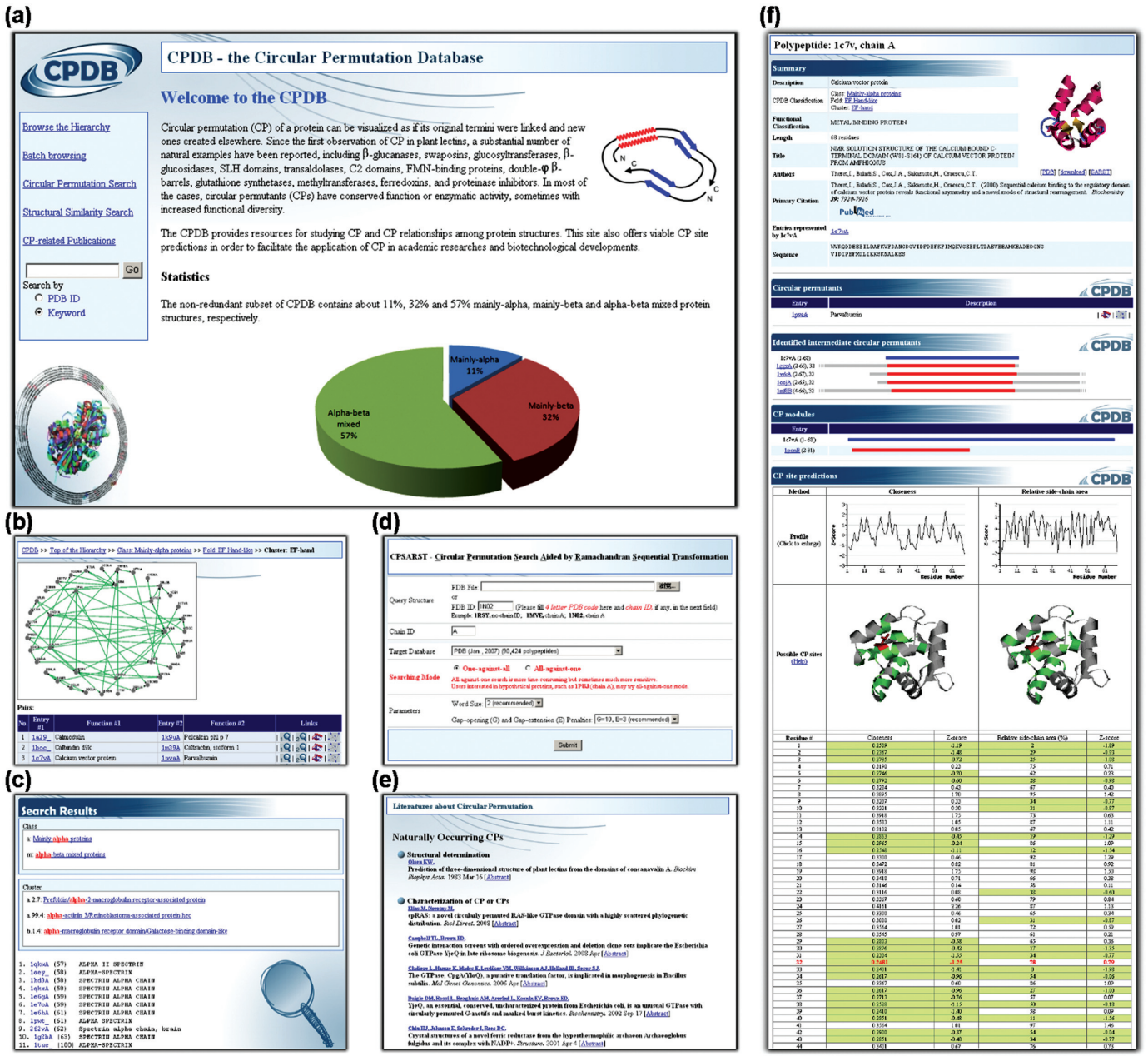

Figure 2. Navigation of the CPDB. (**a**) Home page, (**b**) hierarchy browsing page, (**c**) search results page, (**d**) structural similarity search tools, (**e**) literature list page and (**f**) protein page. See Figure 1a for an example of the alignment pages.

(the unit of CP is the whole protein). However, partial CP (the CP is within a partial region of the protein) also exists in nature, even if some scientists consider it as 'swap' rather than CP (24). We have planned to enhance the ability of CPSARST to identify partial CPs by modifying its strategy and then update CPDB with the retrieved data. Once the information of partial CP is sufficient, a deeper understanding of the effects, importance and evolutionary mechanisms of CP shall be achievable. Besides, including these data will result in a larger training pool that is useful to develop more accurate predictors for viable circular permutants.

SUPPLEMENTARY DATA

Supplementary Data are available at *NAR* Online.

ACKNOWLEDGEMENTS

We thank Dr Margaret Dah-Tsyr Chang, Institute of Molecular and Cellular Biology, NTHU, for her insightful suggestions for the development of CPDB. We also thank Yu-Kwei Chang and Chun-Ting Yeh for their help in manually examining the raw data of CP pairs.

FUNDING

National Science Council, Taiwan, R.O.C. [grant numbers 96-3112-B-007-006, 97-2752-B-007-003-PAE]. Funding for open access charge: National Science Council, Taiwan, R.O.C. [grant number 97-3112-B-007-007].

Conflict of interest statement. None declared.

REFERENCES

1. Jeltsch,A. (1999) Circular permutations in the molecular evolution of DNA methyltransferases. *J. Mol. Evol.*, **49**, 161–164.
2. Weiner,J. III, Thomas,G. and Bornberg-Bauer,E. (2005) Rapid motif-based prediction of circular permutations in multi-domain proteins. *Bioinformatics*, **21**, 932–937.
3. 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.
4. 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.
5. 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.
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. 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.
8. Heinemann,U. and Hahn,M. (1995) Circular permutations of protein sequence: not so rare? *Trends Biochem. Sci.*, **20**, 349–350.
9. Lindqvist,Y. and Schneider,G. (1997) Circular permutations of natural protein sequences: structural evidence. *Curr. Opin. Struct. Biol.*, **7**, 422–427.
10. Vogel,C. and Morea,V. (2006) Duplication, divergence and formation of novel protein topologies. *Bioessays*, **28**, 973–978.
11. 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.
12. Chen,J., Wang,J. and Wang,W. (2004) Transition states for folding of circular-permuted proteins. *Proteins*, **57**, 153–171.
13. 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.
14. 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.
15. 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.
16. Todd,A.E., Orengo,C.A. and Thornton,J.M. (2002) Plasticity of enzyme active sites. *Trends Biochem Sci*, **27**, 419–426.
17. 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.
18. 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.
19. 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.
20. Ostermeier,M. (2005) Engineering allosteric protein switches by domain insertion. *Protein Eng. Des. Sel.*, **18**, 359–364.
21. Galarneau,A., Primeau,M., Trudeau,L.E. and Michnick,S.W. (2002) Beta-lactamase protein fragment complementation assays as in vivo and in vitro sensors of protein protein interactions. *Nat. Biotechnol.*, **20**, 619–622.
22. 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.
23. Jung,J. and Lee,B. (2001) Circularly permuted proteins in the protein structure database. *Protein Sci.*, **10**, 1881–1886.
24. Uliel,S., Fliess,A. and Unger,R. (2001) Naturally occurring circular permutations in proteins. *Protein Eng.*, **14**, 533–542.
25. Russell,R.B. and Ponting,C.P. (1998) Protein fold irregularities that hinder sequence analysis. *Curr. Opin. Struct. Biol.*, **8**, 364–371.
26. Chen,L., Wu,L.Y., Wang,Y., Zhang,S. and Zhang,X.S. (2006) Revealing divergent evolution, identifying circular permutations and detecting active-sites by protein structure comparison. *BMC Struct. Biol.*, **6**, 18.
27. Uliel,S., Fliess,A., Amir,A. and Unger,R. (1999) A simple algorithm for detecting circular permutations in proteins. *Bioinformatics*, **15**, 930–936.
28. Vesterstrom,J. and Taylor,W.R. (2006) Flexible secondary structure based protein structure comparison applied to the detection of circular permutation. *J. Comput. Biol.*, **13**, 43–63.
29. 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.
30. 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.
31. Paszkiewicz,K.H., Sternberg,M.J. and Lappe,M. (2006) Prediction of viable circular permutants using a graph theoretic approach. *Bioinformatics*, **22**, 1353–1358.
32. 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.
33. Zhu,J. and Weng,Z. (2005) FAST: a novel protein structure alignment algorithm. *Proteins*, **58**, 618–627.
34. Weiner,J. III and Bornberg-Bauer,E. (2006) Evolution of circular permutations in multidomain proteins. *Mol. Biol. Evol.*, **23**, 734–743.
35. Ponting,C.P. and Russell,R.B. (1995) Swaposins: circular permutations within genes encoding saposin homologues. *Trends Biochem Sci.*, **20**, 179–180.
36. Peisajovich,S.G., Rockah,L. and Tawfik,D.S. (2006) Evolution of new protein topologies through multistep gene rearrangements. *Nat. Genet.*, **38**, 168–174.
37. Jain,A.K. and Dubes,R.C. (1988) *Algorithms for Clustering Data*. Prentice Hall, New Jersey.
38. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
39. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
40. Breitkreutz,B.J., Stark,C. and Tyers,M. (2003) Osprey: a network visualization system. *Genome Biol*, **4**, R22.
41. Lu,G. (2000) Top: a new method for protein structure comparisons and similarity searches. *J. Appl. Cryst.*, **33**, 176–183.
42. Case,D.A., Cheatham,T.E., III, Darden,T., Gohlke,H., Luo,R., Merz,K.M. Jr, Onufriev,A., Simmerling,C., Wang,B. and Woods,R.J. (2005) The Amber biomolecular simulation programs. *J. Comput. Chem.*, **26**, 1668–1688.