Acta Cryst. (1999). B**55**, 1044–1056

Visualization and characterization of non-covalent networks in molecular crystals: automated assignment of graph-set descriptors for asymmetric molecules

W. D. Samuel Motherwell,* Gregory P. Shields and Frank H. Allen

Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, England.
E-mail: motherwell@ccdc.cam.ac.uk

(*Received 17 December 1998; accepted 7 May 1999*)

Abstract

A method of visualizing intermolecular networks (for example, hydrogen-bonded networks) in the crystalline state has been developed, based on the concept of link atoms, *i.e.* those atoms deemed to be in contact with each unique molecule or ion in the crystal chemical unit (CCU). Extension of a structure using each of these primary links can be achieved, enabling the generation and investigation of extended networks. Algorithms have been developed for the automatic assignment of graph-set notation for patterns up to second level, *i.e.* those involving one or two crystallographically independent non-covalent bonds, in the absence of internal crystallographic symmetry in the unique molecules of the CCU. The self, ring, chain and discrete motifs may be displayed by highlighting the atoms and bonds comprising the pattern. These methodologies have been implemented in the Cambridge Structural Database program *PLUTO*.

1. Introduction

It is well known that the robust, hence reproducible, intermolecular motifs found in organic systems (Jeffrey, 1997) can be used to direct the synthesis of supramolecular complexes, *e.g.* in crystal engineering (Panuto *et al.*, 1987; Etter & Frankenbach, 1989; Etter, 1991; Jones *et al.*, 1996; Garcia-Tellado *et al.*, 1991; Bernstein *et al.*, 1994; Desiraju, 1995; Aakeröy & Seddon, 1993). The conceptual relationship between crystal engineering and organic synthesis has led to the term *supramolecular synthon* (Desiraju, 1995) being proposed for structure-directing motifs involving non-covalent bonds. A number of early studies explored the nature of hydrogen-bonded motifs in classes of compounds with particular functional groups, *e.g.* carbohydrates (Jeffrey & Takagi, 1978), carboxylic acids (Leiserowitz & Schmidt, 1969) and amides (Leiserowitz, 1976; Leiserowitz & Tuval, 1978). Similar motifs in inorganic systems have been recognized previously by Wells (1962).

In order to provide a systematic notation for the topology of hydrogen-bonded motifs and networks, a graph-set approach has been suggested (Kuleshova & Zorkii, 1980; Etter *et al.*, 1990; Etter, 1990; Bernstein *et al.*, 1995). This provides a description of hydrogen-bonding patterns in terms of chains (**C**), rings (**R**), discrete complexes (**D**) and intramolecular (self-associating) rings (**S**). The degree of the pattern (n, the number of atoms comprising the pattern), together with the number of donors (d) and the number of acceptors (a), are combined to form the *quantitative* graph-set descriptor $X_d^a(n)$ (Bernstein *et al.*, 1995). For convenience, the alternative notation $Xa,d(n)$ has been adopted in this paper.

This descriptor does not distinguish between patterns of the same degree that have different numbers of bonds in the covalent fragments comprising the pattern, nor does it differentiate alternative arrangements of the donors and acceptors. Furthermore, it is necessary to assume that the H atoms are not disordered between donor and acceptor sites, *i.e.* $X{-}H\cdots Y$ *vs* $X\cdots H{-}Y$, and where disorder does occur it is necessary to make an arbitrary donor and acceptor assignment. Despite such limitations, these purely topological descriptors have proved useful in decoding differences between the packing arrangements adopted in polymorphic systems, *e.g.* in L-glutamic acid (Bernstein, 1991) and iminodiacetic acid (Bernstein *et al.*, 1990, 1995). The graph-set nomenclature provides a basic description of hydrogen-bonded synthons and can aid the identification of preferred motifs. Recently a systematic general search for $R_2^2(8)$ motifs has been performed, to explore the chemical diversity of the functional groups which adopt this topology (Bernstein & Davis, 1999). It should be noted that such graph sets cannot describe three-dimensional structure, *e.g.* network interpenetration or the geometric disposition of graph-set patterns in a structure.

Crystallographically equivalent hydrogen bonds are considered as being a *type* of hydrogen bond and each type may be identified by a convenient label (**a**, **b**, **c** *etc.*). Patterns are distinguished on the basis of their *level*, the first level being the sets {**a**}, {**b**} *etc.* of patterns formed by one type of hydrogen bond, the second level sets involving two types of hydrogen bond {**a**,**b**} and similarly for higher levels. Chemically equivalent patterns may become apparent at different levels depending on the presence or otherwise of crystallographic symmetry.

More recently, the mathematical basis of this type of analysis has been established, placing the procedure on a rigorous graph-theoretical (Harary, 1969) footing. In this approach (Bernstein *et al.*, 1997; Grell *et al.*, 1999), the extended crystal structure (*e.g.* benzamide, Fig. 1*a*) is considered as an *array* of vertices (atoms) linked by *covalent edges* and *hydrogen (bond) edges* (H-edges), the latter being labelled according to their crystallographic equivalence as previously. Networks may be described more clearly using a *constructor graph*, in which the molecules have been collapsed to points and only the H-edges are shown (Fig. 1*b*).

Paths in the constructor graph may be described by a *directed label sequence*, where the direction of the H-edges ($D > A$ or $A > D$) is indicated in the vector notation $\overrightarrow{a}$ or $\overleftarrow{a}$ respectively, *e.g.* $\overrightarrow{a}\,\overrightarrow{a}\,\overrightarrow{b}\,\overleftarrow{a}\,\overleftarrow{a}\,\overleftarrow{b}$. If there is only one H-edge from each molecule (vertex in the constructor graph) with the same directed label (*i.e.* there are no crystallographically equivalent hydrogen bonds which both start from a donor or both start from an acceptor in the same molecule), then the path is defined uniquely by the directed label sequence (*i.e.* in mathematical terms it is a *significant* labelling). In practice, this condition is obeyed where no donor or acceptor atoms or molecules lie on crystallographic special positions and H-edges emanating from the same donor or acceptor (*e.g.* for three- or four-centre hydrogen bonding) then have distinct labels.

These directed label sequences may be combined with a designator *S*, *D*, *R* or *C* to form a *qualitative descriptor*, *e.g.* $C(\overrightarrow{a}\,\overrightarrow{a}\,\overrightarrow{b}\,\overleftarrow{a}\,\overleftarrow{a}\,\overleftarrow{b})$. Information on the internal molecular structure (the covalent edges), which is needed for deriving the path length and number of donors and acceptors for the quantitative descriptors, is contained in the *covalent distance table*, as the shortest covalent path length between all pairs of donor and acceptor atoms within molecules. These new concepts are described in detail in the accompanying paper by Grell *et al.* (1999) and applied to three polymorphs of iminodiacetic acid for illustration.

The graph sets $\{F\}$ do not necessarily have a finite number of *representants* (patterns which are members of the set), although they may have in special cases (*e.g.* where $\{F\}$ is an empty set). For example, if graph set representants $(\overrightarrow{a}\,\overrightarrow{b})$ and $(\overrightarrow{a}\,\overleftarrow{b})$ exist, $(\overrightarrow{a}_m\overrightarrow{b}_n)$ and $(\overrightarrow{a}_p\overleftarrow{b}_q)$ [where $m, n, p, q = 1, 2, 3. .. , e.g.$ $(\overrightarrow{a}\,\overrightarrow{b}\,\overrightarrow{a}\,\overleftarrow{b}\,\overleftarrow{b})$ for $p = 3, q = 2$] are also members of the graph set {**ab**}. It is usually convenient to select a subset of $\{F\}$ which obeys additional conditions. For example, *M*-simple paths are those which pass through a molecule M_i at most once: Bernstein *et al.* (1997) suggested that only representants with this property should be considered.

Previously, the assignment of graph-set descriptors was undertaken by visual inspection of the intermolecular network, a process that is prone to human error. We present here a methodology for making the assignment in an automated, systematic manner, which should enable consistent descriptors to be generated. The methodology has been implemented in the program *PLUTO* and we give examples of its use.

2. Methodology

2.1. *The crystal chemical unit*

The starting point for investigation of intermolecular networks in molecular crystal structures is the crystal chemical unit (CCU), which contains the *complete* discrete molecule(s) and ion(s) that comprise the crystal structure (Allen *et al.*, 1974). It is synonymous with the asymmetric unit, except where molecular and crystallographic symmetry coincide: here, any fractional

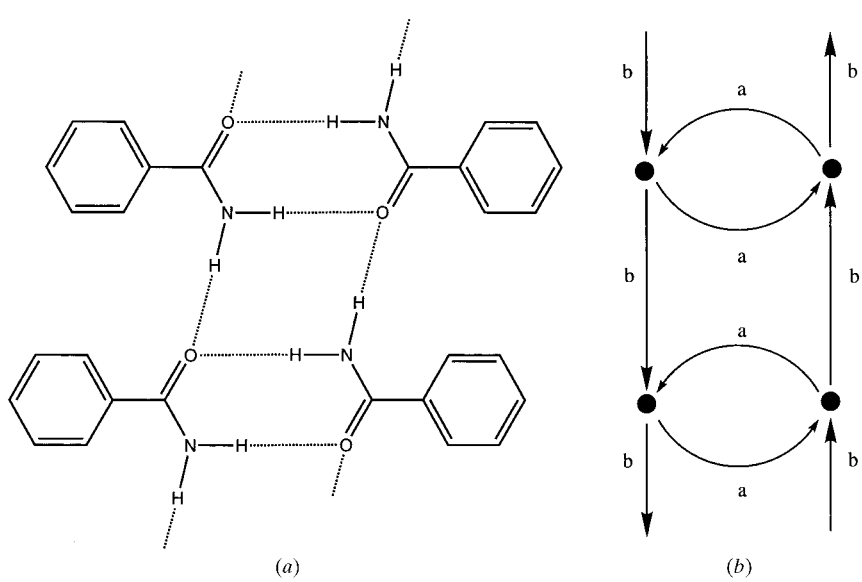

Fig. 1. (*a*) Conventional chemical diagram and (*b*) constructor graph representations of the hydrogen-bond network in the crystal structure of benzamide.

molecules are expanded by symmetry to form the CCU prior to investigation of intermolecular interactions. Each atom and bond is considered as being a part of one and only one of the unique molecules. The centroids of these unique molecules do not necessarily lie in the range (0−1,0−1,0−1) on the crystal axes in published crystal structure determinations, and as stored in the Cambridge Structural Database (CSD: Allen & Kennard, 1993): it may be convenient to transform the coordinates of the original molecules to achieve this. Each atom in the unique molecule(s) is characterized by an original atom number and the symmetry operator relating it to the original atom from which it is generated. This relationship may be expressed conveniently as a packed code of the type first used in the *ORTEP* program (Johnson, 1965), where 2456 represents a transformation by symmetry operator 2 and translations of $x = -1$, $y = 0$, $z = 1$. Where no original molecules possess internal crystallographic symmetry, all the atoms in the original molecules will be original atoms having identity symmetry operators. If a symmetry-generated atom lies on a special position, the symmetry operator relating it to its original position is not unique; in this case, an arbitrary choice of symmetry operator may be made.

2.2. *Display of symmetry relationships within and between molecules*

Understanding of packing diagrams may be enhanced by colour-coding the atoms and bonds according to the type of symmetry operator generating the atom from the corresponding atom in the crystallographic asymmetric unit. Alternatively, each molecule can be labelled with a symbol to display the generating operator. A colour key classifies broad symmetry-operator types as follows: *T* translation; *I* inversion centre; *G* glide plane; *M* mirror plane; *nS n*-fold screw axis; *nR n*-fold rotation axis; *nI n*-fold rotation/inversion axis. Fig. 2 shows molecules in the *xz* plane in the crystal structure of benzamide coloured in this manner.

For this purpose the symmetry operations of the space group alone are considered and additional translations are neglected. This maintains translational periodicity of the colour scheme in packing diagrams and allows internal crystallographic symmetry in molecules, and symmetry relationships in packing diagrams, to be investigated. Where the molecules possess atoms lying on special positions, the colouring of some atoms will not be unique. The type of symmetry operator is identified by consideration of the trace and determinant of their rotation matrices and the cumulative effect of their inherent translational components. A detailed description of this methodology has been provided elsewhere (Fischer & Koch, 1983).

2.3. *Location of intermolecular bonds*

Intermolecular non-covalent contacts ('bonds') may be found efficiently by the algorithm described by Rollett (1965). Each pairwise atom combination is considered as a potential contact and all positions of one atom relative to another are investigated with axial translations of the acceptor by $(x,y,z) = (-3,-3,-3) \rightarrow$

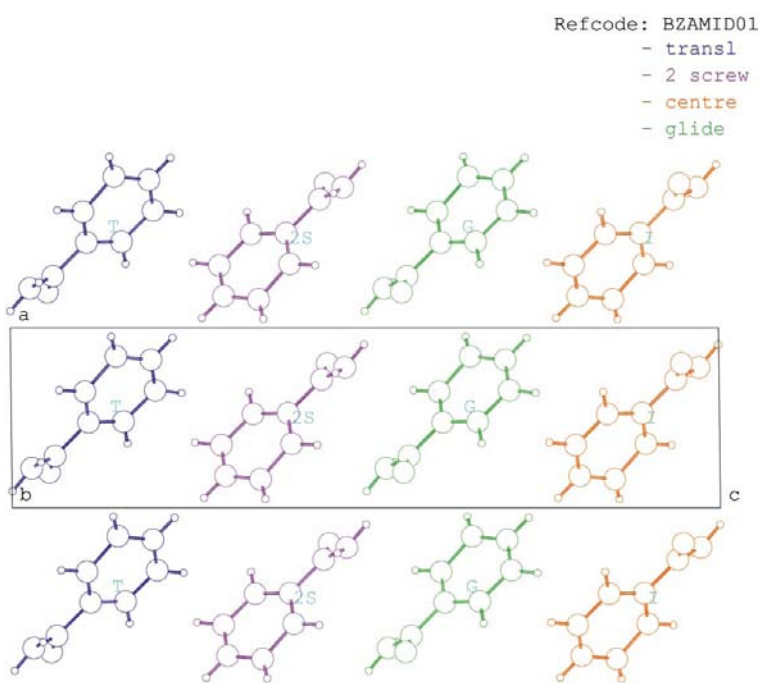

Fig. 2. Projection of the crystal structure of benzamide down the *y* axis, showing molecules coloured according to symmetry operator relating them to the original molecule, which belongs to the set labelled *T*.

(3,3,3), and under the operation of each of the symmetry elements of the space group. It is only necessary to evaluate the actual distance between the two atoms when products of the fractional coordinate displacements *dx*, *dy*, *dz* and the plane spacings *d*(100), *d*(010), *d*(001), respectively, are all less than the distance limit. This limit may be set either by: (*a*) a specific user – defined value for the atom types concerned, or (*b*) a van der Waals' radius criterion with a specified tolerance. The search may be restricted to particular element types as appropriate. Alternatively, the search can be restricted to contacts within a specified absolute distance of a particular atom (symmetry-related contacts are not added automatically, which has implications for the expansion of networks and the assignment of graph sets). Contacts to several atoms may be added cumulatively in the latter approach.

There are chemical situations where it is convenient to use dummy points rather than atoms for contact searching and a notional radius can be assigned to them. The facility may be used, for example, to define the centroid of a six-membered aromatic ring or the midpoint of an alkyne triple bond when studying D—H$\cdots\pi$ bonds.

The intermolecular contact search incorporated in the CSD system program *QUEST3D* (Allen & Kennard, 1993) has been refined for the specific case of hydrogen bonds, D—H$\cdots A$, and this modified version was used throughout this work. Potential hydrogen-bond donors (D) and acceptors (A) are first identified on the basis of atom type. Since the nature of the substituents on the D or A atoms is important, a more specific Mol2 atom-type designation (Clark *et al.*, 1989) was adopted, which describes the chemical environment of the atom as well as the element type. This permits more specific searches to be made and eliminates some potential contacts prior to the non-bonded search itself (*e.g.* quaternary nitrogen may not function as a hydrogen-bond acceptor). The atom types for the elements most relevant to hydrogen-bonding studies are defined below with Mol2 atom types in parentheses.

Carbon: sp^3 (C.3); sp^2 (C.2); *sp* (C.1); aromatic (C.ar); cationic $[C(NR_2)_3]^+$ (C.cat);

Nitrogen: sp^3 trigonal pyramidal (N.3); sp^2 trigonal planar (N.pl3); sp^3 tetrahedral [cationic] (N.4); sp^2 two-coordinate (N.2); *sp*(N.1); aromatic (N.ar); amide (N.am);

Oxygen: sp^3 (O.3); sp^2 (O.2); carboxylate/phosphate O (O.co2);

Sulfur: sp^3 (S.3); sp^2 (S.2); sulfoxide (S.o); sulfone (S.o2);

Phosphorus: sp^3 (P.3).

Routines for evaluating Mol2 atom types were available within other CSD programs. A user-defined atom is taken as matching a donor or acceptor string in the contact definition if the characters typed for the atom definition match the corresponding leading characters in the assigned Mol2 atom type. This match is not case-sensitive and a period (.) is implied after a single-character element type. Thus N.a would match both the atom types N.ar and N.am, and O would match any oxygen (but not Os) *etc*. This provides the ability to search for both general and specific chemical interactions, *e.g.* aromatic C—H$\cdots$O=C contacts (D = C.ar, A = O.2).

For each donor–acceptor combination, distance limits are applied as described above. To enable systematic comparisons of hydrogen-bond geometry, the hydrogen positions should be neutron-normalized (Jeffrey & Lewis, 1978), *i.e.* hydrogen is moved along the D—H vector to a position corresponding to the mean D—H bond length obtained in neutron studies (Allen *et al.*, 1987). A minimum allowed D—H$\cdots A$ angle may be defined for each contact type. In addition, intramolecular contacts may be identified, as in a conventional *QUEST3D* non-bonded contact search, and limits on the intramolecular bond-path length between the acceptor atom and the H donor atom can be imposed.

2.4. *Visualization of link atoms and link bonds*

Intermolecular contacts are represented graphically by link atoms and link bonds (*e.g.* represented by dashed lines). The link atoms associated with each molecule in the CCU are the atoms which are in contact with that molecule, as defined in a previous intermolecular contact search, and the link bonds represent the contact vectors themselves. These link atoms are considered as being an extension of the molecule with which they are in contact, and thus a property of that molecule. Link atoms may coincide with original atoms in another molecule of the CCU, where unique molecules are in contact; however, the link and original atoms will not be associated with the same molecule. In order to identify the type of H atom in hydrogen-bond searches, the donor as well as the H atom itself may be added as part of a link group, bonded to the contact atom H itself by an intramolecular bond. These concepts are illustrated in Fig. 3.

It is necessary to avoid the addition of duplicate link atoms and bonds, which could arise if more than one atom in a unique molecule has a contact to the same atom in a symmetry-generated molecule, or where the unique molecule(s) contains symmetry-generated atoms. It is convenient to derive the symmetry operator $\mathbf{S}_{\rm link}$, which relates the link atom to an original atom in a unique molecule, in situations where the link has been derived by operator $\mathbf{S}_j$ from an atom in a unique molecule symmetry-generated by operator $\mathbf{S}_i$, by combining the two operators $\mathbf{S}_i$ and $\mathbf{S}_j$. A potential link atom may then be compared with those previously found. If the two link atoms are associated with the same molecule and derived from the same original atom, but the symmetry operators differ, it is necessary to also

compare the fractional coordinates since the symmetry operators are not unique for atoms which lie on special positions. Applying the $\mathbf{S}_{\rm link}$ for the potential link atom to each of these, and comparison of each result with the operators for the existing link atoms, allows duplicate links to be identified. Similarly, link bonds are considered as being equivalent where the atoms at each end of the contact are derived from the same original atom, are the property of the same molecule and have equal coordinates within a suitable tolerance (*e.g.* dx, dy, $dz < 0.0001$ Å).

Each crystallographically independent contact $A\cdots B$ is added at least twice, starting from both A and B in the unique molecules, and from any atoms A^* and B^* symmetry-related to A and B in the unique molecules. This is useful when analysing a network to identify symmetry-equivalent contacts by colour and particularly when assigning graph sets. Contacts may be considered as being symmetry-equivalent where the link atoms are derived from the same pair of original atoms and are separated by an intermolecular bond vector of equal length, within an appropriate tolerance. If desired, the contacts may be emphasized by increasing the radius of the intermolecular bond cylinders in a diagram, to provide a convenient means of displaying the network (Fig. 3).

2.5. *Network expansion*

Link atoms may be considered as growth points for network expansion. If the symmetry-generated molecule containing the link atom is added to the crystal structure diagram, the intermolecular network may be extended. This process may be repeated until the desired crystal structure unit (*e.g.* ring, chain, sheet) is displayed, since the link atoms associated with each symmetry-generated molecule are also added. The symmetry operators relating each atom and link atom in the additional molecule are calculated as the expansion proceeds. Each molecule is characterized by the number of the unique molecule from which it was generated and by the *ORTEP* code for the symmetry operation. In order to avoid duplicates being added where unique molecules lie on a symmetry element, the unit-weighted centroid of each unique molecule is calculated and transformed by the appropriate symmetry operator for each symmetry-generated molecule. Molecules are then considered to be equivalent if they are generated from the same unique molecule and have a common centroid. Duplicates are eliminated in this manner as a network is extended.

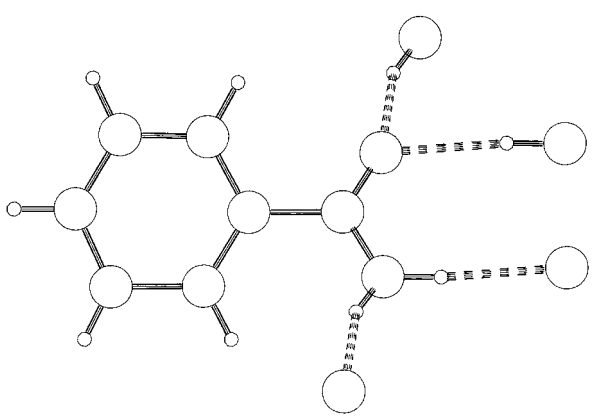

Fig. 3. Benzamide molecule with hydrogen-bond contacts shown as link groups (X—H units), link atoms and link bonds (dashed lines).

As the expansion proceeds, it is necessary to determine which link atoms remain active and which are now inactive, *i.e.* those for which the molecule overlapping the link atom has already been added. Link atoms are considered as being active provided there is not a non-link atom with the same original atom number (but the property of another molecule) in the same position. The relative initial positions of the unique molecules may be such that some link atoms are inactive, *i.e.* if the unique molecules have intermolecular contacts between them when in their original positions.

Network expansion may be started from the currently displayed molecules or from an individual molecule. Link atoms may be displayed with sequential link numbers (coloured cyan), only active link atoms (those at which new molecules may be added) being numbered. Expansion may proceed by selecting an individual link atom in the graphical display or by using all currently active links, enabling the network to be built up by adding a shell of molecules around the current set. This function may be restricted to links *to* molecules generated from a particular unique molecule. It is also possible to expand on chosen links from a packing diagram bounded by cell axial limits, which allows parallel non-intersecting chains to be investigated. Fig. 4 shows two non-intersecting chains of benzamide molecules produced in this way, the molecules being coloured according to symmetry operator.

The expansion process may be reversed, removing molecules stepwise until the point is reached at which the expansion was last (re)started. This allows alternative paths to be explored conveniently. Alternatively, unwanted molecules may be omitted from the plot simply by selecting any atom in the molecule: this may have the effect of re-activating some link atoms. If the expansion was started from the unique molecules, the expansion may be undone until the point is reached at which only the unique molecules remain. At the end of an expansion sequence, the link atoms and/or bonds may be removed from the display for presentation purposes.

2.6. *Graph-set assignment*

The first stage in graph-set analysis involves defining which atoms are considered as acceptors and which as

donors, since the treatment can clearly be extended to contacts other than hydrogen bonds (although for convenience contact vectors may still be called *H-edges*). It is not necessary for all intra- and intermolecular bonds established with the contact search to be considered in the graph-set analysis; furthermore, the graph-set terminology implicitly assumes that the interactions are directional, such that atoms may be either donors or acceptors but not both. Centroids may also be defined as being either donors or acceptors: they are considered as being bonded to the same atoms as the atoms defining the centroid for the purpose of deriving intramolecular path lengths.

Two possible approaches have been suggested for graph-set analysis. The first involves selecting a portion of the extended crystal structure which is sufficiently large to include all patterns up to a specified level and degree of complexity. The network is then analysed for intra- and intermolecular paths, which may be described with directed label sequences (*e.g.* $\overrightarrow{a}\ \overrightarrow{b}\ \overleftarrow{a}\ \overleftarrow{b}$). Typically, this has been performed by colouring symmetry-independent bonds and tracing possible paths on hardcopy plots (Bernstein *et al.*, 1995). The concept of a *constructor graph* (Grell *et al.*, 1999) is based on this approach. In an alternative approach only the unique molecules and their contacts are considered, the network being described fully by the application of the symmetry elements of the space group and the unit-cell translations. In this approach the directed label sequences are derived by an analysis of the effect of propagating the path through application of the symmetry operators relating molecules involved in the H-edges. The latter approach has been adopted here.

The assignment of graph sets is more straightforward where no atoms or molecules lie on a crystallographic symmetry element. In the simplest case of one unique molecule in the CCU, there are pairs of symmetry-related H-edges (link bonds) associated with the molecule and these may be distinguished by their directionality. Thus, each contact to the molecule (H-edge) has a unique *significant* label. Where there are contacts between different unique molecules, each molecule will have only one contact, $\overrightarrow{a}$ or $\overleftarrow{a}$, respectively. Intramolecular contacts may be considered as being both $\overrightarrow{a}$ and $\overleftarrow{a}$.

2.7. *First-level patterns*

The pattern designator (*S*, *D*, *C*, *R*) for a first-level representant of the graph set {**a**}, *i.e.* a pattern composed only of H-edges **a**, may be determined by considering the relationship between the donor and acceptor molecules connected by the H-edge **a**. Bernstein *et al.* (1997) have shown that for a significant labelling of H-edges, any H-edge **a** belongs to at least one first-level graph set {**a**}. Where the donor and acceptor are in molecules

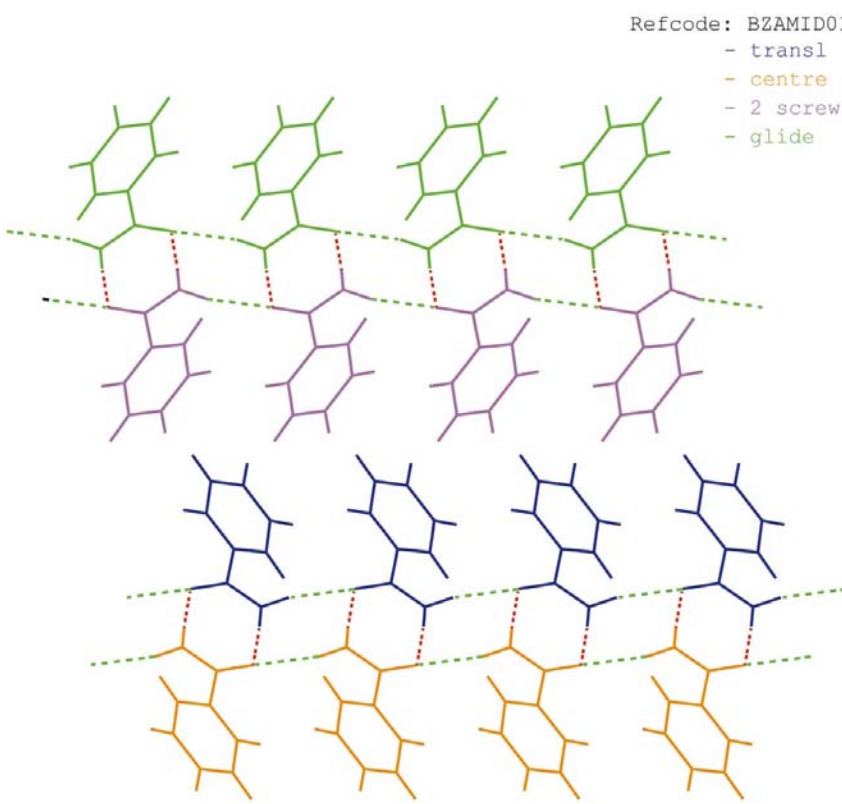

Fig. 4. Anti-parallel chains of benzamide dimers in the crystal structure, coloured according to symmetry operator.

derived from the same original molecule, analysis of the symmetry operator relating the donor and acceptor molecules (including the integral unit-cell translation component) indicates whether the motif is an infinite chain **C** (Fig. 5*a*), where the two molecules are related by a translational symmetry element (translation, glide plane, *n*-fold screw axis) or an intermolecular ring *R* (Fig. 5*b*) when the two molecules are related by a non-translational operator (inversion centre, *n*-fold rotation or inversion axis, mirror plane). The paths ($\overrightarrow{a}$) and ($\overleftarrow{a}$) in reverse directions have the same quantitative descriptors, since the path length is independent of the direction along which the entire path is followed, and thus only paths ($\overrightarrow{a}$) need be considered.

In terms of qualitative descriptors, the *M*-simple ring and chain representants (*i.e.* those describing paths passing through each molecule M_i at most once) with a path repeat unit of one H-edge which could occur in the set {**a**}, are of the form $R(\overrightarrow{a})_m$ (m = 2, 3, 4, 6) and $C(\overrightarrow{a})_\infty$. The paths are propagated by repeated application of the same symmetry operator of the crystallographic space group and the period is the intra- and intermolecular path length in the unit which, when propagated by repeated application of this symmetry operator, generates the extended path. Where no atom or molecule lies on a crystallographic special position, the H-edge repeat unit is apparent from the directed label sequence, *viz.* the shortest portion of the sequence which, when repeated, generates the whole path. In the case of rings, the *M*-simple path is finite and the multiplicity *m* of a ring motif (*i.e.* the number of repeat units, *periods*, in the cycle) is a finite integer. This may be indicated as a subscript to the qualitative descriptor as a convenient shorthand rather than the descriptor being expressed in full, *i.e.* in Fig. 5(*b*) $R(\overrightarrow{a}\,\overrightarrow{a}\,\overrightarrow{a}\,\overrightarrow{a}) \equiv R(\overrightarrow{a})_4$. The multiplicity *m* is equal to the multiplicity of the symmetry operator whose action on the path repeat unit (*e.g.* $\overrightarrow{a}$) generates the cyclic path (*e.g.* $\overrightarrow{a}\,\overrightarrow{a}\,\overrightarrow{a}\,\overrightarrow{a}$) and is thus restricted to (m = 2, 3, 4, 6) in crystals.

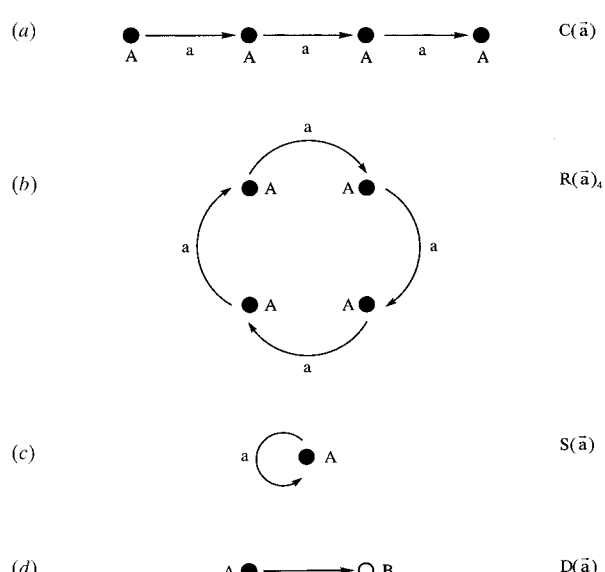

Fig. 5. Schematic constructor graphs for first-level graph-set patterns.

Where the acceptor and donor are in the same molecule, the contact is intramolecular (self, *S*, Fig. 5*c*). Where the donor and acceptor molecules are derived from different unique molecules, the motif (*a*) is finite (discrete, *D*, Fig. 5*d*), since there is only one path from each unique molecule to the other. In the absence of internal symmetry, first-level motifs necessarily include one donor and one acceptor, and the degree *n* of discrete motifs is always 2, hence the shorthand *D* has been adopted for a *D*1,1(2) motif. Here, the number of nodes (atoms) rather than the number of edges in the graph are counted, following Bernstein *et al.* (1995); Bernstein *et al.* (1997) suggested that the degree should be determined by the number of bonds. The node and edge counts are the same for cyclic and chain patterns so the descriptions are equivalent in these cases.

The degree *n* of an **S** motif is equal to the shortest covalent bond path between the donor and the acceptor plus 1 for the H-edge: there may be more than one such path of the same length. These shortest intramolecular paths may be summarized in a *covalent distance table* (Grell *et al.*, 1999); this is done in Table 1 for *o*-amino-benzoic acid, form II (Boone *et al.*, 1977; CSD refcode AMBACO03, see also Fig. 10). Note that zero indicates that the path enters and leaves the molecule at the same donor or acceptor atom (*e.g.* in Table 1, the zero in the first line and last column indicates that the H-edges *a* and *c* share the same acceptor). If no covalent path exists between the H-edges, *i.e.* acceptor and donor belong to different original molecules, this is indicated with a blank (or ∞) in the table. The covalent distance table may be derived before the contacts are assigned or individual elements may be evaluated as they are needed. The (equal-) shortest paths are derived by considering all paths starting from the donor of a given length and increasing this length until a path is found which leads to the acceptor. Similarly, the repeat unit of a first-level ring or chain is the shortest path between the acceptor and donor plus 1 for the H-edge; this is equivalent to the degree *n* for a chain. For *R* motifs, the path repeat unit must be multiplied by the multiplicity *m* of the symmetry operator (= 2 for twofold axes, mirror planes and inversion centres; m = 3 for threefold axes, m = 4 for fourfold and $\bar{4}$ axes, and m = 6 for sixfold axes and for $\bar{3}$ and $\bar{6}$ axes).

2.8. *Second-level graph sets*

Bernstein *et al.* (1997) demonstrated that for a set of H-edges {**a**,**b**}, the representants are either rings and/or chains *or* selfs *or* discretes *or* there are none (the set is empty). Where {**a**,**b**} comprises rings or chains, the

Table 1. *Covalent distance table for o-aminobenzoic acid, form II, giving the covalent bond paths between H-edges* ***p*** *and* ***q*** *(after Bernstein et al., 1997)*

$q\backslash p$	$\vec{a}$	$\overleftarrow{a}$	$\vec{b}$	$\overleftarrow{b}$	$\vec{c}$	$\overleftarrow{c}$
$\vec{a}$	3	0	5	0	5	0
$\overleftarrow{a}$	0	3	6	3	6	3
$\vec{b}$	3	0	5	0	5	0
$\overleftarrow{b}$	6	5	0	5	2	5
$\vec{c}$	3	0	5	0	5	0
$\overleftarrow{c}$	6	5	2	5	0	5

number of representants is not finite. There are two possible unique paths with a repeat unit of two H-edges which must be considered when deriving the quantitative descriptors for rings and chains, which may be expressed as $(\vec{a}\ \vec{b})_m$ and $(\vec{a}\ \overleftarrow{b})_m$. The path $(\vec{a}\ \vec{b})_m$ is equivalent to $(\vec{b}\ \vec{a})_m$ with a different starting point and $(\overleftarrow{a}\ \overleftarrow{b})_m$ and $(\overleftarrow{b}\ \overleftarrow{a})_m$ describe the same path in the opposite direction. Other paths, *e.g.* $(\vec{a}\ \vec{a}\ \vec{b})_m$, $(\vec{a}\ \vec{a}\ \vec{b}\ \vec{b})_m$ or $(\vec{a}\ \vec{b}\ \overleftarrow{a}\ \overleftarrow{b})_m$ could be considered, up to a defined H-edge repeat unit, if required. We have concentrated on deriving the paths with the shortest H-edge repeat unit systematically, rather than those with the shortest period (total intra- and intermolecular path repeat) or with the shortest total path length, in contrast to the approach of Bernstein *et al.* (1995, 1997). The possibilities for second-level representants of the set **{a,b}** are shown in Fig. 6(*a*)–(*e*), which illustrate the systematic combination of H-edges *a* and *b* in the different possible first-level motifs *C*, *R*, *S* and *D*.

2.8.1. *Case I.* $\vec{a}$, $\overleftarrow{a}$, $\vec{b}$ and $\overleftarrow{b}$ all start in the same original molecule *A* and thus $X(\vec{a})_m$ and $X(\vec{b})_m$

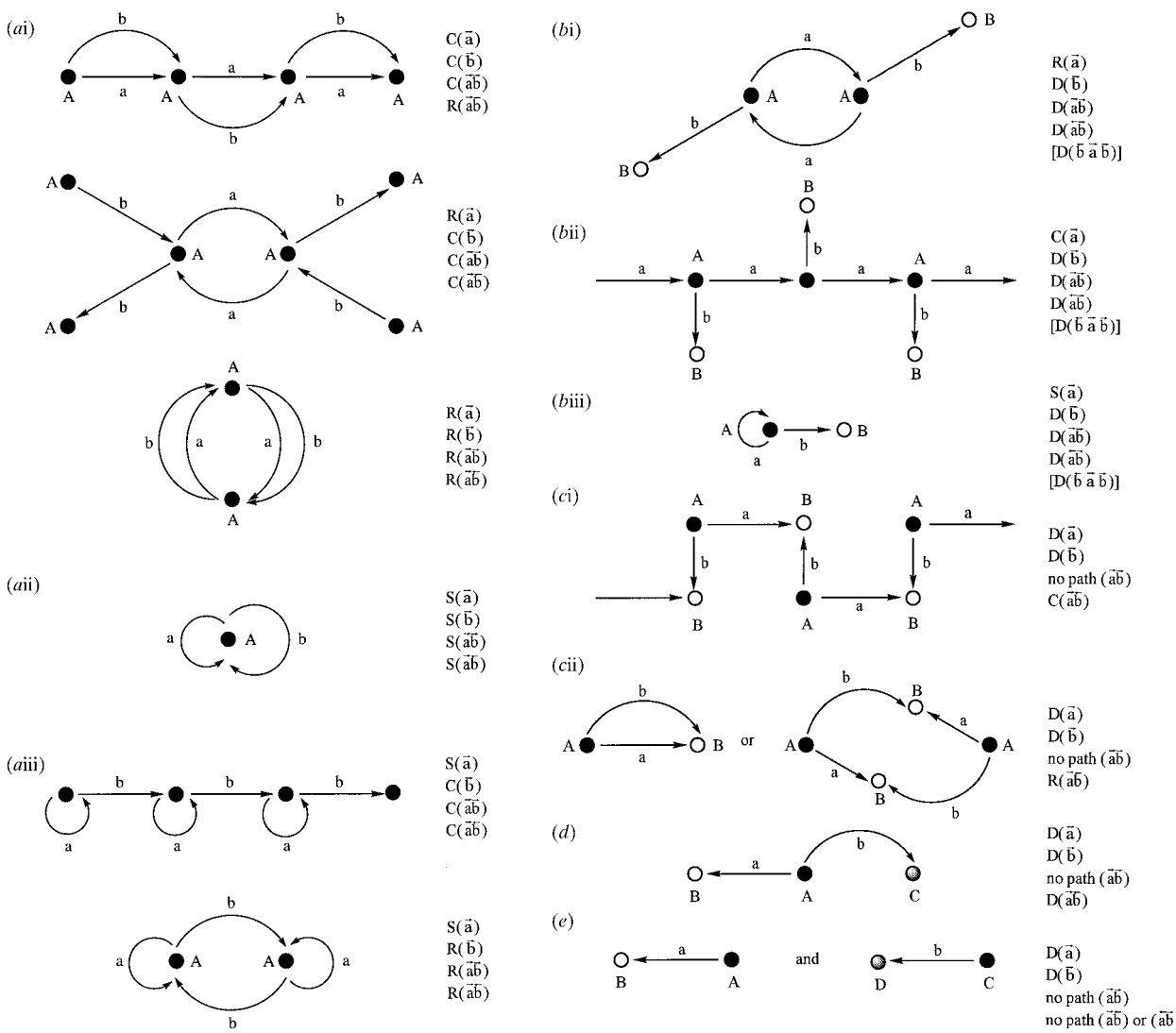

Fig. 6. Schematic constructor graphs for second-level graph-set patterns.

cannot describe *D* motifs. Both combinations $\overrightarrow{a}\,\overrightarrow{b}$ and $\overrightarrow{a}\,\overleftarrow{b}$ are possible and both must be considered. In Fig. 6(*a*i) the first level motifs $X(\overrightarrow{a})_m$ and $X(\overrightarrow{b})_m$ have the designators *R* or *C*, and $X(\overrightarrow{a}\,\overrightarrow{b})_m$ and $X(\overrightarrow{a}\,\overleftarrow{b})_m$ both describe rings and/or chains. In Fig. 6(*a*ii) the first-level representants are $S(\overrightarrow{a})$ and $S(\overrightarrow{b})$ and the second-level patterns $S(\overrightarrow{a}\,\overrightarrow{b})$ and $S(\overrightarrow{a}\,\overleftarrow{b})$ are also both **S** since both intramolecular contacts are in the same molecule. However, the significance of such motifs is debatable. In Fig. 6(*a*iii) the motif $X(\overrightarrow{a})$ is *S* and $X(\overrightarrow{b})_m$ is *R* or *C* (or *vice versa*); both $\overrightarrow{a}$ and $\overleftarrow{a}$ provide alternative paths between the *b* donor and acceptor in molecule *A*, and the resulting pattern is of the same type, *R* or *C*, as that of *b*.

2.8.2. *Case II*. $\overrightarrow{a}$, $\overleftarrow{a}$, $\overrightarrow{b}$ start in original molecule *A*, $\overleftarrow{b}$ in original molecule *B*. $X(\overrightarrow{b})$ is a discrete motif, $X(\overrightarrow{a})$ is *R*, *C* or *S*. Both $D(\overrightarrow{a}\,\overrightarrow{b})$ and $D(\overleftarrow{a}\,\overrightarrow{b})$ exist and are distinct discrete motifs; the paths $X(\overrightarrow{b}\,\overrightarrow{a})$ and $X(\overrightarrow{b}\,\overleftarrow{a})$, respectively, are not equivalent and do not exist, since an *a* path may only start or end in molecule *A*). As a result, four directed label sequences must be considered in this case. The path conventionally chosen is the shortest which starts and ends in two molecules *M* and *M*′, to which only one H-edge in the set {**a**,**b**} is incident. *M* and *M*′ correspond to two symmetry-related molecules *B* in Fig. 6(*b*). Where $X(\overrightarrow{a})$ is not a self, the path chosen is *M*-simple, *e.g.* the path $D(\overleftarrow{b}\,\overrightarrow{a}\,\overrightarrow{a}\,\overrightarrow{b})$ in Fig. 6(*b*i) is not a valid representant of the subset of {**a**,**b**} obeying this condition. To be systematic, we have chosen to describe the path with the smallest H-edge repeat unit, *i.e.* directed label sequence, the composite pattern $D(\overleftarrow{b}\,\overrightarrow{a}\,\overrightarrow{b})$.

2.8.3. *Case III*. $\overrightarrow{a}$, $\overrightarrow{b}$ (or $\overrightarrow{a}$, $\overleftarrow{b}$) start in molecule *A*, $\overleftarrow{a}$, $\overleftarrow{b}$ (or $\overleftarrow{a}$, $\overrightarrow{b}$) start in molecule *B*: the first-level motifs are $D(\overrightarrow{a})$ and $D(\overrightarrow{b})$. Motif $X(\overrightarrow{a}\,\overleftarrow{b})_m$ [$\equiv X(\overleftarrow{b}\,\overrightarrow{a})_m$] may be *C* or *R*, whereas $X(\overrightarrow{a}\,\overrightarrow{b})$ does not exist, *i.e.* there is only one path with an H-edge repeat unit (*ab*), Fig. 6(*c*).

2.8.4. *Case IV*. $\overrightarrow{a}$ and $\overrightarrow{b}$ start in molecule *A*, $\overleftarrow{a}$ starts in molecule *B* and $\overleftarrow{b}$ starts in molecule *C*. Discrete $D(\overleftarrow{b}\,\overrightarrow{a})$ exists, $X(\overrightarrow{a}\,\overrightarrow{b})$, $X(\overrightarrow{b}\,\overrightarrow{a})$, $X(\overrightarrow{a}\,\overleftarrow{b})$ do not: each of these four combinations are not equivalent.

2.8.5. *Case V*. No single molecule is associated with *a* and *b* in either direction, there is no second level pattern involving contacts *a* and *b*, *i.e.* the set {**a**,**b**} is empty, although the molecules related by *a* and *b* may be part of the same extended network *via* other contacts, Fig. 6(*e*).

In practice, these considerations may be implemented by taking all pairwise combinations of motifs $\overrightarrow{a}$, $\overleftarrow{a}$, $\overrightarrow{b}$ and $\overleftarrow{b}$; only those which start in the same molecule *A* need be considered further. The classification of a motif as *R* or *C* is achieved by considering the operation **T** relating the two molecules *B*′ and *B*″ in contact with the original molecule *A*. If molecules *B*′ and *B*″ are related to the original molecule *B* by operations **S**′ and **S**″, respectively (where **S**′ and/or **S**″ may be the identity operator), the resulting operator **T** is given by $\mathbf{T} = \mathbf{S}'^{-1}\mathbf{S}''$.

The *inverse symmetry operation* may be found as follows: consider an operation **S** which transforms an atom at position **x** to **x**′. This symmetry operation may be expressed as a combination of a translation **t** and rotation/inversion **R**

$$\mathbf{Sx} = \mathbf{Rx} + \mathbf{t} = \mathbf{x}'.$$

Hence the inverse operation **S**′ which maps **x**′ back to **x** is given by

$$\mathbf{R}'\mathbf{x}' + \mathbf{t}' = \mathbf{R}^{-1}(\mathbf{x}' - \mathbf{t}) = \mathbf{R}^{-1}\mathbf{Rx} = \mathbf{x},$$

i.e. a rotation $\mathbf{R}' = \mathbf{R}^{-1}$ and translation $\mathbf{t}' = \mathbf{R}^{-1}(-\mathbf{t})$.

In order to identify the inverse symmetry operator it is necessary to compare the rotation matrix **R**′ with those for the symmetry operations stored for the space group and similarly for the non-integral part of the translation **t**′. For convenience, the translation may be divided into the part included in the symmetry operator $\mathbf{t}'_{\rm sym}$ and that due to additional translations of an integral number of cell lengths $\mathbf{t}'_{xyz}$. The appropriate rotational component **R**′ may be identified by either inverting the matrix **R** or pre-multiplying **R** with that of each symmetry element until one is found which generates the identity matrix.

The degree of the {**a**,**b**} *R* and *C* patterns are evaluated by summing the covalent bond paths between donor and acceptor in adjacent molecules in the ring or chain and adding 1 for each edge: where the path length in a molecule is zero, the (*ab*) sequence involves only one acceptor or donor. As for first level motifs, the degree of an *R* pattern is multiplied by the multiplicity *m* of the operator generating the cyclic path. For *D* patterns, only the bond path in the original molecule *A* is considered, and incremented by two for the link atoms in molecules *B* and *B*′ or *C*.

Conventionally, only the representant of lowest degree is retained where $X(\overrightarrow{a}\,\overrightarrow{b})$ and $X(\overrightarrow{a}\,\overleftarrow{b})$ are either both *R*, both *C* or both *S*. However, it may be more useful for systematic comparisons to retain both paths rather than make a selection based on path length, as this may preserve similarities in the graph-set descriptions of compounds which differ in terms of their intramolecular path lengths, but not in their hydrogen-bonding network. As a consequence, both $\overrightarrow{a}\,\overleftarrow{b}$ and $\overrightarrow{a}\,\overrightarrow{b}$ paths are identified where present.

2.9. *Higher-level graph sets*

The approach described above could, in principle, be extended to third- and higher-level graph sets by considering combinations of three or more independent H-edges, *e.g.* $(\overrightarrow{a}\,\overrightarrow{b}\,\overrightarrow{c})$, $(\overrightarrow{a}\,\overleftarrow{b}\,\overrightarrow{c})$ *etc*. However, the number of possible combinations increases rapidly with the level of graph set and the methodology has not been implemented yet for graph sets higher than second level,

Table 2. *(a) Initial graph-set matrix, describing first-level graph sets (p) along the leading diagonal, second-level graph sets ($\overrightarrow{p}\ \overrightarrow{q}$) in upper-right and ($\overrightarrow{p}\ \overleftarrow{q}$) in the lower left, and (b) final graph-set matrix for o-aminobenzoic acid, form II*

	a	*b*	*c*
(*a*)			
a	*R*2,2(8)	*R*4,2(16)	*C*2,2(10)
b	*R*4,4(20)	*S*1,1(6)	*C*2,2(12)
c	*C*1,2(8)	*C*1,2(4)	*C*1,1(6)
(*b*)			
a	*R*2,2(8)		
b	*R*4,2(16)	*S*1,1(6)	
c	*C*1,2(8) [*R*2,2(8)]	*C*1,2(4) [*S*1,1(6)]	*C*1,1(6)

Table 3. *(a) Initial and (b) final graph-set matrix for 8-acetamido-5,6,7,8-tetrahydro-2-naphthoic acid N-isopropylamide*

	a	*b*	*c*	*d*
(*a*)				
a	*D*1,1(2)	*D*2,2(6)	*R*2,2(20)	*D*2,2(6)
b	*D*2,2(11)	*R*2,2(16)	*D*2,2(6)	
c		*D*2,2(11)	*D*1,1(2)	*D*2,2(6)
d	*D*2,2(11)		*D*2,2(11)	*R*2,2(16)
(*b*)				
a	*D*1,1(2)			
b	*D*3,3(15)	*R*2,2(16)		
c	*R*2,2(20)	*D*3,3(15)	*D*1,1(2)	
d	*D*3,3(15)		*D*3,3(15)	*R*2,2(16)

although work in this area is in progress. In many cases the chemically significant sequences are those involving only one or two independent H-edges, although this is not always true of more complex examples.

2.10. *Graph-set matrix*

This initial graph-set description, of first and second-level patterns, may be summarized conveniently in a square matrix, the first-level motifs on the leading diagonal, $\overleftarrow{a}\ \overrightarrow{b}$, $\overrightarrow{a}\ \overleftarrow{b}$ motifs in the lower left triangle and $\overrightarrow{a}\ \overrightarrow{b}$, $\overleftarrow{a}\ \overleftarrow{b}$ in the upper right. The initial graph-set matrix for form II of *o*-aminobenzoic acid is given in Table 2(*a*). Where both second level discrete patterns of the form $\overrightarrow{a}\ \overrightarrow{b}$ and $\overrightarrow{a}\ \overleftarrow{b}$ exist, they are not combined in the initial matrix (although they are in the plot key list and expressed as *a*&*b*).

The conventional lower-triangular matrix (Bernstein *et al.*, 1995; Grell *et. al.*, 1999: Table 2*b*) may then be derived from this matrix. The diagonal elements representing the first-level motifs are identical to those in the initial matrix. Of the second-level patterns, chain motifs are considered first and only the pattern $\overrightarrow{a}\ \overrightarrow{b}$ or $\overrightarrow{a}\ \overleftarrow{b}$ of shortest degree is retained. Where both $\overrightarrow{a}\ \overrightarrow{b}$ and $\overrightarrow{a}\ \overleftarrow{b}$ have the same degree, the pattern with the smaller number of donors and/or acceptors is selected. *S* and *R* motifs are selected similarly where both $\overrightarrow{a}\ \overrightarrow{b}$ and $\overleftarrow{a}\ \overrightarrow{b}$ describe the same motif type. If $\overrightarrow{a}\ \overrightarrow{b}$ forms a ring and $\overrightarrow{a}\ \overleftarrow{b}$ a chain, or *vice versa*, the ring motif is displayed after the chain motif in square brackets. Similarly, first-level *R* motifs involving contacts $\overrightarrow{a}$ and $\overrightarrow{b}$ are included in square brackets after the $\overrightarrow{a}\ \overrightarrow{b}$ chain if present (the 'chain of rings' pattern). Where both $\overrightarrow{a}\ \overrightarrow{b}$ and $\overleftarrow{a}\ \overrightarrow{b}$ discrete patterns exist, these are combined to form the complex graph set $\overrightarrow{a}\ \overrightarrow{b}\ \overleftarrow{a}$, the degree $n_c = n_{\vec{a}\vec{b}} + n_{\bar{a}\vec{b}} - 2$, $d_c = d_{\vec{a}\vec{b}} + d_{\bar{a}\vec{b}} - 1$ and $a_c = a_{\vec{a}\vec{b}} + a_{\bar{a}\vec{b}} - 1$, since the contact atom pair $X\cdots Y$ occurs in both $\overrightarrow{a}\ \overrightarrow{b}$ and $\overrightarrow{b}\ \overleftarrow{a}$. These principles are illustrated for the more complex examples 8-acetamido-5,6,7,8-tetrahydro-2-naphthoic acid *N*-isopropylamide (Ernest *et al.*, 1990: Table 3) and *N*-acetyl-dehydroalanine *N*′-methylamide (Palmer *et al.*, 1992:

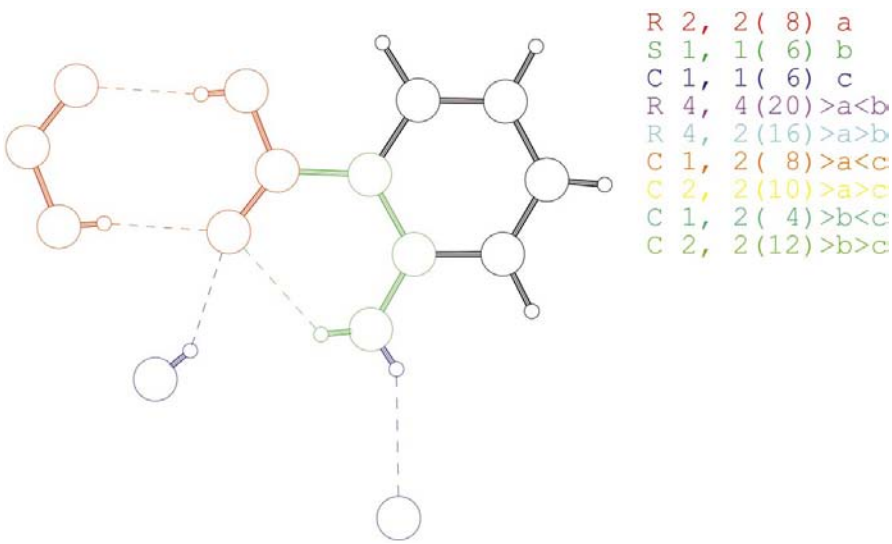

Fig. 7. Visual representation of first- and second-level graph sets for form II of *o*-aminobenzoic acid.

Table 4. *(a) Initial and (b) final graph-set matrix for N-acetyl-dehydroalanine N′-methylamide*

	a	*b*	*c*	*d*	*e*	*f*
(*a*)						
a	*S*1,1(5)	*D*2,2(7)	*C*2,2(8)	*D*2,2(7)		
b	*D*2,1(3)	*D*1,1(2)	*D*2,2(6)	*R*2,2(10)	*D*2,1(3)	*D*2,2(6)
c	*C*2,2(12)	*D*2,2(8)	*C*1,1(7)	*D*2,2(6)		
d	*D*1,2(3)		*D*2,2(8)	*D*1,1(2)	*D*1,2(3)	*D*2,2(6)
e		*D*2,2(7)		*D*2,2(7)	*S*1,1(5)	*C*2,2(8)
f		*D*2,2(8)		*D*2,2(8)	*C*2,2(12)	*C*1,1(7)
(*b*)						
a	*S*1,1(5)					
b	*D*3,2(8)	*D*1,1(2)				
c	*C*2,2(8) [*S*1,1(5)]	*D*3,3(12)	*C*1,1(7)			
d	*D*2,3(8)	*R*2,2(10)	*D*3,3(12)	*D*1,1(2)		
e		*D*3,2(8)		*D*2,3(8)	*S*1,1(5)	
f		*D*3,3(12)		*D*3,3(12)	*C*2,2(8) [*S*1,1(5)]	*C*1,1(7)

Table 4), both of which have such complex discrete graph sets at second level.

2.11. *Visualization of graph sets, examples*

In addition to the matrix representation, graph sets are displayed on screen with colour-key coding in the style $X\,a,d(n)$ and the atoms are highlighted by colour in the diagram (*e.g.* Fig. 7 highlights the first-level graph sets of polymorph II of *o*-aminobenzoic acid, *cf.* Table 2). The key lists the first-level motifs for each of the H-edges p (where p is an integer label) and the second-level patterns formed by pairs of these H-edges, labelled $>p>q$ (for $\overrightarrow{p}\ \overrightarrow{q}$ patterns), $>p<q$ ($\overrightarrow{p}\ \overleftarrow{q}$), and/or $p\&q$ (discrete motifs $\underrightarrow{p\ q}\ \overrightarrow{p}$ and $\underrightarrow{q\ p}\ \overleftarrow{q}$). Although the current implementation is limited to nine independent hydrogen bonds, the methodology is applicable to any number of independent H-edges.

Particular motifs may be highlighted, by colouring the atoms and bonds forming the path, and duplicate intramolecular bond paths are included (the colours used for first-level patterns are the same as those of the H-edge describing the motif). The motifs are identified by number: p for first-level, pq for second-level patterns $>p>q$ or $p\&q$ and $-pq$ for second-level patterns $>p<q$. Several motifs may be specified on the command line and if no arguments are given all motifs are coloured. Where motifs overlap, the colour of the first motif in the list of those to be displayed takes precedence. The motif colours correspond to the graph-set colour key and the first-level motif colours to the colours associated with the corresponding intermolecular contact. This display method is illustrated by the graph sets for 8-acetamido-5,6,7,8-tetrahydro-2-naphthoic acid *N*-isopropylamide (Fig. 8, Table 3, CSD refcode JEPHOB) and *N*-acetyl-dehydroalanine *N′*-methylamide (Fig. 9), both of which have two molecules in the CCU. In the second example (Fig. 9, Table 4, CSD refcode JUDZEN) ladders of centrosymmetric dimers are formed, the two molecules in the CCU alternating along the ladder in such a manner that the predominant chain pattern appears at the fourth level ($\overrightarrow{a}\ \overrightarrow{b}\ \overrightarrow{c}\ \overrightarrow{d}$). This illustrates that only considering graph sets up to second level is not always sufficient to describe the chemically significant patterns.

These motifs can be explored in the extended crystal structure by the network expansion process described above. Where two or more contacts are formed to the same adjacent molecule, the atoms comprising intramolecular shortest paths in this molecule are also added

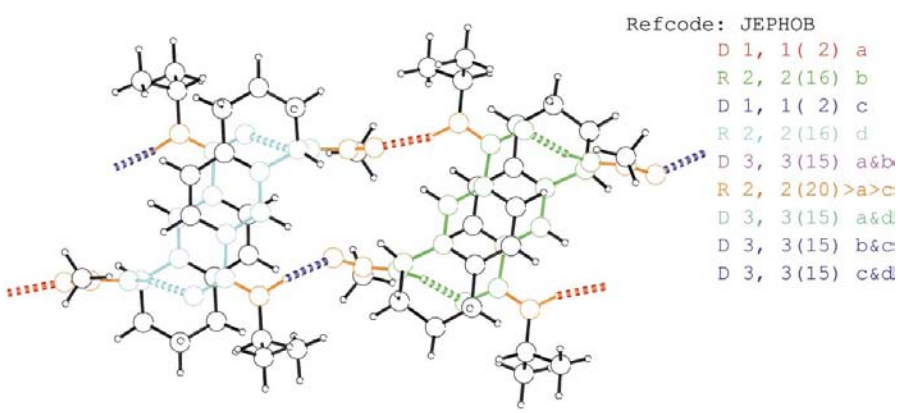

Fig. 8. Graph sets for 8-acetamido-5,6,7,8-tetrahydro-2-naphthoic acid *N*-isopropylamide, highlighting centrosymmetric first-level *R*2,2(16) motifs formed by each independent molecule in the CCU, joined by second-level *R*2,2(20) rings.

as extra link atoms. Fig. 10 shows the C1,2(8) chain in form II of anthranilic acid propagated in this manner.

2.12. *Graph sets where intramolecular symmetry is present*

The situation is more complex where one or more molecules possesses internal crystallographic symmetry. In this case, it is not possible to enumerate the contacts uniquely, as more than one symmetry-related contact may emanate from the same molecule in the same direction. A unique enumeration may be achieved, however, if the structure is treated as if it were in a lower-symmetry subgroup of the space group, in which the operator generating the symmetry atoms was absent. An alternative approach is possible, in which the structure is treated in the original space group and a complete molecule may be related to the same

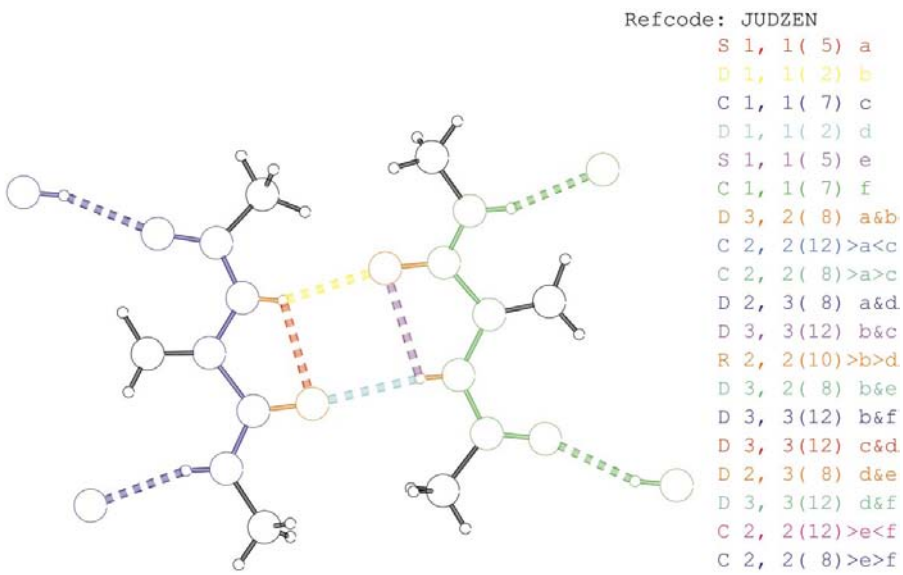

Fig. 9. Graph sets for *N*-acetyldehydroalanine *N*′-methylamide, showing first-level chain motifs formed by two crystallographically independent molecules, linked by a second-level *R*2,2(8) pattern.

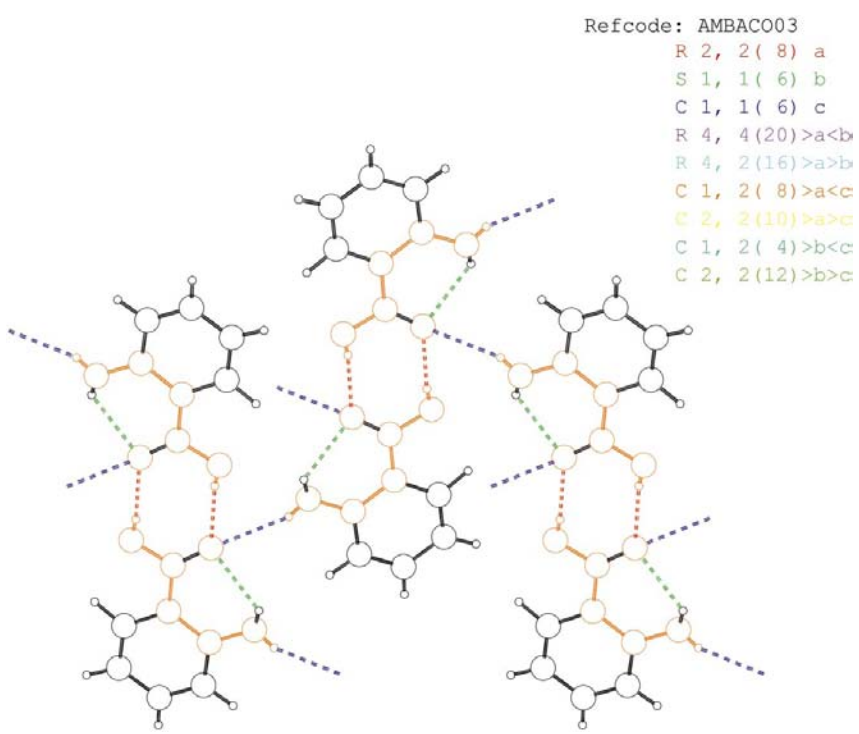

Fig. 10. Part of the extended crystal structure of form II of *o*-aminobenzoic acid, highlighting the *C*1,2(8) chain pattern.

symmetry-generated neighbour by more than one symmetry operator. Work in this area is in progress and will be reported later.

3. Conclusions

A convenient and intuitive expansion method has been developed for visualizing crystal structures, which enables networks of non-covalent bonds to be investigated more easily than with traditional unit-cell packing diagrams. An algorithm has been developed for the automatic computation of graph-set descriptors, up to second level, for asymmetric molecules, based on the symmetry relationships between the molecules linked by non-covalent interactions. Graph sets have previously proved useful for the comparison of networks in different crystals, particularly in polymorphic systems and in series of closely related compounds. The availability of an automated procedure for graph-set derivation should ensure consistency and facilitate the routine use of graph-set descriptors in structural studies. It is expected that these tools will be invaluable for crystal structure comparison, prediction and modelling applications.

The program *PLUTO* described in this paper may be obtained free of charge from the CCDC internet site http://www.ccdc.cam.ac.uk/, email: pluto@ccdc.cam.uk.

References

Aakeröy, C. B. & Seddon, K. R. (1993). *Chem. Soc. Rev.* **22**, 397–407.

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

Allen, F. H., Kennard, O., Motherwell, W. D. S., Town, W. G., Watson, D. G., Scott, T. J. & Larson, A. C. (1974). *J. Appl. Cryst.* **7**, 73–78.

Allen, F. H., Kennard, O., Watson, D. G., Brammer, L., Orpen, A. G. & Taylor, R. (1987). *J. Chem. Soc. Perkin Trans. 2*, pp. S1–S19.

Bernstein, J. (1991). *Acta Cryst.* B**47**, 1004–1010.

Bernstein, J. & Davis, R. E. (1999). *Implications of Molecular and Materials Structure for New Technologies*, edited by J. A. K. Howard, F. H. Allen & G. P. Shields, pp. 275–290. Dordrecht: Kluwer Academic Publishers.

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

Bernstein, J., Etter, M. C. & Leiserowitz, L. (1994). *Structure Correlation*, edited by H.-B. Bürgi and J. D. Dunitz, Vol. 2, pp. 431–507. Weinhein: VCH.

Bernstein, J., Etter, M. C. & MacDonald, J. M. (1990). *J. Chem. Soc. Perkin Trans. 2*, pp. 695–704.

Bernstein, J., Ganter, B., Grell, J., Hengst, U., Kuske D. & Pöschel, R. (1997). Technische Universität Dresden Preprint.

Boone, C. G. D., Derissen, J. L. & Schoone, J. C. (1977). *Acta Cryst.* B**33**, 3205–3206.

Clark, M., Cramer III, R. D. & van Opdenbosch, N. (1989). *J. Comput. Chem.* **10**, 982–1012.

Desiraju, G. R. (1995). *Angew. Chem. Int. Ed. Engl.* **34**, 2311–2327.

Ernest, I., Kalvoda, J., Rihs, G. & Mutter, M. (1990). *Tetrahedron Lett.* **31**, 4011.

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

Etter, M. C. (1991). *J. Phys. Chem.* **95**, 4601–4610.

Etter, M. C. & Frankenbach, G. M. (1989). *Materials*, **1**, 10–12.

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

Fischer, W. & Koch, E. (1983). *International Tables for Crystallography*, edited by Th. Hahn, Vol. A, ch. 11, pp. 788–792. Birmingham: Kynoch Press. (Present distributor Kluwer Academic Publishers, Dordrecht.)

Garcia-Tellado, F., Geib, S. J., Goswani, S. & Hamilton, A. D. (1991). *J. Am. Chem. Soc.* **113**, 9265–6269.

Grell, J., Bernstein, J. & Tinhofer, G. (1999). *Acta Cryst.* B**55**, 1030–1043.

Harary, F. (1969). *Graph Theory*. Reading, Massachusetts: Addison-Wesley.

Jeffrey, G. A. (1997). *An Introduction to Hydrogen Bonding*. New York, NY: Oxford University Press.

Jeffrey, G. A. & Lewis, L. (1978). *Carbohydrate Res.* **60**, 179–182.

Jeffrey, G. A. & Takagi, S. (1978). *Acc. Chem. Res.* **11**, 264.

Johnson, C. K. (1965). *ORTEP*. Report ORNL-3794. Oak Ridge National Laboratory, Tennessee, USA.

Jones, W., Pedireddi, V. R., Chorlton, A. P. & Docherty, R. (1996). *Chem. Commun.* pp. 997–998.

Kuleshova, L. N. & Zorkii, P. M. (1980). *Acta Cryst.* B**36**, 2113–2115.

Leiserowitz, L. (1976). *Acta Cryst.* B**32**, 775–802.

Leiserowitz, L. & Schmidt, G. M. J. (1969). *J. Chem. Soc. A*, pp. 2372–2382.

Leiserowitz, L. & Tuval, M. (1978). *Acta Cryst.* B**34**, 1230–1247.

Palmer, D. E., Pattaroni, C., Nunami, K., Chadha, R. K., Goodman, M., Wakamiya, T., Fukase, K., Horimoto, S., Kitazawa, M., Fujita, H., Kubo, A. & Shiba, T. (1992) *J. Am. Chem. Soc.* **114**, 5634–5642.

Panunto, T. W., Urbánczyk-Lipkowska, Z., Johnson, R. B. & Etter, M. C. (1987). *J. Am. Chem. Soc.* **109**, 7786–7797.

Rollett, J. S. (1965). In *Computing Methods in Crystallography*, edited by J. S. Rollett. Oxford: Pergamon Press.

Wells, A. F. (1962). *Structural Inorganic Chemistry*. Oxford University Press.