J Mol Model (2011) 17:429–442
DOI 10.1007/s00894-010-0733-5

ORIGINAL PAPER

Metals in proteins: cluster analysis studies

Juan A. C. Tamames · Maria João Ramos

Received: 11 November 2008 / Accepted: 23 April 2010 / Published online: 21 May 2010

Abstract We have conducted a prospective analysis of the Protein Data Bank in order to study certain constituents of proteins: elements that are neither halogens nor phosphorus nor part of the biological amino acid set. A sample of 5749 structures was analyzed and classified according to the 56 elements encountered. Fifteen metals (Na, Mg, K, Ca, Mn, Fe, Co, Ni, Cu, Zn, As, Mo, Cd, W, Hg) are involved in almost half of the structures, with each metal figuring in more than 100 structures. We analyzed this subsample in more detail by computing the amino acid residues occurring within a coordination sphere of 5 Å centered on the element, and using methods of cluster analysis to group the elements. The analyses undertaken here are able to distinguish between real components of proteins and elements inserted by artefacts of the crystallization process or experimental techniques.

Keywords PDB · Protein · Statistical analysis · Cluster analysis

Introduction

The Protein Data Bank [1] (PDB) is currently an essential reference source for anyone investigating or simply studying any matter relating to proteins. Right from its inception, this data bank has grown at a dramatic rate. The actual rate of novel structure deposition is about 20 files per day; therefore, while you are reading this text, half a structure will be inserted into the PDB.

As a result, any study concerning the contents of the PDB is inevitably a study of a snapshot of the data "this" day and hour, and is thus almost immediately dated. On the other hand, the large number of structures that are readily available in the PDB allows us to assume that it is a statistically valid sample of the proteins (or more exactly, our knowledge of proteins), and conclusions based on the statistical properties of the available data should be valid for a reasonable amount of time. We have analyzed 58737 files of the PDB (its complete content in June 2009).

Figure 1 shows the number of structures deposited yearly in total and by method. In fact, even though many experimental methods are used to determine structures nowadays, only three have yielded a significant number of the structures in the PDB: X-ray diffraction, with 50125 structures (oldest deposition is from 1972 [2]); solution nuclear magnetic resonance (NMR), with 7883 structures (first deposition in 1988); and electron microscopy, with 243 structures (first deposition in 1996). Although the total number of structures resolved with electron microscopy is still fairly small, the number of depositions associated with this method is increasing rapidly.

The increase in the number of structures deposited yearly has been nearly exponential, although it has slightly decelerated recently, suggesting the possibility that it is approaching a logistic sigmoid curve. Obviously, the availability of new technologies may change this trend at any moment. It is interesting to observe in Fig. 2 the relation between the structural resolution afforded by X-ray diffraction and the deposition date.

In Fig. 2a, it is clear that the structural resolution of X-ray diffraction has increased over time, but that the average structural resolution has remained at around 2.1 Å. In

J. A. C. Tamames · M. J. Ramos (✉)
Requimente, Departamento de Química,
Faculdade de Ciências, Universidade do Porto,
Rua do Campo Alegre, 687,
4169-007 Porto, Portugal
e-mail: mjramos@fc.up.pt

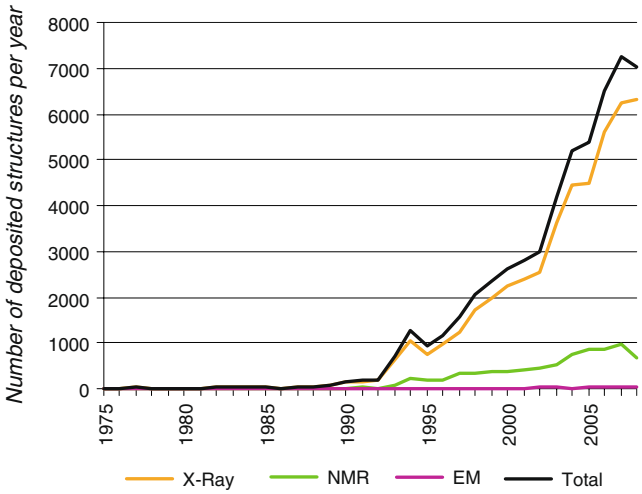


Fig. 1 Temporal evolution of structure deposition in the PDB with respect to method of structural determination (data for the three most commonly used methods are shown, as well as the total number of structures deposited per year)

Fig. 2b, we can see that the worst resolutions of deposited structures are so large that they have to be depicted on a different scale to that used in Fig. 2a. However, it is interesting to note that, among all of the structures in the PDB, only 500 (~0.9%) are depicted with resolutions of ≥3.5 Å.

A similar pattern can be found if we examine the sizes of the deposited X-ray diffraction structures. In the last few years, very large structures have increasingly been placed in the PDB (the largest recorded structure has a mass of 2153315.72 D: structure 1ml5 [3], deposited in 2003), and the same can be said of relatively small structures (around 500 D or less). These trends can be observed in Fig. 3.

Note that 1% of the biological species represented in the PDB comprise ~60% of its structures, with *Homo sapiens* and *Escherichia coli* being the two champions in this regard, since they are associated with 32% of the structures. *H. sapiens* is interesting for obvious reasons, and *E. coli* because of its easy manipulation and culture standardization. About 3900 species and variants are represented in the PDB, and all of the principal taxonomic phyla are associated with some structures.

This work developed from other studies on coordination distances in metalloproteins [4]. However, the number of different elements present in the proteins was much larger than our initial expectations, and so an investigation of the global aspects of their presence in protein structure appeared interesting.

There are many publications that explore the large amount of information that is contained in the PDB: one field of special interest is the study of metal ions in proteins. These may focus on either the crystallographic aspects, emphasizing the geometric dispositions of ligands, or on the statistical aspects, such as residues and distance variations.

As early as 1973, Kretsinger and Nockolds [5] described a motif consisting of two nearly symmetric pairs of helix segments in the active centers of proteins that bind Ca^{2+}; they called this motif a "calcium hand."

Kirberger et al. [6] carried out a statistical analysis of calcium-binding proteins, identifying the main characteristics of the calcium-binding sites as well as different coordination numbers and coordination distances.

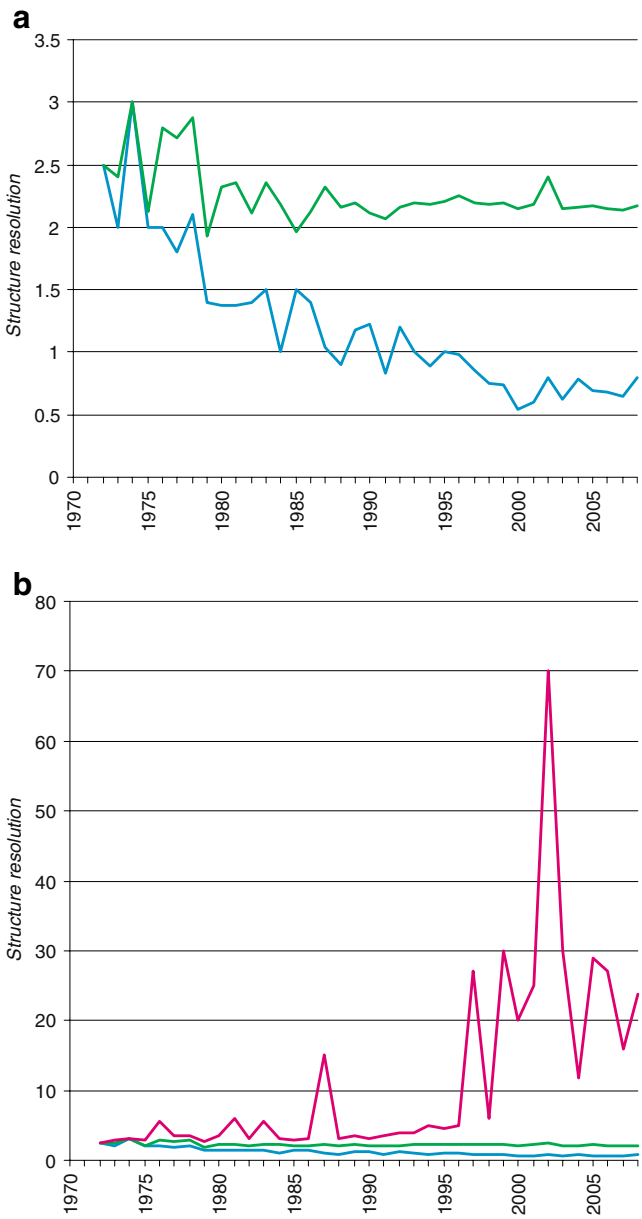


Fig. 2 **a** Temporal evolution of the structural resolution afforded by X-ray diffraction (*blue*), and the average structural resolution by date (*green*). **b** Temporal evolution of the best (*blue*) and worst (*red*) structural resolution afforded by X-ray diffraction, along with the average structural resolution by date (*green*)

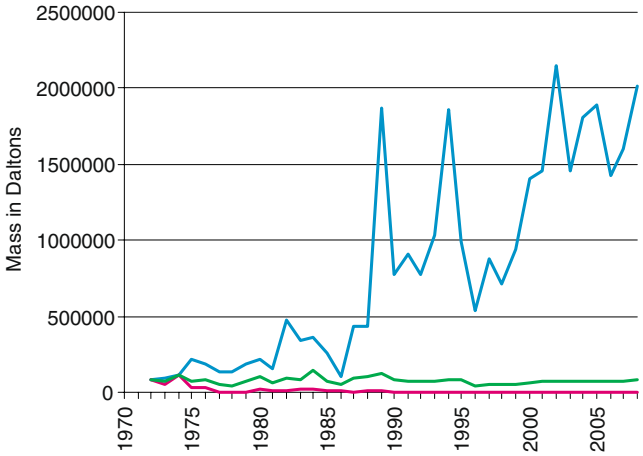


Fig. 3 Temporal evolution of the largest (*blue*) and smallest (*red*) molecular masses of structures determined by X-ray diffraction and deposited in the PDB, as well as the average structural molecular mass by date (*green*)

Harding studied [7–11] a representative sample of PDB structures with respect to Ca, Mg, Mn, Fe, Cu, Zn, Na and K, paying special attention to the crystallographic aspects of different metals in each publication. These metals frequently occur as components of functional enzymes, and are therefore also very common in the PDB.

Dokmanic et al. [12] presented a study of the correlation between metal, coordination number, and the corresponding residues involved. They used the same set of metals as Harding, as well as Cd. One interesting contribution is the recently implemented MESPEUS database [13], which contains data on coordination number, geometry and distances in association with PDB file references. This database also shows spatial models of metal sites.

A thorough bibliographic search unearths many publications that study one or a limited number of metals in proteins in detail, focusing on the crystallographic and/or the functional aspects.

All these data were used in this work to build a protein-oriented periodic table and then to draw conclusions from its contents. We also used cluster analysis to identify concealed patterns and spatial structures, thus evaluating the reliability of this technique for the study of metalloproteins.

Methodology

Structure selection

All of the statistical analyses presented in the previous section were performed using the information present in the whole PDB.

We then selected files that refer only to proteins. This was achieved by using PDB queries relating to "molecule type," thus eliminating references to nucleic acids or hybrid molecules. As a second step, we located all records of "HETATM" from among the PDB files, selecting only the structures of proteins with ligands. We directed our study towards elements that are neither halogens nor phosphorus nor constituents of the biological amino acid set.

In the studies involving quantitative distance evaluations, we imposed some additional restraints by suppressing all files that involved mutant structures, which could confuse the statistics. Furthermore, we considered only those structures with sequence similarities of <90%, as well as those with the best image quality (measured as 1/resolution – R-value) and the most recent deposition dates. Finally, we eliminated all structures with X-ray diffraction resolution limits of >2 Å. All NMR-based structures were selected.

Therefore, we were left with a database of nonredundant proteins with the best structures available in the PDB.

Statistical analysis

Statistical analyses were performed using the software MS Access, MS Excel [14] and NTSYS-pc [15].

Distances between specific ligand atoms and specific atoms in the protein chain were calculated with our own software, written in C++. This program scans a PDB file, searching for atoms of a given element, and computes all distances to atoms of oxygen, nitrogen and sulfur up to a maximum radius introduced as a parameter. The program output—the symbol of the scanned element, its location (i.e., residue) in the file, all of the atoms close to it, PDB residue identification, as well as the code and distance of the chain—can easily be imported by MS Excel or MS Access.

Our analyses focused on the occurrence of the residue in the environment of the studied element, considering not only coordination distances but also an extended neighborhood, in order to characterize the long-range interaction area for the different elements studied.

Cluster analysis

To perform these analyses, we began with a data matrix corresponding to the atoms of all amino acids within a sphere of 5 Å centered on each element studied: thus, we ended up with a matrix with 15 columns corresponding to the elements Na, Mg, K, Ca, Mn, Fe, Co, Ni, Cu, Zn, As, Mo, Cd, W and Hg (operational taxonomic units or "OTUs" in cluster analysis jargon), and 20 lines corresponding to the actual amino acid (characteristics). Therefore, each value in the matrix is the occurrence of a given residue in the neighborhood of a given element. We standardized each line of the data matrix to a mean of zero and a standard deviation of one. Thus, the data were measured in standard deviation units, which made them comparable.

The similarity between any two elements can be measured using the Euclidean distance coefficient:

$$d(x_i, x_j) = \frac{1}{n}\sqrt{\sum_{k=1}^{n}(x_{i,k} - x_{j,k})^2}$$

xi and *xj* are points (elements), x_{ik} is the k^{th} coordinate of the i^{th} point, and n is the number of characteristics. Therefore, two items with identical coordinates coincide with a distance coefficient of zero. There is no theoretical upper limit on the distance between two items (although in standardized data the value is seldom greater than 2). The lower the distance coefficient, the more similar the two items are. Repeating this operation for each pair of items, we obtained a symmetrical distance matrix, to which we applied the sequential, agglomerative, hierarchical, non-overlapping (SAHN) method [16].

Within the SAHN method, the items to be classified were selected by minimum distance, and a specific algorithm was used to compute the distance between each associated pair of elements and the remaining elements in the matrix.

We performed the analyses using two algorithms: *single linkage* and *pair group with unweighted average* (UPGMA) [16].

Single linkage computes the distance between two groups as the minimum distance between items of both groups; the result emphasizes the chained clusters but has the effect of contracting the space around the clusters.

UPGMA computes the distance by averaging the distances between the two groups. The results are fairly space-conservative, and groups with small gaps among them can be detected.

This agglomerative process is shown in a tree-like graphic (a phenogram), in which the item pairs are linked by a fork-shaped line, the height of which is the distance level at which the association occurs. Each phase of association may be either between items, between groups from a former association, or between both.

As in all processes that summarize information, there is an associated error, which grows as the successive associations accumulate calculated distances between clusters. Therefore, the clusters associated at the first levels are more reliable than those associated at the last levels.

The quality of a phenogram can be evaluated by the cophenetic correlation [16] coefficient: we can compute a new distance matrix from the tree (ultrametric, so it is more constrained than the metric of the original multidimensional space) and compare it with the original distance matrix using the correlation coefficient to evaluate how similar they are. This correlation does not have the same meaning as the correlation used in statistics; rather, it can be viewed as a measure of resemblance [17, 18].

Results and discussion

The elements studied here may occur in the proteins in the PDB for a variety of reasons:

a) *The metal provides an effective functional link, such as in a prosthetic group or cofactor.* Typical examples of these elements are zinc and manganese [7, 19, 20].

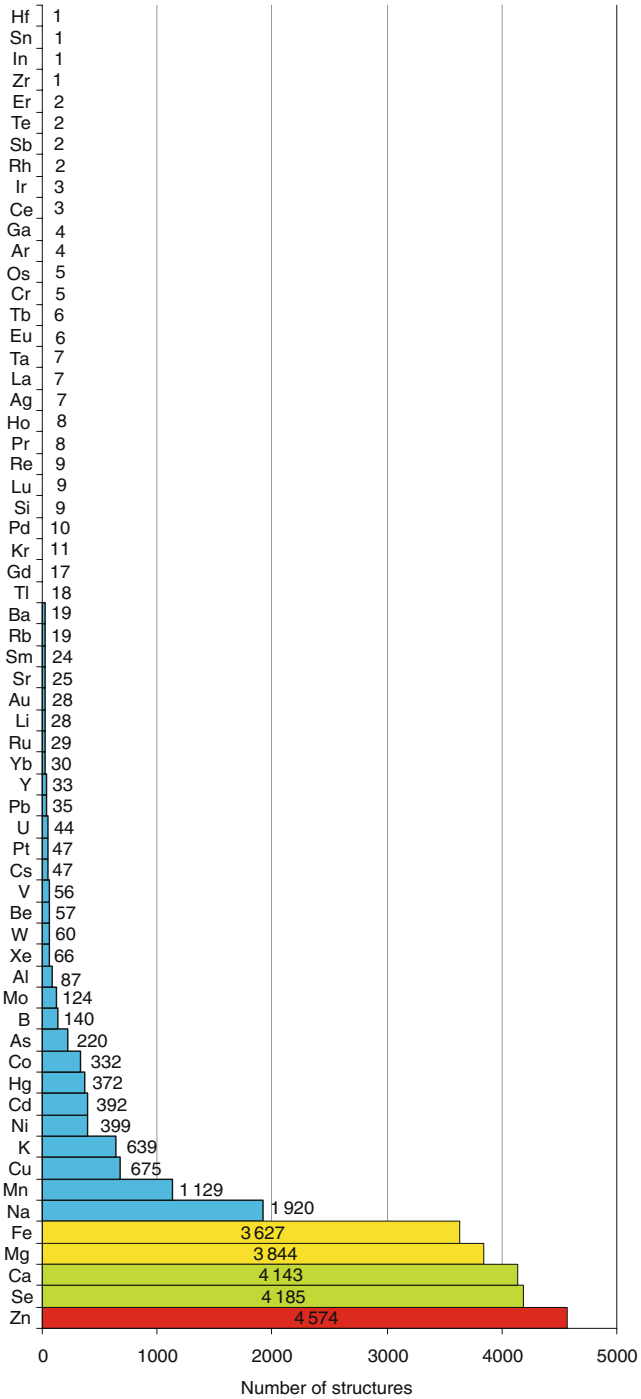


Fig. 4 Histogram showing the number of occurrences of the different elements considered in the selected structures from the PDB

b) *Metabolic processing of the element by a protein, as in detoxifying proteins*. Some occurrences of Cd, As and Hg in our sample fall within this group [21–23].

c) *An artefact caused by crystallization techniques*. Elements may be either intentionally or unintentionally inserted into the protein structure during crystallization or another preparative process. This may be the case for some calcium or potassium ions.

d) *The experimental insertion of an element in order to mark some specific structure or the active site of a protein with an atom that is more easily tracked by X-ray diffraction, or which exhibits better or specific behavior in crystallization procedures.* This is the case for some unusual elements, such as ytterbium [24, 25], which is a good reference replacement for other metals, or the xenon found in some structures, which is used to locate hydrophobic cavities by high-pressure injection [26, 27].

Our initial sample consisted of the protein-only files in the PDB that had been deposited before the end of 2008, and which had ligands containing elements that were neither halogens nor phosphorus nor part of the biological amino acid set (i.e., carbon, nitrogen, oxygen and sulfur). It totaled 47030 different files.

Figure 4 shows the number of occurrences of each element in ligands within the selected structures in the PDB. In order to classify the elements, we divided all of the structures studied into four sets (quartiles). The first 25% of the structures (the first quartile) are shown in blue, the second quartile is in yellow, the third quartile in green, and the fourth quartile in red. The median of the distribution is located on magnesium in the third quartile, which coincides with a significant increase in the number of structures in the chart for each element. Sodium marks the limit between the third and fourth quartiles, which is signaled by another visible increase in the number of structures in the chart for each element.

All of the elements typically associated with physiologic processes correspond to the last sixteen shown in Fig. 4. These include, somewhat unexpectedly, arsenic, mercury and cadmium, which will be considered later.

One special case is selenium, which exhibits a very large number of occurrences (only surpassed by zinc), but almost all of these relate to residues of selenomethionine, which are integrated into the protein chain; only 150 (around 0.25% of the occurrences of selenium) are present in other complexes. Therefore, we decided to include selenium in the set of amino acid component elements.

If the classification used in Fig. 4 is transferred to the periodic table of the elements, as shown in Table 1, we obtain an alternative understanding of the elements as they occur in proteins. In this table, we include the number of structures obtained for each element, and color the elements

Table 1 Periodic table presented from a protein point of view. Elements denoted in gray are not included in this study

H																	He
Li 28	Be 57											B 140	C	N	O	F	Ne
Na 1920	Mg 3844											Al 87	Si 9	P	S	Cl	Ar 4
K 639	Ca 4143	Sc	Ti	V 56	Cr 5	Mn 1129	Fe 3627	Co 332	Ni 399	Cu 675	Zn 4574	Ga 4	Ge	As 220	Se 4185	Br	Kr 11
Rb 19	Sr 25	Y 33	Zr 1	Nb	Mo 124	Tc	Ru 29	Rh 2	Pd 10	Ag 7	Cd 392	In 1	Sn 1	Sb 2	Te 2	I	Xe 66
Cs 47	Ba 19		Hf 1	Ta 7	W 60	Re 9	Os 5	Ir 3	Pt 47	Au 28	Hg 372	Tl 18	Pb 35	Bi	Po	At	Rn
Fr	Ra		Rf														
		La 7	Ce 3	Pr 8	Nd	Pm	Sm 24	Eu 6	Gd 17	Tb 6	Dy	Ho 8	Er 2	Tm	Yb 30	Lu 9	
		Ac	Th	Pa	U 44	Np	Pu										

1st Quartile | 2nd Quartile | 3rd Quartile | 4th Quartile | Absent

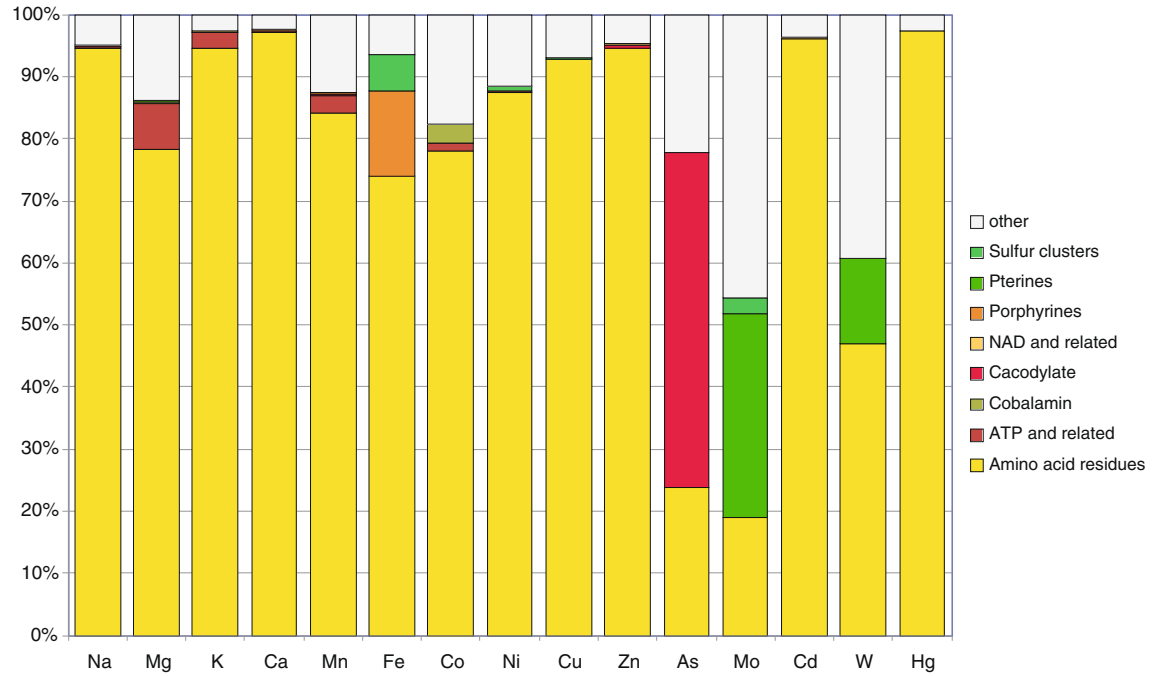


Fig. 5 Percentage of metal atoms bound to nonprotein structures in comparison to those coordinated to residues (*yellow*)

using the same code as in Fig. 4: blue for the first quartile, yellow for the second, green for the third, and red for the fourth quartile. Elements that are left uncolored are absent from all proteins in the PDB. Elements denoted in gray are not included in this study: these are members of the biological amino acid set, the halogens, or phosphorus.

Coordination distances

The immediate challenge is to draw some conclusions based on the presence of all these elements in the proteins.

To perform a more detailed statistical analysis, we selected a set of elements based on both statistical and biological criteria: elements that occur commonly enough to allow for meaningful statistical conclusions to be drawn, and which also have biological significance. Therefore, we chose to study the elements that figured in more than 100 structures, which reduced the set of studied elements to 15 "metals:" Na, Mg, K, Ca, Mn, Fe, Co, Ni, Cu, Zn, As, Mo, Cd, W and Hg.

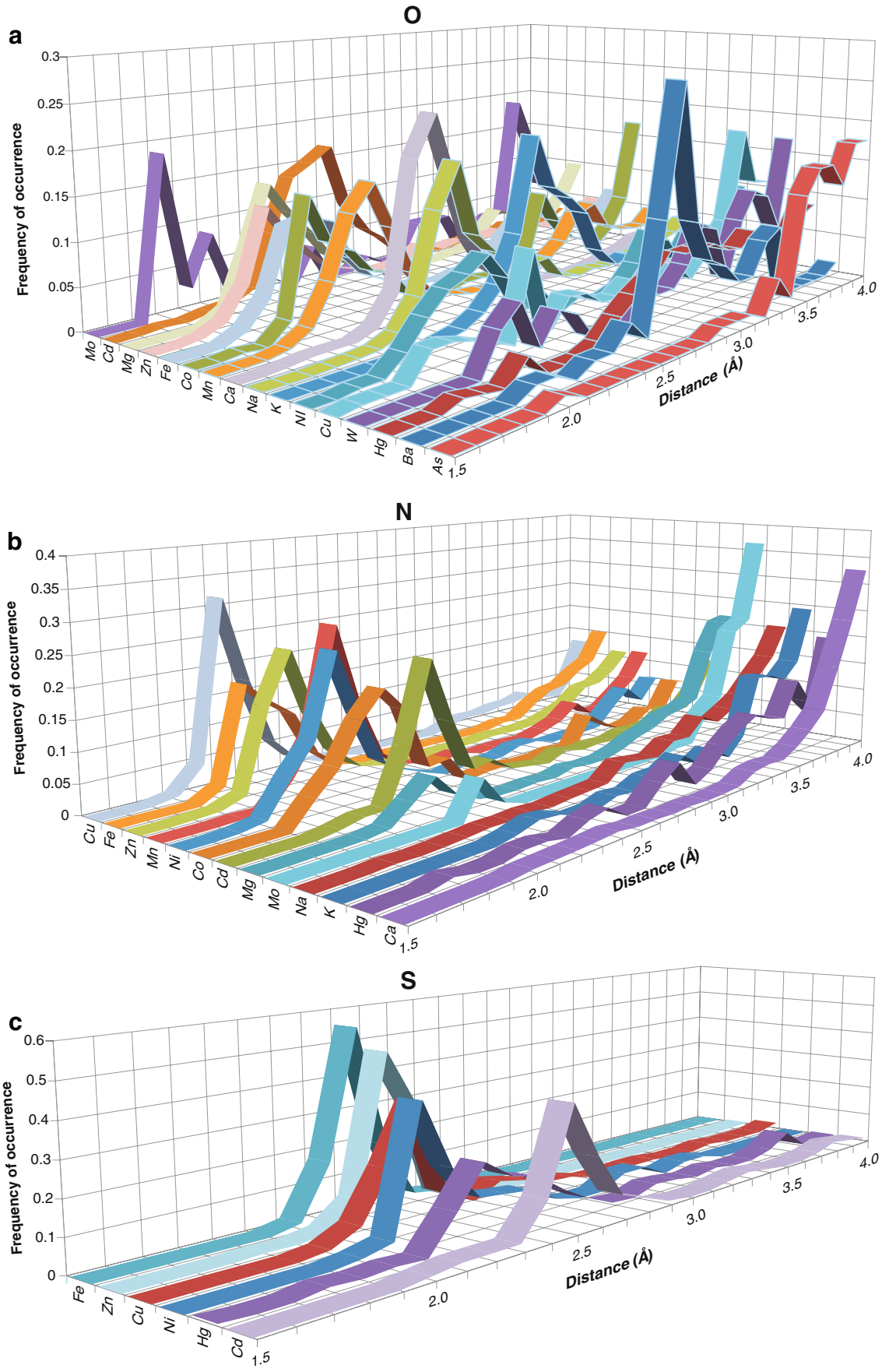
Another relevant question is the way in which the element is associated with the protein. Some elements, such as Zn, are usually associated directly with some amino acid residues in the structure, while others, such as Fe, are often associated with complexes containing sulfur or porphyrin heme groups. Figure 5 shows, as percentages, how often these metals are bound to nonprotein structures in comparison to how often they are coordinated to residues (in yellow). Arsenic and molybdenum are the elements that interact the most with atoms pertaining to a cofactor aside from atoms from (amino acid) residues in protein chains.

We evaluated the coordination of these metals to proteins by calculating their specific distances to some reactive atoms of the protein or cofactor. Thus, we used the program PdbDist (written in-house) to calculate the distances between the studied metals and oxygen, nitrogen and sulfur atoms in the protein amino acid residues (or other molecules associated with the protein) located less than 9 Å from the metal. These distances were then grouped in intervals of 0.1 Å, and all occurrences within each interval were counted. We considered multiple interactions with the same residue or cofactor molecule (such as the heme's four nitrogen atoms) to be one hit, thus avoiding an artificial overload in the occurrence of some cofactors.

Selecting only distances between the metal and the amino acid residues of the protein, we then created plots of occurrence vs. distance (see Fig. 6), which allowed us to check if there is any preferential distance between the metal and the amino acid residue atoms. If the distance is small (i.e., close interaction), the metal probably plays a functional role within the protein; greater distances may indicate a "casual" position of the atom in an unoccupied cavity of the protein.

A basic pattern has been identified for most of the elements studied: a peak is normally detected between 1.5

Fig. 6 Frequency of occurrence vs. distance between various metals and O (**a**), N (**b**) and S (**c**) atoms from amino acid residues present in the proteins of the PDB structures. All distances are in Å ▶


a
O
Frequency of occurrence
0.3
0.25
0.2
0.15
0.1
0.05
0
Mo
Cd
Mg
Zn
Fe
Co
Mn
Ca
Na
K
Ni
Cu
W
Hg
Ba
As
1.5
2.0
2.5
3.0
3.5
4.0
Distance (Å)
b
N
Frequency of occurrence
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Cu
Fe
Zn
Mn
Ni
Co
Cd
Mg
Mo
Na
K
Hg
Ca
1.5
2.0
2.5
3.0
3.5
4.0
Distance (Å)
c
S
Frequency of occurrence
0.6
0.5
0.4
0.3
0.2
0.1
0
Fe
Zn
Cu
Ni
Hg
Cd
1.5
2.0
2.5
3.0
3.5
4.0
Distance (Å)

and 2.5 Å, and in some cases there is a second peak around 3 Å. Other peaks can be seen at greater distances as more unrelated atoms are included in the sphere considered.

It is tricky to compare the profiles shown by the different elements using the absolute occurrence frequencies. The relevant features here are not the peak heights but rather their positions. Therefore, we normalized all curves according to the maximum value observed in the range 0–4 Å, and then scaled the ordinate for each metal curve in order to obtain a set of curves with similar amplitudes.

Figure 6a–c show frequency of occurrence curves for the distances between the studied metals and nitrogen, oxygen and sulfur, respectively. In all figures, it is clear that the curves for cobalt, copper, iron, manganese, nickel and zinc are very similar in terms of their distances from N, O or S, respectively. Accordingly, these graphs show two main peaks, one around 2 Å and a second around 3.5 Å. Both peaks are extremely clear for O, with the second becoming less sharp and less well defined as we travel from N to S.

The first peak, in all graphs, refers to the most populated area and represents all of the bonds established between the element in question and the N, O or S atoms, respectively. However, this main peak can be an overlap of two peaks, one of which represents a monocoordinated ligand and most other ligands in which a single atom interacts with the element, while the other accounts for bicoordinated ligands that present longer bond lengths than normal and intermediate cases. Zinc is a well-researched case in which this situation occurs in its binding to oxygen, resulting in what is known as the carboxylate shift [4, 28].

The second peaks in the curves shown by cobalt, copper, iron, manganese, nickel and zinc in relation to their distances to N, O or S, respectively, occur in nonbonding regions, and this peak relates to all of the N, O or S atoms that exist in the vicinity of the given element without being bound to it. The graphs therefore show that there are always N atoms present in such situations (as in the cases of Arg or Lys, for example), as well as some O atoms (as in the cases of Glu or Asp), but that S atoms appear to be far less abundant in such situations.

If we consider the metals studied here, it is well known that small amounts of cobalt are essential to many living organisms. Even though cobalt is one of the elements that we have studied in detail here, it is less common in proteins than metals such as manganese, iron or zinc. Analyzing the results obtained in our search, we noticed that, in many cobalt proteins, the metal is frequently part of a cofactor (as in vitamin B_{12}), but that there are also many cases in which cobalt is directly linked to the protein structure and has a preference for residues such as histidine, glutamate and aspartate (as in methionine aminopeptidase [29], integrin [30, 31], and many other cases).

Copper is an essential element in all plants and animals, and this element is found in a great variety of enzymes, including the very important superoxide dismutase [19, 20, 32] and the blue copper proteins [32, 33], in which copper is directly linked to the protein structure.

Iron is an element that is required by almost all living organisms, and is often incorporated in heme prosthetic groups, so it is not directly bound to the protein structure as such. However, it also occurs very commonly in the central regions of metalloproteins (for example in many transferrines and forms of superoxide dismutase), often in sulfur complexes (such as Fe_3S_4 and Fe_4S_4), even though it is not found as commonly in such regions as within a heme.

The classes of enzymes that have manganese cofactors are broad, ranging from oxidoreductases to isomerases, ligases, lectins and integrins. The reverse transcriptases of many retroviruses also contain manganese. It is thus not surprising that manganese is a metal that occurs in all forms of life.

Zinc is one of the most abundant transition elements in living organisms; it is an essential component of a very large number of enzymes, including those deriving from each of the six classes established by the International Union of Biochemistry [34]. This metal is extremely important in biology, and has been the focus of recent detailed studies [4] that provide valuable guidelines for the study of biological Zn systems.

Calcium, potassium, magnesium and sodium each present a small peak near 2 or 2.5 Å, but most of their distance values are associated with the second set of graph peaks, near 4 or 4.5 Å. In fact, there are some cases in which all of those elements are directly coordinated to the protein structure, and, especially in the case of magnesium, the resulting metalloproteins can be extremely important; for example farnesyltransferase [28, 35, 36], geranylgeranyl transferase [37] and integrase [38, 39], to name but a few. These are the cases that account for the small peak near 2 or 2.5 Å. However, the distances associated with the second set of graph peaks, near 3 or 3.5 Å, are too large to have any chemical meaning, and may be either large molecule cofactor components or artefacts from crystallization techniques. The pattern shown by cadmium is somewhat different, since it presents peaks that are still within the coordination radius, but which occur at relatively long distances. One interesting case is a carbonic anhydrase of marine diatoms, which can switch between using zinc or cadmium at its active site depending on the natural availability of each metal at a given moment. This enzyme is more efficient when it uses zinc, but it is also completely functional with cadmium [40]. Some cadmium- and mercury-binding proteins perform detoxification [23, 41] functions. Some proteins with arsenic are also detoxifying proteins, but the presence of As is usually associated with the cacodylate ion (from dimethyl arsinic acid,

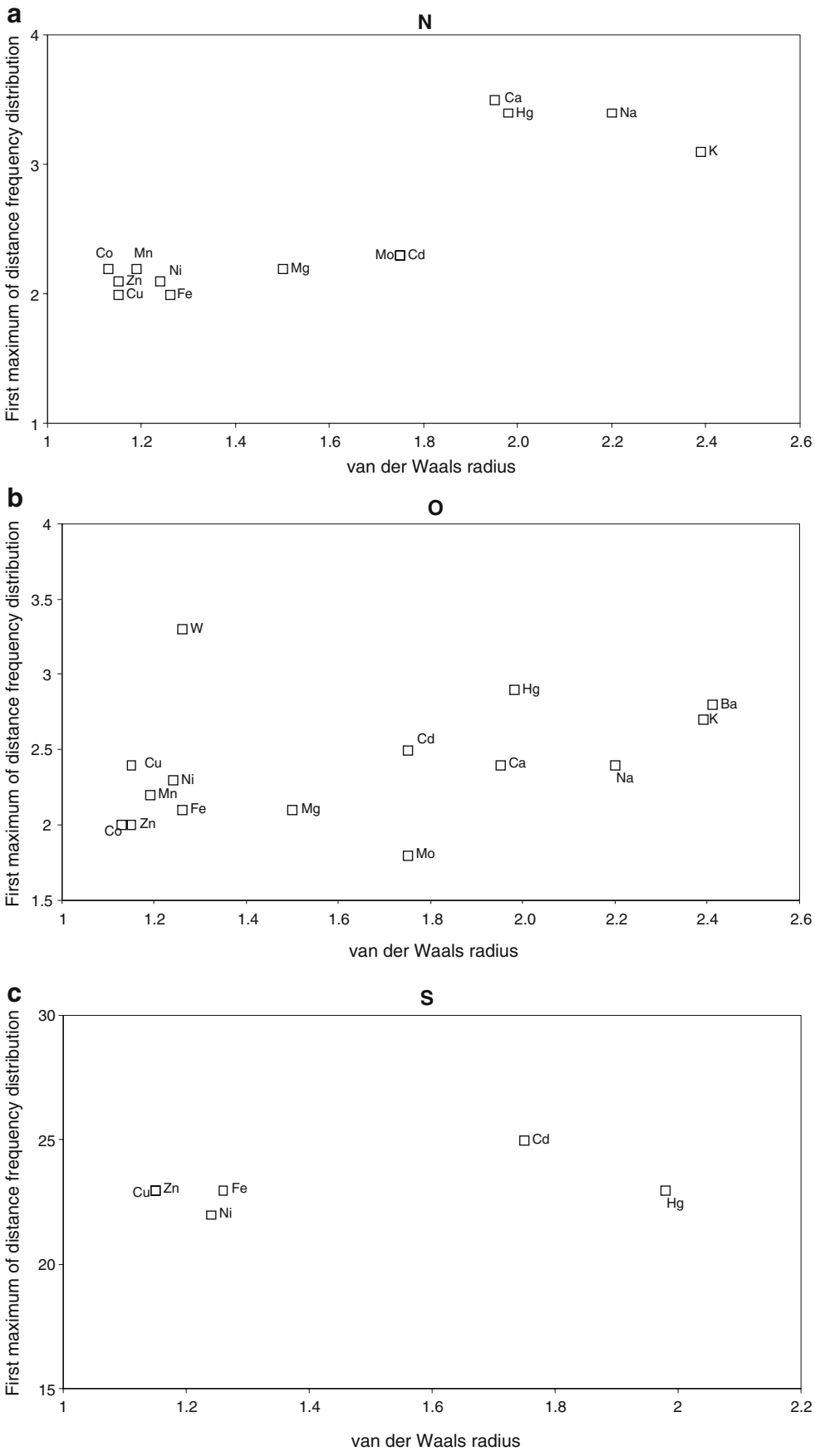


Fig. 7 Atomic radius (Å) vs. first maximum of the distance frequency distribution for distances to N, O and S atoms from protein residues

$(CH3)_2AsO_2H$), which probably derives from a buffer solution [42]. However, in many proteins with structures in the PDB, these elements occur as experimental or methodological insertions, often substituting for calcium or zinc ions. This is done either to facilitate cation identification [43] or to study ion substitution in order to elucidate the role of zinc in normal enzyme activity [44].

Coordination distances may be correlated with atomic radius; accordingly, plots of atomic radius versus the first maximum of the distance frequency for distances to N, O and S atoms from protein residues are shown in Fig. 7.

These graphs were created based on the values shown in Fig. 6, using the first peak in the distance frequency up to 3.5 Å only as the ordinate. Therefore, although these peak values point to a definite trend, they cannot be taken as absolutely correct values.

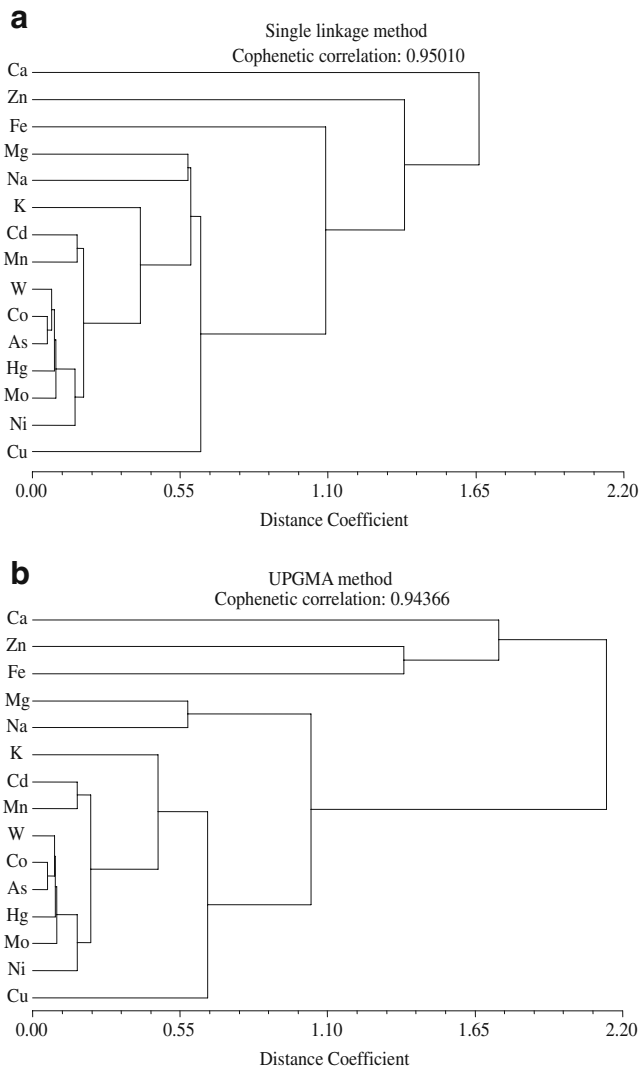


Fig. 8 Phenograms resulting from the number of residues encountered in the neighborhood of each metal, and obtained utilizing the **a** single linkage method, **b** UPGMA method

Table 2 Principal component loads for axes I, II and III

	I	II	III
ALA	0.9518	0.1767	0.1754
ARG	0.9020	−0.3189	−0.1179
ASN	0.7985	0.5720	0.0687
ASP	0.8011	0.5541	−0.0988
CYS	0.6447	−0.7386	−0.0781
GLN	0.9912	0.0789	−0.0818
GLU	0.9559	0.1554	−0.1405
GLY	0.9340	0.2563	0.1783
HIS	0.6371	−0.7517	−0.0484
ILE	0.9721	0.0045	−0.0744
LEU	0.9806	−0.0645	−0.0483
LYS	0.9015	−0.1385	−0.3923
MET	0.5204	−0.5024	0.6516
PHE	0.9132	−0.2406	−0.2618
PRO	0.8985	0.0565	0.4101
SER	0.9591	0.2106	0.0030
THR	0.8778	0.4031	−0.0438
TRP	0.9465	0.0451	0.2595
TYR	0.9711	−0.1182	0.0896
VAL	0.9532	−0.1698	−0.2083

A set consisting of Co, Cu, Fe, Mn, Ni and Zn, all largely known as cofactors or protein components, is apparent, as well as Cd (more unexpectedly). There is, however, a clear distinction between the two sets of elements (Fe, Ni, Co, Zn, Mn and Cd) and (Hg, Mg, Na, Ca and K), with a clear gap in peak distance between them.

Metals and residues

A second approach to this metal classification is to consider the relationship between the metals studied and the residues

Table 3 Principal component scores for axes I, II and III

	Ca	Zn	Fe	Mg	Na
I	1.8668	1.8145	0.9268	0.4125	0.3196
II	0.8665	−0.7268	−0.5451	0.3366	0.1455
III	0.1247	−0.4670	0.5449	−0.1868	0.0957
	K	Cu	Cd	Mn	W
I	−0.2067	−0.3720	−0.4875	−0.4859	−0.6464
II	0.0914	−0.2639	−0.0223	0.0526	0.0238
III	0.0321	0.3712	−0.0915	−0.0858	−0.0537
	Ni	Hg	Mo	Co	As
I	−0.5702	−0.6176	−0.6315	−0.6499	−0.6725
II	−0.0126	0.0102	0.0049	0.0209	0.0184
III	−0.0382	−0.0592	−0.0690	−0.0576	−0.0598

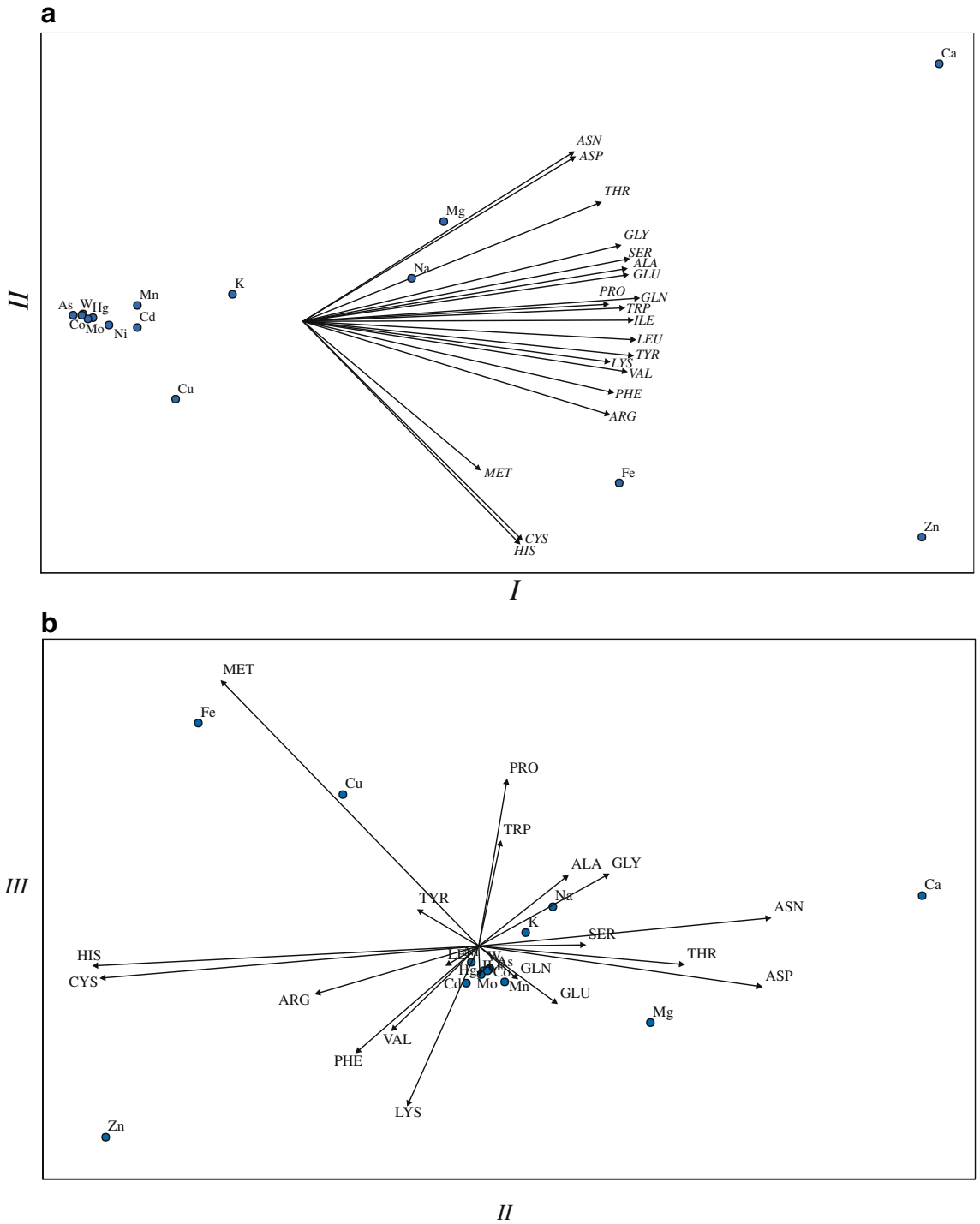


Fig. 9 Biplot graphics showing elements (*dots*) and frequency of occurrence of residues (*vectors*) plotted on the first three principal components (**a**: axes I and II, **b**: axes II and III). The behavior of a metal is determined by the vectors that are aligned with its position (but do not necessarily point to it)

in the proteins. For this analysis, we used all residues within 5 Å of the metal, a distance chosen taking into consideration the average dimension of a residue. Therefore, residues located further away than this are not assumed to directly affect the metal.

Note that this analysis is not directly related to that presented previously, because in that analysis we did not use distances but rather counted the number of residues that overlapped with a sphere centered on the studied metal.

The results obtained in this analysis were arranged in a matrix of elements vs. residues. Thus, the matrix values were simply the frequencies of occurrence of each residue within 5 Å of each studied element.

We applied cluster analysis techniques to these data in order to classify the metals according to the SAHN method described in the methodology section. We obtained values of 0.95 for the single linkage method and 0.94 for the UPGMA method, both of which may be considered very good fits. The high value obtained in the single linkage approach suggests that there are fairly distinct clusters.

The results obtained using both algorithms (single linkage and UPGMA) are shown in Fig. 8.

Analysis of the single linkage phenogram suggests the existence of a well-defined cluster consisting of Co, As, W and Mo, associated at a distance coefficient of 0.16. At a higher level (0.19), this cluster associates with Ni, Mn and Cd, and at a still higher level with Cu, K Na and Mg. On the other hand, Zn, Ca, and Fe are outliers that each show particular associations with residues. The topology of the UPGMA phenogram is very similar, differing only in the association order of Ca, Zn and Fe.

It is worth noting the close association of Co, As, W and Mo (0.201), which indicates that they show similar behaviour in relation to protein residues.

Ordination in reduced space

Another approach to clustering is to perform an ordination in reduced space. Although each element (column of the matrix) is a point in a space with 20 dimensions (one for each residue), we have tried to visualize this spatial structure in a reduced number of dimensions.

We utilized the principal component analysis method, using its geometrical properties more than its statistical ones. A correlation matrix between pairs of lines in the data matrix (i.e., vectors of the frequency of occurrence of each residue for each element) was computed. We extracted the eigenvectors from the correlation matrix; each of these represents an orthogonal direction of maximum variance in the original space. The precise amount of variance explained by a vector is the fraction of its eigenvalue with respect to the total. This set of vectors may be used to project the original points onto this new base. This results in a spatial rotation that facilitates a projection in a two- or three-dimensional plot, which is easier to interpret.

In our data, the first three eigenvectors represent 96.5% of the total variance. Therefore, we can draw the points on these three axes without losing much information regarding the overall spatial structure. The loads for the first three axes are shown in Table 2. The scores are the points on the new rotated axes, and these are shown in Table 3.

Figure 9a and b are biplots that simultaneously show vectors (residue frequencies within the 5 Å sphere) and metals projected onto a rotated axis. The position of each element is related to the absolute value of its score in the projected matrix. Therefore, the vectors that are aligned with (but do not necessarily point to) each point are the residues that characterize the behavior of the metal in some way when they are in its neighborhood.

In the projection of axes I and II, a "size effect" for all of the factors with positive values on axis I can be seen. Even more interestingly, the projection of axes II and III exhibits contrasting behavior between some amino acid residues.

The association of Zn with His and Cys is readily apparent, in perfect agreement with other findings reported in the literature [4], emphasizing the reliability of the methodology used here. Fe exhibits strong associations with Cys and His too, as well as with Met. The set consisting of Mn, Co, W, As, Ni, Hg and Cd is centrally placed, with no special preference for any residue, and these elements have negative values on the first axis in Fig. 9a. Cu shows some association with Met, but presents negative loads from His and Cys, which are strongly associated with Fe and Zn.

K and Mg exhibit eclectic behavior, locating themselves near the center, and are therefore not associated specifically with any residue. Ca appears to be positively aligned with Asn and Asp, and negatively with His and Cys, indicating that these residues are found in the neighborhood of this metal relatively infrequently.

As with the SAHN method, we can see that Mn, Co, W, As, Ni, Hg and Cd show similar behavior.

In order to validate this spatial configuration, we computed a new distance coefficient matrix using the coordinates of the points projected onto the first two principal axes, and compared this with the original distance matrix. We obtained a value of 0.999, indicative of an excellent representation (Fig. 10).

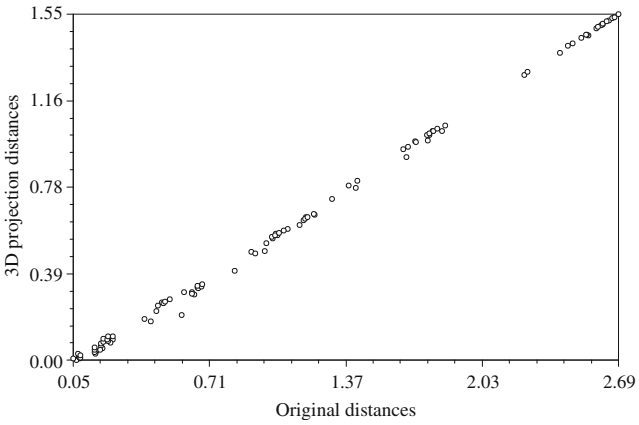


Fig. 10 Comparison of the distances between points (metals) in Fig. 9 (abscissa) and the distance coefficients from the original matrix (ordinate). Correlation is 0.999

Conclusions

The number of elements associated with proteins in the PDB is unexpectedly large. Nevertheless, the majority of the elements are involved in a small number of structures in the PDB. Only fifteen of them are present in a significant number of representations, and it is interesting to note that, among these elements, those that are most commonly found in the structures in the PDB are far more abundant than the others. These data were used to build a protein-oriented periodic table.

All of the analyses suggest that metals traditionally associated with biological enzymatic activities (Mn, Fe, Co, Mo, Ni, Cu and Zn) exhibit different behavior compared to other metals, which are inserted into the protein structure at a later stage, either intentionally or not. It was somewhat surprising to find 37 structures with noble gases, which were deliberately inserted under high pressure in order to study hydrophobic cavities in the protein structure [26, 27]. Less unusual is the presence of some metals that are complexed with the protein structure but are further away from it than usual. This is frequently the case for Na, K and Mg. These appear as eclectic elements with no definite preference for some residues over others.

Cd and Hg are interesting cases. In the analysis of the frequency of residue occurrence in the neighborhood of Cd, the metal seems to adopt a behavior similar to that of Mn, and similarly Hg seems to follow the behavior of Co and Ni. The most dramatic behavior is shown by Zn, which appears as an outlier due to its very strong association with Cys and His. These findings, in agreement with what is known from the literature, point to the reliability of the cluster analysis techniques. Principal component and other ordination methods can uncover some hidden relations between metals and specific residues.

Finally, we must remark that the PDB is not really a sample of protein structure, but rather a sample of our knowledge and interest in proteins. As an example, 2414 structures have the word “thermophilic” in their title or among their keywords, even though thermophilic organisms comprise a very small part of the biosphere in terms of biomass and taxonomic significance. Thus, the statistical distribution of biological species in the PDB is very different from that in nature.

Some tantalizing thoughts, which cannot currently be statistically corroborated, suggest the possibility of continuing and deepening this study in the future.

References

1. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28:235–242
2. Alden RA, Birktoft JJ, Kraut J, Robertus JD, Wright CS (1971) Atomic coordinates for subtilisin BPN' (or Novo). Biochem Biophys Res Commun 45(2):337–344
3. Klaholz BP, Pape T, Zavialov AV, Myasnikov AG, Orlova EV, Vestergaard B, Ehrenberg M, van Heel M (2003) Structure of the *Escherichia coli* ribosomal termination complex with release factor 2. Nature 421:90–94
4. Tamames B, Sousa SS, Tamames JAC, Fernandes PA, Ramos MJ (2007) Analysis of zinc ligand bond lengths in metalloproteins: trends and patterns. Proteins 69:466–475
5. Kretsinger RH, Nockolds CE (1973) Carp muscle calcium-binding protein. 2. Structure determination and general description. J Biol Chem 248(9):3313–3326
6. Kirberger M, Wang X, Deng H, Yang W, Chen G, Yang JJ (2008) Statistical analysis of structural characteristics of protein Ca^{2+}-binding sites. J Biol Inorg Chem 13(7):1169–1181
7. Harding MM (2004) The architecture of metal coordination groups in proteins. Acta Crystallogr D 60(Pt 5):849–859
8. Harding MM (2002) Metal–ligand geometry relevant to proteins and in proteins: sodium and potassium. Acta Crystallogr D 58(Pt 5):872–874
9. Harding MM (2001) Geometry of metal–ligand interactions in proteins. Acta Crystallogr D 57(Pt 3):401–411
10. Harding MM (1999) The geometry of metal–ligand interactions relevant to proteins. Acta Crystallogr D 55(Pt 8):1432–443
11. Harding MM (2006) Small revisions to predicted distances around metal sites in proteins. Acta Crystallogr D 62:678–682
12. Dokmanic I, Sikic M, Tomic S (2008) Metals in proteins: correlation between the metal-ion type, coordination number and the amino-acid residues involved in the coordination. Acta Crystallogr D 64(Pt 3):257–263
13. Hsin K, Sheng Y, Harding MM, Taylor P, Walkinshaw MD (2008) MESPEUS: a database of the geometry of metal sites in proteins. J Appl Cryst 41:963–968
14. Microsoft Corporation (2007) Access and Excel 2007. Microsoft Corporation, Redmond
15. Rohlf FJ (2004) NTSYSpc (numerical taxonomy system), v.2.2. Applied Biostatistics, Inc., Port Jefferson
16. Sneath PHA, Sokal RR (1973) Numerical taxonomy: the principles and practice of numerical classification. WH Freeman and Co., San Francisco
17. Rohlf FJ, Fisher DL (1968) Test for hierarchical structure in random data sets. Systematic Zool 17:407–412
18. Lapointe FJ, Legendre P (1992) Statistical significance of the matrix correlation-coefficient for comparing independent phylogenetic trees. Syst Biol 41:378–384
19. Branco RJF, Fernandes PA, Ramos MJ (2006) Cu, Zn superoxide dismutase: distorted active site binds substrate without significant energetic cost. Theor Chem Acc 115(1):27–31
20. Branco RJF, Fernandes PA, Ramos MJ (2006) Molecular dynamics simulations of the enzyme Cu, Zn superoxide dismutase. J Phys Chem B 110(33):16754–16762
21. Aposhian HV, Zakharyan RA, Avram MD, Sampayo-Reyes A, Wollenberg ML (2004) A review of the enzymology of arsenic metabolism and a new potential role of hydrogen peroxide in the detoxication of the trivalent arsenic species. Toxicol Appl Pharmacol 198:327–335
22. Murphy JN, Saltikov CW (2009) The ArsR repressor mediates arsenite-dependent regulation of arsenate respiration and detoxification operons of *Shewanella* sp. strain ANA-3. J Bacteriol 191:6722–6731
23. Steele RA, Opella SJ (1997) Structures of the reduced and mercury-bound forms of MerP, the periplasmic protein from the bacterial mercury detoxification system. Biochemistry 36(23):6885–6895
24. Burling FT, Weis WI, Flaherty KM, Brunger AT (1996) Direct observation of protein solvation and discrete disorder with

experimental crystallographic phases. Science 271(5245): 72–77
25. Tornaselli S, Zanzoni S, Ragona L, Gianolio E, Aime S, Assfalg M, Molinari H (2008) Solution structure of the supramolecular adduct between a liver cytosolic bile acid binding protein and a bile acid-based gadolinium(III)-chelate, a potential hepatospecific magnetic resonance imaging contrast agent. J Med Chem 51:6782–6792
26. Quillin ML, Breyer WA, Griswold IJ, Matthews BW (2000) Size versus polarizability in protein–ligand interactions: binding of noble gases within engineered cavities in phage T4 lysozyme. J Mol Biol 302:955–977
27. Olia AS, Casjens S, Cingolani G (2009) Structural plasticity of the phage P22 tail needle gp26 probed with xenon gas. Protein Sci 18 (3):537–548
28. Sousa SF, Fernandes PA, Ramos MJ (2007) The carboxylate shift in zinc enzymes: a computational study. J Am Chem Soc 129 (5):1378–1385
29. Lowther WT, Zhang Y, Sampson PB, Honek JF, Matthews BW (1999) Insights into the mechanism of *Escherichia coli* methionine aminopeptidase from the structural analysis of reaction products and phosphorus-based transition-state analogues. Biochemistry 38(45):14810–14809
30. Emsley J, Knight CG, Farndale RW, Barnes MJ, Liddington RC (2000) Structural basis of collagen recognition by integrin alpha2beta1. Cell 101(1):47–56
31. Smith C, Estavillo D, Emsley J, Bankston LA, Liddington RC, Cruz MA (2000) Mapping the collagen-binding site in the I domain of the glycoprotein Ia/IIa (integrin alpha(2)beta(1)). J Biol Chem 275(6):4205–4209
32. Branco RJF, Fernandes PA, Ramos MJ (2005) Density-functional calculations of the Cu, Zn superoxide dismutase redox potential: the influence of active site distortion. J Mol Struct 729(1–2):141–146
33. Paraskevopoulos K, Sundararajan M, Surendran R, Hough MA, Eady RR, Hillier IH, Hasnain SS (2006) Active site structures and the redox properties of blue copper proteins: atomic resolution structure of azurin II and electronic structure calculations of azurin, plastocyanin and stellacyanin. Dalton Trans 25:3067–3076
34. Vallee BL, Auld DS (1990) Active-site zinc ligands and activated H_2O of zinc enzymes. Proc Natl Acad Sci USA 87(1):220–224
35. Sousa SF, Fernandes PA, Ramos MJ (2005) Unraveling the mechanism of the farnesyltransferase enzyme. J Biol Inorg Chem 10(1):3–10
36. Sousa SF, Fernandes PA, Ramos MJ (2009) The search for the mechanism of the reaction catalyzed by farnesyltransferase. Chemistry 15(17):4243–4247
37. Taylor JS, Reid TS, Terry KL, Casey PJ, Beese LS (2003) Structure of mammalian protein geranylgeranyltransferase type-I. EMBO J 22(22):5963–5974
38. Delelis O, Carayon K, Saib A, Deprez E, Mouscadet JF (2008) Integrase and integration: biochemical activities of HIV-1 integrase. Retrovirology 5:114
39. Jaskolski M, Alexandratos JN, Bujacz G, Wlodawer A (2009) Piecing together the structure of retroviral integrase, an important target in AIDS therapy. FEBS J 276(11):2926–2946
40. Xu Y, Feng L, Jeffrey PD, Shi Y, Morel FM (2008) Structure and metal exchange in the cadmium carbonic anhydrase of marine diatoms. Nature 452(7183):56–61
41. Hennig HF (1986) Metal-binding proteins as metal pollution indicators. Environ Health Perspect 65:175–187
42. Maksimainen M, Timoharju T, Kallio JM, Hakulinen N, Turunen O, Rouvinen J (2009) Crystallization and preliminary diffraction analysis of a beta-galactosidase from *Trichoderma reesei*. Acta Crystallogr F 65:767–769
43. Hall DR, Kemp LE, Leonard GA, Marshall K, Berry A, Hunter WN (2003) The organization of divalent cations in the active site of cadmium *Escherichia coli* fructose-1,6-bisphosphate aldolase. Acta Crystallogr D 59(Pt 3):611–614
44. Zhang FL, Fu HW, Casey PJ, Bishop WR (1996) Substitution of cadmium for zinc in farnesyl:protein transferase alters its substrate specificity. Biochemistry 35(25):8166–8171