376–381 *Nucleic Acids Research, 1998, Vol. 26, No. 1*

PROMISE: a database of information on prosthetic centres and metal ions in protein active sites

K. N. Degtyarenko*, A. C. T. North, D. N. Perkins and J. B. C. Findlay

School of Biochemistry and Molecular Biology, University of Leeds, Leeds LS2 9JT, UK

Received August 21; Revised and Accepted September 24, 1997

ABSTRACT

The PROMISE (Prosthetic centres and metal ions in protein active sites) database aims to gather together comprehensive sequence, structural, functional and bibliographic information on proteins which possess prosthetic centres, with an emphasis on active site structure and function. The database is available on the World Wide Web at http://bioinf.leeds.ac.uk/promise/

BACKGROUND

Of the more than 5000 proteins of known three-dimensional structure, at least half contain metal ions or other non-protein prosthetic groups in their active sites; such prosthetic groups often themselves containing metal ions. Nature has made abundant use of the wide range of chemical properties of the elements, exemplified conspicuously by the complex coordination states of the transition metals. The presence of metal ions and other prosthetic groups confers vitally important properties on the proteins concerned, while the protein environments of the groups modulate the chemistry of the ions in subtle ways. The 'tuning' of the chemical reactivity of catalytic centres to the needs of the organism by this interplay of the protein and non-protein components is clearly highly sensitive to the local geometry of the active sites and further modified by the overall molecular structure.

The genome sequencing and other structural biology projects are accumulating ever more data, but the need to systematise and analyse the data becomes correspondingly more demanding. In addition to the rapidly growing 'primary' databanks of nucleic acid and protein sequences (corresponding already to >150 000 different proteins) and the Protein Databank of three-dimensional (3-D) structures, a number of 'secondary' databases have been derived, covering a diverse array of specialised areas. These include classification of proteins in terms of invariant active site groups (3), of 'fingerprints' of complex sequence motifs (1) or of patterns of chain fold (13,19). The PROMISE database is uniquely focused on protein active site structure and on the relationships between protein molecules and *non-protein* prosthetic centres, *combining* the relevant sequence, 3-D structural and physico-chemical information (7).

The concept of *bioinorganic motif* (14), a structural feature peculiar for metalloproteins and other complex proteins, may be useful in further discussion. Bioinorganic motifs, some of which are exemplified in Figure 1, endow a protein with function(s) possessed by neither the apoprotein nor prosthetic group alone. These motifs are quite different from sequence motifs (basically strings of text) or protein fold motifs (such as βαβ or HTH motifs). Although there are many families of proteins sharing both fold and bioinorganic motif, there are also evolutionarily unrelated but functionally analogous systems which have similar bioinorganic motif and, vice versa, there are proteins which share the same fold but have distinct active site structures. Bioinorganic motifs often possess unique spectroscopic properties which make it possible to study them relatively independently from the rest of the protein matrix.

OBJECTIVES

The PROMISE database has several objectives:

(i) To classify protein families on the basis of bioinorganic motifs.
(ii) To present a concise description of these protein families.
(iii) To provide pertinent links of each family to the various primary and secondary molecular biology databases.
(iv) To compile a comprehensive bibliography on these protein families, principally of their structural and spectroscopic properties.
(vi) To keep the database up-to-date, i.e. to supplement existing entries by new structural and bibliographic information as it becomes available.

DATABASE CONTENT AND STRUCTURE

The PROMISE database is being created in the form of hypertext (html) documents. This allows the incorporation of text, tables and graphics into database entries. Each entry forms a separate html document. Database entries follow a hierarchy: at the top level are major **groups** (such as iron–sulphur proteins or mononuclear iron proteins), which consist of **classes** (e.g. $Fe[Cys]_4$ proteins), which in turn consist of **families** (e.g. rubredoxins). For most **families**, the 3-D 'portraits' of representative **members** were created with the help of the MOLSCRIPT program (15). The classification is not rigid and some intermediate or alternative levels could easily be introduced. Moreover, since some proteins contain more than one type of prosthetic group, the classification may be best described as a 'network' rather than a 'tree'. As at 15 September 1997, PROMISE version 1.5 contained five major groups (diiron-carboxylate proteins,

*To whom correspondence should be addressed. Tel: +44 113 233 2592; Fax: +44 113 233 3167; Email: bmbknd@sgi.leeds.ac.uk

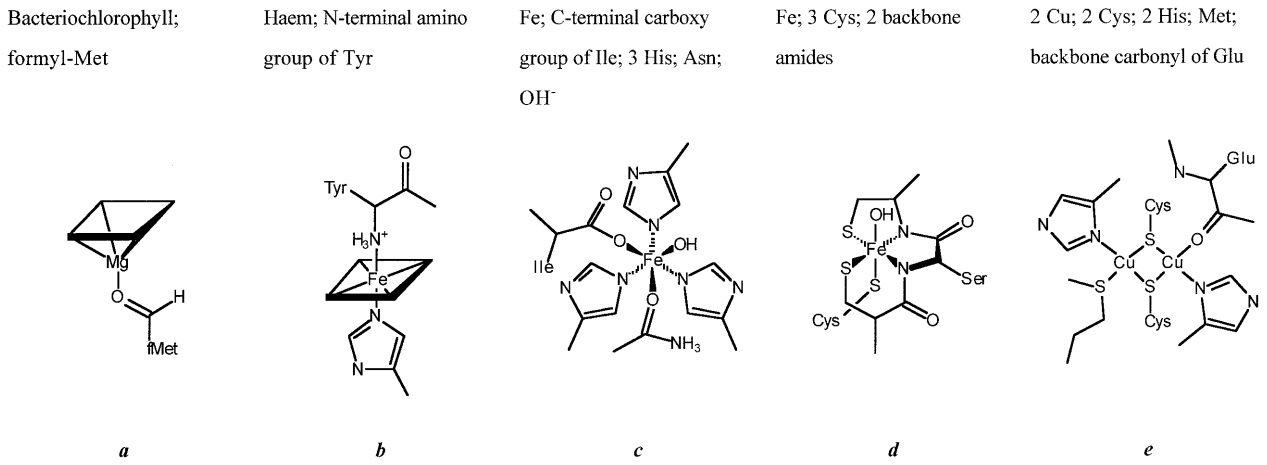


Figure 1. Examples of bioinorganic motifs with 'unusual' amino acid–metal coordination: *a*, light-harvesting complex II from *Rhodopseudomonas acidophila* (16); *b*, cytochrome *f* (17); *c*, lipoxygenase (6); *d*, nitrile hydratase (12); *e*, cytochrome *c* oxidase (23).

haem proteins, iron–sulphur proteins, mononuclear iron proteins and chlorophyll-containing proteins) comprising nine intermediate class entries and 43 protein family entries, each with an associated bibliographic entry (a total of 2552 references). Table 1 presents a full list of entries in PROMISE.

In a typical PROMISE **family** entry, a group of proteins sharing a certain bioinorganic motif is analysed and the structure of the bioinorganic motif is depicted, a concise description is presented of the proteins' functions and a short bibliography provided, and numerous hypertext links are established (Fig. 2). Via external hypertext links, organised in several tables, the relevant entries from a variety of on-line molecular biology databases are accessible (Table 2). Via internal hypertext links, the user may find the way to other relevant PROMISE entries, both of higher (e.g. **class**) or lower hierarchy (e.g. bibliography on structural studies of proteins from this family or MOLSCRIPT images of a representative **member** of the family). In contrast to such derived databases as SCOP (13) and CATH (19), PROMISE thus categorises proteins on the basis of their active/binding site geometry and environment, rather than on the basis of their chain folds.

PROMISE contains two types of **bibliography** entries: 'structural studies' and 'reviews'. As the names suggest, the former entries include references to original papers on crystallographic and spectroscopic studies of proteins from the given family whereas the latter contain citations of reviews and other secondary literature sources. Where available, the hot-links to *Entrez*-MEDLINE (4) abstracts and other on-line bibliographic sources (e.g. the *Current Biology* publications) are included.

Every **family** and **bibliography** entry is revised at least bimonthly; new entries and updates are released and the database is re-indexed weekly. Currently, the update procedure includes scanning a number of on-line databases (Table 2) and is performed 'manually'. With the expected growth of the database, the update will become more time-consuming so we intend to implement a semi-automatic search routine to facilitate this work. It is important to stress that, in contrast to computer-generated derived databases, a great deal of human involvement is necessary for the creation and updating of PROMISE entries.

AVAILABILITY AND SEARCH OF DATABASE

PROMISE is available on the World Wide Web from URL: http://bioinf.leeds.ac.uk/promise/ . Use of Netscape version 2.0 or higher is recommended. To provide simple and fast textual searching through the whole PROMISE database, the SRS browser (8) has been installed. The user can thereby search simultaneously for the (combination of maximum four) keywords through the whole entries (AllText) or in one of the following fields:

- ID—Unique identifier (name of the entry);
- Definition—Short title;
- DatabaseLinks—The external and internal hypertext links that are organised into tables;
- Comment—A concise description of the proteins' functions;
- Reference—Bibliography information.

NOVELTY

The previous lack of such a database is largely due to the complexity of the problem and to the difficulty of extracting and interpreting the relevant information. The amino acid sequence alone clearly does not provide sufficient information about both the nature and co-ordination mode of a protein's prosthetic centre. Biosynthesis of such prosthetic groups as tetrapyrrroles, flavins or the iron–molybdenum cofactor (FeMoco) is carried out by a number of enzymes and does not depend directly on the folding of apoproteins. Apart from the 'classic' amino acid ligands that commonly form bioinorganic motifs (side-chains of cysteine, histidine, tyrosine, methionine, glutamic and aspartic acids), various 'novel' ligands have been reported during the last few years (Fig. 1). While comparisons of sequence data between species can assist identification of the amino acids that are involved in the centre and can in some cases hint at the possible function of 'novel' proteins (i.e. proteins whose sequences have been translated from gene sequences and whose functions have not been determined biochemically) we know that homologous proteins sometimes possess different prosthetic groups (e.g. mammalian and bacterial haem catalases). If the 3-D structure of

Table 1. Contents of PROMISE version 1.5

PROMISE groups, classes and families	Bibliography files: Structural studies[a]	Bibliography files: Reviews[a]	3-D example
Diiron-carboxylate proteins			
Class I: Ribonucleotide reductase R2-type proteins	91	42	1RIB
Class II: Ferrooxidase			
Bacterioferritin (cytochrome b_1; cytochrome b_{557})	12	} 44	1BCF, 1BFR
Ferritin	52		
Rubrerythrin	9		
Class III: Haemerythrin family	41		1HMO
Class IV: Purple acid phosphatase	14		1KBP
Haem proteins			
Catalases	38		1CAF
Cytochrome *c* oxidase	259	49	1OCC
Cytochromes			
Cytochromes *b*			
Cytochrome b_5 family	59		3B5C
Soluble cytochrome b_{562}	16		256B
Cytochromes *c*			
Class I cytochromes *c*	201		1CRY
Class II cytochromes *c*	33		1BBH
Class III cytochromes *c*	66		2CY3
Class IV cytochromes *c*	19		1PRCC
Cytochromes c_1	32		
Cytochromes *f*	7		1CTM
Globins	527	60	1BBB, 1MBO
Haem peroxidases			
Animal haem peroxidases	19	17	1MHL
Fungal, plant and bacterial haem peroxidases	149	35	2CYP
Haem-thiolate proteins			
Chloroperoxidase	36		1CPO
P450 proteins	85	99	3CPP
Iron-sulphur proteins		24	
$Fe(Cys)_4$ proteins			
Desulforedoxin-type $Fe(Cys)_4$ proteins	8		
Rubredoxin-type $Fe(Cys)_4$ proteins	30		1CAD
Fe_2S_2 proteins			
Adrenodoxin-type ferredoxins	24		1PUT
Plant-type ferredoxins	60	12	1FRR
Rieske iron-sulphur proteins	37		1RIE
Fe_4S_4 / Fe_3S_4 proteins			
Aconitase family	36	25	7ACN
Bacterial-type mono-, di- and polycluster ferredoxins	108	6	1FXR
Endonuclease III family	5		2ABK
High potential iron proteins (HiPIPs)	44		1ISU
Nickel-iron hydrogenase	28		1FRV
Nitrogenase component I (MoFe protein)	48	} 28	1MIN
Nitrogenase component II (Fe protein)	19		1NIP
Sirohaem-Fe_4S_4 enzymes	25		1GEP
Trimethylamine dehydrogenase	11		2TMD
Fe_6S_6 proteins	8		
Mononuclear iron proteins			
Extradiol aromatic-ring cleavage monooxygenases	19	} 9	1HAN
Intradiol aromatic-ring cleavage monooxygenases	28		2PCD
Isopenicillin N synthase	10	10	
Lipoxygenases	54	19	2SBL
Chlorophyll-containing proteins			
Bacteriochlorophyll *a* protein	18		3BCL
Light-harvesting complex II of purple bacteria	43	13	1LGH
Photosynthetic reaction centre of purple bacteria	124	14	1PRC
Total	2552	506	

[a]The numbers in the columns correspond to the numbers of literature references in the associated bibliographic entries.

a representative member of a certain protein family is available, homology modelling may be used to derive the structures of other family members, since there is a high probability that members share the same *overall* 3-D fold; nevertheless, *details* of their conformation may differ, giving rise to subtle differences in chemical properties whose structural basis may remain obscure until a high-resolution X-ray or NMR structure becomes available. However, experimental results from other, mainly spectroscopic, techniques also yield substantial information about prosthetic group/active site structures (even in the absence of sequence data) and it is therefore essential that a database on prosthetic centres and metal ions in protein active sites should contain such spectroscopic information.

To illustrate the extent of the gap which exists in the field of protein informatics, let us consider the problem of protein structure prediction from amino-acid sequence. Despite great and continuing efforts throughout the last three decades, this problem remains unsolved in the general case. Although homology modelling of monodomain globular proteins has become an almost routine task, a number of problems appear still to be beyond the conventional '*in silico* protein folding' methodology. In particular, the following vital features of protein molecules are largely ignored by protein structure 'predictors':

- disulphide bonds and redox-active thiol groups
- membrane-bound domains
- modified amino acid residues
- non-standard amino-acid residues
- non-standard (e.g. locally distorted) secondary structure
- prosthetic centres.

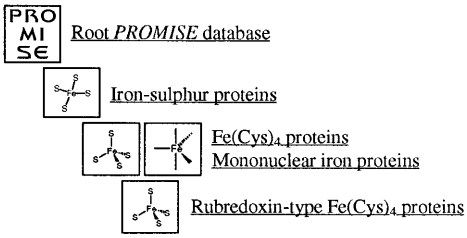
At least some of these features already are (or will be) incorporated in PROMISE. This does not mean that the problem is solved; rather, PROMISE makes a researcher aware of such features and provides him/her with known examples.

APPLICATIONS

Our goal is to create both *comprehensive* and *comprehensible* information source on prosthetic centres and metal ions in protein active sites. Our intended areas of application include biological inorganic chemistry, biophysics, bioenergetics, crystallography, spectroscopy and molecular pharmacology, as well as biological education; there are clearly important areas of industrial application in the pharmaceutical, agro-chemical, food and biotechnology industries. From the outset, we have therefore aimed at world-wide access to this resource and have designed the database to be interrogated via a World Wide Web front end, making extensive use of hypertext links both within PROMISE itself and to external databases. By providing such a familiar mode of access and because of its interdisciplinary approach, we hope that PROMISE will become a 'meeting place' for specialists from these different fields, as well as a powerful facility for novices.

ACKNOWLEDGEMENTS

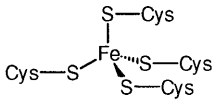
This work has been supported by Astra Charnwood, Glaxo Wellcome, Pfizer, Zeneca and the BBSRC under the U.K. DTI/BBSRC Biotechnology LINK Programme.



Created: 28 May 1996
Last modified: 2 September 1997

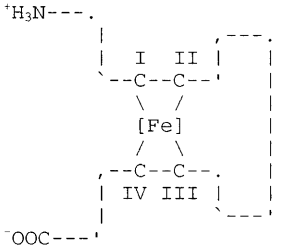
Rubredoxin-type $Fe(Cys)_4$ proteins

- Prosthetic group features
- Rubredoxins in motif databases
- Rubredoxins in alignment databases
- Rubredoxins in 3-D databases
- References

Iron-sulphur cluster	Formal oxidation states
 $Fe(S^{\gamma}_{Cys})_4$	$Fe(Cys)_4^{1-}$; $Fe(Cys)_4^{2-}$

Rubredoxin is a low molecular weight iron-containing bacterial protein involved in electron transfer, sometimes replacing ferredoxin as an electron carrier [1]. Rubrerythrin is a fusion protein containing a C-terminal domain homologous to rubredoxin [2].

The 3-D structures of a number of rubredoxins have been solved [3, 4]. The fold belongs to the α+β class, with 2 α-helices and 2–3 β-strands. Its active site contains an iron ion which is co-ordinated by the sulphurs of four conserved cysteine residues forming an almost regular tetrahedron [4]. The conserved cysteines reside on two loops, which are the most conserved regions of the protein:



In addition, a ring of acidic residues in the proximity of the $[Fe(Cys)_4]$ centre is also well-conserved [4].

The 3-D structure of rubrerythrin has been solved [5]. The structure reveals a tetramer of two-domain subunits. In each monomer, the N-terminal 146 residues form a four-helix bundle containing the diiron-oxo site, and the C-terminal 45 residues form a rubredoxin-like FeS_4 domain.

■

Figure 2. (Above and opposite) A sample entry from PROMISE. Underlined text indicates the hypertext links.

REFERENCES

1 Attwood,T.K., Beck,M.E., Bleasby,A.J., Degtyarenko,K.N., Michie,A.D. and Parry-Smith,D.J. (1997) *Nucleic Acids Res.*, **25**, 212–217 [see also this issue (1998) *Nucleic Acids Res.* **26**, 304–308].
2 Bairoch,A. (1996) *Nucleic Acids Res.*, **24**, 221–222.
3 Bairoch,A., Bucher,P. and Hofmann,K. (1997) *Nucleic Acids Res.*, **25**, 217–221.
4 Benson,D.A., Boguski,M.S., Lipman,D.J. and Ostell,J. (1997) *Nucleic Acids Res.*, **25**, 1–6 [see also this issue (1998) *Nucleic Acids Res.* **26**, 1–7].
5 Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F.,Jr, Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) *J. Mol. Biol.*, **112**, 535–542.
6 Boyington,J.C., Gaffney,B.J. and Amzel,L.M. (1993) *Science*, **260**, 1482–1486.
7 Degtyarenko,K.N., North,A.C.T. and Findlay,J.B.C. (1997) *Protein Engng.*, **10**, 183–186.
8 Etzold,T., Ulyanov,A. and Argos,P. (1996) *Methods Enzymol.*, **266**, 114–128.
9 Hendrickson,W. and Wüthrich,K. (eds) (1991–1996) *Macromolecular Structures.* Current Biology Ltd, London.
10 Hendlich,M., Rippmann,F. and Barnickel,G., in preparation.
11 Henikoff,J.G., Pietrokovski,S. and Henikoff,S. (1997) *Nucleic Acids Res.*, **25**, 222–225 [see also this issue (1998) *Nucleic Acids Res.* **26**, 309–312].
12 Huang,W., Jia,J., Cummings,J., Nelson,M., Schneider,G. and Lindqvist,Y. (1997) *Structure*, **5**, 691–699.

Rubredoxin-type Fe(Cys)$_4$ proteins in motif databases

PRINTS ID	PRINTS AC	PROSITE/BLOCKS ID	PROSITE AC	BLOCKS AC
RUBREDOXIN	PR00163	RUBREDOXIN	PS00202	BL00202

■

Rubredoxin-type Fe(Cys)$_4$ proteins in alignment databases

Protein Superfamily	Protein Homology Domain	Pfam	LPFC 3-D alignment
0033.0; rubredoxin 0034.0; *Pseudomonas* rubredoxin I 0035.0; *Pseudomonas* rubredoxin II	00142; rubredoxin	PF00301; rubredoxin	rub

■

Rubredoxin-type Fe(Cys)$_4$ proteins in 3-D databases

All rubredoxins contain single iron ion (see Figure 1CAD) except for * (Zn-substituted) containing zinc ion.

PDB	scop	BSM	RELI Base	Header	MACROMOLECULAR STRUCTURES
1caa	1caa	1caa	1caa	Rubredoxin (oxidised); *Pyrococcus furiosus*	MMS93186
1cad	1cad	1cad	1cad	Rubredoxin (reduced); *Pyrococcus furiosus*	MMS93186
1irn	1irn	1irn	1irn	Rubredoxin (oxidised); *Clostridium pasteurianum*	–
1iro	1iro	1iro	1iro	Rubredoxin (oxidised); *Clostridium pasteurianum*	–
1rdg	1rdg	1rdg	1rdg	Rubredoxin; *Desulfovibrio gigas*	–
1zrp*	1zrp*	1zrp*	1zrp*	Rubredoxin (Zn-substituted); *Pyrococcus furiosus*	MMS93185*
4rxn	4rxn	4rxn	4rxn	Rubredoxin (oxidised) (unconstrained model); *Clostridium pasteurianum*	–
5rxn	5rxn	5rxn	5rxn	Rubredoxin (oxidised) (constrained model); *Clostridium pasteurianum*	–
6rxn	6rxn	6rxn	6rxn	Rubredoxin; *Desulfovibrio desulfuricans*, strain 27774	MMS91112
7rxn	7rxn	7rxn	7rxn	Rubredoxin (complex with sulphate); *Desulfovibrio vulgaris*	–
8rxn	8rxn	8rxn	8rxn	Rubredoxin (complex with sulphate); *Desulfovibrio vulgaris*	–
–	–	–	–	Rubredoxin (racemic); *Desulfovibrio desulfuricans*	MMS94233

[1] Macromolecular Structures abstract. Full text is available to BioMedNet Members

■

References

1. Lee, W.Y., Brune, D.C., LoBrutto, R. and Blankenship, R.E. (1995) Isolation, characterization, and primary structure of rubredoxin from the photosynthetic bacterium, *Heliobacillus mobilis*. *Arch. Biochem. Biophys.* **318**, 80–88.
2. Van Beeumen, J.J., Van Driessche, G., Liu, M.Y. and LeGall, J. (1991) The primary structure of rubrerythrin, a protein with inorganic pyrophosphatase activity from *Desulfovibrio vulgaris*. Comparison with hemerythrin and rubredoxin. *J. Biol. Chem.* **266**, 20645–20653.
3. Day, M.W., Hsu, B.T., Joshua-Tor, L., Park, J.B., Zhou, Z.H., Adams, M.W.W. and Rees, D.C. (1992) X-ray crystal structures of the oxidized and reduced forms of the rubredoxin from the marine hyperthermophilic archaebacterium *Pyrococcus furiosus*. *Protein Science* **1**, 1494–1507.
4. Frey, M., Sieker, L., Payan, F., Haser, R., Bruschi, M., Pepe, G. and LeGall, J. (1987) Rubredoxin from *Desulfovibrio gigas*. A molecular model of the oxidized form at 1.4 Å resolution. *J. Mol. Biol.* **197**, 525–441.
5. deMaré, F., Kurtz, D.M., Jr. and Nordlund, P. (1996) The structure of *Desulfovibrio vulgaris* rubrerythrin reveals a unique combination of rubredoxin-like FeS_4 and ferritin-like diiron domains. *Nature Struct. Biol.* **3**, 539–546.

■

Bibliography on structural studies of rubredoxins

Table 2. On-line databases linked to PROMISE

Database	WWW address	Reference
Function		
ENZYME	http://expasy.hcuge.ch/	[2]
LIGAND	http://www.genome.ad.jp/	[22]
Protein motifs		
PRINTS	http://bioinf.leeds.ac.uk/prints/	[1]
PROSITE	http://expasy.hcuge.ch/	[3]
BLOCKS	http://www.blocks.fhcrc.org/	[11]
Multiple alignments		
PROTFAM	http://www.mips.biochem.mpg.de/	[18]
Pfam	http://www.sanger.ac.uk/Software/Pfam/	[21]
LPFC	http://www-camis.stanford.edu/projects/helix/LPFC/	[20]
3-D structures		
PDB (EBI mirror)	http://www2.ebi.ac.uk/pdb/	[5]
SCOP	http://scop.mrc-lmb.cam.ac.uk/scop/	[13]
CATH	http://www.biochem.ucl.ac.uk/bsm/cath/	[19]
RELIBase	http://www2.ebi.ac.uk:8081/home.html	[10]
Macromolecular Structures	http://biomednet.com/library/mms	[9]
Bibliography		
Entrez-MEDLINE	http://www.ncbi.nlm.nih.gov/Entrez/medline.html	[4]

13 Hubbard,T.J.P., Murzin,A.G., Brenner,S.E. and Chothia,C. (1997) *Nucleic Acids Res.*, **25**, 236–239.
14 Karlin,K.D. (1993) *Science*, **261**, 701–708.
15 Kraulis,P.J. (1991) *J. Appl. Crystallogr.*, **24**, 946–950.
16 McDermott,G., Prince,S.M., Freer,A.A., Hawthornthwaite-Lawless,A.M., Papiz,M.Z., Cogdell,R.J. and Isaacs,N.W. (1995) *Nature*, **374**, 517–521.
17 Martinez,S.E., Huang,D., Szczepaniak,A., Cramer,W.A. and Smith,J.L. (1994) *Structure*, **2**, 95–105.
18 Mewes,H.W., Albermann,K., Heumann,K., Liebl,S. and Pfeiffer,F. (1997) *Nucleic Acids Res.*, **25**, 28–30 [see also this issue (1998) *Nucleic Acids Res.* **26**, 33–37].
19 Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) *Structure* **5**, 1093–1108.
20 Schmidt,R., Gerstein,M. and Altman,R.B. (1997) *Protein Sci.*, **6**, 246–248.
21 Sonnhammer,E.L.L., Eddy,S.R. and Durbin,R. (1997) *Proteins*, **28**, 405–420.
22 Suyama,M., Ogiwara,A., Nishioka,T. and Oda,J. (1993) *Comp. Appl. Biosci.*, **9**, 9–15.
23 Tsukihara,T., Aoyama,H., Yamashita,E., Tomizaki,T., Yamaguchi,H., Shinzawa-Itoh,K., Nakashima,R., Yaono,R. and Yoshikawa,S. (1996) *Science*, **272**, 1136–1144.