BMC Structural Biology

BioMed Central

Research article

Open Access

COMe: the ontology of bioinorganic proteins

Kirill Degtyarenko* and Sergio Contrino

Address: European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambs CB10 1SD, United Kingdom

Email: Kirill Degtyarenko* - kirill@ebi.ac.uk; Sergio Contrino - contrino@ebi.ac.uk

* Corresponding author

Published: 27 February 2004

BMC Structural Biology 2004, **4**:3

Received: 18 September 2003
Accepted: 27 February 2004

This article is available from: http://www.biomedcentral.com/1472-6807/4/3

Abstract

Background: Many characterised proteins contain metal ions, small organic molecules or modified residues. In contrast, the huge amount of data generated by genome projects consists exclusively of sequences with almost no annotation. One of the goals of the structural genomics initiative is to provide representative three-dimensional (3-D) structures for as many protein/domain folds as possible to allow successful homology modelling. However, important functional features such as metal co-ordination or a type of prosthetic group are not always conserved in homologous proteins. So far, the problem of correct annotation of bioinorganic proteins has been largely ignored by the bioinformatics community and information on bioinorganic centres obtained by methods other than crystallography or NMR is only available in literature databases.

Results: COMe (Co-Ordination of Metals) represents the ontology for bioinorganic and other small molecule centres in complex proteins. COMe consists of three types of entities: 'bioinorganic motif' (BIM), 'molecule' (MOL), and 'complex proteins' (PRX), with each entity being assigned a unique identifier. A BIM consists of at least one centre (metal atom, inorganic cluster, organic molecule) and two or more endogenous and/or exogenous ligands. BIMs are represented as one-dimensional (1-D) strings and 2-D diagrams. A MOL entity represents a 'small molecule' which, when in complex with one or more polypeptides, forms a functional protein. The PRX entities refer to the functional proteins as well as to separate protein domains and subunits. The complex proteins in COMe are subdivided into three categories: (i) metalloproteins, (ii) organic prosthetic group proteins and (iii) modified amino acid proteins. The data are currently stored in both XML format and a relational database and are available at http://www.ebi.ac.uk/come/.

Conclusion: COMe provides the classification of proteins according to their 'bioinorganic' features and thus is orthogonal to other classification schemes, such as those based on sequence similarity, 3-D fold, enzyme activity, or biological process. The hierarchical organisation of the controlled vocabulary allows both for annotation and querying at different levels of granularity.

Background

Many characterised proteins contain metal ions, small organic molecules or modified residues. In contrast, the huge amount of data generated by genome projects consists exclusively of sequences with almost no annotation. One of the goals of the structural genomics initiative is to provide representative three-dimensional (3-D) structures for as many protein/domain folds as possible to allow successful homology modelling [1]. However, important functional features such as metal co-ordination or the type of prosthetic group are not always conserved in homologous proteins.

So far, the problem of correct annotation of bioinorganic proteins has been largely ignored by the bioinformatics community. The only comprehensive database of metal sites in proteins, Metalloprotein Database and Browser (MDB) [2], is automatically built from the structures available at the Protein Data Bank (PDB) [3]. Although crystallography is the single most informative method for studying protein structure, it has a number of limitations as far as bioinorganic chemists are concerned [4,5]. Many structures deposited at the PDB contain metal ions and molecules that are not present in native proteins. On the other hand, the information on bioinorganic/small molecule centres obtained by other (mostly spectroscopic) methods is not available to the scientific community in any form apart from literature databases.

'Ontology' is a formal definition of concepts (such as entities and relationships) of a given area of knowledge, described in a standardised form [6]. It can be organised as a structured vocabulary in the form of a directed acyclic graph or a network in which each term may be a 'child' of one or more 'parent' [7]. In this paper, we describe COMe (Co-Ordination of Metals), the ontology for bioinorganic proteins and their features.

A previous version of this manuscript was made available before peer review at http://preprint.chemweb.com/biochem/0307002/.

Results

COMe version 4.01 contains data on 1280 'bioinorganic proteins', 470 'bioinorganic motifs' and 174 'molecules'. The data exist in two formats: as a collection of XML files and in a relational database. This relational database implementation is complete with a web-based interface and provides an easy way to navigate the ontology.

The data in COMe are gathered from the literature. Every COMe entry is manually edited and each ontological relationship is manually assigned; thus the pitfalls of automatically generated datasets are avoided (e.g. the centres containing non-native metals are not included). COMe does not aim to list all known bioinorganic proteins but rather to provide a controlled vocabulary and classification to allow better annotation of them in comprehensive databases. Representative examples (instances) of every protein family are included. Each instance has a cross-reference either to literature citation or to a publicly available database.

Data types

There are three types of entries in COMe: 'bioinorganic protein' (PRX), 'bioinorganic motif' (BIM), and 'molecule' (MOL). Here, 'bioinorganic protein' is any complex protein, such as a metal-binding protein, an organic molecule-binding protein, a protein containing post-translational modifications, or a combination of any of these classes. Likewise, the original definition of 'bioinorganic motif' [8] is extended to include organic prosthetic groups and modified amino acids. Bioinorganic motif is now defined as a common structural feature shared by functionally related, but not necessarily homologous, proteins, and consisting of either

(i) metal atom(s) and first coordination shell ligands, linked to polypeptide-derived groups by covalent or ionic bonds;

(ii) organic molecule, linked to polypeptide-derived group(s) by covalent bond(s);

(iii) covalently modified amino acid residue(s), or

(iv) combination of any of the above.

As mentioned before, the data in COMe are derived from the literature. Thus, identification of ligands and binding mode in a BIM is based on assessment of the authors and the curator, and *not* on distance thresholds, like in automatically generated collection (see discussion of Example 4.1 below).

'Molecule' can form a permanent part of a complex protein, either directly (if no covalent or ionic bond is defined between the amino acid residue and the molecule; e.g. non-covalently bound FAD is a permanent part of a flavoprotein) or otherwise as a part of a BIM (e.g. covalently bound phycobilin is a permanent part of a phycobiliprotein).

Data structure

Any COMe entry (PRX, BIM or MOL) minimally includes an *identifier* (*ID*) and a *term*. Each entry is related to at least one other entry (see ***Entry relationships***). In addition, external cross-references (Xref) to a number of other databases are provided (Table 1).

Since COMe contains classes as well as individual entities, we take care to provide the most suitable cross-reference for a given level. For example, a protein homology family will be cross-referenced to the corresponding InterPro family [13]; a protein subunit to a Swiss-Prot [11] or TrEMBL [12] sequence (SPTR); a functional multisubunit enzyme to an EC number [14]; an instance of a metalloprotein in a particular state to a PDB entry [3], etc.

Protein

The protein (PRX) entity refers to the functional protein as well as to separate protein domains and subunits. A typical low-level PRX entity is shown in Table 2. The PRX

Table 1: On-line databases cross-referenced in COMe

Database	URL used in COMe web interface	Reference
Chemical compounds		
COMPOUND	http://srs.ebi.ac.uk/	[9]
chemPDB	http://www.ebi.ac.uk/msd-srv/chempdb/	
RESID	http://srs.ebi.ac.uk/	[10]
NIST Chemistry WebBook	http://webbook.nist.gov/chemistry/	
Protein sequence		
SPTR	http://srs.ebi.ac.uk/	[11,12]
Protein structure		
PDB	http://oca.ebi.ac.uk/	[3]
Protein families		
InterPro	http://www.ebi.ac.uk/interpro/	[13]
MEROPS	http://www.merops.ac.uk/	
Protein function		
ENZYME	http://srs.ebi.ac.uk/	[14]
GO	http://www.ebi.ac.uk/ego/	[7]
Bibliography		
PubMed/MEDLINE	http://srs.ebi.ac.uk/	[15]

Table 2: Example of protein entry

	Term	Species	State	Centre	External reference	ID
Protein	monomeric periplasmic nitrate reductase					PRX001226
Instance	periplasmic nitrate reductase	*Desulfovibrio desulfuricans*	Mo6+, hydroxo	Moco	PDB:2NAP	

entity minimally consists of *ID* and *term*. It also may include *instance* as well as MOL, BIM and other PRX entities. The instance always has a *species* attribute. The role of instance is to provide the external evidence that the protein in question does exist in a particular organism. Currently, instance has no separate ID, but external Xref(s) should be provided (while Xref is not always available for the parent term). The other attributes of instance are *centre* (for proteins containing more than one bioinorganic/ small molecule centre) and *state* (e.g. "reduced", "CO-bound", etc.).

Molecule

MOL is an entity representing a 'small molecule' (as opposed to a macromolecule) or atom, which, in complex with one or more polypeptides, forms a functional protein. An example MOL entry is shown in Table 3. As a rule, there are no systematic names or chemical formulae. Instead, the terms in MOL entries are cross-referenced to chemical or bibliography databases (Table 1).

Bioinorganic motif

As mentioned before, a Bioinorganic Motif (BIM) can include both metallic and non-metallic centres. In COMe representation, every BIM consists of at least one centre and two or more ligands. The complete lists of centres, ligands and polyhedral symbols are given in the Additional Material.

Neither the existing coordination chemistry nomenclature [16] nor the IUPAC Recommendations on bioinorganic terms [17] provides a suitable guide for describing bioinorganic centres in proteins. We have developed a 1-D 'shorthand' representation for intrinsically 2-D structures such as BIMs. This qualitative representation is intuitively straightforward since it is based on similar descriptions employed by bioinorganic chemists in the literature [4]. A 1-D string has also the advantage of allowing quick text searches. In the rest of this section, we illustrate the use of this 'BIM language' with a number of examples (Table 4).

In a BIM for a *mononuclear* centre, the central atom (metal) is written first, followed by the endogenous ligands (amino acid residues), and then the exogenous ligands, e.g. water. If the ligating atom needs to be indicated to avoid ambiguity, the symbol for this is separated from the ligand symbol by a dot, e.g. NE.His stands for the N^{ε} atom

Table 3: Example of molecule entry

Term	External reference	ID
CoM-S-S-CoB		MOL000133
O-phosphono-N-{(2E)-7-[(2-sulfoethyl)dithio]hept-2-enoyl}-L-threonine	chemPDB:SHT	
coenzyme M 7-mercaptoheptanoylthreonine-phosphate heterodisulfide	COMPOUND:C04832	

Table 4: Examples of bioinorganic motifs

	Mononuclear metal centre	
4.1	[*TBPY*-5] Cu(ND.His)2(O.Gly)(SD.Met)(SG.Cys)	BIM000095
4.2	[*T*-4] Zn(ND.His)2{k2-(O,SG.Cys)}	BIM000176
	Dinuclear metal centre	
4.3	{Cu(ND.His)(SD.Met)}{Cu(ND.His)(O.Glu)}{μ-(SG.Cys)}2	BIM000104
4.4	{Fe(ND.His)(OE.Glu)(OH2)}{Fe(ND.His)(OE.Glu)}{μ-(OE.Glu)}{μ-(OE,OE.Glu)}{μ-(OH2)}	BIM000080
4.5	{Fe(ND.His)(OE.Glu)(OH2)}{Fe(ND.His)(OE.Glu)2}{μ-(OE,OE.Glu)}{μ-(OH2)}{μ-(OH)}	BIM000081
	Polynuclear metal centre	
4.6	{Fe4(μ3-S)4}(NE.His)(SG.Cys)3	BIM000059
	Metal – organic group complex	
4.7	[*SPY*-5] Co(Crn)(NE.His)	BIM000209
4.8	{(Fe4S4)(SG.Cys)3}{Fe(por)}{μ-(SG.Cys)}	BIM000026
4.9	[*TPR*-6] Mo{k2-(S,S.dtpp)}2(SeG.Sec)(OH)	BIM000181
	Organic prosthetic group centre	
4.10	(C6.FMN)(SG.Cys)	BIM000137

of a His residue. This is a simplistic description that does not take into account the stereochemistry at the metal atom. The polyhedral symbol is not mandatory for it may be unknown. It also does not make sense for polynuclear metal centres (see below).

Example 4.1 in Table 4 shows a mononuclear centre found in the blue copper protein azurin. In this centre, one copper atom is coordinated by the N^{δ} atoms of two His residues, one mainchain oxygen derived from Gly, one S^{δ} atom of a Met residue and one S^{γ} atom of a Cys residue. The coordination geometry is trigonal bipyramidal (*TBPY*-5; the polyhedral symbols used are as in Table S3 in Additional file 3).

It can be observed that the Met and mainchain O ligands from the azurin copper centre are in the BIM, even if they are positioned further away than some conventional cut-off distance (e.g. 3 Å). This is because it is known from the literature that a protein family is characterised by a metal atom surrounded by specific ligands forming a recurrent structural motif, and so the groups referred to as 'ligands' are included in the BIM.

In Example 4.2, the zinc atom is tetrahedrally coordinated [*T*-4] by the N^{δ} atoms of two His residues and by two atoms of a Cys residue, the mainchain oxygen and S^{γ} (the k2 prefix designates didentate binding).

In *dinuclear* metal centres such as Cu_A copper centre (Example 4.3), each metal atom and its unique ligands are enclosed in braces and followed by the bridging ligand(s) designated by a μ prefix.

Some dinuclear metal proteins, notably diiron–carboxylate proteins, undergo a change of metal coordination by the carboxylate ligands upon oxidation/reduction (the so-called carboxylate shift) [18]. Therefore, two or more BIMs can be assigned to the same protein, as in methane monooxygenase hydroxylase (Examples 4.4 and 4.5 in Table 4). Note that the protonation is explicitly stated, i.e. OH_2, OH^- and O are different ligands in BIMs. This information is taken from the literature and not PDB.

There are no central atoms in *polynuclear* metal centres. Therefore, a cluster (such as the cubane iron-sulphur unit) takes the place of the central atom (Example 4.6).

In a BIM for a centre containing a *metal–organic group complex*, first comes the metal, then the organic group, the amino acid residues, and finally the exogenous ligands (e.g. CO). Examples 4.7 and 4.8 show BIMs for metal–tetrapyrrole (haem, chlorophyll, cobalamin) proteins, where Crn = corrin, por = porphyrin. Tetrapyrrole compounds are assumed always to be tetradentate. For pyranopterin-containing centres such as molybdenum cofactor [19], BIMs look like the one in Example 4.9 (dtpp = ene-dithiol pyranopterin).

Finally, for a purely *organic prosthetic group* such as FMN covalently attached to the polypeptide (Example 4.10), the same approach is used (except that the metal is absent and the organic group now takes the first place).

Entry relationships

The relationships between entities are not made explicit in XML, but can be deduced using the set of rules. The relational implementation, however, provides explicit relationships. Several examples in Table 5 show fragments of COMe ontology.

IsA

The Example 5.1 in Table 5 illustrates the *IsA* (child to parent) relationship (also known as the *IsKindOf* relationship). 'Fe2S2 protein' is kind of 'iron-sulphur protein' which is kind of 'iron protein' which is kind of 'metalloprotein'. This relationship occurs between entities of the same class (PRX to PRX, BIM to BIM, MOL to MOL).

An important feature of this relationship is inheritance. For example, all proteins belonging to the 'Fe2S2 protein' class inherit the substructure (Fe2S2)(SG.Cys)2* (BIM000063).

An entity may have more than one parent. In the example (Figure 1), carbon monoxide oxidase inherits features from each subunit, viz. two different Fe_2S_2 clusters, molybdenum cofactor (Mo-pyranopterin complex) and FAD.

Examples 5.2. and 5.3 show how *IsA* relationships are used to create ontologies of BIM and MOL entities. Note the asterisk (wildcard) in BIM000025 and MOL000041!

From Example 5.2, one can get an impression that every observed exogenous ligand can give rise to a separate BIM. However, this is not the case. Although PDB contains numerous instances of enzyme-inhibitor complexes or substituted metalloproteins, these will not be included in COMe. Since the entries are manually annotated, only biologically relevant motifs (e.g. cited or confirmed as such in the literature) are included.

IsPartOf

This relationship can occur between entities of the same class (BIM to BIM, MOL to MOL) or different classes (MOL to BIM, BIM to PRX). Example 5.4 in Table 5 illustrates the *IsPartOf* relationship (BIM to BIM).

In this example, each of the two mononuclear centres is part of the dinuclear metal centre. For the mononuclear centres, it is possible to indicate their coordination geometry. Note the different representation of mono- or didentate (bidentate) coordination modes: monodentate in Zn(OE.Glu)4 and didentate in Fe{k2-(OE,OE.Glu)}. On the other hand, BIM000352 is not just a sum of BIM000353 and BIM000354. Nothing in BIM000353 or BIM000354 indicates which ligands bridge the two metal atoms.

IsPartOf and *IsA* can be used alternatively. Multisubunit proteins illustrate the difference between the two approaches. The 'Cellular Component' part of Gene Ontology [7] contains macromolecular complexes, with the relationship 'subunit A *IsPartOf* complex C'. In COMe, multisubunit proteins follow the pattern 'complex C *IsA* subunit A'. The reasoning is that a complex inherits all the properties of its constituents, as in our example (Figure 1). To the bioinorganic chemist, it is made clear that carbon monoxide oxidase is a molybdenum iron–sulphur flavoprotein. In this respect, protein subunits are completely analogous to protein domains.

IsBoundTo

This special relationship occurs in the case MOL to PRX. It is used because the molecule which *IsBoundTo* a protein can be changed chemically and, strictly speaking, become a different entity. For instance (Example 5.5 in Table 5), one can say that the protein binds pyridoxal 5'-phosphate (MOL000108), but the resulting substructure has no aldehyde group and therefore is different from 'free' pyridoxal 5'-phosphate, and so MOL000108 is *not* part of PRX000808. However, BIM000270 *IsPartOf* PRX000808.

Table 5: Examples of entry relationships in COMe

	Fragment of ontology of PRX entities
5.1	complex protein PRX000001 ↲ *IsA* metalloprotein PRX000002 ↲ *IsA* iron protein PRX000004 ↲ *IsA* iron-sulphur protein PRX000007 ↲ *IsA* Fe2S2 protein PRX000053 ↲ *IsA* Fe2S2 ferredoxin PRX000058 ↲ *IsA* plant and mammalian-type ferredoxin PRX000963 ↲ *IsA* plant-type ferredoxin domain PRX000062 ↲ *IsA* plant-type ferredoxin PRX000063
	Fragment of ontology of BIM entities
5.2	Fe(por)(NE.His)* BIM000025 ↲ *IsA* Fe(por)(NE.His) BIM000007 ↲ *IsA* Fe(por)(NE.His)(NO) BIM000018 ↲ *IsA* Fe(por)(NE.His)(OH) BIM000019 ↲ *IsA* Fe(por)(NE.His)(O2) BIM000021 ↲ *IsA* Fe(por)(NE.His)(CO2) BIM000022 ↲ *IsA* Fe(por)(NE.His)(CO) BIM000023 ↲ *IsA* Fe(por)(NE.His)(O2)-BIM000024 ↲ *IsA* Fe(por)(NE.His)(CN) BIM000198
	Fragment of ontology of MOL entities
5.3	FAD* MOL000041 ↲ *IsA* FAD MOL000039 ↲ *IsA* FADH2 MOL000040 ↲ *IsA* FADH. MOL000042
	IsPartOf *relationship*
5.4	Zn(OE.Glu)2 [Fe(ND.His) k2-(OE,OE.Glu)] μ-(OE,OE.Glu)2 BIM000352 ↲ *IsPartOf* [*OC*-6] Fe(ND.His)(OE.Glu)2 k2-(OE,OE.Glu) BIM000353 ↲ *IsPartOf* [T-4] Zn(OE.Glu)4 BIM000354
	IsBoundTo *and* IsPartOf *relationships*
5.5	pyridoxal 5'-phosphate protein PRX000808 ↲ *IsBoundTo* pyridoxal 5'-phosphate MOL000108 ↲ *IsPartOf* (C4A.Pxy)(NZ.Lys) BIM000270
	IsPartOf *and* IsA *relationships*
5.6	bacterioferritin PRX000159 ↲ *IsPartOf* Fe(por)(SD.Met)2 BIM000015 ↲ *IsA* haem b bacterioferritin PRX001170

Table 5: Examples of entry relationships in COMe *(Continued)*

	↲ *IsA* Fe-coproporphyrin III bacterioferritin PRX001171
	IsBoundTo *and* IsA *relationships*
5.7	cytochrome b PRX000153 ↲ *IsBoundTo* haem b MOL000013 ↲ *IsA* haem-bis-His cytochrome b PRX000674 ↲ *IsA* haem-His-Met cytochrome b PRX000675 ↲ *IsA* haem b bacterioferritin PRX001170

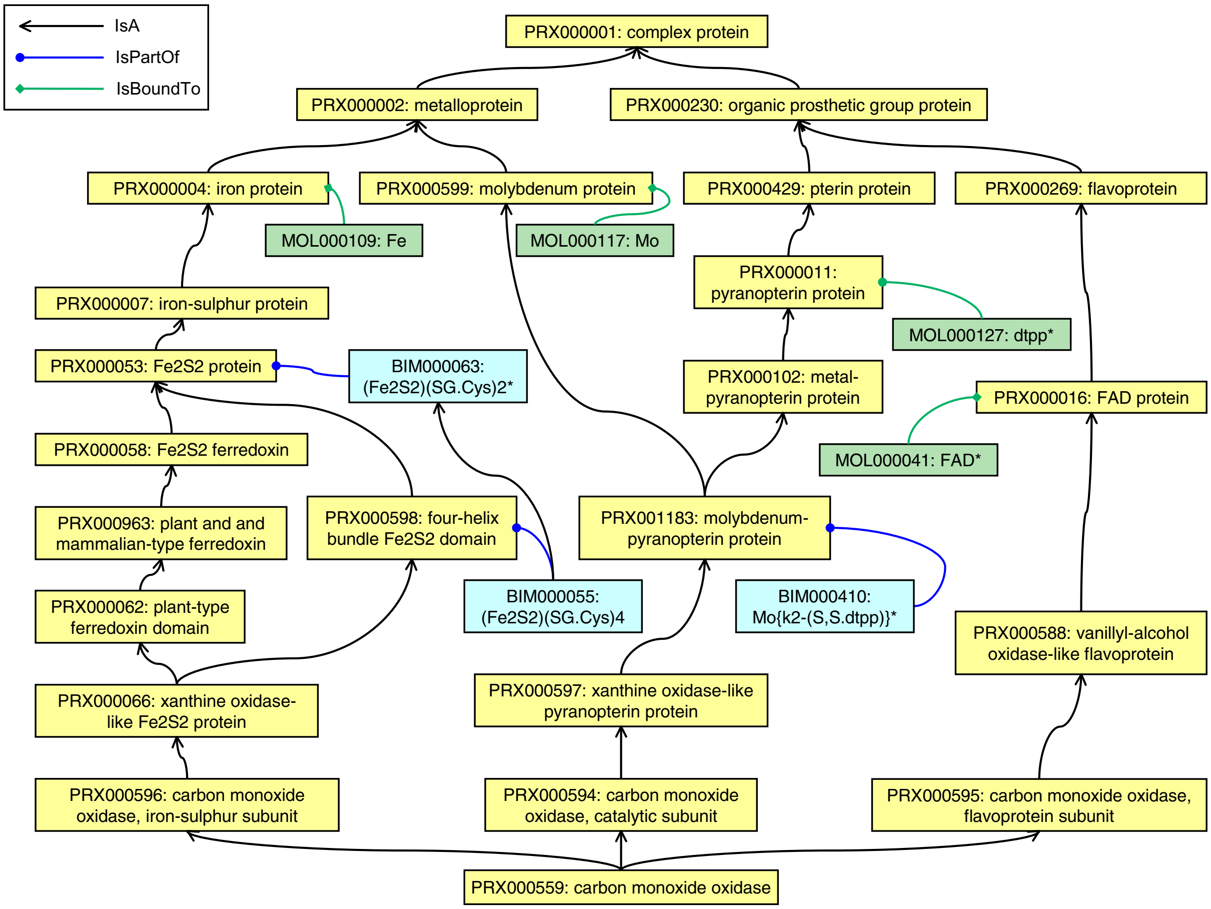
IsA
IsPartOf
IsBoundTo
PRX000001: complex protein
PRX000002: metalloprotein
PRX000230: organic prosthetic group protein
PRX000004: iron protein
PRX000599: molybdenum protein
PRX000429: pterin protein
PRX000269: flavoprotein
MOL000109: Fe
MOL000117: Mo
PRX000011: pyranopterin protein
PRX000007: iron-sulphur protein
MOL000127: dtpp*
PRX000053: Fe2S2 protein
BIM000063: (Fe2S2)(SG.Cys)2*
PRX000102: metal-pyranopterin protein
PRX000016: FAD protein
PRX000058: Fe2S2 ferredoxin
MOL000041: FAD*
PRX000963: plant and and mammalian-type ferredoxin
PRX000598: four-helix bundle Fe2S2 domain
PRX001183: molybdenum-pyranopterin protein
PRX000062: plant-type ferredoxin domain
BIM000055: (Fe2S2)(SG.Cys)4
BIM000410: Mo{k2-(S,S.dtpp)}*
PRX000588: vanillyl-alcohol oxidase-like flavoprotein
PRX000066: xanthine oxidase-like Fe2S2 protein
PRX000597: xanthine oxidase-like pyranopterin protein
PRX000596: carbon monoxide oxidase, iron-sulphur subunit
PRX000594: carbon monoxide oxidase, catalytic subunit
PRX000595: carbon monoxide oxidase, flavoprotein subunit
PRX000559: carbon monoxide oxidase

Figure 1
Fragment of ontology for carbon monoxide oxidase.

The same ligation pattern may apply to different prosthetic groups and vice versa. The combined use of both BIM and MOL entries to characterise such bioinorganic/small molecule centres and classify the bioinorganic proteins is illustrated in the following examples. In Example 5.6, bacterioferritins have different prosthetic groups but share the co-ordination mode of haem iron. Example 5.7 shows that the same prosthetic group can have different axial ligands.

It is important to stress that the data relationships in COMe are *not* automatically derived from any primary database. Each one is either a statement found in literature or a curator's judgement that, for example, the protein family PRX *y* is characterised by BIM *x*, with the curator assigning the logical relationship BIM *x* *IsPartOf* PRX *y*.

A summary of the relationships is presented in Table 6.

Table 6: Relationships in COMe

Relationship	Inheritance	Applies to	Inverse relationship
IsA	all attributes	PRX to PRX BIM to BIM MOL to MOL	*Includes*
IsPartOf	-	BIM to BIM MOL to MOL MOL to BIM BIM to PRX	*Contains*
IsBoundTo	-	MOL to PRX	*Binds*

Search tools

COMe has a web-based query interface that utilises Java Servlets technology [20]. Several basic queries are available: by COMe identifier, by text (both case insensitive and case sensitive), and by external reference identifier. The textual searches are also utilised to build a set of predefined queries of general interest (by residue, by a restricted vocabulary of keywords, by chemical element, vitamins and enzyme cofactor). In addition, it is possible to query the database for all the possible paths through a particular entry.

Graphical representations

The 1-D 'shorthand' representation of a BIM is not always unambiguous. Work is in progress on providing every BIM and MOL with a 2-D diagram.

An active graphical map of all paths through every entry of the ontology is also available.

Conclusions

There is little interaction between genomics and bioinformatics, on the one hand, and bioinorganic chemistry, on the other. Bioinorganic protein chemistry deals with at least three types of objects: metalloproteins and other complex proteins; naturally occurring 'small molecules' which can have different functional roles (e.g. prosthetic group, substrate, inhibitor); and bioinorganic models, which are artificial mimics of protein active sites. Neither 'small molecules' nor bioinorganic models occupy a central place in bioinformatics, while in the absence of experimental evidence the features of complex proteins are assigned on the basis of sequence similarity. The situation is further exacerbated by the absence of a definitive terminology shared by scientists in these fields.

COMe (Co-Ordination of Metals) provides a manually edited ontology for bioinorganic proteins and their features. The main groups of proteins in COMe are (i) metalloproteins, (ii) organic prosthetic group proteins and (iii) modified amino acid proteins. The classification of proteins according to these features is orthogonal to other classification schemes, such as those based on sequence similarity [13], 3-D fold [21], enzyme activity [14], or biological process [7]. The organisation of the controlled vocabulary allows both for annotation and querying at different levels of granularity. The controlled vocabulary can be used for structural and functional annotation of proteins, e.g. in sequence databases. The data are currently stored in both XML format and a relational database and are available at http://www.ebi.ac.uk/come/.

An intuitive nomenclature for 1-D representation of a 2-D bioinorganic motif (BIM) has been developed. This 'shorthand' representation of a BIM is not always unambiguous (for example, no stereochemistry data at the metal centre is included), but it is useful for quick searches. In future, the nomenclature could be extended, e.g. an explicit definition of every ligand at every position of a coordination polyhedron can be given while the 'shorthand' description of the BIM could be generated 'on the fly'.

Methods

The main source of data in COMe is the literature. Every COMe entry is manually edited as a separate XML file and ontological relationships are assigned via references to other XML files. The relational implementation of COMe is built on the original XML version. The conversion utilises a SAX parser [22] and some loading scripts. The parser also builds a table of relationships between pairs of COMe entries. COMe has the typical ontological structure of a Directed Acyclic Graph (DAG) [7]. In practice this means that each node of the ontology (apart from the root) has one or more parents. This structure is represented in COMe with a table containing all the possible paths (from the root, the 'complex protein' entity, to all the leaves) in the ontology. This table is filled by a program that reads the table of relationships between the pairs of entities generated by the parser.

First, all the leaves (nodes without children) are selected, then the DAG is explored starting from each leaf and ascending to the root. If there is branching in the graph, the partial 'common' path is stored and the program will in turn explore all the branches, and so on. This structure allows a quick retrieval of the path information. It is also used to create an active graphical map for each path in the DAG allowing easy navigation through COMe. The maps are built with the GraphViz package [23].

List of abbreviations

1-D, one-dimensional

2-D, two-dimensional

3-D, three-dimensional

BIM, bioinorganic motif

DAG, Directed Acyclic Graph

MDB, Metalloprotein Database and Browser

NMR, nuclear magnetic resonance

PDB, Protein Data Bank

SPTR, Swiss-Prot/TrEMBL database

XML, Extensible Markup Language

Xref, cross-reference

Authors' contributions

KD: concept, research and data curation. SC: all software and database implementation. Both authors read and approved the final manuscript.

Glossary

Bridging ligand, atom or chemical group linking two or more different metal atoms in polynuclear centres; indicated by the symbol μ.

Corrin, a macrocycle containing four pyrrole rings. It differs from porphyrin in that one of the single carbon bridges is replaced by a direct C–C bond. Naturally occurring complexes of corrin derivatives (corrinoids) with cobalt include cobalamin (vitamin B_{12}).

Coordination geometry, arrangement of the ligands around the central atom.

Coordination shell (first coordination shell), the collective name for the ligands surrounding the central atom(s).

Didentate (bidentate), containing two binding sites for a single metal atom.

Diiron-carboxylate proteins, a group of proteins characterised by a dinuclear iron centre bridged by carboxylate group(s) of Asp or Glu and oxide/hydroxide group(s).

Dinuclear (binuclear), containing two or more metal atoms within a single coordination shell.

Directed Acyclic Graph (DAG), a graph with one-way edges where no path starts and ends at the same node.

Endogenous, polypeptide-derived.

Enzyme, a protein catalyst.

Exogenous, not derived from a polypeptide.

Haem, an iron-porphyrin complex.

Homology, common evolutionary ancestry.

IsA (*IsKindOf*), semantic relationship of subsumption. If term **A** *IsA* term **B**, then **A** has a more specific meaning. The inverse relationship is *Includes*.

IsBoundTo, relationship between 'small molecule' **A** and macromolecule **B** in functional complex. Since both molecules may change chemically upon complex formation, *IsBoundTo* is not identical with *IsPartOf*. The inverse relationship is *Binds*.

IsPartOf, part/whole semantic relationship. The inverse relationship is *Contains*.

Ligand (in coordination chemistry), one of the atoms or chemical groups bound to the metal atom, usually by the donation of a lone-pair of electrons.

Molybdenum cofactor (Moco), the metal (Mo or W) complex of pyranopterin. Moco functions as the prosthetic group of a number of oxidoreductases.

Monodentate, containing a single binding site for a metal atom.

Mononuclear, containing one metal atom within a coordination shell.

Polydentate, containing two or more binding sites for a single metal atom.

Polynuclear, containing two or more metal atoms within a single coordination shell.

Porphyrin, a macrocycle containing four pyrrole rings each linked by single carbon atom bridges. Naturally occurring porphyrins form tight complexes with metal ions, such as Fe (haems), Mg (chlorophylls) and Ni (F430).

Prosthetic group, a non-polypeptide compound that conveys specific biological function to a protein. Single metal ions, inorganic compounds, organic compounds and metal-organic complexes all may function as prosthetic groups.

Additional material

Additional File 1
Table S1. Ligands in bioinorganic motifs.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1472-6807-4-3-S1.pdf]

Additional File 2
Table S2. Inorganic and organic centres in bioinorganic motifs.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1472-6807-4-3-S2.pdf]

Additional File 3
Table S3. Coordination polyhedra.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1472-6807-4-3-S3.pdf]

Acknowledgements

We thank Gillian Adams, Marcus Ennis and John S. Garavelli for their helpful comments and suggestions on the manuscript.

References

1. Vitkup D, Melamud E, Moult J, Sander C: **Completeness in structural genomics.** *Nature Struct Biol* 2001, **8:**559-566.
2. **MDB: the Metalloprotein Database and Browser** [http://metallo.scripps.edu/]
3. **Protein Data Bank** [http://www.pdb.org/]
4. Holm RH, Kennepohl P, Solomon EI: **Structural and functional aspects of metal sites in biology.** *Chem Rev* 1996, **96:**2239-2314.
5. Müller P, Köpke S, Sheldrick GM: **Is the bond-valence method able to identify metal atoms in protein structures?** *Acta Crystallogr D* 2003, **59:**32-37.
6. Carugo O, Pongor S: **The evolution of structural databases.** *Trends Biotechnol* 2002, **20:**498-501.
7. **Gene Ontology Consortium** [http://www.geneontology.org/]
8. Degtyarenko KN: **Bioinorganic motifs: towards functional classification of metalloproteins.** *Bioinformatics* 2000, **16:**851-864.
9. **LIGAND** [http://www.genome.ad.jp/ligand/]
10. **RESID** [http://home.earthlink.net/~jsgaravelli/RESIDInfo.HTML]
11. **Swiss-Prot** [http://www.ebi.ac.uk/swissprot/]
12. **TrEMBL** [http://www.ebi.ac.uk/trembl/]
13. **InterPro** [http://www.ebi.ac.uk/interpro/]
14. **ENZYME** [http://www.expasy.org/enzyme/]
15. **PubMed** [http://www.ncbi.nlm.nih.gov/PubMed/]
16. Leigh GJ, Ed: **Nomenclature of Inorganic Chemistry, Recommendations 1990.** *Oxford: Blackwell Scientific Publications*; 1990.
17. **Glossary of terms used in bioinorganic chemistry** [http://www.chem.qmw.ac.uk/iupac/bioinorg/]
18. Rosenzweig AC, Nordlund P, Takahara PM, Frederick CA, Lippard SJ: **Geometry of the soluble methane monooxygenase catalytic diiron center in two oxidation states.** *Chem Biol* 1995, **2:**409-418.
19. Fischer B, Enemark JH, Basu P: **A chemical approach to systematically designate the pyranopterin centers of molybdenum and tungsten enzymes and synthetic models.** *J Inorg Biochem* 1998, **72:**13-21.
20. **Java Servlet Technology** [http://java.sun.com/products/servlet/]
21. **SCOP** [http://scop.mrc-lmb.cam.ac.uk/scop/]
22. **SAX Project** [http://www.saxproject.org/]
23. **Graph Visualization Project** [http://www.graphviz.org/]