doi:10.1016/j.jmb.2009.02.052 J. Mol. Biol. (2009) 388, 356–380

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Structural Analysis of Metal Sites in Proteins: Non-heme Iron Sites as a Case Study

Claudia Andreini[1,2], Ivano Bertini[1,2], Gabriele Cavallaro[1,2], Rafael J. Najmanovich[3] and Janet M. Thornton[3]*

[1]*Magnetic Resonance Center (CERM) – University of Florence, Via L. Sacconi 6, 50019 Sesto Fiorentino, Italy*

[2]*Department of Chemistry – University of Florence, Via della Lastruccia 3, 50019 Sesto Fiorentino, Italy*

[3]*EMBL - European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK*

Received 27 November 2008; received in revised form 19 February 2009; accepted 19 February 2009
Available online 2 March 2009

Edited by M. Guss

In metalloproteins, the protein environment modulates metal properties to achieve the required goal, which can be protein stabilization or function. The analysis of metal sites at the atomic level of detail provided by protein structures can thus be of benefit in functional and evolutionary studies of proteins. In this work, we propose a structural bioinformatics approach to the study of metalloproteins based on structural templates of metal sites that include the PDB coordinates of protein residues forming the first and the second coordination sphere of the metal. We have applied this approach to non-heme iron sites, which have been analyzed at various levels. Templates of sites located in different protein domains have been compared, showing that similar sites can be found in unrelated proteins as the result of convergent evolution. Templates of sites located in proteins of a large superfamily have been compared, showing possible mechanisms of divergent evolution of proteins to achieve different functions. Furthermore, template comparisons have been used to predict the function of uncharacterized proteins, showing that similarity searches focused on metal sites can be advantageously combined with typical whole-domain comparisons. Structural templates of metal sites, finally, may constitute the basis for a systematic classification of metalloproteins in databases.

Keywords: metalloprotein; metal; iron; protein structure; structural bioinformatics

Introduction

A significant fraction of known proteins require metal ions to carry out their biological function, and are thus termed metalloproteins. The reasons for this requirement are manifold. Metals can play a structural role, when they are needed to stabilize the tertiary and/or the quaternary structure of the protein, or can serve a more directly functional role, when they are involved in the catalytic mechanism of enzymes or constitute electron transfer centers. Furthermore, metals are essential to bind oxygen in oxygen carriers such as hemoglobin, and can function as regulatory elements in signaling proteins such as calmodulin.[1,2] To accomplish this diversity of functions, evolution has selected different metals and has engineered a broad variety of protein sites, where the physico-chemical properties of the metals are modulated by the local properties of the protein matrix to achieve the required functionality.[3] Therefore, metal sites in proteins can be regarded as structural and functional units composed of the metal and the protein environment, whose properties as a whole are optimized for function.

Despite the crucial importance of metals for many proteins, the attention given to metalloproteins by bioinformatics has been limited, and general information sources such as MDB[4] and PROMISE[5] have been discontinued in the last few years. At present, therefore, there is a lack of up-to-date databases centered on metalloproteins, as well as of tools for the structural and functional analysis of metal sites in proteins. This situation can be due, at least in part, to the difficulty in establishing formal, consistent criteria to describe metal sites in proteins, which are

*Corresponding author. E-mail address: thornton@ebi.ac.uk.

Present address: R. J. Najmanovich, Département de Biochimie, Université de Sherbrooke, Sherbrooke J1H 5N4, Canada.

0022-2836/$ - see front matter © 2009 Published by Elsevier Ltd.

required as the basis for their comparative analysis and classification. Several descriptions are possible, ranging from one-dimensional patterns to three-dimensional templates.[6] For instance, the ProFunc server[7] employs templates that contain the spatial coordinates of the metal ligands (first coordination sphere). However, given the considerations above, this representation may not be sufficiently accurate to describe the protein environment of the metal, in that its functionality is known to be affected also by protein residues interacting with the metal ligands (second coordination sphere).[8,9]

In this work, we extend the three-dimensional representation of metal sites in proteins used in ProFunc by defining structural templates that include (i) the metal, (ii) the first coordination sphere, and (iii) the second coordination sphere. These templates are extracted automatically from Protein Data Bank (PDB)[10] structures using a tool that we have developed for this purpose. Using non-heme iron as a case study, we show that this representation can be advantageously used to analyze metal sites in proteins from several points of view. First, we classify iron sites based on CATH[11] and SCOP[12] databases, identifying the protein structures with which these sites are most often associated. Second, we examine the amino acid composition of iron sites, retrieving trends and regularities associated with different iron functions. Third, we carry out pairwise comparisons of iron sites by structural alignment of the templates. These comparisons reveal common structural motifs in structurally and evolutionarily unrelated proteins, and highlight fine differences in the active sites of proteins that have similar structures but different functions. In both cases, it is possible to obtain hints as to the possible mechanisms of evolution of iron sites and of the proteins that contain them. Also, these comparisons provide functional hints for iron proteins with unknown function by detecting similarities between their sites and sites of iron proteins with known function.

In perspective, this description can provide the basis for a systematic, structure-based classification of all known metal sites, which can be useful as an organized source of information.

Results and Discussion

Grouping and annotation of non-heme iron sites

The approach described here can be applied to the analysis of all metal sites in proteins, including the comparison of sites binding different metals. To illustrate the potential of the approach, we chose to apply it to non-heme iron sites for three main reasons. First is the biological importance of iron, which is the dominant redox metal in biological systems, occurring most commonly as Fe(II) ($[Ar]3d^6$ electronic configuration) and Fe(III) ($[Ar]3d^5$ electronic configuration), although other oxidation states can occur.[13,14] As such, it is essential to virtually all life, with the possible exceptions of certain microbial parasites such as *Borrelia burgdorferi*.[15] Second, non-heme iron displays versatile coordination properties and is found in a large variety of active centers of proteins that have different functions.[16] In iron–sulfur proteins, for instance, two or more iron ions are coordinated with tetrahedral geometry by thiol ligands from cysteine residues and bridging sulfide ions, forming so-called iron–sulfur clusters with well known geometry (e.g., see Supplementary Data Fig. S1). In other proteins, iron is preferably coordinated by six ligands with octahedral geometry, or by five ligands with trigonal bipyramidal geometry. In these cases, the coordination of iron can involve open coordination sites for binding of substrates and/or exogenous ligands, as well as significant distortions from symmetrical geometries (e.g., see Supplementary Data Fig. S1). Third, a wealth of structural information is available for iron proteins. In December 2007, PDB contained 1274 experimental structures of proteins having a total of 2034 non-heme iron sites (referred to here as Fe-sites). We applied our approach to the analysis of these 2034 Fe-sites.

As the first step of our approach, we grouped together Fe-sites located in protein domains that are classified in the same CATH or SCOP superfamily, and we selected a single representative Fe-site for each group. CATH and SCOP provide hierarchical classification of protein domain structures, in which superfamilies comprise protein domains that may have low levels of sequence identity but have structural and functional similarity that suggests a common evolutionary origin.[11,12] This grouping allowed us to reduce the size of the data set from over 2000 to only 86 Fe-sites (Table 1). This large reduction highlights the redundancy of PDB with respect to CATH and SCOP superfamilies,[17] some of which, such as the 4Fe-4S (CATH code 3.30.70.20) and the 2Fe-2S ferredoxins (CATH code 3.10.20.30), include more than 100 PDB structures. Because more than one Fe-site can be present in the same domain (e.g., see Supplementary Data Fig. S2), the 86 representative Fe-sites are found in 69 different superfamilies (referred to here as Fe-superfamilies) and in nine other domains that could not be included in any CATH or SCOP superfamily.

In Table 1, we have annotated Fe-sites with respect to their type, i.e. iron–sulfur (47 sites), mononuclear (27 sites), dinuclear (10 sites), and polynuclear (1 site). We annotated as iron-sulfur + dinuclear the complex site of iron hydrogenases, which is termed H-cluster and consists of a Fe_4S_4 cluster bridged by a Cys thiol to a di-iron center.[18] When available from the literature, Fe-sites were assigned one general function among electron transfer (26 sites), redox catalysis (26 sites), non-redox catalysis (7 sites), sensing (3 sites) and structural (3 sites). Electron transfer is most commonly carried out by iron–sulfur centers, whereas various types of Fe-sites can be involved in catalysis (Fig. 1). Other selected properties of representative Fe-sites, including oxidation and spin state (when available from the literature),

Table 1. The groups of Fe-sites resulting from structure-based classification

Domain	Representative site	PDB description	Type	Function
1.10.1060.10	1gte-SF4_1027	Dihydropyrimidine dehydrogenase	FeS	Electron transfer
1.10.150.120	1n62-FES_3907	CO dehydrogenase	FeS	Electron transfer
1.10.1670.10	1kg2-SF4_300	Adenine glycosylase (MutY)	FeS	Structural
1.10.3140.10 + 1.20.140.10	1u8v-SF4_491	4-Hydroxybutyryl-CoA dehydratase	FeS	Non-redox catalysis
1.10.569.10	1aor-FE_1	Aldehyde-ferredoxin oxidoreductase	Mononuclear	Structural
1.10.569.10 + 1.10.599.10	1b25-SF4_700	Aldehyde-ferredoxin oxidoreductase	FeS	Electron transfer
1.10.620.20	1mxr-FE_1001_FE_ 1002	Ribonucleotide reductase R2	Dinuclear	Redox catalysis
1.10.645.10	1wui-NFC_1004	Ni-Fe hydrogenase	Dinuclear	Redox catalysis
1.10.800.10	1ltz-FE_400	Phenylalanine hydroxylase	Mononuclear	Redox catalysis
1.20.1090.10	1o2d-FE_900	Alcohol dehydrogenase	Mononuclear	Non-redox catalysis
1.20.1130.10	1jb0-SF4_3001	Photosynthetic reaction center	FeS	Electron transfer
1.20.120.50	2mhr-FEO_119	Hemerythrin	Dinuclear	Sensing
1.20.1260.10	1yuz-FE_301_FE2_302	Nigerythrin	Dinuclear	Redox catalysis
1.20.1270.20	1gnl-SF4_1544	Hybrid cluster protein	FeS	Unknown
1.20.1270.30_1	1oao-SF4_1675	CO dehydrogenase	FeS	Electron transfer
1.20.1270.30_2	1su8-SF4_637	CO dehydrogenase	FeS	Electron transfer
1.20.245.10	1f8n-FE2_840	Lipoxygenase	Mononuclear	Redox catalysis
1.20.58.220	1sum-FE_202_ FE_203_FE_204	PhoU homologue	Polynuclear	Unknown
1.20.85.10	2j8c-FE_1307	Photosynthetic reaction center	Mononuclear	Electron transfer
1.20.910.10	1rcw-FE_401_FE_402	Chlamydia protein CADD	Dinuclear	Unknown
2.10.240.10	1ep3-FES_503	Dihydroorotate dehydrogenase	FeS	Electron transfer
2.102.10.10	1jm1-FES_501	Rieske protein	FeS	Electron transfer
2.20.28.10	2dsx-FE_501	Rubredoxin	Mononuclear	Electron transfer
2.60.120.10	2b5 h-FE_501	Cysteine dioxygenase	Mononuclear	Redox catalysis
2.60.120.330	1odm-FE2_1336	Isopenicillin N synthase	Mononuclear	Redox catalysis
2.60.130.10	1dmh-FE_400	Catechol 1,2-dioxygenase	Mononuclear	Redox catalysis
2.60.40.730_1	1vzi-FE2_1131	Superoxide reductase	Mononuclear	Electron transfer
2.60.40.730_2	1vzi-FE2_1132	Superoxide reductase	Mononuclear	Redox catalysis
3.10.180.10	1kw3-FE2_301	2,3-Dihydroxybiphenyl dioxygenase	Mononuclear	Redox catalysis
3.10.20.30_1	1n62-FES_3908	CO dehydrogenase	FeS	Electron transfer
3.10.20.30_2	1feh-SF4_584	Fe-only hydrogenase	FeS	Electron transfer
3.20.20.140	1k6w-FE_501	Cytosine deaminase	Mononuclear	Non-redox catalysis
3.20.20.70_1	1o94-SF4_1732	Trimethylamine dehydrogenase	FeS	Electron transfer
3.20.20.70_2	1r30-FES_402	Biotin synthase	FeS	Redox catalysis
3.20.20.70_3	1r30-SF4_401	Biotin synthase	FeS	Redox catalysis
3.30.413.10	1aop-SF4_575	Sulfite reductase	FeS	Redox catalysis
3.30.428.10	1gup-FE_351	Nucleotidyl transferase	Mononuclear	Structural
3.30.499.10	2b3y-SF4_1000	Aconitase	FeS	Non-redox catalysis
3.30.70.20	1iqz-SF4_82	Ferredoxin	FeS	Electron transfer
3.40.190.10_1	1ryo-FE_329	Transferrin	Mononuclear	No function
3.40.190.10_2	1si0-FE_320	Ferric iron-binding protein	Mononuclear	No function
3.40.30.10_1	1m2a-FES_201	Thioredoxin-like ferredoxin	FeS	Electron transfer
3.40.30.10_2	2ht9-FES_1001	Glutaredoxin	FeS	Unknown
3.40.470.10	1ui0-SF4_210	Uracil-DNA glycosylase	FeS	Unknown
3.40.50.1400_1	2hrc-FES_1501	Ferrochelatase	FeS	Sensing
3.40.50.1400_2	2h1w-FE2_900	Ferrochelatase	Mononuclear	No function
3.40.50.1780 + 3.40.950.10 + 4.10.260.20	1hfe-SF4_424_FE2_ 426_FE2_427	Fe-only hydrogenase	FeS + dinuclear	Redox catalysis
3.40.50.1980_1	1m1n-CFN_6496	Nitrogenase	FeS	Redox catalysis
3.40.50.1980_2	1m1n-CLF_6498	Nitrogenase	FeS	Electron transfer
3.40.50.2030	1su8-NFS_639	CO dehydrogenase	FeS	Redox catalysis
3.40.50.300	1cp2-SF4_290	Nitrogenase	FeS	Electron transfer
3.40.50.700	1wui-SF4_1001	Ni-Fe hydrogenase	FeS	Electron transfer
3.40.50.970	2c42-SF4_2235	Pyruvate- ferredoxin oxidoreductase	FeS	Electron transfer
3.40.640.10	1ohv-FE_800	4-Aminobutyrate aminotransferase	FeS	Unknown
3.40.830.10	1b4u-FE_501	Protocatechuate 4,5-dioxygenase	Mononuclear	Redox catalysis
3.40.970.20	1oao-SF4_1730	CO dehydrogenase	FeS	Redox catalysis
3.60.130.10	1ds1-FE2_341	Clavaminate synthase	Mononuclear	Redox catalysis
3.60.15.10	2obw-FE_252_FE_253	Glyoxalase II	Dinuclear	Redox catalysis
3.60.20.10	1ao0-SF4_466	Glutamine phosphoribosyl-pyrophosphate amidotransferase	FeS	Sensing
3.60.21.10	1ute-FEO_501	Purple acid phosphatase	Dinuclear	Non-redox catalysis
3.60.21.30	1t71-FE_301_FE_302	Mycoplasma phosphatase	Dinuclear	Unknown
3.90.330.10	2cz1-FE_300	Nitrile hydratase	Mononuclear	Non-redox catalysis
3.90.380.10	2bmo-FE_1441	Nitrobenzene dioxygenase	Mononuclear	Redox catalysis
3.90.45.10	1lm4-FE_603	Peptide deformylase	Mononuclear	Non-redox catalysis
3.90.460.10	1dj7-SF4_120	Ferredoxin-thioredoxin reductase	FeS	Redox catalysis
4.10.480.10_1	1wui-SF4_1002	Ni-Fe hydrogenase	FeS	Electron transfer

Table 1 (*continued*)

Domain	Representative site	PDB description	Type	Function
4.10.480.10_2	1wui-F3S_1003	Ni-Fe hydrogenase	FeS	Electron transfer
4.10.490.10	1iua-SF4_84	High potential iron-sulfur protein	FeS	Electron transfer
a.2.11 + d.44.1	1bsm-FE_202	Superoxide dismutase	Mononuclear	Redox catalysis
a.211.1	2pq7-FE_1_FE_2	Hydrolase	Dinuclear	Redox catalysis
c.55.1_1	1hux-SF4_290	2-Hydroxyglutaryl-CoA dehydratase	FeS	Unknown
c.55.1_2	2ivp-FE2_1326	O-Sialoglycoprotein endopeptidase	Dinuclear	Unknown
c.66.1	1uwv-SF4_1432	Ribosomal RNA 5-methyluridine methyltransferase	FeS	Unknown
c.81.1	1kqf-SF4_800	Formate dehydrogenase	FeS	Electron transfer
d.145.1	1rm6-SF4_910	4-Hydroxybenzoyl-CoA reductase	FeS	Electron transfer
e.59.1_1	2fiy-FE_401	Pseudomonas FdhE protein	Mononuclear	Unknown
e.59.1_2	2fiy-FE_405	Pseudomonas FdhE protein	Mononuclear	Unknown
Unclassified_1	1x0g-FES_500	Iron-sulfur cluster assembly protein IscA	FeS	No function
Unclassified_2	2hu9-FES_131	Archaeoglobus CopZ	FeS	Unknown
Unclassified_3	2biw-FE_1492	Carotenoid oxygenase	Mononuclear	Redox catalysis
Unclassified_4	2cb2-FE_1310	Sulfur oxygenase reductase	Mononuclear	Redox catalysis
Unclassified_5	2fug-SF4_8	NADH dehydrogenase	FeS	Electron transfer
Unclassified_6	2goy-SF4_301	Adenosine phosphosulfate reductase	FeS	Unknown
Unclassified_7	2jh3-SF4_650	Deinococcus DR2441 protein	FeS	Unknown
Unclassified_8	2qh7-FES_500	Mitochondrial mitoNEET protein	FeS	Unknown
Unclassified_9	2z1d-SF4_501	Ni-Fe hydrogenase maturation protein HypD	FeS	Unknown

The table reports: (i) the CATH (or SCOP) domain(s) where the Fe-site is located; numerical suffixes identify distinct sites found in the same domain; (ii) the representative Fe-site of the group, identified by PDB code, residue name and number; note that Fe atoms of dinuclear and polynuclear sites can be identified by different residues in PDB structures; (iii) the description of the representative PDB structure; (iv) the type of Fe-site; and (v) the general function of the Fe-site.

iron–ligand distances and coordination geometry are collected in Supplementary Data Table S3.

Distribution of non-heme iron sites across protein domain structures

The structure-based grouping of Fe-sites described above allowed us to analyze their association with different types of protein domain structures, based upon the CATH classification. In the CATH classification,[11] superfamilies (see above) are grouped into folds based on the orientations and connectivity of the secondary structure elements of the domain. Folds are then grouped into architectures, in which the orientations but not the connectivity of the secondary structure elements are taken into account. Finally, architectures are grouped into four main classes according to the gross secondary structure content of the domain, i.e. mainly alpha, mainly beta, alpha-beta, and low secondary structure content.

Fe-superfamilies represent 2.9% of all the superfamilies included in the CATH database, and are found in all of the four main classes in the CATH hierarchy. Their occurrence in these classes substantially resembles that of all CATH superfamilies (Fig. 2a), indicating that, at this broad level of classification, Fe-sites are distributed over the entire space of protein structures. When considering the second level in the CATH hierarchy, Fe-superfamilies are found in 13 different architectures, which, though representing only 32.5% of all CATH architectures, altogether comprise about 87% of all CATH

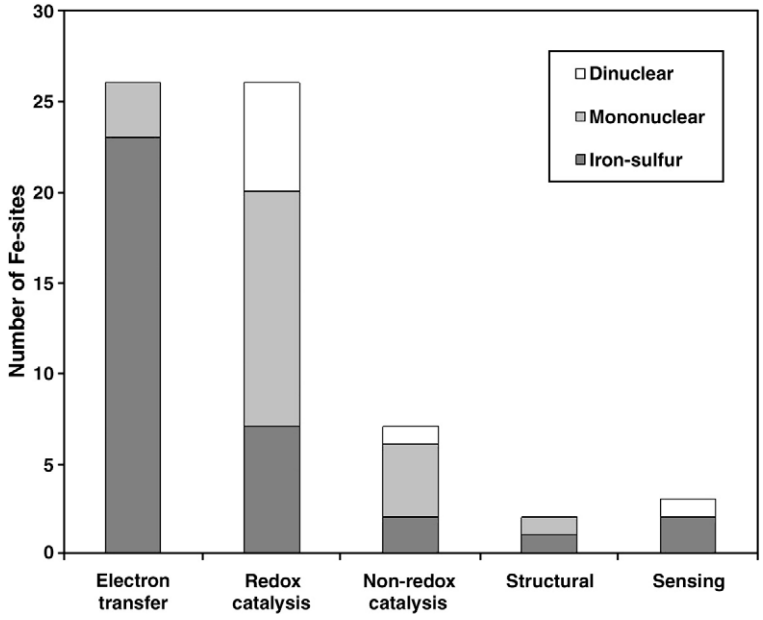

Fig. 1. Number and types of Fe-sites performing the different functions annotated in Table 1.

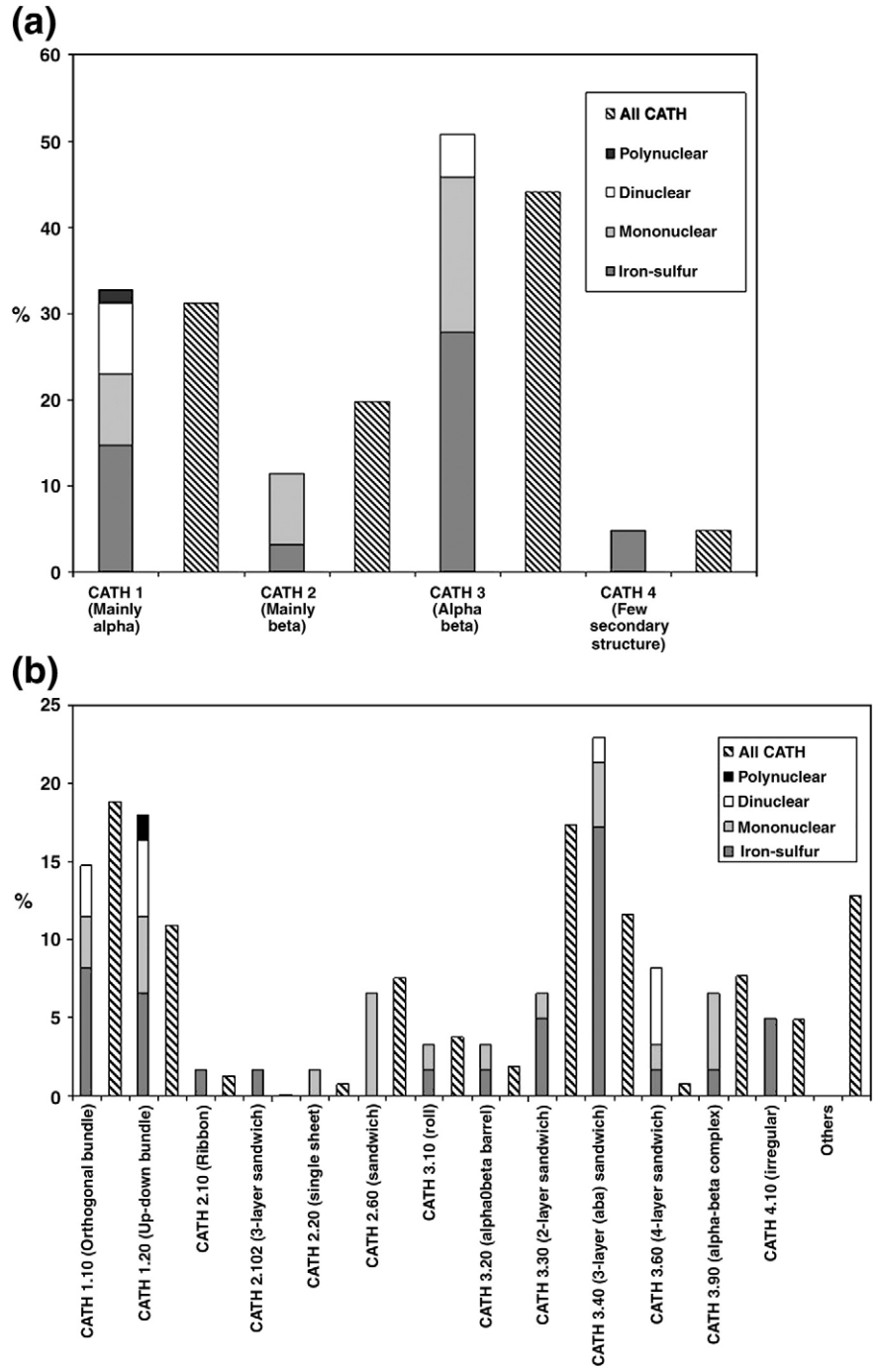

Fig. 2. (a) Percentage of the four CATH classes in CATH superfamilies containing Fe-sites, and in the whole CATH database. (b) Percentage of the 13 CATH architectures in CATH superfamilies containing Fe-sites, and in the whole CATH database.

superfamilies. Out of these architectures, helix bundles (CATH codes 1.10, 1.20) and multiple-layer sandwiches (CATH codes 3.40, 3.60) are the most populated, accounting for about 64% of Fe-superfamilies, and are associated with all the types of Fe-site (Fig. 2b). This indicates that Fe-sites are associated preferentially with these architectures. In particular, the occurrence of multiple-layer sandwiches is much higher (respectively double and more than ten times higher for 3.40 and 3.60 architectures) in Fe-superfamilies than in the whole CATH database. Similarly, there are folds that are closely associated with Fe-sites. In particular, 31 of the 51 folds to which Fe-superfamilies belong do not contain other superfamilies. On the other hand, some of these folds are very common and are adopted by several different superfamilies in addition to Fe-superfamilies. In particular, Fe-superfamilies are found in six large fold groups referred to as superfolds,[17] including the Rossmann fold (CATH code 3.40.50), which is adopted by seven Fe-superfamilies, the jelly roll (CATH code 2.60.120) and the

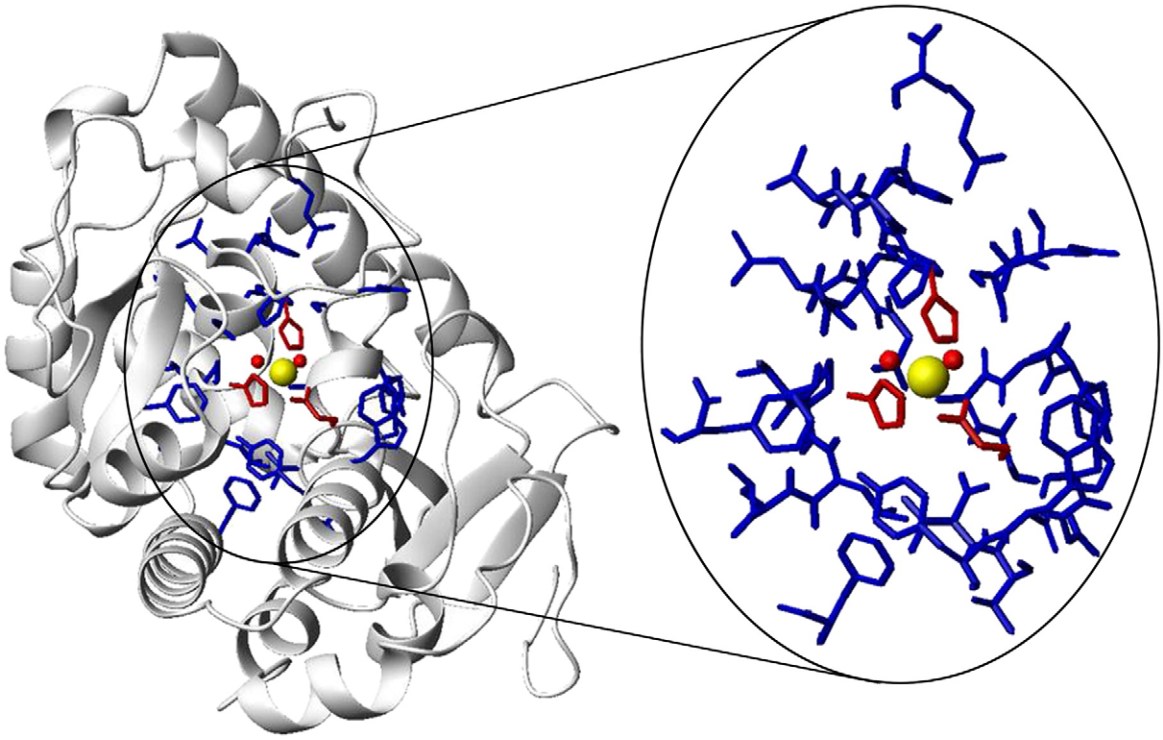

Fig. 3. An example of a structural template representing the mononuclear iron site in phenylalanine hydroxylase (PDB code 1ltz). Iron is depicted as a yellow sphere. Residues forming the first coordination sphere (see the text for definition) include His138, His143, Glu184 (shown as red sticks) and two water molecules (shown as red spheres). Residues forming the second coordination sphere (see the text for definition) are shown as blue sticks.

TIM barrel (CATH code 3.20.20), each adopted by two Fe-superfamilies, and the immunoglobulin-like (CATH code 2.60.40), the alpha-beta plait (CATH code 3.30.70) and the four-helix bundle (CATH code 1.20.120), each adopted by one Fe-superfamily.

Amino acid composition of the first and second coordination spheres of non-heme iron sites

In our approach, we represent metal sites using structural templates that contain the spatial coordinates of the residues forming the first and the second coordination sphere of the metal, as taken from PDB structures (Fig. 3). Therefore, the analysis of the templates for the 86 representative Fe-sites allowed us to investigate the amino acid composition of the first and the second coordination sphere of these sites.

The most common residues in the first coordination sphere of Fe-sites are, in order, Cys, His and Asp or Glu, followed by the relatively rare Tyr and Asn (Fig. 4a). Cys is by far the dominant ligand of iron–sulfur clusters, whereas it is rare in non-iron–sulfur sites. In iron–sulfur clusters, Cys is seldom replaced by His (the Rieske proteins being the best known case) and by Arg in the exceptional case of biotin synthase, where the role of this ligand in catalysis is debated.[19,20] In non-iron–sulfur sites, Cys occurs only in rubredoxin-like $Fe(Cys)_4$ mononuclear centers, which perform electron transfer, and in a few enzymes with quite distinctive properties or functions, such as superoxide reductase (CATH code 2.60.40.730), nitrile hydratase (CATH code 3.90.330.10), and peptide deformylase (CATH code 3.90.45.10). Nitrile hydratase, for instance, is a photoreactive enzyme that catalyses the conversion of nitriles to amides and contains a unique low-spin Fe(III) coordinated by two backbone amide nitrogens and three cysteines, two thiolates of which are post-translationally modified to sulfinate and sulfenate, respectively.[21] Outside iron–sulfur clusters, therefore, Cys appears to be employed only for peculiar catalytic mechanisms, in which theoretical studies suggest that the presence of the highly covalent interaction with sulfur is necessary for iron to function in substrate activation.[22,23] Cys is associated with all the functions annotated in Table 1 (Fig. 4b) because the iron–sulfur clusters have various functions in addition to their typical role in electron transfer.[24,25]

N- and O-donor ligands (mostly His and Asp or Glu) are prevalent in non-iron–sulfur sites, where they are typically (in 26 of 38 cases) found in combination. These sites function mainly as catalytic sites (Fig. 4b) and, in particular, to promote dioxygen chemistry.[26] On average, the ratio between N- and O-donors decreases when going from mononuclear (where it is about 1.5) to dinuclear (where it is about 1) to polynuclear sites, where the site in the only available example (the PhoU protein, CATH code 1.20.58.220) is formed exclusively by O-donors. This observation depends on the bidentate character of Asp and Glu, which can act as bridging ligands for

two iron ions. Besides the polynuclear site of PhoU, there is only one other site formed only by O-donors, that of bacterial periplasmic iron-binding protein (3.40.190.10_2), which contains Fe(III) coordinated by three Tyr residues. Tyr is useful to stabilize Fe(III) with respect to Fe(II) because its phenolate group is a strong donor ligand,[3] and in fact Fe(III) is the stable oxidation state of iron also in the other proteins where Tyr is a ligand, including intradiol dioxygenases (2.60.130.10), transferrins (3.40.190.10_1) and purple acid phosphatase (3.60.21.10). Comparison of mononuclear Fe(II) and Fe(III) sites (Fig. 4c) shows that O-donors are preferred to N-donors in Fe(III) sites with respect to Fe(II) sites, in agreement with the well-known tendency of Fe(II) and Fe(III) to form coordination complexes with N-donor and O-donor ligands, respectively.[27]

The analysis of the composition of the second coordination sphere brings two general observations. First, there is a clear trend for Fe-sites to be enriched in aromatic residues and particularly in Trp and Tyr, whose frequencies are about 2.5 times and 1.5 times more than their average frequencies in proteins (Fig. 5a). This trend is especially visible in the sites that perform redox catalysis (Fig. 5c), whereas it is relatively less significant in the other cases, including sites that perform non-redox catalysis (Fig. 5d). Second, Fe-sites tend to be depleted of charged residues, with Asp, Glu and Lys having frequencies reduced by 30–40% with

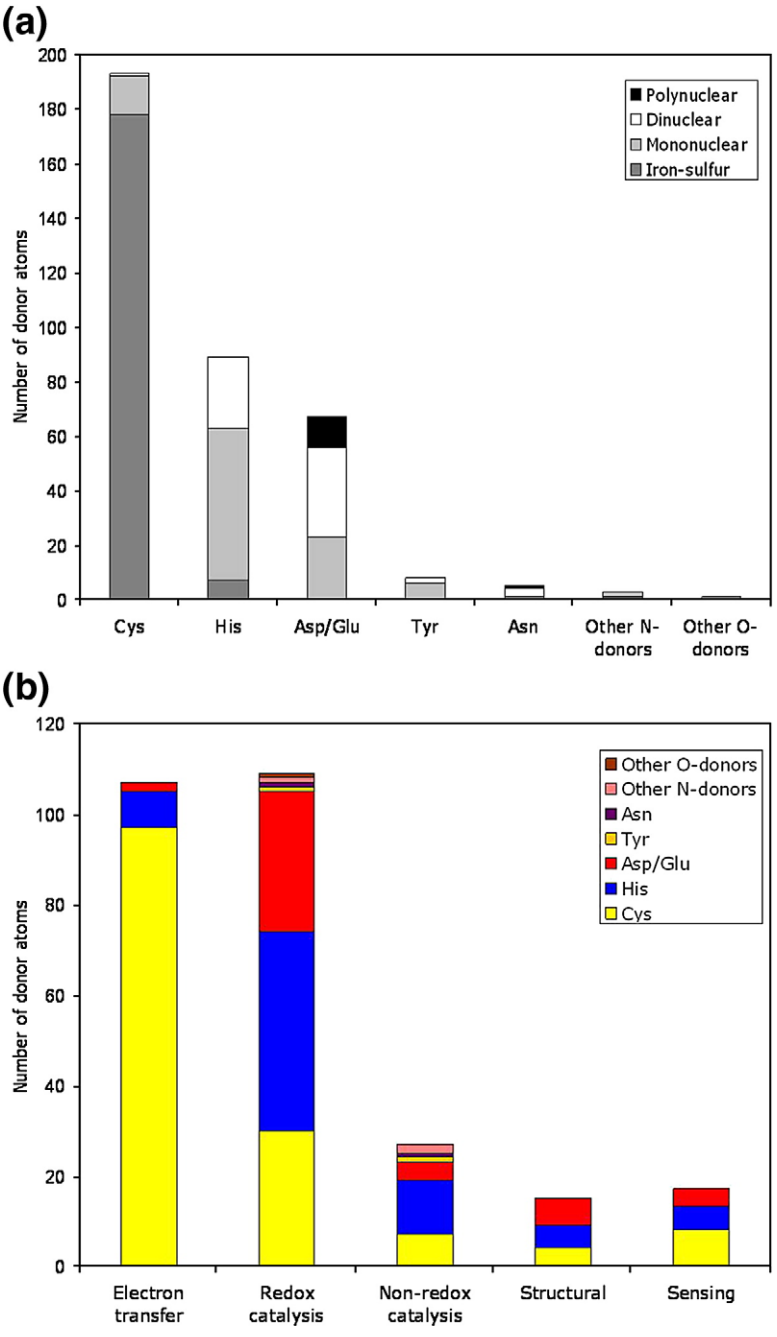

Fig. 4. (a) Occurrence of different protein residues in the first coordination sphere of the 86 representative Fe-sites. Cys, His, Asp/Glu, Tyr and Asn residues all coordinate iron with side chain atoms (sulfur for Cys, nitrogen for His, and oxygen for Asp/Glu, Tyr and Asn). Other N-donors include the guanidine group of Arg side chain and backbone amide groups, while other O-donors are carboxylates located at the C-terminus of the polypeptide chain. (b) Protein residues in the first coordination sphere of representative Fe-sites performing the different functions annotated in Table 1. (c) Protein and non-protein ligands in representative mononuclear Fe-sites containing Fe(II) (top) and Fe(III) (bottom). Pie charts on the left take into account all Fe(II) and Fe(III) sites; pie charts on the right take into account only Fe(II) and Fe(III) sites that do not involve changes in the oxidation state of iron. O-donors, N-donors, S-donors and C-donors are shown in red, blue, yellow, and white, respectively.

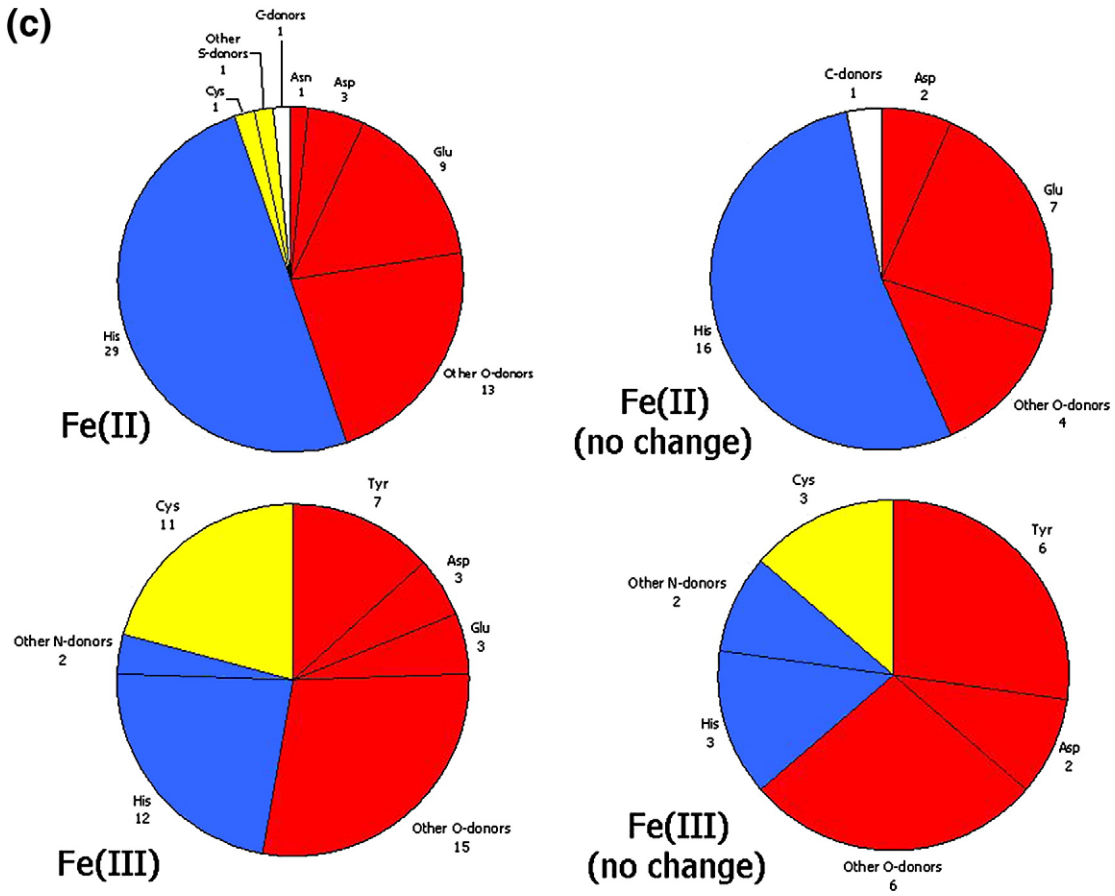

Fig. 4 *(legend on previous page)*

respect to their average frequencies (Fig. 5a). The rationale for these observations is not obvious. It has long been known that a common qualitative feature of metal sites in proteins is that the hydrophilic layer containing the metal ligands is embedded within a larger hydrophobic layer.[28] This notion can justify the general reduction observed for charged residues; however, it does not explain the distinctive enrichment of aromatic residues that is not observed for other hydrophobic residues. A possible functional requirement for aromatic residues in Fe-sites may be due to their ability to mediate electron transfer reactions, which makes them most suitable for tunneling electrons to/from redox sites.[29-31] It has been suggested that the interaction between ligands coordinated to a metal ion and aromatic residues can represent a kind of cation–π interaction,[32] which may be important for stabilizing the conformation of metal sites, and for the mechanisms of certain enzymatic reactions occurring at the metal center.[33]

Structural comparison of non-heme iron sites

The representation of metal sites as structural templates allows their comparison in a highly automated fashion by using the available programs for structure alignment, such as FAST. The comparison of metal sites found in different superfamilies is aimed at detecting common metal-binding structural motifs in structurally and evolutionarily unrelated metalloproteins. These motifs can be used to classify metal sites by integrating and extending beyond CATH and SCOP classifications. We carried out the all-*versus*-all comparison of the 86 representative Fe-sites and we have clustered them using three different thresholds for high, medium and low similarity (see Methods). This procedure allowed us to recognize five clusters of Fe-sites which, though belonging to different superfamilies, share common structural motifs. At the high, medium and low similarity level, these clusters include a total of 15 (17%), 26 (30%) and 32 (37%) Fe-sites, respectively (Table 2). Therefore, although the majority of the representative Fe-sites are found to be dissimilar to any other Fe-site of the set, the structural variation across Fe-sites even at the high similarity level is reduced by 17% with respect to CATH classification.

The examination of the above-mentioned clusters showed that they can provide hints about the function and evolution of the sites included therein. As discussed in more detail below, the small clusters 1 and 2 provide clear examples of Fe-sites that, though located in different structural contexts, share the same structural and functional features, and can thus result from convergent evolution of unrelated proteins to reach similar solutions. Similar indications can be gleaned from the larger cluster 3, where Fe-sites display relatively larger variations of a common structural motif. In cluster 4, instead, a

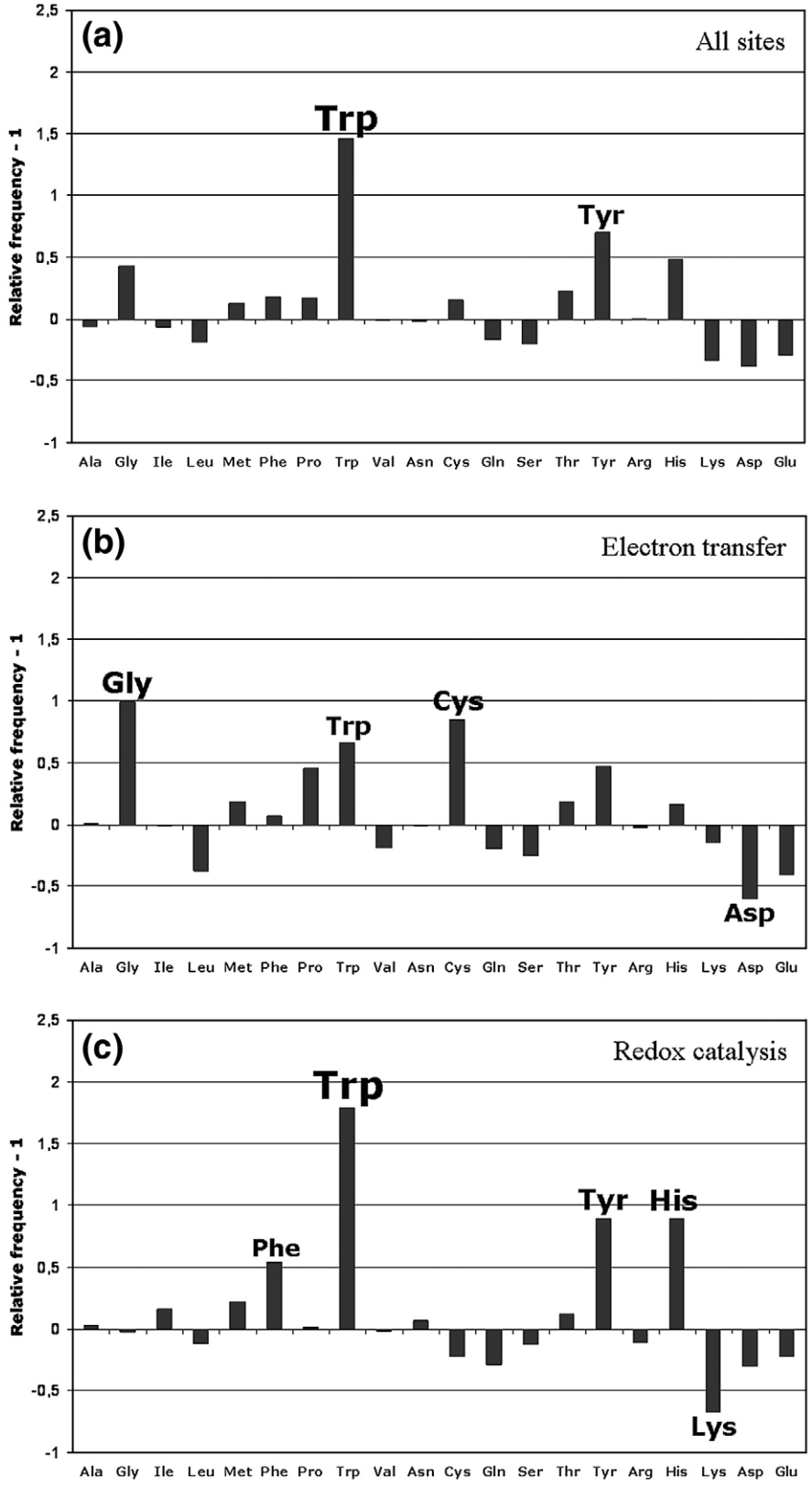

Fig. 5. Frequencies of protein residues in the second coordination sphere of the 86 representative Fe-sites against their average frequencies in proteins, as calculated from the UniProtKB/Swiss-Prot database (http://www.expasy.ch/sprot/relnotes/relstat.html). Frequencies are calculated for: (a) all sites; (b) sites performing electron transfer; (c) sites performing redox catalysis; and (d) sites performing non-redox catalysis. Residues with frequencies higher by more than 50% or lower by more than 50% with respect to their average frequencies in proteins are labeled.

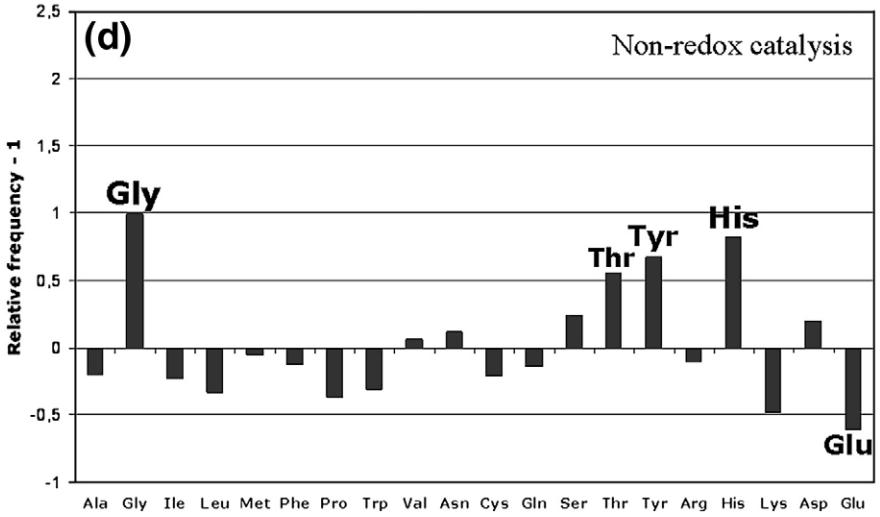

Fig. 5 *(legend on previous page)*

common structural motif is easily recognizable only at the high similarity level. This indicates that metal sites clustered at the medium and the low similarity level must be analyzed with more caution. Cluster 5, finally, includes Fe-sites of proteins that belong to three different superfamilies but share the same fold, and may thus all result from a process of divergent evolution.

Our templates can be used also to compare metal sites that belong to the same superfamily. This kind of comparison is aimed at detecting fine differences across structurally and evolutionarily related metal sites, which can provide hints on the mechanisms of divergent evolution acting on them. An example of this analysis is given in the discussion of cluster 5, which contains Fe-sites found in a large superfamily of enzymes with different functions.

Cluster 1

Cluster 1 includes two iron–sulfur sites. One is the Fe_2S_2 site of dihydroorotate dehydrogenase B (PDB code 1ep3),[34] and the other is the Fe_2S_2 site of carbon monoxide dehydrogenase (PDB code 1n62),[35] which represents the centers of 2Fe-2S ferredoxins.[36,37] Both of these enzymes are flavoproteins; however, they appear to be unrelated in terms of both structure and sequence: the chains containing the Fe_2S_2 centers, in particular, are classified in different main classes in both CATH (2.10.240.10 for 1ep3 and 3.10.20.30 for 1n62) and SCOP hierarchies (c.25.1 for 1ep3 and d.15.4 for 1n62), and their sequence identity is no higher than 12%. Although the two Fe_2S_2 centers are both coordinated by four cysteines embedded in closely similar sequence motifs (CX(4) CX(2)CX(14)C for 1ep3 and CX(4)CX(2)CX(11)C for 1n62), sequence alignment is not able to detect correspondence between the two patterns, because 1ep3 has the binding site at the C-terminus, and 1n62 has the binding site at the N-terminus. On the other hand, structural comparison of the two sites shows that they are nicely superimposable, although they are found in very different structural contexts. This similarity correlates with a similarity of function, because both clusters are involved in electron transfer to/from flavin cofactors nearby. The Fe-sites of this cluster may thus be an example of convergent evolution. Figure 6a shows that the two sites share a common structural motif formed by a two-stranded, antiparallel β-sheet followed by a loop structured to bind the Fe_2S_2 cluster. The superimposition (see Fig. 6b) suggests that a conserved lysine residue (Lys247 in 1ep3 and Lys60 in 1n62) is important for maintaining the structure of the site by forming a hydrogen bond with the carbonyl oxygen of a residue located between the first two coordinating cysteines (Gly229 in 1ep3 and Ser45 in 1n62).

Cluster 2

Cluster 2 is composed of three Fe-sites. One (PDB code 1jm1)[38] is a Fe_2S_2 site representing the centers of the so-called Rieske domains found, for instance, in bc_1 and b_6f complexes.[39] Rieske centers are coordinated by two cysteines and two histidines. The other two are mononuclear sites consisting of an iron ion coordinated by four cysteines in a tetrahedral configuration. One (PDB code 2dsx) represents the centers of rubredoxins,[40] which are small bacterial proteins that can be found as domains of fusion proteins such as rubrerythrins.[41] The other (PDB code 2fiy) represents the two centers, possibly originated from duplication, found in a protein of unknown function annotated as a homologue of FdhE, a protein required for the assembly of formate dehydrogenase in certain bacteria.[42] Rubredoxin and Rieske domains, both of which function in electron transfer, are classified in different CATH architectures (2.20.28.10 and 2.102.10.10, respectively) and different SCOP main classes (g.41.5 and b.33.1, respectively). The 2fiy protein is not classified in CATH, and is the only member of the FdhE-like superfamily in SCOP (e.59.1). The average sequence identity of the chains containing these sites is 9±3%.

Table 2. Clusters of representative Fe-sites obtained at high, medium and low levels of similarity

	Cluster 1				Cluster 2				Cluster 3				Cluster 4				Cluster 5			
High	1ep3	2.10.240.10	FeS	ET	1jm1	2.102.10.10	FeS	ET	1kg2	1.10.1670.10	FeS	Str	1f8n	1.20.245.10	M	RC	1odm	2.60.120.330	M	RC
	1n62	3.10.20.30	FeS	ET	2fiy	e.59.1	M	Unk	1feh	3.10.20.30	FeS	ET	1sum	1.20.58.220	P	Unk	2b5h	2.60.120.10	M	RC
													1mxr	1.10.620.20	D	RC				
													1yuz	1.20.1260.10	D	RC				
													2j8c	1.20.85.10	M	ET				
													1rcw	1.20.910.10	D	Unk				
													2mhr	1.20.120.50	D	Sen				
Medium									1gte	1.10.1060.10	FeS	ET	1o2d	1.20.1090.10	M	NC	1ds1	3.60.130.10	M	RC
									1iqz	3.30.70.20	FeS	ET	2bmo	3.90.380.10	M	RC				
									1o94	3.20.20.70	FeS	ET	1lm4	3.90.45.10	Di	RC				
													2cb2	UNCLASS.	M	NC				
													1aor	1.10.569.10	M	Str				
													1ltz	1.10.800.10	M	RC				
													1u8v	1.10.3140.10+ 1.20.140.10	FeS	NC				
Low					2dsx	2.20.28.10	M	ET	1uwv	c.66.1	FeS	Unk	1bsm	a.2.11+d.44.1	M	RC				
									1wui	4.10.480.10	FeS	ET	2pq7	a.211.1	D	RC				
									2z1d	UNCLASS.	FeS	Unk								

For each Fe-site, the table reports (i) the PDB code, (ii) the CATH or SCOP domain, (iii) the type (FeS, iron-sulfur; M, mononuclear; D, dinuclear; P, polynuclear), and (iv) the function (ET = electron transfer; RC, redox catalysis; NC, non-redox catalysis; Str, structural; Sen, sensing; Unk, unknown).

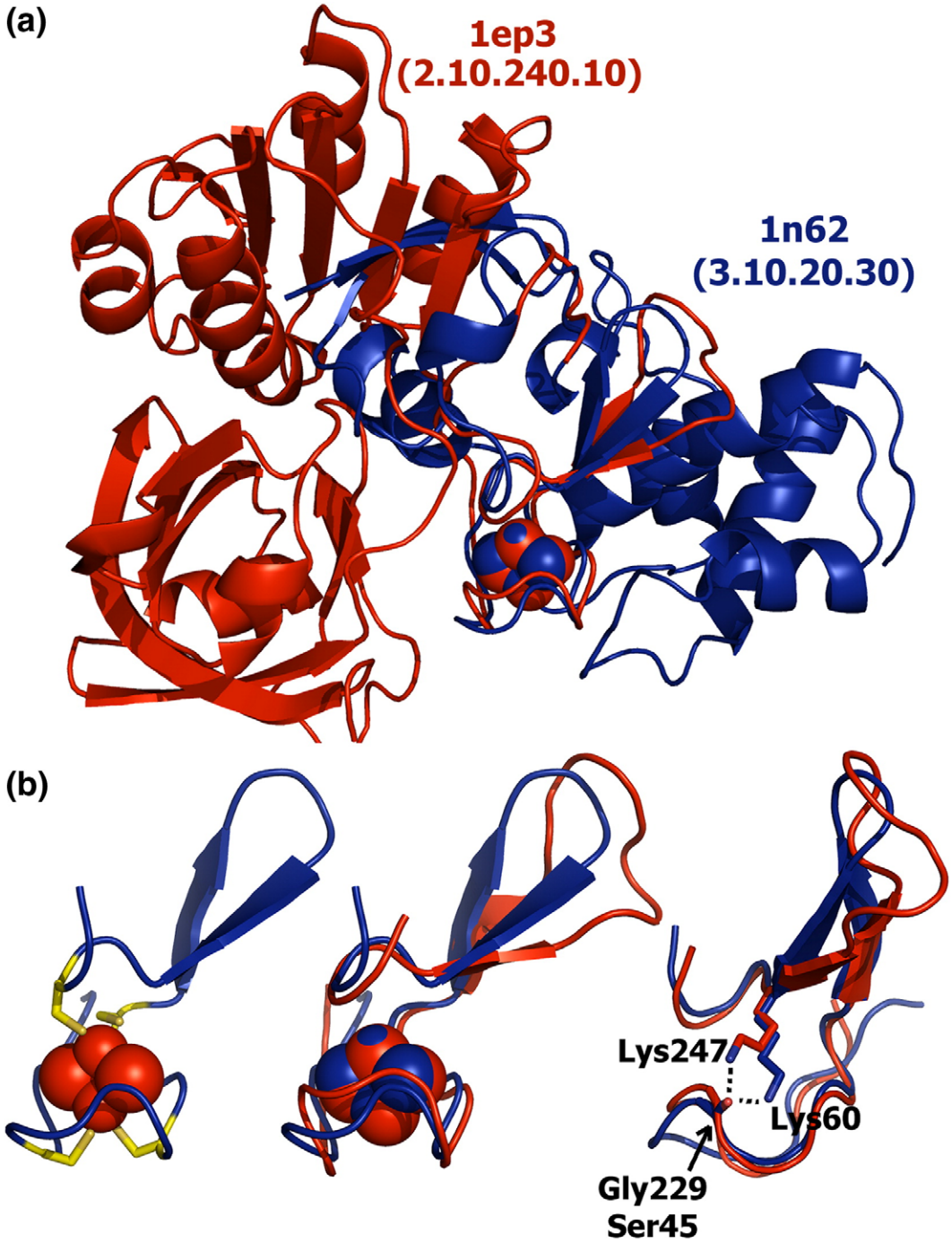

Fig. 6. Superimposition of the two Fe-sites included in cluster 1. (a) The cartoon structure of dihydroorotate dehydrogenase B (PDB code 1ep3) is shown in red, and that of carbon monoxide dehydrogenase (PDB code 1n62) in blue. The Fe_2S_2 clusters are depicted as spheres. (b) Common structural motif around the Fe-site. Left: The Cys ligands are shown as yellow sticks (only 1n62 is shown for clarity). Middle: superimposition of the structural motif (colors as in a). Right: hydrogen bond formed between Lys247 and Gly229 in 1ep3 and between the equivalent Lys60 and Ser45 in 1n62. Lysine side chains and backbone carbonyl groups are shown as sticks.

The superimposition of the Fe-sites of this cluster shows that they share a common structural motif formed of a three-stranded β-sheet and two loops containing two iron ligands each (Fig. 7a). This motif can be regarded as composed of two half-sites consisting of (i) a short loop flanked by two strands of the β-sheet (2 and 3 in Fig. 7) and (ii) a relatively longer loop followed by the third strand of the β-sheet (1 in Fig. 7). In both loops, the conformation is stabilized by two conserved hydrogen bonds between the N-terminal iron ligand and the two residues that follow the C-terminal iron ligand (Fig. 7b). It has been noticed previously that rubredoxins assume a fold around their metal center similar to that of Rieske proteins.[43,44] On this basis, it has been suggested that Rieske domains evolved from an

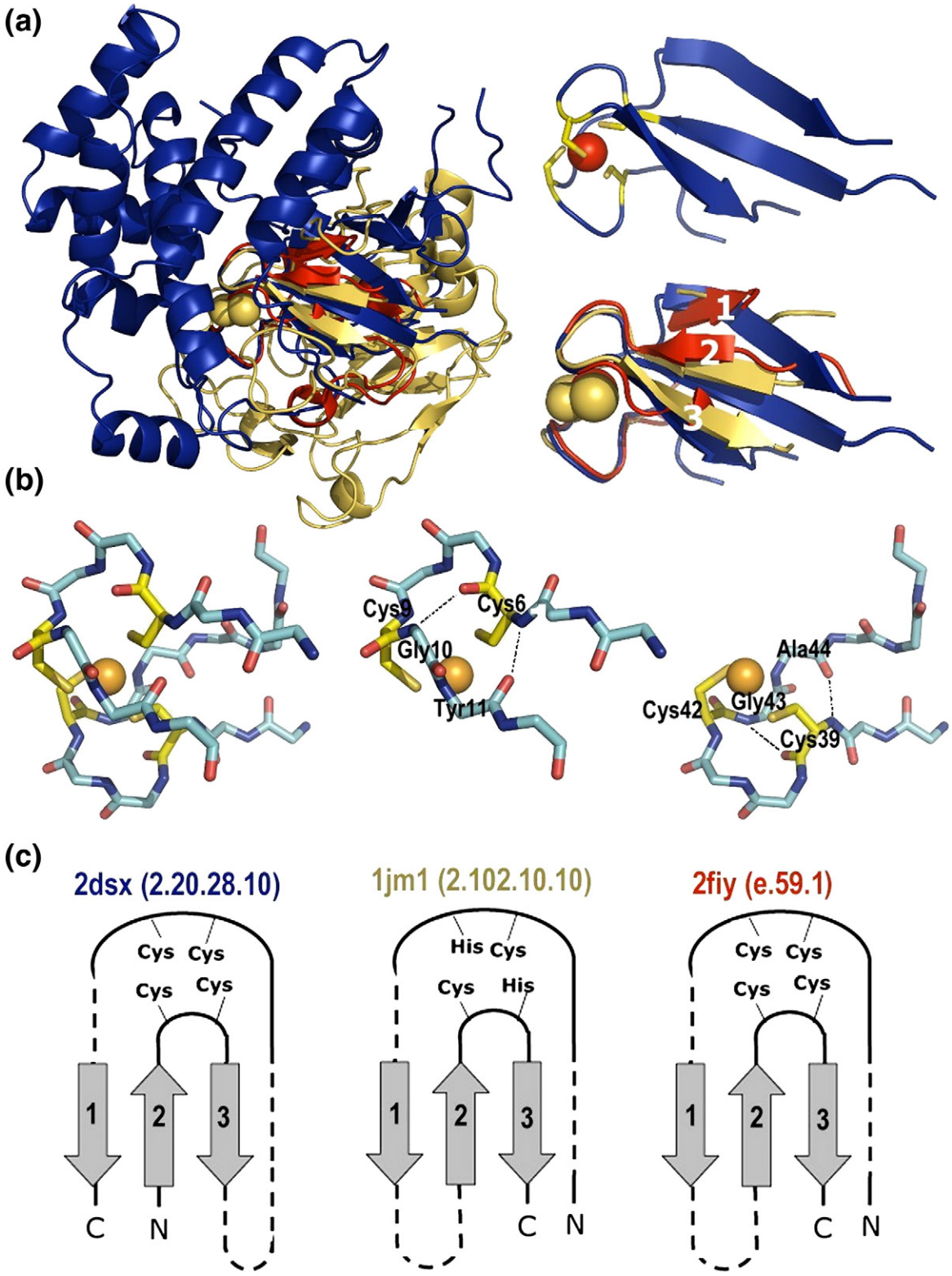

Fig. 7. Superimposition of the three Fe-sites included in cluster 2. (a) The cartoon structure of rubredoxin (PDB code 2dsx) is shown in blue, that of the Rieske protein SoxF from *Sulfolobus acidocaldarius* (PDB code 1jm1) is shown in gold, and that of the FdhE homolog protein from *Pseudomonas aeruginosa* (PDB code 2fiy) is shown in red. (b) Hydrogen bonds formed in 2dsx between Cys6 (located in the loop between the strands numbered 2 and 3 in a) and Gly10 and Tyr11, and between Cys39 (located in the loop preceding the strand numbered 1 in a) and Gly43 and Ala44. These hydrogen bonds are conserved also in 1jm1 and 2fiy. (c) Topology of the β-sheet characterizing the motif binding the Fe-sites of this cluster. Strands are numbered as in a.

early rubredoxin-like module through minimal changes of the protein scaffold.[45] However, the sequential arrangement of the β-strands in the motif is different in the two cases. In rubredoxins, half-site (i) precedes (ii), whereas the opposite occurs in Rieske domains (Fig. 7c). Therefore, if rubredoxin and Rieske domains originated from a common ancestral domain, a rearrangement of the two half-sites must have occurred. In this hypothesis, it is interesting to note that the 2fiy motif may represent an evolutionary intermediate between the two, in that it binds a mononuclear iron like rubredoxin, yet its topology resembles that of Rieske domains (Fig. 7c). In any case, it is conceivable that the two centers in 2fiy function in electron transfer, similarly to the other two centers of this cluster.

Cluster 3

Cluster 3 is composed of eight Fe-sites, all of which are of the Fe_4S_4 type. They are found in a remarkable variety of protein domains belonging to

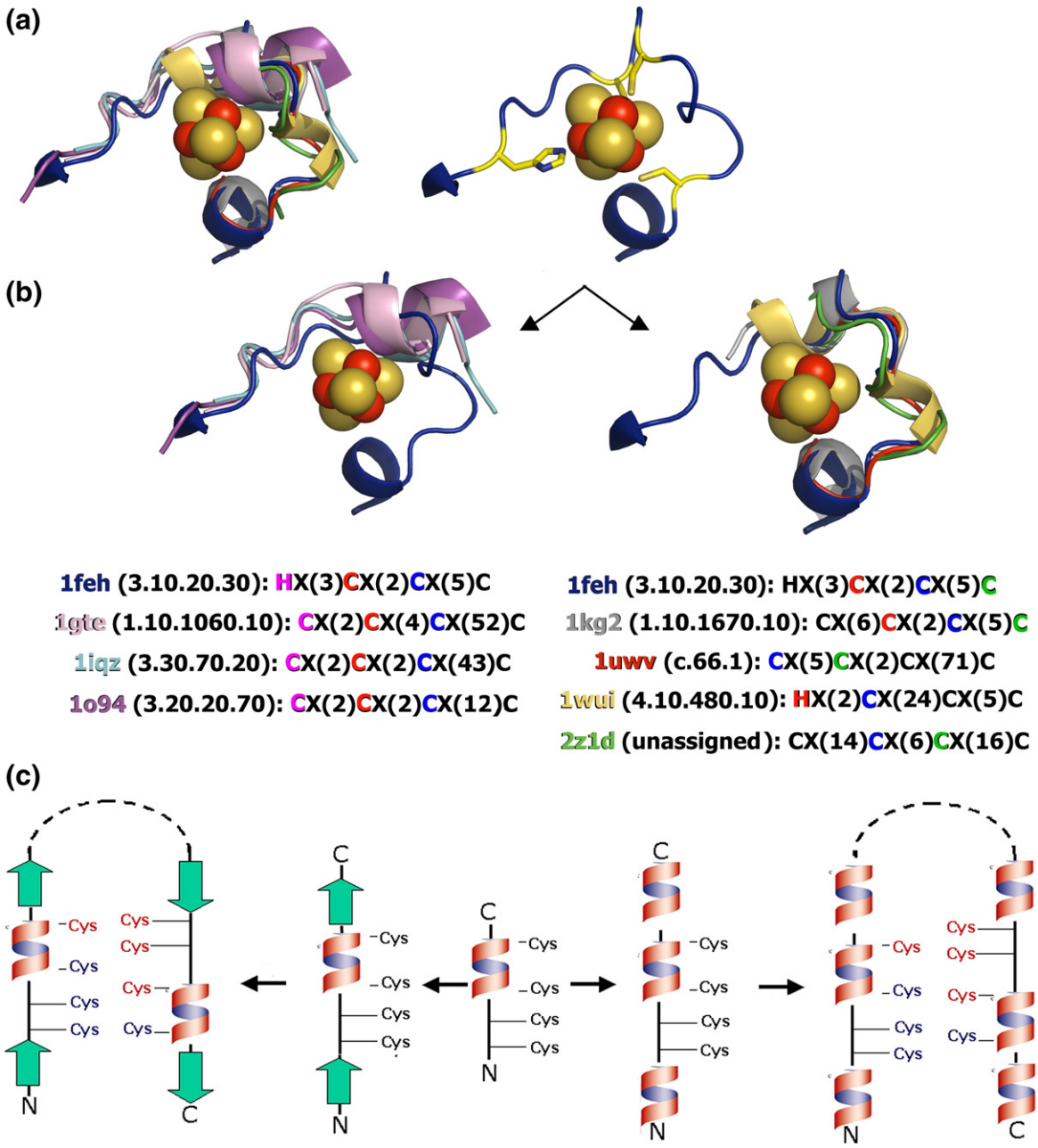

Fig. 8. Superimposition of the eight Fe-sites included in cluster 3. (a) Cartoon structure of 1feh, 1kg2, 1gte, 1iqz, 1o94, 1uwv, 1wui and 2z1d (left) and of 1feh with iron ligands (sticks) around the Fe_4S_4 cluster (spheres). (b) Superimposition of 1gte, 1iqz and 1o94 with the N-terminal part of 1feh (left) and of 1kg2, 1uwv, 1wui and 2z1d with the C-terminal part of 1feh (right). Superimposed residues are shown in the same color in the sequence motifs containing the iron ligands. (c) A model for the evolution of archetypical ferredoxin sites, involving the duplication of a loop-helix segment containing four iron ligands that is preceded and followed by either a β-strand (left) or a α-helix (right).

three of the four main classes in the CATH hierarchy (Table 2) and having an average sequence identity of 9±3%. At the high similarity level, the cluster includes two sites that are contained in the DNA repair enzyme MutY[46] (PDB code 1kg2) and in Fe-only hydrogenase (PDB code 1feh), respectively.[18] These sites possibly represent the most compact structural motifs that can bind a Fe_4S_4 cluster in a protein. In these motifs, which have been referred to as Fe_4S_4 cluster loops or FCLs,[47] the Fe_4S_4 cluster is bound entirely within a loop by four ligands that are spaced only a few residues apart (Fig. 8).

At the medium similarity level, the cluster includes three additional sites that are contained in dihydropyrimidine dehydrogenase (PDB code 1gte),[48] trimethylamine dehydrogenase (PDB code 1o94),[49] and 4Fe-4S ferredoxin (PDB code 1iqz).[50] These sites are related to the two discussed above because they have three sequentially close iron ligands that superimpose with the three N-terminal ligands of 1feh (Fig. 8). The structures of these three sites can all be described by a model that involves the duplication of a motif consisting of a loop-helix segment that contains four iron ligands and is preceded and followed by the same secondary structure element, which can be either a β-strand (as in 1iqz) or an α-helix (as in 1gte and 1o94) (Fig. 8c). This motif, which can bind two Fe_4S_4 clusters, can be the archetype of all the ferredoxin sites found in single-domain ferredoxins like 1iqz and in multi-domain oxidoreductases like 1gte and 1o94, implying that sites binding a single Fe_4S_4 cluster result from the disruption of one of the two Fe_4S_4-binding sites. In some cases the disruption is limited to the loss of part or all of the iron ligands, such as in 1iqz, whereas in others the structure may have undergone major changes, including the loss of entire secondary structure elements, such as in 1o94.

Finally, at the low similarity level, three further sites are included in the cluster. These sites, which are contained in RNA 5-methyluridine methyltransferase (PDB code 1uwv),[51] Ni-Fe hydrogenase (PDB code 1wui),[52] and in the Ni-Fe hydrogenase maturation protein HypD (PDB code 2z1d),[53] also share a partial similarity with 1feh and 1kg2, but, at variance with those included at the medium similarity level, are related more closely to the C-terminal part rather than to the N-terminal part of the 1feh and 1kg2 sites (Fig. 8).

Cluster 4

With respect to the clusters discussed previously, cluster 4 includes a relatively high number of Fe-sites, going from seven at the high similarity level to 14 and 16 at the medium and the low similarity level, respectively (Table 2). However, in this case a major structural similarity is observed only for Fe-sites that are included at the high similarity level. These latter sites are found in protein chains whose average identity is 10±3%, and include the dinuclear centers of ribonucleotide reductase R2 (PDB code 1mxr),[54] nigerythrin (PDB code 1yuz),[55] hemerythrin (PDB code 2mhr),[56] and of the CADD protein from *Chlamydia trachomatis* (PDB code 1rcw),[57] the mononuclear centers of lipoxygenase-1 (PDB code 1f8n),[58] and of the photosynthetic reaction center from *Rhodobacter sphaeroides* (PDB code 2j8c),[59] and the multinuclear center of a PhoU protein homologue from *Thermotoga maritima* (PDB code 1sum).[60] All of these Fe-sites share a common structural scaffold consisting of a bundle of four helices which, however, can be connected in different ways and can be supplied by different polypeptide chains (Fig. 9). Previous analyses suggest that they have evolved independently in a convergent manner, even when the topology of the bundle is the same.[61] In most cases, this architecture can be described as resulting from the duplication of a helix-loop-helix motif, where each helix contains either one or two closely spaced iron-binding residues. In 2j8c, instead, two helix pairs from two different protein subunits are packed together to form the Fe-site. It is possible that a similar helix packing occurs in 1sum, where the iron cluster is exposed on the surface of the protein, by way of dimerization or oligomerization. In fact, this is also the case for the dimeric aldehyde ferredoxin oxidoreductase (PDB code 1aor),[62] which binds an iron ion at the dimer interface. Also in 1aor, iron is coordinated by two helix pairs from the two monomers, yet the packing of the helices is different to the extent that this Fe-site was included in the cluster only at the medium similarity level.

The other Fe-sites included in this cluster at the medium and the low similarity levels have part or all of the iron ligands located in helices, which, however, are not arranged in bundles related to those discussed above. In most cases, the location of such sites in this cluster is due only to the presence of two close iron ligands (most commonly histidines) within a helix, while the other ligands are found in a variety of structural contexts. Examples include phenylalanine hydroxylase (PDB code 1ltz),[63] peptide deformylase (PDB code 1lm4),[64] and sulfur oxygenase reductase (PDB code 2cb2),[65] all of which have a HX(3-4)H iron coordination motif embedded in a helix. This suggests that the use of two ligands closely spaced in a helix is a relatively common solution adopted by various proteins to support the structure of iron sites.

Cluster 5

Cluster 5 includes the three mononuclear Fe-sites of clavaminate synthase (PDB code 1ds1),[66] isopenicillin N synthase (PDB code 1odm),[67] and cysteine dioxygenase (PDB code 2b5h).[68] These sites represent the three CATH superfamilies 3.60.130.10, 2.60.120.330 and 2.60.120.10, respectively, and are found in protein chains whose average sequence identity is 12±2%. However, all the proteins classified in these CATH superfamilies belong in SCOP to the same fold (b.82), which is called double-stranded β-helix (DSBH) and consists of a β-

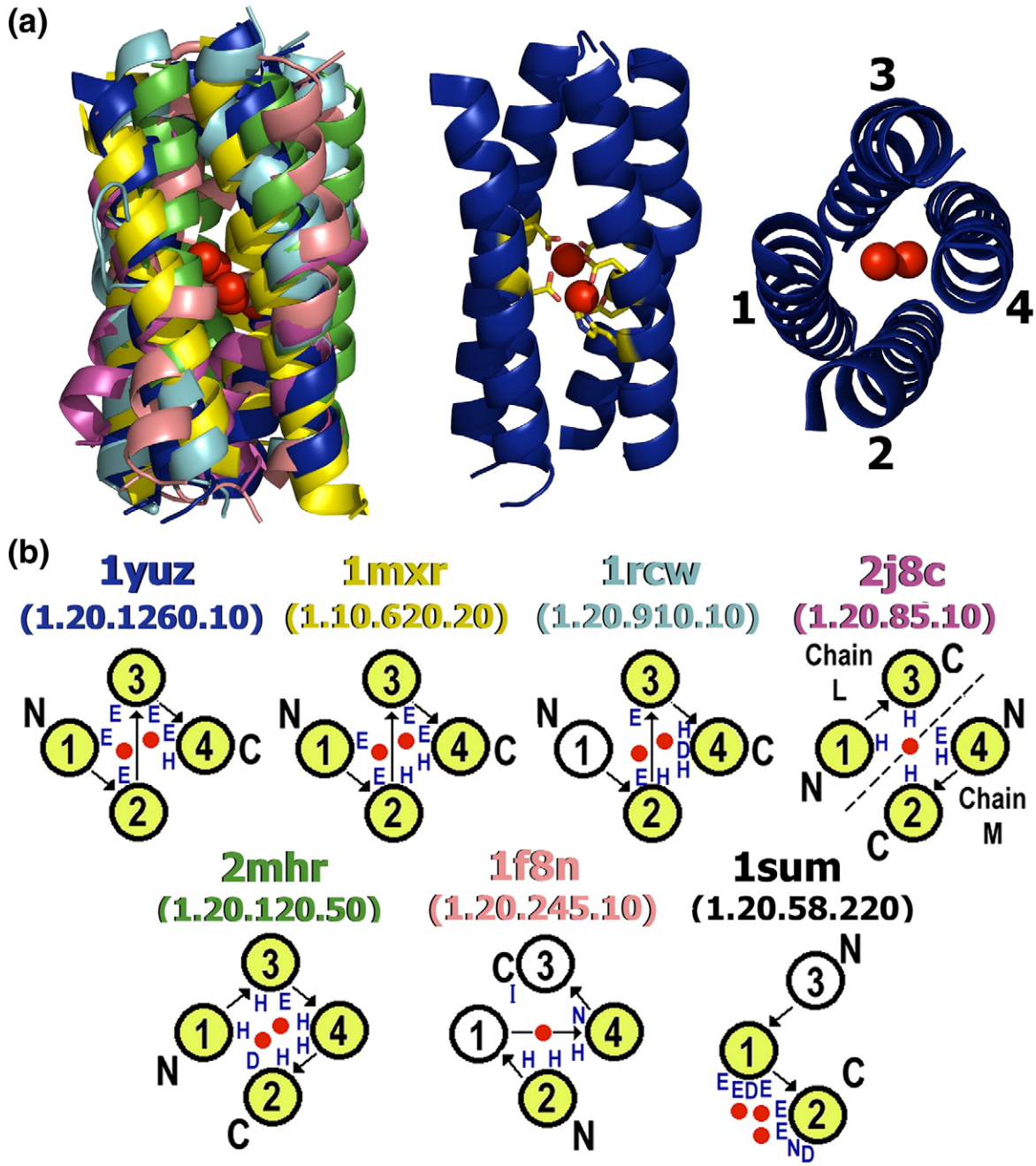

Fig. 9. Superimposition of the seven Fe-sites included in cluster 4 at the high similarity level. (a) Cartoon structure of the helix bundles wrapping the iron site (shown as red spheres) in 1yuz, 1mxr, 1rcw, 2j8c, 2mhr and 1f8n (left) and in 1yuz only (middle, side view with iron ligand shown as sticks, and right, top view). (b) Scheme of the connectivity of helices in the helix bundles. Helices are numbered as in a, and are depicted in yellow when providing at least one iron ligand (also shown in blue one-letter code). Iron ions are shown as red spheres. Iron ligands are shown using the one-letter code (D, Asp; E, Glu; H, His; I, Ile; N, Asn).

sandwich formed of two four-stranded, anti-parallel β-sheets (see Fig. 10a). The DSBH fold is common to many different proteins and enzymes that have very diverse overall structures and functions, and in iron-dependent enzymes it provides the scaffold for iron binding.[69] The iron site is invariably located at one end of the DSBH core (Fig. 10a); therefore, the similarity observed between the Fe-sites of this cluster is not unexpected. However, it is interesting to look at the variations in iron coordination displayed by the different iron enzymes included in these superfamilies, all of which may have arisen *via* a process of divergent evolution from an ancestral iron enzyme.

The CATH superfamilies 3.60.130.10 and 2.60.120.330 contain various oxygenases catalyzing the oxidation of a wide range of substrates, most of which, but not all, require 2-oxoglutarate (2OG) as a co-substrate. In all these enzymes, iron is coordinated by two His residues located on two adjacent

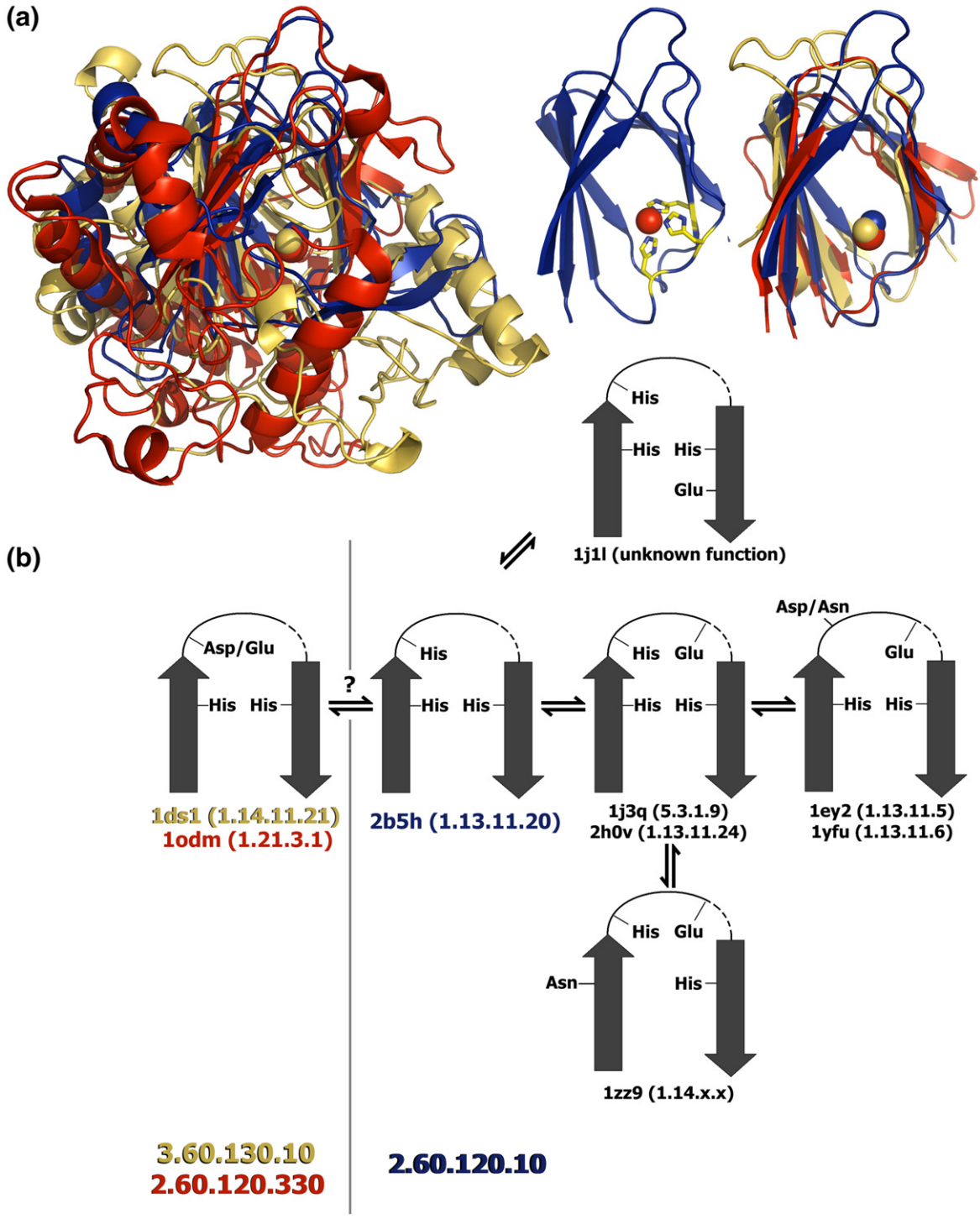

Fig. 10. Superimposition of the three Fe-sites included in cluster 5. (a) Left, cartoon structure of clavaminate synthase (PDB code 1ds1, green), isopenicillin N synthase (PDB code 1odm, red) and cysteine dioxygenase (PDB code 2b5h, blue). Right, DSBH cores of the three proteins (same color code). (b) A representation of the Fe-sites included in the CATH superfamilies 2.60.120.330, 3.60.130.10 and 2.60.120.10, showing the residues present in the first coordination sphere. These are the same in all the Fe-sites of 2.60.120.330 and 3.60.130.10, whereas they vary among different proteins and enzymes (identified by their EC number) in 2.60.120.10. Fe-sites differing by a single ligand are separated by double arrows.

β-strands and by one acidic (Asp or Glu) residue located in a loop two positions from the N-terminal His (Fig. 10b, left). This is in fact the case for both clavaminate synthase, which is a 2OG-dependent enzyme, and isopenicillin N synthase, which is not. The enzymes classified in the CATH superfamily 2.60.120.10, instead, display a relatively higher variability in iron coordination, which is generally

different in enzymes catalyzing different reactions. As shown in Fig. 10b, the representative Fe-site of cysteine dioxygenase is related by one or two ligand changes to the Fe-sites of phosphoglucose isomerase (PDB code 1j3q),[70] quercetin dioxygenase (PDB code 2h0v),[71] hydroxypropylphosphonic acid epoxidase (PDB code 1zz9),[72] homogentisate dioxygenase (PDB code 1ey2),[73] hydroxyanthranilate dioxygenase (PDB code 1yfu),[74] and pirin (PDB code 1j1l).[75] Pirin is a nuclear protein for which a catalytic activity in human cells has been reported.[76] It is worth noting that none of the proteins included in 2.60.120.10 coordinate iron by Asp or Glu in the position where Asp or Glu occurs in the Fe-sites of 2.60.120.330 and 3.60.130.10. Asp is in fact present in that position in the Fe-site of hydroxyanthranilate dioxygenase, but does not coordinate iron (Fig. 10b). Therefore, we hypothesize that the usage of 2OG in 2OG-dependent enzymes requires the coordination of iron by Asp or Glu in that position.

Fe(II) *versus* Fe(III)

As noted above, the value of iron to biological systems resides largely in the two stable oxidation states, Fe(II) and Fe(III), that determine the versatile reactivity of this metal. Fe(II) and Fe(III) have different properties, which may also depend on their spin state; for example, the ionic radius of iron in octahedral complexes varies from 0.55 Å for low-spin Fe(III) to 0.78 Å for high-spin Fe(II).[77] The specific chemical properties of Fe(II) and Fe(III) are reflected in their differential use by iron-dependent enzymes: in mononuclear Fe-sites, for example (see Supplementary Data Table S3), iron can be employed as Fe(II) to promote dioxygen activation (e.g., in extradiol-cleaving dioxygenases such as 2,3-dihydroxybiphenyl dioxygenase), or as Fe(III) to promote substrate activation (e.g., in intradiol-cleaving dioxygenases such as catechol 1,2-dioxygenase).[13] The analysis of the first coordination sphere of these sites (Fig. 4c) indicates that the relative stabilities of Fe(II) and Fe(III) can be modulated by the type of coordinating ligands (e.g., Tyr favoring the oxidized form). Both Fe(II) and Fe(III), on the other hand, are nearly always found in the high-spin state (the aforementioned nitrile hydratase being the only exception), and are invariably antiferromagnetically coupled in dinuclear sites.

Within our dataset of representative Fe-sites, in only 20 cases (about 23%) the functionality of iron does not appear to involve oxidation state changes (see Supplementary Data Table S3). Not unexpectedly, therefore, iron is used mostly as a redox-active center, switching between two (or more) redox forms in electron transfer and enzymatic catalytic processes. Upon redox changes, iron coordination in a Fe-site may also change to various extents. For example, upon reduction of the dinuclear site of rubrerythrin from the Fe(III)-Fe(III) state to the Fe(II)-Fe(II) state, one iron has been observed to switch from a Glu to a His ligand, yet maintaining a 6-fold coordination.[78] PDB structures are available for both the oxidized (PDB code 1lkm) and the reduced form (PDB code 1lko); however, our grouping procedure (see Methods) did not separate the corresponding Fe-sites into different subgroups. This is in agreement with the observation that the coordination change occurring in rubrerythrin is in fact due to the redox-induced movement (approximately 1.8 Å) of the iron, whereas the protein does not undergo appreciable structural rearrangement.[78] In general, within our dataset there are no Fe-sites of the same protein but with different oxidation states that are clustered into different subgroups, indicating that oxidation state changes typically determine only a minimal reorganization of the protein fold around the metal site. Details of such reorganization can be investigated by comparing the Fe-sites in the oxidized and the reduced state, when both structures are available: as an example, the abovementioned case of rubrerythrin is shown in Supplementary Data Fig. S5.

Hints for functional assignment

Structural comparison of metal sites can be used to obtain hints on the function of uncharacterized metalloproteins. Structural templates of sites with known function can be employed as datasets against which structural templates of sites with unknown function can be compared systematically. As an example of this application of our approach, we compared the Fe-sites of 15 structural genomics (SG) proteins of unknown function contained in our ensemble of non-heme iron proteins against all the other Fe-sites of our dataset. For each SG protein we selected the protein with known function yielding the best FAST score, and we used the functions of the best-scoring proteins to infer those of the SG proteins, on the assumption that similar active sites imply similar functionality for the proteins (Table 3). In all cases except one, the best FAST score was higher than 2.25, corresponding to the high similarity threshold used to cluster the representative Fe-sites (see Methods). For those SG proteins for which a functional annotation is available in the PDB, our predictions are in agreement with those annotations (Table 3).

In almost all cases, the best-scoring protein belongs to the same CATH or SCOP superfamily of the SG protein (Table 3). This indicates that the comparison of site templates can provide information equivalent to that obtainable from the comparison of entire domain or protein structures for identifying proteins that are possibly functionally related. However, the usage of site templates can result in more accurate predictions, in particular when SG proteins belong to superfamilies that comprise several proteins with different functions. For example, the SG proteins 1vme and 2gcu both belong to the large superfamily of metallo-β-lactamases (corresponding to 3.60.15.10 in CATH); however, the Fe-site of the former most closely resembles that of the nitric oxide reductase 1ycf,[79] whereas the Fe-site of the latter is most similar to

Table 3. Results of the comparison of the Fe-sites of structural genomics (SG) proteins with the Fe-sites of proteins with known function

SG protein	SG annotation	Best hit protein	Best hit annotation	FAST score
1jr7 (3.60.130.10)	Hypothetical 37.4 kDa protein in IleY-GabD intergenic region	2og6 (3.60.130.10)	Asparagine oxygenase	3.48
1t71 (3.60.21.30)	Phosphatase	-	-	-
1vme (3.60.15.10)	Flavoprotein	1ycf (3.60.15.10)	Nitric oxide reductase	9.21
1vrb (3.60.130.10)	Putative asparaginyl hydroxylase	1mzf (3.60.130.10)	Asparaginyl hydroxylase	2.82
2amu (2.60.40.730)	Putative superoxide reductase	1dqk (2.60.40.730)	Superoxide reductase	8.48
2csg (3.60.130.10)	Putative cytoplasmic protein	2cgo (3.60.130.10)	Asparaginyl hydroxylase	3.42
2gcu (3.60.15.10)	Putative hydroxyacylglutathione hydrolase	2obw (3.60.15.10)	Hydroxyacylglutathione hydrolase	4.93
2gm6 (2.60.120.10)	Cysteine dioxygenase type I	2ic1 (2.60.120.10)	Cysteine dioxygenase type I	6.17
2gyq (1.20.1260.10)	YcfI, putative structural protein	1lkp (1.20.1260.10)	Rubrerythrin	5.99
2hvb (2.60.40.730)	Superoxide reductase	1do6 (2.60.40.730)	Superoxide reductase	7.93
2itb (1.20.1260.10)	tRNA-(ms(2)io(6)a)-hydroxylase, putative	1mxr (1.10.620.20)	Ribonucleotide reductase R2	8.03
2o08 (a.211.1)	Bh1327 protein	2huo (a.211.1)	Inositol oxygenase	3.59
2oc5 (1.10.620.20)	Hypothetical protein	2av8 (1.10.620.20)	Ribonucleotide reductase R2	4.38
2ogi (a.211.1)	Hypothetical protein sag1661	2huo (a.211.1)	Inositol oxygenase	3.34
2p17 (2.60.120.10)	Pirin-like protein	2cgo (3.60.130.10)	Asparaginyl hydroxylase	2.53

The table reports (i) the PDB code and the CATH or SCOP assignment (in parentheses) of the SG protein; (ii) the annotation of the SG protein in the PDB; (iii) the PDB code and the CATH or SCOP assignment (in parentheses) of the best hit protein, i.e. the protein that contains the Fe-site most similar to the Fe-site of the SG protein; (iv) the annotation of the best hit protein in the PDB, which represents the functional prediction for the SG protein; and (v) the FAST score of the comparison between the Fe-site of the SG protein and the Fe-site of the best hit protein.

that of the hydroxyacylglutathione hydrolase 2obw (Table 3).[80] The superimposition of the Fe-sites of 1vme and 1ycf shows that in addition to having identical first coordination spheres, they share very similar residues in the second coordination sphere that have been implicated in the interaction with the NO substrate (Fig. 11, top panel).[79] Similarly, the superimposition of the Fe-sites of 2gcu and 2obw reveals their close relationship, which includes a number of H-bonding interactions between first-shell and second-shell residues[80] (Fig. 11, middle panel). Another example is given by the SG protein 2itb, which belongs to a large superfamily of ferritin-like proteins (corresponding to 1.20.1260.10 in CATH). The Fe-site of 2itb is most similar to that of the ribonucleotide reductase R2 1mxr,[54] which in fact belongs to a different (though structurally related, see Fig. 9) CATH superfamily (1.10.620.20). The superimposition of these Fe-sites shows that they share important features such as H-bonding interactions and, importantly, a tyrosine residue that in 1mxr is known to harbor a stable radical playing a central role in the catalysis (Fig. 11, bottom panel),[54] suggesting that a similar radical mechanism may be active in 2itb.

Concluding remarks

Here, we described an approach to the study of metal sites of proteins for which structural information is available. To show the potential of the approach, we applied it to non-heme iron, but it could be used to investigate any other metal. The core of the approach is to represent metal sites as three-dimensional templates that include the first and the second coordination sphere of the metal, so as to incorporate the characteristics of the protein environment around the metal. These templates are automatically extracted from PDB structures and compared with structure alignment programs to identify and evaluate similarities between metal sites in a highly automated fashion. Since metals bound to proteins typically have essential roles, this comparative analysis of metal sites can provide useful hints on protein function and evolution that add to the information obtainable by the analysis of whole protein structures, as contained in CATH and SCOP databases. On one hand, in fact, proteins that are structurally and evolutionarily unrelated according to CATH and SCOP classifications can contain similar metal sites, possibly resulting from convergent evolution. In the case of non-heme iron, our approach revealed that at least 17% of the sites found in unrelated proteins are highly similar. On the other hand, proteins that share a common structure and evolutionary origin according to CATH and SCOP classifications can contain non-identical metal sites, which can account for functional differences resulting from divergent evolution. As an example, we showed that functional variation across a large superfamily of iron-dependent enzymes is associated with fine differences in iron coordination in the active site. Furthermore, function prediction of unannotated metalloproteins based on standard similarity searches that use the whole protein structure can be usefully complemented by similarity searches that employ our structural templates, in that they focus on the metal site that is, in many cases, responsible for function. In the case of non-heme iron, we obtained functional hints for 14 of 15 proteins with unknown function, which were especially useful when the whole protein structure was common to many proteins with different functions. Although we found no example in non-heme iron proteins, we expect that these searches can help when the whole protein structure bears no similarity to known proteins, because, as

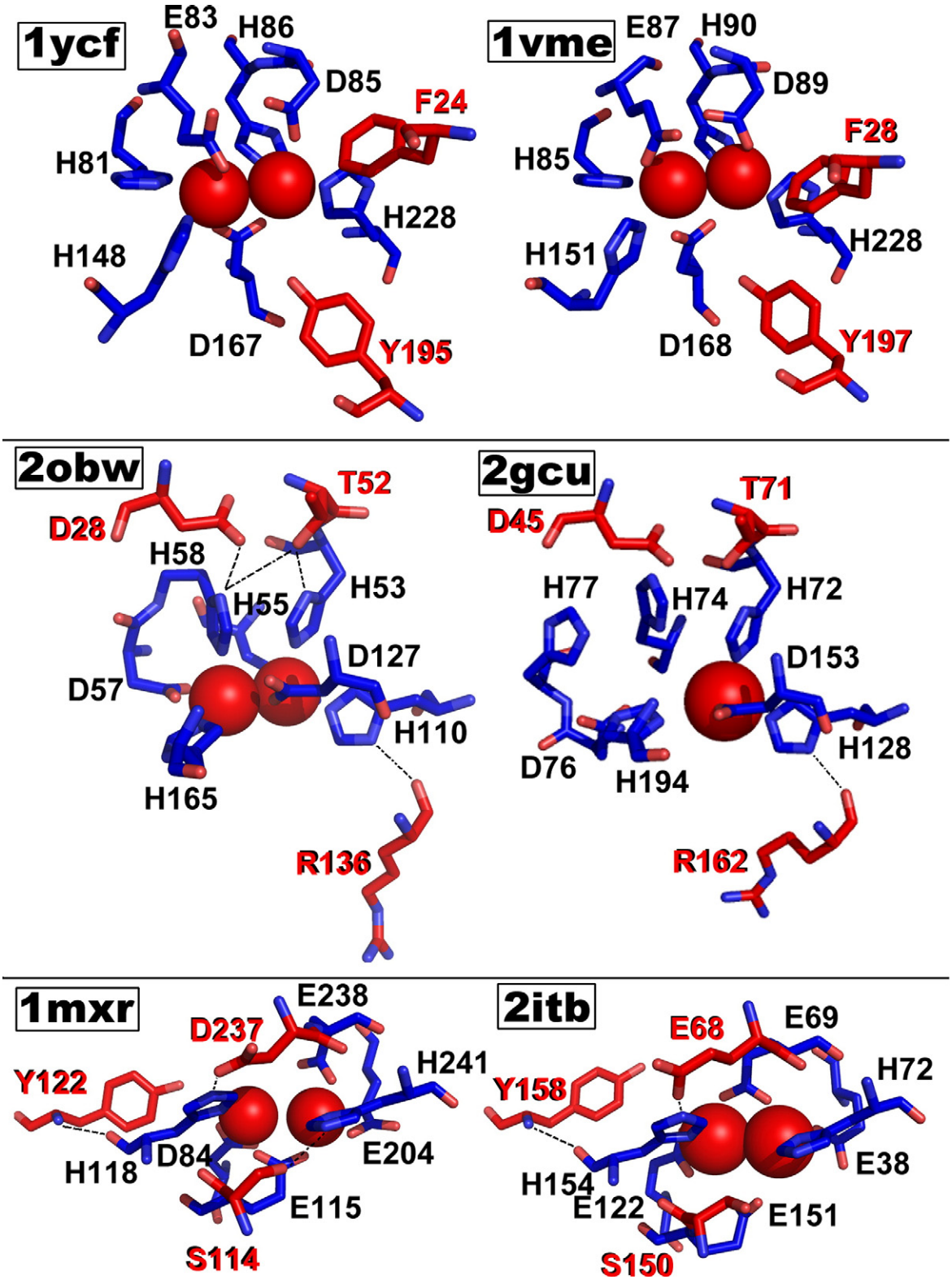

Fig. 11. Examples of Fe-sites of proteins with known function (left) that are most similar to Fe-sites of structural genomics proteins (right), shown in the same orientation. Residues of the first and the second coordination sphere are shown as blue sticks and red sticks, respectively. Iron ions are shown as red spheres.

noted above, similar metal sites can be present in structurally unrelated proteins. Finally, we note that our templates can be used to organize and classify metal sites in proteins in a systematic way. In this perspective, the dataset on non-heme iron sites compiled in this work can be regarded as the initial core of a comprehensive information resource for metalloproteins.

Methods

Selection of the PDB structures of non-heme iron proteins

The PDB Chemical Component Dictionary† describes all the chemical components found in PDB structures, which are generally called residues and are identified by three-character ID codes. As of December 2007, this dictionary included 84 residues containing one or more iron atoms (Supplementary Data Table S1). We used the information available at Ligand Expo‡ to select the 42 residues containing non-heme iron (Supplementary Data Table S1), excluding, in addition to heme groups, siderophores (such as pyochelin in the PDB structure 1xkw) and synthetic iron-containing compounds (such as the ferrocene derivative in the PDB structure 1a3l). We then downloaded from the PDB all the experimentally determined protein structures containing at least one of the selected residues, thus obtaining a set of 1274 structures.

Generation of the structural templates for non-heme iron sites from the PDB structures

We used an in-house Perl program to generate from the PDB structures of non-heme iron proteins the structural templates that represent the non-heme iron sites present in those proteins. For every residue containing non-heme iron, as identified in the previous step (see above), the program extracted the PDB coordinates of (i) all the iron atoms in the residue, (ii) all the atoms in protein and non-protein residues having a non-hydrogen atom distant <3.0 Å from any atom in (i), and (iii) all the atoms in protein residues not included in (ii) and having a non-hydrogen atom distant <5.0 Å from any non-hydrogen atom in (ii). The sets (ii) and (iii) define the first and the second coordination sphere, respectively (Fig. 3). This procedure resulted in the generation of a total of 2034 structural templates for non-heme iron sites.

Structure-based classification of non-heme iron sites

We used the CATH database (version 3.1.0)[11] as the basis to classify the 2034 non-heme iron sites identified as above according to the protein domain in which they are found. First, we classified the sites for which we were able to map all the protein residues included in the first coordination sphere onto a structural domain classified in CATH, using the distinct CATH codes associated to these domains to designate the sites. For example, in the 1ltz structure (Fig. 3) the iron ligands His138, His143, and Glu184 are all located in the same CATH domain with code 1.10.800.10, therefore we classified this iron site as 1.10.800.10. Because not all the PDB structures are included in the CATH database, 590 sites remained unassigned.

Then, we repeated the same procedure using the SCOP database (release 1.73)[12] as the reference for classification, and by comparing SCOP and CATH assignments we picked out SCOP codes corresponding (within our dataset) to a single CATH code. We were thus able to assign a CATH code to further 176 sites, which despite being in structures not included in the CATH database were assigned a SCOP code univocally associated with that CATH code. Moreover, we used SCOP codes to classify an additional 100 sites.

To classify the 314 sites still unassigned, we generated from the PDB structures of non-heme iron proteins a new set of structural templates, one for every non-heme iron site, both assigned and non-assigned. Each of these templates includes the PDB coordinates of all the atoms in protein residues having an atom distant <22 Å from the non-heme iron (whose coordinates are averaged over the iron atoms in the site), and roughly represents the protein domain containing the site. This definition is based on the simplifying assumptions that (i) the non-heme iron site is located approximately at the center of the domain, and that (ii) the volume of the domain is equal to that of a randomly packed spherical assembly consisting of N residues of the same size, which is given by:

$$V = \left(4N\pi a^3\right)/3\alpha$$

where a (the radius of a residue) is 3.5 Å and α (the packing ratio) is 0.64.[81] We set $N=159$, the average number of residues in a CATH domain,[82] thus obtaining $V=44618$ Å^3, which corresponds to a radius for the domain of 22 Å. We then compared each domain template centered on an unassigned site to every domain template centered on an assigned (either to a CATH or to a SCOP code) site using the program FAST,[83] which is a program for aligning protein structures that uses a directionality-based scoring scheme to compare the intramolecular residue-residue relationships in two structures, and was shown to be more robust and much faster than other widely used algorithms.[83] It yields similarity scores that are normalized against the numbers of residues in the two proteins and indicate significant structural similarity when >1.5§. We attributed to an unassigned site the code of the site that yielded the highest similarity score in pair-wise FAST comparisons between the domain template centered on that unassigned site and all domain templates centered on an assigned site, if the score was >1.5. This approach allowed us to assign a further 287 sites, leaving out only 27. Finally, we used FAST to carry out an all-*versus*-all comparison of the domain templates centered on these residual 27 sites, which we grouped into 23 unclassified domains using a single linkage clustering strategy and 1.5 as the threshold similarity score for clustering.

Grouping of non-heme iron sites

We first grouped the non-heme iron sites according to their associated code, representing the protein domain in which they are found. We then used FAST (see above) to carry out an all-*versus*-all comparison of the site templates within each group, and we divided them into subgroups using a single linkage clustering strategy and 1.5 as the threshold similarity score for clustering. As a result of this procedure, distinct sites found in the same protein domain were placed in different subgroups. For example, the tetrahedral Fe(Cys)$_4$ and the square pyramidal Fe(Cys)(His)$_4$ center of desulfoferrodoxin were placed in the 2.60.40.730_1 and in the 2.60.40.730_2 subgroup, respectively (Supplementary Data Fig. S2). We manually analyzed the composition of each subgroup to identify those comprising non-physiological non-heme iron sites, based on data in the literature. The clustering of sites into

† http://deposit.pdb.org/het_dictionary.txt
‡ http://ligand-expo.rcsb-org

§ http://biowulf.bu.edu/FAST

subgroups makes the detection of non-physiological sites such as those due to metal ions that remain attached to the protein surface during crystallization, easier in that such sites are typically spotted out as the only members of distinct subgroups. We identified a total of 128 non-physiological sites, which we removed from our dataset. We thus obtained a final set of 1906 non-heme iron sites, grouped into 86 subgroups, all of which are given in Supplementary Data Table S2.

Selection, analysis and comparison of representative non-heme iron sites

For every subgroup of non-heme iron sites identified as above, we selected a single representative template by choosing the PDB structure with the best X-ray resolution. NMR structures were discarded in favor of X-ray structures, unless they were the only representatives for a given subgroup. We checked that the representative templates did not contain engineered mutations.

We used the representative templates to evaluate the occurrence of different amino acid residues in the first and the second coordination sphere of non-heme iron sites (see above for their definition). For the first coordination sphere, we grouped all the protein residues involved in iron coordination by amino acid identity and ligand atom type, and we calculated the relative frequency of each group with respect to the total number of protein residues involved in iron coordination. For the second coordination sphere, we grouped all the protein residues present in that sphere by amino acid identity, and we calculated the relative frequency of each amino acid with respect to the total number of protein residues present in the second coordination spheres. Then, we divided these frequencies by the corresponding average amino acid frequencies in proteins, as determined from the complete UniProtKB/Swiss-Prot database‖.[84]

We performed an all-*versus*-all comparison of the representative site templates using the FAST program (see above). The distribution of all pair-wise similarity scores obtained from FAST and a heat map representation of them are shown in Supplementary Data Figs. S3 and S4, respectively. We used the 99th, 97th and 95th percentiles (i.e., the FAST scores below which 99%, 97%, and 95% of non-zero scores fall in the distribution) to define three different thresholds for high, medium, and low similarity, respectively, and we used these thresholds (2.25, 1.30, and 1.02, respectively) for clustering the representative Fe-sites by applying single-linkage clustering.

We used the CLUSTAL W program[85] to perform pair-wise sequence alignments of the protein domains containing the representative sites included in each cluster identified as above.

Acknowledgements

This work was supported by the Ministero Italiano dell'Università e della Ricerca (MIUR) through the FIRB Project RBLA032ZM7, by the European Union through the EU-NMR Contract 026145, and by Ente Cassa di Risparmio di Firenze. R.J.N. was supported by NIH grant GM62414, US DOE under contract (W-31-109-ENG38). We acknowledge support from the EMBL.

‖ http://www.expasy.ch/sprot/relnotes/relstat.html

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2009.02.052

References

1. Frausto da Silva, J. J. R. & Williams, R. J. P. (2001). *The Biological Chemistry of the Elements: the Inorganic Chemistry of Life.* Oxford University Press, New York, NY.
2. Bertini, I., Gray, H. B., Stiefel, E. I. & Valentine, J. S. (2006). *Biological Inorganic Chemistry.* University Science Books, Sausalito, CA.
3. Holm, R. H., Kennepohl, P. & Solomon, E. I. (1996). Structural and functional aspects of metal sites in biology. *Chem. Rev.* **96**, 2239–2314.
4. Castagnetto, J. M., Hennessy, S. W., Roberts, V. A., Getzoff, E. D., Tainer, J. A. & Piquet, M. E. (2002). MDB: the metalloprotein database and browser at the Scripps Research Institute. *Nucleic Acids Res.* **30**, 379–382.
5. Degtyarenko, K. N., North, A. C., Perkins, D. N. & Findlay, J. B. (1998). PROMISE: a database of information on prosthetic centres and metal ions in protein active sites. *Nucleic Acids Res.* **26**, 376–381.
6. Degtyarenko, K. (2000). Bioinorganic motifs: towards functional classification of metalloproteins. *Bioinformatics*, **16**, 851–864.
7. Laskowski, R. A., Watson, J. D. & Thornton, J. M. (2005). ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res.* **33**, W89–W93.
8. Dudev, T., Lin, Y. L., Dudev, M. & Lim, C. (2003). First-second shell interactions in metal binding sites in proteins: a PDB survey and DFT/CDM calculations. *J. Am. Chem. Soc.* **125**, 3168–3180.
9. Dudev, T. & Lim, C. (2008). Metal binding affinity and selectivity in metalloproteins: insights from computational studies. *Annu. Rev. Biophys.* **37**, 97–116.
10. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al.* (2000). The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242.
11. Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). CATH- a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.
12. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540.
13. Costas, M., Mehn, M. P., Jensen, M. P. & Que, L., Jr (2004). Dioxygen activation at mononuclear nonheme iron active sites: enzymes, models, and intermediates. *Chem Rev.* **104**, 939–986.
14. Nam, W. (2007). High-valent iron(IV)-oxo complexes of heme and non-heme ligands in oxygenation reactions. *Accts Chem. Res.* **40**, 522–531.
15. Posey, J. E. & Gherardini, F. C. (2000). Lack of a role for iron in the Lyme disease pathogen. *Science*, **288**, 1651–1653.

16. Bertini, I., Sigel, A. & Sigel, H. (2001). *Handbook on Metalloproteins.* Marcel Dekker, New York, NY.
17. Orengo, C. A. & Thornton, J. M. (2005). Protein families and their evolution – a structural perspective. *Annu. Rev. Biochem.* **74**, 867–900.
18. Peters, J. W., Lanzilotta, W. N., Lemon, B. J. & Seefeldt, L. C. (1998). X-ray crystal structure of the Fe-only hydrogenase (CpI) from Clostridium pasteurianum to 1.8 Å resolution. *Science*, **282**, 1853–1858.
19. Berkovitch, F., Nicolet, Y., Wan, J. T., Jarrett, J. T. & Drennan, C. L. (2004). Crystal structure of biotin synthase, an S-adenosylmethionine-dependent radical enzyme. *Science*, **303**, 76–79.
20. Broach, R. B. & Jarrett, J. T. (2006). Role of the $[2Fe\text{-}2S]^{2+}$ cluster in biotin synthase: mutagenesis of the atypical metal ligand arginine 260. *Biochemistry*, **45**, 14166–14174.
21. Nagashima, S., Nakasako, M., Dohmae, N., Tsujimura, M., Takio, K., Odaka, M. *et al.* (1998). Novel non-heme iron center of nitrile hydratase with a claw setting of oxygen atoms. *Nat. Struct. Biol.* **5**, 347–351.
22. Wu, X. H., Quan, J. M. & Wu, Y. D. (2007). Theoretical study of the catalytic mechanism and metal-ion dependence of peptide deformylase. *J. Phys. Chem. B*, **111**, 6236–6244.
23. Dey, A., Jenney, F. E., Jr, Adams, M. W., Johnson, M. K., Hodgson, K. O., Hedman, B. & Solomon, E. I. (2007). Sulfur K-edge X-ray absorption spectroscopy and density functional theory calculations on superoxide reductase: role of the axial thiolate in reactivity. *J. Am. Chem. Soc.* **129**, 12418–12431.
24. Beinert, H., Holm, R. H. & Munck, E. (1997). Iron-sulfur clusters: nature's modular, multipurpose structures. *Science*, **277**, 653–659.
25. Johnson, D. C., Dean, D. R., Smith, A. D. & Johnson, M. K. (2005). Structure, function, and formation of biological iron-sulfur clusters. *Annu. Rev. Biochem.* **74**, 247–281.
26. Koehntop, K. D., Emerson, J. P. & Que, L., Jr (2005). The 2-His-1-carboxylate facial triad: a versatile platform for dioxygen activation by mononuclear non-heme iron(II) enzymes. *J. Biol. Inorg. Chem.* **10**, 87–93.
27. Greenwood, N. N. & Earnshaw, A. (1984). *Chemistry of the Elements.* Pergamon Press, Oxford.
28. Yamashita, M. M., Wesson, L., Eisenman, G. & Eisenberg, D. (1990). Where metal ions bind in proteins. *Proc. Natl Acad. Sci. USA*, **87**, 5648–5652.
29. Winkler, J. R. (2000). Electron tunneling pathways in proteins. *Curr. Opin. Chem. Biol.* **4**, 192–198.
30. Gray, H. B. & Winkler, J. R. (2005). Long-range electron transfer. *Proc. Natl Acad. Sci. USA*, **102**, 3534–3539.
31. Shih, C., Museth, A. K., Abrahamsson, M., Blanco-Rodriguez, A. M., Di Bilio, A. J., Sudhamsu, J. *et al.* (2008). Tryptophan-accelerated electron flow through proteins. *Science*, **320**, 1760–1762.
32. Ma, J. C. & Dougherty, D. A. (1997). The cation-pi interaction. *Chem. Rev.* **97**, 1303–1324.
33. Zaric, S. D., Popovic, D. M. & Knapp, E. W. (2000). Metal ligand aromatic cation-pi interactions in metalloproteins: ligands coordinated to metal interact with aromatic residues. *Chemistry*, **6**, 3935–3942.
34. Rowland, P., Norager, S., Jensen, K. F. & Larsen, S. (2000). Structure of dihydroorotate dehydrogenase B: electron transfer between two flavin groups bridged by an iron-sulphur cluster. *Structure*, **8**, 1227–1238.
35. Dobbek, H., Gremer, L., Kiefersauer, R., Huber, R. & Meyer, O. (2002). Catalysis at a dinuclear [CuSMo(=O)OH] cluster in a CO dehydrogenase resolved at 1.1-Å resolution. *Proc. Natl Acad. Sci. USA*, **99**, 15971–15976.
36. Matsubara, H. & Saeki, K. (1992). Structural and functional diversity of ferredoxins and related proteins. *Adv. Inorg. Chem.* **38**, 223–280.
37. Sticht, H. & Rosch, P. (1998). The structure of iron-sulfur proteins. *Prog. Biophys. Mol. Biol.* **70**, 95–136.
38. Bonisch, H., Schmidt, C. L., Schafer, G. & Ladenstein, R. (2002). The structure of the soluble domain of an archaeal Rieske iron-sulfur protein at 1.1 Å resolution. *J. Mol. Biol.* **319**, 791–805.
39. Smith, J. L., Zhang, H., Yan, J., Kurisu, G. & Cramer, W. A. (2004). Cytochrome bc complexes: a common core of structure and function surrounded by diversity in the outlying provinces. *Curr. Opin. Struct. Biol.* **14**, 432–439.
40. Sieker, L. C., Stenkamp, R. E. & LeGall, J. (1994). Rubredoxin in crystalline state. *Methods Enzymol.* **243**, 203–216.
41. deMare, F., Kurtz, D. M., Jr & Nordlund, P. (1996). The structure of Desulfovibrio vulgaris rubrerythrin reveals a unique combination of rubredoxin-like FeS4 and ferritin-like diiron domains. *Nature Struct. Biol.* **3**, 539–546.
42. Luke, I., Butland, G., Moore, K., Buchanan, G., Lyall, V., Fairhurst, S. A. *et al.* (2008). Biosynthesis of the respiratory formate dehydrogenases from Escherichia coli: characterization of the FdhE protein. *Arch. Microbiol.* **190**, 685–696.
43. Iwata, S., Saynovits, M., Link, T. A. & Michel, H. (1996). Structure of a water soluble fragment of the 'Rieske' iron-sulfur protein of the bovine heart mitochondrial cytochrome bc1 complex determined by MAD phasing at 1.5 Å resolution. *Structure*, **4**, 567–579.
44. Meyer, J., Gagnon, J., Gaillard, J., Lutz, M., Achim, C., Munck, E. *et al.* (1997). Assembly of a $[2Fe\text{-}2S]^{2+}$ cluster in a molecular variant of Clostridium pasteurianum rubredoxin. *Biochemistry*, **36**, 13374–13380.
45. Iwasaki, T., Kounosu, A., Tao, Y., Li, Z., Shokes, J. E., Cosper, N. J. *et al.* (2005). Rational design of a mononuclear metal site into the archaeal Rieske-type protein scaffold. *J. Biol. Chem.* **280**, 9129–9134.
46. Au, K. G., Clark, S., Miller, J. H. & Modrich, P. (1989). Escherichia coli mutY gene encodes an adenine glycosylase active on G-A mispairs. *Proc. Natl Acad. Sci. USA*, **86**, 8877–8881.
47. Thayer, M. M., Ahern, H., Xing, D., Cunningham, R. P. & Tainer, J. A. (1995). Novel DNA binding motifs in the DNA repair enzyme endonuclease III crystal structure. *EMBO J.* **14**, 4108–4120.
48. Dobritzsch, D., Ricagno, S., Schneider, G., Schnackerz, K. D. & Lindqvist, Y. (2002). Crystal structure of the productive ternary complex of dihydropyrimidine dehydrogenase with NADPH and 5-iodouracil. Implications for mechanism of inhibition and electron transfer. *J. Biol. Chem.* **277**, 13155–13166.
49. Leys, D., Basran, J., Talfournier, F., Sutcliffe, M. J. & Scrutton, N. S. (2003). Extensive conformational sampling in a ternary electron transfer complex. *Nature Struct. Biol.* **10**, 219–225.
50. Fukuyama, K., Okada, T., Kakuta, Y. & Takahashi, Y. (2002). Atomic resolution structures of oxidized [4Fe-4S] ferredoxin from Bacillus thermoproteolyticus in two crystal forms: systematic distortion of [4Fe-4S] cluster in the protein. *J. Mol. Biol.* **315**, 1155–1166.
51. Lee, T. T., Agarwalla, S. & Stroud, R. M. (2004). Crystal structure of RumA, an iron-sulfur cluster containing

E. coli ribosomal RNA 5-methyluridine methyltransferase. *Structure*, **12**, 397–407.
52. Ogata, H., Hirota, S., Nakahara, A., Komori, H., Shibata, N., Kato, T. *et al.* (2005). Activation process of [NiFe] hydrogenase elucidated by high-resolution X-ray analyses: conversion of the ready to the unready state. *Structure*, **13**, 1635–1642.
53. Watanabe, S., Matsumi, R., Arai, T., Atomi, H., Imanaka, T. & Miki, K. (2007). Crystal structures of [NiFe] hydrogenase maturation proteins HypC, HypD, and HypE: insights into cyanation reaction by thiol redox signaling. *Mol. Cell*, **27**, 29–40.
54. Hogbom, M., Galander, M., Andersson, M., Kolberg, M., Hofbauer, W., Lassmann, G. *et al.* (2003). Displacement of the tyrosyl radical cofactor in ribonucleotide reductase obtained by single-crystal high-field EPR and 1.4-Å X-ray data. *Proc. Natl Acad. Sci. USA*, **100**, 3209–3214.
55. Iyer, R. B., Silaghi-Dumitrescu, R., Kurtz, D. M., Jr. & Lanzilotta, W. N. (2005). High-resolution crystal structures of Desulfovibrio vulgaris (Hildenborough) nigerythrin: facile, redox-dependent iron movement, domain interface variability, and peroxidase activity in the rubrerythrins. *J. Biol. Inorg. Chem.* **10**, 407–416.
56. Sheriff, S., Hendrickson, W. A. & Smith, J. L. (1987). Structure of myohemerythrin in the azidomet state at 1.7/1.3 Å resolution. *J. Mol. Biol.* **197**, 273–296.
57. Schwarzenbacher, R., Stenner-Liewen, F., Liewen, H., Robinson, H., Yuan, H., Bossy-Wetzel, E. *et al.* (2004). Structure of the Chlamydia protein CADD reveals a redox enzyme that modulates host cell apoptosis. *J. Biol. Chem.* **279**, 29320–29324.
58. Tomchick, D. R., Phan, P., Cymborowski, M., Minor, W. & Holman, T. R. (2001). Structural and functional characterization of second-coordination sphere mutants of soybean lipoxygenase-1. *Biochemistry*, **40**, 7509–7517.
59. Koepke, J., Krammer, E. M., Klingen, A. R., Sebban, P., Ullmann, G. M. & Fritzsch, G. (2007). pH modulates the quinone position in the photosynthetic reaction center from *Rhodobacter sphaeroides* in the neutral and charge separated states. *J. Mol. Biol.* **371**, 396–409.
60. Liu, J., Lou, Y., Yokota, H., Adams, P. D., Kim, R. & Kim, S. H. (2005). Crystal structure of a PhoU protein homologue: a new class of metalloprotein containing multinuclear iron clusters. *J. Biol. Chem.* **280**, 15960–15966.
61. Nordlund, P. & Eklund, H. (1995). Di-iron-carboxylate proteins. *Curr. Opin. Struct. Biol.* **5**, 758–766.
62. Chan, M. K., Mukund, S., Kletzin, A., Adams, M. W. & Rees, D. C. (1995). Structure of a hyperthermophilic tungstopterin enzyme, aldehyde ferredoxin oxidoreductase. *Science*, **267**, 1463–1469.
63. Erlandsen, H., Kim, J. Y., Patch, M. G., Han, A., Volner, A., Abu-Omar, M. M. & Stevens, R. C. (2002). Structural comparison of bacterial and human iron-dependent phenylalanine hydroxylases: similar fold, different stability and reaction rates. *J. Mol. Biol.* **320**, 645–661.
64. Kreusch, A., Spraggon, G., Lee, C. C., Klock, H., McMullan, D., Ng, K. *et al.* (2003). Structure analysis of peptide deformylases from *Streptococcus pneumoniae*, *Staphylococcus aureus*, *Thermotoga maritima* and *Pseudomonas aeruginosa*: snapshots of the oxygen sensitivity of peptide deformylase. *J. Mol. Biol.* **330**, 309–321.
65. Urich, T., Gomes, C. M., Kletzin, A. & Frazao, C. (2006). X-ray structure of a self-compartmentalizing sulfur cycle metalloenzyme. *Science*, **311**, 996–1000.
66. Zhang, Z., Ren, J., Stammers, D. K., Baldwin, J. E., Harlos, K. & Schofield, C. J. (2000). Structural origins of the selectivity of the trifunctional oxygenase clavaminic acid synthase. *Nature Struct. Biol.* **7**, 127–133.
67. Elkins, J. M., Rutledge, P. J., Burzlaff, N. I., Clifton, I. J., Adlington, R. M., Roach, P. L. & Baldwin, J. E. (2003). Crystallographic studies on the reaction of isopenicillin N synthase with an unsaturated substrate analogue. *Org. Biomol. Chem.* **1**, 1455–1460.
68. Simmons, C. R., Liu, Q., Huang, Q., Hao, Q., Begley, T. P., Karplus, P. A. & Stipanuk, M. H. (2006). Crystal structure of mammalian cysteine dioxygenase. A novel mononuclear iron center for cysteine thiol oxidation. *J. Biol. Chem.* **281**, 18723–18733.
69. Clifton, I. J., McDonough, M. A., Ehrismann, D., Kershaw, N. J., Granatino, N. & Schofield, C. J. (2006). Structural studies on 2-oxoglutarate oxygenases and related double-stranded beta-helix fold proteins. *J. Inorg. Biochem.* **100**, 644–669.
70. Jeong, J. J., Fushinobu, S., Ito, S., Jeon, B. S., Shoun, H. & Wakagi, T. (2003). Characterization of the cupin-type phosphoglucose isomerase from the hyperthermophilic archaeon Thermococcus litoralis. *FEBS Lett.* **535**, 200–204.
71. Bowater, L., Fairhurst, S. A., Just, V. J. & Bornemann, S. (2004). Bacillus subtilis YxaG is a novel Fe-containing quercetin 2,3-dioxygenase. *FEBS Lett.* **557**, 45–48.
72. Higgins, L. J., Yan, F., Liu, P., Liu, H. W. & Drennan, C. L. (2005). Structural insight into antibiotic fosfomycin biosynthesis by a mononuclear iron enzyme. *Nature*, **437**, 838–844.
73. Titus, G. P., Mueller, H. A., Burgner, J., Rodriguez, D. C., Penalva, M. A. & Timm, D. E. (2000). Crystal structure of human homogentisate dioxygenase. *Nature Struct. Biol.* **7**, 542–546.
74. Zhang, Y., Colabroy, K. L., Begley, T. P. & Ealick, S. E. (2005). Structural studies on 3-hydroxyanthranilate-3,4-dioxygenase: the catalytic mechanism of a complex oxidation involved in NAD biosynthesis. *Biochemistry*, **44**, 7632–7643.
75. Pang, H., Bartlam, M., Zeng, Q., Miyatake, H., Hisano, T., Miki, K. *et al.* (2004). Crystal structure of human pirin: an iron-binding nuclear protein and transcription cofactor. *J. Biol. Chem.* **279**, 1491–1498.
76. Adams, M. & Jia, Z. (2005). Structural and biochemical analysis reveal pirins to possess quercetinase activity. *J. Biol. Chem.* **280**, 28675–28682.
77. Shannon, R. D. (1976). Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides. *Acta Crystallogr. A*, **32**, 751–767.
78. Jin, S., Kurtz, D. M., Jr, Liu, Z. J., Rose, J. & Wang, B. C. (2002). X-ray crystal structures of reduced rubrerythrin and its azide adduct: a structure-based mechanism for a non-heme diiron peroxidase. *J. Am. Chem. Soc.* **124**, 9845–9855.
79. Silaghi-Dumitrescu, R., Kurtz, D. M., Jr, Ljungdahl, L. G. & Lanzilotta, W. N. (2005). X-ray crystal structures of Moorella thermoacetica FprA. Novel diiron site structure and mechanistic insights into a scavenging nitric oxide reductase. *Biochemistry*, **44**, 6492–6501.
80. Campos-Bermudez, V. A., Leite, N. R., Krog, R., Costa-Filho, A. J., Soncini, F. C., Oliva, G. & Vila, A. J. (2007). Biochemical and structural characterization of Salmonella typhimurium glyoxalase II:

new insights into metal ion selectivity. *Biochemistry*, **46**, 11069–11079.
81. Shen, M. Y., Davis, F. P. & Sali, A. (2005). The optimal size of a globular protein domain: a simple sphere-packing model. *Chem. Phys. Lett.* **405**, 224–228.
82. Greene, L. H., Lewis, T. E., Addou, S., Cuff, A., Dallman, T., Dibley, M. *et al.* (2007). The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Res.* **35**, D291–D297.
83. Zhu, J. & Weng, Z. (2005). FAST: a novel protein structure alignment algorithm. *Proteins: Struct. Funct. Genet.* **58**, 618–627.
84. The universal protein resource (UniProt). *Nucleic Acids Res.* **36**. (2008)., D190–D195.
85. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680.