Acta Crystallographica Section B
Structural Science
ISSN 0108-7681

Structures of dipeptides: the head-to-tail story

Carl Henrik Görbitz

Department of Chemistry, University of Oslo, Norway

Correspondence e-mail: c.h.gorbitz@kjemi.uio.no

Received 1 October 2009
Accepted 10 December 2009

The hydrogen-bonding patterns in crystal structures of unprotected, zwitterionic dipeptides are dominated by head-to-tail chains involving the *N*-terminal amino groups and the *C*-terminal carboxylate groups. Patterns that include two concomitant chains, thus generating a hydrogen-bonded layer, are of special interest. A comprehensive survey shows that dipeptide structures can conveniently be divided into only four distinct patterns, differing by definition in the symmetry of the head-to-tail chains and amide hydrogen-bonding type, but also in other properties such as peptide conformation and the propensity to include solvent water or various organic guest molecules. Upon crystallization, the choice of pattern for a specific dipeptide is not random, but follows from the amino acid sequence.

1. Introduction

The concept of infinite head-to-tail hydrogen-bonded chains was originally introduced by Suresh & Vijayan (1983) to describe sequences of the type $\cdots NH_3^+-CHR-COO^-\cdots NH_3^+-CHR-COO^-\cdots NH_3^+-CHR-COO^-\cdots$ consistently observed in the crystal structures of simple amino acids. Soon after the same authors analysed the crystal structures of linear peptides in a similar manner (Suresh & Vijayan, 1985, abbreviated to S&V hereafter), focusing on head-to-tail sequences involving the *N*-terminal amino group and the *C*-terminal carboxylate group, or *C*(8) chains in graph-set terminology (Etter *et al.*, 1990; Grell *et al.*, 2002). From the available experimental material at the time, consisting of 27 dipeptide structures (as well as 11 tri-, tetra- and pentapeptides), S&V found that the symmetry elements available for propagation of a head-to-tail chain were limited to translation and twofold screw axes, giving rise to *S* (straight) and *Z* (zigzag) sequences. In an unprecedented manner, the authors then constructed a series of theoretical hydrogen-bonding arrangements with particular focus on patterns with two co-existing head-to-tail sequences ($N = 2$) as in Fig. 1.

Two generalized peptide conformations were considered: extended (*E*) and folded (*F*). Combining symmetry (*S* or *Z*) with conformation (*E* or *F*) and carefully considering potential steric conflict, the authors derived seven plausible idealized crystalline patterns for dipeptide aggregation, six two-dimensional and one three-dimensional. Fig. 1 shows an *EZA* pattern, the last letter in the code (*A*) distinguishing between patterns with the same molecular geometry and the same type of sequence (there is also an idealized antiparallel pattern called *EZB*). Of the 16 dipeptide structures (out of 25 in total) with $N = 2$ or $N = 3$, 12 fitted the idealized patterns, two were considered intermediates between two classes, while two

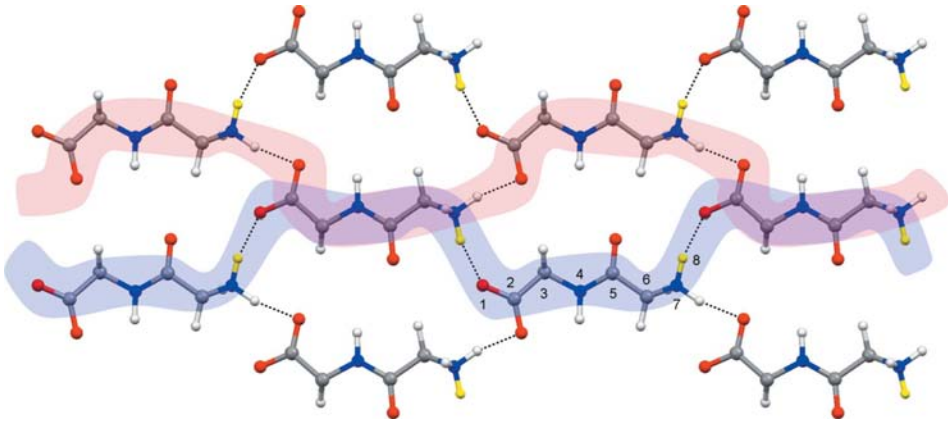

Figure 1
The simultaneous presence of two crystallographically independent hydrogen-bonded chains, highlighted in blue (with yellow H atoms) and red, in a dipeptide structure. Peptide side chains have been omitted.

structures, including Ala–Ala (Fletterick *et al.*, 1971), clearly fell outside the classification scheme.

Since S&V carried out their investigation, numerous new structures have been published. There is thus ample room for improved statistics in a new survey, with the inclusion of packing arrangements that were not known in 1985 such as nanotubular dipeptides. Furthermore, knowing that the primary sequence of a protein contains a complete set of instructions for chain folding, it was of interest to carry out a chemical interpretation of dipeptide structures, that is how amino-acid sequence affects or even dictates the crystal-structure properties of a small peptide. In doing so the traditional *C*(8) chains clearly had to be considered, but also hydrogen-bonding sequences involving the peptide >N—H group. Two main-chain acceptors are available for this donor, either the *C*-terminal carboxylate group, leading to a *C*(5) chain, or the peptide carbonyl group, leading to a *C*(4) chain, Fig. 2.

As is evident from Fig. 2, the >N—H···O hydrogen bond in a *C*(4) chain is accompanied by two C^{α}—H···O interactions, generating a much stronger, tape-like hydrogen-bonding motif.

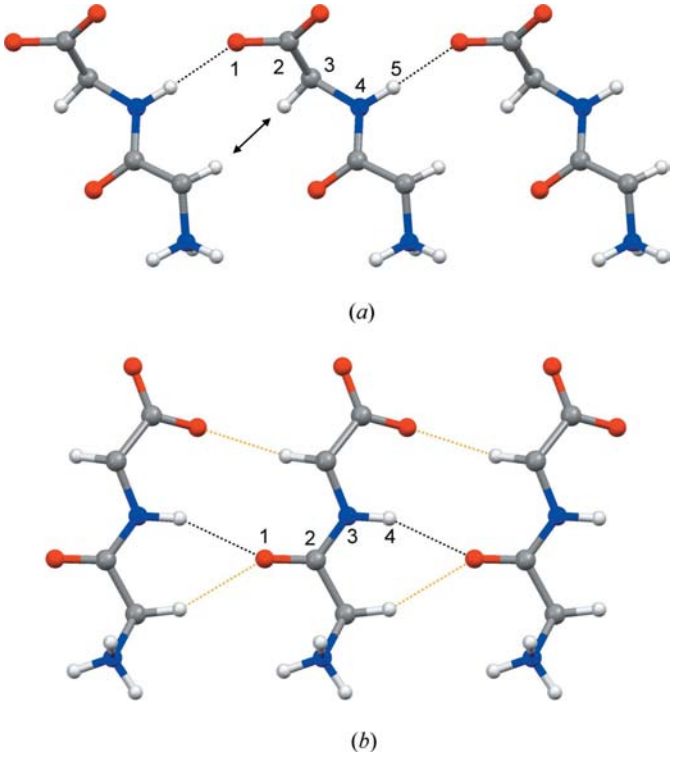

Figure 2
(*a*) *C*(5) and (*b*) *C*(4) hydrogen-bonded chains involving peptide >N—H donors. C^{α}—H···O hydrogen bonds are coloured in orange. The H···H distance indicated with an arrow in (*a*) is normally > 2.8 Å and does not represent steric conflict.

2. Methodology

2.1. Retrieval of structures from the Cambridge Structural Database

The Cambridge Structural Database (CSD, Version 5.30 of November 2008; Allen, 2002) was searched for unprotected dipeptide structures with zwitterionic main chains. The peptides were allowed to include uncommon amino acids like those depicted in Fig. 3.

H-atom distances were normalized to default values with N—H (amide and amine) = 1.009 Å and C—H = 1.083 Å. For entries devoid of H-atom coordinates H atoms were introduced in theoretical positions.

Some compounds have been co-crystallized with a number of different solvent or guest molecules, the binding modes and conformations of the peptides themselves being almost indistinguishable from one structure to the next. For statistics not to be overly biased by such pseudopolymorphic structural families, only those entries were retained that are different in at least one of the following aspects: chirality, space group, cell dimensions (within approximately 2 Å for the longest axis) and the value of *Z*. For the two by far the largest families, fGly-fGly and nGly-fGly (see Fig. 3), the numbers of entries were consequently reduced from 16 and 19 to 8 and 5, respectively, while *e.g.* Leu-Ser was reduced from 5 to 1.

2.2. Dataset

The database for the present survey consisted of a total of 159 dipeptide structures retrieved from the CSD together with an unpublished structure (Görbitz, 2010). The majority, 139 peptides, are constructed solely from the 20 common amino acids, while 21 incorporate one or two of the uncommon residues shown in Fig. 3. As far as charge is concerned, 152 out of 160 dipeptides have a net charge of 0, either as normal

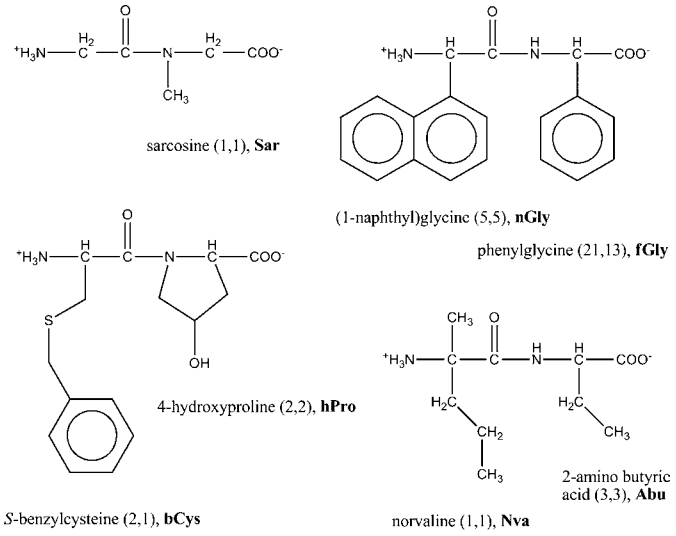

Figure 3
Uncommon amino acids included in the investigation, with three- or four-letter code. The numbers in parenthesis give the total number of residues of this type and the total number of peptides in which they occur.

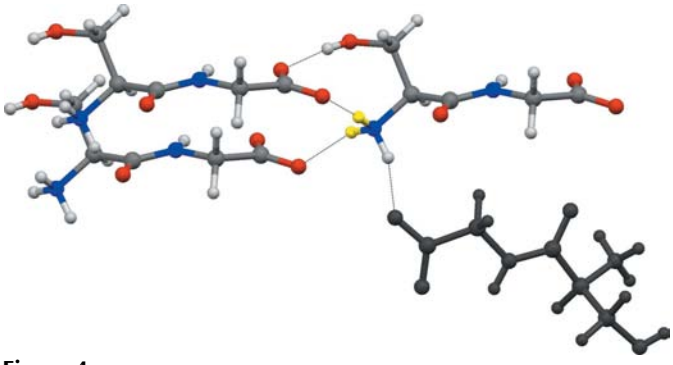

Figure 4
Detail from the crystal structure of Ser-Gly (Jones *et al.*, 1978*a*) with N = 3. The two amino H atoms coloured in yellow give rise to a two-dimensional layer, the third $C(8)$ chain (to a black molecule) cannot be combined with any of the other two $C(8)$ chains to generate a layer.

zwitterions (153) or as double zwitterions (7) with two charged side chains such as Arg-Glu dihydrate (Pandit *et al.*, 1983), while eight have a net charge of +1. An *N*-terminal Pro residue, which introduces a special main-chain conformation and limits the maximum number of $C(8)$ chains to two rather than three, is found for seven dipeptides.

Considering the chirality of the two residues, the 160 structures can be divided into three groups: 128 with two chiral residues, 27 with one chiral residue (the other one being Gly) and five achiral Gly-Gly structures.[1] The first group can be further subdivided into three groups: 122 L-L or D-D peptides, four L-D or D-L peptides and two L-D/D-L racemates (there are no L-L/D-D racemates). The second group consists of 23 L or D structures (Gly-L-Xaa, Gly-D-Xaa, L-Xaa-Gly or D-Xaa-Gly, Xaa = any amino acid) and four L/D racemates. Below, stereochemical indicators are included only for D enantiomers (Ala-D-Leu = L-Ala-D-Leu).

2.3. Hydrogen bonds and the identification of layers

It is usually straightforward to establish whether a hydrogen bond is present or not, but a 2.5 Å distance limit for the H···O distance was applied when required. Three-centre hydrogen bonds involving carboxylate acceptors require some special attention, and for interactions involving the peptide >N—H donor it is often necessary to consider much longer contacts (see below). Hydrogen-bonded layers are easily recognized by the fact that two of the amino H atoms are donated to peptide molecules that are related to each other by translation along a crystallographic axis, as in Fig. 1, or by pseudotranslational symmetry when $Z' > 1$. Notably, for structures with $N = 3$ there is always a unique combination of two $C(8)$ chains that generates a layer equivalent to one seen in structures with N = 2; the third $C(8)$ chain merely adds to this pattern as illustrated for Ser-Gly (Jones *et al.*, 1978*a*) in Fig. 4.

3. Results and discussion

3.1. *C*(4), *C*(5) and *C*(8) chains

The presence of $C(8)$ head-to-tail chains in each of the 160 structures in the database was first established through structure searches with *ConQuest* (Bruno *et al.*, 2002) and subsequently verified by manual scrutiny. The number of such chains varies between 3 and 0: N = 3: 25, N = 2: 93, N = 1: 27, N = 0: 15. Only the 118 structures with N = 2 and N = 3 may form the various types of patterns studied by S&V, see the example shown in Fig. 1. Approximately two thirds of the structures contain regular chains involving the amide >N—H donor. $C(5)$ chains occur in 71 structures with an average H···O distance of 1.96 Å and range 1.70–2.34 Å. $C(4)$ chains are not only less abundant, occurring in 37 structures, but are also significantly longer with an average H···O distance of 2.49 Å and a 1.99–3.15 Å range that extends well beyond the normally applied limits for recognizing the presence of a hydrogen bond. It is nevertheless important to consider these weak interactions in order to not overlook packing similarities among peptide structures. Additional statistics on these hydrogen bonds are available as supplementary material.[2]

3.2. The four basic aggregation patterns

S&V used a classification scheme for dipeptide patterns based on chain symmetry and peptide conformation, while a new terminology will be introduced here. As before, two different peptide properties are considered in distinguishing between structure types, and the first, $C(8)$ chain symmetry, remains the same although new abbreviations will be used: **T** for translation and **S** for screw axis. Peptide conformation,

[1] Gly-Gly, with α- (Kvick *et al.*, 1977) and β-forms (Hughes & Moore, 1949), is incidentally still the only dipeptide for which two true polymorphs are known (in distinction to structures that differ in solvent content, so-called pseudo-polymorphs).

[2] Supplementary data for this paper are available from the IUCr electronic archives (Reference: RY5028). Services for accessing these data are described at the back of the journal.

Table 1
Observed packing patterns in 93 dipeptide structures with $N = 2$ and 25 structures with $N = 3$.

Pattern	$N = 3$	$N = 2$
T4	1 + 1†	6 + 1†
S4	1	18 + 1†
T5	5 + 3†	20
S5	2 + 1†	23
Other parallel	5	2
Antiparallel	2	6
No layer, nanotubular	0	11‡
No layer, other	4	5

† Modified pattern, $C(4)$ or $C(5)$ chain missing. ‡ Val-Ala class (Görbitz, 2003*a*).

used by S&V as the second property, will be abandoned in favour of amide hydrogen-bonding type. Accordingly, **4** is used to indicate the presence of $C(4)$ chains, while **5** is used for $C(5)$ chains. The combination of chain symmetry (**T** or **S**) and hydrogen-bonding pattern (**4** or **5**) thus gives rise to four basic aggregation patterns: **T4**, **S4**, **T5** and **S5**. These are illustrated in Fig. 5.

A summary of observed hydrogen-bonding patterns in the 118 structures with $N = 2$ and $N = 3$ is given in Table 1.

The four regular patterns dominate the experimental material, and for $N = 2$ structures the nanotubular Val-Ala class of structures (Görbitz, 2003*a*, 2007) is the only other major group. In structures indicated to have modified patterns in Table 1, amino–carboxylate interactions are essentially unperturbed, but the carbonyl or carboxylate acceptor of the $C(4)$ or $C(5)$ chains has been replaced by an acceptor in a cocrystallized solvent molecule, in a side chain or by another main-chain carboxylate group (when $N = 3$). Examples are provided as part of the supplementary material.

3.3. The carboxylate binding mode

A closer inspection of Fig. 5 shows that the amino H atoms involved in $C(8)$ chains can be accepted by two different carboxylate O atoms, as for **T5** and **S5**, or both by a single O atom, as for **T4**. These binding modes will be called **A** and **B**, respectively, Fig. 6(*a*).

One hydrogen bond in the **S4** structure in Fig. 5 appears as three-centred. In the following, a bond will be operationally treated as three-centred (abbreviated **c**) when $|d(\mathrm{H}\cdots\mathrm{O}') - d(\mathrm{H}\cdots\mathrm{O}'')| < 0.50$ Å, Fig. 6(*b*). **Ac** and **Bc** notations are used as shown in Fig. 6(*c*) for three-centred transition states between the pure **A** and **B** modes. A summary of carboxylate binding modes for the four basic patterns is given in Table 2 (including 'modified' structures in Table 1).

The **A** mode dominates for all patterns except **T4**, but **S5** includes all modes, as evident from Fig. 6. In general, it will be sufficient to use only the simplified pattern designators **S4**, **T4**, **S5** and **T5**, but whenever appropriate the code may be extended to also indicate the carboxylate binding mode, *e.g.* **S4A** or **S4Bc**.

3.4. Hybrid patterns and other parallel patterns

The seven structures of the 'other parallel' category in Table 1 can be divided into two subgroups, one group of three with individual and unique patterns, and a group of four hybrid structures with $Z' = 2$ or $Z' = 3$ where different molecules show characteristics of different basic patterns. Two examples are Ala-Met hemihydrate (Görbitz, 2003*b*) and Ala-Abu (Görbitz, 2005*a*). In the former one molecule has **S4** connectivity while the second has **T5** connectivity, in the latter two molecules have **T5** connectivity while the third has **T4** connectivity. Details are given in the supplementary material.

3.5. Antiparallel patterns

The eight antiparallel structures in Table 1 constitute a heterogeneous group with seven different hydrogen-bonding patterns. Two regular and related patterns, which are also the most interesting ones in relation to the work performed by

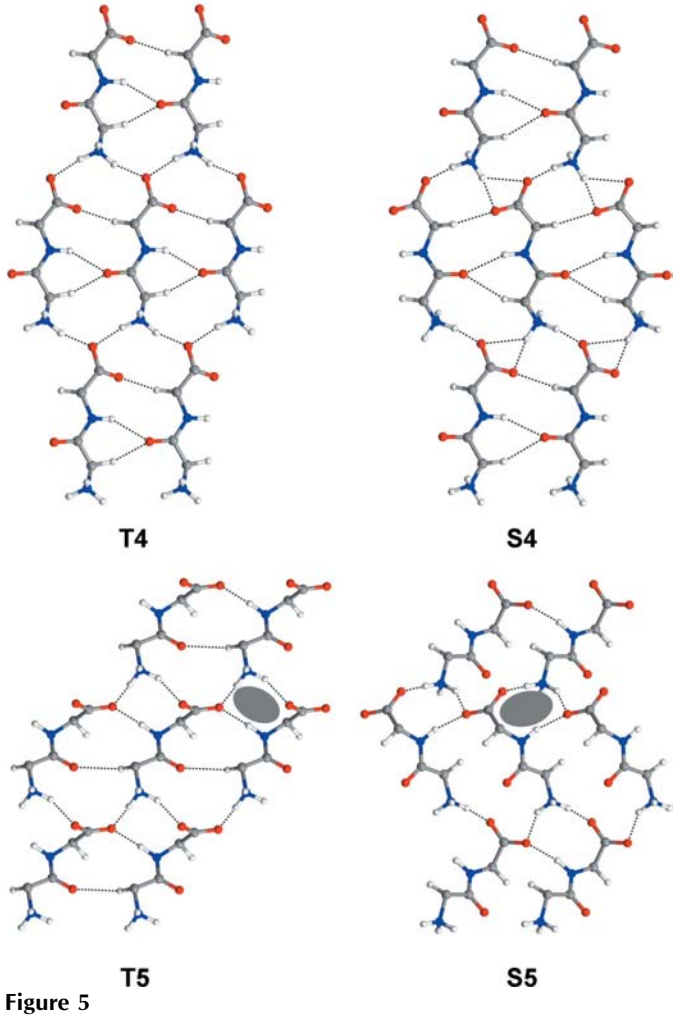

Figure 5
The four basic dipeptide aggregation patterns compatible with the concomitant existence of (at least) two head-to-tail $C(8)$ chains. Ellipses for the **T5** and **S5** patterns highlight characteristic rings with three N—H···O hydrogen bonds and third-level graph set $R_3^2(9)$ (Etter *et al.*, 1990; Grell *et al.*, 2002). The extra C^{α}—H···O(carbonyl) interaction for the **T5** pattern compared with Fig. 2 is quite common. Such contacts may occasionally also be found for **S5** patterns.

Table 2
Carboxylate binding mode as a function of basic hydrogen-bonding pattern.

Pattern	**A**	**Ac**	**Bc**	**B**	**AA**†	**AB**†	**BBc**†	**AAAA**‡	**ABAB**‡
T4	–	–	2	6	–	–	1	–	–
S4	12	3	1	–	1	2	–	–	1
T5	22	2	–	–	3§	–	–	1	–
S5	17	1¶	3¶	2	2	–	–	–	1

† $Z' = 2$. ‡ $Z' = 4$. § **AAc**. ¶ Illustrated in Fig. 6.

S&V, are shown in Fig. 7. Other patterns are described in the supplementary material.

The structure of α-Gly-Gly was among the 27 structures studied by S&V and corresponds to one of their idealized patterns called *ESZA*. S&V also constructed a second antiparallel pattern with the code *EZB*. In 1985 there were no known structures of this type, but amazingly, 14 years later, Akazome *et al.* (1999) found the predicted pattern in the structure of an inclusion complex between (*R*)-nGly-(*R*)-fGly and an organic ester (obtained as an alcohol hydrate with $Z' = 2$), Fig. 7(*b*). The recent structure of Ala-His ethanol solvate hemihydrate (Cheng *et al.*, 2005) provided a second example of a *EZB* structure. The low abundance of antiparallel, β-sheet-like structures, as shown in Fig. 7, is a good indication that peptides do not behave like proteins; hydrogen bonds on the charged termini dominate the structures of dipeptides, but are insignificant in proteins where antiparallel β-sheets involving amide donors and acceptors constitute one of the two important types of secondary structures (with parallel sheets being observed much less frequently).

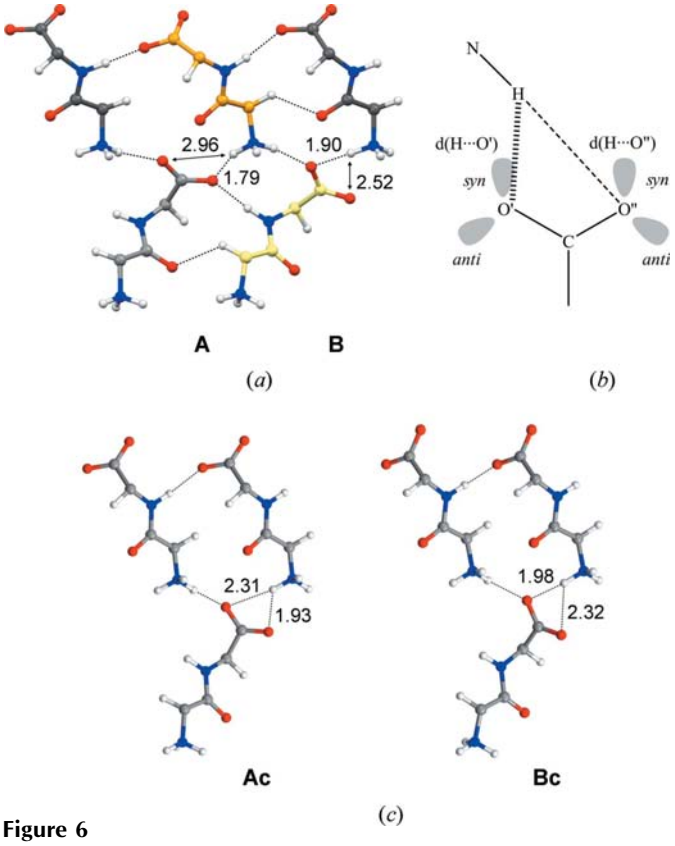

Figure 6
(*a*) The **S5** crystal structure of Leu-Val·C_2H_5OH (Görbitz & Torgersen, 1999) with $Z' = 4$ (C atoms coloured differently) displays two different carboxylate binding modes **A** and **B**. (*b*) H···O distances for a *syn–syn* carboxylate contact. (*c*) The related structures of Leu-Val·C_3H_7OH (Görbitz & Torgersen, 1999) (bottom, left) and bCys-bCys (Capasso *et al.*, 1975) (bottom, right) with intermediate **Ac** and **Bc** modes. H···O distances are indicated hydrogen bonds (dotted) as well as contacts not considered to be such in describing **A** and **B** modes (↔).

3.6. A new twist: hydrogen-bonded tubes

The use of the word 'layer' to describe infinite hydrogen-bonded patterns suggests that these are always two-dimensional entities. A surprising result of the current investigation is that this is not always the case. The structure of Ala-Ala (space group $I4$; Fletterick *et al.*, 1971) may serve as an example, Fig. 8.

When viewed along the tetragonal axis, Fig. 8(*a*), it may readily be dismissed as a 'layered' structure, but a closer inspection of the hydrogen-bonding pattern, provided in Fig. 8(*b*), is quite revealing. It turns out that two $C(8)$ chains generate exactly the same pattern as observed for a normal **T5** structure like Gly-Leu (Patthabi *et al.*, 1974), Fig. 8(*c*), the essential difference being that the two-dimensional layer of the latter has been turned into a one-dimensional tube for Ala-Ala, akin to the way graphite may conceptually be converted into carbon nanotubes. Out of the nine **T4** structures in Table 1, one is nanotubular (Thr-Ala, space group $P4_2$; Görbitz, 2005*b*), while 10 out of the 18 **T5** structures are nanotubular, including Ala-Ala (Fletterick *et al.*, 1971), Leu-

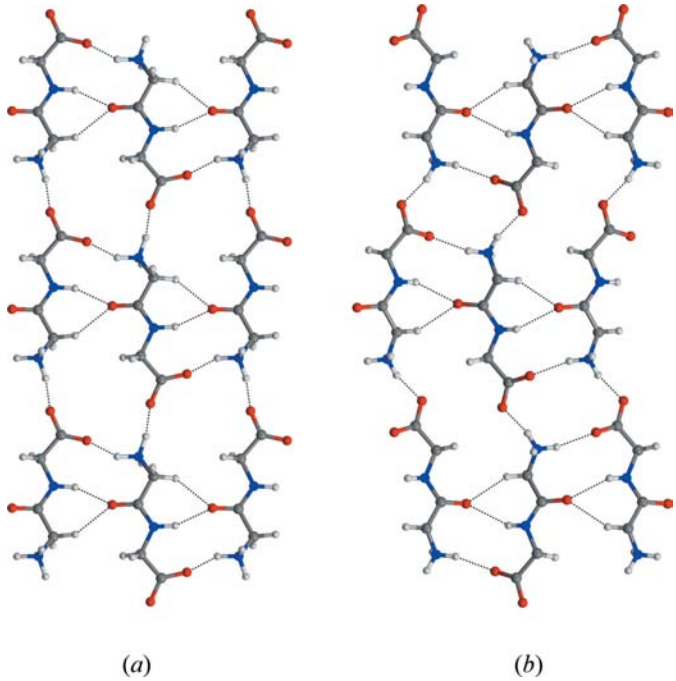

Figure 7
Hydrogen bonding in the structures of (*a*) Gly-Gly (α-form; Kvick *et al.*, 1977) and (*b*) an inclusion complex of (*R*)-nGly-(*R*)-fGly (Akazome *et al.*, 1999). Side chains and solvent molecules have been removed. Both patterns include C^{α}—H···O(carbonyl) interactions.

Table 3
Dipeptide aggregation pattern descriptors used by Suresh & Vijayan (1985) compared with codes used in the present investigation for 16 structures with $N = 2$ or $N = 3$.

Code	*ESA*	*FSA*	*FSB*	*EZA*	*EZB*†	*ESZA*†	*FZA*	*EZA-EZB*‡	*EZA-FZA*‡	none
T4	0 + 2§									
S4				3				1		
T5		4								1
S5							2	1		
anti¶						1				
None										1

† Antiparallel. ‡ Intermediate between two patterns. § Modified, see above. ¶ Further specification needed.

Ser [*C*(5) chain modified; Görbitz *et al.*, 2005] and Phe-Phe (Görbitz, 2001) as well as other members of the Phe-Phe class of structures (Görbitz, 2007). No nanotubular **S4** and **S5** patterns have been observed as they involve crystallographic screw operations that are most likely incompatible with the formation of hydrogen-bonded ring systems (Ala-Ala) or helices (Phe-Phe class and others).

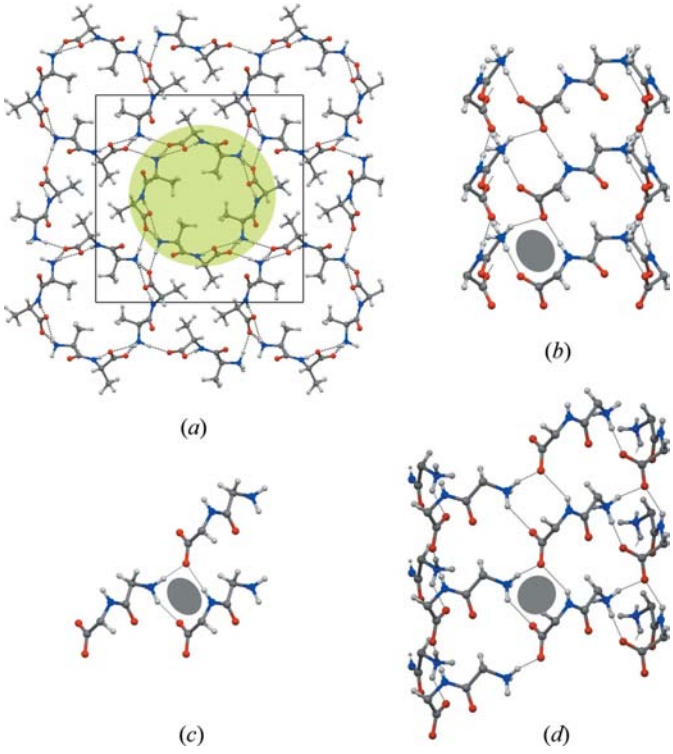

Figure 8
(*a*) The structure of Ala-Ala viewed along the tetragonal axis (Fletterick *et al.*, 1971). The highlighted hydrogen-bonded tube is rotated 90° along the *x* axis and shown in detail in (*b*). (*c*) Hydrogen-bonding detail in the structure of Gly-Leu (Patthabi *et al.*, 1974). (*d*) A hydrogen-bonded tube in the crystal structure of Phe-Phe (Görbitz, 2001; the drawing actually shows the D-Phe-D-Phe structure in order to emphasize similarities with the other structures). Amino-acid side chains for (*b*), (*c*) and (*d*) as well as the rear side of the tubes for (*b*) and (*d*) have been removed to eliminate overlap. $R_3^2(9)$ (Etter *et al.*, 1990; Grell *et al.*, 2002) hydrogen-bonded rings are highlighted by ellipses as in Fig. 5.

3.7. Classification scheme: old and new

A comparison between the original classification scheme by S&V and the new terminology is provided in Table 3 using the original 16 structures with $N = 2$ or 3 as the database.

It can be seen that *ESA* corresponds to **T4**, but that S&V in a sense were a bit unfortunate in not having any unmodified *ESA*/**T4** structures in their experimental material. *FSA* belongs to the **T5** pattern, but so does Ala-Ala (Fletterick *et al.*, 1971), which S&V did not explain in terms of the idealized patterns. *EZA* structures belong to the **S4** pattern, as does the structure of Ala-Ser (Jones *et al.*, 1978*b*), which S&V described as resembling 'in part the *EZA* as well as the *EZB* patterns'. *FZA* structures belong to the **S5** pattern, which also includes the structure of Ala-Asp (Eggleston & Hodgson, 1983) with an arrangement with 'characteristics intermediate between those of *EZA*- and *FZA* type arrangements, ...' according to S&V. Finally, the *ESZA* structure of α-Gly-Gly (Kvick *et al.*, 1977) is just classified as '*anti*' in Table 3. The complexity of this small group (eight structures, seven patterns) suggests that an independent classification scheme is not required; rather it is recommended to name a pattern after the first compound for which it is observed, *e.g.* the α-Gly-Gly pattern.

Overall, the new hydrogen-bond-based terminology offers several advantages over the previous conformation-based descriptors:

(i) Only four common parallel patterns are considered.

(ii) Being based on readily identified hydrogen-bonding patterns, there is no need to consider peptide conformations as defined by the torsion angles.

(iii) The use of hydrogen bonds for classification means that structures not considered to be related by S&V are nevertheless grouped together, including nanotubular structures with unusual dipeptide conformations.

(iv) Few or no borderline cases between patterns.

(v) Easy, readily applicable codes. The final letter (*A* or *B*) of S&V's codes are not self-explicatory, and it is not trivial to understand (or remember) how the peptide conformation actually affects crystal-packing arrangements.

It may be considered a disadvantage that antiparallel patterns are not described in any detail with the new code, but as discussed above, these are so few and diverse that individual descriptions are required in most cases anyway (five out of eight antiparallel patterns in Table 1 do not fit an S&V idealized pattern).

3.8. Structures with $N = 1$ or $N = 0$

S&V noted that some structures, even though one or even both head-to-tail chains had been broken, still resembled the idealized patterns. This is illustrated in Fig. 9.

Figs. 9(*a*) and (*b*) show Ser-Tyr hydrate (Görbitz & Hartviksen, 2008), a typical example of a structure with $N = 1$. The hydrogen-bonding pattern is related to the **S4** pattern shown in Fig. 5, but due to the presence of the bulky Tyr side chain the separation between individual $C(4)$ tapes is increased to the extent that one of the head-to-tail $C(8)$ chains is lost compared with the idealized pattern, Fig. 9(*b*). This generates an **S4*** pattern, where the asterisk in the code indicates a missing $C(8)$ chain compared with the parent pattern. For Tyr-Trp the side chains are even bulkier, and in the **T5**-derived crystal structure shown in Fig. 9(*d*) (Görbitz & Hartviksen, 2008) $C(5)$ chains are forced apart so that no $C(8)$ chain remains ($N = 0$). The resulting **T5**** pattern is shown in Fig. 9(*c*). Amino and carboxylate groups here are involved in hydrogen bonds to cocrystallized water solvent molecules and side chain groups rather than to each other.

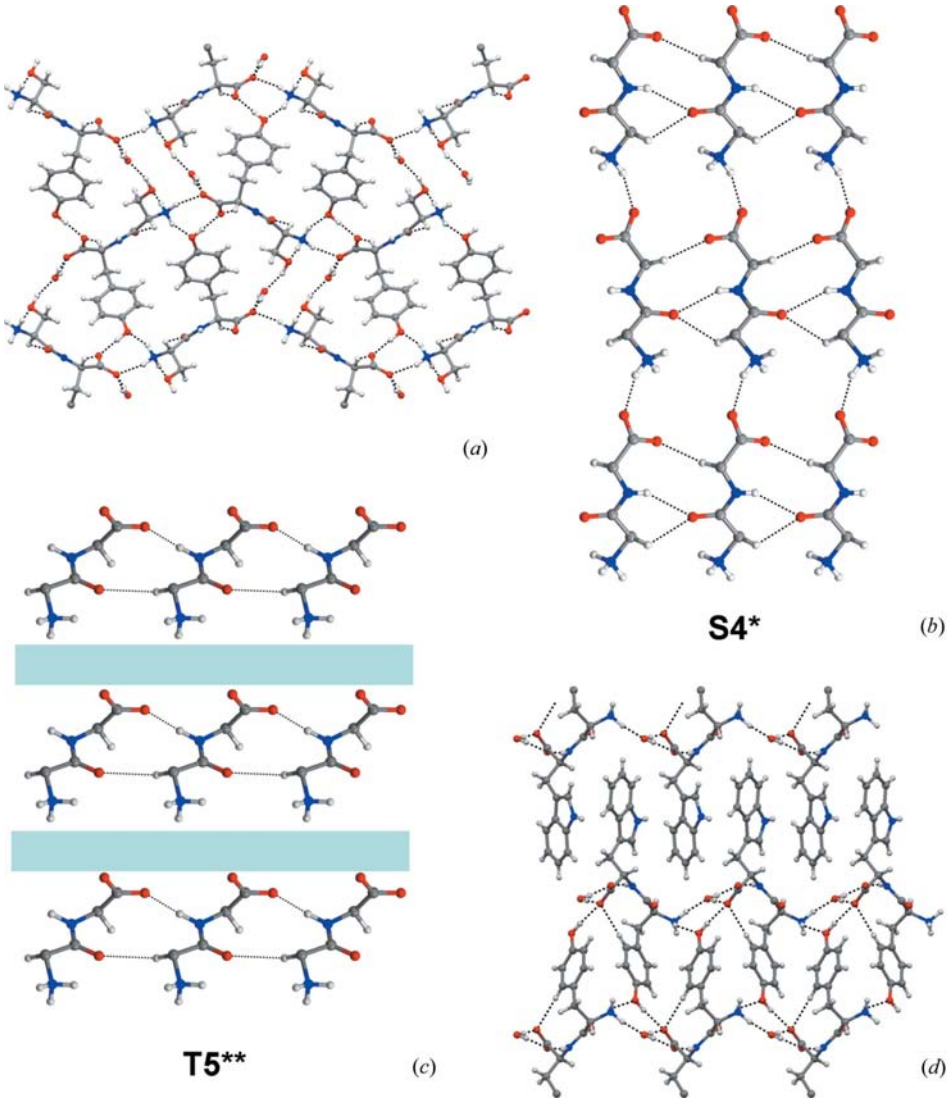

Figure 9
The crystal structures of (*a*) Ser-Tyr hydrate and (*d*) Tyr-Trp hydrate (Görbitz & Hartviksen, 2008) with hydrogen bonding shown in (*b*) and (*c*). The grey rectangles in (*c*) separate $C(5)$ chains that would be in direct contact through $C(8)$ chains in a regular **T5** pattern.

3.9. Less obvious consequences of hydrogen-bonding pattern

The choice of hydrogen-bonding pattern has a more profound impact on the general build-up of the crystal lattice than is immediately realised from Fig. 5.

The characteristic construction of a **T4** structure is exemplified in Fig. 10 by Val-Ser trihydrate (Johansen *et al.*, 2005). Between peptide main-chain layers there is a layer that includes contributions from the side chains of both residues (and usually water molecules, as here), one coming from the hydrogen-bonded layer above (Ser in Fig. 10) and one from the layer below (Val in Fig. 10). **S4** patterns, observed for Leu-Ala benzyl methyl sulfoxide clathrate (Akazome *et al.*, 2005; Fig. 10), and **S5** patterns (not shown) also generate just one type of side chain/solvent region, but it can be easily distinguished from **T4** as the screw symmetry along the $C(8)$ chains with alternating side chain orientations means that either side chain enters the region from both sides. The **T5** arrangement, in its layered version as for His-Leu (Krause *et al.*, 1993) in Fig. 10, is radically different in that the two types of side chains define their own, independent regions in the crystals. The only exceptions to this observation are three structures with *N*-terminal Gly residues where direct contact between neighbouring main-chain layers is achieved (see supplementary material).

3.10. Structure from sequence

In the extended polypeptide chains of proteins the local secondary structure is determined by residue type, a connection used extensively by structure prediction programs. For dipeptides the correlation between sequence and the experimental hydrogen-bonding patterns as well as solvent inclusion is presented in Fig. 11.

3.10.1. The four basic dipeptide aggregation patterns. **T4** is clearly associated with a polar *C*-terminal residue, while the nature of the *N*-terminal residue could be either polar or nonpolar. The tubular structure of Thr-Ala (Görbitz, 2005*b*) is an oddball in this group. **S4** and **S5** patterns occur frequently for dipeptides with two nonpolar or aromatic residues (as inclusion complexes), but to some extent also for nonpolar–polar dipeptides.

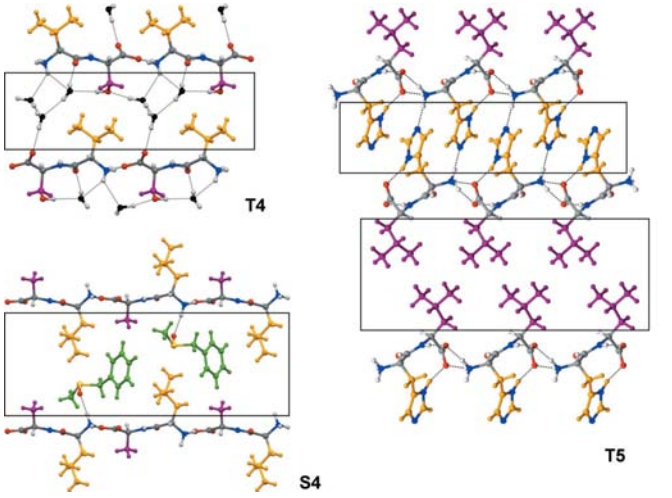

Figure 10
The crystal structures of Val-Ser trihydrate (Johansen *et al.*, 2005) with **T4** patterns, His-Leu (Krause *et al.*, 1993) with **T5** patterns and Leu-Ala benzyl methyl sulfoxide clathrate (Akazome *et al.*, 2005) with **S4** patterns. Side chain C and H atoms in residue 1 and 2 have been coloured in orange and violet, respectively, while C and H of the sulfoxide cocrystallized with Leu-Ala (**S4**) are green. O atoms in solvent water molecules of Val-Ser (**T4**) appear in black. Layers of hydrogen-bonded peptide main chains are seen edge-on, boxes indicate combined side chain/solvent regions.

Reversing the sequence to polar–nonpolar shifts the hydrogen-bonding preference to **T5**, which is also the choice of dipeptides with a *N*-terminal Pro residue. The large group of nanotubular **T5** structures is dominated by entries with bulky hydrophobic or aromatic residues.

As found by S&V, peptide main-chain conformations vary between patterns and it follows that as patterns are sequence-dependent, so are conformations, even in the absence of the traditional intramolecular hydrogen bonds required for folding the polypeptide chain of a protein. Accordingly, **T5** and **S5** patterns lead to rather folded peptide conformations, while extended conformations are found in **T4** and **S4** structures.

The propensity for the formation of hydrates is another highly structure-dependent property. All six layered **T4** structures, including Val-Ser trihydrate (Johansen *et al.*, 2005) in Fig. 10, were obtained as hydrates. For **T5** structures water inclusion depends on structure type: the nanotubular structures of the Phe-Phe class (Görbitz, 2007) and dipeptides with *N*-terminal Pro residues contain water, other dipeptides, such as the polar–nonpolar group, as a rule do not (15 out of 16 structures). **S4** and **S5** structures are not likely to include water; there are just three **S5** hydrates and two **S4** hydrates and none at all for the special nonpolar–polar sequences.[3]

The fact that the **S4** structure in Fig. 10 contains a cocrystallized organic molecule is also not a coincidence, in fact 13 out of the 19 regular **S4** structures (Table 1) contain co-crystallized organic molecules or ions, as do 12 out of the 25 **S5** structures. Remarkably, among **T4**, **T5** and any other group of structures, there are no examples of organic solvent inclusion

[3] The structure of Ala-Ser (Jones *et al.*, 1978*b*) is referred to as a 'hydrate' under CSD refcode LALLSE, but in fact is not.

Pattern	Residue 1	Residue 2: Gly	Np	Ar	Po	A/B	Pro
T4	Gly	0+1[a]					
	Np				**3**		
	Ar						
	Po	0+1	1[b]		**2**		
	A/B					**1**	
S4	Gly						
	Np		*7*	*2*	2		
	Ar			*6*			
	Po		0+1		1		
	A/B				**1**[c]		
T5	Gly		2	1			
	Np		**1**				
	Ar			0+1	0+1		
	Po		5	1	1		
	A/B						
	Pro	**1**	**3**	**1**			
T5[b]	Np		**4**	**1**	0+1		
	Ar	**1**	**1**	**2**			
S5	Gly				**2**		
	Np	1	*6*	*2*	4+1		
	Ar		*1*	*4*			
	Po	1	**1**				
	A/B		3[c]				
O.l.[d]	Gly	1/0/0[d]	0/0/1	1/0/0			0,0,**1**
	Np		*1*/**3**/0	0/0/1	*1*/0/0		
	Ar		0/0/**1**	*2*/0/0			
	Po						
	A/B					**1**/0/0	
	Pro	**1**[c]/0/0					
N.l.[f]	Gly		0/2[f]		0/1		0/1
	Np		**10/3**		1/0		
	Ar						
	Po	0/**1**	0/**1**				
***N* = 1**	Gly	3[g]		**2**	**3**	2[c]	
	Np	**1**	**2**	**1**	**1**		
	Ar		**1**	*2*	**2**		**1**
	Po	**1**		**1**			
	A/B	1[c]				**2**	
***N* = 0**	Gly				**1**		
	Np			**3**			
	Ar		**3**	**3**			
	Po						
	A/B					**3**	
	Pro				**1**	**1**[c]	

Figure 11
Hydrogen-bonding patterns as a function of amino-acid composition. Amino-acid codes: Np (nonpolar) = Ala, Val, Leu, Ile, Met, Abu, Nva; Ar (aromatic) = Phe, Tyr, Trp, fGly, nGly, bCys; Po (polar) = Ser, Thr, Asn, Gln, Asp, Glu, His (uncharged); A/B (acid/base) = Lys, Arg, Asp, Glu, His (charged); Pro = Pro, hPro. Numbers in **bold** and *italic* typeface indicate that $\geq$ 50% of the structures are hydrates or organic cocrystals (generally solvates), respectively; when underlined this is extended to all structures. Boxes highlight structure clusters discussed in the text. [a]Structure with modified pattern. [b]Tubular structure. [c]Overall peptide charge = +1, anion present. [d]Number of antiparallel/hybrid/unique structures for other types of layers. [e]Pro-Sar (Kojima *et al.*, 1980). [f]Number of Val-Ala class (Görbitz, 2003*a*)/others for structures without layers. [g]Three structures with cocrystallized metal salts.

(except the two entries for $N = 1$). This means that any cocrystallization with an organic molecule or organic ion forces introduction of screw symmetry along the head-to-tail chains, a surprising result indeed.

A fine detail in Fig. 11 is the fact that six out of seven structures with modified basic patterns, indicated by +1 in Table 1, occur for sequences with no regular patterns (0 + 1 entries). This suggests that modifications are normally the result of peptides adapting to patterns that are unusual and initially less favourable for their particular sequence.

3.10.2. Other types of structures. Structures with layers other than the basic four (**O.l.** in Fig. 11) constitute a diverse group, while packing arrangements without layers (**N.l.**) are dominated by nonpolar peptides belonging to the Val-Ala class of nanotubular structures (Görbitz, 2003*a*, including Val-Ser trifluoroethanol solvate hydrate, Görbitz, 2005*c*). Other $N = 2$ or 3 structures without layers are, with the exception of Asn-Val hydrate (with $Z' = 3$, Bonge *et al.*, 2005), limited to dipeptides with a Gly residue as the lack of a regular side chain renders arrangements possible that would otherwise be prohibited due to steric conflict.

Out of 27 structures with $N = 1$ only six structures show no sign of being divided into layers. At least 11 of the 21 layered structures can be derived from the four basic patterns, with **S4** being the parent pattern of six structures including Ser-Tyr hydrate, Fig. 9(*a*) (Görbitz & Hartviksen, 2008). As many as 13 dipeptides include a Gly residue, while five incorporate a His residue. All but two structures include either cocrystallized water molecules, anions, metal salts or a combination of these, such as Ala-Gly·LiBr dihydrate (Declercq *et al.*, 1971). Details are given as part of the supplementary material.

While the $N = 1$ group is rather heterogeneous, the 15 structures with $N = 0$ can be readily divided into three separate subgroups:

(i) Nine highly hydrated structures of dipeptides with two bulky side chains. All except Tyr-Tyr dihydrate (Cotrait *et al.*, 1984) are clearly divided into layers, derived from **S5** or more commonly from **T5**, as shown for Tyr-Trp hydrate in Fig. 9(*d*) (Görbitz & Hartviksen, 2008).

(ii) Three unlayered structures of double zwitterions hydrates, such as Arg-Glu dihydrate (Pandit *et al.*, 1983).

(iii) Two structures with *N*-terminal Pro residues.

Only the structure of Gly-His dihydrate (Cheng *et al.*, 2005) does not fit into this pattern.

4. Conclusion

In a database of 160 dipeptide structures, 118 structures contain two or three $C(8)$ head-to-tail hydrogen-bonded chains. Out of these, 97 fit into a new classification scheme with four basic patterns (or hybrids thereof). These are not limited to two-dimensional hydrogen-bonded layers, but also encompass one-dimensional hydrogen-bonded cylinders or tubes. The remaining 21 structures are mostly nanotubular hydrophobic dipeptides. This means that regular head-to-tail patterns are more common than previously thought, and in fact even a majority of the 42 structures with only one or even no head-to-tail chains are structurally related to the four basic patterns, retaining chains of molecules linked by amide >N—H···O—C carbonyl/carboxylate interactions. A careful analysis of the connection between crystal structure and amino acid composition shows that, as for proteins, the outcome of a dipeptide crystallization, not in terms of chain folding but regarding the intermolecular hydrogen-bonding pattern, solvent or guest inclusion and even peptide conformation, is dictated by the amino-acid sequence. The fact that dipeptides behave in a much more rational, or even predictable manner than has been realised in the past paves the way for more structure-directed investigations with short peptides as tools in molecular engineering research.

References

Akazome, M., Hirabayashi, A., Takaoka, K., Nomura, S. & Ogura, K. (2005). *Tetrahedron*, **61**, 1107–1113.
Akazome, M., Takahashi, T. & Ogura, K. (1999). *J. Org. Chem.* **64**, 2293–2300.
Allen, F. H. (2002). *Acta Cryst.* B**58**, 380–388.
Bonge, H. T., Rosenberg, M. L., Riktor, M. & Görbitz, C. H. (2005). *Acta Cryst.* E**61**, o524–o527.
Bruno, I. J., Cole, J. C., Edgington, P. R., Kessler, M., Macrae, C. F., McCabe, P., Pearson, J. & Taylor, R. (2002). *Acta Cryst.* B**58**, 389–397.
Capasso, S., Mattia, C., Zagari, A. & Puliti, R. (1975). *Acta Cryst.* B**31**, 2466–2469.
Cheng, F., Sun, H., Zhang, Y., Mukkamala, D. & Oldfield, E. (2005). *J. Am. Chem. Soc.* **127**, 12544–12554.
Cotrait, M., Bideau, J.-P., Beurskens, G., Bosman, W. P. & Beurskens, P. T. (1984). *Acta Cryst.* C**40**, 1412–1416.
Declercq, J. P., Meulemans, R., Piret, P. & Van Meerssche, M. (1971). *Acta Cryst.* B**27**, 539–544.
Eggleston, D. S. & Hodgson, D. J. (1983). *Int. J. Pept. Protein Res.* **21**, 288–295.
Etter, M. C., MacDonald, J. C. & Bernstein, J. (1990). *Acta Cryst.* B**46**, 256–262.
Fletterick, R. J., Tsai, C. & Hughes, R. E. (1971). *J. Phys. Chem.* **75**, 918–922.
Görbitz, C. H. (2001). *Chem. Eur. J.* **7**, 5153–5159.
Görbitz, C. H. (2003*a*). *New J. Chem.* **27**, 1789–1793.
Görbitz, C. H. (2003*b*). *Acta Cryst.* C**59**, o730–o732.
Görbitz, C. H. (2005*a*). *Acta Cryst.* E**61**, o3735–o3737.
Görbitz, C. H. (2005*b*). *Acta Cryst.* E**61**, o2012–o2014.
Görbitz, C. H. (2005*c*). *CrystEngComm*, **7**, 670–673.
Görbitz, C. H. (2007). *Chem. Eur. J.* **13**, 1022–1031.
Görbitz, C. H. (2010). In preparation.
Görbitz, C. H. & Hartviksen, L. M. (2008). *Acta Cryst.* C**64**, o171–o176.
Görbitz, C. H., Nilsen, M., Szeto, K. & Tangen, L. W. (2005). *Chem. Commun.* pp. 4288–4290.
Görbitz, C. H. & Torgersen, E. (1999). *Acta Cryst.* B**55**, 104–113.
Grell, J., Bernstein, J. & Tinhofer, G. (2002). *Crystallogr. Rev.* **8**, 1–56.
Hughes, E. W. & Moore, W. J. (1949). *J. Am. Chem. Soc.* **71**, 2618–2623.
Johansen, A., Midtkandal, R., Roggen, H. & Görbitz, C. H. (2005). *Acta Cryst.* C**61**, o198–o200.
Jones, P. G., Falvello, L. & Kennard, O. (1978*a*). *Acta Cryst.* B**34**, 2379–2381.
Jones, P. G., Falvello, L. & Kennard, O. (1978*b*). *Acta Cryst.* B**34**, 1939–1942.
Kojima, T., Kido, T., Itoh, H., Yamane, T. & Ashida, T. (1980). *Acta Cryst.* B**36**, 326–331.

Krause, J. A., Baures, P. W. & Eggleston, D. S. (1993). *Acta Cryst.* B**49**, 123–130.
Kvick, Å., Al-Karaghouli, A. R. & Koetzle, T. F. (1977). *Acta Cryst.* B**33**, 3796–3801.
Pandit, J., Seshadri, T. P. & Viswamitra, M. A. (1983). *Acta Cryst.* C**39**, 1669–1672.
Patthabi, V., Venkatesan, K. & Hall, S. R. (1974). *J. Chem. Soc. Perkin Trans. 2*, pp. 1722–1727.
Suresh, C. G. & Vijayan, M. (1983). *Int. J. Pept. Protein Res.* **22**, 129–143.
Suresh, C. G. & Vijayan, M. (1985). *Int. J. Pept. Protein Res.* **26**, 311–328.