Acta Crystallographica Section B
Structural Science
ISSN 0108-7681

Application of graph theory to detect disconnected structures in a crystallographic database: copper oxide perovskites as a case study

Yuri Kotliarov[a,b]* and Shuichi Iwata[a]

[a]Research into Artifacts, Center for Engineering, University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 156, Japan, and [b]Institute of Inorganic Chemistry, Siberian Branch, Russian Academy of Sciences, 3 Lavrent'ev Avenue, Novosibirsk 630090, Russia

Correspondence e-mail:
yuri@race.u-tokyo.ac.jp

Received 2 February 2000
Accepted 4 September 2000

Every crystal structure can be described as a graph with atoms as vertices and bonds as edges. Although such a graph loses the space arrangement of atoms and symmetry elements, it can mathematically represent the connectivity between atoms. This topological approach was used to develop a new method for detecting disconnected structures, in which individual atoms or structural fragments are located too far from each other, forming impossibly large gaps. Approximately 2300 perovskite-related crystal structures have been extracted from the Inorganic Crystal Structure Database (in 1999) and the maximum disconnecting distances, and the relations between them and the ionic radii of elements, have been analysed. Several disconnected structures, which are erroneous by our definition, have been revealed. Conventional tests for crystallographic data checking did not detect those entries.

1. Introduction

The description of a crystal structure is characterized by a specific set of variables which are related to each other by rules that must satisfy physical and chemical laws. Without full and proper data testing and with even one incorrect parameter, the whole crystal structure may differ significantly from the correct one. For instance, simple misprinting of unit-cell parameters or atomic coordinates or loss of several atoms can cause individual atoms or structural fragments to be separated and the crystal structure to be disjointed (Fig. 1).

The new method presented in this paper will help to reveal such erroneous structures in the crystallographic database. The procedure is automated and does not require structure visualization (however, the latter can be used for final confirmation). All the calculations are by computer program written in ANSI C with the help of the SgInfo crystallographic library (Grosse-Kunstleve, 1999). Test data were retrieved from Inorganic Crystal Structure Database (release 1999/1) in CRYSTIN format.

2. Application of graph theory

Connectivity in a system is a fundamental factor of graph theory. Crystal structure, for its part, is represented as a set of atoms connected by bonds. Therefore, it can be described as a graph with atoms as vertices and bonds as edges, and can be called a *structure graph*. It should be noted that by the term '*bonds*' we do not mean the real chemical bonds, but the topological closeness of atoms in the space of crystal structure. A similar approach has been used for the study of three-dimensional framework structures such as zeolites (Grosse-Kunstleve *et al.*, 1996).

To make the graph finite, only atoms inside one or several unit cells may be included in the graph. Although such a graph loses the spatial arrangement of atoms and symmetry elements, it still mathematically describes the relationships between the atoms in terms of their connections.

A distance matrix **D** (edge-weighted in terms of graph theory) can be built after calculation of the distances between all the atoms in a graph. It is a symmetric $N \times N$ matrix, where N is the number of atoms included in the graph. Each element of the matrix is defined as d_{ij} if $i \neq j$, or as zero, if $i = j$, where d_{ij} is the shortest distance between the corresponding atoms.

Two atoms are considered as being connected if the distance d_{ij} between them is shorter than some critical distance d_c, which can be defined in advance. (Possible criteria for choosing the proper critical distance will be discussed later.) The connectivity matrix **C** (or adjacency matrix in terms of graph theory), which contains information about the internal connectivity of atoms in the structure graph, can be constructed from the distance matrix. The elements of **C** are defined as

$$\begin{aligned} c_{ij} &= 1, \text{ if } i \neq j \text{ and } d_{ij} < d_c \\ c_{ij} &= 0, \text{ if } i = j \text{ or } d_{ij} > d_c. \end{aligned}$$

As an example, the crystal structure of caesium chloride, CsCl, and its distance and connectivity ($d_c = 4.2$ Å) matrices (for one unit cell including the atoms on the faces) are illustrated in Fig. 2.

Accordingly, a non-zero entry appears in the connectivity matrix if atoms i and j are connected.

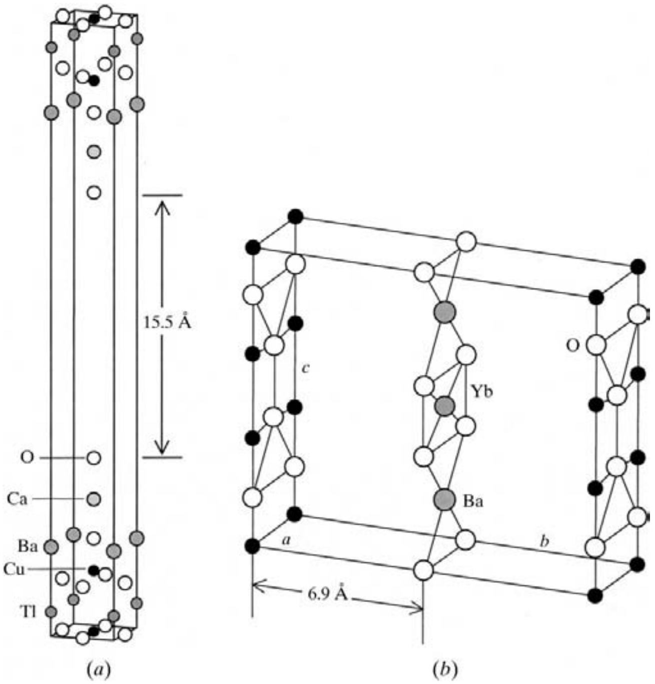
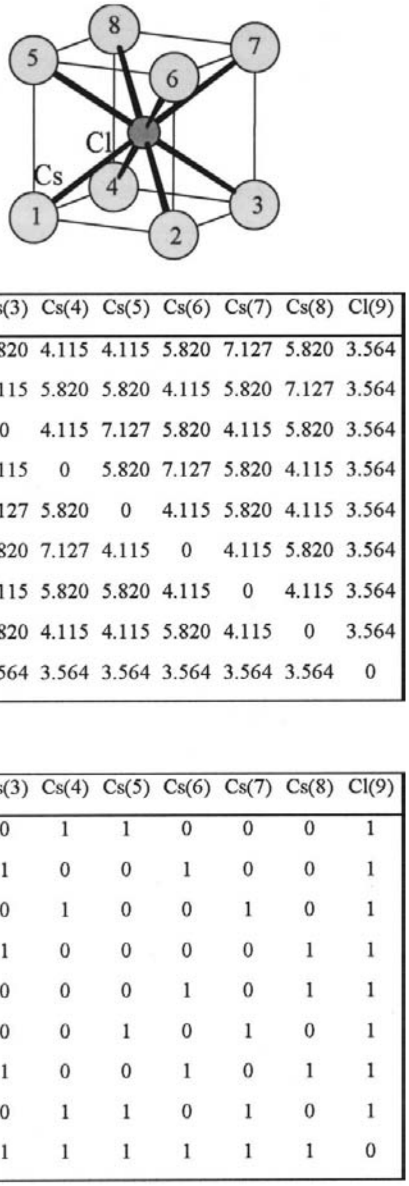


Figure 1
Disconnected crystal structures detected in ICSD. (*a*) $Tl_{1.5}Ca_2Ba_2Cu_{2.10}O_{8.8}$ (collection code 71342); (*b*) $YbBa_2Cu_3O_{6.952}$ (collection code 67645).

The structure graph is called *connected*, if a path exists between every pair of atoms. The connectivity criterion d_c can be chosen in such a way that a disconnected graph must mean a disjointed invalid structure.

It is usually enough to include only the atoms of a single unit cell, including all those located on its faces, for the determination of the erroneous structures. However, we found, that in some structures atoms, which are disconnected in a single unit cell, are connected *via* atoms in the adjacent cells. On the other hand, connected atoms in a single unit cell can be disconnected from atoms in an adjacent cell. Therefore, it is better if the structure graph contains atoms of eight adjacent cells ($2a \times 2b \times 2c$), excluding the atoms located on the external faces ($-1 < x < 1$, $-1 < y < 1$, $-1 < z < 1$).



D	Cs(1)	Cs(2)	Cs(3)	Cs(4)	Cs(5)	Cs(6)	Cs(7)	Cs(8)	Cl(9)
Cs(1)	0	4.115	5.820	4.115	4.115	5.820	7.127	5.820	3.564
Cs(2)	4.115	0	4.115	5.820	5.820	4.115	5.820	7.127	3.564
Cs(3)	5.820	4.115	0	4.115	7.127	5.820	4.115	5.820	3.564
Cs(4)	4.115	5.820	4.115	0	5.820	7.127	5.820	4.115	3.564
Cs(5)	4.115	5.820	7.127	5.820	0	4.115	5.820	4.115	3.564
Cs(6)	5.820	4.115	5.820	7.127	4.115	0	4.115	5.820	3.564
Cs(7)	7.127	5.820	4.115	5.820	5.820	4.115	0	4.115	3.564
Cs(8)	5.820	7.127	5.820	4.115	4.115	5.820	4.115	0	3.564
Cl(9)	3.564	3.564	3.564	3.564	3.564	3.564	3.564	3.564	0

C	Cs(1)	Cs(2)	Cs(3)	Cs(4)	Cs(5)	Cs(6)	Cs(7)	Cs(8)	Cl(9)
Cs(1)	0	1	0	1	1	0	0	0	1
Cs(2)	1	0	1	0	0	1	0	0	1
Cs(3)	0	1	0	1	0	0	1	0	1
Cs(4)	1	0	1	0	0	0	0	1	1
Cs(5)	1	0	0	0	0	1	0	1	1
Cs(6)	0	1	0	0	1	0	1	0	1
Cs(7)	0	0	1	0	0	1	0	1	1
Cs(8)	0	0	0	1	1	0	1	0	1
Cl(9)	1	1	1	1	1	1	1	1	0

Figure 2
Crystal structure, distance matrix (**D**) and connectivity matrix (**C**) for CsCl, $d_c = 4.2$ Å.

The following algorithm, related to DFS (depth first search), is used for the connectivity test:

(i) The structure graph is constructed as described above and its distance matrix **D** is calculated.

(ii) The connectivity matrix **C** is derived from the distance matrix **D** using the given connectivity criterion d_c.

(iii) Two virtual arrays for atoms are created, all atoms included in the graph first being put in array 1.

(iv) The first atom and all atoms connected to it (corresponding to non-zero elements of **C**) are then moved to array 2.

(v) The atoms from array 1 that are connected to the next atom from array 2 are moved to array 2.

(vi) Steps (iv) and (v) are repeated for all atoms from array 2 until no atoms left in array 1 (*i.e.* graph is connected) or no further atoms from array 1 can be moved to array 2 (*i.e.* graph is disconnected).

The connectivity test diagram in our example structure CsCl is shown in Fig. 3. (For simplicity only one unit cell is used to build the graph.) Obviously, we could complete the test in one cycle if we took the Cl atom as the first atom. However, in the case of more complex structures the gain in time is not large enough to justify the algorithmic complexity. For instance, the analysis of the graph with about 600 atoms takes less than a second on a modern computer.

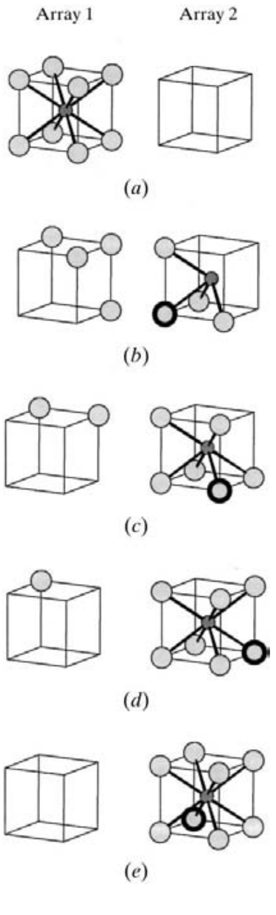


Figure 3
Connectivity test diagram for CsCl: (*a*) before the test; (*b*)–(*d*) first three cycles – tested atom is outlined; (*e*) final cycle – no atom is left; graph is connected.

3. Choice of critical distance–maximum disconnecting distance–maximum disconnecting ratio

A proper choice of the critical distance d_c is the key point of the new approach described above. Disconnected erroneous structures will be hidden in the case of a critical distance which is too long, while for too short a distance even correct structures will be qualified as erroneous. Several ways of defining d_c can be proposed, such as using an absolute distance (or range of distances) on the basis of the statistical distribution of distances, the ratio of the distance to the sum of the ionic radii of the atoms, possible coordination numbers for the particular element and so on. The important task of this work was to find the distribution of the smallest value of d_c that leads to a connected graph.

We define the *Maximum Disconnecting Distance* (*MDD*) as the smallest value of d_c that leads to a connected structure.

In other words, if we sort in ascending order all distances in the structure graph and then check the graph connectivity using those distances consecutively as d_c (starting from the shortest distances), at some value of d_c the graph becomes connected. That distance is the MDD.

From the chemical point of view, the interatomic distances are less meaningful if the size of the connected atoms is not taken into account. The ratio (*R*) of the distance to the sum of the ionic radii of the elements can be used as an alternative to the interatomic distances for the connectivity test. This ratio can be considered as the distance normalized to the size of the atoms. The *Maximum Disconnecting Ratio* (*MDR*) can be calculated by analogy to the MDD using normalized distances rather than real distances as the components for the distance matrix **D**. It should be emphasized that MDR is not exactly the ratio of MDD to the sum of ionic radii because it is calculated on the basis of a different distance matrix.

In this work, the structure graphs have been constructed for the 2294 crystal structures (perovskite-related copper oxide compounds have been chosen for this study) and their MDDs and MDRs have been calculated. The ionic radii of the elements (with regards to their oxidation state) have been taken from the Appendix in the ICSD's User Manual (1995).

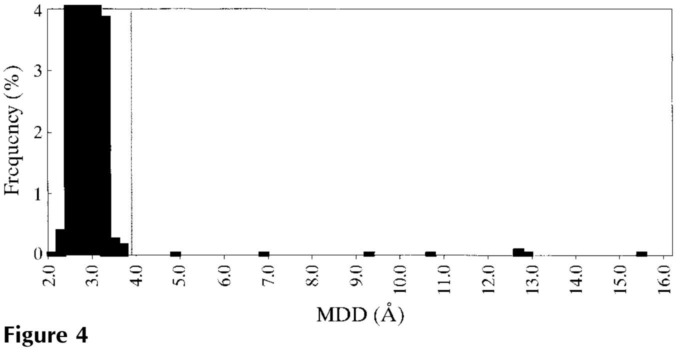


Figure 4
The distribution of maximum disconnecting distances (MDD) in the perovskite-related structures.

Table 1
Erroneous entries with the largest MDD and MDR.

ICSD ID	MDD (Å)	Corresponding atoms (with oxidation state)	Sum of IR (Å)	MDR	Corresponding atoms (with oxidation state)	Sum of IR (Å)	Compound formula
63428	4.957	$Cu^{2+}-O^{2-}$	1.83	2.091	$Ba^{2+}-O^{2-}$	2.55	$NdBa_2Cu_3O_{6.85}$
67645†	6.933	$Cu^{2+}-O^{2-}$	1.83	2.821	$Ba^{2+}-O^{2-}$	2.55	$YbBa_2Cu_3O_{6.952}$
69238	9.351	$Sr^{2+}-O^{2-}$ ($O^{2-}-O^{2-}$)	2.33 (2.42)	3.864	$O^{2-}-O^{2-}$	2.42	$Bi_{1.36}Pb_{0.18}Sr_{1.4}Ca_2Cu_{2.94}O_{8.7}$
71341	10.730	$O^{2-}-O^{2-}$	2.42	4.434	$O^{2-}-O^{2-}$	2.42	$Tl_{1.04}Ca_{0.96}BaCa_{1.54}Cu_{2.24}O_{6.96}$
66305	12.75	$O^{2-}-O^{2-}$	2.42	5.269	$O^{2-}-O^{2-}$	2.42	$Tl_{2.16}Ca_{0.72}Ba_2Cu_2O_8$
66307	12.793	$O^{2-}-O^{2-}$	2.42	5.286	$O^{2-}-O^{2-}$	2.42	$Tl_{1.94}Ca_{0.84}Ba_2Cu_2O_8$
66306	12.835	$O^{2-}-O^{2-}$	2.42	5.304	$O^{2-}-O^{2-}$	2.42	$Tl_{1.95}Ca_{0.8}Ba_2Cu_2O_8$
71342†	15.534	$O^{2-}-O^{2-}$	2.42	6.419	$O^{2-}-O^{2-}$	2.42	$Tl_{1.5}Ca_2Ba_2Cu_{2.10}O_{8.8}$

† Structures illustrated in Fig. 1.

The distribution histograms of MDDs and MDRs for these compounds are shown in Figs. 4 and 5, respectively. The MDDs of 99.6% of the oxides studied lie between 2 and 3.8 Å, while in the case of MDR 99.7% lie between 0.9 and 2.2.

The eight structures with the largest MDDs and MDRs, which are separated from the normal distribution, have been analysed individually and erroneous entries are summarized in Table 1. Two of these structures are illustrated in Fig. 1. Entry 63428 ($NdBa_2Cu_3O_{6.85}$) is not actually rejected by the MDR criterion of 2.2. However, visual analysis showed that this entry has a *b* parameter which is too large: 9.9 Å rather than ~3.8 Å as in the related $YBa_2Cu_3O_7$ (well known as the 123 phase). The ratio of the distance to the sum of the ionic radii is smaller for the $Ba^{2+}-O^{2-}$ connection, although the actual $Cu^{2+}-O^{2-}$ distance is shorter because of the significant difference between the sizes of Cu^{2+} (IR = 0.62 Å) and Ba^{2+} (IR = 1.34 Å).

The commonly occurring errors among the remaining seven structures are one of the unit-cell parameters being too small or an incorrect atomic coordinate. (In 67645, for example, the error seems to be simple mistyping in the parameter b = 13.8658 Å rather than 3.8658 Å.)

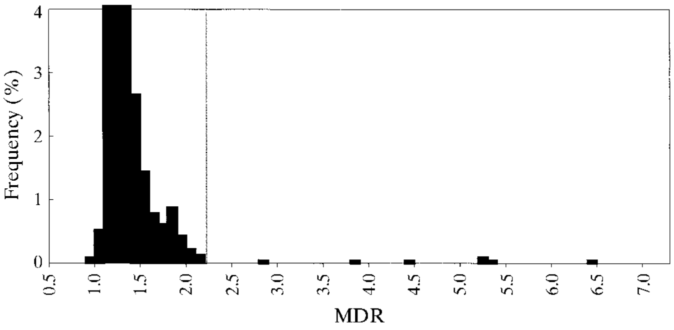


Figure 5
The distribution of MDR in the perovskite-related structures.

Finally, the eight erroneous structures have been checked by the program for the conventional crystallographic test, which did not reveal the error. This fact shows the importance of developing new procedures for checking crystallographic data.

4. Conclusions

In this paper we propose a new topological approach to determine disconnected erroneous crystal structures with large gaps between structural fragments using graph theory and connectivity criteria. Analysis of the distribution of the maximum disconnecting distances (MDD) and the maximum disconnecting ratios of distances to the sum of ionic radii (MDR) illustrated that the critical values for erroneous and suspicious structures can be clearly chosen. The ICSD database was used to introduce this approach, and, as a result, eight erroneous entries (from the 2294 analysed entries) have been identified.

Finally, it should be mentioned that the value of the critical distance d_c can be varied depending on the information needed from the structure graph. Coordination polyhedra and other structural fragments can be extracted and systematized, and can be used for data quality control.

References

Grosse-Kunstleve, R. W. (1999). *Acta Cryst.* A**55**, 383–395 (see also http://www.kristall.ethz.ch/LFK/software/sginfo/).

Grosse-Kunstleve, R. W., Brunner, G. O. & Sloane, N. J. A. (1996). *Acta Cryst.* A**52**, 879–889.

Inorganic Crystal Structure Database (1995). *User's Manual.* Fachinformationszentrum Fiz Karlshruhe, Germany.