Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Is the bond-valence method able to identify metal atoms in protein structures?

Peter Müller,[a]* Sinje Köpke[b] and George M. Sheldrick[b]

[a]UCLA-DOE Laboratory of Structural Biology and Molecular Medicine, 205 MBI, Box 95157, Los Angeles, CA 90095-1570, USA, and [b]Lehrstuhl für Strukturchemie der Universität Göttingen, Tammannstrasse 4, 37077 Göttingen, Germany

Correspondence e-mail: peterm@mbi.ucla.edu

Received 20 May 2002
Accepted 1 October 2002

The proper assignment of metal ions in X-ray structures of proteins is not always easy, but in many cases this knowledge can be important, *e.g.* for an understanding of enzyme mechanism. In this publication, the bond-valence approach is assessed critically. A simplified version, the calcium bond-valence sum (CBVS), is proposed for the convenient analysis of the geometric environment of potential sites with a view to metal-ion assignment. The bond-valence approach is found to be more reliable for structures determined from high-resolution data (1.5 Å or better).

1. Introduction

Many proteins bind metal ions specifically or unspecifically as part of the active site or to stabilize the fold. Frequently, metal atoms are artificially introduced into protein crystals (*e.g. via* soaking) as an aid to solving the phase problem. Other possible sources of metal ions are buffers and other chemicals used during the purification and crystallization process. An informative review about metal binding to proteins is given by Glusker (1991). In principle, metal sites can be assigned unambiguously using anomalous differences measured at wavelengths on each side of the absorption edges of all the elements that come into question, but this is not always practicable. In small-molecule structure determinations, the scattering power of an atom may enable identification in favourable cases, but this may be difficult where isoelectronic ions are involved. In the past, several approaches have been made to gain information about the nature of a metal atom from its coordination geometry (*e.g.* Nayal & Di Cera, 1994, 1996; Harding, 1999), but these have tended to focus on particular species and symmetrical environments. In chemical crystallography and mineralogy, the bond-valence method (Brown & Shannon, 1973; Brown, 1977, 1992; O'Keeffe, 1989; O'Keeffe & Brese, 1991) is a powerful tool to estimate oxidation states of atoms (Süsse & Tilmann, 1987; Palenik, 1997; Shields *et al.*, 2000), to distinguish between *e.g.* H_2O, HO^- and O^{2-} (Hawthorne, 1994) and sometimes to distinguish between elements of similar scattering power (*e.g.* Al and Si in zeolites). Since the coordination polyhedra of metal ions are, in general, less well determined in protein structures than in small-molecule structures, it is not clear whether the bond-valence method will be applicable to proteins; this paper attempts a critical assessment and introduces a procedure suitable for automating the assignment of metal ions.

2. The bond-valence method

The bond-valence method is a quantitative generalization of Pauling's second rule (Pauling, 1929, 1947). The valence (bond

order) v_{ij} of a bond between two atoms i and j is assumed to be a function of the bond length d_{ij},

$$v_{ij} = \exp[(d_0 - d_{ij})/b]. \quad (1)$$

Here, d_0 is the so-called 'bond-valence parameter' describing the expected bond length of a ideal single bond between the atoms i and j. The factor b is usually taken to be a 'universal constant' equal to 0.37 Å (Brown & Altermatt, 1985). The bond valences of all bonds from an atom i sum up to the valency V_i of that atom. For a metal cation the valency is the same as the positive charge,

$$V_i = \sum_j v_{ij}. \quad (2)$$

Suitable values of d_0 have been tabulated by Brown & Altermatt (1985) and Brese & O'Keeffe (1991); the latter were used throughout this work. For example, d_0(CaO) is given by Brese & O'Keeffe (1991) as 1.967 Å, so if Ca^{2+} is symmetrically coordinated by six O atoms, v_{ij} is 2/6 and the Ca—O distance would be predicted to be $1.967 - 0.37\ln(0.3333) = 2.373$ Å. Since d_0(KO) is 2.13 Å, if a Ca^{2+} ion surrounded by six O atoms at equal distances of 2.373 Å was erroneously interpreted as the isoelectronic K^+ (the expected electron density would be very similar), the valency of the ion would be calculated as $6\exp[(2.13 - 2.373)/0.37] = 3.11$, which is quite different from the value of 1.00 expected for K^+, clearly exposing the wrong assignment. However, this example can also be used to illustrate a potential pitfall. Suppose that the actual distances were all measured to be 2.50 Å instead of 2.373 Å, either as a result of experimental error and low-resolution data or of accidentally applied 'anti-bumping' restraints; the calculated valency would then be 2.21 for K^+ and 1.42 for Ca^{2+}, confusing the assignment. A missing (water) ligand would have a similar undesirable effect. Although in these examples the distances to the ligands were all the same, the real strength of the bond-valence method lies in its ability to handle less regular coordination polyhedra effectively.

3. Bond-valence method for protein structures

In order to verify the applicability of the bond-valence method to protein structures, we searched the PDB (Berman *et al.*, 2000) for structures with a reported resolution of 1.8 Å or

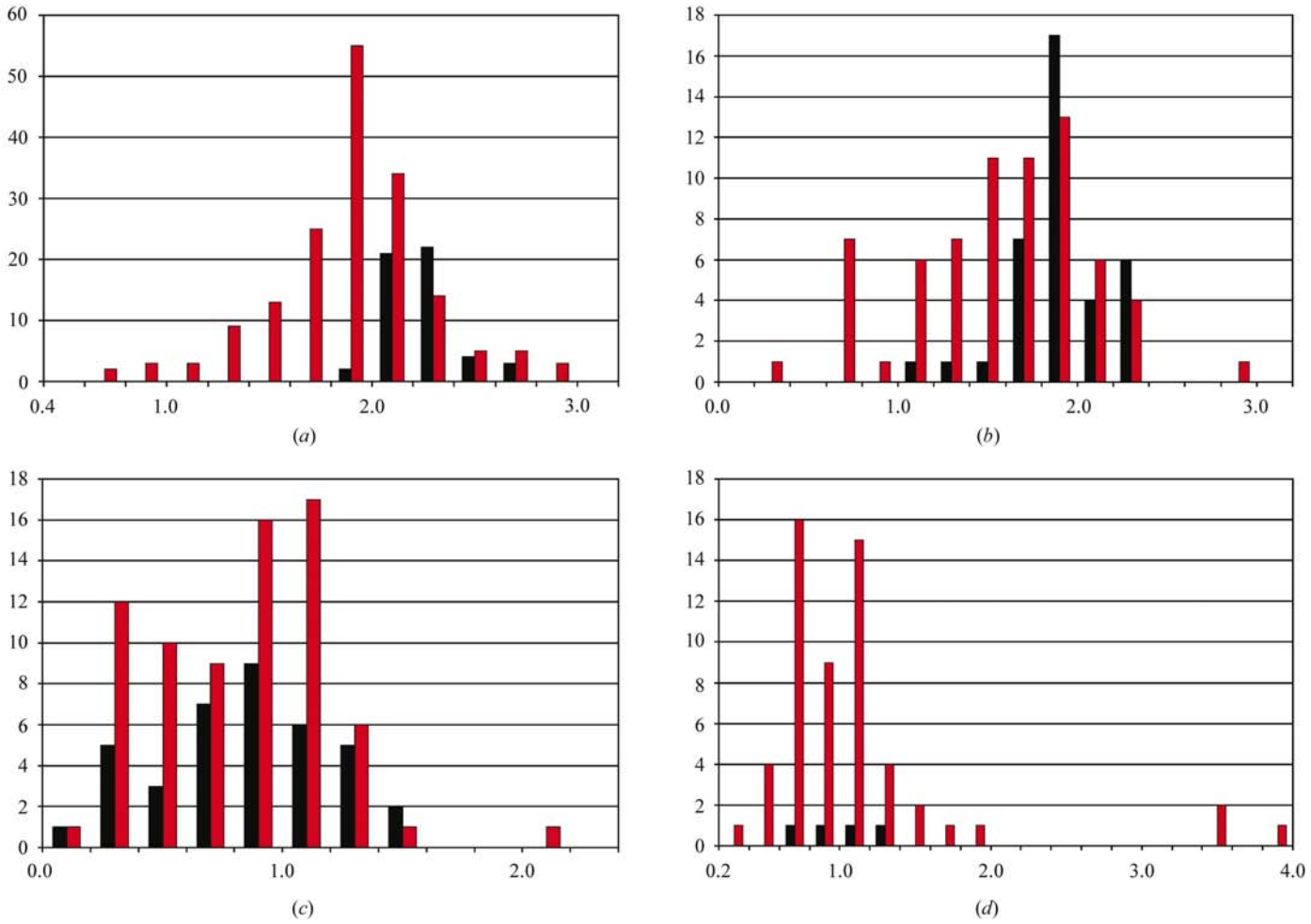

Figure 1
V_{ij} distributions for (*a*) Ca^{2+}, (*b*) Mg^{2+}, (*c*) Na^+ and (*d*) K^+. The vertical axes give the number of structures and the horizontal axes the calculated V_{ij} values; the red bars correspond to structures with a resolution between 1.8 and 1.5 Å and the black bars correspond to structures with a resolution of 1.5 Å or better.

Table 1
Ca—X bond-valence parameters (Å) for some ligands as given by Brese & O'Keeffe (1991).

These values were used to calculate the CBVS values in (3).

	$d_0^{Ca}(MX)$
N	2.14
O	1.967
S	2.45
Cl	2.37
Br	2.49

better and $R_1 = \sum |F_o - F_c|/|F_o| \leq 20\%$ containing fully occupied metal-ion sites. Wherever several isostructural entries for one protein were found, the structure with the better resolution was chosen. The ENVI instruction in the program *XP* (Sheldrick, 2001) was used to generate the environment of each metal atom, including symmetry equivalents. For all bonds, the bond valences were calculated using (1). Each v_{ij} value was multiplied by the occupancy of the corresponding ligand atom. The resulting bond valences v_{ij} were summed according to (2) to give the valencies V_i.

Many structures deposited with the PDB contain Ca atoms, about 200 of which were suitable for our calculations. Fig. 1(*a*) shows the resulting V_i values. Structures corresponding to data with a resolution of 1.5 Å or better (black bars) show a relatively sharp distribution around a mean value of 2.23, which is slightly higher than the theoretically expected value of 2.0. Structures corresponding to data with a resolution between 1.5 and 1.8 Å (red bars) show a much broader distribution of the V_i values around a mean value of 1.89. This can be explained by the inferior data-to-parameter ratio of these structures, which may possibly have led to some ligands being overlooked (*e.g.* half-occupied water molecules). On the whole, the mean values for the valency of the calcium ions lie close enough to the expected value of 2.0 to show that the bond-valence method would be suitable for the identifications of calcium ions in high-resolution protein structures.

Analogous calculations were performed for magnesium (mean V_i = 2.10), sodium (mean V_i = 0.80) and potassium (mean V_i = 0.95) (Figs. 1*b*, 1*c* and 1*d*). Although slightly more ambiguous, the results show that the bond-valence concept can at least give a strong indication as to the nature of a cation.

4. The calcium bond-valence sum (CBVS)

By summing the bond valences of an unidentified atom that is coordinated to O, N or other electronegative atoms, one derives the hypothetical valency of that atom. Since it is somewhat inconvenient to perform this calculation with the d_0 values for each possible cation, we introduce the calcium bond-valence sum (CBVS). The CBVS is calculated as a normal bond-valence sum, assuming that the cation is calcium, *i.e.* using the d_0 values for Ca—X interactions as given in Table 1. Provided the coordination sphere of the cation is complete, the CBVS should possess values of about 2 for the case that the ion is indeed calcium and other values for other elements.

Table 2
Bond-valence parameters, M—O distances (Å) and estimated CBVS values for some cations.

	$d_0(MO)$†	$d(MO)$‡	$d(MO)$§	CBVS¶
NH_4^+	2.219	2.88††	—	0.51††
K^+	2.13	2.79	2.84	0.64
Na^+	1.80	2.46	2.42	1.57
Ca^{2+}	1.967	2.37	2.38	2.00
Mn^{2+}	1.790	2.20	2.19	3.23
Fe^{2+}	1.734	2.14	2.12	3.75
Zn^{2+}	1.705	2.11	2.11	4.07
Mg^{2+}	1.693	2.10	2.07	4.19
Fe^{3+}	1.759	2.02	2.06	5.26

† Bond-valence parameters from García-Rodríguez *et al.* (2000) for NH_4^+ and from Brese & O'Keeffe (1991) for all other cations. ‡ Calculated from the first column using (1) and (2) assuming symmetrical octahedral coordination. § Mean experimental values (in general for octahedral coordination) from the Cambridge Structural Database (CSD; Allen *et al.*, 1991; Allen & Kennard, 1993) compiled by Harding (1999, 2002). ¶ Calculumn using (3). †† Assuming tetrahedral coordination (more reasonable for ammonium) gives $d(MO)$ = 2.73 and CBVS = 0.51. As this illustrates, the assumption of particular coordination has very little effect on the predicted CBVS value.

To take into account the effect of disordered ligands, the calculated valences are multiplied by the occupancies p_j of the ligands prior to summation,

$$\mathrm{CBVS}_i = \sum_j \left[\exp\left(\frac{d_0^{Ca} - d_{ij}}{b}\right) p_j\right]. \qquad (3)$$

For most of the metals frequently present in protein structures only one oxidation state is reasonable (*e.g.* 1 for potassium or sodium and 2 for calcium, magnesium or zinc). Therefore, the CBVS values are characteristic for the element type and it is relatively easy to compare the calculated CBVS value for a cation in a given crystal structure with a table containing the expected values for the different atom types. We assume that the ligand elements can be assigned unambiguously from chemical considerations, although it should be noted that

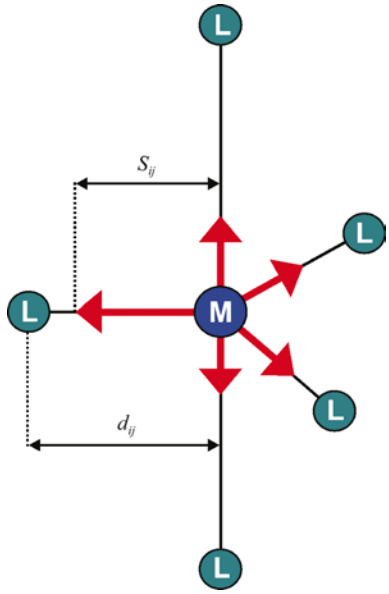

Figure 2
The VECSUM for the case of trigonal bipyramidal geometry. Short distances result in long vectors and long distances in short ones, proportional to the strengths of the interactions.

water molecules that are partially substituted by halides may be difficult to identify except possibly by using anomalous dispersion (Dauter & Dauter, 2001). By assuming octahedral coordination and equal bond distances to oxygen, we can calculate M—O distances from the bond-valence parameters and so estimate the CBVS values that the more common cations in protein structures would give. The results are given in Table 2, which also shows good general agreement between the estimated M—O distances and their average values from small-molecule structures. It can be seen that although the CBVS values cover an appreciable range and so should provide useful discriminatory power, there are cases (*e.g.* Mg^{2+} and Zn^{2+}) that will be difficult to distinguish, although a lower coordination number than six is much more common for Zn^{2+} than Mg^{2+} and so would provide an alternative test.

5. The VECSUM concept

A vital premise for a successful calculation of the correct values for V_i or CBVS is the completeness of the coordination sphere of the metal ion. If the coordination sphere is incomplete, the summed number of interactions is too low and so is the resulting calculated valency of the ion. In order to assess the completeness of a coordination sphere, we introduce the concept of the vector sum of bond valences (VECSUM). This involves the assumption that vectors drawn from a central atom to its ligands, with lengths proportional to the corresponding bond valences (multiplied by the ligand occupancies if these are not unity), will sum to approximately zero. This is consistent with simple bonding models both for covalent and for ionic bonding: a more strongly bonded ligand tends to subtend a greater solid angle at the central atom. This assumption will not be valid if the central atom has an unsymmetrical electron distribution caused by the presence of a stereochemically active lone pair of electrons, *e.g.* in the case of Pb^{2+} or Tl^{+}, but such ions are rare in protein structures; square-pyramidal five-coordinated Cu^{2+} could also be an exception.

As shown in Fig. 2, the VECSUM is the sum over all vectors $\mathbf{f}_{ij}$ within the coordination sphere of one atom, normalized by the overall valence V_i of the central atom.

$$s_{ij} = v_{ij}p_j, \tag{4}$$

$$\mathbf{r}_{ij} = \frac{\mathbf{d}_{ij}}{d_{ij}}, \tag{5}$$

$$\mathbf{f}_{ij} = s_{ij}\mathbf{r}_{ij}, \tag{6}$$

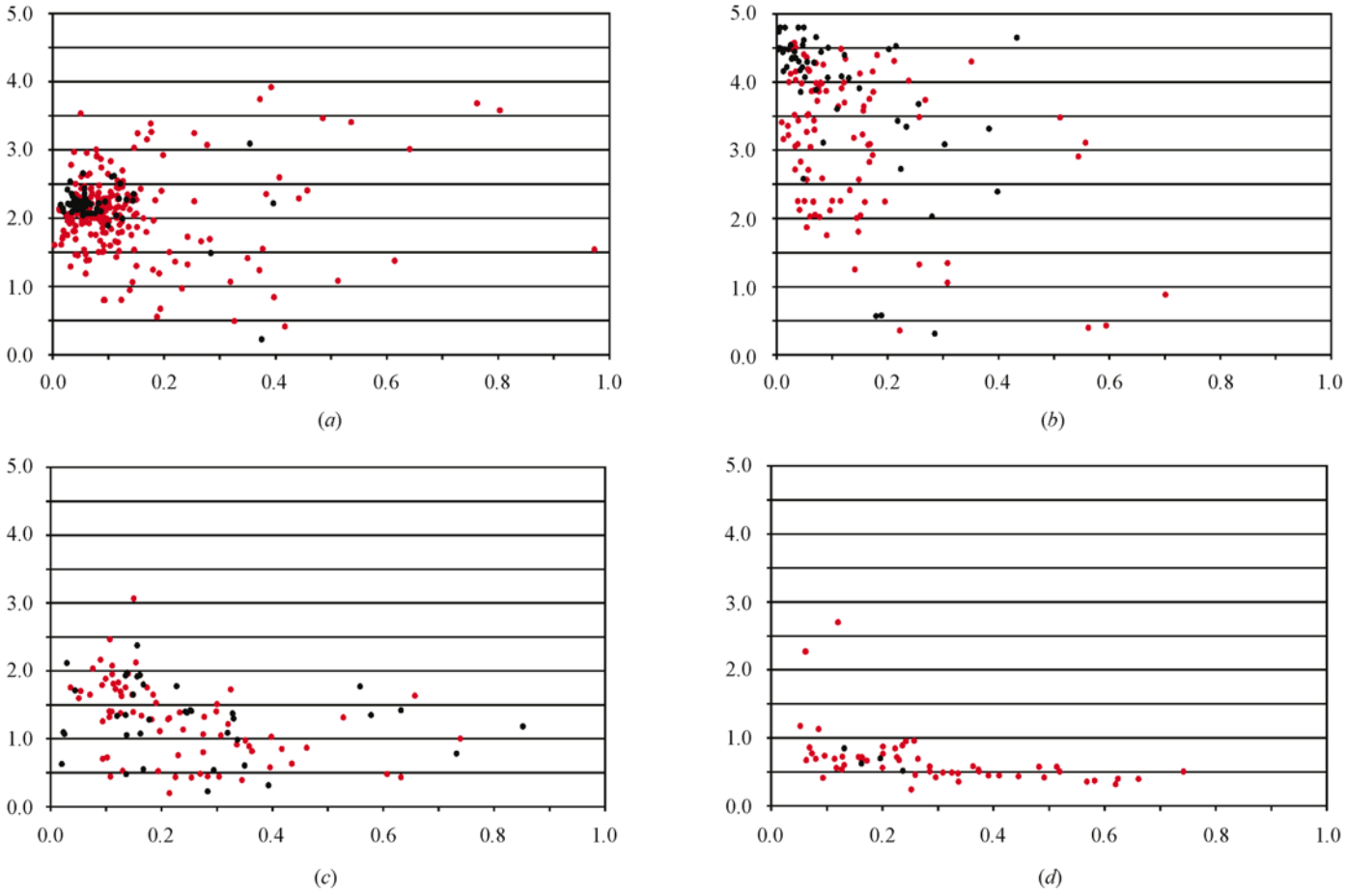

Figure 3
CBVS distributions (vertical axes) for (*a*) Ca^{2+}, (*b*) Mg^{2+}, (*c*) Na^{+} and (*d*) K^{+} plotted against the VECSUM (horizontal axes); the red dots correspond to structures with a resolution between 1.8 and 1.5 Å and the black dots correspond to structures with a resoultion better than 1.5 Å.

$$\mathrm{VECSUM}_i = \frac{\left|\sum_j \mathbf{f}_{ij}\right|}{V_i}. \qquad (7)$$

Here, $\mathbf{r}_{ij}$ is a vector of unit length along the bond between i and j, d_{ij} is the bond distance and $\mathbf{d}_{ij}$ is a vector corresponding to the bond. For practical reasons we use again the d_0 values for Ca$-X$ distances to calculate the normalizing factor V_i.

The condition that VECSUM should be close to zero is a necessary but not sufficient condition for a complete coordination sphere; for example, the lack of two ligands from geometrically opposite sites of the central atom would also result in a zero VECSUM.

6. Derivation of the expected CBVS values

In order to derive a list of the expected CBVS values, we calculated both VECSUM and CBVS according to (3) and (7) for structures from the PDB. Again, we chose only structures with a resolution of 1.8 Å or better and $R_1 \leq 20\%$ containing fully occupied metal ions. Figs. 3(*a*)–3(*d*) show the calculated CBVS values against the corresponding VECSUM for the metals calcium, magnesium, sodium and potassium. One can see clearly that the CBVS values are somewhat untrustworthy for a VECSUM larger than 0.2. This is not surprising, since an incomplete coordination sphere is incompatible with the bond-valence concept. In contrast, for very small VECSUM values, at least for structures corresponding to high-resolution data (1.5 Å and better; black dots), the points tend to cluster about a particular CBVS value, as can be seen as well in Fig. 4.

Calcium shows the sharpest peak in the plot (Fig. 4, blue line) and as expected the CBVS value is close to 2. The CBVS value for magnesium (red line) lies between 3.5 and 4.5 and for potassium (yellow line) it is less than 1. Sodium (green line) shows no clear tendency, possibly because sodium is frequently the 'cation of choice' when the true identity is unknown and in some cases a water molecule may have been inadvertently assigned as sodium.

Particularly interesting is the graph for magnesium (red). In addition to the broad maximum between 3.5 and 4.5 there is another clear maximum at about 2.2, which could well correspond to some calcium ions that were erroneously refined as magnesium.

Analogous calculations have been performed for the ions I^-, Br^-, Cl^- and NH_4^+, as well as for water molecules. The results are shown in Fig. 5. Calculated CBVS values for the three halides are close to zero for I^-, around 0.05 for Br^- and between 0 and 0.44 for Cl^-. Water and ammonium both show values between 0.45 and 0.55 and so cannot be distinguished by this method.

7. Reliability, convenience and benefit of the method

To avoid bias in the application of the bond-valence method, it is important that the distances to the metal ions were not restrained in the refinement; this could happen unintentionally if anti-bumping restraints are applied uncritically. Since the bond valences are sensitive to small errors in the bond distances, the accuracy of the structures plays an important role. This requires a good data-to-parameter ratio, which means that fairly high resolution is required. On the basis of the work reported here, the minimal required resolution seems to lie somewhere between 1.5 and 1.8 Å. However, the distinction between the isoelectronic K^+ and Ca^{2+} is relatively clear even at the lower end of this resolution range and similarly Na^+ can be distinguished well from the isoelectronic Mg^{2+}.

The CBVS concept is rather a simple one and the calculations can be performed easily by hand. All equations and the expected CBVS values have also been included into the latest version of *SHELXPRO* (Sheldrick, 2002). The program calculates the VECSUM and CBVS values for all metal atoms, assuming atoms within a radius of 3.5 Å around the metal to be ligands. The question posed by the title of this paper can be answered with a qualified 'yes'; the CBVS approach provides a numerical basis for cation identification, suitable for auto-

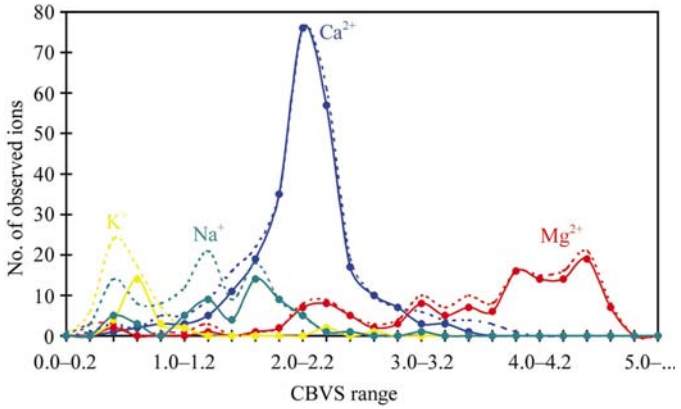

Figure 4
CBVS distributions for Ca^{2+}, Mg^{2+}, Na^+ and K^+. The number of observed ions (vertical axis) is plotted against the CBVS range (horizontal axis). The dotted lines represent all structures selected for a particular cation, while the solid lines refer only to structures with a VECSUM less than 0.2.

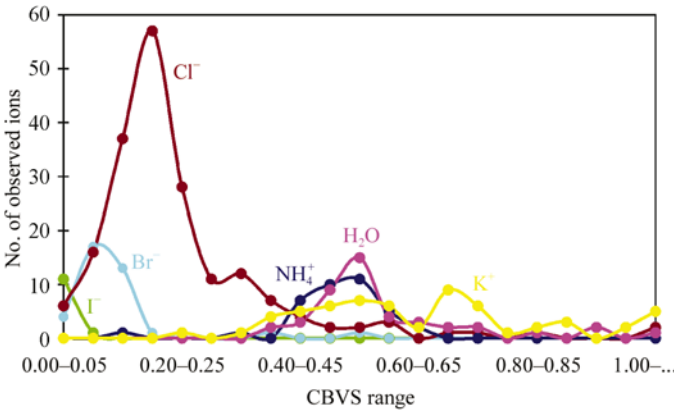

Figure 5
CBVS distributions for I^-, Br^-, Cl^-, NH_4^+, water and K^+. As in Fig. 3, the number of observed ions (vertical axis) is plotted against the CBVS range on the horizontal axis.

mated structure validation. However, other evidence may also need to be taken into account; the crystallographer should always have the last word!

References

Allen, F. H., Davies, J. A., Galloy, J. J., Johnson, O., Kennard, O., Macrae, C. F., Mitchell, E. M., Mitchell, G. F., Smith, J. M. & Watson, D. G. (1991). *J. Chem. Inf. Comput. Sci.* **31**, 187–204.

Allen, F. H. & Kennard, O. (1993). *Chem. Des. Autom. News*, **8**, 31–37.

Berman, H. W., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.

Brese, N. E. & O'Keeffe, M. (1991). *Acta Cryst.* B**47**, 192–197.

Brown, I. D. (1977). *Acta Cryst.* B**33**, 1305–1310.

Brown, I. D. (1992). *Acta Cryst.* B**48**, 553–572.

Brown, I. D. & Altermatt, D. (1985). *Acta Cryst.* B**41**, 244–247.

Brown, I. D. & Shannon, R. D. (1973). *Acta Cryst.* A**29**, 266–282.

Dauter, Z. & Dauter, M. (2001). *Structure*, **9**, R21–R26.

García-Rodríguez, L., Rute-Perez, A., Pinero, J. R. & Gonzalez-Silgo, C. (2000). *Acta Cryst.* B**56**, 565–569.

Glusker, J. P. (1991). *Adv. Protein Chem.* **42**, 1–76.

Harding, M. M. (1999). *Acta Cryst.* D**55**, 1432–1443.

Harding, M. M. (2002). *Acta Cryst.* D**58**, 872–874.

Hawthorne, F. C. (1994). *Acta Cryst.* B**50**, 481–510.

Nayal, M. & Di Cera, E. (1994). *Proc. Natl Acad. Sci. USA*, **91**, 817–821.

Nayal, M. & Di Cera, E. (1996). *J. Mol. Biol.* **256**, 228–234.

O'Keeffe, M. (1989). *Struct. Bonding*, **71**, 161–191.

O'Keeffe, M. & Brese, N. E. (1991). *J. Am. Chem. Soc.* **113**, 3226–3229.

Palenik, G. J. (1997). *Inorg. Chem.* **36**, 122.

Pauling, L. (1929). *J. Am. Chem. Soc.* **51**, 1010–1026.

Pauling, L. (1947). *J. Am. Chem. Soc.* **69**, 542–553.

Sheldrick, G. M. (2001). *XP*, v. 6.12. Bruker AXS Inc., Madison, Wisconsin, USA.

Sheldrick, G. M. (2002). *SHELXPRO*. Universität Göttingen, Germany.

Shields, G. P., Raithby, P. R., Allen, F. R. & Motherwell, W. D. S. (2000). *Acta Cryst.* B**56**, 455–465.

Süsse, P. & Tilmann, B. (1987). *Z. Kristallogr.* **179**, 323–334.