Acta Cryst. (1999). B**55**, 1030–1043

Graph-set analysis of hydrogen-bond patterns: some mathematical concepts†

J. Grell,[a] J. Bernstein[b]* and G. Tinhofer[c]

[a]*Institute of Freshwater Ecology and Inland Fisheries, Department of Ecohydrology, Rudower Chaussee 6a, D-12489 Berlin, Germany,* [b]*Department of Chemistry, Ben-Gurion University of the Negev, PO Box 653, Beer Sheva 84105, Israel, and* [c]*Zentrum Mathematik, Technische Universität München, D-80290 München, Germany.*
E-mail: yoel@bgumail.bgu.ac.il

(*Received 9 January 1999; accepted 20 May 1999*)

Abstract

To provide a foundation for further theoretical and software development of the application of graph sets to patterns of hydrogen bonding and other intermolecular interactions a number of mathematical concepts and tools are defined, developed and demonstrated. Following a review of the basic definitions and uses of graph sets, the directional properties of hydrogen bonds are now included in the treatment. The concepts of a *constructor graph* and *covalent distance* matrix have been developed to aid in the generation of a *qualitative descriptor* for the straightforward, consistent and ultimately automatic (with appropriate software) definition of patterns. An additional mathematical tool, the *arrowed T-labeling*, has been developed to deal with situations in which pattern-forming moieties are located on crystallographic special positions. To demonstrate the utility and various features of these concepts they are applied in detail to two particular structures, polymorphic iminodiacetic acid [*N*-(carboxymethyl)glycine] and *trans*-tetraamminedinitrocobalt(III) acetate. To facilitate the application and use of graph sets many of these developments have already been incorporated into the software of the Cambridge Structural Database, as described in the accompanying paper.

1. Introduction

Since Etter introduced graph-set notation for the characterization and analysis of hydrogen-bond patterns (Etter, 1990; Etter *et al.*, 1990), the method has enjoyed increasing use (Bernstein *et al.*, 1995, and references therein). The recent use of the notation in the trade literature (Anon, 1997) indicates that it is becoming part of the language of structural chemists, in a manner similar to the way that the notation for reaction mechanisms (S_N2, E_1 *etc.*) is part of the language of organic chemists. This mode of thinking is very much in keeping with the philosophical underpinning of Etter's notions about hydrogen-bonding patterns.

Although there certainly was a fundamental mathematical basis for the graph-set notation (Etter rests on Harary, 1967; Merrifield & Simmons, 1989), most of the notions were based on chemical intuition and chemical experience. While there have been improvements (Bernstein *et al.*, 1995) and many still remain to be made, the fact that those original notions still form the basis of the approach is a tribute to Peggy Etter's insight and her deep understanding of hydrogen bonding.

The increased use of the graph-set tools in the chemical community for describing hydrogen-bonding patterns coupled with our desire to incorporate the assignment of graph sets into structural databases (Motherwell *et al.*, 1999) and extend the use beyond taxonomy to correlation, analysis and prediction has led us to attempt to put the method on a firmer mathematical foundation. Such a foundation also serves as a springboard for creating the software for these extended uses. We have recently made some progress in establishing that mathematical foundation (Bernstein *et al.*, 1997), and the software developments and potential uses are described in the following paper (Motherwell *et al.*, 1999).

The purpose of this paper is to present these mathematical concepts in 'layman crystallographer's terms', to help lower some of the barriers that exist between chemists and mathematicians. We will first put the current situation into context by reviewing briefly the graph-set notation, and its application to a polymorphic system, iminodiacetic acid (Boman *et al.*, 1974; Bernstein, 1979), which we treated with the original Etter notions (Bernstein *et al.*, 1990, 1995). We will then introduce and employ the mathematical concepts in the context of the iminodiacetic acid structures and the existing notation. Some of the mathematical formalism, which of course can be developed independent of any specific example, will be presented in §5; the remainder will appear in an appropriate mathematical publication. The ultimate intention here, as stated in the title, is to outline and to present the mathematical concepts underlying the graph-set methods so that they can be understood by the working structural chemist. On the one hand, the same individual can pursue the mathematical definitions, on the other hand, when these are

† Work supported by grant No. I-0333-263.06/93 from the GIF, the German–Israeli Foundation for Scientific Research and Development.

Table 1. *Summary of the hydrogen bonds in the three polymorphs of* ***IMDA***

The acceptor molecule refers to the molecule containing the acceptor atom.

Designation	Hydrogen bond	Location of acceptor molecule	Description of symmetry
IMDA1: space group $P2_1/c$, $Z = 4$			
a	O1—H1···O3	$(x+1, \bar{y}+\frac{1}{2}, z-\frac{1}{2})$	c-glide and translation
b	N—H2···O3	$(\bar{x}-1, y-\frac{1}{2}, \bar{z}+\frac{1}{2})$	2_1-screw
c	N—H3···O4	$(\bar{x}, \bar{y}, \bar{z})$	Inversion
IMDA2: space group $Pbc2_1$, $Z = 8$			
a	O1—H1··· O3	$(\bar{x}+2, y-\frac{1}{2}, z)$	b-glide and translation
a′	O1′—H1′···O3′	$(\bar{x}'+1, y'+\frac{1}{2}, z')$	b-glide and translation
b	N—H3···O4	$(x-1, y, z)$	Translation
b′	N′—H3′···O4′	$(x'+1, y', z')$	Translation
c	N—H2···O4′	$(x'+1, y', z')$	Res. 1 ··· res. 2 connection
c′	N′—H2′···O4	$(x-1, y, z)$	Res. 2 ··· res. 1 connection
IMDA3: space group $P2_1/n$, $Z = 4$			
a	O1—H1···O3	$(\bar{x}+\frac{1}{2}, y-\frac{1}{2}, \bar{z}+\frac{3}{2})$	2_1-screw
b	N—H2···O4	$(x-1, y, z)$	Translation
d	N—H3···O2	$(x+\frac{1}{2}, \bar{y}+\frac{1}{2}, z+\frac{1}{2})$	n-glide

employed in implementing the concepts in the appropriate software the idea is that the mathematical underpinnings will be transparent to the user, if he or she so chooses.

2. Brief review of graph-set notations

2.1. *The polymorphic system of iminodiacetic acid*

Iminodiacetic acid, $C_4H_7NO_4$ (Fig. 1), is known to be at least trimorphic (Boman *et al.*, 1974; Bernstein, 1979). The three known polymorphs are denoted as **IMDA1**, **IMDA2** and **IMDA3**.

In **IMDA1** and **IMDA3** there is one molecule in the asymmetric unit, while in **IMDA2** there are two molecules in the asymmetric unit. Each molecule has three strong hydrogen-bond donors and four potential hydrogen-bond acceptors. In fact, each molecule acts as a donor for three hydrogen bonds and as an acceptor for three. The labeling and crystallographic symmetry elements that generate these hydrogen bonds are summarized in Table 1 and correspondingly labeled in Fig. 2.

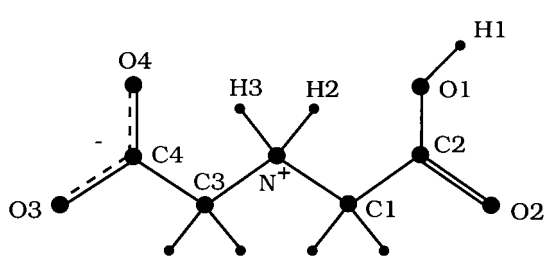

Fig. 1. The iminodiacetic acid molecule, zwitterionic form. The numbers, designated only for the non-H atoms and H atoms that participate in hydrogen bonding, are consistent for the three polymorphs.

2.2. *Graph-set notation and examples*

As developed by Etter and coworkers, the graph-set description of hydrogen-bond patterns involves the assignment of a pattern designator or a combination of pattern designators to each pattern (Etter, 1990; Etter *et al.*, 1990). Four designators were defined: rings (**R**), chains (**C**), intramolecular hydrogen-bonding patterns, termed selfs (**S**), and other finite patterns, termed discretes (**D**). The designator also includes a subscript d to denote the number of hydrogen-bond donored H atoms in the pattern, and a superscript a to denote the number of hydrogen-bond acceptors in the pattern. The smallest number of bonds required to define the pattern (termed the degree by Etter) is included in parentheses to complete the definition of the pattern.† As special situations not considered by Etter have arisen some modifications to this notation have been made (Bernstein *et al.*, 1995, 1997). We now intend to term this designator the *quantitative descriptor* of the hydrogen-bond pattern. Examples of the use of these quantitative descriptors are given in Fig. 2.

The graph-set notation is particularly useful for comparing, for instance, the hydrogen-bonding patterns of different crystal structures (polymorphic structures), a family of similar molecules or molecules containing a hydrogen-bonding functional group or a number of similar hydrogen-bonding functional groups. To facilitate this comparison we have suggested using a common labeling for all hydrogen-bond donors and acceptors. These are labeled **a**, **b**, **c** *etc.* in Table 1. These then form the rows and columns in a matrix-type table which is

† In Etter's original definition the number of *atoms* constituted the degree of the pattern. In the interest of mathematical rigor we now believe that the number of *bonds* is a preferred definition of this parameter. When $d = a = 1$ these are considered as the default values and are generally not specifically noted.

Table 2. *Quantitative graph-set descriptors for the first and second levels for the three **IMDA** polymorphs*

L	**a**	**b**	**c**
IMDA1			
a	**C**(8)		
b	$\mathbf{R}_4^2(14)$	**C**(5)	
c	$\mathbf{C}_2^2(9)$	$\mathbf{C}_2^2(6)$	$\mathbf{R}_2^2(10)$

L	**a**	**b**	**d**
IMDA 3			
a	**C**(8)		
b	$\mathbf{C}_2^2(9)$	**C**(5)	
d	$\mathbf{R}_4^4(18)$	$\mathbf{C}_2^2(10)$	**C**(5)

L	**a**	**a′**	**b**	**b′**	**c**	**c′**
IMDA2						
a	**C**(8)					
a	-	**C**(8)				
b	$\mathbf{C}_2^2(9)$	-	**C**(5)			
b′	-	$\mathbf{C}_2^2(9)$	-	**C**(5)		
c	$\mathbf{D}_3^3(12)$	$\mathbf{D}_3^3(12)$	$\mathbf{D}_3^3(9)$	$\mathbf{D}_3^2(7)$	**D**	
c′	$\mathbf{D}_3^3(12)$	$\mathbf{D}_3^3(12)$	$\mathbf{D}_3^2(7)$	$\mathbf{D}_3^3(9)$	$\mathbf{R}_2^2(10)$	**D**

used to summarize the patterns. The diagonal elements of this table contain the quantitative descriptors for the patterns containing only one type of hydrogen bond. These were designated *motifs* by Etter and in combination constitute the *first* or *primary level* hydrogen-bonding pattern. Pairwise combinations of the individual patterns are given in the off-diagonal entries to this matrix; they constitute the *second* or *binary level* of hydrogen bonding (Bernstein *et al.*, 1995). The appropriate matrices for the three forms of **IMDA** are given in Table 2.†

Examination of the three matrices in Table 2 readily permits a comparison of the hydrogen-bonding patterns in the three structures. The motifs of **a** and **b** are identical in **IMDA1** and **IMDA3** [**C**(8) and **C**(5), respectively]. It is also seen that the chemically equivalent hydrogen bonds in **IMDA2** form the same patterns. **IMDA1** and **IMDA3** differ in the motifs for **c** (**IMDA1**) and **d** (**IMDA3**), the former creating an $\mathbf{R}_2^2(10)$ ring while the latter is a **C**(5) chain. In **IMDA2 c** and **c′** are simple discrete patterns at the first level, but at the second level the $\mathbf{R}_2^2(10)$ pattern appears (for the **c**, **c′** matrix element). A simple and easily recognized distinction in the hydrogen-bonding patterns between **IMDA1** and **IMDA2** may be found in the **a**, **b** matrix element $\mathbf{R}_4^2(14)$ for the former, while at the fourth level {**b**, **c′**, **b′**, **c**} there is a characteristic pattern $\mathbf{C}_2^2(9)$ for the latter (see Fig. 2).

† In the original publication (Etter *et al.*, 1990) the two molecules in the asymmetric unit for **IMDA2** were not considered separately. This treatment is somewhat lacking crystallographic rigor and the approach was modified in a subsequent publication (Bernstein *et al.*, 1995). Table 2 now contains the entries according to the modified treatment. The added rigor usually results in the need to define more patterns and the relevant chemical or crystallographic information appears at higher levels of the graph set, but none of this information is lost.

3. The G-array and the constructor graph of a crystal

3.1. *Encoding a crystal structure as a G-array*

With these examples of **IMDA** as a background for the chemical crystallographic aspects of the use of graph sets we now wish to describe the development of a mathematical model for the graph-set analysis.

The traditional model for representing a chemical molecule is the structural formula; in mathematical terms, this is an undirected graph consisting of points and lines connecting some of these points. Each atom is represented within this graph by a point, usually termed a *vertex*, and each covalent bond between two atoms is represented by a line between the respective vertices, usually termed an *edge*. To arrive at a graphical model for a crystal we start with a set of isolated undirected graphs – one for each molecule in the crystal – and add a second type of edge – those which correspond to hydrogen bonds. The additional edges are drawn as lines between the donored hydrogen and the acceptor atom of a hydrogen bond. In this way, a crystal can be thought of as a (huge) connected undirected graph with two different types of edges.

Edges representing covalent bonds we henceforth will refer to as *covalent edges*, while edges representing an interaction $H\cdots A$ between a hydrogen and an acceptor atom of a hydrogen bond will be termed *hydrogen edges* or *H-edges* for short. Information concerning the types of atoms and types of bonds is included in the graph by means of (the chemist's traditional atomic) labels which distinguish vertices representing hydrogen and other atoms, and by additional edge labels which also allow the distinction between covalent and H-edges. In graph-set analysis the resulting labeled graph is termed the *array* of the crystal structure (Etter *et al.*, 1990).

The labels in the array have to be chosen such that they express the existing crystallographic equivalences according to the space group *G* of the crystal. Equal labels of vertices indicate that the corresponding atoms are crystallographically equivalent, and analogously, equal labels of covalent or H-edges indicate that the corresponding bonds are equivalent. The labels for H-edges thereby take over the role of the designators used for hydrogen bonds in §2.

An ideal crystal structure consists of an infinite number of molecules. Accordingly, the array of an ideal crystal structure is an infinite graph. Now inevitably the question arises as to how we can label and handle an infinite graph. There is, as yet, no general answer to this question. However, in our case the infinite graph is a periodic array. Therefore, we need only to identify the smallest possible part of that array whose repetition can then be described by the translations of the unit cell and the other symmetry operations of the space group.

We denote the space group of a crystal by *G*. Since no representation of the array of a crystal can be given without the knowledge of *G* we propose to term this a *G-array*, rather than just an array.

Fig. 3 illustrates the arrays of the three **IMDA** forms. For **IMDA1** and **IMDA3** the part shown corresponds to the single molecule of an asymmetric unit. For **IMDA2** the two molecules in an asymmetric unit are given. H atoms which are not involved in H-edges are omitted for clarity. For the same reason, labels of covalent edges, which are irrelevant in our context, are not indicated.

3.2. *The constructor graph*

For identifying hydrogen-bond patterns in the first place it is not the inner structure of molecules which is of

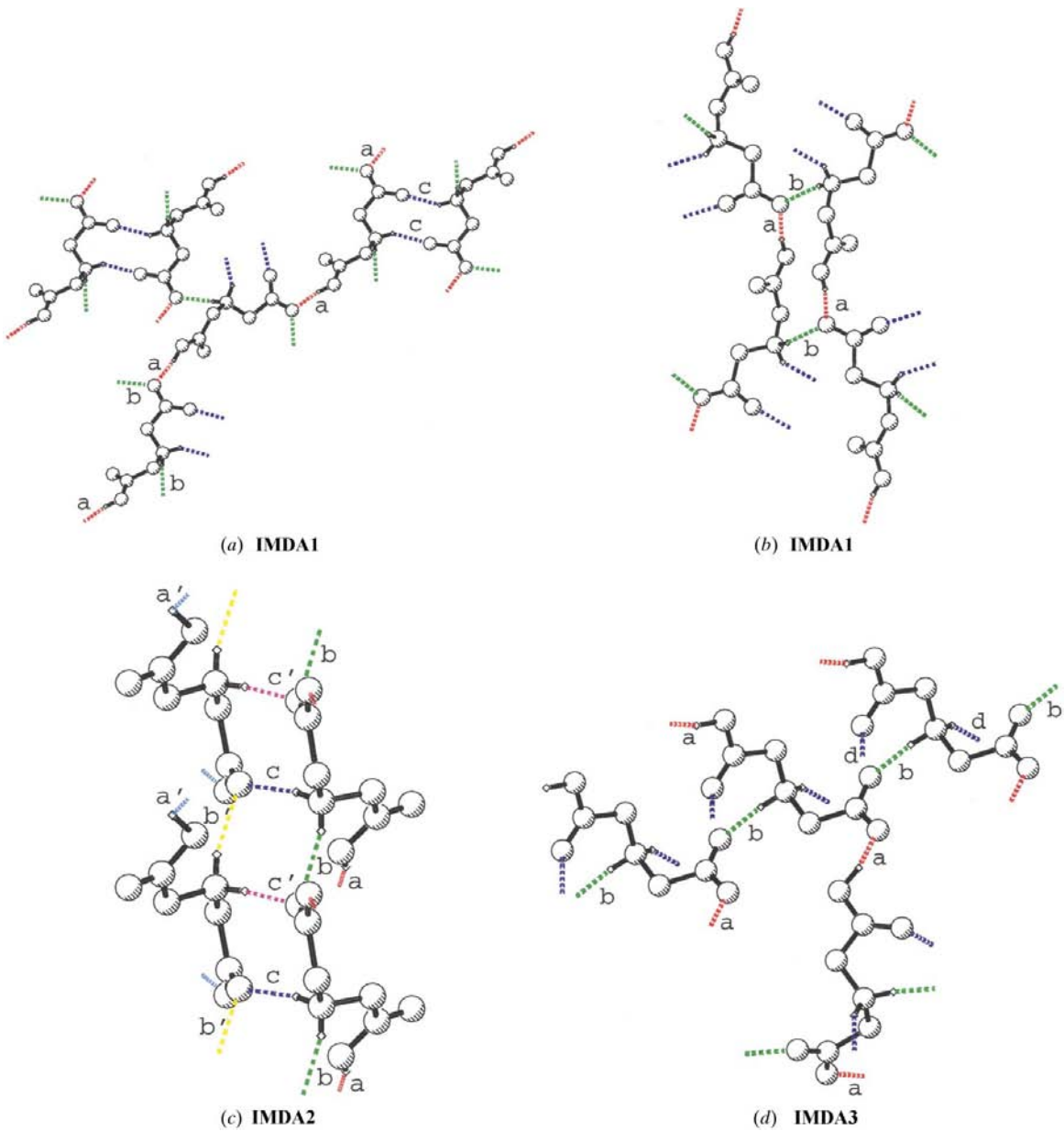

Fig. 2. Representative examples of hydrogen-bond patterns and the associated graph sets found in the three polymorphs of iminodiacetic acid, specified as **IMDA1**, **IMDA2** and **IMDA3**, respectively. The figures have been generated with the version of *PLUTO* (Allen & Kennard, 1993) in the October 1998 release of the CSD, described in the accompanying paper (Motherwell *et al.*, 1999). The labeling of the hydrogen bonds specified in {} is consistent with that in Tables 1 and 2. Each pattern may be identified by the colors of its hydrogen bonds, given in square brackets. (*a*) **IMDA1**: {**a**}: **C**(8) [red], {**b**}: **C**(5) [green], {**c**}: $\mathbf{R}_2^2(10)$ [blue]; (*b*) **IMDA1**: {**a**, **b**}: $\mathbf{R}_4^2(14)$ [red, green]; (*c*) **IMDA2**: {**a**}: **C**(8) [red], {**a**′}: **C**(8) [light blue], {**b**}: **C**(5) [green], {**b**′}: **C**(5) [yellow], {**c**}: **D** [dark blue], {**c**′}: **D** [lilac], {**b**, **c**′ **b**′, **c**}: $\mathbf{R}_4^2(8)$ [green, lilac, yellow, dark blue]; (*d*) **IMDA3**: {**a**}: **C**(8) [red], {**b**}: **C**(5) [green], {**d**}: **C**(5) [blue]; {**a**, **b**}: $\mathbf{C}_2^2(9)$ [red, green].

interest, but rather how these molecules are connected *via* hydrogen bonds to each other. Furthermore, when visualizing hydrogen-bond patterns, large or complex molecules tend to detract from the clarity of the picture and hinder one's ability to define and visualize the hydrogen-bonding patterns. For this reason, we now go a step further (already suggested *e.g.* in Etter, 1990) and create from the *G*-array a graph where the vertices are whole molecules rather than individual atoms. Such a reduction will of course result in a loss of chemical information (all covalent edges in a molecule disappear and are hidden into a black box, the vertex of the molecule); however, as we shall see, this apparent drawback can be overcome in a convenient way.

Consider any H-edge of the G-array. If this H-edge connects two different molecules, then one of them contains the vertex representing the donored H atom (and hence the donor atom). Let us term this the *donor molecule*. The other molecule contains the vertex which represents the acceptor atom. Call this the *acceptor molecule*. After reducing molecules of the *G*-array to single vertices we will indicate for H-edges which molecules are their donor and acceptor molecules by assigning an *orientation* to them. It is a matter of taste which one of the two possible orientations we choose. We decided to choose the orientation *from the donor to the acceptor molecule*. In the case of an *intramolecular* H-edge the donor molecule is the same as the acceptor molecule. In this case the corresponding H-edge becomes a loop attached to its molecule. Here an orientation is superfluous, even meaningless.

The structure which results from the two operations mentioned above (reduction of molecules to single vertices and orienting H-edges) is termed the *constructor graph* of the *G*-array. It has

the molecules as vertices,
the set of directed hydrogen edges as edges and
the same H-edge labels as the *G*-array.

Let us reconsider the graph-set representatives shown in Fig. 2. By the reduction process just described above they are reduced to the pictures shown in Fig. 4. Hydrogen edges are drawn as continuous arrows from donor to acceptor molecules. Note that a full or empty circle now represents an entire molecule from which hydrogen bonds emanate (arrows leaving signifying that the hydrogen-bond donor is on that molecule) or at which hydrogen bonds enter (arrows entering signifying that the hydrogen-bond acceptor is on that molecule). Note the difference from Fig. 3, where empty and full circles distinguish between non-G-equivalent molecules.

The constructor graph may have multiple edges as well as loops. This is, however, not the case with the three forms of **IMDA**. Fig. 5 shows representative parts of the constructor graphs of the three **IMDA** polymorphs.

As can be seen from Fig. 4, graph-set representatives appear in the constructor graph as simple subgraphs such as paths (Figs. 4 *a*, *d* and *e*) or cycles (Figs. 4*b*, *c* and *f*). Any cycle or path in the constructor-graph indicates a ring or a chain, respectively, and therefore can be considered as a candidate for a graph-set representative in constructor-graph representation and *vice versa*. A more detailed description of the correspondence between graph-set representatives, and paths and cycles in the constructor graph will be given in §5. Another

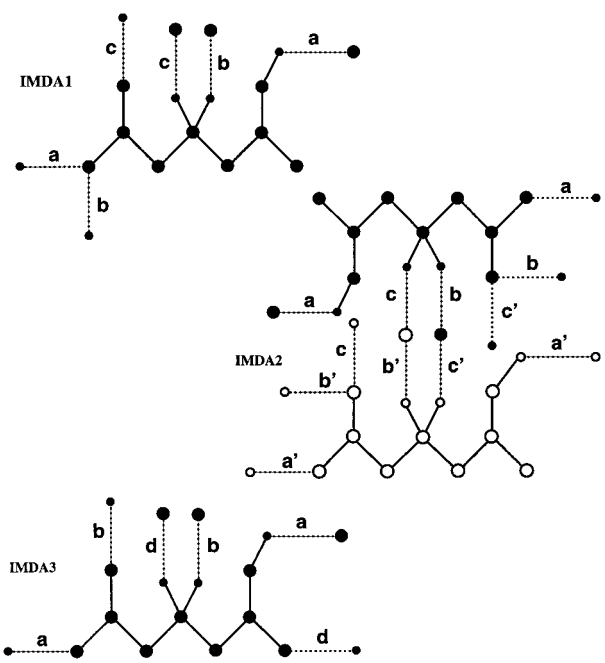

Fig. 3. Molecules and hydrogen bonds of the three **IMDA** polymorphs as part of their arrays. Small circles represent H atoms, large circles non-H atoms. In **IMDA2**, vertices representing atoms belonging to crystallographically non-equivalent molecules are additionally distinguished by full or empty circles. Covalent edges are drawn as full lines, H-edges as dotted lines with a label indicating the respective hydrogen-bond type corresponding to Table 1. H atoms not participating in hydrogen bonds are omitted.

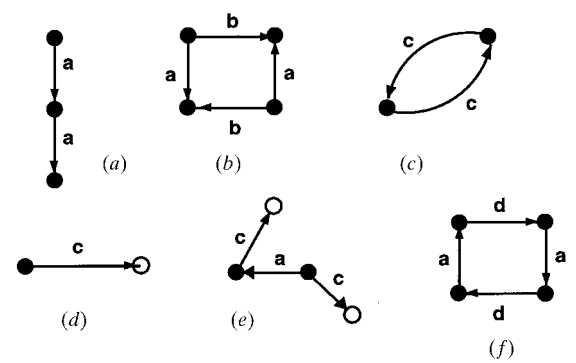

Fig. 4. Some of the graph-set representatives of **IMDA** portrayed using the constructor-graph representation. **IMDA1**: (*a*) {**a**}: **C**(8); (*b*) {**a**, **b**}: $\mathbf{R}_4^2(14)$; (*c*) {**c**}: $\mathbf{R}_2^2(10)$. **IMDA2**: (*d*) {**c**}: **D**; (*e*) {**a**, **c**}: $\mathbf{D}_3^3(12)$, **IMDA3**: (*f*) {**a**,**d**}: $\mathbf{R}_4^4(18)$.

question is how we can obtain the descriptors of graph sets from their representatives in the constructor graph. This question will be treated in §4 in the context of our example with the three **IMDA** polymorphs.

4. Qualitative descriptors for graph sets in the three IMDA polymorphs

4.1. *Qualitative descriptors*

In the *G*-array of a crystal graph sets are detected in the form of finite open (chain period, discrete) or closed (ring, self) paths. A finite path is a subgraph of the array, the vertices and edges of which can be arranged as an alternating sequence

$$v_0, e_1, v_1, e_2, \ldots, v_{n-1}, e_n, v_n,$$

where the v_is are vertices and the e_is are edges connecting v_{i-1} and v_i. In this paper we have to deal with so-called simple paths only, *i.e.* we always assume that all v_i vertices are different, except in a closed path, of course, where the last vertex v_n equals the first vertex v_0. For instance, the chain in Fig. 2(*a*) we have denoted until now as **C**(8) has a period which is given by the path

$$\mathrm{O3}-\mathrm{C4}-\mathrm{C3}-\mathrm{N}-\mathrm{C1}-\mathrm{C2}-\mathrm{O1}-\mathrm{H1}\cdot\overset{\mathbf{a}}{\cdot\cdot}\cdot\,\mathrm{O3}^{*}.$$

Here $\cdots$ denotes an H-edge, — a covalent edge and we have used the atom labels of Fig. 1 to denote vertices; O3* is the acceptor atom at the beginning of the next (8-atom) period. The H-edge has the label **a**, whereas the labels of the covalent edges, not relevant in this context, are not shown. Now starting at the H atom and moving along this path we encounter exactly one H-edge, which we cross in the direction from the donored H atom to the acceptor atom. This is also the direction which is assigned to this H-edge in the constructor graph. We characterize this situation by the symbol $\overrightarrow{\mathbf{a}}$. Let us combine this with the symbol **C** for a chain to give $\mathbf{C}(\overrightarrow{\mathbf{a}})$. This is what we propose to term the *qualitative descriptor* of the chain under consideration. Its usefulness will be shown below.

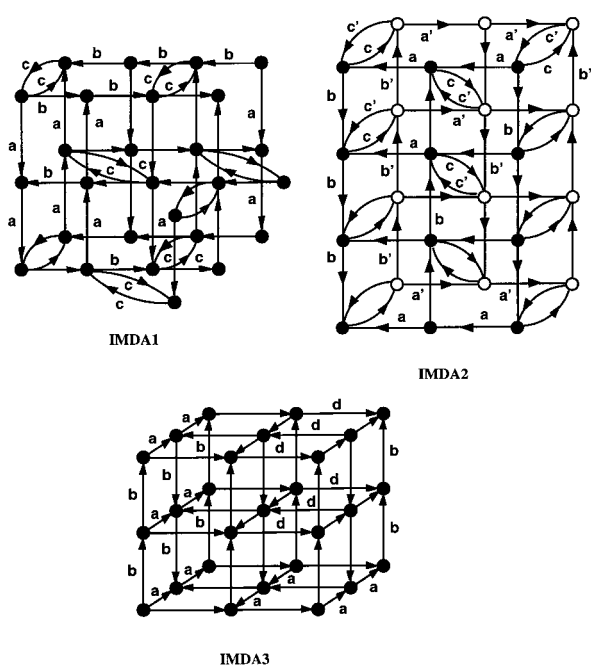

Fig. 5. The crystal structures of the three polymorphs of **IMDA** using constructor-graph representations. Not all labels are indicated. Parallel and equally directed arrows have equal labels.

There is also a chain with qualitative descriptor $\mathbf{C}(\overrightarrow{\mathbf{b}})$ in **IMDA1**, namely the chain with the period

$$\mathrm{O3}-\mathrm{C4}-\mathrm{C3}-\mathrm{N}-\mathrm{H2}\cdot\overset{\mathbf{b}}{\cdot\cdot}\cdot\,\mathrm{O3}^{*}.$$

The three different chains of **IMDA3** noted in Table 2 have qualitative descriptors $\mathbf{C}(\overrightarrow{\mathbf{a}})$, $\mathbf{C}(\overrightarrow{\mathbf{b}})$ and $\mathbf{C}(\overrightarrow{\mathbf{d}})$. There are four chains of **IMDA2** with qualitative descriptors $\mathbf{C}(\overrightarrow{\mathbf{a}})$,$\mathbf{C}(\overrightarrow{\mathbf{a}'})$,$\mathbf{C}(\overrightarrow{\mathbf{b}})$ and $\mathbf{C}(\overrightarrow{\mathbf{b}'})$, respectively.

Let us turn to second-level chains; such chains are also listed in Table 2. From Fig. 5 we can immediately see that **IMDA1** contains a chain with qualitative descriptor $\mathbf{C}(\overrightarrow{\mathbf{a}}\overleftarrow{\mathbf{c}})$ [$C_2^2(9)$ in Table 2]. In this descriptor the symbol $\overleftarrow{\mathbf{c}}$ appears. We use it to note the fact that moving along a period of the chain we encounter (after having passed the first H-edge with label **a** *in* its direction) an H-edge with label **c**, which we have to pass *against* its direction, *i.e.* from acceptor to donored H-atom.

From Fig. 5 we also see that **IMDA1** contains a chain with qualitative descriptor $\mathbf{C}(\overrightarrow{\mathbf{a}}\overleftarrow{\mathbf{b}}\overleftarrow{\mathbf{a}}\overrightarrow{\mathbf{b}})$. (Start at the leftmost molecule on the top and move down vertically from vertex to vertex, turn right, go down again and turn right once more. Repeating this movement periodically will lead us along a chain.) In this descriptor the symbols $\overleftarrow{\mathbf{b}}$ and $\overleftarrow{\mathbf{a}}$ appear. They indicate the fact that moving along a period of the chain we encounter (after having passed an H-edge with label **a** in its direction) an H-edge with label **b** and then one with label **a**, which we have to pass against their direction. The last H-edge on this period is labeled **b** again and is passed in its direction. This chain is not listed in Table 2.

Rings and discretes will be treated in a similar manner. In **IMDA1** there appears, for instance, the ring

$$\mathrm{H1}^{(1)}\cdot\overset{\mathbf{a}}{\cdot\cdot}\cdot\,\mathrm{O3}^{(2)}\overset{\mathbf{b}}{-}\mathrm{H1}^{(3)}\cdot\overset{\mathbf{a}}{\cdot\cdot}\cdot\,\mathrm{O3}^{(4)}\cdot\overset{\mathbf{b}}{\cdot\cdot}\cdot\,\mathrm{H2}^{(2)}-\mathrm{H1}^{(1)}$$

[Fig. 2*b*, $\mathbf{R}_4^2(14)$; upper indices in brackets distinguish the four molecules involved, ——— indicates a shortest path between H atoms in a molecule]. The corresponding qualitative descriptor is $\mathbf{R}(\overrightarrow{\mathbf{a}}\overleftarrow{\mathbf{b}}\overrightarrow{\mathbf{a}}\overleftarrow{\mathbf{b}})$.

The sequence of atoms in a ring can be started at any atom in the ring. If we start at the atom $\mathrm{O3}^{(2)}$ then the first H-edge we encounter moving along the ring is one labeled **b** and we pass it against its direction. Thus, $\mathbf{R}(\overleftarrow{\mathbf{b}}\overrightarrow{\mathbf{a}}\overleftarrow{\mathbf{b}}\overrightarrow{\mathbf{a}})$ is also a qualitative descriptor of the same ring. Further, if we choose to move along the ring

Table 3. *Qualitative graph-set descriptors for the first and second levels for the three* ***IMDA*** *polymorphs*

L	**a**	**b**	**c**
IMDA1			
a	$\mathbf{C}(\overrightarrow{\mathbf{a}})$		
b	$\{\mathbf{R}\}(\overrightarrow{\mathbf{a}}\ \overleftarrow{\mathbf{b}}\ \overrightarrow{\mathbf{a}}\ \overleftarrow{\mathbf{b}})$	$\mathbf{C}(\overrightarrow{\mathbf{b}})$	
c	$\mathbf{C}(\overrightarrow{\mathbf{a}}\ \overleftarrow{\mathbf{c}})$	$\mathbf{C}(\overrightarrow{\mathbf{b}}\ \overleftarrow{\mathbf{c}})$	$\mathbf{R}(\overrightarrow{\mathbf{c}}\ \overrightarrow{\mathbf{c}})$

L	**a**	**b**	**d**
IMDA3			
a	$\mathbf{C}(\overrightarrow{\mathbf{a}})$		
b	$\mathbf{C}(\overrightarrow{\mathbf{a}}\ \overleftarrow{\mathbf{b}})$	$\mathbf{C}(\overrightarrow{\mathbf{b}})$	
d	$\mathbf{R}(\overrightarrow{\mathbf{a}}\ \overrightarrow{\mathbf{d}}\ \overrightarrow{\mathbf{a}}\ \overrightarrow{\mathbf{d}})$	$\mathbf{C}(\overrightarrow{\mathbf{b}}\ \overrightarrow{\mathbf{d}})$	$\mathbf{C}(\overrightarrow{\mathbf{d}})$

L	**a**	**a′**	**b**	**b′**	**c**	**c′**
IMDA2						
a	$\mathbf{C}(\overrightarrow{\mathbf{a}})$					
a′	–	$\mathbf{C}(\overrightarrow{\mathbf{a}'})$				
b	$\mathbf{C}(\overrightarrow{\mathbf{a}}\ \overleftarrow{\mathbf{b}})$	–	$\mathbf{C}(\overrightarrow{\mathbf{b}'})$			
b′	–	$\mathbf{C}(\overrightarrow{\mathbf{a}'}\overleftarrow{\mathbf{b}'})$	–	$\mathbf{C}(\overrightarrow{\mathbf{b}'})$		
c	$\mathbf{D}(\overleftarrow{\mathbf{c}}\ \overrightarrow{\mathbf{a}}\ \overrightarrow{\mathbf{c}})$	$\mathbf{D}(\overrightarrow{\mathbf{c}}\ \overrightarrow{\mathbf{a}'}\overleftarrow{\mathbf{c}})$	$\mathbf{D}(\overleftarrow{\mathbf{c}}\ \overrightarrow{\mathbf{b}'}\ \overrightarrow{\mathbf{c}})$	$\mathbf{D}(\overrightarrow{\mathbf{c}}\ \overrightarrow{\mathbf{b}'}\overleftarrow{\mathbf{c}})$	$\mathbf{D}(\overrightarrow{\mathbf{c}})$	
c′	$\mathbf{D}(\overrightarrow{\mathbf{c}'}\ \overrightarrow{\mathbf{a}}\ \overrightarrow{\mathbf{c}'})$	$\mathbf{D}(\overrightarrow{\mathbf{c}'}\ \overrightarrow{\mathbf{a}'}\overrightarrow{\mathbf{c}'})$	$\mathbf{D}(\overrightarrow{\mathbf{c}'}\ \overrightarrow{\mathbf{b}'}\ \overrightarrow{\mathbf{c}'})$	$\mathbf{D}(\overrightarrow{\mathbf{c}'}\ \overrightarrow{\mathbf{b}'}\ \overrightarrow{\mathbf{c}'})$	$\mathbf{R}(\overrightarrow{\mathbf{c}}\ \overrightarrow{\mathbf{c}'})$	$\mathbf{D}(\overrightarrow{\mathbf{c}'})$

in the opposite direction, then the order in which the H-edges are met is reversed and also the direction in which we pass them is inverted. Thus, $\mathbf{R}(\overrightarrow{\mathbf{b}}\overleftarrow{\mathbf{a}}\overrightarrow{\mathbf{b}}\overleftarrow{\mathbf{a}})$ and $\mathbf{R}(\overleftarrow{\mathbf{a}}\overrightarrow{\mathbf{b}}\overleftarrow{\mathbf{a}}\overrightarrow{\mathbf{b}})$ are two more qualitative descriptors of the same ring. It turns out that for the purpose of these descriptors it does not matter which one we choose. They are all equivalent. The same observation is also valid for qualitative descriptors of chains introduced above and for those of discretes which we consider next.

In **IMDA2** the first-level discretes H2$\cdots$O3′ and H2′$\cdots$O3 and, among others, the second-level discrete

$$\mathrm{H2}\cdot\overset{\mathbf{c}}{\vdots}\cdot\mathrm{O4'}^{*}\cdot\overset{\mathbf{b}'}{\vdots}\cdot\mathrm{H3'}-\mathrm{N'}-\mathrm{C3'}-\mathrm{C4'}-\mathrm{O4'}\cdot\overset{\mathbf{c}}{\vdots}\cdot\mathrm{H2}^{*}$$

appear, to which we can assign the qualitative descriptors $\mathbf{D}(\overrightarrow{\mathbf{c}})$, $\mathbf{D}(\overrightarrow{\mathbf{c}'})$ and $\mathbf{D}(\overrightarrow{\mathbf{c}}\overrightarrow{\mathbf{b}'}\overleftarrow{\mathbf{c}})$, respectively. The existence of discretes with such qualitative descriptors follows directly from the constructor graph in Fig. 5.

To provide still more examples for qualitative descriptors we add here Table 3, in which all qualitative descriptors for the graph sets listed in Table 2 are given. Each entry in this table corresponds to the quantitative descriptor in Table 2, which is in the same position of the corresponding array.

We invite the reader to convince himself/herself that the existence of all the graph sets for which the qualitative descriptors are given in Table 3 can be conveniently checked using the constructor graph representations of the three **IMDA** polymorphs in Fig. 5.

4.2. *Covalent distance tables of IMDA*

The idea is to search the constructor graph, which gives a clear picture of the hydrogen-bond pattern of a crystal structure, for representatives of graph sets, to establish their qualitative descriptors and to derive from them the corresponding quantitative ones. This is possible, and in fact convenient, as we intend to show in this subsection.

Let us start with an example and consider the ring $\mathbf{R}(\overrightarrow{\mathbf{a}}\overleftarrow{\mathbf{b}}\overrightarrow{\mathbf{a}}\overleftarrow{\mathbf{b}})$ of **IMDA1**. In order to find the quantitative descriptor $\mathbf{R}_d^a(n)$, we have to determine the number of edges n, the number of acceptors a and the number of donored H atoms d, which participate in forming the ring.

We deal first with the number of edges n on the ring. The number of H-edges is evident: it is simply equal to the number of arrowed letters appearing in the qualitative descriptor. But how many covalent edges are between two consecutive H-edges, for example, how many are between $\overrightarrow{\mathbf{a}}$ and $\overleftarrow{\mathbf{b}}$? These covalent edges belong to a single molecule of **IMDA1**. There is exactly one directed H-edge with label **a** and exactly one directed H-edge with label **b** entering this molecule. Therefore, the combination $\overrightarrow{\mathbf{a}}\overleftarrow{\mathbf{b}}$ uniquely determines two atoms in this molecule: the acceptor atoms of these two H-edges (see Fig. 3). The number of covalent edges we are looking for equals the *covalent distance*, *i.e.* the length of a shortest path between these two atoms. A similar observation is true for the next combination $\overleftarrow{\mathbf{b}}\overrightarrow{\mathbf{a}}$. The number of covalent edges between these two H-edges equals the covalent distance between the donored H-atom of the first and the donored H-atom of the second H-edge. The covalent distances for these and other combinations may be found by simple counting in Fig. 3.

The four individual tables in Table 4 contain the covalent distances of all combinations of two directed H-edges which may appear in a qualitative descriptor of a graph set in any of the three **IMDA** polymorphs. To find the entry in a table which tells us the covalent

Table 4. *Covalent distance tables of the three* ***IMDA*** *polymorphs*

The entry cd indicates *covalent distance.*

cd	$\vec{a}$	$\overleftarrow{a}$	$\vec{b}$	$\overleftarrow{b}$	$\vec{c}$	$\overleftarrow{c}$
IMDA1						
$\overleftarrow{a}$	0	7	5	7	5	7
$\vec{a}$	7	0	4	0	4	2
$\overleftarrow{b}$	5	4	0	4	2	4
$\vec{b}$	7	0	4	0	4	2
$\overleftarrow{c}$	5	4	2	4	0	4
$\vec{c}$	7	2	4	2	4	0

cd	$\vec{a}$	$\overleftarrow{a}$	$\vec{b}$	$\overleftarrow{b}$	$\overleftarrow{c'}$	$\vec{c}$
IMDA2, residue 1						
$\overleftarrow{a}$	0	7	5	7	7	5
$\vec{a}$	7	0	4	2	2	4
$\overleftarrow{b}$	5	4	0	4	4	2
$\vec{b}$	7	2	4	0	0	4
$\vec{c'}$	7	2	4	0	0	4
$\overleftarrow{c}$	5	4	2	4	4	0

cd	$\vec{a'}$	$\overleftarrow{a'}$	$\vec{b'}$	$\overleftarrow{b'}$	$\overleftarrow{c}$	$\vec{c'}$
IMDA2, residue 2						
$\overleftarrow{a'}$	0	7	5	7	7	5
$\vec{a'}$	7	0	4	2	2	4
$\overleftarrow{b'}$	5	4	0	4	4	2
$\vec{b'}$	7	2	4	0	0	4
$\vec{c}$	7	2	4	0	0	4
$\overleftarrow{c'}$	5	4	2	4	4	0

cd	$\vec{a}$	$\overleftarrow{a}$	$\vec{b}$	$\overleftarrow{b}$	$\vec{d}$	$\overleftarrow{d}$
IMDA3						
$\overleftarrow{a}$	0	7	5	7	5	3
$\vec{a}$	7	0	4	2	4	6
$\overleftarrow{b}$	5	4	0	4	2	4
$\vec{b}$	7	2	4	0	4	6
$\overleftarrow{d}$	5	4	2	4	0	4
$\vec{d}$	3	6	4	6	4	0

distance between $\vec{a}$ and $\overleftarrow{b}$, for example, we have to look at row $\vec{a}$ and column $\overleftarrow{b}$. For **IMDA1** we find there the entry 0, which indicates that there is no covalent edge on a shortest path between these two H-edges on the ring; they have the same acceptor atom. The entry in row $\overleftarrow{b}$ and column $\vec{a}$, however, is 5. Therefore, since in the ring $\mathbf{R}(\vec{a}\,\overleftarrow{b}\,\vec{a}\,\overleftarrow{b})$ there are four covalent distances, from $\vec{a}$ to $\overleftarrow{b}$, from $\overleftarrow{b}$ to $\vec{a}$, again from $\vec{a}$ to $\overleftarrow{b}$ and back again from $\overleftarrow{b}$ to $\vec{a}$ (from the last to the first H-edge), and since there are four H-edges, we obtain the result $n = 0 + 5 + 0 + 5 + 4 = 14$ by referring to Table 4.

Next consider the number of acceptor atoms *a.* Since every H-edge ends in an acceptor molecule, there can be as many acceptors as there are H-edges. However, as we could observe above, it may happen that two consecutive H-edges in a ring (or in any other graph set) share a single acceptor atom. Each occurrence of such a sharing reduces the number of acceptors by one (compared with the maximum number possible). Hence, the number of acceptors a equals the number of H-edges minus the number of pairs of H-edges which share a single acceptor.

Sharing of acceptors can also be detected using Table 4. This occurs when the label arrows are head-to-head (such as $\rightarrow\leftarrow$) and the entry in the covalent distance table which corresponds to the two arrowed labels is zero. For instance, this happens in the trivial cases $\vec{a}\,\overleftarrow{a}$, $\vec{b}\,\overleftarrow{b}$, and so on, which are of no relevance here. However, it also happens to $\vec{a}\,\overleftarrow{b}$, $\vec{b}\,\overleftarrow{a}$ (**IMDA1**), $\vec{b}\,\overleftarrow{c'}$, $\overleftarrow{c}\,\vec{b}$ (**IMDA2**, residue 1), and $\vec{b'}\,\overleftarrow{c}$ and $\vec{b}\,\overleftarrow{b'}$ (**IMDA2**, residue 2).

In our example $\mathbf{R}(\vec{a}\,\overleftarrow{b}\,\vec{a}\,\overleftarrow{b})$, there are two occurrences of $\vec{a}\,\overleftarrow{b}$ and therefore we obtain $a = 4 - 2 = 2$.

Finally, consider the number of donored H-atoms d. This number equals the number of H-edges minus the number of pairs of consecutive H-edges which share a single donored H-atom. A donored H-atom is shared by two H-edges when the label arrows are tail-to-tail (such as $\leftarrow\rightarrow$) and the entry in the covalent distance table which corresponds to the two arrowed labels is zero. This happens, for instance, in the case with $\overleftarrow{a}\,\vec{a}$, $\overleftarrow{b}\,\vec{b}$, and so on. In the ring $\mathbf{R}(\vec{a}\,\overleftarrow{b}\,\vec{a}\,\overleftarrow{b})$, no such case occurs. Hence, here we obtain $d = 4$.

In this manner we can determine the quantitative descriptor $\mathbf{R}_4^2(14)$ for the ring $\mathbf{R}(\vec{a}\,\overleftarrow{b}\,\vec{a}\,\overleftarrow{b})$, as obtained earlier by the relatively simple, but often confusing, method of picking and counting.

The consideration of some more examples will demonstrate additional features of this process. In a first-level chain such as $\mathbf{C}(\vec{a})$ in **IMDA1** or **IMDA3** we have only one distance to look up, namely the covalent distance between an H-edge with the label $\vec{a}$ and one with the same label in the next molecule of the chain. In both polymorphs this distance is 7. Hence, $n = 7 + 1 = 8$. No sharing of acceptors or donored H-atoms occurs. Hence, in both cases $d = a = 1$, giving the quantitative descriptor $\mathbf{C}(8)$.

Next consider the chain $\mathbf{C}(\vec{b}\,\vec{d})$ in **IMDA3**. Consulting the corresponding covalent distance table in Table 3 we find twice the distances 4 (corresponding to $\vec{b}\,\vec{d}$ and $\vec{d}\,\vec{b}$). Hence, $n = 4 + 4 + 2 = 10$. No sharing of acceptors or donored H-atom occurs. Therefore, $a = d = 2$ and we obtain the quantitative descriptor $\mathbf{C}_2^2(10)$.

Consider the discrete $\mathbf{D}(\overleftarrow{c}\,\vec{a}\,\overleftarrow{c})$ in **IMDA2**. In a discrete with three H-edges we have only two covalent distances, here according to $\overleftarrow{c}\,\vec{a}$ and $\vec{a}\,\overleftarrow{c}$, which are 5 and 4. This gives $n = 5 + 4 + 3 = 12$. No sharing of acceptors or donored H-atoms occurs. Therefore, $a = d = 3$ and the quantitative descriptor is $\mathbf{D}_3^3(12)$.

As a final example, consider the discrete $\mathbf{D}(\vec{c'}\,\vec{b}\,\overleftarrow{c'})$ in **IMDA2**. The covalent distances according to $\vec{c'}\,\vec{b}$ and $\vec{b}\,\overleftarrow{c'}$ are 4 and 0. This yields $n = 4 + 0 + 3 = 7$. The H-edges with labels $\vec{b}$ and $\vec{c'}$ share the acceptor atom, therefore, $a = 3 - 1 = 2$. No sharing of donored H-atoms occurs, so $d = 3$. Thus, we obtain the quantitative descriptor $\mathbf{D}_3^2(7)$.

The examples treated here demonstrate that the constructor graph is a very convenient tool in graph-set analysis. With the three **IMDA** polymorphs, graph sets in constructor graph representation are easily detected and their qualitative descriptors can readily be transformed into their quantitative descriptors using the covalent distance tables. The question arises: *Is this always true? Can we proceed in this way also with other crystal structures*? The answer is: *Yes, if the crystal structure is such that there are no crystallographically equivalent H-edges emanating from the same donor molecule or entering the same acceptor molecule*. Crystallographically equivalent H-edges can occur when a molecule lies in a special position, *i.e.* on a crystallographic symmetry element.

A crystal structure with the property formulated in the above answer allows us to establish well defined covalent distance tables in advance, *i.e.* before we start the true graph-set analysis. Further, with such a crystal structure the sequence of arrowed labels in a qualitative descriptor defines uniquely a sequence of molecules and hydrogen bonds, when the first individual molecule is fixed (Bernstein *et al.*, 1997). If the crystal structure does not possess this property, then one or even both of these consequences may be wrong. In such cases we may use the constructor graph only after this tool has undergone some appropriate refinement. This refinement can be achieved in a rather elegant way, as described in a slightly more mathematical discussion of this subject in the next section.

5. Some simple but useful mathematical considerations

5.1. *G-equivalence*

In §3 we discussed the *G*-array of a crystal structure. It is an undirected graph with vertex labels and edge labels. In essence the vertex labels denote the crystallographic equivalence of atoms; the edge labels do the same for covalent and hydrogen bonds. Hence, we may also speak of equivalent vertices and equivalent edges. We will term two vertices equivalent if the atoms for which they stand are crystallographically equivalent.† Analogously, we will term two covalent edges or two H-edges equivalent if the bonds they denote or specify are crystallographically equivalent.

Having the *G*-array in hand, the equivalence or non-equivalence of vertices or edges can be verified by looking at their labels. However, in dealing with graph sets we may wish to be able to determine the equivalence or non-equivalence of larger parts of the graph which consist of several vertices and several edges (as for instance subgraphs representing molecules). We will say that two different parts of the G-array are equivalent if and only if there is a symmetry operation g in the space group G which maps one part onto the other. For those readers who prefer a rigorous definition we reformulate this idea in a somewhat more formal manner. In the context of graph sets it suffices to restrict ourselves to the consideration of open or closed paths. No other parts of the *G*-array (except, of course, single vertices, edges and subgraphs representing molecules) will be involved in our considerations.

Consider a path P of length n in the *G*-array, given by an alternating sequence

$$v_0, e_1, v_1, e_2, \ldots, v_{n-1}, e_n, v_n$$

as in §4.1, with vertices v_i and edges e_i. Denote the vertex onto which a vertex v_i is mapped under the symmetry operation g by v_i^g.‡ Analoguously, denote the edge onto which an edge e_i is mapped under g by e_i^g. Then a path $\overline{P}$ of the G-array is termed G-equivalent to P if and only if there is a symmetry operation g in G such that $\overline{P}$ is given by

$$v_0^g, e_1^g, v_1^g, e_2^g, \ldots, v_{n-1}^g, e_n^g, v_n^g,$$

i.e. if P is mapped by g onto $\overline{P}$.

From now on we deliberately use the notation *G*-equivalent in order to stress the fact that this equivalence is caused by the space group G. In the sequel this notation will be used generally, in particular also in the above-considered case of single vertices and single edges which are paths of length 0 and 1, respectively. Hence, for vertices and edges equivalent means *G*-equivalent.

Our definition of *G*-equivalence also applies to graph sets. Rings and selfs are closed paths, periods of chains and discretes are open paths. Hence, for any representative of a graph set, say a chain **C**, a ring **R**, a discrete **D** or a self **S**, there is a whole infinite set of *G*-equivalent representatives (which motivates the use of the term graph set). In terms of logic, a chain (a ring, a discrete or a self) is an abstract notion meaning the set of all representatives, each representative being a possible realisation of this abstract notion.

5.2. *T-equivalence*

There are crystal structures in which several *G*-equivalent H-edges emanate from the same donor molecule and/or enter the same acceptor molecule.

We consider a structure with molecules in a special position, **ACNACO** [*trans*-tetraamminedinitrocobalt(III) acetate; Cagnon *et al.*, 1978]. **ACNACO** crystallizes in space group *Cmcm* (No. 63). The cation and anion each lie on a crystallographic special position with site symmetry *m*2*m*. **ACNACO** consists of crystal-

† In terms of the *G*-array this means that one of the vertices is mapped onto the other by an operation g of G.

‡ The reader may think of v_i as given, for instance, by the triple (x_i, y_i, z_i) of atomic coordinates of the atom which is represented by v_i. Hence, for example, if g is a *c*-glide reflection plus translation along the *a*-axis (as in the first row of Table 1, Section 2.1), then v_i^g is the vertex representing the atom with atomic coordinates $\left(x_i + 1, \overline{y}_i + \frac{1}{2}, z_i - \frac{1}{2}\right)$.

lographically equivalent layers which are isolated from each other, *i.e.* neither covalent edges nor H-edges exist between different layers. Therefore, each graph-set representative of **ACNACO** is situated completely within a single layer.

As already mentioned at the end of §4, more than one *G*-equivalent H-edge entering or emanating from a molecule creates an ambiguous situation. For instance, let us enter the cation in Fig. 6 from the right-most atom O1B^* *via* an H-edge with the label **a** and leave this molecule *via* an H-edge with the label **b**. There are eight different possibilities to do this, involving the shortest covalent paths of length 2 or 4. Therefore, the pair $\overleftarrow{\mathbf{a}}\,\overrightarrow{\mathbf{b}}$, when appearing in the arrowed label sequence of some qualitative descriptor, does not give us unique information for describing a graph set, and similar situations are met with other combinations of arrowed letters. For example, in Fig. 6 we find the chains with periods

$$\mathrm{H2}B^* \cdot\overset{\mathbf{a}}{\cdot}\cdot \mathrm{O1}B\mathrm{—C1—O1} \cdot\overset{\mathbf{a}}{\cdot}\cdot \mathrm{H2—N1—Co—N1}B\mathrm{—H2}B$$

and

$$\mathrm{H2}B^* \cdot\overset{\mathbf{a}}{\cdot}\cdot \mathrm{O1}B\mathrm{C1—O1} \cdot\overset{\mathbf{a}}{\cdot}\cdot \mathrm{H2}N\mathrm{—N1}N\mathrm{—Co—N1}L\mathrm{—H2}L.$$

Both have a period with directed label sequence $\overrightarrow{\mathbf{a}}\,\overleftarrow{\mathbf{a}}$. The first is mapped onto itself only by (multiples of) a translation along the direction $\overrightarrow{a} = (1, 0, 0)$ (not to be confused with the arrowed label $\overrightarrow{\mathbf{a}}$), the second only by (multiples of) a glide reflection perpendicular to direction $\overrightarrow{c} = (0, 0, 1)$ with translation part $(1, 0, 0)$. They are crystallographically **not** equivalent, *i.e.* not *G*-equivalent. However, indicating the *G*-labels in the respective qualitative descriptor they would become the same. Furthermore, there is also a ring

$$\mathrm{H2} \cdot\overset{\mathbf{a}}{\cdot}\cdot \mathrm{O1} \cdot\overset{\mathbf{a}}{\cdot}\cdot \mathrm{H2}N\mathrm{—Co—N1—H2}$$

with the same label sequence $\overrightarrow{\mathbf{a}}\,\overleftarrow{\mathbf{a}}$, which is obviously not crystallographically equivalent to a period of any chain.

In such situations, in order to work with the constructor graph a method must be found to eliminate these ambiguities.

One way to meet this requirement is to refine the H-edge labels appearing in the *G*-array with an additional index. Instead of using **a** and **b**, we shall use different labels **a**1, **a**2, . . ., and **b**1, **b**2, . . ., for the H-edges incident with a molecule. This provides the required distinction. However, this label refinement has to be performed consistently at different, but G-equivalent, molecules. Otherwise, we are not able to retrieve a graph-set representation from its qualitative descriptor. Such a process requires a guiding principle for performing this label refinement. We shall first discuss such a principle in a general way, and then return to the specific example of **ACNACO**.

Our approach to a refinement of the labeling follows the treatment in Bernstein *et al.* (1997). It is based on the subgroup T of the space group G, which consists of all translations of the crystal structure. This subgroup is used in order to add the notion of T-equivalence, or translational equivalence, as it is often called in crystallography, to the earlier defined notion of G-equivalence. Two parts of the G-array of a crystal structure are termed *translationally equivalent*, or *T-equivalent* for short, if and only if there is some translation $t \in T$, which maps one part onto the other. In particular, two paths P and $\overline{P}$, where P is described by the sequence $v_0, e_1, v_1, e_2, \ldots, v_{n-1}, e_n, v_n$, are T-equivalent if and only if there is a translation $t \in T$ such that $\overline{P}$ is given by

$$v_0^t, e_1^t, v_1^t, e_2^t, \ldots, v_{n-1}^t, e_n^t, v_n^t,$$

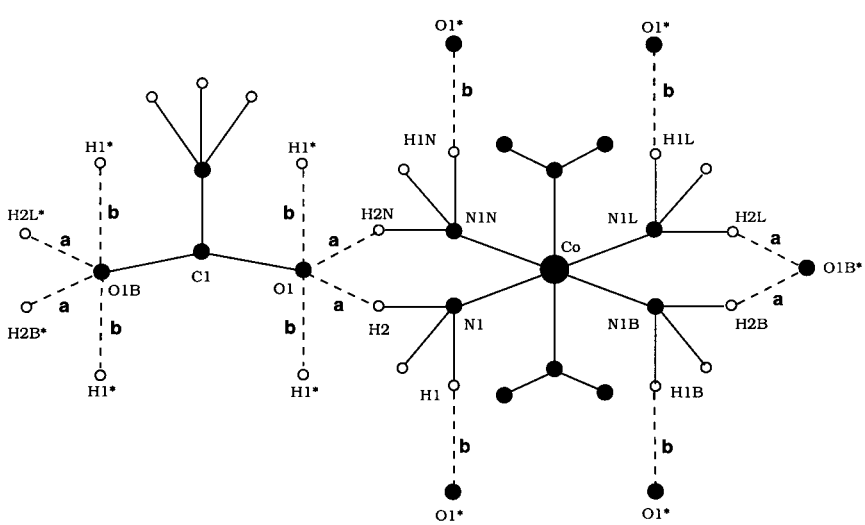

Fig. 6. The cation and anion of the crystal structure of **ACNACO** shown as part of the G-array; empirical formulae are $[Co(NO_2)_2(NH_3)_4]^+$ and $[C_2H_3O_2]^-$. Atom labeling is consistent with that in the CSD entry for **ACNACO**. H-edges are indicated by broken lines, starred atom labels such as O1*, H1* *etc.* indicate that the corresponding atoms belong to another molecule, not necessarily all to the same. Atoms not involved in a graph set are not labeled. H-atoms are indicated by empty circles, non-H atoms by full circles.

i.e. if P is mapped onto $\overline{P}$ by a translation t.†

In our context, T-equivalence has no *a priori* crystallographic relevance. It is an auxiliary notion which turns out to be very convenient in graph-set analysis, as we intend to show in this section. A first motivation for the use of this notion comes from the following

Fact. No two H-edges emanating from the same donor molecule or entering the same acceptor molecule are T-equivalent.

This is clear by the following consideration. Since a translation (except the identity) never belongs to a special position, it does not fix a molecule. Therefore, no two H-edges emanating from or entering the same molecule can be such that one is mapped onto the other by a proper translation.

We now propose to use the following rule for labeling the G-array and the constructor graph of a crystal structure

Refine the labels of the H-edges in the G-array in such a way that H-edges are T-equivalent if and only if they have the same refined label. Use the resulting labeling also for the edges of the constructor graph.

To have a compact working term, a labeling of this type will be termed *T-labeling*.

Returning now from general considerations to the example of **ACNACO** we have prepared in Fig. 7 a T-labeling of some part of its G-array. The space group of **ACNACO** is *Cmcm*. Since each graph-set representative is situated entirely within a single layer, for graph-set analysis it is sufficient to restrict our considerations to one layer. All layers of **ACNACO** are translationally equivalent. They are parallel to the xz plane. Any layer is transformed into the consecutive one by the C-face centering translation along vector $\frac{1}{2}\left(\overrightarrow{a}+\overrightarrow{b}\right)=\left(\frac{1}{2},\frac{1}{2},0\right)$. The layer group of an **ACNACO** layer is *Pm*(*n*)*m* (Weber #46; see Weber, 1929, or Grell *et al.*, 1988). The direction perpendicular to the layer plane – here the y-direction – is in parentheses. The layer group which we need here can be derived from the space group *Pmmn* (IT #59), taking into account a coordinate transformation $(x, y, z)\mapsto(z, x, y)$.‡

The crystal class (*i.e.* the corresponding point group) of *Pm*(*n*)*m* is *mmm*, which consists of eight elements. Hence, each class of G-equivalent objects is decomposed into eight T-equivalence classes for each type **a** and **b** of hydrogen bonds. In the T-labeling for **ACNACO** (Fig. 8) we use the labels **a1**, ..., **a8** and **b1**, ..., **b8** to refer to these eight classes. G-equivalent but not T-equivalent H-edges will have the same letter, but different numbers. The numbering used here corresponds to the numbering of symmetry operations of *Pmmn*, origin choice 2 (IT #59, p. 282), after the coordinate transformation $(x, y, z)\mapsto(z, x, y)$.

Each element $g \in G$ maps classes of T-equivalent objects (vertices, H-edges, molecules *etc.*) onto classes of

Table 5. *Symmetry operations of the group G within a unit cell of a layer of* ***ACNACO*** *and the corresponding permutations of the classes of translational equivalent positions*

Layer group *Pm*(*n*)*m*, origin choice 2.

No. of position	Symmetry operation $x, y, z\mapsto$	Induced permutation on the classes of translationally equivalent positions
1	x, y, z	Identity
2	$\overline{x}+\frac{1}{2}, y, \overline{z}+\frac{1}{2}$	(12)(34)(56)(78)
3	$x+\frac{1}{2}, \overline{y}, \overline{z}$	(13)(24)(57)(68)
4	$\overline{x}, \overline{y}, z+\frac{1}{2}$	(14)(23)(58)(67)
5	$\overline{x}, \overline{y}, \overline{z}$	(15)(26)(37)(48)
6	$x+\frac{1}{2}, \overline{y}, z+\frac{1}{2}$	(16)(25)(38)(47)
7	$\overline{x}+\frac{1}{2}, y, z$	(17)(28)(35)(46)
8	$x, y, \overline{z}+\frac{1}{2}$	(18)(27)(36)(45)

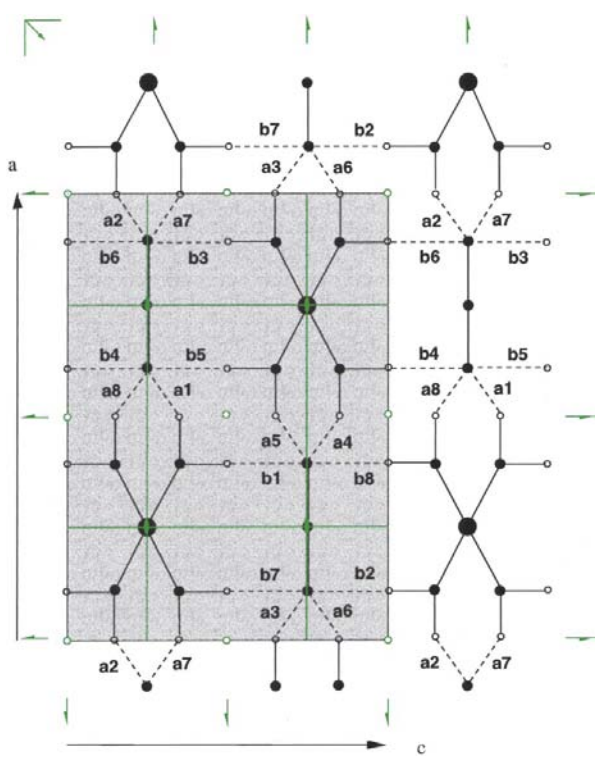

Fig. 7. Part of the G-array for **ACNACO** showing T-labels and symmetry elements. The shaded area indicates a unit cell with respect to *Pm*(*n*)*m*, origin at $\bar{1}$ at 2_1n2_1 at $-\frac{1}{4}$, 0, $-\frac{1}{4}$ from *m*2*m*. Atoms which are not relevant for graph sets are not drawn.

† T-equivalence is *finer* than G-equivalence in the sense that two paths (vertices, edges, molecules) which are T-equivalent, are also G-equivalent, but not necessarily *vice versa*. Or in more formal terms: each T-equivalence class of objects is fully contained in some G-equivalence class of the same objects.

‡ The difference between space group *Pmmn* and layer group *Pm*(*n*)*m* is that the translation group of the first is three-dimensional, while the translation group of the latter is two-dimensional.

Table 6. *Covalent distance tables for* ***ACNACO***

Cation $[Co(NO_2)_2(NH_3)_4]^+$} at special position $(\frac{1}{4}, y, \frac{1}{4})$

cd	$\vec{a1}$	$\vec{a2}$	$\vec{a7}$	$\vec{a8}$	$\vec{b1}$	$\vec{b2}$	$\vec{b7}$	$\vec{b8}$
$\overleftarrow{a1}$	0	4	4	4	2	4	4	4
$\overleftarrow{a2}$	4	0	4	4	4	2	4	4
$\overleftarrow{a7}$	4	4	0	4	4	4	2	4
$\overleftarrow{a8}$	4	4	4	0	4	4	4	2
$\overleftarrow{b1}$	2	4	4	4	0	4	4	4
$\overleftarrow{b2}$	4	2	4	4	4	0	4	4
$\overleftarrow{b7}$	4	4	2	4	4	4	0	4
$\overleftarrow{b8}$	4	4	4	2	4	4	4	0

Anion $[C_2H_3O_2]^-$ at special position $(\frac{3}{4}, y, \frac{1}{4})$

cd	$\overleftarrow{a1}$	$\overleftarrow{a2}$	$\overleftarrow{a7}$	$\overleftarrow{a8}$	$\overleftarrow{b3}$	$\overleftarrow{b4}$	$\overleftarrow{b5}$	$\overleftarrow{b6}$
$\vec{a1}$	0	2	2	0	2	0	0	2
$\vec{a2}$	2	0	0	2	0	2	2	0
$\vec{a7}$	2	0	0	2	0	2	2	0
$\vec{a8}$	0	2	2	0	2	0	0	2
$\vec{b3}$	2	0	0	2	0	2	2	0
$\vec{b4}$	0	2	2	0	2	0	0	2
$\vec{b5}$	0	2	2	0	2	0	0	2
$\vec{b6}$	2	0	0	2	0	2	2	0

Cation $[Co(NO_2)_2(NH_3)_4]^+$ at special position $(\frac{3}{4}, \overline{y}, \frac{3}{4})$

cd	$\vec{a3}$	$\vec{a4}$	$\vec{a5}$	$\vec{a6}$	$\vec{b3}$	$\vec{b4}$	$\vec{b5}$	$\vec{b6}$
$\overleftarrow{a3}$	0	4	4	4	2	4	4	4
$\overleftarrow{a4}$	4	0	4	4	4	2	4	4
$\overleftarrow{a5}$	4	4	0	4	4	4	2	4
$\overleftarrow{a6}$	4	4	4	0	4	4	4	2
$\overleftarrow{b3}$	2	4	4	4	0	4	4	4
$\overleftarrow{b4}$	4	2	4	4	4	0	4	4
$\overleftarrow{b5}$	4	4	2	4	4	4	0	4
$\overleftarrow{b6}$	4	4	4	2	4	4	4	0

Anion $[C_2H_3O_2]^-$ at special position $(\frac{1}{4}, \overline{y}, \frac{3}{4})$.

cd	$\overleftarrow{a3}$	$\overleftarrow{a4}$	$\overleftarrow{a5}$	$\overleftarrow{a6}$	$\overleftarrow{b1}$	$\overleftarrow{b2}$	$\overleftarrow{b7}$	$\overleftarrow{b8}$
$\vec{a3}$	0	2	2	0	2	0	0	2
$\vec{a4}$	2	0	0	2	0	2	2	0
$\vec{a7}$	2	0	0	2	0	2	2	0
$\vec{a6}$	0	2	2	0	2	0	0	2
$\vec{b1}$	2	0	0	2	0	2	2	0
$\vec{b2}$	0	2	2	0	2	0	0	2
$\vec{b7}$	0	2	2	0	2	0	0	2
$\vec{b8}$	2	0	0	2	0	2	2	0

T-equivalent objects. Thus, the group G permutes the eight indices $1, \ldots, 8$ as shown in Table 5.

Consider an arbitrary path P in the G-array and the corresponding path in the constructor graph. Let us assume that its arrowed label sequence is, for instance, say $\vec{a1}\overleftarrow{a7}\vec{a1}\overleftarrow{a7}$. Then, according to Table 5, there is a path which is G-equivalent to P and has an arrowed label sequence that we obtain from this sequence by exchanging numbers according to one of the permutations in the third column in Table 5. Thus,

$\vec{a1}\overleftarrow{a7}\vec{a1}\overleftarrow{a7}$, $\vec{a4}\overleftarrow{a6}\vec{a4}\overleftarrow{a6}$, $\vec{a3}\overleftarrow{a5}\vec{a3}\overleftarrow{a5}$, $\vec{a8}\overleftarrow{a2}\vec{a8}\overleftarrow{a2}$ *etc.*

all belong to G-equivalent paths.

5.3. *Graph sets of* ***ACNACO***

We are now able to remove the ambiguities mentioned in the previous subsections.

Since now (after the label refinement) all H-edges emanating from a molecule and all H-edges entering a molecule have different labels, we are able to establish the covalent distance tables of the two non-equivalent molcules of **ACNACO**. The result is shown in Table 6.

There are eight molecules – four cations and four anions – in a unit cell of *Cmcm*. There are four classes of translationally non-equivalent molecules and there are only two classes of crystallographically non-equivalent molecules. In principle, it is sufficient to have two covalent distance tables (one for the cation and one for the anion) and to generate the remaining two tables applying *e.g.* permutation (13)(24)(57)(68) to the labels (which is induced by the twofold screw rotation $(x, y, z) \mapsto (x + \frac{1}{2}, \overline{y}, \overline{z}$, # 3 in Table 5).

Let us look now for the first-level graph sets of **ACNACO**.

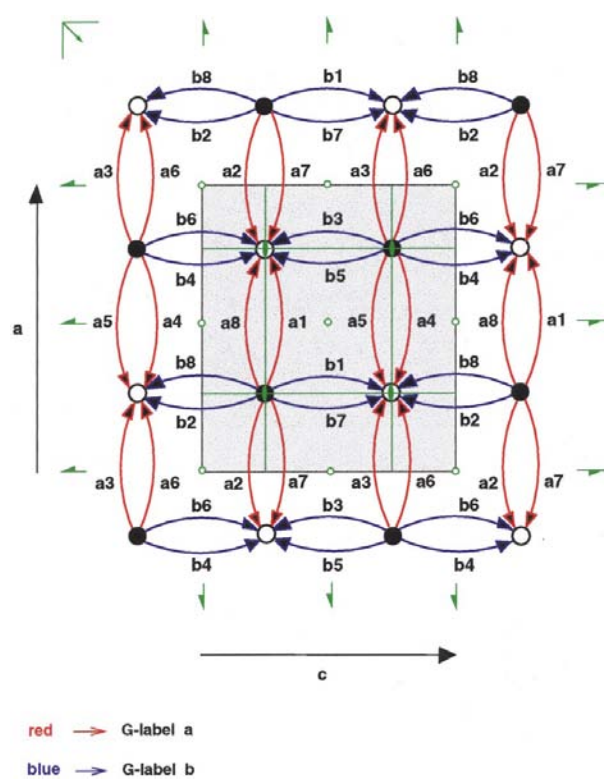

Fig. 8. The constructor graph of **ACNACO**. Cations are represented by full circles, anions by empty circles. The unit cell and origin are chosen in accordance with Fig. 7.

† Note that for this property of G it is essential that T is a normal subgroup, *i.e.* that gtg^{-1} is again a translation, for all t from T and all g from G.

Table 7. *Some of the first- and second-level graph sets of* ***ACNACO***

L	**a**	**b**
a	$\mathbf{R}_2^1(6)$, $\mathbf{C}_2^2(8)$	
b	$\mathbf{R}_4^2(8)$	$\mathbf{R}_2^2(8)$, $\mathbf{C}_2^1(6)$

Since there are no intramolecular hydrogen bonds, there are no selfs. There are also no discretes in the crystal structure of **ACNACO**, since each molecule in the *G*-array is incident with eight H-edges, four with *G*-label **a** and four with with *G*-label **b**.

Let us use the constructor graph for finding chains and rings. To find all existing motifs (first-level graph sets) we have to restrict the search to objects with qualitative descriptors which either contain only symbols **a1**, **a2**, ..., **a8** or only symbols **b1**, **b2**, ..., **b8**.

In the constructor graph in Fig. 8 we recognize the rings

$$\mathbf{R}(\overrightarrow{\mathbf{a1}}\overleftarrow{\mathbf{a8}}),\ \mathbf{R}(\overrightarrow{\mathbf{a2}}\overleftarrow{\mathbf{a7}}),\ \mathbf{R}(\overrightarrow{\mathbf{a3}}\overleftarrow{\mathbf{a6}}),\ \mathbf{R}(\overrightarrow{\mathbf{a4}}\overleftarrow{\mathbf{a5}}),$$

which are all *G*-equivalent. Using the covalent distance tables we find

$$n = 0 + 4 + 2 = 6,\ \ d = 2,\ \ a = 2 - 1 = 1.$$

Hence the quantitative descriptor is $\mathbf{R}_2^1(6)$.

Similarly, we find the *G*-equivalent rings

$$\mathbf{R}(\overrightarrow{\mathbf{b1}}\overleftarrow{\mathbf{b7}}),\ \mathbf{R}(\overrightarrow{\mathbf{b2}}\overleftarrow{\mathbf{b8}}),\ \mathbf{R}(\overrightarrow{\mathbf{b3}}\overleftarrow{\mathbf{b5}}),\ \mathbf{R}(\overrightarrow{\mathbf{b4}}\overleftarrow{\mathbf{b6}}),$$

and their quantitative descriptor $\mathbf{R}_2^2(8)$.

In the constructor graph we also find the chains

$$\mathbf{C}(\overrightarrow{\mathbf{a1}}\overleftarrow{\mathbf{a7}}),\ \mathbf{C}(\overrightarrow{\mathbf{a1}}\overleftarrow{\mathbf{a7}}|\overrightarrow{\mathbf{a8}}\overleftarrow{\mathbf{a2}}).$$

Both have *G*-period $\overrightarrow{\mathbf{a1}}\overleftarrow{\mathbf{a7}}$. In the second chain, this is indicated by inserting | after the *G*-period. For the first chain the *G*-period and *T*-period coincide; for the second chain they do not. Again using the covalent distance tables we find for both chains the quantitative descriptor $\mathbf{C}_2^2(8)$. Further, there are the chains

$$\mathbf{C}(\overrightarrow{\mathbf{b1}}\overleftarrow{\mathbf{b8}}),\ \mathbf{C}(\overrightarrow{\mathbf{b1}}\overleftarrow{\mathbf{b8}}|\overrightarrow{\mathbf{b7}}\overleftarrow{\mathbf{b2}})$$

with equal *G*-periods $\overrightarrow{\mathbf{b1}}\overleftarrow{\mathbf{b8}}$ and, therefore, equal quantitative descriptors $\mathbf{C}_2^1(6)$. Again, | indicates the end of the *G*-period. The net result is that there are two non-*G*-equivalent chains of the same type with the same quantitative, but different qualitative descriptor.

For an example of a second-level graph set consider once more the constructor graph in Fig. 8. This contains the ring $\mathbf{R}(\overrightarrow{\mathbf{a1}}\overleftarrow{\mathbf{b5}}\overrightarrow{\mathbf{a4}}\overleftarrow{\mathbf{b1}})$. *G*-equivalent rings are

$$\mathbf{R}(\overrightarrow{\mathbf{a2}}\overleftarrow{\mathbf{b6}}\overrightarrow{\mathbf{a6}}\overleftarrow{\mathbf{b2}}),\ \mathbf{R}(\overrightarrow{\mathbf{a3}}\overleftarrow{\mathbf{b7}}\overrightarrow{\mathbf{a7}}\overleftarrow{\mathbf{b3}}),\ \mathbf{R}(\overrightarrow{\mathbf{a4}}\overleftarrow{\mathbf{b8}}\overrightarrow{\mathbf{a8}}\overleftarrow{\mathbf{b4}})\ \textit{etc}.$$

For its quantitative descriptor we find $\mathbf{R}_4^2(8)$.

The graph sets found so far are listed in Table 7.

5.4. *Remarks*

(*a*) The sequence of arrowed *T*-labels of a path in the constructor graph describes this path uniquely up to *T*-equivalence of its H-edges. Therefore, we can use this sequence as a descriptor of this path and speak, for instance, of the path $\overrightarrow{\mathbf{b1}}\overleftarrow{\mathbf{a3}}\overrightarrow{\mathbf{a5}}\overleftarrow{\mathbf{b8}}$, the path $\overrightarrow{\mathbf{a3}}\overleftrightarrow{\mathbf{b1}}\mathbf{a1}$, and so on. Implicitly, a sequence of arrowed *T*-labels also determines a path in the *G*-array uniquely up to *T*-equivalence and up to the choice of shortest paths through molecules.† However, different sequences of arrowed *T*-labels can belong to *G*-equivalent paths, as we have observed several times in the last subsection. To keep *G*-equivalence recognizable the *T*-labeled constructor graph is not sufficient. In addition, we need the information on how *T*-labels change under the symmetry operations *g*, which are not translations. For this reason we have prepared Table 5, which shows for the crystal structure of **ACNACO** how elements of the space group *G* permute the orbits of the translation group. A similar table would be required in other cases. This problem does not arise when all molecules are in a general position and, therefore, *G*-equivalence is sufficient to distinguish all H-edges incident with a molecule, in which case there is no need for a label refinement.

(*b*) Each infinite periodic path in the constructor graph corresponds to a infinite periodic path in the *G*-array (and *vice versa*) and, hence, determines a chain. The period of this chain under the group *G* may be shorter than its period under the group *T*. This was the case, for instance, with the chain $\mathbf{C}(\overrightarrow{\mathbf{a1}}\overleftarrow{\mathbf{a7}}|\overrightarrow{\mathbf{a8}}\overleftarrow{\mathbf{a2}})$ found in *ACNACO*. The true period with respect to *G* has to be found on the basis of the information on how *T*-labels change under operations $g \in G$ (Table 5). Differences between G-periods and *T*-periods, as seen in the crystal structure of **ACNACO**, are not observed in the crystal structures of the **IMDA** polymorphs. The reason for this phenomenon is that in **ACNACO** we are dealing with molecules (cation and anion) with non-trivial and even high molecular symmetry and crystallographic site symmetry: *m*(2)*m* – group of four elements.

(*c*) Rings in the constructor graph correspond to rings in the *G*-array (and *vice versa*). Discretes can be observed in the constructor graph as well. Selfs in the *G*-array correspond to loops in the constructor graph. However, for selfs we do not need either the constructor graph or the *G*-array.

6. Concluding remarks

Although they incorporated the fundamental philosophy and concepts of graph theory, Etter's original applications to hydrogen-bonding patterns and subse-

† Note that the shortest paths through molecules between atoms belonging to H-edges need not be unique. For graph-set analysis this is not relevant: graph sets are distinguished only by equivalence or non-equivalence of their sets of H-edges.

quent developments relied more on chemical and structural concepts than on mathematical principles. In this paper we have presented some basic mathematical concepts for assigning and treating graph sets to patterns of hydrogen bonds. For the most part the original definitions and notions are entirely consistent with these mathematical principles. The development of the mathematical tools required the definition of some additional concepts, among them the ideas of

a *quantitative descriptor* to describe the graph set,

a *G-array* to describe the interplay between the crystallographic space group and the hydrogen-bond patterns,

directed hydrogen edges to specify the directional properties of hydrogen bonds,

the *constructor graph* to obtain a general view over the existing hydrogen-bond patterns and thus to help summarize numerically the number and directionality of the bonds (edges) involved in a pattern,

a *qualitative descriptor* to summarize the types and directionality of combinations of hydrogen-bond patterns,

the *covalent distance tables* to derive quantitative descriptors from qualitative ones and

the *arrowed T-labeling* to handle molecules lying on crystallographic symmetry elements.

These mathematical tools may be readily incorporated into software algorithms for detecting, presenting and analysing hydrogen-bond patterns, as has been described by Motherwell *et al.* (1999) in the accompanying paper and demonstrated both here briefly and there in more detail.

In this paper the mathematical terms and definitions have been couched very much in crystallographic language, in order to facilitate crystallographers' familiarity and the eventual use of these ideas. The tools are there, but for the 'casual user' they can remain transparent, as they are for users of the CSD software (Motherwell *et al.*, 1999). While the initial treatment has been for hydrogen-bonded systems, the mathematical basis is sufficiently broad and the tools are sufficiently versatile that virtually any other pattern-forming interaction may be treated with the same formalism. There are still many chemical crystallographic and mathematical questions and challenges in defining and interpreting hydrogen-bond patterns using graph theory. Among these are the proper treatment of ladders and branched chains, the definition of homodromic and heterodromic patterns, the detailed treatment of structures in which molecules lie on crystallographic symmetry elements, which means in positions with non-trivial site symmetry, prioritizing 'degenerate' patterns (*e.g.* those with the same degree or number of edges) *etc*. We hope that the foundations laid here will serve in meeting those challenges and answering those questions.

This work was supported by a grant from the German–Israel Foundation for Scientific Research and Development, which is gratefully acknowledged. We are also grateful for technical assistance from Arkady Ellern and Oshrit Navon at Ben-Gurion University of the Negev (BGU, Israel), and for many stimulating discussions with Reinhard Pöschel (Technical University Dresden, Germany), Mikhail Klin (BGU) and Jan-Olav Henck (University of Innsbruck, Austria). JG wants to thank the coworkers of the Chemistry and Mathematics Departments of the BGU where she spent 11 months of a study visit to Israel, and the Computing Center of the Chemistry Department of the Freie Universität Berlin (Germany) for using the facilities of a CSD access. JB wishes to thank the members of the Cambridge Crystallographic Data Center for their warm hospitality and collegiality for a sabbatical visit during which portions of this work were carried out.

References

Allen, F. H. & Kennard, O. (1993). *Chem. Des. Autom. News*, **8**, 1, 31–37.

Anon (1997). *IUCr Newsl.* **5**, 14.

Bernstein, J. (1979). *Acta Cryst.* B**35**, 360–366.

Bernstein, J., Davis, R. E., Shimoni, L. & Chang, N. (1995). *Angew. Chem. Int. Ed. Engl.* **34**, 1555–1573.

Bernstein, J., Etter, M. & MacDonald, J. C. (1990). *J. Chem. Soc. Perkin Trans. 2*, pp. 695–698.

Bernstein, J., Ganter, B., Grell, J., Hengst, U., Kuske, K. & Pöschel, R. (1997). Report Math-AL-17-1997. Technische Universität Dresden, Germany.

Boman, C.-E., Herbertsson, H. & Oskarsson, Å. (1974). *Acta Cryst.* B**30**, 378–382.

Cagnon, C., Matalon, J. R. & Beauchamp, A. L. (1978). *Acta Cryst.* B**34**, 1128–1130.

Etter, M. C. (1990). *Acc. Chem. Res.* **23**, 120–126.

Etter, M. C., MacDonald, J. C. & Bernstein, J. (1990). *Acta Cryst.* B**46**, 256–262.

Grell, H., Krause, Ch. & Grell, J. (1988). Report No. 2. Institut für Informatik und Rechentechnik, Akademie der Wissenschaften der DDR, Germany.

Harary, F. (1967). *Graph Theory and Theoretical Physics*, edited by F. Harary. New York: Academic Press.

Merrifield, R. E. & Simmons, H. E. (1989). *Topological Methods in Chemistry*. New York: Wiley.

Motherwell, W. D. S., Shields, G. & Allen, F. H. (1999). *Acta Cryst.* B**55**, 1044–1056.

Weber, L. (1929). *Z. Kristallogr.* **70**, 309–327.