Chapter 17

BIOINFORMATICS TOOLS FOR THE NEXT GENERATION OF METAL BIOTECHNOLOGY

Hideki Nakayama
Faculty of Environmental Studies, Nagasaki University,
1-14 Bunkyo-machi, Nagasaki 852-8521, Japan
nakayamah@nagasaki-u.ac.jp

17.1 Introduction

At present, above 5900 genomes of living organisms have been completely sequenced and stored as text-based data files through the Internet-based database systems. Most database systems were made freely available to the public through the World Wide Web-based systems, such as the National Center for Biotechnology Information (NCBI) Entrez,[1] the Kyoto Encyclopedia of Genes and Genomes (KEGG),[2] and the Genomes OnLine Database (GOLD).[3] It is estimated that more than one third of all known proteins are metalloproteins, which require metals to maintain their structures and functions in living organisms.[4] Obviously, there is not enough experimental evidence to elucidate metal-selection mechanisms of the complete set of metalloproteins encoded by genomes of living

Handbook of Metal Biotechnology: Applications for Environmental Conservation and Sustainability
Edited by Michihiko Ike, Mitsuo Yamashita, and Satoshi Soda

www.panstanford.com

organisms. Based on Irving–Williams series of relative stabilities of complexes formed by metal ions ($Ca^{2+} < Mg^{2+} < Mn^{2+} < Fe^{2+} < Co^{2+} < Ni^{2+} < Cu^{2+} > Zn^{2+}$),[5] metalloproteins would bind most strongly to divalent (cupric) copper, and to a lower strength to other metal ions. To maintain life systems, however, metalloproteins must selectively bind to specific metal ions required for their function. Recently, it is proposed that metal-binding selectivity of metalloproteins is determined by spatiotemporal folding of the proteins in the cells, and not only by the basis of the nature, number, and geometric arrangement of the binding residues, or the size and charge of the metal-binding pocket.[6]

In the field of metal biotechnology research, it is important to discover and engineer metalloproteins with improved functions involving recognition or sensing of metals, chelating or binding of metals, and reduction or oxidation of metals, which are useful for various types of metal biotechnologies, such as metal-biosensors for monitoring of metal-pollution, bioleaching of metals from their ores, and either bioadsorption or biomineralization for recovering ionic metals from environmental water systems, as shown in other chapters of this book. However, discovery of novel and superior metalloproteins is still difficult due to a lack of reliable high-throughput experimental procedures. Therefore, bioinformatics tools are required as a primary screening tool to predict and identify a group of candidate metalloproteins for further identification and confirmation. For example, based on nucleotide and amino acid sequences, putative metalloproteins, such as zinc-finger proteins or metalloenzymes, can be predicted by the structural similarity with known metalloproteins, the presence of specific metal-binding sites, or metal-binding domains. Subsequently, the predicted metalloproteins can be analyzed for their functional relationship with specific metals in the cells, and the superior metalloproteins identified by this process can be used for development of metal biotechnology.

17.2 Public Protein Database as a Gold Mine

As the number of completely sequenced genomes increases in public databases, international research communities are now refocusing on collecting information about all the proteins encoded in these genomes. Emerging public protein databases allow us to access and rationalize large amount of protein data.

17.2.1 *Universal Protein Resource*

The Universal Protein Resource (UniProt; http://www.uniprot.org/) is a comprehensive resource for protein sequence and annotation data.[7] UniProt is produced by the UniProt Consortium which consists of groups from the European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB), and the Protein Information Resource (PIR). UniProt is composed of four major components, each optimized for different uses: the UniProt Archive (UniParc), the UniProt Knowledgebase (UniProtKB), the UniProt Reference Clusters (UniRef), and the UniProt Metagenomic and Environmental Sequence Database (UniMES), as follows.

(a) *UniParc* is the most comprehensive publicly accessible non-redundant protein sequence database available, providing links to all underlying sources and versions of these sequences. Researchers can instantly find out whether a sequence of interest is already in the public domain and, if not, identify its closest relatives.

(b) *UniProtKB* is used to access functional information on proteins. The UniProtKB consists of two sections: Swiss-Prot, which is manually annotated and reviewed, and TrEMBL, which is automatically annotated and is not reviewed. Every UniProtKB entry contains the amino acid sequence, protein name or description, taxonomic data, and citation information. In addition, UniProtKB contains further annotation that includes widely accepted biological ontologies, classifications and cross-references, as well as clear indications on the quality of annotation in the form of evidence attribution to experimental and computational data.

(c) *UniRef* provides clustered sets of sequences from UniProtKB and selected UniParc records. UniRef90 and UniRef50 yield a database size reduction of approximately 40% and

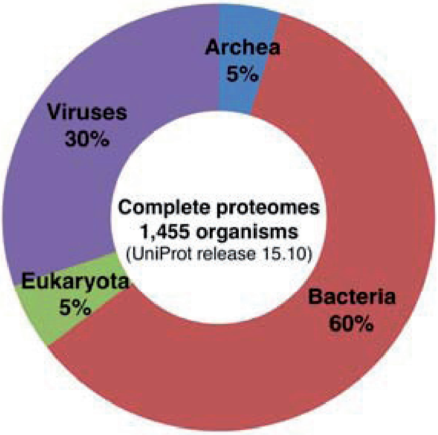

Figure 17.1. Distributions of organisms in the complete proteomes of UniProt. To be included in the complete proteomes, an organism must have a completely sequenced genome, i.e., fully closed and exhibiting either good gene prediction models or good quality transcriptome/proteome data. Therefore, for bacterial and archaeal genomes, whole-genome shotguns (WGS) and draft sequences are not included.

65%, respectively, providing significantly faster sequence searches.

(d) *UniMES* is a repository specifically for metagenomic and environmental data.

At present, UniProt (release 15.10) contains complete proteomes of total 1455 organisms including 68 archea, 872 bacteria, 76 eukariota, and 439 viruses (Fig. 17.1).

Searching UniProtKB under the field of Gene Ontology (GO) for entry whose molecular function is metal ion binding (GO:0046872), a total of 772,892 entries are retrieved. These entries are considered to be putative metalloproteins. The detail results for each metal ion are shown in Table 17.1. Currently, 16 types of metalloproteins are found based on their possible binding to 16 different types of metals including 3 alkali metals (Li, Na, and K), 2 alkaline earth metals (Mg and Ca), 10 transition metals (V, Mn, Fe, Co, Ni, Cu, Zn, Mo, Cd, Hg), and 1 other metal (Pb).

Table 17.1. Sixteen types of putative metalloproteins listed in UniProtKB.

Metal ions	Number of metalloproteins	GO IDs
Iron (Fe) ion	324,953	GO:0005506
Zinc (Zn) ion	229,453	GO:0008270
Copper (Cu) ion	117,795	GO:0005507
Magnesium (Mg) ion	79,617	GO:0000287
Calcium (Ca) ion	37,669	GO:0005509
Manganese (Mn) ion	23,548	GO:0030145
Potassium (K) ion	9,597	GO:0030955
Molybdenum (Mo) ion	9,444	GO:0030151
Nickel (Ni) ion	8,038	GO:0016151
Cobalt (Co) ion	6,469	GO:0050897
Sodium (Na) ion	4,028	GO:0031402
Mercury (Hg) ion	660	GO:0045340
Cadmium (Cd) ion	132	GO:0046870
Lithium (Li) ion	36	GO:0031403
Vanadium (V) ion	18	GO:0051212
Lead (Pb) ion	2	GO:0032791

17.2.2 *Worldwide Protein Data Bank*

The Worldwide Protein Data Bank (wwPDB; http://www.wwpdb.org/) consists of organizations that act as deposition, data processing, and distribution centers for PDB data.[8] The members are RCSB PDB (USA), PDBe (Europe), and PDBj (Japan). The mission of the wwPDB is to maintain a single PDB Archive of macromolecular structural data that is freely and publicly available to the global community. Currently, total number of wwPDB depositions is more than 56,000 structures.

In addition, the chemical component dictionary is found in wwPDB as an external reference file describing all residue and small molecule components. This dictionary contains detailed chemical descriptions for standard and modified amino acids/nucleotides, small molecule ligands, and solvent molecules, which include metal ions and metal-containing ligand molecules. The PDBeChem (http://www.ebi.ac.uk/msd-srv/chempdb/), one of the PDBe services, offers possibilities for searching and exploring the dictionary. The chemical component dictionary currently contains 224 different

Table 17.2. Metal-containing ligand molecules in PDB entries.

Core metals	Number of metal ions and metal-containing ligand molecules	Number of linked PDB entries
Fe	54	1,759
Cu	22	717
Mo	19	111
Ni	16	426
Hg	16	395
Zn	13	5,320
Mg	13	4,901
Co	13	459
Mn	12	1,327
Ca	11	4,671
Na	11	2,457
V	10	64
K	5	867
Cd	4	484
Pb	3	40
Li	2	34
Total	224	24,032

types of metal ions and metal-containing ligand molecules, which are linked to 24,032 PDB entries (Table 17.2).

17.3 Mining of Protein Domains

Proteins are generally composed of one or more core functional elements, commonly termed domains. Different combinations of domains allow evolving the diverse range of proteins with various functions in nature. The identification of domains that occur within proteins can provide valuable insights into their functions. As metal-binding domains are core functional elements found in metalloproteins, the bioinformatics tools for mining of metal-binding domains would be advantageous to all researchers in the field of metal biotechnology.

17.3.1 *InterPro: The Integrative Protein Signature Database*

InterPro (http://www.ebi.ac.uk/interpro/) is a database of protein families, domains, regions, repeats, sites, and posttranslational modifications (PTMs) of known proteins.[9] InterPro (release 23.1) contains 19,269 entries, representing 82 active sites, 58 binding sites, 546 conserved sites, 5466 domains, 11,537 families, 1302 regions, 23 PTMs, and 255 repeats.

17.3.1.1 Protein family and domain databases integrated with InterPro

Now InterPro is integrated with member databases including UniProtKB and other 11 protein family and domain databases shown in Table 17.3.

(a) *Pfam* is a comprehensive collection of protein domains and families, represented as multiple sequence alignments and as profile Hidden Markov Models (HMMs).[10] The Pfam (release 24.0) contains 11,912 protein families. The Pfam is

Table 17.3. Members of protein-family and domain database integrated in InterPro (release 23.1) consortium.

Member database	Integrated version	Signatures*	Integrated signatures**
Pfam	23.0	10,340	10,329
TIGRFAMs	8.0	3,603	3,581
PIRSF	2.70	2,742	2,691
PANTHER	6.1	30,128	2,234
PROSITE	20.52	2,168	2,141
PRINTS	39.0	1,950	1,927
SUPERFAMILY	1.69	1,538	1,090
Gene3D	3.0.0	2,147	1,026
ProDom	2006.1	1,894	992
SMART	6.0	809	804
HAMAP	280509	1,633	502
UniProtKB	15.10	8,123,918	7,813,392

*Some signatures may not have matches to UniProtKB proteins.
**Not all signatures of a member database may be integrated at the time of InterPro release.

available on the web from the consortium members using the web sites in the UK (http://pfam.sanger.ac.uk/), the USA (http://pfam.janelia.org/), and Sweden (http://pfam.sbc.su.se/), as well as from mirror sites in France (http://pfam.jouy.inra.fr/) and South Korea (http://pfam.ccbb.re.kr/).

(b) *TIGRFAMs* (http://www.jcvi.org/cms/research/projects/tigrfams/) are a collection of protein families featuring curated multiple sequence alignments, HMMs, and associated information designed to support the automated functional identification of proteins by sequence homology.[11]

(c) *PIRSF* (Protein Information Resource SuperFamily) classification system (http://pir.georgetown.edu/pirsf/) reflects evolutionary relationships of full-length proteins and domains.[12] The primary PIRSF classification unit is the homeomorphic family, whose members are both homologous (evolved from a common ancestor) and homeomorphic (sharing full-length sequence similarity and a common domain architecture). PIRSF families are curated systematically based on literature review and integrative sequence and functional analysis, including sequence and structure similarity, domain architecture, functional association, genome context, and phyletic pattern.

(d) *PANTHER* (Protein ANalysis THrough Evolutionary Relationships) classification system (http://www.pantherdb.org/) classifies genes by their functions, using published scientific experimental evidence and evolutionary relationships to predict function even in the absence of direct experimental evidence.[13] Proteins are classified by expert biologists into families and subfamilies of shared function, which are then categorized by molecular function and biological process ontology terms. For an increasing number of proteins, detailed biochemical interactions in canonical pathways are captured and can be viewed interactively.

(e) *PROSITE* (http://www.expasy.org/prosite/) is a protein domain database for functional characterization and annotation.[14] The PROSITE (release 20.54) contains 1308 patterns, 863 profiles, and 869 ProRules.

(f) *PRINTS* (http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/) is a compendium of protein fingerprints.[15] A fingerprint is a group of conserved motifs used to characterize a protein family; its diagnostic power is refined by iterative scanning of a SWISS-PROT/TrEMBL composite. Usually the motifs do not overlap, but are separated along a sequence, though they may be contiguous in 3D-space. Fingerprints can encode protein folds and functionalities more flexibly and powerfully than can single motifs, full diagnostic potency deriving from the mutual context provided by motif neighbors.

(g) *SUPERFAMILY* (http://supfam.cs.bris.ac.uk/SUPERFAMILY/) is a database of structural and functional annotation for all proteins and genomes.[16] The SUPERFAMILY annotation is based on a collection of HMMs, which represent structural protein domains at the Structural Classification of Proteins (SCOP)[17] superfamily level. The annotation is produced by scanning protein sequences from over 1200 completely sequenced genomes against the HMMs.

(h) *Gene3D* (http://gene3d.biochem.ucl.ac.uk/) provides accurate structural domain family assignments for over 1100 genomes and nearly 10 million proteins.[18] A HMM library, constructed from the manually curated CATH (Class, Architecture, Topology, Homology) structural domain hierarchy,[19] is used to search UniProt, RefSeq, and Ensembl[20] protein sequences. The resulting matches are refined into simple multi-domain architectures using a recently developed algorithm, DomainFinder 3 (ftp://ftp.biochem.ucl.ac.uk/pub/gene3d_data/DomainFinder3/). The domain assignments are integrated with multiple external protein function descriptions (e.g. GO and KEGG), structural annotations (e.g. coiled coils, disordered regions, and sequence polymorphisms) and family resources (e.g. Pfam and eggNog[21]) and displayed on the Gene3D website.

(i) *ProDom* (http://prodom.prabi.fr/) is a comprehensive set of protein domain families automatically generated from the UniProtKB.[22] The ProDom (release 2006.1) contains 1,716,114 domain families.

(j) *SMART* (Simple Modular Architecture Research Tool; http://smart.embl.de/) is an online tool for the identification and annotation of protein domains.[23] It provides a user-friendly platform for the exploration and comparative study of domain architectures in both proteins and genes. The SMART (release 6.0) contains manually curated models for 784 protein domains. The underlying protein database is based on completely sequenced genomes of 630 species. The interaction network view is available for more than 2 million proteins.

(k) *HAMAP* (High-quality Automated and Manual Annotation of microbial Proteomes) system (http://www.expasy.org/sprot/hamap) is composed of two databases, the proteome database and the family database.[24] The proteome database comprises biological and sequence information for each completely sequenced microbial proteome. The family database currently comprises more than 1600 manually curated orthologous protein families that belong to one of the HAMAP families.

17.3.1.2 Functional molecules for metal ion binding found in InterPro

Searching InterPro under the field of GO for entry whose molecular function is metal ion binding (GO:0046872), a total of 789 entries are found as functional molecules for metal-ion binding among known proteins. The results of each GO ID are extracted in Table 17.4. Among of the known metalloproteins, protein families or domains binding to Zn or Fe ions are well-characterized at present, while no entries are found for protein families or domains binding to Li or Pb ions in InterPro.

17.3.2 *Bioinformatics Tools for Prediction of Metal-Binding Domains*

Since number of sequences and structures of proteins with unknown biological function are continually accumulating in public databases, sophisticated and efficient tools for metal-binding domain

Table 17.4. Numbers of functional molecules as metal ion binding in InterPro.

Metals	Number of InterPro entry	GO IDs
Zn	228	GO:0008270
Fe	129	GO:0005506
Ca	88	GO:0005509
Mg	83	GO:0000287
Cu	47	GO:0005507
Ni	28	GO:0016151
Mn	18	GO:0030145
Mo	14	GO:0030151
Co	6	GO:0050897
K	4	GO:0030955
Hg	3	GO:0045340
Na	2	GO:0031402
Cd	1	GO:0046870
V	1	GO:0051212
Li	0	GO:0031403
Pb	0	GO:0032791

prediction are needed for further progress of metal biotechnology. At present, several publicly available bioinformatics tools have been developed to predict metal binding residues from sequence data. These tools, listed below, can be used for finding and design novel metal-binding domains.

(a) *MDB* (Metalloprotein Database and Browser; http:// metallo. scripps.edu/) is a web-accessible resource for metalloprotein research.[25] It includes quantitative information on geometrical parameters of metal-binding sites in protein structures available from the wwPDB.

(b) *MetSite* (http://bioinf.cs.ucl.ac.uk/MetSite/) represents a fully automatic approach for the detection of metal-binding residue clusters applicable to protein models of moderate quality.[26] The method involves using sequence profile information in combination with approximate structural data. MetSite allows users to scan query structures using one of the six metal type (Ca, Zn, Mg, Fe, Cu, and Mn) classifiers.

(c) *FEATURE metal scanning* (http://feature.stanford.edu/ metals/) is a currently developing tool for identification of metal binding sites in proteins with no exist sequence similarity to known structures.[27] At present, only zinc binding sites could be identified by this tool.

(d) *MetalDetector* (http://metaldetector.dsi.unifi.it/) is a classifier that predicts transition-metal binding for Cys and His residues in protein; for Cys it also predicts disulfide bonding bridges.[28]

(e) *CHED server* (http://ligin.weizmann.ac.il/ched/) uses the "CHED" algorithm to predict 3D intra-chain protein binding sites for transition metals (Zn, Fe, Mn, Cu, Ni, Co), and for Ca and Mg sites that can be replaced by a transition metal.[29] The algorithm searches for a triad of amino acids composed of four residue types (Cys, His, Glu, Asp; CHED) having ligand atoms within specific distances.

(f) *SeqCHED server* (http://ligin.weizmann.ac.il/seqched/) is a sequence-based prediction server that enables the user to analyze a translated gene sequence for transition metals (Zn, Fe, Ni, Cu, Co, Mn), and for Ca and Mg binding sites.[30] The application checks for homology of your target sequence to PDB template sequences and then models the target side chains in 3D (using SCCOMP[31]) on the backbone of the selected template. A metal binding prediction algorithm (based on the CHED procedure) is then applied to the 3D model to identify any putative binding sites and their ligating CHED residues.

17.4 The Next Generation of Metal Biotechnology

Using the current bioinformatics tools, researchers could obtain lists of putative metalloproteins and metal-binding domains. However, for further innovation for next generation of metal biotechnology, we need to develop and establish high-throughput experimental methods for analyzing the function of metalloproteins and metal-binding domains on those lists.

Recent progress on ionome,[32,33] metalloproteomes,[34] and metallomics[35] are brought on by the use of efficient high-throughput analytical machineries and sophisticated bioinformatics tools.

Researchers are focusing now on the continually accumulating data obtained from these studies to elucidate interactions between metal ions and biomolecules including genome, proteome, and metabolome. These omics studies would clarify the role of metal ions, metalloproteins, and metal-binding domains in living organisms and provide crucial ideas for innovation of the next generation of metal biotechnology. Alternatively, combinatorial bioengineering integrated with cell-surface display system[36] would be a promising approach to create and screen novel function of metalloproteins and metal-binding domains. Although it is still challenging, functional design of metalloproteins has been initiated.[37] Further progression of both metalloproteins and bioinformatics studies would open gateway to the future of metal biotechnology.

References

1. Baxevanis, A. D., Searching NCBI databases using Entrez. *Curr. Protoc. Bioinformatics*, Chapter 1, pp. Unit 1 3 (2008).
2. Aoki-Kinoshita, K. F. and Kanehisa, M., *Methods Mol. Biol.* **396**, 71–91 (2007).
3. Bernal, A., Ear, U., and Kyrpides, N., *Nucleic Acids Res.* **29**, 126–127 (2001).
4. Tainer, J. A., Roberts, V. A., and Getzoff, E. D., *Curr. Opin. Biotechnol.* **3**, 378–387 (1992).
5. Irving, H. and Williams, R. J. P., *Nature* **162**, 746–747 (1948).
6. Tottey, S., Waldron, K. J., Firbank, S. J., Reale, B., Bessant, C., Sato, K., Cheek, T. R., Gray, J., Banfield, M. J., Dennison, C., and Robinson, N. J., *Nature* **455**, 1138–1142 (2008).
7. The UniProt Consortium, *Nucleic Acids Res.* (2009).
8. Berman, H., Henrick, K., Nakamura, H., and Markley, J. L., *Nucleic Acids Res.* **35**, D301–D303 (2007).
9. Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., Finn, R. D., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Laugraud, A., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., Mitchell, A., Mulder, N., Natale, D., Orengo, C., Quinn, A. F., Selengut, J. D., Sigrist,

C. J., Thimma, M., Thomas, P. D., Valentin, F., Wilson, D., Wu, C. H., and Yeats, C., *Nucleic Acids Res.* **37**, D211–D215 (2009).

10. Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H. R., Ceric, G., Forslund, K., Eddy, S. R., Sonnhammer, E. L., and Bateman, A., *Nucleic Acids Res.* **36**, D281–D288 (2008).
11. Haft, D. H., Selengut, J. D., and White, O., *Nucleic Acids Res.* **31**, 371–373 (2003).
12. Nikolskaya, A. N., Arighi, C. N., Huang, H., Barker, W. C., and Wu, C. H., *Evol. Bioinform. Online* **2**, 197–209 (2006).
13. Mi, H., Guo, N., Kejariwal, A., and Thomas, P. D., *Nucleic Acids Res.* **35**, D247–D252 (2007).
14. Sigrist, C. J., Cerutti, L., de Castro, E., Langendijk-Genevaux, P. S., Bulliard, V., Bairoch, A., and Hulo, N., *Nucleic Acids Res.* (2009).
15. Attwood, T. K., Beck, M. E., Flower, D. R., Scordis, P., and Selley, J. N., *Nucleic Acids Res.* **26**, 304–308 (1998).
16. Wilson, D., Madera, M., Vogel, C., Chothia, C., and Gough, J., *Nucleic Acids Res.* **35**, D308–D313 (2007).
17. Andreeva, A., Howorth, D., Chandonia, J. M., Brenner, S. E., Hubbard, T. J., Chothia, C., and Murzin, A. G., *Nucleic Acids Res.* **36**, D419–D425 (2008).
18. Lees, J., Yeats, C., Redfern, O., Clegg, A., and Orengo, C., *Nucleic Acids Res.* (2009).
19. Cuff, A. L., Sillitoe, I., Lewis, T., Redfern, O. C., Garratt, R., Thornton, J., and Orengo, C. A., *Nucleic Acids Res.* **37**, D310–D314 (2009).
20. Hubbard, T. J., Aken, B. L., Ayling, S., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Clarke, L., Coates, G., Fairley, S., Fitzgerald, S., Fernandez-Banet, J., Gordon, L., Graf, S., Haider, S., Hammond, M., Holland, R., Howe, K., Jenkinson, A., Johnson, N., Kahari, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Megy, K., Meidl, P., Overduin, B., Parker, A., Pritchard, B., Rios, D., Schuster, M., Slater, G., Smedley, D., Spooner, W., Spudich, G., Trevanion, S., Vilella, A., Vogel, J., White, S., Wilder, S., Zadissa, A., Birney, E., Cunningham, F., Curwen, V., Durbin, R., Fernandez-Suarez, X. M., Herrero, J., Kasprzyk, A., Proctor, G., Smith, J., Searle, S., and Flicek, P., *Nucleic Acids Res.* **37**, D690–D697 (2009).
21. Muller, J., Szklarczyk, D., Julien, P., Letunic, I., Roth, A., Kuhn, M., Powell, S., von Mering, C., Doerks, T., Jensen, L. J., and Bork, P., *Nucleic Acids Res.* (2009).

22. Bru, C., Courcelle, E., Carrere, S., Beausse, Y., Dalmar, S., and Kahn, D., *Nucleic Acids Res.* **33**, D212–D215 (2005).
23. Letunic, I., Doerks, T., and Bork, P., *Nucleic Acids Res.* **37**, D229–D232 (2009).
24. Lima, T., Auchincloss, A. H., Coudert, E., Keller, G., Michoud, K., Rivoire, C., Bulliard, V., de Castro, E., Lachaize, C., Baratin, D., Phan, I., Bougueleret, L., and Bairoch, A., *Nucleic Acids Res.* **37**, D471–D478 (2009).
25. Castagnetto, J. M., Hennessy, S. W., Roberts, V. A., Getzoff, E. D., Tainer, J. A., and Pique, M. E., *Nucleic Acids Res.* **30**, 379–382 (2002).
26. Sodhi, J. S., Bryson, K., McGuffin, L. J., Ward, J. J., Wernisch, L., and Jones, D. T., *J. Mol. Biol.* **342**, 307–320 (2004).
27. Ebert, J. C. and Altman, R. B., *Protein Sci.* **17**, 54–65 (2008).
28. Lippi, M., Passerini, A., Punta, M., Rost, B., and Frasconi, P., *Bioinformatics* **24**, 2094–2095 (2008).
29. Babor, M., Gerzon, S., Raveh, B., Sobolev, V., and Edelman, M., *Proteins* **70**, 208–217 (2008).
30. Levy, R., Edelman, M., and Sobolev, V., *Proteins* **76**, 365–374 (2009).
31. Eyal, E., Najmanovich, R., McConkey, B. J., Edelman, M., and Sobolev, V., *J. Comput. Chem.* **25**, 712–724 (2004).
32. Williams, L. and Salt, D. E., *Curr. Opin. Plant Biol.* **12**, 247–249 (2009).
33. Eide, D. J., Clark, S., Nair, T. M., Gehl, M., Gribskov, M., Guerinot, M. L., and Harper, J. F., *Genome Biol.* **6**, R77 (2005).
34. Andreini, C., Bertini, I., and Rosato, A., *Acc. Chem. Res.* (2009).
35. Mounicou, S., Szpunar, J., and Lobinski, R., *Chem. Soc. Rev.* **38**, 1119–1138 (2009).
36. Kondo, A. and Ueda, M., *Appl. Microbiol. Biotechnol.* **64**, 28–40 (2004).
37. Lu, Y., Yeung, N., Sieracki, N., and Marshall, N. M., *Nature* **460**, 855–862 (2009).