Nucleic Acids Research Advance Access published November 21, 2013

Nucleic Acids Research, 2013, 1–5
doi:10.1093/nar/gkt1189

ArchDB 2014: structural classification of loops in proteins

Jaume Bonet[1], Joan Planas-Iglesias[1], Javier Garcia-Garcia[1], Manuel A. Marín-López[1], Narcis Fernandez-Fuentes[2,*] and and Baldo Oliva[1,*]

[1]Structural Bioinformatics Lab (GRIB-IMIM), Universitat Pompeu Fabra, Barcelona Research Park of Biomedicine (PRBB), Barcelona, Catalonia, 08950, Spain and [2]Institute of Biological, Environmental and Rural Sciences, Aberystwyth University, SY23 3DA Aberystwyth, Ceredigion, UK

Received August 8, 2013; Revised October 9, 2013; Accepted November 1, 2013

ABSTRACT

The function of a protein is determined by its three-dimensional structure, which is formed by regular (i.e. β-strands and α-helices) and non-periodic structural units such as loops. Compared to regular structural elements, non-periodic, non-repetitive conformational units enclose a much higher degree of variability—raising difficulties in the identification of regularities, and yet represent an important part of the structure of a protein. Indeed, loops often play a pivotal role in the function of a protein and different aspects of protein folding and dynamics. Therefore, the structural classification of protein loops is an important subject with clear applications in homology modelling, protein structure prediction, protein design (e.g. enzyme design and catalytic loops) and function prediction. ArchDB, the database presented here (freely available at http://sbi.imim.es/archdb), represents such a resource and has been an important asset for the scientific community throughout the years. In this article, we present a completely reworked and updated version of ArchDB. The new version of ArchDB features a novel, fast and user-friendly web-based interface, and a novel graph-based, computationally efficient, clustering algorithm. The current version of ArchDB classifies 149,134 loops in 5739 classes and 9608 subclasses.

INTRODUCTION

The three-dimensional (3D) structure of a protein is key to determine its function (1,2). In order to exploit this relationship, proteins have been divided and classified according to their fold in databases such as SCOP (3). Structural similarity inferred from these classifications has been used, with different degrees of success, to predict protein functions (4) and interactions (5). Most of these techniques are based on mapping domains over protein sequences via assignation or protein structure modelling (1,3). However, protein domains are also composed of a finite number of secondary structure elements that fit together in a limited number of supersecondary structures (4,6). Supersecondary structures have been used to exploit the structure–function relationship for function and structure prediction (7,8), which has motivated the creation of fragment-based databases such as BriX (9) or SuperLooper (10), protein block identification methods (11,12) and structural alphabets like SA-Mot (13).

Most fragment-based databases split structure fragments according to the number of amino acids involved (i.e. length) and cluster them by means of structural similarity (9). Thus, clusters are limited to fragments of the same length, which allows very little flexibility. On the other hand, methods based on the geometrical relation between two secondary structures have shown a high performance in modelling the aperiodic structure, i.e. loops, connecting them (7,8,14,15).

In a previous work we used the density search (DS) algorithm to combine the geometrical relationship between two secondary structures and the conformation of their linking loop to obtain an automated classification (16). Based on that classification of loops, we have developed ArchDB 2014, which includes super-secondary structures with 3_{10} helices, and a new clustering method that relies on the Markov Clustering (MCL) algorithm (17). This new release of the database still preserves the DS classification in order to maintain consistency with previous database releases. The new database has increased by 5-fold the number of classified loops (from 34 685 to 149 134). Additionally, we have provided a new and intuitive web interface to access the data. We expect this new database to be more useful for the scientific community, in particular for modelling and predicting loop structure and function in proteins. Furthermore, as we

*To whom correspondence should be addressed. Tel: +44 1970 621 680; Fax: +44 1970 622 350; Email: narcis.fernandez@gmail.com
Correspondence may also be addressed to Baldo Oliva. Tel: +34 933 160 509; Fax: +34 933 160 550; Email: baldo.oliva@upf.edu

have recently showed, the classification of loops can also be employed to predict protein–protein interactions (8,18). Consequently, we expect that this new classification will contribute to improve and extend the prediction of new interactions.

DATABASE CONTENT

ArchDB classifies loops based on their flanking secondary structures and geometry. The types of secondary structures considered are: β-strands (E), α-helices (H) and 3_{10} helices (G). The geometry of a loop is defined by the distance and the angles hoist, packing and meridian as described in our previous work (14,15). The ontology of a given loop in the classification is therefore defined by its bracing secondary structures (e.g. α-helix–β-strand), its length and its geometry (16).

Obtaining the loops

Loops were extracted from a non-redundant set of PDB (19) structures with a resolution better than 2.5 Å. Redundancy was removed at 40% sequence identity between PDB chains using CD-HIT (20). The secondary structure of each protein was defined using DSSP (21). Secondary structure was mapped on the corresponding PDB chain sequence when a minimum number of consecutive residues were defined with the same secondary structure type: two, three and four residues for E, G and H, respectively. By this procedure 252 895 different loops were obtained.

Clustering

The new ArchDB contains two independent classifications based on two different clustering algorithms: DS and MCL. In the previous classification, we used DS to classify loops with similar, but not identical length (using a potential deviation of 1 or 2 amino acids). The large increase of protein structures in the PDB makes the implementation of DS clustering of different-length loops computationally unfeasible. However, a classification of loops that takes into account the flexibility in the definition of the hydrogen-bonding network is very useful for loop modelling. Therefore, we have grouped loops according to their length into four different categories (short, medium, long and extra-long) and we have applied the new clustering algorithm, MCL, to each one of those groups. Furthermore, clustering loops with different lengths allows us to bypass the fact that boundaries of secondary structures are difficult to delineate. For instance, automatic algorithms such as DSSP may fail to accurately define the limits of secondary structures, particularly α-helices (22). The DS clustering has been maintained for consistency with previous releases of the database (16), but this was applied only to classify loops with the same length. See Supplementary Material Methods 1 and 2 for further details on the clustering algorithms.

Building the classification

A full independent classification is built for each clustering method, i.e. DS and MLC. Each classification is composed of four levels forming a tree-like hierarchy. At the top of the hierarchy, loops are grouped into 'loop types', which are defined by its bracing secondary structures (see Obtaining the loops section). Consequently, the first level is composed of 10 loop types: alpha–alpha (HH), alpha–beta (HE), beta–alpha (EH), beta–beta hairpin (BN), beta–beta link (BK), beta–helix3_{10} (EG), helix3_{10}–beta (GE), helix3_{10}–helix (GH), helix–helix3_{10} (HG) and helix3_{10}–helix3_{10} (GG). The second level of hierarchy, in descending order, groups the loops by their length. The MCL clustering approach allows a variation of the loop length (see Clustering section), and thus the length of the cluster is defined by the shortest loop(s). The third level is the class, which is defined by grouping all the clusters with a common conformation of the loop region plus the first two amino acid residues in the bracing secondary structures [defined by the (ϕ,ψ) space and referred as Ramachandran consensus]. The lowest level in the hierarchy is the subclass, which corresponds to the individual clusters (Figure 1). Thus, subclasses within the same class share the same loop conformation but have different geometry. Codes for classes and subclasses are assigned by size (number of loops). This means that the most populated class in a given length will have assigned the code '1' and, similarly, the most populated subclass within a class will be the first one. For example, a subclass labelled as 'DS.*HH.1.1.1*' is composed of alpha–alpha (HH) super-secondary structures linked by a loop of one residue, belonging to the most populated class among HH loops of length one and the most populated cluster obtained with the DS approach within this class. The loop classification can be browsed and downloaded through an efficient and user-friendly interface (see Database access section).

Database statistics

A total of 252 895 loops were extracted from a set of 13 238 non-redundant proteins (see Obtaining the loops section). Loops are unevenly distributed among the different types, and only ~50% of them could be classified with each method. The highest percentage of loops classified had short or medium lengths. Two different reasons can be identified as probable causes for this behaviour: (i) the larger number of loops accumulated at shorter lengths and (ii) the smaller number of degrees of freedom in the conformational space of short or medium length loops (Table 1, Figure 2). This observation also agrees with our previous work showing the saturation of loop conformations for short and medium loops (24).

The clustering of loops is RMSD-independent and, thus, this measure can be used *a posteriori* as an indication of the quality of the clustering. The RMSD values of the loops of each cluster were obtained with a structural alignment using STAMP (25). The distribution of RMSD as a function of the loop length is shown in Figure 3 (see Supplementary Figures S1 and S2 for details on each type of loop). The MCL algorithm clusters loops of

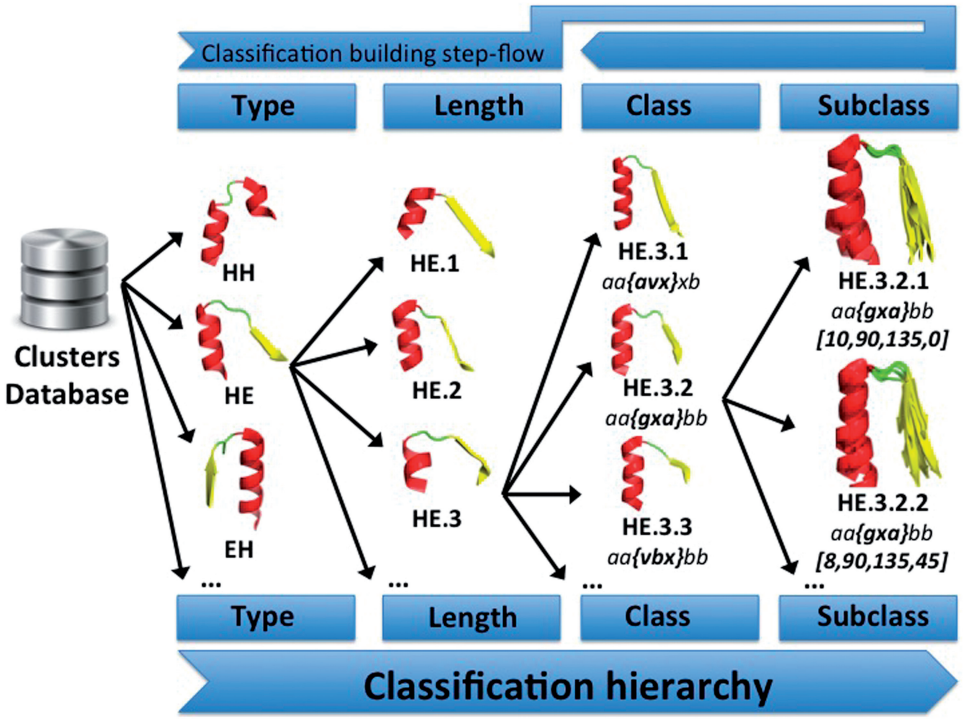

Figure 1. Classification pipeline. Two different methods are applied to build the loop clusters (DS and MCL, see Clustering section and Supplementary Material). Shown within brackets in each subclass is the consensus geometry of the clustered loops, i.e. distance, hoist angle, packing angle and meridian angle [see definitions for loop geometry in the supplementary material, FAQs and in (23)].

Table 1. The different loop types according to their flanking secondary structure

Type	Type description	All	DS (%)	MCL (%)
BK	β-link	28 418	11 777 (41.4)	6054 (21.3)
BN	β-hairpin	35 616	27 995 (78.6)	22 536 (63.3)
EG	β-helix3_{10}	18 349	6950 (37.8)	8531 (46.5)
EH	beta–alpha helix	42 442	23 364 (55.0)	19 661 (46.3)
GE	helix3_{10}–beta	16 478	6829 (41.4)	7731 (46.9)
GG	helix3_{10}–helix3_{10}	3498	704 (20.1)	23 (0.6)
GH	helix3_{10}–α-helix	16 249	7537 (46.9)	10 141 (62.4)
HE	α-helix–β	42 079	24 870 (59.1)	23 327 (55.4)
HG	α-helix–helix3_{10}	14 472	5689 (39.3)	9133 (63.1)
HH	α-helix–α-helix	35 294	18 200 (51.5)	19 503 (55.2)

The total number for each type as well as the number of each type that has been classified is also shown.

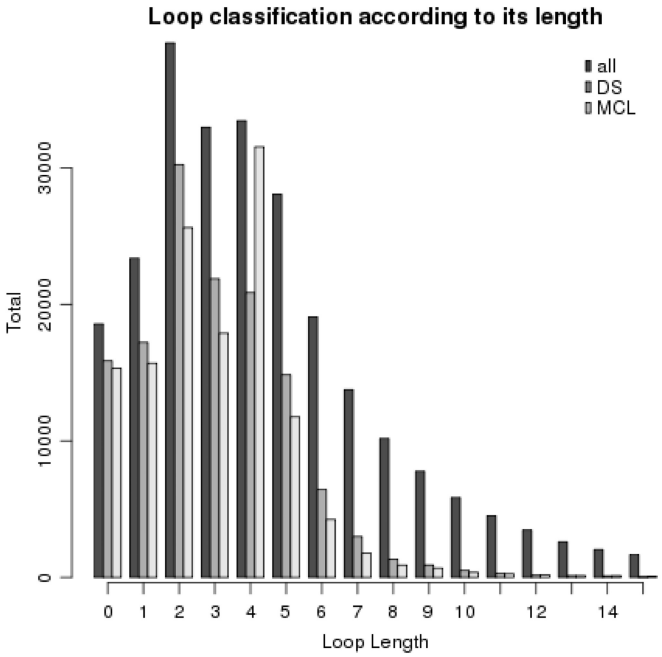

Figure 2. Distribution of classified loops for each of the clustering method as a function of loop length.

different lengths, resulting in slightly higher RMSD measures than the ones obtained using the DS algorithm. Still, the average RMSD is below 1.5 Angstroms. Even with different loop lengths, the distribution of RMSDs when using the MCL algorithm is similar to the distribution obtained with DS algorithm using fixed loop lengths (Figure 3, Supplementary Figures S1 and S2).

Applications of the database

The previous ArchDB classification of loops was used as gold standard to develop new methods for loop prediction [e.g. (26)], as a test set in support-vector-machine methods for the identification of β-hairpins (27), to search templates for protein modelling (15), for function prediction (28), evolutionary conservation (29) and, more recently, to understand and predict protein–protein interactions (8,18). The new database provides new insights useful for researchers focused on the structural/functional features of protein loops [see Example 1 on the P-loop in Supplementary Material; (30)] and improves the

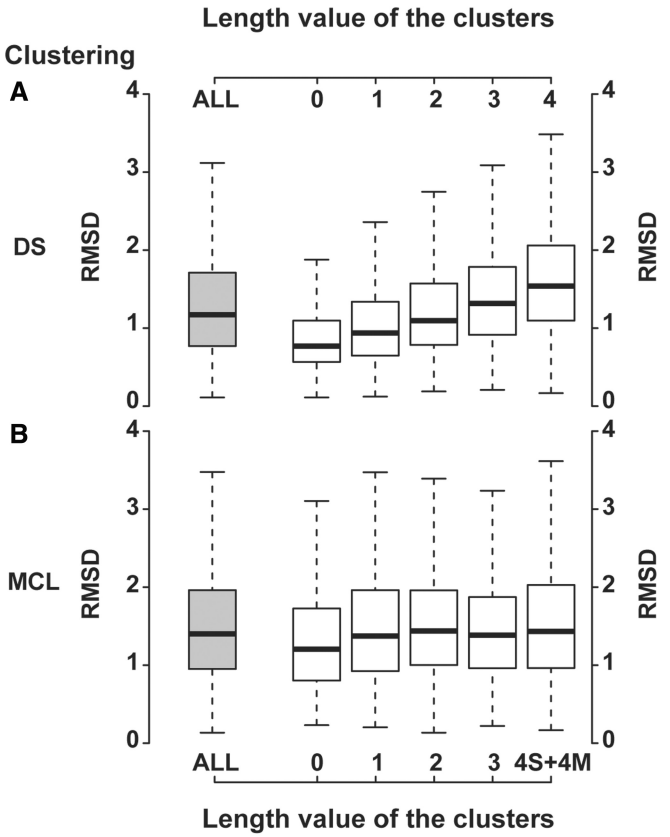

Figure 3. RMSD distribution of the five most populated loop lengths (from 0 to 4) for all loop types. Distribution using DS clustering (top). Distribution using MCL clustering (bottom; this includes two types of subclasses 4S and 4M at length 4). See Supplementary Figures S1 and S2 for a detailed analysis of the RMSD distribution by type-length.

prediction of the structural conformation of loops (by increasing the coverage of loop conformations and the possibility to search among different loop-lengths). Moreover, the annotation of external databases to the classes and subclasses of loops, such as SCOP (3), GO (31), ENZYME (32) or DrugBank (33), and the analysis of interacting heteroatoms and known PDB sites, will help researchers on the annotation of protein function. Finally, the extension of the database of loops will also help to improve the coverage on predictions of protein–protein interactions, the detection of enabling/disabling loops (7) and the annotation of binding sites.

DATABASE ACCESS

The database is available in the form of a user-friendly web interface at http://sbi.imim.es/archdb. The classification is accessible through a composed panel, which allows users to visualize the entire hierarchy, i.e. loop type, loop length, class and subclass, while the selected data is shown in the main section of the web page. There are different visualization modes for every step of the classification. Clustering, type and length views offer useful statistics of the loops included at each level, while class and subclass views offer detailed information that defines such levels. The alignment of the sequence, the secondary structure calculated with DSSP, and the ($\phi\psi$) angles defining the conformation of each loop [in codes as in (16)] is provided in the details of the subclass. External annotations of databases, functional sites from PDB and heteroatoms found at distance shorter than 6 Å from the atoms of the loops, are also shown in the detailed information of the subclass. The enrichment of functions [in GO terms (31) and ENZYME EC codes (32)], drug targets [defined by DrugBank (33)] and SCOP domains (3) provides a useful mechanism to annotate the subclass and infer a putative relationship between function and local structure. Additionally, a downloadable section provides the user with a tab-formatted file containing the most relevant data of the classification for local use. Finally, a Frequent Asked Questions section provides guidance on browsing and understanding the database. In some relevant views (loop and subclass), the web provides 3D visualizations both for each individual loop and for the structural superposition [build with STAMP (25)] and visualization of loops within the subclass.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online, including [34].

Funding

Spanish Ministry of Science and Innovation (MICINN) [FEDER BIO2008-0205, FEDER BIO2011-22568, EUI2009-04018]; FI-DGR 2012 fellowship from 'Generalitat de Catalunya' (to M.A.M.L.). Funding for open access charge: Spanish Ministry of Science and Innovation.

Conflict of interest statement. None declared.

REFERENCES

1. Garcia-Garcia,J., Bonet,J., Guney,E., Fornes,O., Planas-Iglesias,J. and Oliva,B. (2012) Networks of protein–protein interactions: from uncertainty to molecular details. *Mol. Inform.*, **31**, 342–362.
2. Tyagi,M., Hashimoto,K., Shoemaker,B.A., Wuchty,S. and Panchenko,A.R. (2012) Large-scale mapping of human protein interactome using structural complexes. *EMBO Rep.*, **13**, 266–271.
3. Andreeva,A., Howorth,D., Brenner,S.E., Hubbard,T.J.P., Chothia,C. and Murzin,A.G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.*, **32**, D226– D229.
4. Lee,D., Redfern,O. and Orengo,C. (2007) Predicting protein function from sequence and structure. *Nat. Rev. Mol. Cell Biol.*, **8**, 995–1005.
5. Mosca,R., Pons,T., Ceol,A., Valencia,A. and Aloy,P. (2013) Towards a detailed atlas of protein-protein interactions. *Curr. Opin. Struct. Biol.*, July 26 (doi:10.1016/j.sbi.2013.07.005).
6. Fitzkee,N.C., Fleming,P.J., Gong,H., Panasik,N., Street,T.O. and Rose,G.D. (2005) Are proteins made from a limited parts list? *Trends Biochem. Sci.*, **30**, 73–80.
7. Akiva,E., Itzhaki,Z. and Margalit,H. (2008) Built-in loops allow versatility in domain-domain interactions: lessons from self-interacting domains. *Proc. Natl Acad. Sci. USA*, **105**, 13292–13297.
8. Planas-Iglesias,J., Bonet,J., Garcia-Garcia,J., Marín-López,M.A., Feliu,E. and Oliva,B. (2013) Understanding Protein-Protein

Interactions Using Local Structural Features. *J. Mol. Biol.*, **425**, 1210–1224.
9. Vanhee,P., Verschueren,E., Baeten,L., Stricher,F., Serrano,L., Rousseau,F. and Schymkowitz,J. (2011) BriX: a database of protein building blocks for structural analysis, modeling and design. *Nucleic Acids Res.*, **39**, D435– D442.
10. Hildebrand,P.W., Goede,A., Bauer,R.A., Gruening,B., Ismer,J., Michalsky,E. and Preissner,R. (2009) SuperLooper–a prediction server for the modeling of loops in globular and membrane proteins. *Nucleic Acids Res.*, **37**, W571– W574.
11. Moon,H.S., Bhak,J., Lee,K.H. and Lee,D. (2005) Architecture of basic building blocks in protein and domain structural interaction networks. *Bioinformatics*, **21**, 1479–1486.
12. de Brevern,A.G., Etchebest,C. and Hazout,S. (2000) Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins*, **41**, 271–287.
13. Regad,L., Saladin,A., Maupetit,J., Geneix,C. and Camproux,A.-C. (2011) SA-Mot: a web server for the identification of motifs of interest extracted from protein loops. *Nucleic Acids Res.*, **39**, W203– W209.
14. Fernandez-Fuentes,N., Querol,E., Avilés,F.X., Sternberg,M.J.E. and Oliva,B. (2005) Prediction of the conformation and geometry of loops in globular proteins: testing ArchDB, a structural classification of loops. *Proteins*, **60**, 746–757.
15. Fernandez-Fuentes,N., Oliva,B. and Fiser,A. (2006) A supersecondary structure library and search algorithm for modeling loops in protein structures. *Nucleic Acids Res.*, **34**, 2085–2097.
16. Espadaler,J., Fernandez-Fuentes,N., Hermoso,A., Querol,E., Avilés,F.X., Sternberg,M.J.E. and Oliva,B. (2004) ArchDB: automated protein loop classification as a tool for structural genomics. *Nucleic Acids Res.*, **32**, D185–D188.
17. Van Dongen,S. (2008) Graph clustering via a discrete uncoupling process. *SIAM. J. Matrix Anal. Appl.*, **30**, 121–141.
18. Planas-Iglesias,J., Marín-López,M.A., Bonet,J., Garcia-Garcia,J. and Oliva,B. (2013) iLoops: a protein-protein interaction prediction server based on structural features. *Bioinformatics*, **29**, 2360–2362.
19. 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.
20. Fu,L., Niu,B., Zhu,Z., Wu,S. and Li,W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.
21. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
22. Carter,P., Andersen,C.A.F. and Rost,B. (2003) DSSPcont: Continuous secondary structure assignments for proteins. *Nucleic Acids Res.*, **31**, 3293–3295.
23. Oliva,B., Bates,P.A., Querol,E., Avilés,F.X. and Sternberg,M.J. (1997) An automated classification of the structure of protein loops. *J. Mol. Biol.*, **266**, 814–830.
24. Fernandez-Fuentes,N. and Fiser,A. (2006) Saturating representation of loop conformational fragments in structure databanks. *BMC Struct. Biol.*, **6**, 15.
25. Russell,R.B. and Barton,G.J. (1992) Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins*, **14**, 309–323.
26. Regad,L., Martin,J., Nuel,G. and Camproux,A.-C. (2010) Mining protein loops using a structural alphabet and statistical exceptionality. *BMC Bioinformatics*, **11**, 75.
27. Hu,X.Z. and Li,Q.Z. (2008) Prediction of the beta-hairpins in proteins using support vector machine. *Protein J.*, **27**, 115–122.
28. Espadaler,J., Querol,E., Avilés,F.X. and Oliva,B. (2006) Identification of function-associated loop motifs and application to protein function prediction. *Bioinformatics*, **22**, 2237–2243.
29. Fernandez-Fuentes,N., Hermoso,A., Espadaler,J., Querol,E., Avilés,F.X. and Oliva,B. (2004) Classification of common functional loops of kinase super-families. *Proteins*, **56**, 539–555.
30. Saraste,M., Sibbald,P.R. and Wittinghofer,A. (1990) The P-loop–a common motif in ATP- and GTP-binding proteins. *Trends Biochem. Sci.*, **15**, 430–434.
31. The Gene Ontology Consortium. (2012) The Gene Ontology: enhancements for 2011. *Nucleic Acids Res.*, **40**, D559–D564.
32. Bairoch,A. (2000) The ENZYME database in 2000. *Nucleic Acids Res.*, **28**, 304–305.
33. Knox,C., Law,V., Jewison,T., Liu,P., Ly,S., Frolkis,A., Pon,A., Banco,K., Mak,C., Neveu,V. *et al.* (2011) DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.*, **39**, D1035– D1041.
34. Everitt,B., Landau,S. and Leese,M. (2001) *Cluster Analysis*, 4 edn. Oxford University Press, New York, pp. 142–144.