Nucleic Acids Research, 2006, Vol. 34, Web Server issue ***W89–W94***
doi:10.1093/nar/gkl235

ProKware: integrated software for presenting protein structural properties in protein tertiary structures

Jui-Hung Hung[1], Hsien-Da Huang[1,2,3,*] and Tzong-Yi Lee[1]

[1]Institute of Bioinformatics and [2]Department of Biological Science and Technology, National Chiao Tung University, Hsin-Chu 300, Taiwan and [3]Core Facility for Structural Bioinformatics, National Chiao Tung University, Hsin-Chu, Taiwan

Received February 14, 2006; Revised March 6, 2006; Accepted March 28, 2006

ABSTRACT

Protein tertiary structure plays an essential role in deciphering protein functions, especially protein structural properties, including domains, active sites and post-translational modifications. These properties typically yield useful clues for understanding protein functions. This work presents an integrated software, named ProKware, that presents protein structural properties in protein tertiary structures, such as domains, functional sites, families, active sites, binding sites, post-translational modifications and domain–domain interaction. Using this web-based and Windows-based interface, users can manipulate and visualize three-dimensional protein structures, as well as the supported structural properties that are curated in the protein knowledge database. ProKware is an effective and convenient solution for investigating protein functions and structural relationships. This software can be accessed on the internet at http://ProKware.mbc.nctu.edu.tw/.

INTRODUCTION

Protein tertiary structure plays a crucial role in unraveling protein functions. The protein structural properties, including domains, active sites and post-translational modifications, contribute useful information when investigating protein functions. Previously developed software, such as Rasmol (http://www.umass.edu/microbio/rasmol/), PyMol (http://pymol.sourceforge.net/) and the Swiss-PDB Viewer (1), provide an effective approach for visualizing and manipulating protein structures. However, these tools are inconvenient for users who intend to annotate and present protein structural properties collected in various biological databases against protein tertiary structures. Software with enhanced effectiveness and convenience is required for investigating protein structural properties and tertiary protein structures. Thus, this work presents an integrated solution for easily manipulating and demonstrating the protein structural properties in protein tertiary structures.

The previously developed program, PdbMotif (2), automatically identifies protein motifs with PDB protein structures and generates a scripting file that can be imported directly into the molecular rendering program RasMol. Investigated motifs can be highlighted automatically when visualizing protein structure. Motif3D (3), which is a web-based protein structure viewer that allows users to input sequence motifs, visually presents the domain/motifs along protein tertiary structures. 3MOTIF (4), which is a web application that visualizes discrete sequence conservation data by freely available Chime plugin and interface with other bioinformatics resources.

To facilitate annotation of protein structural properties, which can be imported into a protein knowledge database, an integrated program called ProKware was developed to present protein structural properties, such as domains, functional sites, families, active sites, binding sites, post-translational modifications and domain–domain interaction, in protein tertiary structures. Using the web-based and Windows-based interface, users can manipulate and visualize three-dimensional (3D) protein structures and their supported structural properties that are maintained in the back-end protein knowledge database. Consequently, ProKware is an effective and convenient tool for presenting protein structural properties in protein structures for investigating protein functions and structural relationships.

OVERVIEW

The proposed software consists of two principal components: a protein knowledge database and a web-based and a Windows-based application. Figure 1 depicts the ProKware

To whom correspondence should be addressed. Tel: +886 3 571 2121, ext. 56952; Fax: +886 3 572 9288; Email: bryan@mail.nctu.edu.tw

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

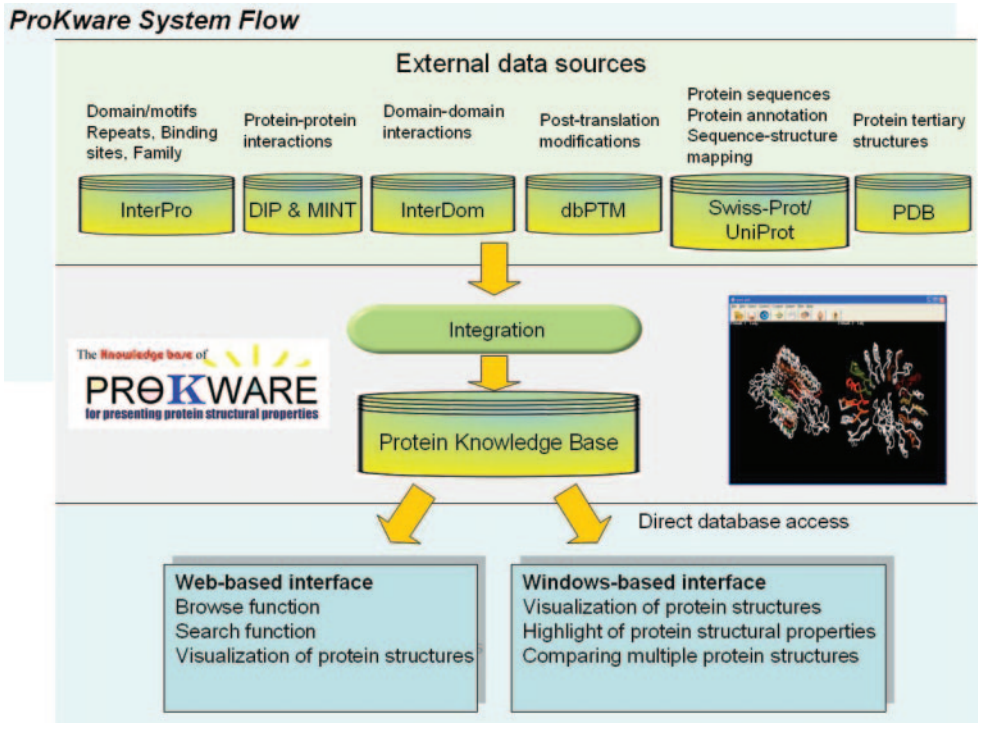

Figure 1. ProKware system flow.

Table 1. The external data sources integrated into the ProKware protein knowledge database

Knowledge category	Number of entries	Database name	URL	Reference
Protein domains/motifs[a]	11 007	InterPro	http://www.ebi.ac.uk/interpro/	(6)
Post-translational modifications	786 211	dbPTM[b]	http://dbptm.mbc.nctu.edu.tw/	(11)
Protein tertiary structure	26 260	PDB	http://www.rcsb.org/pdb/	(8)
Protein sequence	176 469	Swiss-Prot	http://www.expasy.org/sprot/	(5)
Sequence-structure mapping	40 648	UniProt	http://www.pir.uniprot.org/	(15)
Domain–domain interaction	60 850	InterDom	http://interdom.lit.org.sg/	(7)
Protein-protein interaction	55 732	DIP	http://dip.doe-mbi.ucla.edu/	(9,10)
	70 036	MINT	http://mint.bio.uniroma2.it/mint/	

[a]This category contains domains, active sites, binding sites, repeats, functional sites and families from diverse databases (6).
[b]dbPTM contains 14 057 known PTM sites and 772 154 putative PTM sites.

system flow. The protein knowledge database was first constructed to compile protein sequences, protein tertiary structures and protein structural properties, obtained in a variety of biological databases in the public domain. The protein knowledge database utilizes a MySQL database to support the data access of the web-based and Windows-based application.

Second, the Windows-based application presents essential protein structural properties, such as domains, binding sites and post-translational modification on protein structural coordinates. The application can connect to the protein knowledge database when users require information in the protein knowledge database. After retrieving the required data, the application can be run as a stand-alone program. To access the functional sites information, users must choose the desired functional protein sites. Once the specific functional sites and corresponding PDB files are chosen, the software merges the functional site information accessed from the protein knowledge database.

Users can retrieve numerous protein structures with completed functional regional information or putative domain–domain interactions. The visualization pipeline then accommodates this additional reinforcement and transfers the protein structural information for 3D rendering.

PROTEIN KNOWLEDGE DATABASE

The protein sequences, domain/motifs, protein–protein interacting domains, protein tertiary structures and post-translational modifications were obtained from Swiss-Prot (5), InterPro (6), InterDom (7), PDB (8), DIP (9), MINT (10) and dbPTM (11), respectively (Table 1).

The Protein Data Bank (PDB) (8) is a worldwide database that deposits the 3D coordinates of protein structures. Numerous databases, such as Pfam (12), PROSITE (13), ProDom (14) and InterPro (6), have been developed for collecting protein functional sites, domains/motifs and protein families. Pfam is a large collection of sequence alignments and two profile-hidden Markov models (HMMs) that represent protein functional domains and families. PROSITE incorporates multiple sequence alignments and HMMS that define protein

Table 2. Characteristics of ProKware

Comparing features	ProKware	Motif3D	3MOTIF	PdbMotif	Descriptions
Direct access	Yes	Yes	Yes	—	Connecting the database directly
Database supported	Knowledge Database[a]	PRINTS	PROSITE	PROSITE	Annotation according to specific database
Web-based interface	Yes	Yes	Yes	—	Supporting viewing the protein properties on web
Window-based application	Yes	—	—	—	Supporting viewing the protein properties of line
RasMol script	Yes	—	Yes	Yes	Script for RasMol viewer
PyMol script	Yes	—	—	—	Script for PyMol viewer
Full domain display	Yes	—	—	—	Capable of displaying the full definition of all functional sites within specific protein structures
DD interaction display	Yes		—	No	Capable of displaying the multiple protein having interaction
PTM sites support	Yes	—	—	—	Showing the PTM small molecule with PTM sites
Split windows	Yes	—	—	—	At max 4 separated frames showing diverse contents
Graphic presentation	Rich[b]	Mono.(Java applet)	By Chime plugin	By RasMol	3D graphic performance and structure presentation
Motif management	Yes	—	—	—	Management of different protein with protein properties
Cross platform	Not yet	Yes	Yes	Yes	The portability of software

DD, domain–domain.
[a]Swiss-Prot, InterPro, InterDom, PDB and dbPTM.
[b]ProKware supports cartoon and strand views for easy verification of protein secondary structures. OpenGL pipeline provides efficient viewing performance.

conserved regions. ProDom is a comprehensive set of protein domain families automatically generated from the Swiss-Prot and TrEMBL (5) sequence databases. To integrate these protein domain databases into a single resource, InterPro merges roughly 30 databases of protein families, domains and functional sites by assigning an accession number to each domain/motif or functional site from distinct databases and presents these numbers via a web-based interface. InterPro integrates motif, domain and protein family information from Prosite, PRINTS, SMART, pFAM, TIGR FAMS, Prodom, UniProt, PANTHER, MSD, SUPERFAMILY and SCOP. ProKware integrates the comprehensive and systematic compilation of InterPro into the proposed protein knowledge database. InterDom (7) compiles putative protein domain–domain interactions that can be utilized as supporting evidence when determining protein interactions.

Protein post-translational modification (PTM) is an extremely important cellular regulatory mechanism that affects protein physical and chemical properties, folding, conformation distribution, stability, activity, and consequently, protein functions. In the previous work by the authors of this study, dbPTM (11) comprehensively annotated PTM information for Swiss-Prot proteins. The PTM of proteins, which are structural properties, can be presented on the corresponding protein tertiary structures.

The biological databases above provide useful information of protein structural properties for determining protein functions. However, an integrated resource must be developed that collects and manages these protein structural properties. The InterPro (Release 8.0) contains 11 007 entries and InterDom contains 60 850 domain–domain interactions. Combining this information with InterPro provides a macroscopic understanding of protein properties and biological processes.

Moreover, to examine the regulatory mechanisms of protein activity and the involvement in biological process and metabolic pathways, the PTM is critical information and extremely important in the proposed knowledge database. The dbPTM adequately identifies the known and putative PTM sites for phosphorylation, glycosylation, and other types of PTMs. In total, 14 057 known PTM sites and 772 154 putative PTM sites were compiled.

Consequently, to present all functional regions, these regions must be mapped with the PDB files. Parts of the achievements of this task are developed utilizing sequence–structure mapping information extracted from UniProt—UniProt has enhanced integration with structural databases by utilizing residue level mapping of sequences from the PDB entries onto corresponding UniProt entries. In total, 40 648 entries exist for such mapping. After mapping functional regions (including dbPTM data) on protein structures, domain–domain interaction was added to establish the connection between primary sequences and tertiary structures.

VISUALIZATION INTERFACE

Combining structural properties and tertiary structures is required for the proposed software to maximize exploitation of the protein structural information. When a merged file is input, the parser recognizes the specific header and stores the additional information. Users can choose available PDB files for protein tertiary structures, functional regions and different definitions using the function 'Motif manager' in ProKware. The 3D rendering software in ProKware is OpenGL-based, which at present only runs on WinOS. ProKware has numerous convenient features that enhance insight gained into protein properties and biological processes (Table 2).

In addition, ProKware also contains direct protein viewing on web. This trait is written based on ActiveX techniques and embedded in ProKware web server to provide convenient structure viewing (must be installed).

The website for the ProKware consists of three search categories: domain, protein and PTM.

Domain. When users input an InterPro entry, ProKware retrieves the general information for this domain—including the full and short domain names, and the abstract catalogued in InterPro database—the relative PDB file—including PDB files that contain this domain and the corresponding Swiss-Prot accession number—definitions in different databases—for instance PROSITE and Pfam that use different approaches of searching putative LRR domain sites (12,13)—and the putative domain–domain interaction of this domain—including

the domain name in Pfam and corresponding InterPro entry. The web interface also contains a button that directly activates the ProKware web-based interface to graphically present the 3D domain structure on the web (users must install ProKware) (Fig. S1, see Supplementary Data). Domains contained within the target PDB protein structure are highlighted.

Protein. When users input a PDB entry, the software retrieves the Swiss-Prot accession number corresponding to each chain in the PDB file —including name and sequence. ProKware also identifies all possible domains and PTM sites contained within this PDB file and allows users to observe the structure of these domains via the ProKware web-based interface.

PTM. Searching with PTM differs from two other categories described above. Users must input a keyword of a specific PTM, for instance, a user can input 'phos' to search all related PTMs (Fig. S2, see Supplementary Data), and with the help of the ProKware web-based interface, the modified residue of the PTM can be displayed on the web. If users prefer using other platforms or other visualization software, the proposed website provides the scripts that can be read by RasMol and PyMol.

The proposed ProKware provides a powerful and useful visual interface for presenting protein structural properties or elements in protein tertiary structures. To simplify the process of obtaining the required data, ProKware provides an interactive dialog that is activated by mouse clicks throughout the entire process. Four scenarios (Figure 2a) are possible when investigating protein structural properties underlying the protein tertiary structures: (i) users extract a specified protein domain/motif in a specific protein structure; (ii) users retrieve all protein domain/motifs for specific protein structures; (iii) users extract protein structures that have domains interacting with another specific domain and (iv) users retrieve post-translation modifications for a specific protein structure.

Figure 2a presents a 3D visualization of protein structural properties and protein tertiary structures corresponding to a user's selection.

Each of the categories results in a particular presentation (Figure 2b). In the domain category, aided by direct access to ProKware protein knowledge database, users can obtain different definitions for specific functional protein regions in various databases. In the full domain category, users can see all functional regions with different definitions in multiple 3D protein structures. To efficiently manage these diverse functional regions, ProKware used the Motif manager, which handles domain labeling in four frames. In the domain–domain interaction category, by inputting the target domain and identifying the appropriate interacting domain with corresponding proteins, users can compare spatial conformations of these interacting domains and understand their biological processes. In the case of PTM, ProKware presents all known and putative PTM sites within a selected protein structure and illustrates the modified residue to users graphically.

In addition, ProKware also utilizes a variety of functions that facilitate and simplify visualizing protein tertiary structures. Multi-frame visualization allows users to compare multiple protein structures simultaneously. Furthermore, direct

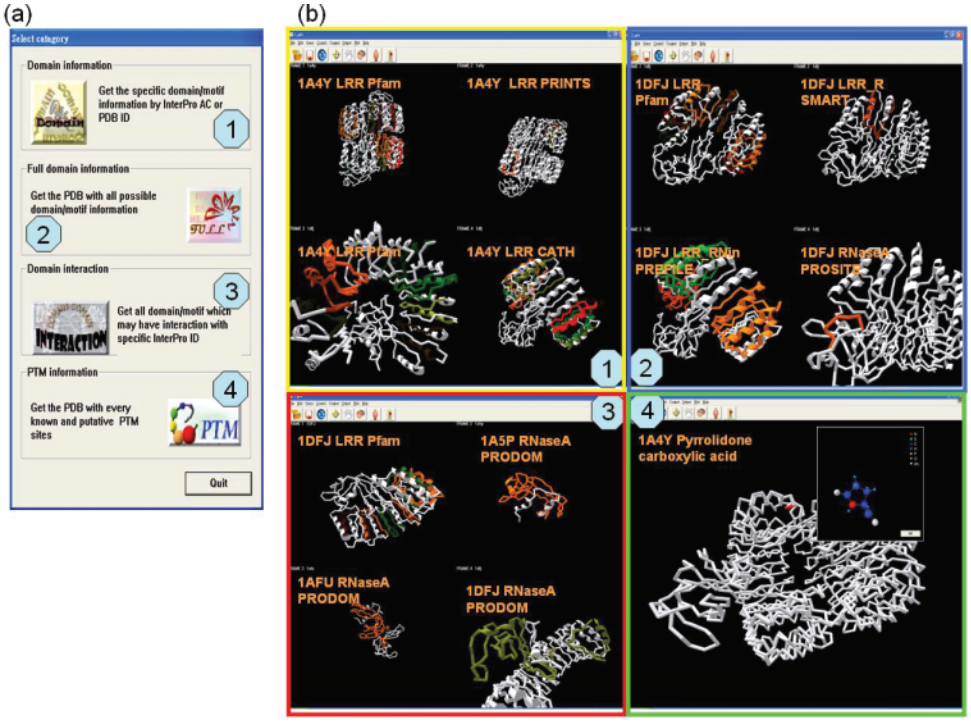

Figure 2. ProKware application. (**a**) Users select the following scenarios to investigate the protein structural properties (1) extracting a specific protein domain/motif in a specified protein structure; (2) retrieving all protein domain/motifs in a specific protein structure; (3) extracting protein structures that have domains interacted with another specified domain; (4) retrieving post-translational modification for a specific protein structure. (**b**) The 3D visualization of the protein structural properties and protein tertiary structures corresponding to a user's selection in (a).

selection of protein residues is supported to allow users to click on a particular residue and highlight it with a particular color. To present a 3D image for a protein structure, ProKware presents the protein structure in backbone, strand, and cartoon views.

IMPLEMENTATION

The principal achievement in enhancing the utility in understanding functional regions and domain–domain interaction is primarily based on direct access to the ProKware protein knowledge database and active interaction with users. Direct access to the ProKware protein knowledge database is the most intuitive and efficient method for understanding protein properties, which simplifies complex and time-consuming tasks. For instance, when users intend to figure out what to identify the functional regions in a certain protein, users first need to know to which PDB entry it belongs and visits all databases that provide functional region annotations for the protein. After collecting this information, users must create a script file that highlights these regions on their protein structure viewer. Such tasks are obviously time consuming and complex. The proposed software, ProKware, develops two user-friendly interfaces, such as the web-based interface and Windows-based application, allowing users to retrieve data from the protein knowledge database.

The web-based interface and the Windows-based application are implemented using C++ programming language and Win32 API; the ActiveX technique was exploited to design the web-based application. The 3D graphic interface for the web-based and Windows-based application utilizes an OpenGL pipeline. The environment of the back tier protein knowledge database utilizes a Linux Apache server with a MySQL database system.

DISCUSSION AND CONCLUSION

Table 1 gives the comparisons with other relevant works, ProKware contributes both Window- and web-based application, of which efficient and comprehensive integration provides users the best experience in dissecting protein properties, especially in direct access of knowledge database of diverse protein structural properties.

The ProKware does not currently work for a protein structure with a crystal structure that is determined by NMR. This disadvantage can also be found in RasMol. When users import an NMR structure, the proposed ProKware displays all structures simultaneously, and observing the colored annotation region becomes difficult. Moreover, ProKware in its current version is incapable of presenting nucleic acid molecules (e.g. DNA and RNA); i.e. ProKware ignores the molecules and only displays the part of protein structures. In addition to web-based interface, ProKware currently supports a Windows-based application. In the near future, ProKware will be extended to support additional platforms. During prospective work, it is necessary to continuously update ProKware protein knowledge database.

The proposed software is a convenient and intuitive interface that facilitates the presentation of protein structural properties in protein tertiary structures. The primary contributions of this work are as follows: (i) ProKware comprehensively collects the protein structural properties, protein sequences and protein tertiary structures into a protein knowledge database; (ii) a web-based interface is utilized for information retrieval; (iii) a Windows-based application is alternatively utilized for information retrieval and (iv) an effective and flexible graphical interface for depicting protein tertiary structure as well as the protein structural properties.

AVAILABILITY

The ProKware web server will be continuously maintained and updated. The web server is now freely available at http://ProKware.mbc.nctu.edu.tw/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank the National Science Council of the Republic of China for financially supporting this research under Contract No. NSC 95-3112-E-009-002. Special thanks for the financial supports from National Research Program For Genomic Medicine (NRPGM), Taiwan. This work was also partially supported by MOE ATU. Funding to pay the Open Access publication charges for this article was provided by National Science Council of the Republic of China.

Conflict of interest statement. None declared.

REFERENCES

1. Kaplan,W. and Littlejohn,T.G. (2001) Swiss-PDB Viewer (Deep View). *Brief Bioinform.*, **2**, 195–197.
2. Saqi,M.A. and Sayle,R. (1994) PdbMotif—a tool for the automatic identification and display of motifs in protein structures. *Comput. Appl. Biosci.*, **10**, 545–546.
3. Gaulton,A. and Attwood,T.K. (2003) Motif3D: Relating protein sequence motifs to 3D structure. *Nucleic Acids Res.*, **31**, 3333–3336.
4. Bennett,S.P., Nevill-Manning,C.G. and Brutlag,D.L. (2003) 3MOTIF: visualizing conserved protein sequence motifs in the protein structure database. *Bioinformatics*, **19**, 541–542.
5. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
6. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bradley,P., Bork,P., Bucher,P., Cerutti,L. *et al.* (2005) InterPro, progress and status in 2005. *Nucleic Acids Res.*, **33**, D201–205.
7. Ng,S.K., Zhang,Z., Tan,S.H. and Lin,K. (2003) InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes. *Nucleic Acids Res.*, **31**, 251–254.
8. Deshpande,N., Addess,K.J., Bluhm,W.F., Merino-Ott,J.C., Townsend-Merino,W., Zhang,Q., Knezevich,C., Xie,L., Chen,L., Feng,Z. *et al.* (2005) The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res.*, **33**, D233–237.
9. Salwinski,L., Miller,C.S., Smith,A.J., Pettit,F.K., Bowie,J.U. and Eisenberg,D. (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–451.

10. Zanzoni,A., Montecchi-Palazzi,L., Quondam,M., Ausiello,G., Helmer-Citterich,M. and Cesareni,G. (2002) MINT: a Molecular INTeraction database. *FEBS Lett*., **513**, 135–140.
11. Lee,T.Y., Huang,H.D., Hung,J.H., Huang,H.Y., Yang,Y.S. and Wang,T.H. (2006) dbPTM: an information repository of protein post-translational modification. *Nucleic Acids Res*., **34**, D622–627.
12. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res*., **32**, D138–141.
13. Hulo,N., Sigrist,C.J., Le Saux,V., Langendijk-Genevaux,P.S., Bordoli,L., Gattiker,A., De Castro,E., Bucher,P. and Bairoch,A. (2004) Recent improvements to the PROSITE database. *Nucleic Acids Res*., **32**, D134–137.
14. Bru,C., Courcelle,E., Carrere,S., Beausse,Y., Dalmar,S. and Kahn,D. (2005) The ProDom database of protein domain families: more emphasis on 3D. *Nucleic Acids Res*., **33**, D212–215.
15. Bairoch,A., Apweiler,R., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al*. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res*., **33**, D154–159.