情報学基礎 46－5
（1997. 7. 24）

材料設計のための結晶データマイニングへのアプローチ

コトリャローフ　ユーリ[1,2]、岩田修一[1]

[1]東京大学 人工物工学研究センター 〒153 東京都目黒区駒場4-6-1
[2]RASシベリア支部無機化学研究所 ロシア、630090ノボシビルスク、ラブレンチエフ通り3

材料における化学組成、結晶構造、物性それぞれの関係を体系的な分類に基づいて総合的に理解することは、材料設計の基本である。構造や物性値を数多く格納したデータベースは、上記の関係に対する新しい知識を得るための強力な道具となる。ここでは、データベースを利用して無機材料における結晶構造と構造要素のデータマイニングを行う上での問題点を、結晶学的見地から述べる。層状銅酸化物超伝導材料における、新たな探索方法の開発と、複雑で未だ発見されていない規則性を見い出すことを目的として、二次元積層構造に関する分類を行った。

Approach to crystallographic data mining for the material design

Kotliarov Iouri[1,2], Iwata Shuichi[1]

[1]Research into Artifacts, Center for Engineering, the University of Tokyo
4-6-1, Komaba, Meguro-ku, Tokyo 156
[2]Institute of Inorganic Chemistry, Siberian Branch of RAS, Lavrent'ev pr. 3, Novosibirsk, 630090 Russia

Global understanding of the relations among chemical composition, crystal structure and properties of materials on the basis of suitable classification is a root concept of material design. Large databases accumulating structural and physical properties of materials give the powerful tool for new knowledge discovery. In this paper the problems of data mining of crystal structures and structural primitives in inorganic materials are described from crystallographic viewpoint. The approach, allowing representation of crystal structure as a sequence of stacking 2-dimension layers, is being considered for the development of advanced retrieval procedure, searching more complex and hidden regularities in the layered cuprate superconducting materials.

Introduction

This is a worldwide-accepted statement that the majority of physical properties of solid materials correlate with their crystal structure. Nevertheless it is seems to be very difficult not only to find appropriate correlation expression, but also to understand the phenomenon of such dependencies. Moreover, the principles of the crystal structure standardization, that are based exclusively on lattice symmetry, significantly limit the "data mining" process. There is no doubt that symmetry group theory is a great instrument for crystal structure identification, refinement and analysis. It unambiguously explains a great deal of experimental results and physical phenomena such as piesoelectrisity and ferroelectrisity. However, the real world is more complex. Defects, superstructures, modulations and other kinds of distortion, in some cases, space group choice ambiguity, may result in symmetrically different structures, even if they are close in element composition, atomic arrangement and physical properties. Moreover, a minor change in the position of the atom in a crystal structure can often reduce its symmetry, leading to a new mathematical description and a "new" crystal structure. The task becomes rather difficult, when we are going to find compounds with closely related structures, such as homologous series members, solid solutions, derivative structures, etc. The diagram on Figure 1 represents the situation when, due to the difficulty in query formulation, result is rather far from the desirable one. Intersection of these virtual regions is partially covers user's need, but the real results often contain a lot of mismatches.

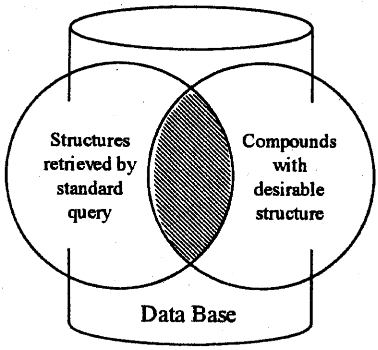


Figure 1. Limitation of standard data search procedure in crystal structural database by symmetrical approach.

We believe the main reason to base in the absence of close connection between crystallographers and people, who design the materials with properties given in advance. One solution of this problem is to take into account not only characteristics of symmetry, but also atoms, their arrangement and interrelation. The atomic environment approach (AE) has been proposed by P.Villars[1)], and it woks well mostly for binary and ternary intermetallic compounds. However, in more complex cases, with mixed atomic environment types, classification becomes very difficult. This paper proposes a new approach, which can be employed for database mining. It bases on the analysis of layered copper oxides, the family with great variety of structures, many of that belongs to compounds with high-temperature superconducting properties.

Analysis of structures of layers in high-temperature and related copper-oxide compounds.

The presence of CuO2 conducting sheets is indispensable for manifestation of superconductivity in copper oxide compounds. Their overall structures are usually described as stacks of 2-dimensional sheets and structural blocks along direction perpendicular to the sheets[2)]. All these compounds comprise similar structural units. Therefore describing their structures in a coherent manner is useful for clarifying their structural features and relationships among various crystal structures. Such structural representation also serves as a tool to devise strategies for search of new compounds that may exhibit interesting electronic and magnetic properties.

Santoro with colleagues[3] has proposed a convenient method of expressing the (ideal) structure of superconducting copper oxides by regarding them to be composed of perovskite and rack-salt-type sheets. This method used for the first systematic analysis of layered structures of copper-oxide superconductors from the viewpoint of layers patterns and stoichiometry.

Layered structure can be described in different ways and at different level of detailing. Despite of their great variety perovskite-like structures allow their description with the help of small set of atomic networks perpendicular to (pseudo)-four fold axis[4]. The composition may be expressed by the chemical formula corresponding to the stoichiometry of each layer, for example A, AX, B_2, BX_2, etc. For the sake of simplicity, generally atomic positions are redefined as A - large size cation position, B - small or middle cation position, X - anion position. Subscript indices represent shift of cell relatively to coordinate origin with translation vector: o=(0,0); c=(1/2,1/2); x=(1/2,0); y=(0,1/2). Square brackets, which are used in some formula, indicate the layers that are contained in one unit cell of the structure. This notation, which is more close to Cava[5], has been chosen because of its simplicity for automation. In other works notations are a little bit different. Furthermore, each pattern can be presented by digital code. Although this code does not describe atomic arrangement, it significantly simplifies computer analysis of the patterns.

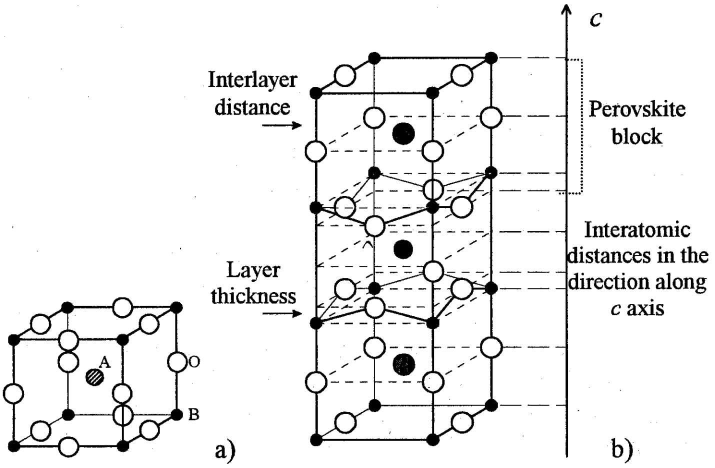


Figure 2. Schematic representation of crystal structures of perovskite (a) and superconducting $YBa_2Cu_3O_{7-x}$ (Y123 phase) with perovskite blocks (b).

For example, if we apply this symbolism to the structure of perovskite $CaTiO_3$ (Figure 2a), we obtain the sequence $[(BX_2)_o(AX)_c]$, where A=Ca, B=Ti and X=O; or more simple - 1o5c. For Y123 structure it is $[(BX_2)_o(AX)_c(BX_2)_o(A)_c(BX_2)_o(AX)_c(BX_2)_o]$ or 1o5c1o8c1o5c1o. As one can see, perovskite block in Y123 phase can be identified easy with the help of the sequence of layers. If necessary (for example, for the analysis of formal electronic state of elements in layers), letters A, B and X can be replaced by real elements symbols as $[(TiO_2)_o(CaO)_c]$ for the perovskite structure.

Taking into account the common principles of such structures, steric factors and limitations of electrostatic character (improper charge balance, according to Pauling's electrovalence rule) practically all possible combinations of considered networks can be defined in advance (Figure 3).

As a result of the present work, a set of computer programs for data mining, for building crystal lattice and for sequencing of layers have been developed[6].

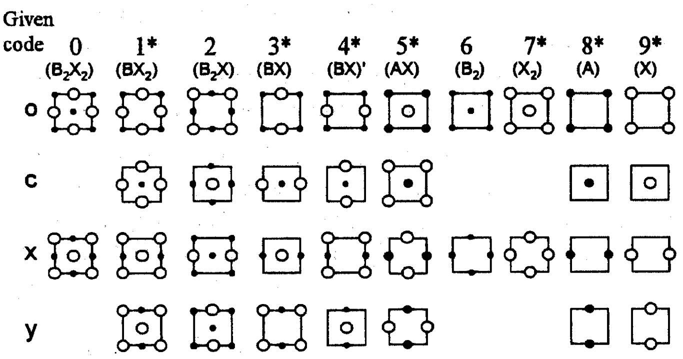


Figure 3. Possible atomic networks for perovskite related compounds. A - large size cation position; B - small or middle cation position; X - anion position.
* - existent types of layers.
Shift from the origin (o): c- center, x- x axis direction, y - y axis direction.

As Figure 2b shows, structural layers in complex perovskite-related structures are not ideally plain, but they may have some thickness. This may cause the situation, when even we have layered structure, we can not separate layers one from another unambiguously, because layers intersection is possible. To show the real layered character of cuprate structures, interatomic distances in them along characteristic direction (axis *c*) have been calculated on the basis of data extracted from Inorganic Crystal Structure Database (more then 1600 records for copper oxides with wide range of metal elements). In the considered case interatomic distance means a distance between two neighboring points, produces by projection of atoms to *c* axis (z coordinates) as shown in Figure 2b. Layer thickness in this context is a distance between minimum and maximum z coordinates of atoms of the same layer. Interlayer distance in its turn is defined as a distance between two neighboring layers, exactly between maximum z coordinate of lower layer and minimum z coordinate of upper layer. Results of calculation are displayed in Figure 4.

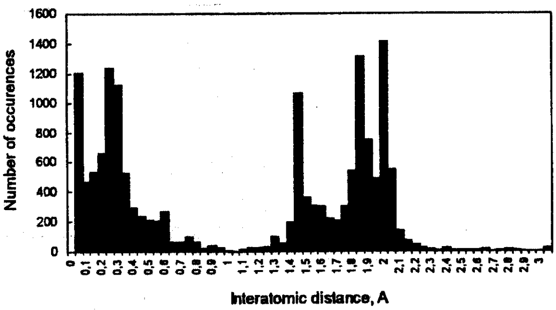


Figure 4. Distribution of interatomic distances along [001] direction (c axis) in cuprates.

On this histogram each column represents number of layers (in this case ideal plain layers are considered) which fall into specific range of interatomic distances. Two clearly separated areas on this graph can be observed, those correspond to layers thickness and real interlayer distances. So, it can be concluded that in the majority of cases there is no ambiguity to separate atoms into two-dimensional layers using characteristic interatomic distances.

Otherwise, in another direction the same distribution of interatomic distances is not observed. Instead, some distances responsible for specific atomic positions and deviation of atoms from ideal model are clearly detected.

Relationship between sequence of layers and 3 dimension primitives in crystal structure

Copper-oxide superconducting structures are most oftenly described by 3-dimensional primitives (perovskite blocks, intermediate blocks and coordination polyhedra) rather than with 2-dimensional layers. They are not only providing convenient structural representation but are also playing important role in several theoretical models. At the same time layers have a close relationships with 3-dimensional primitives. The latter can be easily obtained from sequences of layers. Table 1 shows some examples of such relationships between layers and 3-D primitives in cuprate structures and number of compounds with several 3-D primitives extracted from database.

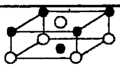
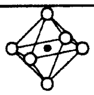
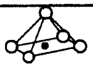
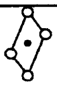
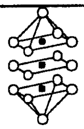
Sequence fragments	5o5c	5o1c5o	8o1c5o	8o1c8o	5o3(4)c5o	5o1'c5o	1o8c1o 8c1o
3D primitives	 Rocksalt shift	 Cu-O octahedron, Perovskite block	 Cu-O square pyramid, defective perovskite	 Cu-O square	 Cu-O square chains	 Cu-O chains	 alternate Cu-O layers
Number of compounds*	Total 298 2 135 3 109 4 54	614	904	45	208	24	Total 877 2 834 3 36 4 7

Table 1. 3D structural primitives as fragments of layers sequence.

* Number of extracted from ICSD compounds, which contain definite structural primitive. For Cu-O and rocksalt alternate layers number of compounds with different number of alternates are listed.

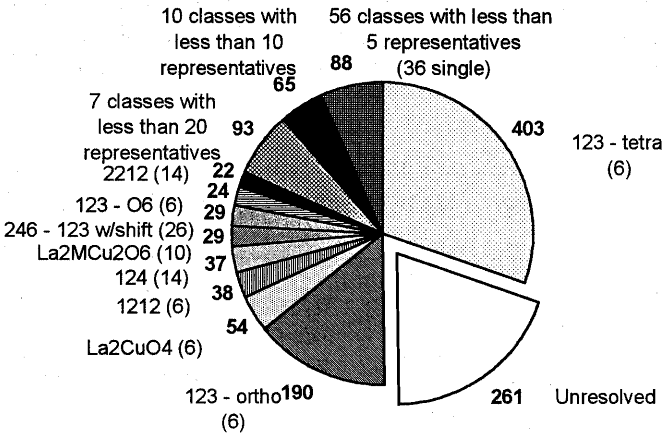
Classification of cuprate structures by sequence of layers

All resulting sequences have been sorted as character strings and as a result 82 different classes have been revealed. Diagram on Figure 5 separates 10 the most representative classes (with more than 20 representatives), others are shown in the groups with less that 20, 10 and 5 representatives. About 19% of structures have been rejected because of their large lattice parameters (cell reduction procedure is still not implemented), or imperfect initial data (positions of some atoms have not been defined), or high disordering.

Among compounds in the last group 36 compounds do not have analogs with the same layers sequence. Several reasons can cause this situation:

- Very rare or new structure.
- Sequences have not been standardized. Structure may be equivalent, but shifted relatively each other in some direction. For example, sequences 1o5c1o8c1o5c1o(1), 1c5o1c8o1c5o1c(2) and 8o1c5o1c5o1c8o(3). Structure (2) shifted relatively (1) in the layers plane to the center (x+1/2, y+1/2); whereas sequence (3) shifted in the direction perpendicular layers planes. To eliminate separation of these structures, normalizing procedure is necessary.
- Layers separation due to large constituent elements, as in case of 5o=8o+9c.
- Various kinds of errors in database.

Several procedures, which will eliminate many of these reasons, are under development now.



Finger 5. The most representative classes of extracted materials combined by their sequences of layers. Number of representatives is given in bold. For accentuated classes number of layers per unit cell (the sequence length) are shown in brackets.

<u>Search for compounds with isomorphic structure</u>

There is also a problem now how to search isomorphic structures for any model structure. It is not a trivial task. Similar structures may appear in different space groups or with a different formula making it difficult to prescreen. Relationships between isomorphic compounds can be often missed without chemical knowledge a priori. On the other hand it is sometimes difficult to separate non-isomorphic structures with similar elemental composition. For example, lets consider 3 compounds from the ICSD (Table 2):

N	Formula	Space group	*a*, Å
1	$Ba_2PrCu_2(Cu_{0.83}Al_{0.17})O_{6.55}$	P4/mmm	3.906
2	$(Cu_{0.95}Zn_{0.05})_2(BaLa_{0.56}Ca_{0.44})(Ca_{0.70}La_{0.30})Cu_{0.78}O_{6.52}$	Pmmm	3.877
3	$(Y_{0.63}Ca_{0.53})Ba_2Hg_{0.63}Cu_{2.21}O_{6.6}$	P4/mmm	3.87

Table 2. Examples of compounds with close element composition but different structure.

All of them have very complex elemental composition with element deficiency and substitution. Is it possible to find some relations between them without 3-D representation? One may find that elemental ratio of these phases is very close to 123 ($YBa_2Cu_3O_{7-?}$):

1. $PrBa_2(Cu_{2.83}Al_{0.17})O_{6.55}$
2. $(Ca_{0.70}La_{0.30})(BaLa_{0.56}Ca_{0.44})[(Cu_{0.95}Zn_{0.05})_2Cu_{0.78}]O_{6.52}$
3. $(Y_{0.63}Ca_{0.53})Ba_2(Hg_{0.63}Cu_{2.21})O_{6.6}$

However, sequences of layers for these compounds:

1. `1o 5c 1o 8c 1o 5c 1o`
2. `1o 5c 1o 8c 1o 5c 1o`
3. `5o 5c 1o 8o 1o 5c 5o`

give us more precise answer. First two compounds have identical layers sequences and belong to Y123 class, whereas the third one is Hg1212 phase.

It seems possible to conclude that this approach to description of layered compounds is very useful for comparing structures, for analysis of their geometrical relationships and for understanding the atomic configuration of complex structures.

Conclusion

In this work the new approach for the definition and isolation of structural primitives has been described and used for the copper-oxide high temperature superconducting and related materials as an example. Similar approach can be applied for other classes of materials by revealing off their common structural features, structure primitives and rules of their organization. As was shown the layer-by-layer model allows not only to retrieve useful information about key elements in crystal structures of currently known superconductors, but also provides with method to find relationships between different structures even if they belong to different structural types and space groups. In addition, the automation of identification of required structures and their dissemination into structural primitives will give us instrument for the advanced data retrieval procedure.

References

---

(1) Villars P., Factors Governing Crystal Structures (Intermetallic Compounds: Principles and Applications), J.Wiley&Sons Ltd. (1994).
(2) Tokura Y., Arima T., Jap.J.Appl.Phys., 29 (1990), 11, pp. 2388-2402.
(3) Santoro A., Beech F., Marezio M., Cava R.J., Physica C, 156 (1988), p. 693-700.
(4) Poole C.P., Datta T., Farach H.A., J.Superconductivity, 2 (1989), p. 369-386.
(5) Cava R.J., Science, 247 (1990), 4943, p. 656-662.
(6) Kotliarov I., Iwata S., to be published