Editorial Manager(tm) for Journal of Molecular Modeling
Manuscript Draft

Manuscript Number: JMMO620R2

Title: The periodic table of the Elements from a protein point of view

Article Type: Original paper

Keywords: PDB; protein; periodic table; statistical analyses; cluster analysis

Corresponding Author: Prof. Maria J Ramos, Ph.D.

Corresponding Author's Institution: Faculty of Sciences, University of Porto

First Author: Juan Tamames, BSc

Order of Authors: Juan Tamames, BSc; Maria J Ramos, PhD

Abstract: We have conducted a prospective analysis of the Protein Data Bank in order to study certain constituents of proteins, i.e. those elements that are neither halogens or phosphorous nor part of the biological amino acid set. A sample of 5749 structures was analyzed and classified according to the 56 elements encountered. Almost half of the structures are represented by a set of twelve elements, each one involved in more than 100 structures. We analysed this subsample with more detail by computing the distance of amino acid residues in a coordination sphere of 5 Å, and studying the corresponding distribution curves of frequency of appearance, as well as using methods of cluster analysis to group the elements. We have therefore constructed a periodic table of the elements from a protein point of view. The analyses undertaken are able to distinguish between real components of proteins and elements inserted by artefacts of crystallization process or experimental techniques.

Response to Reviewers: Dear Editor,

Please find enclosed a reviewed version of the original paper submitted to JMM.
We have answered all the questions raised by the reviewers and very much improved the original version. We have also altered the original title as one of the reviewers remarked that it did not totally match what the study provided. The final version also presents a revised study of the PDB and the conlusions were modified accordingly.
In sum, this version is really much better than the original one.
Kind regards,

Maria Ramos

Manuscript
Click here to download Manuscript: JMM0620R2_final.doc

Metals in proteins – cluster analysis studies

Received: 11.11.2008 / **Accepted**: 23.04.2010

Juan A. C. Tamames and Maria João 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

Abstract

We have conducted a prospective analysis of the Protein Data Bank in order to study certain constituents of proteins, *i.e.* those elements that are neither halogens or phosphorous nor part of the biological amino acid set. A sample of 5749 structures was analyzed and classified according to the 56 elements encountered. In almost half of the structures 15 metals (*i.e.* Na, Mg, K, Ca, Mn, Fe, Co, Ni, Cu, Zn, As, Mo, Cd, W, Hg) are involved, each one figuring in more than 100 structures. We analysed this subsample in more detail by computing the distance of amino acid residues in a coordination sphere of 5 Å, and studying the corresponding occurrences, as well as using methods of cluster analysis to group the elements. The analyses undertaken are able to distinguish between real components of proteins and elements inserted by artefacts of crystallization process or experimental techniques.

Keywords PDB · Protein · Statistical analysis · Cluster analysis

Introduction

The Protein Data Bank[1] (PDB) is today an unavoidable reference for all people who investigate or simply study any matter related with proteins. From the beginning 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 was inserted in the PDB.

As a result, any study concerning the contents of the PDB is inevitably the study of a snapshot of the data "this" day and hour and, consequently, almost immediately dated. On the other hand, the large number of structures that are easily available in the PDB allows for the assumption that it is a statistically valid sample of the proteins (or more exactly, of 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 analysed 58737 files of the PDB, *i.e.* its full contents in June 2009.

Fig. 1 shows the number of structures yearly deposed both in total and by method. In fact, even though many experimental methods to determine structures are used nowadays, only three present a significant number of 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 using this method is increasing rapidly.

<Figure 1>

The increase in the number of structures deposited yearly has been nearly exponential, although with a slight deceleration presently, suggesting the possibility of 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 resolution of the structures determined by X-ray diffraction and the deposition date.

<Figure 2a and 2b>

In Fig. 2a, the evolution of the resolution presents an increasing yearly trend but the average resolution points to a stabilisation around 2.1 Å. In Fig. 2b, we can observe also the worse resolutions of deposited structures depicted in a different scale than those of Fig. 2a, because some values are so large that it is impossible to represent them within the same scale. It is however interesting to notice that, in the whole of the PDB, only 500 structures (near 0.9%) have a resolution with a value that is greater or equal to 3.5 Å.

A similar pattern can be found if we examine the size of deposited X-ray diffraction structures. In the last years the deposition in the PDB of very large structures (the recorded maximum is a mass of 2153315.72 D, structure 1ml5 [3] deposited in 2003) increased, and the same can be said to relatively small structures (around 500 D or less). This can be observed in Fig. 3.

<Figure 3>

Note that 1% of the biological species represented in the PDB accumulate around 60% of its structures, being *Homo sapiens* and *Escherichia coli* the two champions, cumulating 32% of the structures, Clearly, *H. Sapiens* is interesting for obvious reasons and *E. coli* because of its easy manipulation and culture standardisation. About 3900 species and variants are distinguished in PDB and all the principal taxonomic phyla are represented by some structures.

This work has begun as a development of 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 therefore it seems interesting a synthetic approach of the global aspects of their presence in protein structure.

There are many publications that explore the large amount of information that is contained in the PDB: a field with special interest is the study of metal ions in proteins. The focus may be either in the crystallographic aspect, with emphasis in the geometric disposition of ligands or in the statistical one, centring the study in residues and distance variations.

As early as 1973, Kretsinger and Nockolds [5] describe, in active centers of proteins binding the Ca^{2+} ion, a motif constituted by two nearly symmetric pairs of helix segments, which they called calcium hand.

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

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

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

A thorough bibliographic search provides many publications studying in detail one or a limited number of metals in proteins, with emphasis either in the crystallographic and/or in the functional aspects.

In this work all this data was used to build a protein-oriented periodic table and to withdraw conclusions from its contents. We further used cluster analysis in order to identify concealed patterns and spatial structures, thus evaluating the reliability of this technique in the study of metalloproteins.

Methodology

Structure selection

All the statistical analyses presented in the Introduction section of this work were performed using the information contained in the whole PDB.

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

In the studies involving quantitative distance evaluation we imposed some additional restraints by suppressing all those files that involved mutant structures, which could confuse the statistics. Furthermore, we considered only those structures with a sequence similarity inferior to 90% as well as best image quality (measured as 1/resolution – R-value) and more recent deposition date. Finally, we eliminated all structures with a resolution limit larger than 2 Å for the X-ray diffraction structures. All NMR-based structures have been selected.

Therefore, we were left with a database of non-redundant proteins, using 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 of the protein chain were calculated with our own software written in C++. This program scans a PDB file in search 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 residue location in the file, all atoms nearby, residue identification as well as chain code and distance - may be imported easily by MS-Excel or MSAccess.

Our analyses focussed on residue occurrence in the environment of the studied element, considering not only coordination distances but also an extended neighbourhood, aiming at characterizing a 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 centred 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 under study (operational taxonomic units or "OTU's" 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 neighbourhood of a given element. We standardized each line of a matrix of data to a mean of zero and a standard deviation of one. Thus the data was measured in standard deviation units, which made it comparable.

Similarity between any two elements may be measured by using the Euclidean distance coefficient:

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

x_i and x_j are points (elements), x_{ik} is the k^{th} coordinate of 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 for the distance between two items (although in standardized data the value is seldom greater than 2). The lower the distance coefficient the more similar two items are. Repeating this operation for each pair of items, we obtained a symmetrical distance matrix, to which we have applied the sequential, agglomerative, hierarchical, no overlapping (SAHN) method [16] .

Within the SAHN method, the items to be classified were selected by minimum distance, and associated using a specific algorithm in order 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 a distance between two groups as the minimum distance between items of both groups: the result emphasizes the chained clusters, but has an effect of contraction of the space around the clusters.

UPGMA computes a distance by averaging the distances between 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 (phenogram) in which the item pairs are linked by a fork-shaped line whose height is the distance-level in which the association occurs. Each phase of association may be either between items, groups from a former association or both.

As in all processes of information summarizing, there is an associated error, which grows as the successive associations cumulate calculated distances between clusters. So, the clusters associated at first levels are more reliable than the last.

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 more constrained than metric of the original multidimensional space) and compare it with the original distance matrix using the correlation coefficient to evaluate the similarity among them. This correlation has not the same meaning used in statistics, instead it must be viewed as a resemblance measure between matrices[17, 18].

Results and discussion

The presence of the elements studied, present in proteins of the PDB, may result from a variety of causes:

a) The effective functional link as prosthetic group or cofactor. Typical examples of those elements are zinc and manganese. [7, 19, 20]

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

c) An artefact caused by crystallization techniques. Elements may be either intentionally or unintentionally inserted in the protein structure along crystallization or other preparative process. This may be the case of some calcium or potassium ions.

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

Our initial sample consisted of protein-only files of the PDB, deposited until the end of 2008, with ligands containing elements that are neither halogens or phosphorous nor part of the biological amino acid set, *i.e.* carbon, nitrogen, oxygen and sulfur. It basically consisted in 47030 different files.

Fig. 4 shows the element distribution by occurrence in the selected structures. The size of the bars is related to the number of structures in the PDB with the corresponding element as a ligand or included in a ligand. In order to classify the elements, we divided all the structures studied into four sets (quartiles). The first 25% of structures (first quartile) are shown in blue, the second in yellow, the third quartile in green and the fourth one in red. The median of the distribution is located on the cobalt, third quartile, which coincides with a remarkable increase in the number of structures in the chart for one particular element. The limit between the third and fourth quartiles is on the sodium, which is signalled by another visible increase in the number of structures.

All elements typically associated with physiologic processes correspond to the last sixteen shown in Fig. 4. Amongst them, figure arsenic, mercury and cadmium, which are somewhat unexpected and will be considered later.

A special case is selenium presenting a very large number of occurrences, only surpassed by zinc, but almost all of it is in fact incorporated in residues of selenomethionine, integrated in

the protein chain, and only 150 (around 0.25% of occurrences) are present in other complexes. Therefore, we decided to include selenium in the amino acid component elements.

<Figure 4>

If the classification used in Fig. 4 is transferred to our well known Periodic Table, as shown in Table 1, we obtain an alternative understanding of the elements as they occur in proteins. In this table we represent the elements with the number of structures obtained for each one and colored with the same code as in Fig. 4, *i.e.* blue for the 1st quartile, yellow for the 2nd, green for the 3rd and red for 4th quartile. Elements with no colored background are absent in all proteins of the PDB. Elements with a grey symbol are those not included in the study: these are the biological amino acid set, halogens and phosphorus.

<Table 1>

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 in criteria both statistical and biological, *i.e.* elements occurring in a big enough number to allow for meaningful statistical conclusions to be withdrawn and simultaneously with biological significance. Therefore, we chose to study those elements that figured in more than 100 structures, which reduced the set of studied elements to 15 “metals”, *i.e.* 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 to the protein. Some elements, such as Zn, are usually associated directly to some amino acid residues of the structure but others, such as Fe, are often associated to complexes with sulfur or porphyrine heme groups. Fig. 5 shows, in percentage, how often those metals are bound to non-protein structures in comparison to residue coordination (yellow). Arsenic and molybdenum are the elements that interact the most with atoms pertaining to some cofactor, other than atoms from (aminoacid) residues in protein chains.

<Figure 5>

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 calculated with the home written program PdbDist the distances between the studied metals and those of oxygen, nitrogen and sulfur from the protein amino acid residues (or other molecules associated to the protein) located in a neighbourhood of 9 Å from the metal. This distance list was then grouped in intervals of 0.1 Å and all occurrences in each interval were counted. We considered multiple interactions with the same residue or cofactor molecule (such as the heme's four nitrogen atoms) as a unique hit, thus avoiding an artificial load in the occurrence of some cofactors.

By selecting only distances between metal and aminoacid residues of the protein, we have drawn occurrence *vs.* distance graphics (see Fig. 6), being able to check if there is any preferential distance between the metal and the amino acid residue atoms. If the distance falls within a sphere of 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 and 2.5 Å and in some cases there is a second peak around 3 Å. At greater distances other peaks can be seen as more unrelated atoms are included in the sphere considered.

<Figure 6A and 6B and 6C>

It is trickier to compare the profile shown by the different elements using the absolute frequency of occurrence values. The relevant features here are not the height of the peaks but rather their distance position. Therefore, we normalized all curves, selecting the maximum value in the range of 0 to 4 Å, and then scaled the ordinate for each one in order to obtain a set of similar amplitude curves.

Figs. 6.A to 6.C show the frequency of occurrence curves for the distances between the studied metals and nitrogen, oxygen and sulfur, respectively. In all figures, it is noticeable the almost coincidental pattern of the curves of cobalt, copper, iron, manganese, nickel and zinc in their distances to N, O or S, respectively. Accordingly, these graphs show mainly two

peaks each, one around 2 Å and a second one around 3.5 Å. Both peaks are extremely clear for O with the second one 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 the bonds established between the element in question and the N, O or S atom, respectively. However, this main peak can be an overlap of two peaks - one of them represents a mono-coordinated ligand and most other ligands in which a single atom interacts with the element, the second accounts for those bi-coordinated ligands, which present bond lengths higher than normal and for the intermediate cases. Zinc is a well researched case in which this situation takes place regarding its binding to oxygen in what is known as the carboxylate shift [4, 28].

The second peak in Figs. 6.A-6.C, regarding the pattern of the curves shown by cobalt, copper, iron, manganese, nickel and zinc in their distances to N, O or S, respectively, occurs in a non-bonding region and it should take into account all those N, O or S atoms, which 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 that situation (as it would happen in the case of Arg or Lys for example), O atoms do occur also (as it would happen in the case of Glu or Asp) but S atoms appear to be far less abundant.

As regards to the metals under study, it is well known that cobalt in small amounts is essential to many living organisms. Even though cobalt is one of the elements that we have studied in more detail 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 happens in vitamin B_{12} but there are also many cases in which cobalt is directly linked to the protein structure, having a preference for residues such as histidine, glutamate and aspartate, as it happens in methionine aminopeptidase [29], integrin [30, 31] and many other cases.

Copper is an essential element in all plants and animals, which 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 them, copper is directly linked to the protein structure.

Iron is a necessary element used by almost all living organisms, often incorporated in heme prosthetic groups, and therefore not directly bound to the protein structure as such. However, it also turns up very commonly in the centre of metalloproteins, as in many transferrines and

forms of superoxide dismutase, often in sulfur complexes (such as Fe_3S_4 and Fe_4S_4), even though not as commonly as in within a heme.

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

Zinc is one of the most abundant transition elements in living organisms, an essential component of a very large number of enzymes 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 object of further studies that have been reported recently in the literature in detail [4], providing valuable guidelines for the study of biological Zn systems.

Calcium, potassium, magnesium and sodium present a small peak near 2 or 2.5 Å but most of their distance values are in 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, specially in the case of magnesium, the resulting metalloproteins can be of extreme importance, such as happens with farnesyltransferase [28, 35, 36], geranylgeranyl transferase [37] and integrase [38, 39] to name but a few. Those are the cases that account for the small peak near 2 or 2.5 Å. However, the distances relative to 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 due to crystallization techniques. The pattern shown by cadmium is somewhat different, showing peaks still within the coordination radius, but at a slightly larger distance. An interesting case is a carbonic anhydrase of marine diatoms that can switch between zinc and cadmium in its active site using one or another in function of natural availability of each metal in a given moment. The enzyme is more efficient using zinc, however it is completely functional with cadmium [40]. Some cadmium and mercury binding proteins have detoxification [23, 41] functions. Some proteins with arsenic are also detoxifying proteins but the presence of As is in most of cases associated to cacodylate ion (from dimethil arsinic acid, $(CH3)_2AsO_2H$), probably from buffer solutions.[42]. However, in many proteins with structures in the PDB these elements occur as experimental or methodological insertions, often in substitution of calcium or zinc ions either to facilitate

cation identification [43] or to study ion substitution 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* maximum frequency of occurrence of distances to N, O and S atoms, pertaining to residues of protein, are shown in Fig. 7.

These graphs was built with the values shown in Fig. 6, using as ordinate value the first peak of the frequency of occurrence polygon up to 3.5 Å only. Therefore, the maximum frequency of occurrence distances at which they show up in Fig. 6, although pointing to a definite trend, cannot be taken as correct absolute values.

We can observe a set composed by Co, Cu, Fe, Mn, Ni and Zn, all largely known as cofactors or protein components, plus Cd, more unexpectedly. There is, however, a clear distinction between the two sets of elements, *i.e.* (Fe, Ni, Co, Zn, Mn and Cd) and (Hg, Mg, Na, Ca and K), with a clear gap in peak distance between them.

<Figure 7A and 7B and 7C>

Metals and residues

A second approach to this metal classification is the relationship between the studied metals and the residues in the proteins. For this analysis we used all residues within a 5 Å radius neighbourhood, a distance chosen taking into consideration the average dimension of a residue. Therefore, residues farther from this value are assumed as not directly affecting the metal.

We must notice that this analysis isj not directly related to the one presented beforehand, because there we did not use distances but instead we have counted the number of residues overlapping a sphere centred on the studied metal.

The results obtained have been arranged in a matrix of elements *vs.* residues. Therefore, the matrix values are none other than the frequency of occurrence of each residue appearing in within 5 Å of each studied element.

We applied cluster analysis techniques on this data, 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, which may be considered as a *very good fit* either for single linkage or for UPGMA. The high value obtained in single linkage approach suggests that there are fairly distinct clusters.

<Figure 8A and 8B>

These results are shown in Fig. 8 using both algorithms - single linkage and UPGMA.

Analysis of the single linkage phenogram suggests the existence of a well defined cluster constituted by 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 still further up with Cu, K Na and Mg. On the other hand, Zn, Ca, and Fe are outliers, showing a peculiar association with residues specific for each one. The topology of the UPGMA phenogram is very similar, differing only in the association order of Ca, Zn and Fe.

It is noteworthy the close association of Co, As, W and Mo (0.201), indicating a similar behaviour in relation to protein residues.

Ordination in reduced space

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

We utilised the principal component analysis method, using its geometrical properties more than the statistical ones. A correlation matrix between pairs of lines of the data matrix (*i.e*. vectors of residue frequency of occurrence for each element) was computed. We extracted the eigenvectors from the correlation matrix, each one representing 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 from the total. This set of vectors may be used to project the original points onto this new base. This performs a rotation of all space, allowing a projection in a two or three dimension plot, more easily interpretable.

<Table 2>

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

<Table 3>

<Figure 9A and 9B>

Figs. 9.A and 9.B are biplots showing simultaneously vectors (residue frequencies within the 5 Å sphere) and metals projected in a rotated axis. Position of each element is related to absolute values of scores in the projected matrix. Therefore, the vectors aligned (pointing to or not) with each point are the residues whose presence or absence in its neighbourhood characterizes in some way the behaviour of the metal.

In the projection of axes I and II, it is remarkable to witness a “size effect” with all factors with positive value in axis I. More interesting is the projection of axes II and III, which show a contrasting behaviour between some amino acid residues.

It is easy to notice the association of Zn with His and Cys, which is in perfect agreement with other findings reported in the literature[4], emphasizing the reliability of the methodology used here. Fe has also a strong association with Cys and His, but also with Met. The set constituted by Mn, Co, W, As, Ni, Hg and Cd is in a central position, with no special preference for any residue, and can be seen in (Fig. 9.A) with negative values in the first axis. Cu shows some association with Met, but presents negative loads from His and Cys, which are strongly associated to Fe and Zn.

K and Mg appear to have adopted an eclectic behaviour, near the centre, and therefore not associated specifically to any residue. Ca appears positively aligned with Asn and Asp and negatively with His and Cys, indicating that these residues are less frequent than the average in the metal neighbourhood.

As with the SAHN method, we can notice the similarity of behaviour between Mn, Co, W, As, Ni, Hg and Cd.

<Figure 10>

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

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 of the PDB. Only fifteen of them are present in a significant amount of representations, and it is interesting to notice that the first in rank of these elements signals an important increase in abundance relatively to the previous one. All this data was used to build a protein oriented periodic table.

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

Those cases of Cd and Hg are interesting and, in the analysis of residues' neighbourhood frequency of occurrence, Cd seems to adopt a behaviour similar to that of Mn, and similarly Hg seems to follow that of Co and Ni. The most dramatic behaviour 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 may suggest some unidentified relations between metals and specific residues.

Finally, we must remark that the PDB is not really a sample of protein structure but instead a sample of our knowledge and interest about proteins. As an example, 2414 structures have the

word “termophillic” in their title or keywords, being termophillic organisms a very small part of biosphere both in biomass and in taxonomic significance. As referred beforehand, biological species statistical distribution in the PDB is very different that nature’s.

Some tantalizing thoughts, with no possibility of serious statistical corroboration presently, suggest the possibility of continuation and deepening of this study in the future.

References

1. Berman HM, Westbrook J, Feng Z et al (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 Ca2+-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 Biol Crystallogr 60(Pt 5):849-859
8. Harding MM (2002) Metal-ligand geometry relevant to proteins and in proteins: sodium and potassium. Acta Crystallogr D Biol Crystallogr 58(Pt 5):872-874
9. Harding MM (2001) Geometry of metal-ligand interactions in proteins. Acta Crystallogr D Biol Crystallogr 57(Pt 3):401-411
10. Harding MM (1999) The geometry of metal-ligand interactions relevant to proteins. Acta Crystallogr D Biol Crystallogr 55(Pt 8):1432-443
11. Harding MM (2006) Small revisions to predicted distances around metal sites in proteins. Acta Crystallogr D Biol Crystallogr 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 Biol Crystallogr 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 Chryst 41:963-968
14. MS-Access, MS-Excel: Microsoft Corporation; 2007
15. Rohlf FJ (2004) NTSYSpc, Numerical Taxonomy Sistem. 2.2: Applied Biostatistics, Inc

16. Sneath PHA, Sokal RR (1973) Numerical Taxonomy, The principles and practice of numerical classification. 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. Theoretical Chemistry Accounts 115(1):27-31
20. Branco RJF, Fernandes PA, Ramos MJ (2006) Molecular dynamics simulations of the enzyme Cu, Zn superoxide dismutase. Journal of Physical Chemistry 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 et al (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 THEOCHEM 729(1-2):141-146
33. Paraskevopoulos K, Sundararajan M, Surendran R et al (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 H2o 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-Struct Biol Cryst Commun 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 Biol Crystallogr 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

Tables

Table 1 Shown is a periodic table from a protein point of view. Elements with symbol in grey have not been included in the study

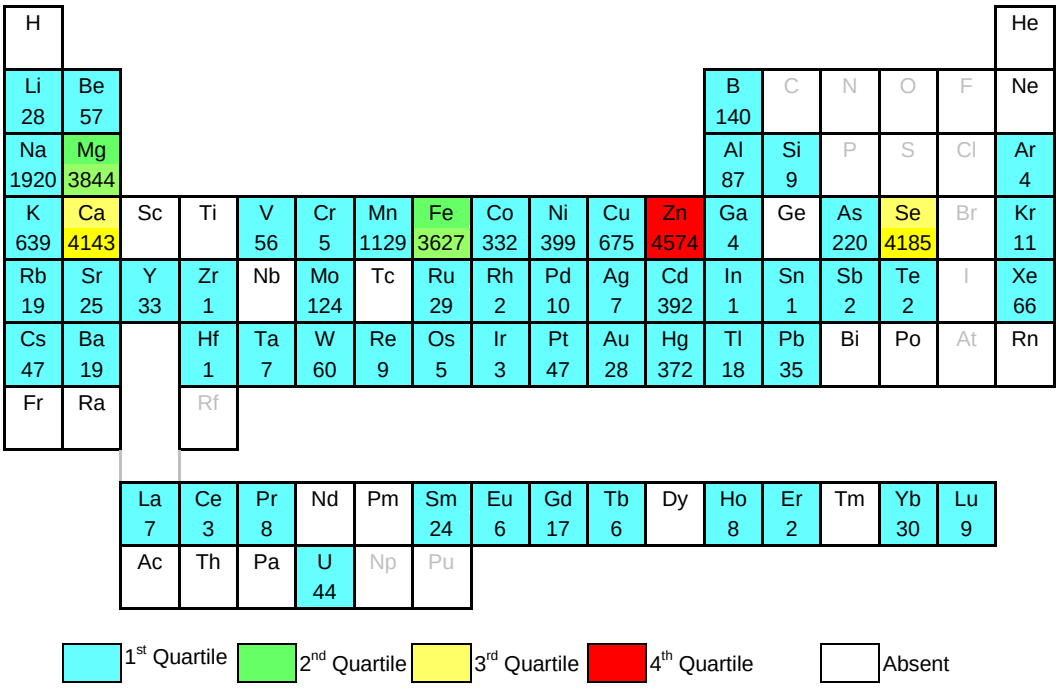

Table 2 Principal component loads for axes I, II and III

	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

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

Figure captions

Fig. 1 The evolution in time of structure deposition in the PDB is shown. The graphic shows both the three most used methods and the total number of deposited structures per year

Fig. (**a**) The evolution in time of structure resolution, determined by X-ray diffraction, is depicted in blue. Average values are shown in green
(**b**)The evolution in time of structure resolution determined by X-ray diffraction is depicted in blue (best values) and in red (worst values). Average values are shown in green

Fig. 3 The evolution in time of the molecular masses of yearly deposited structures in the PDB, determined by X-ray diffraction, is depicted in blue (highest values) and in red (lowest values). Average values are shown in green

Fig. 4 Histogram showing the number of structures in the PDB according to the different elements considered

Fig. 5 Percentage of metal atoms bound to non-protein structures in comparison to residue coordination (yellow)

Fig. 6 Frequency of occurrence *vs*. distance between metal and O (**A**), N (**B**) and S (**C**) atoms from amino acid residues contained in the proteins of the PDB structures. All distances are in Å

Fig. 7 Atomic radius (Å) *vs*. maximum frequency of occurrence of distances to N, O and S atoms, pertaining to residues of protein

Fig. 8 Phenograms resulting from the number of residues encountered in the neighbourhood of the metals with **A** Single linkage method., **B**. UPGMA method

Fig. 9 Biplot graphics showing elements (dots) and frequency of occurrence of residues (vectors) on the first three principal components (**A**: axes I and II, **B**: axes II and III). The position of a metal is determined by the vectors aligned with its position (not necessarily pointing to it)

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

Figure 1

Click here to download line figure: Fig1.eps

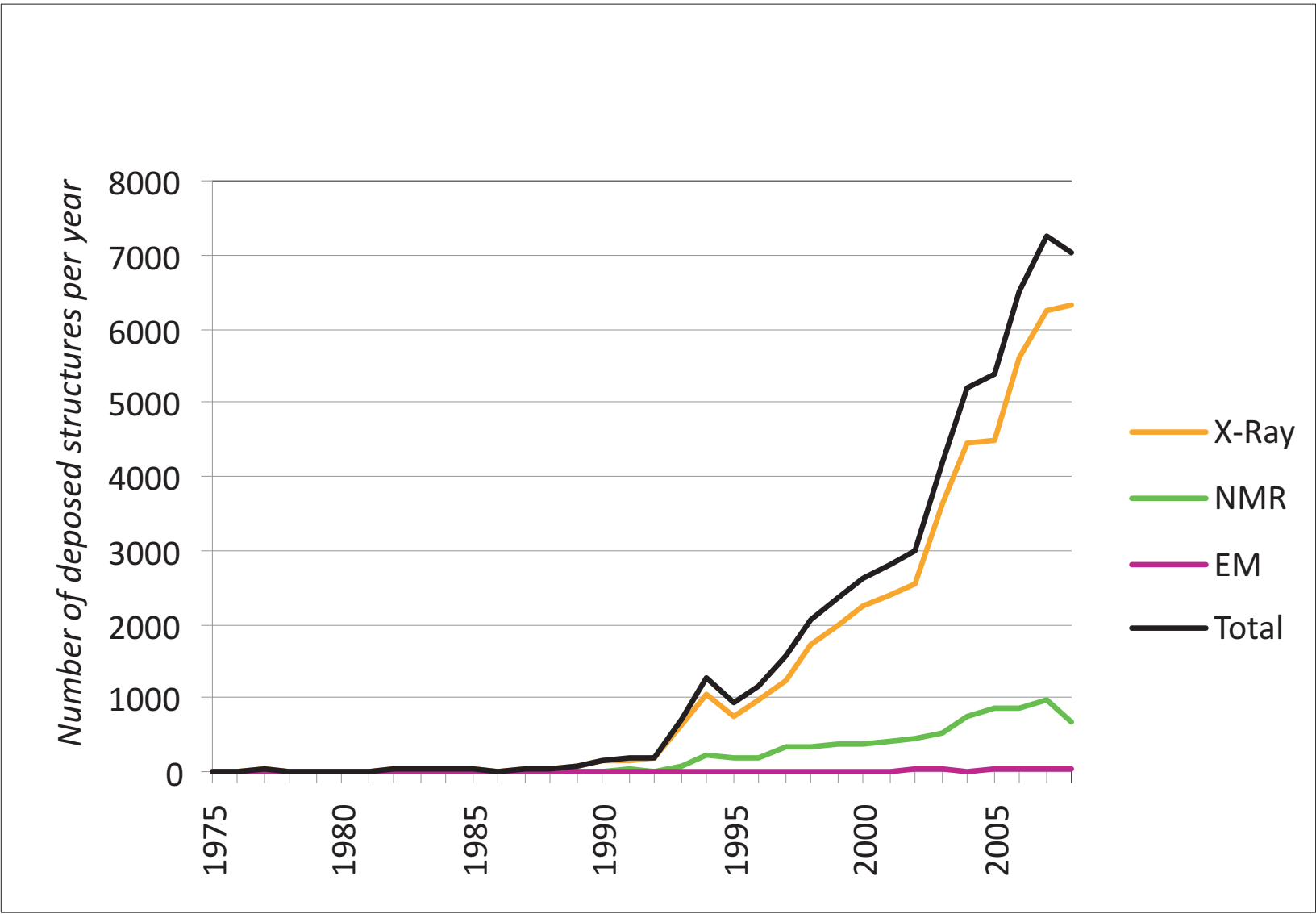

Figure 2a
Click here to download line figure: Fig2a.eps

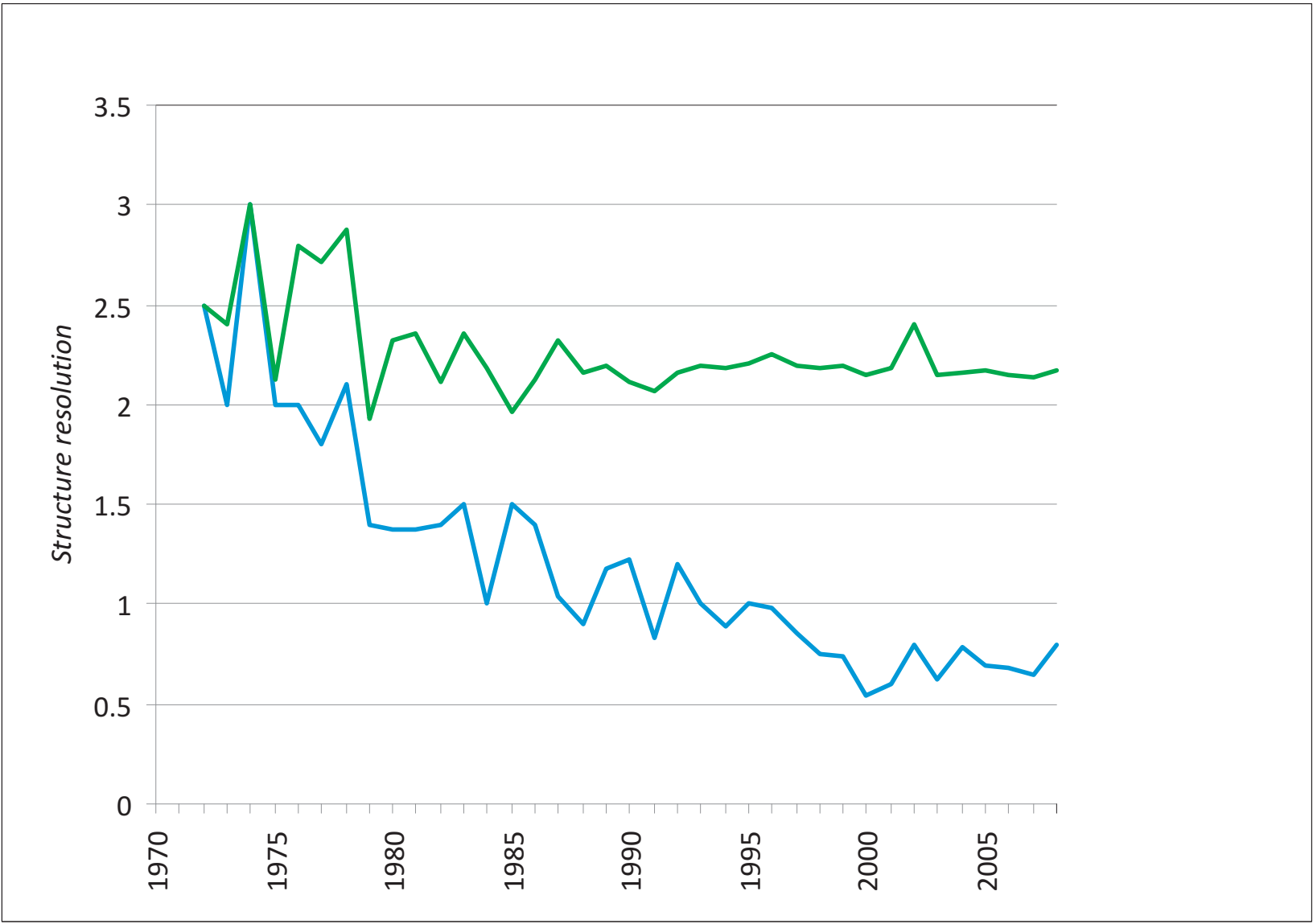

Figure 2b

Click here to download line figure: Fig2b.EPS

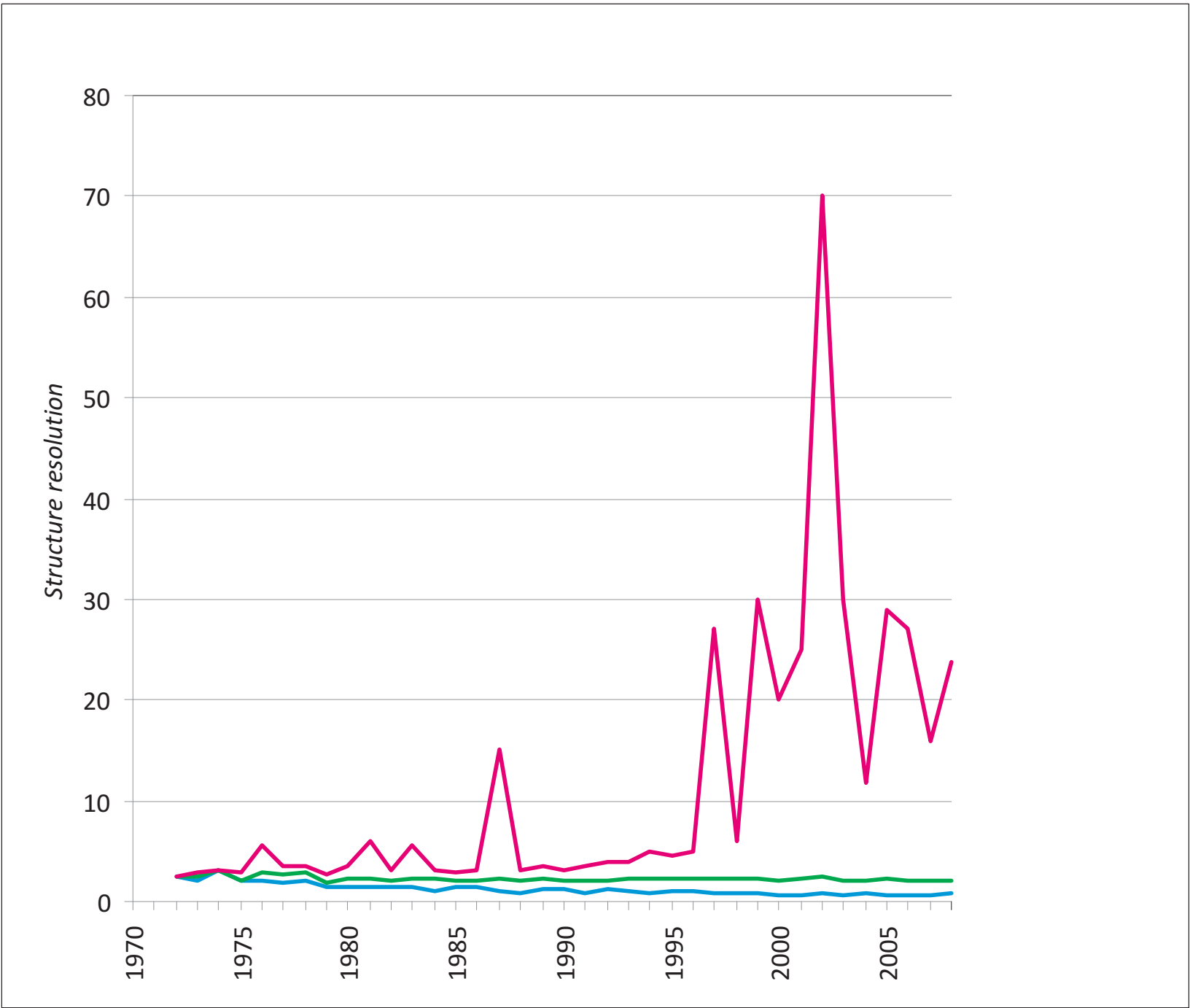

Figure 3

Click here to download line figure: Fig3..eps

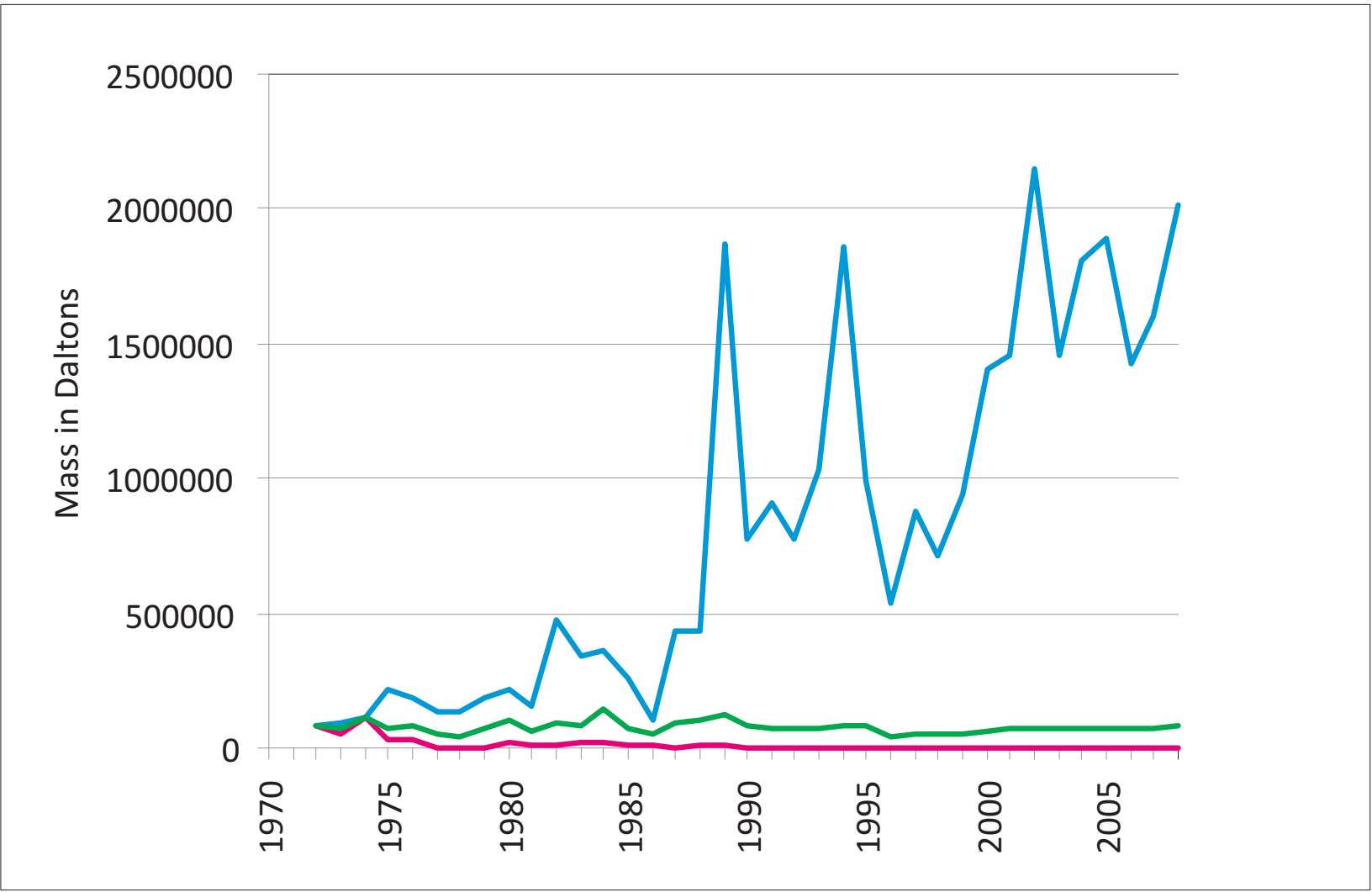

Figure 4

Click here to download line figure: Fig4..eps

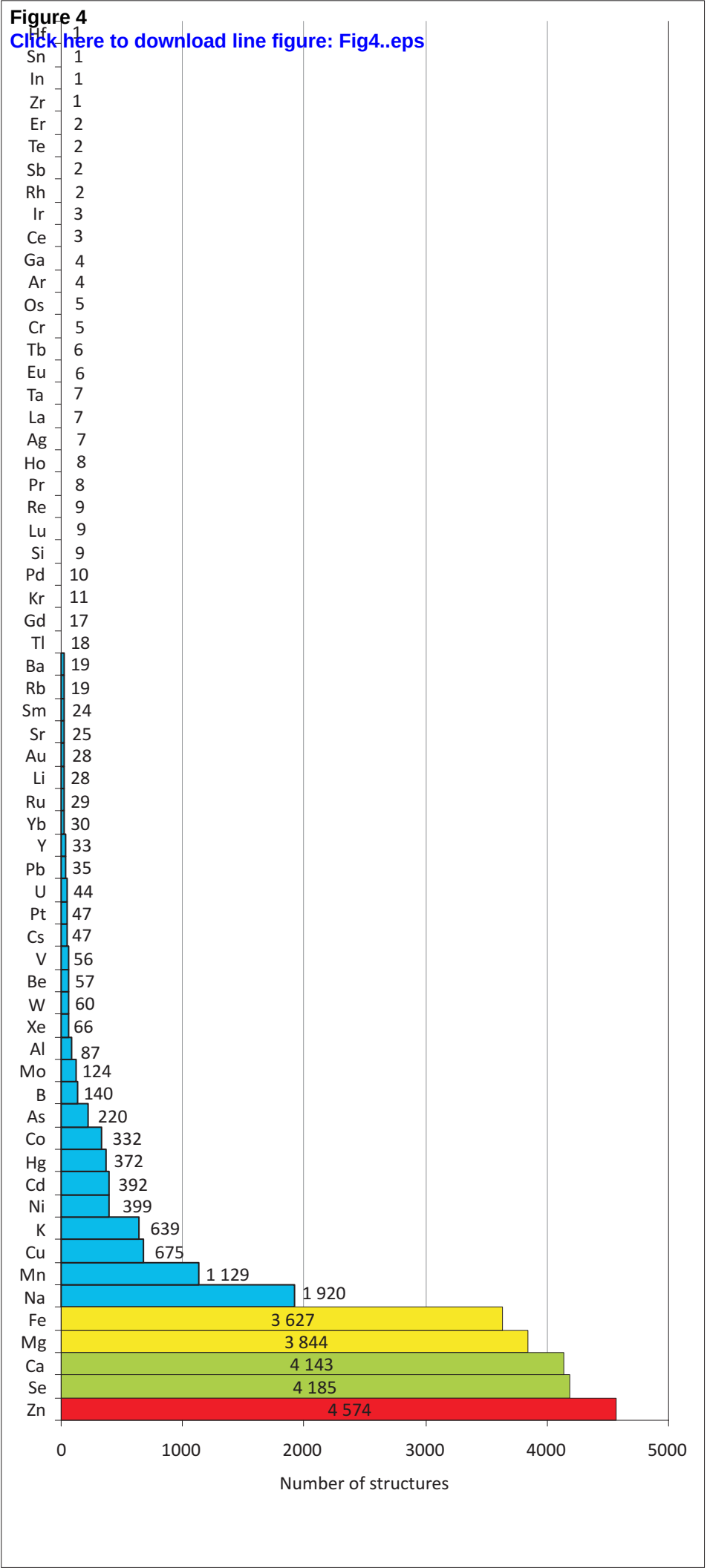

Figure 5

Click here to download line figure: Fig5.eps

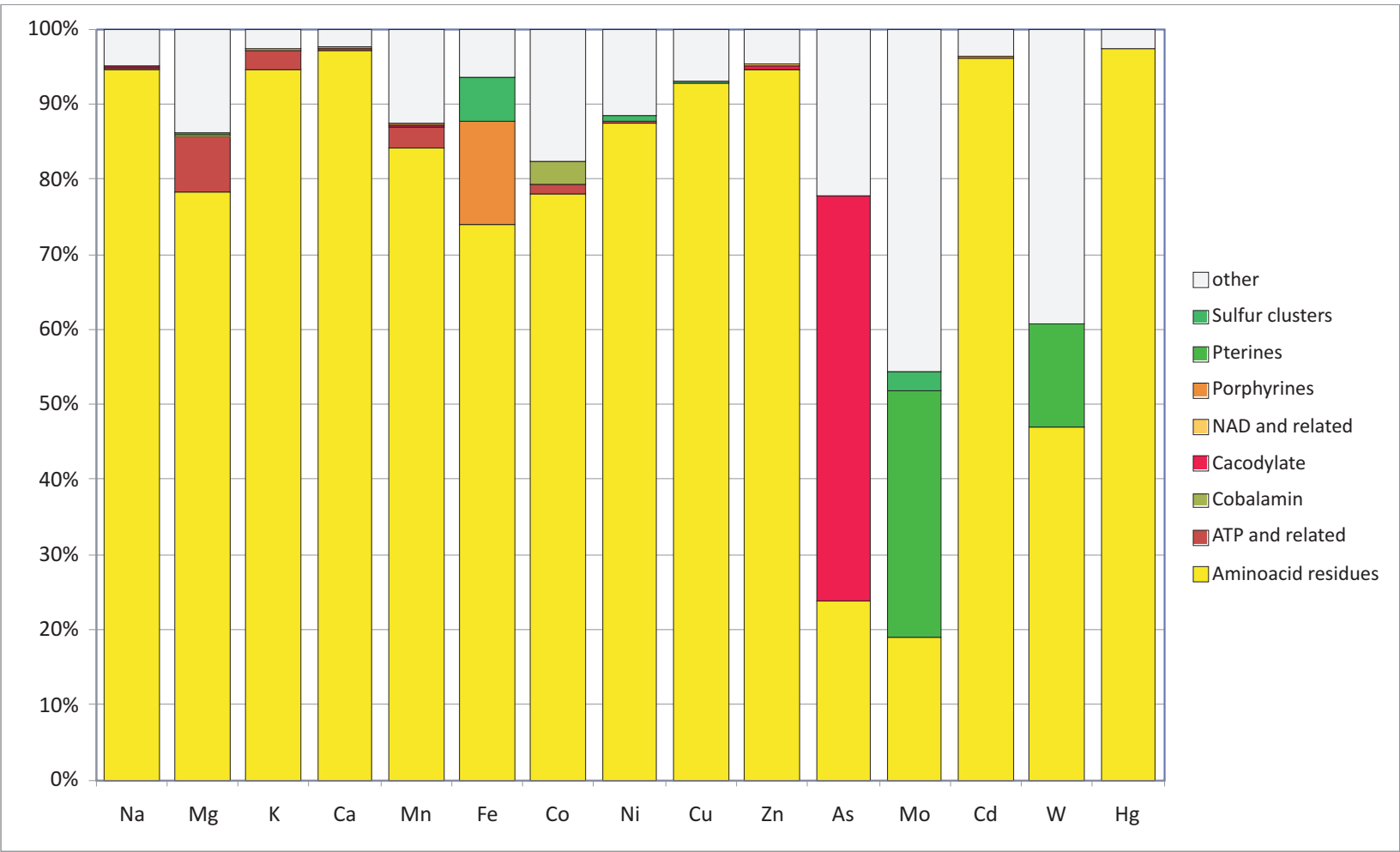

Figure 6a

Click here to download line figure: Fig6a.eps

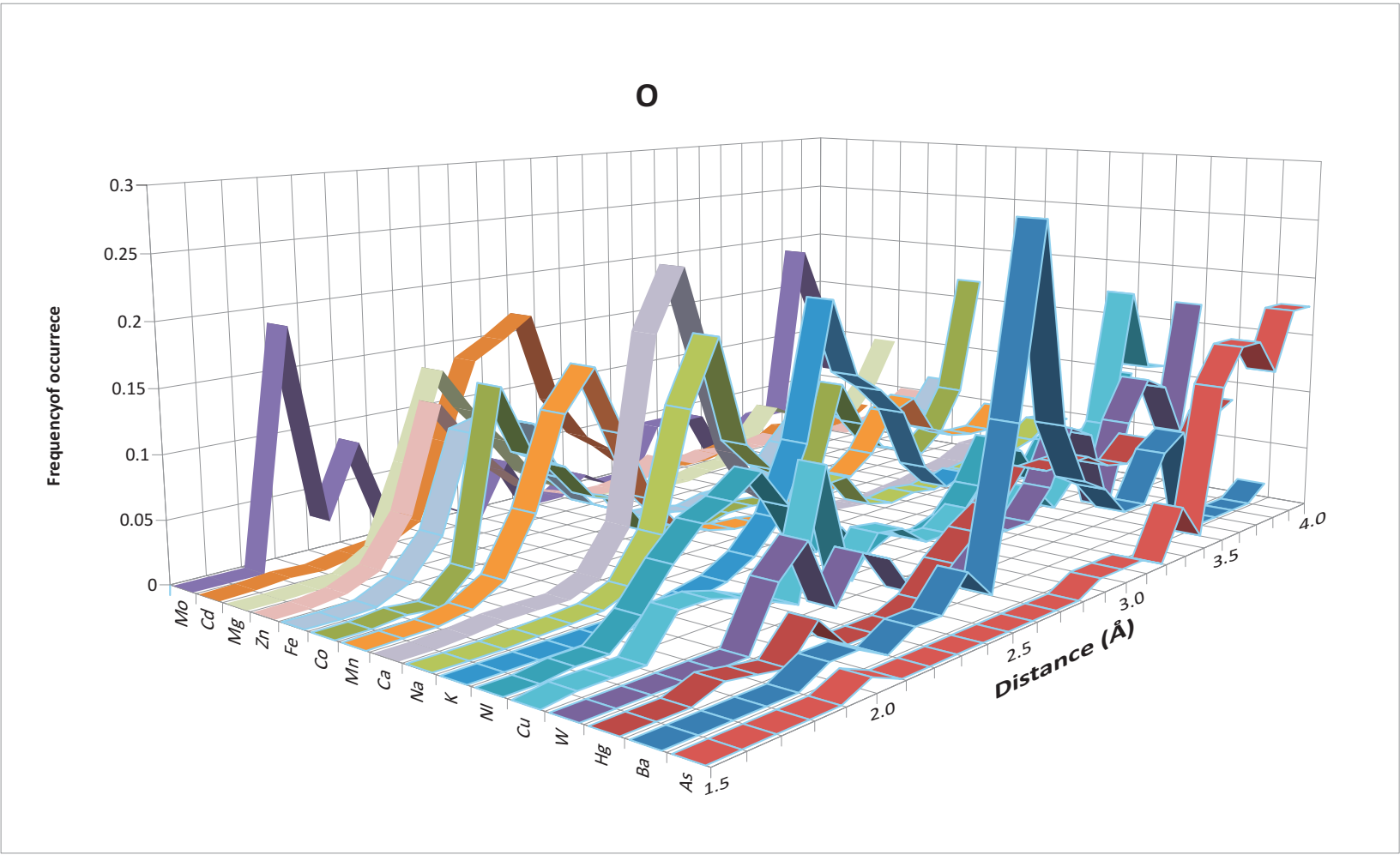

Figure 6b

Click here to download line figure: Fig6b.eps

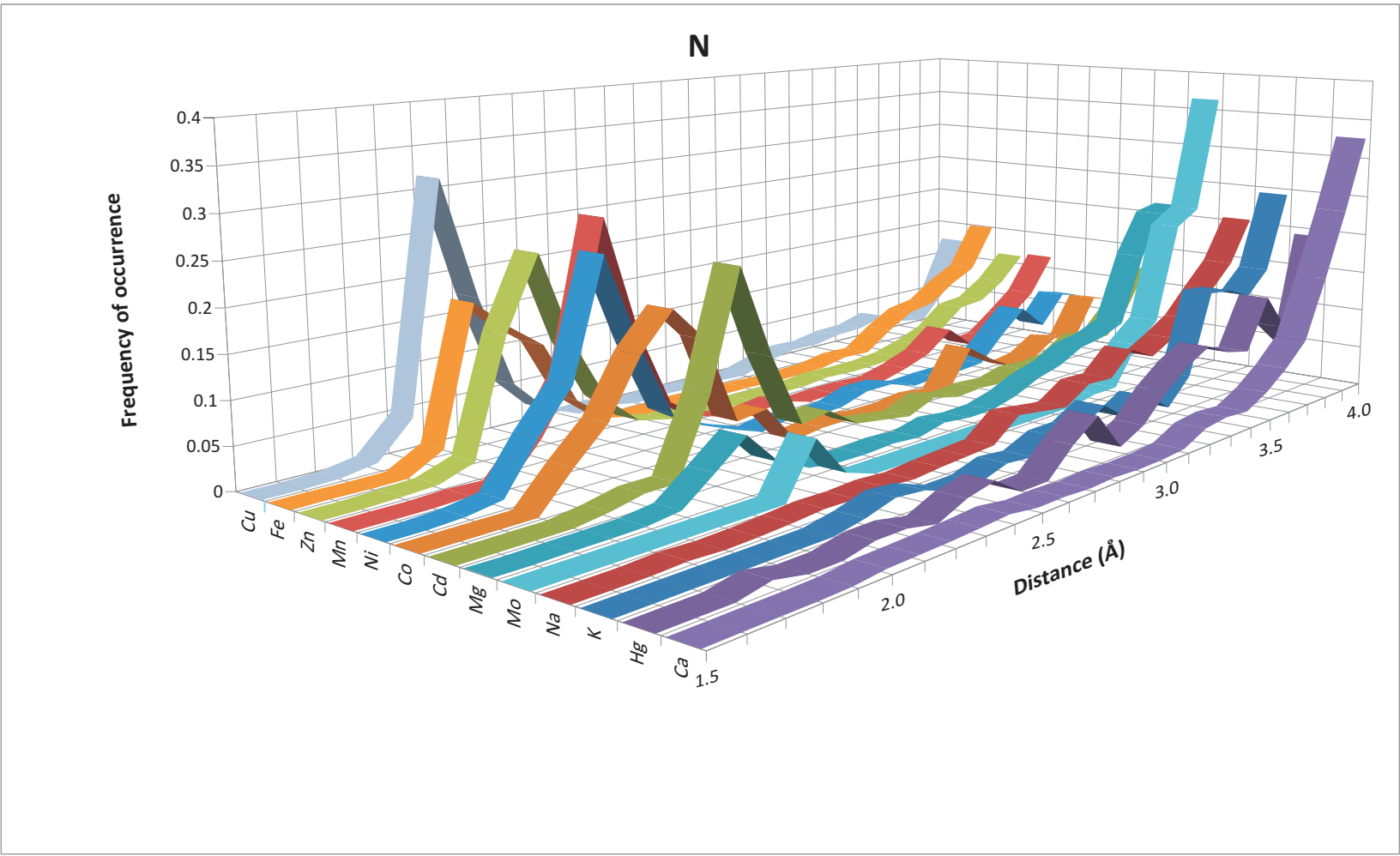

Figure 6c

Click here to download line figure: Fig6c.eps

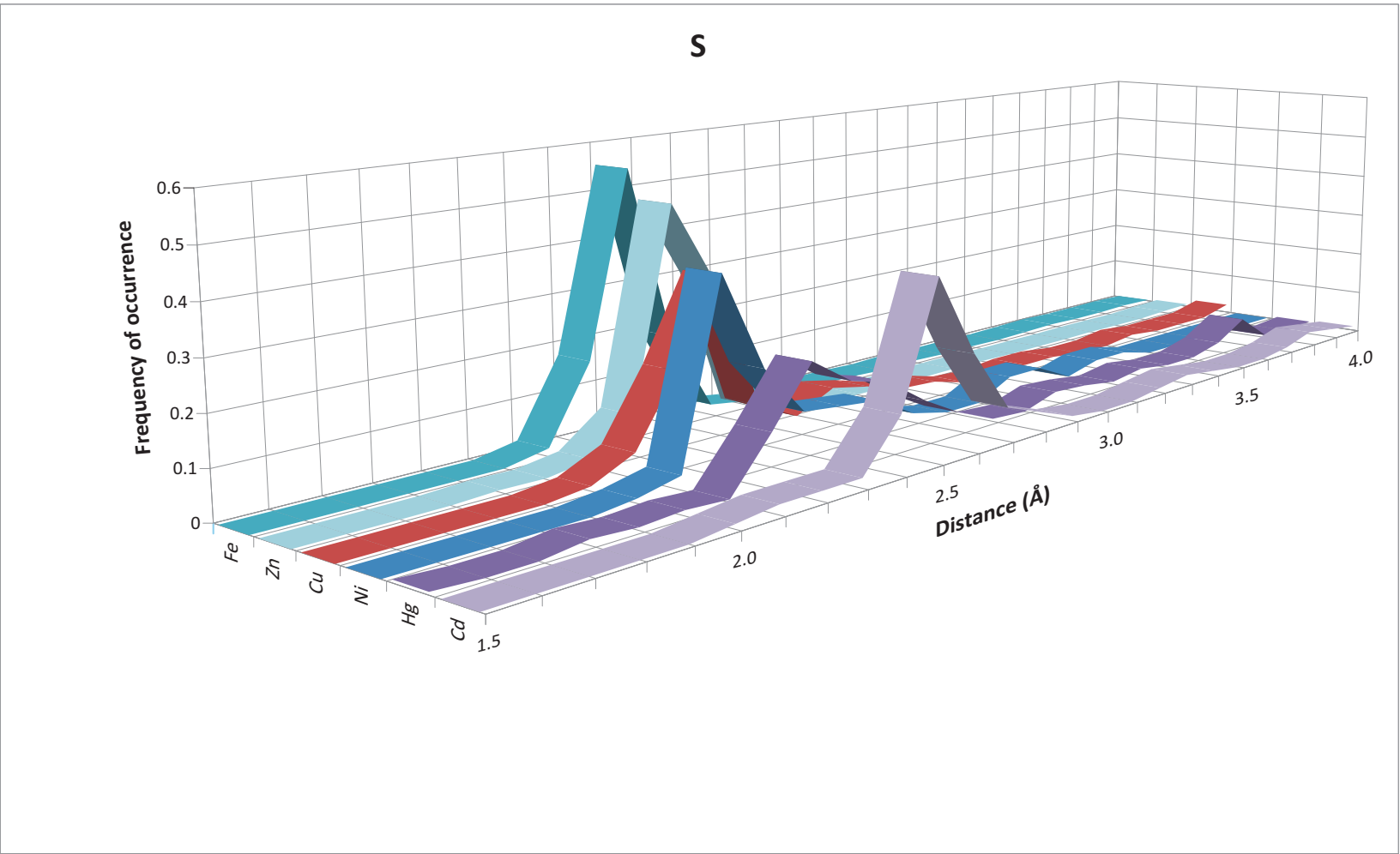

Figure 7a

Click here to download line figure: Fig7a.eps

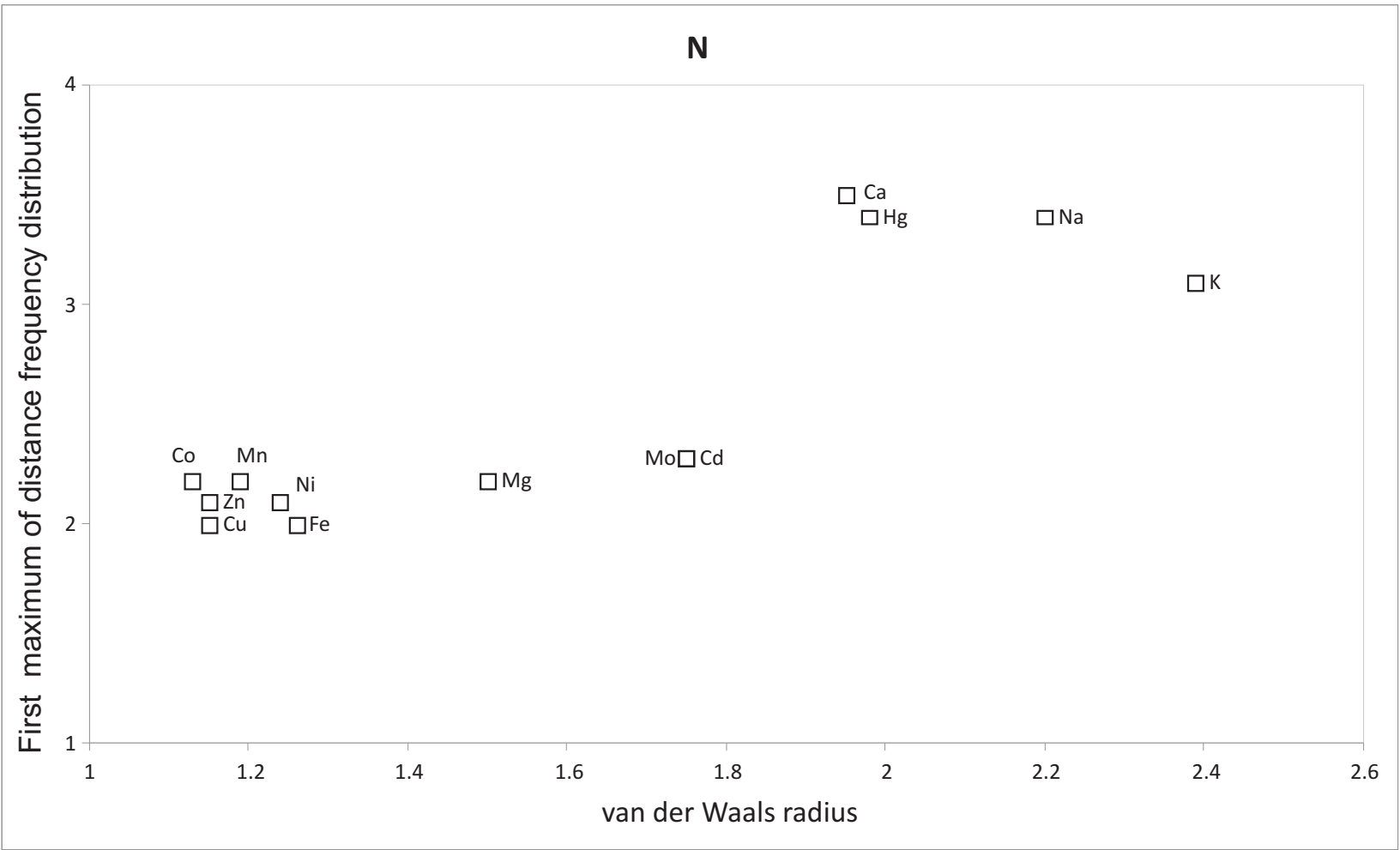

Figure 7b

Click here to download line figure: Fig7b.eps

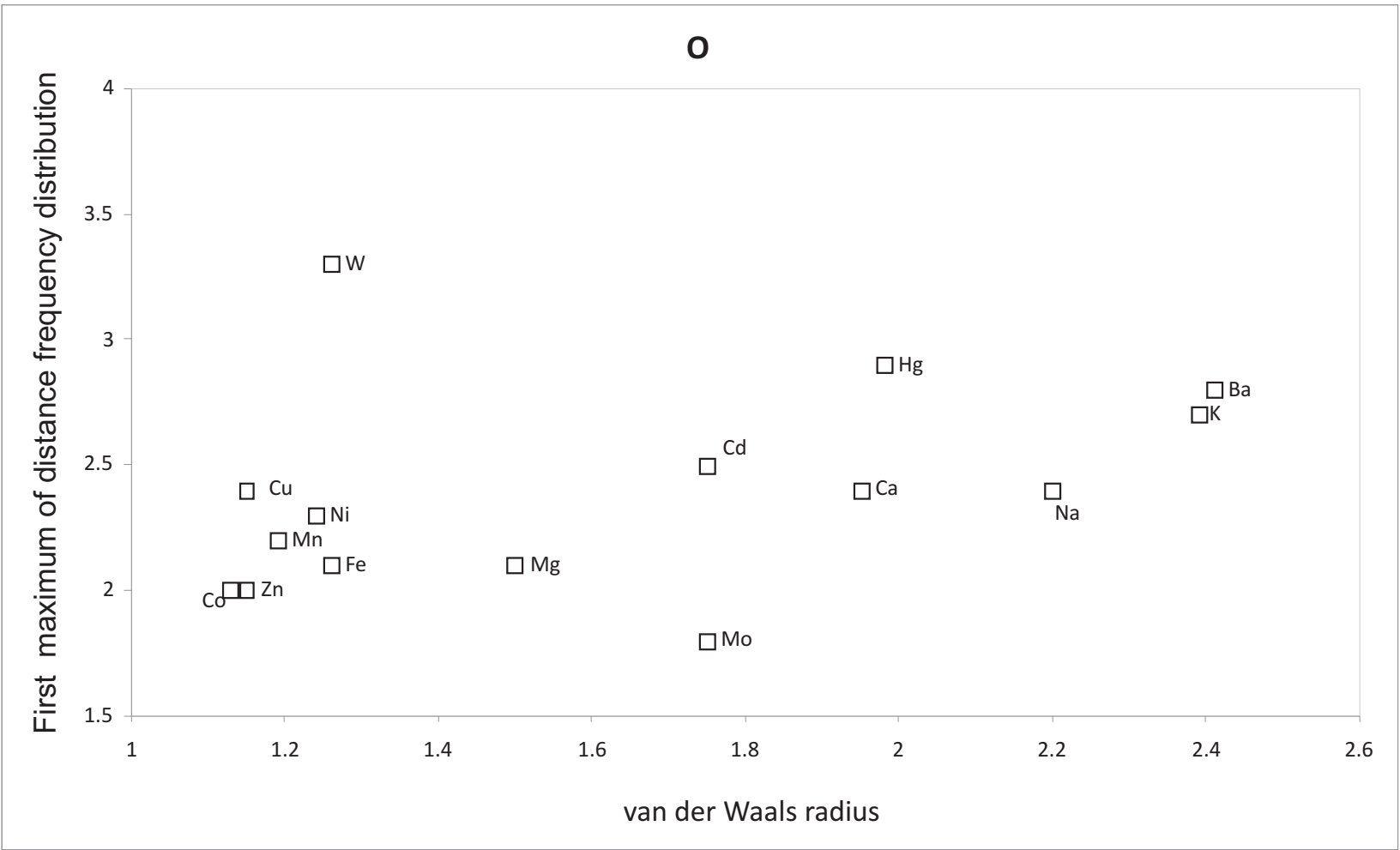

Figure 7c
Click here to download line figure: Fig7c.eps

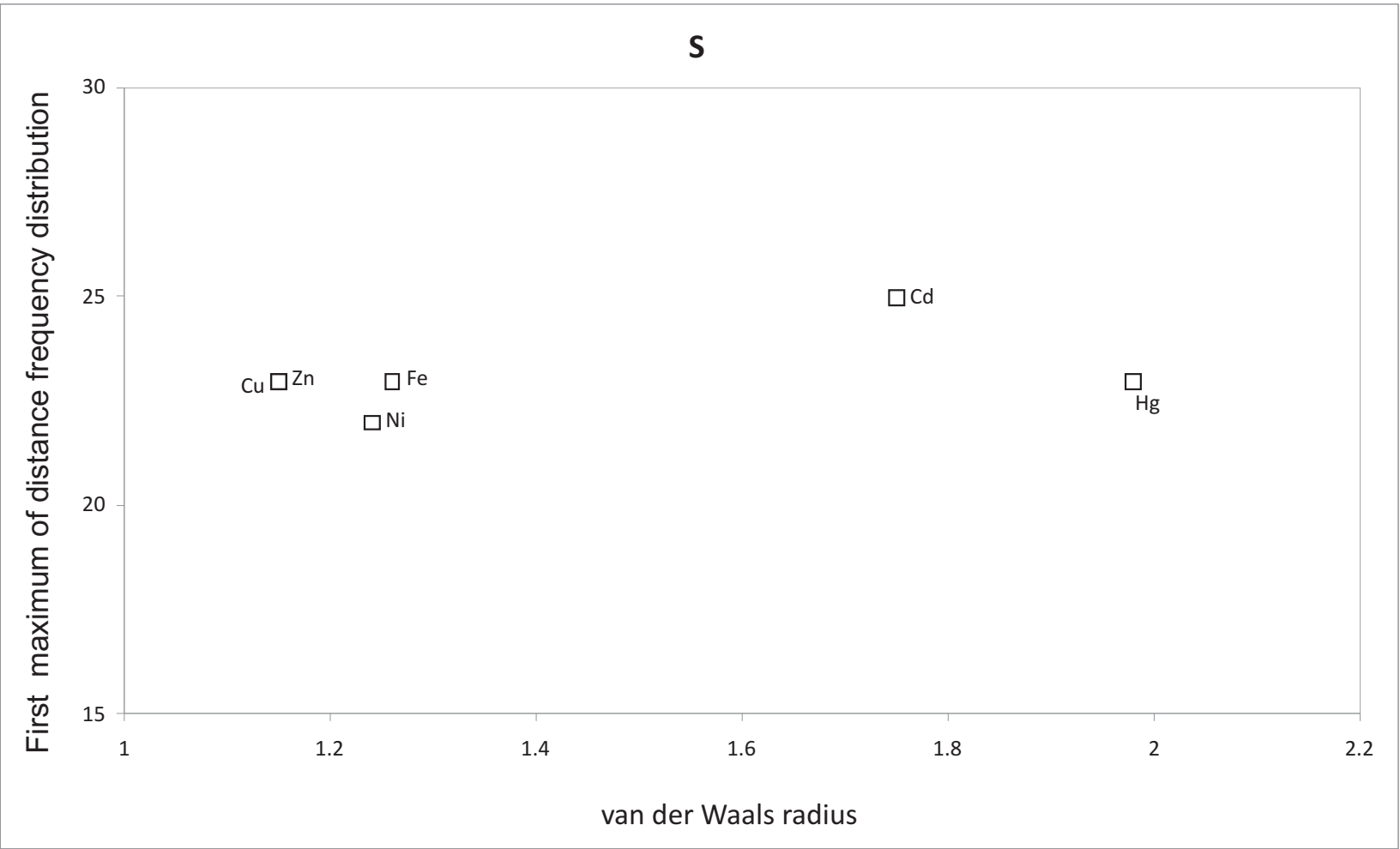

Figure 8a

Click here to download line figure: Fig8a.eps

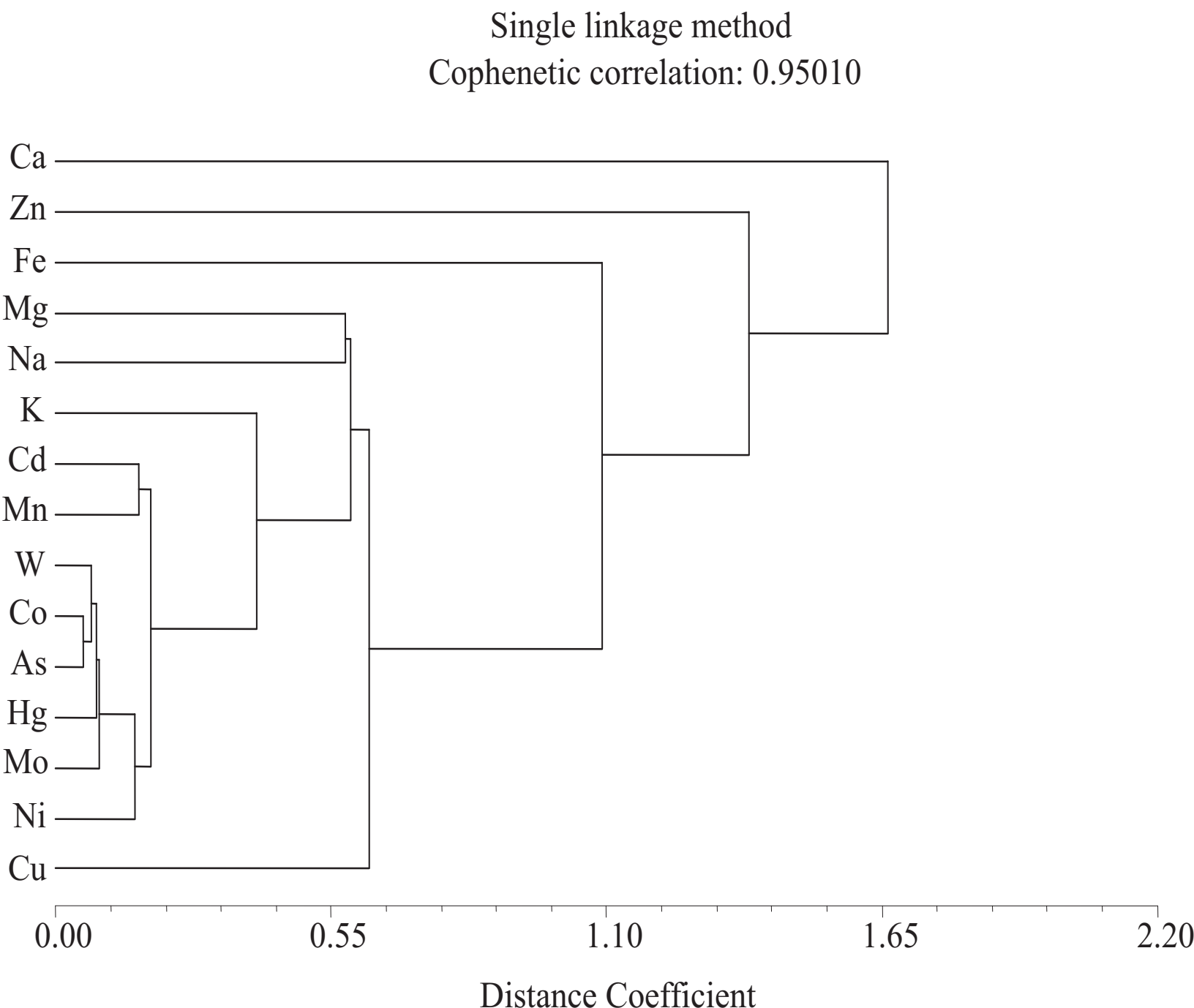

Figure 8b

Click here to download line figure: Fig8b.eps

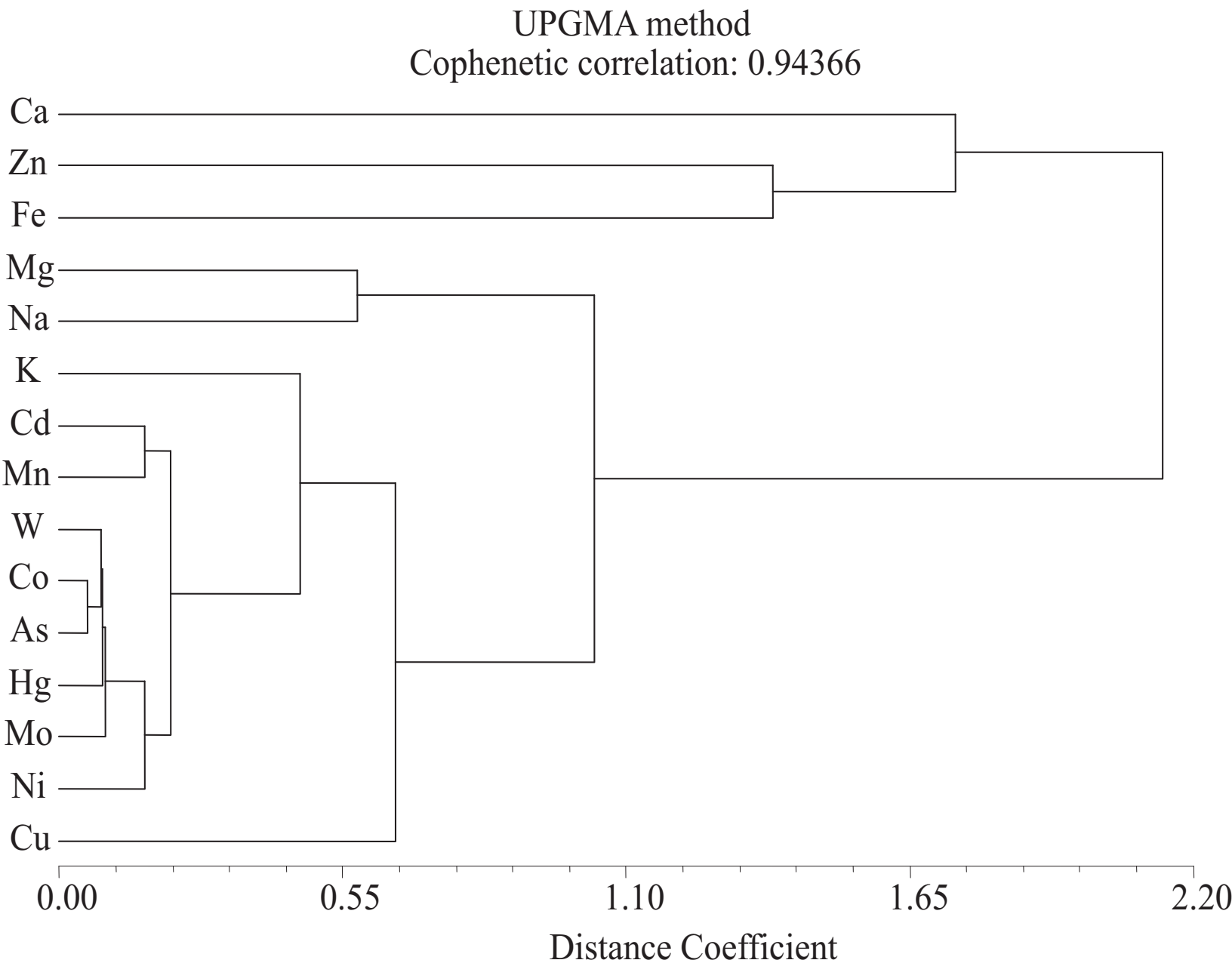

Figure 9a

Click here to download line figure: Fig9a.eps

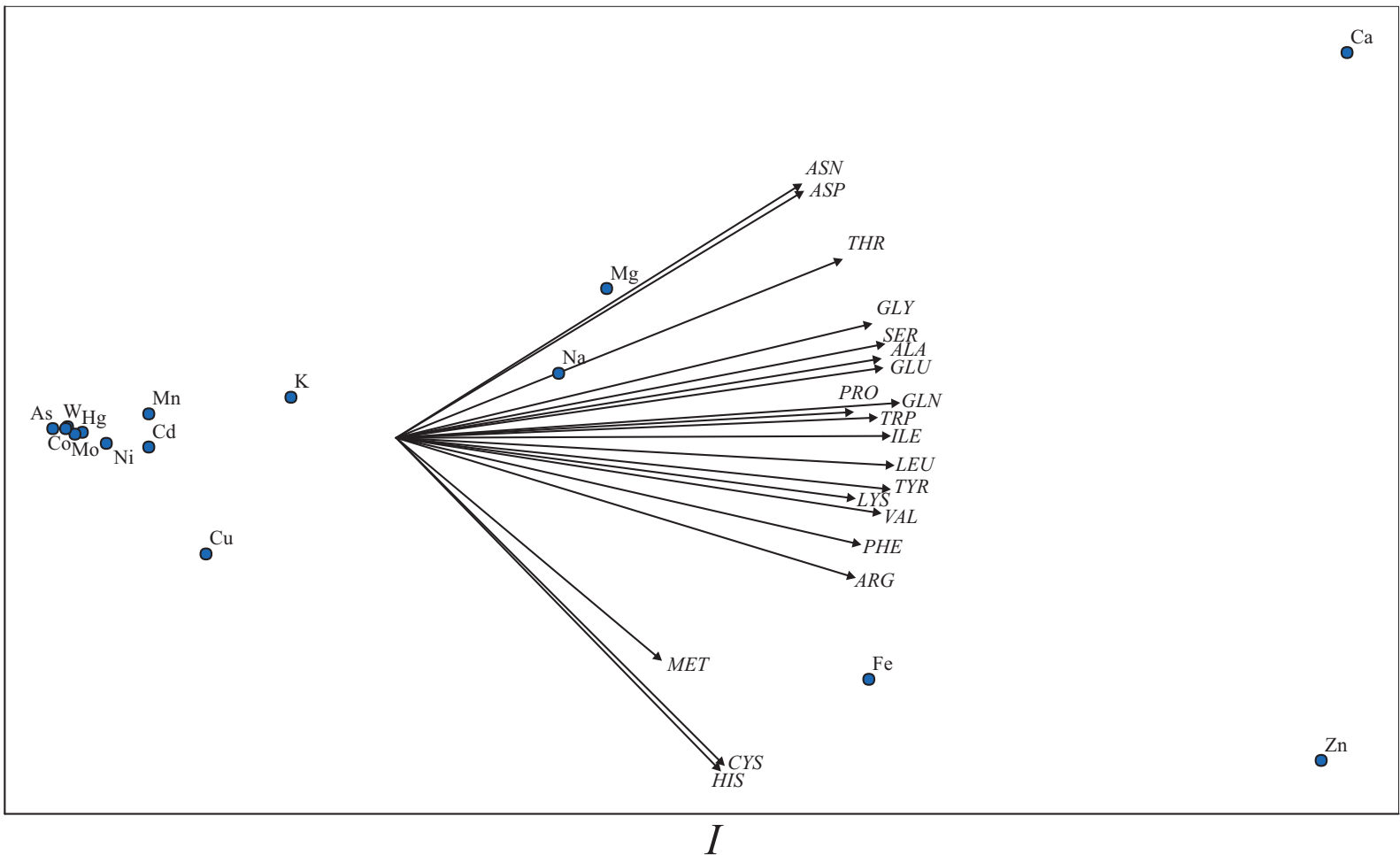

Figure 9b

Click here to download line figure: Fig9b.eps

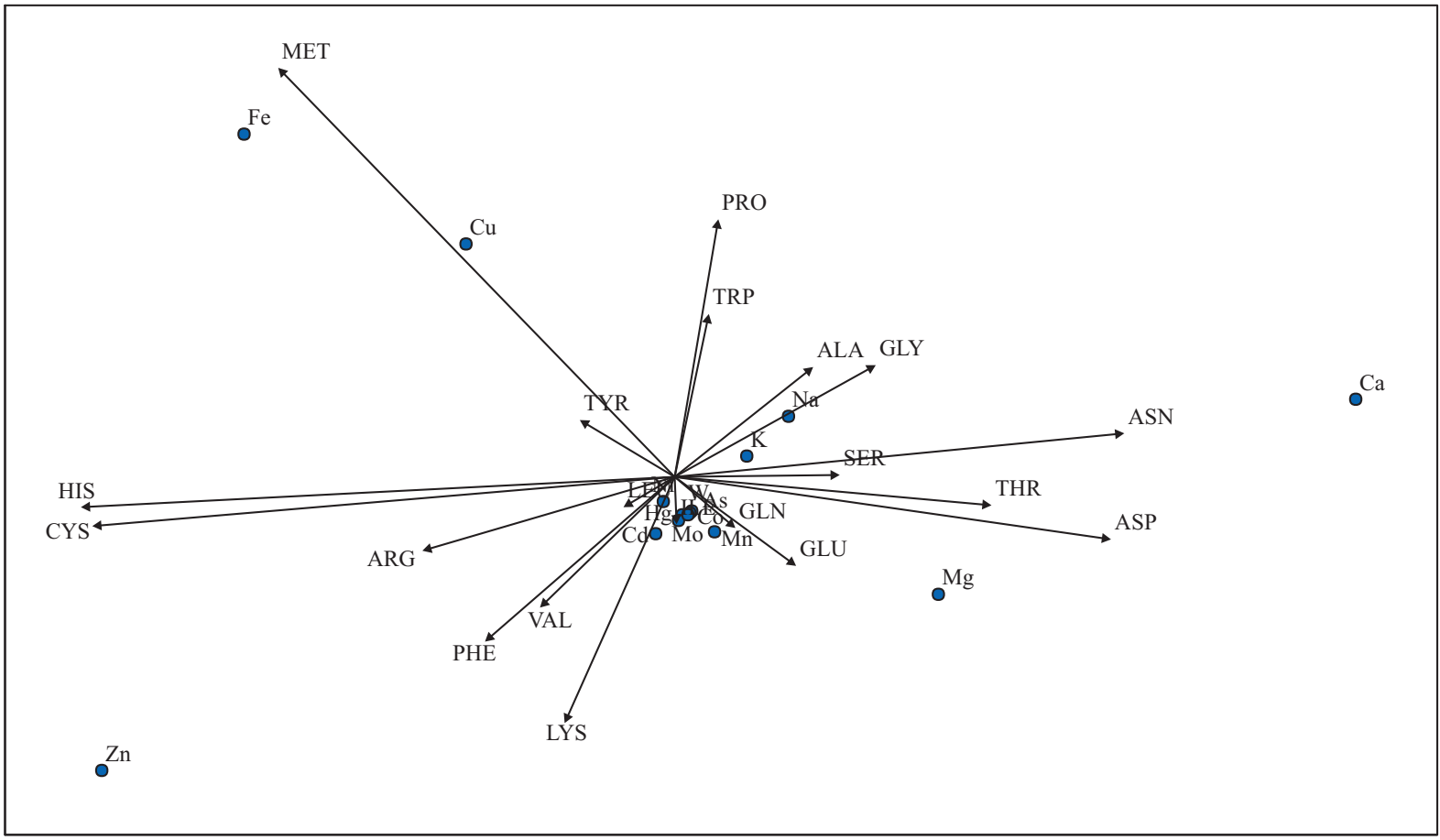

Figure 10

Click here to download line figure: Fig10.eps