Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Recurring main-chain anion-binding motifs in short polypeptides: nests

E. James Milner-White,[a]* J. Willem M. Nissink,[b] Frank H. Allen[b] and William J. Duddy[a]

[a]Division of Biochemistry and Molecular Biology, Institute of Biomedical and Life Sciences, University of Glasgow, Glasgow G12 8QQ, Scotland, and [b]Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, England

Correspondence e-mail: j.milner-white@bio.gla.ac.uk

Received 8 June 2004
Accepted 31 August 2004

A novel tripeptide motif called a nest has recently been described in proteins with the function of binding anionic, or partially anionic, atoms such as carbonyl O atoms. In the present work, a search for nests in small polypeptides stored in the Cambridge Structural Database is reported. 37 unique examples were found: over half form part of hydrogen-bond arrangements resembling those in proteins, such as Schellman/paperclip loop motifs, various types of β-turn and Asx-turns or Ser/Thr-turns, while a third are in novel situations, some involving binding to anionic groups from other molecules within the crystal complex. An example is the antibiotic vancomycin, which incorporates a prominent nest forming part of a peptide-binding site. This nest binds the carboxylate of the C-terminal D-alanine of the bacterial cell-wall precursor peptide, thereby inhibiting the final step of bacterial cell-wall synthesis. As in proteins, a number of nests occur in short peptides with an alternating glycine/L-amino-acid sequence but, uniquely to non-ribosomally synthesized short peptides, several nests within them are constructed from alternating D- and L-amino acids, and such sequences seem to specially favour nests.

1. Glossary

LR and RL: describe the main-chain conformations of the first two amino acids in nests, where R is right-handed and L is left-handed. R signifies a negative φ value, while for L φ is positive. This should not be confused with the configurational isomers of L and D amino acids.

φ, ψ: for amino acid i, φ is the dihedral angle between the C_{i-1}—N_i—C_i^{α}—C_i atoms and ψ is the dihedral angle between the N_i—C_i^{α}—C_i—N_{i+1} atoms ($i-1$ and $i+1$ are the preceding and succeeding residues).

α_R, α_L: α_R is the main-chain conformation of the right-handed α-helix, typically $\varphi = -60$, $\psi = -40^{\circ}$; α_L is the conformation of the left-handed α-helix, typically $\varphi = 60$, $\psi = 40^{\circ}$.

γ_R, γ_L: γ_R is the main-chain conformation $\varphi = -90$, $\psi = 0^{\circ}$; γ_L is the conformation $\varphi = 90$, $\psi = 0^{\circ}$.

Egg: the name given to the anionic or partially anionic atom(s) binding in the nest. Often nests bind two, and sometimes more, such atoms, so there are two or more eggs.

Schellmann loops (also called paperclips): common six-residue hydrogen-bonded motifs (Schellmann, 1980; Milner-White, 1988) that often occur at the C-termini of α-helices. All incorporate nests.

Asx-nest or ST-nest: a type of nest (Watson & Milner-White, 2002) where the first residue is Asp, Asn, Ser or Thr and its side-chain O atom is an egg in the nest.

Compound nest: two or more nests that overlap to form a wider cavity than a simple nest. If residues i, $i + 1$ and $i + 2$ form one nest, residues $i + 1$, $i + 2$ and $i + 3$ form an adjacent one.

2. Introduction

A three-residue recurring motif in proteins has been given the name nest (Watson & Milner-White, 2002*a*,*b*; Pal *et al.*, 2002). Nests are common and between 5 and 8% of amino-acid residues in native proteins in the Protein Data Bank (PDB) are part of one. The characteristic feature, illustrated in Fig. 1(*a*), is a concavity formed by the main-chain NH groups of three successive residues with their H atoms pointing into the nest such that they have the potential to hydrogen bond to one, two or sometimes more anionic or partially negative atoms, especially O atoms. The previous work described their occurrence in proteins. We examine to what extent they are found in small peptides or peptidic molecules, mainly in the Cambridge Structural Database (CSD), and compare them with those in proteins.

In general the PDB stores the three-dimensional structures of biological macromolecules, while the CSD contains those of other chemicals. In the CSD there are many polypeptides, both synthetic and naturally occurring, up to 24 residues long, all of interest to biologists. Being smaller than proteins, their structures are at higher resolution, such that atoms,

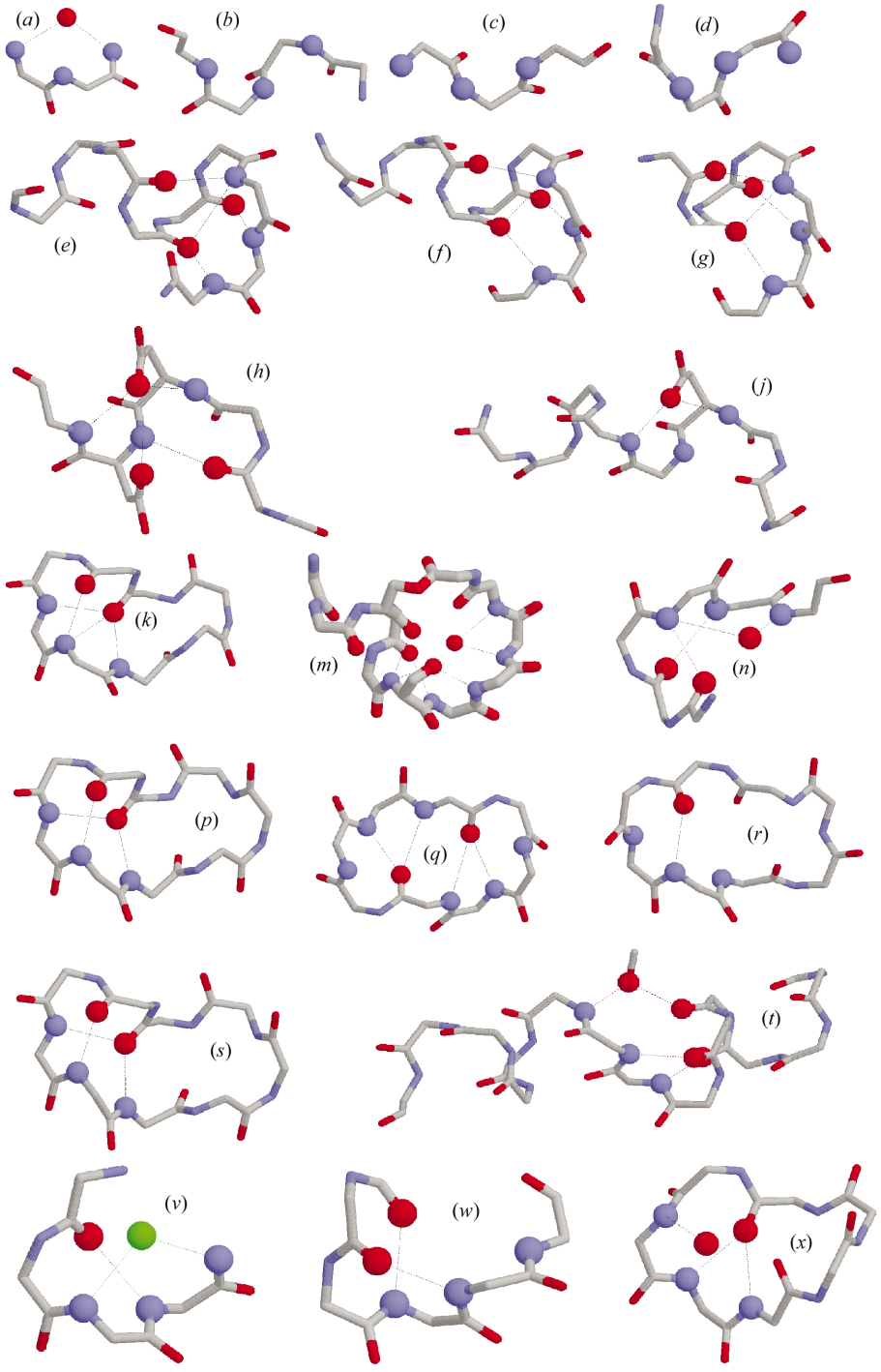

Figure 1
Nest situations within polypeptides. Colour scheme: oxygen, red; nitrogen, blue; carbon, grey; chlorine, green. For each peptide, only the main-chain atoms (N, C^{α}, C and O) are displayed. The N atoms of the nest (three or more) and the atoms hydrogen bonded (or electrostatically attracted) to the nest are shown as spheres. Hydrogen bonds are shown as thin dashed lines. (*a*) A typical RL nest is illustrated by residues 33–35 of protein 1a2p and bound to its egg, the main-chain carbonyl O atom of residue 30. (*b*–*d*) YECDUF, HADFAT and POBJAS are short peptides whose nest NH groups are hydrogen bonded to carbonyl groups from adjacent peptides (not shown). YECDUF is an enkephalin analogue. (*e*–*g*) XESNAK, YIZXIO and NIYXOI mainly consist of α-helices with Schellman loops at their C-termini. (*h*, *j*) SEGXUX and GORVIS incorporate aspartate resides binding to the nest; these are named Asx-nests. (*k*, *p*, *s*) VEYLEQ, JUJHUR and GIPKAR10 are cyclic peptides incorporating Schellmann loops. JUJHUR and GIPKAR10 are variants of cyclinopeptide A. (*m*) GUHGIZ is tensin, a cyclic peptide with a compound LRLR nest. Hydrogen bonded within the nest are a main-chain carbonyl group, a serine side chain forming an ST-nest and a water molecule. (*n*) NUWREC has a nest hydrogen bonded *via* a type I β-turn and a water molecule. (*v*) HIYHAY has a nest with a type I′ β-turn and also a bound chloride ion. (*q*) CUQYUI is a cyclic octapeptide with almost exact dyad symmetry and two nests associated with type I β-turns. (*r*, *x*) CEWCIQ10 and ZORRED are cyclic peptides, each with a nest associated with a β-turn, the ZORRED nest incorporates a water molecule. (*t*) NAHTEV consists of a left-handed α-helix followed by a right-handed one. In between the helices is a nest hydrogen bonded to the O atom of ethanol. (*w*) JOVZAV has a nest associated with a type I β-turn.

Table 1
Nest-containing peptides.

Each row describes a nest-containing peptide. Nest details listed, from left to right, are the peptide's CSD code, the nest type (RL or LR or other), the type of motif with which the nest is associated, the nature of the anionic atom(s) or group bound to the nest (the egg atoms, those binding *via* hydrogen bonding or other electrostatic interactions to the NH groups of the nest cavity; $i-3$ has the meaning that an egg atom is the carbonyl O atom of residue $i-3$ in relation to the nest residues; $i-2$ and $i-1$ are similar), its situation, N-terminal (N) or C-terminal (C), in relation to an α-helix, the peptide's common name (if any) and, at the far right, the peptide's chemical name, with nest residues i and $i+1$ in bold. Amino acids are written with their standard three-letter codes; non-standard three-letter abbreviations for other amino acids are given below. Some amino acids are too complicated to be usefully listed and are designated Xxx. All D-amino acids, however complex, are indicated *via* a D-prefix, while L- or achiral amino acids have no D-prefix. For compound nests (there are two in the table, both of the type LRLR), successive component nests are referred to as *n*1, *n*2 *etc*. Abbreviations for achiral and/or C^{α}-tetrasubstituted amino acids: Aib, α-aminoisobutyryl; Iva, *S*-isovalyl; Dhf, α,β-dehydrophenylalanine; Acp, 1-aminocyclopropane-1-carboxy; Ach, 1-aminocyclohexane-1-carboxy; Dpg, dipropylglycine; Pip, an amino acid incorporating piperidine; Bgl, α-(1,1′-biphenyl-2,2′-dimethylene)glycyl. Abbreviations for chiral amino acids: D-Dmc = D-β,β-dimethylcysteine. Chiral amino acids with complex structures are designated Xxx or D-Xxx. Abbreviations for non-amino acid groups: Boc, *t*-butylcarboxy; Bzoc, benzyloxycarbonyl; Brbz, *p*-bromobenzoyl. More complex non-amino-acid groups are designated Xx.

CSD code†	Nest type	Nest-associated motif	Nest-egg atoms	N or C?	Peptide's common name	Peptide's chemical name (nest amino acids in bold)
Nests in sequences with alternating L- and D-amino acids						
HADFAT	RL	—	Other peptide	—		Boc-**Phe**-**D-Leu**-Thr-O-Me
XESNAK	RL	Schellmann loop	$i-2$, $i-3$, $i-4$	C		Boc-Leu-Aib-Val-Ala-Leu-Aib-**Val**-**D-Ala**-Leu-Aib-O-Me
SEGXUX	LR	LH Asp-nest, type II β-turn	$i-2$, Asp carboxylates	—		Boc-Val-Pro-**D-Asp**-**Asp**-Val-O-Me
VEYLEQ	RL	Schellmann loop	$i-2$, $i-3$	—		Cyclo-(Pro-Pro-Phe-Phe-Ach-**Ile**-**D-Ala**-Val)
GORVIS	LR	Asp-nest	Asp carboxylate	N		Boc-Val-Pro-**D-Asp**-**Aib**-Leu-Aib-Leu-Ala-NH_2
GUHGIZ	LRLR	*n*1: type I′ β-turn, S-nest (cyclized *via* side chains of D-*allo*-Thr and Glu)	$i-2$, $i-3$, $i-4$, HOH, Ser OH	C	Tensin, like SARVIP	Xx-D-Leu-D-Asp-D-*allo*-Thr-D-Leu-D-Leu-**D-Ser**-**Leu**-**D-Gln**-**Leu**-Ile-Glu
NAHTEV	LR	Type I′ β-turn	$i-2$, $i-3$, MeOH	CN		Boc-D-Val-D-Ala-D-Leu-Aib-D-Val-D-Ala-D-Leu-**D-Leu**-**Val**-Ala-Leu-Aib-Val-Ala-Leu-O-Me
JUJHUR	RL	Schellmann loop	$i-3$, $i-2$	—	Cyclinopeptide A1B	Cyclo-(Pro-*cis*-Pro-Phe-Phe-Aib-Aib-**Ile**-**D-Ala**-Val)
CUQYUI	RL	Type I β-turn(s)	$i-2$	—	Two RL nests in a cyclic octapeptide	Cyclo-(Ile-**Thr**-**D-Val**-Aib-Ile-**Thr**-**D-Val**-Aib)
CEWCIQ10	RL	Type I β-turn	$i-2$, HOH	—	(*cf.* JINGAO)	Cyclo-(Gly-Pro-**Phe**-**D-Ala**-Gly-Pro-Phe-D-Ala)
FEPSBC10	LRLR	*n*3: LH Thr-nest	FeO_6 moiety, OH of *allo*-Thr	—	Ferric pseudobactin	Xxx-**D-Xxx**-**Ala**-**D-*allo*-Thr**-**Ala**-D-Xxx
WIPYAV	LR	—	Carboxylate of other peptide	—	Enkephalin analogue (cyclized at disulfide)	Tyr-**D-Dmc**-**Ala**-Phe-D-Dmc
YECDUF	LR	—	Carboxylate of other peptide	—	Enkephalin analogue (cyclized at disulfide)	Tyr-**D-Cys**-**Phe**-D-Dmc
TUCMEJ	LR	—	Carboxylate of acetate	—	Vancomycin	D-Xxx-**D-Xxx**-**Asn**-D-Xxx-D-Xxx-Xxx-Xxx
Nests with an achiral (and/or α-tetrasubstituted) amino acid						
NIVMUA	RL	Schellmann loop-like	'$i-2$, $i-3$, $i-4$'	C		Boc-Leu-Aib-Val-β-alanyl-γ-aminoisobutyrate-**Leu**-**Aib**-Val-O-Me
NUTZUX	RL	Parallel	Other peptide	—		Boc-**Val**-**Dhf**-Ile-O-Me
NUWREC	RL	Type I β-turn	$i-2$, $i-3$, HOH	C		Boc-Gly-Dpg-Gly-**Gly**-**Dpg**-Gly-NH_2
ZOYJUS	RL	Schellmann loop	$i-2$, $i-3$	C		Boc-Val-Val-Dhf-Phe-Ala-Leu-**Ala**-**Dhf**-Leu
POBJAS	LR	Type II β-turn	Other peptide	—		Boc-Pro-**Acp**-**Gly**-NH_2
ZORRED	LR	Type II β-turn	$i-2$, HOH	—		Cyclo-(Leu-**Gly**-**Tyr**-Gly-Pro-Leu-Ile)
PEVJOP	LR	—	Other peptide	—		Boc-**Phe**-**Dhf**-Val-Phe-Dhf-Val-O-Me
KUFKEB	RL	Schellmann loop	$i-2$, $i-3$, $i+4$	C		Boc-Ala-Aib-Ala-Aib-Ala-benzoxyglutamyl-Glu-Ala-Aib-**Ala**-**Aib**-Ala-O-Me
NIYXOI	RL	Schellmann loop	$i-2$, $i-3$, $i-4$	C	(Like TULYOO)	Boc-Leu-Aib-Val-Gly-**Leu**-**Aib**-Val-O-Me
YIZXIO	RL	Schellmann loop	$i-2$, $i-3$, $i-4$	C	(Like JEXSAG, JEXSEG)	Brbz-Aib-Ala-Aib-Ala-Aib-Ala-Aib-**Ala**-**Aib**-Ala-O-Me
VIQBIG	RL	—	Other peptide	—		Bzoc-**Gly**-**Bgl**-Gly-O-Et
HIHYAY	LR	Type I′ β-turn	$i-2$, Cl^-	—	Enkephalin analogue	Tyr-D-Ala-**Gly**-**Phe**-D-Ile
WEHDES	RL	Schellmann loop	$i-2$, $i-3$, $i-4$	C		Boc-Val-Ala-Leu-Aib-Val-Ala-**Leu**-**Gly**-Gly-Leu-Phe-Val-Pro-Gly-Leu-Phe-Val-O-Me
DALSAK	LR	Type I′ β-turn	$i-2$, $i-3$	C		Xx-Ala-Aib-**Pip**-**Ala**-Ala-O-Bu
ZEVBAD	LR	—	Other peptide	—		Boc-Gly-**Dpg**-**Leu**-Val-Aib-Val-O-Me
JOVZAV	RL	Type I β-turn	$i-2$, $i-3$	—	(Iva is a chiral but α-tetra-substituted amino acid)	Boc-Ala-Iva-**Ala**-**Iva**-Ala-O-Me
Nests within sequences with all-L-amino acids						
GIPKAR10	RL	Schellmann loop	$i-2$, $i-3$	—	Cyclinopeptide A (compare JUJHUR)	Cyclo-(Pro-*cis*-Pro-Phe-Phe-Leu-Ile-**Ile**-**Leu**-Val)
ZOHMIS	LR	Type II β-turn	$i-2$	—	Stylopeptide 1	Cyclo-(Pro-**Leu**-**Ile**-Phe-Ser-*cis*-Pro-Ile)
JINGAO	LR	Type II β-turn	$i-2$	—	(Compare CEWCIQ10)	Cyclo-(Gly-Pro-**Phe**-**Ala**-Gly-Pro-**Phe**-**Ala**)

† HADFAT, Doi *et al.* (1993); XESNAK, Aravinda *et al.* (2000); SEGXUX, Fabiola *et al.* (1997); VEYLEQ, Saviano *et al.* (2000); GORVUS, Dhanasekharan *et al.* (1999); GUHGIZ and TULYOO, Henriksen *et al.* (2000); NAHTEV, Banerjee *et al.* (1996); JUJHUR, DiBlasio *et al.* (1992); CUQYUI, Cusack *et al.* (2000); CEWCIQ10, Kopple *et al.* (1986); FEPSBC10, Teintze *et al.* (1981); WIPYAV, Collins *et al.* (1996); YECDUF, Lomize *et al.* (1994); TUCMEJ, Loll *et al.* (1997); NIVMUA, Karle, Pramanik *et al.* (1997); NUTZUX, Dey *et al.* (1996); NUWREC, Karle, Kaul *et al.* (1997); ZOYJUS, Rajashankar *et al.* (1996); POBJAS, Fabiano *et al.* (1993); ZORRED, Morita *et al.* (1995); PEVJOP, Padmanabhan & Singh (1993); KUFKEB, Peersen *et al.* (1992); NIYXOI, Datta *et al.* (1997); YIZXIO, DiBlasio *et al.* (1994); VIQBIG, Formaggio *et al.* (2000); HIHYAY, Deschamps *et al.* (1996); WEHDES, Karle *et al.* (2000); DALSAK, Toniolo *et al.* (1999); ZEVBAD, Karle *et al.* (1995); JOVZAV, Nebel *et al.* (1991); GIPKAR10, DiBlasio *et al.* (1989); ZOHMIS, Pettit *et al.* (1995); JINGAO, Bhandary & Kopple (1991); SARVIP, DiBlasio *et al.* (1992); JEXSAG, Benedetti *et al.* (1990); SARVIP, DiBlasio *et al.* (1992); JEXZAG, Benedetti *et al.* (1990).

including some H atoms, are well resolved. Also, crystal packing has a larger role in determining structure than for proteins, where the interiors at least are little affected. Another difference is the solvent used, which is not always aqueous. The CSD includes certain non-genetically encoded amino acids (Toniolo *et al.*, 2001), notably D-amino acids and the achiral α-amino isobutyrate, abbreviated Aib, with two side chains, like a combination of L- and D-alanine.

Nests can be defined by two alternative enantiomeric (mirror-image) main-chain polypeptide conformations with four characteristic main-chain dihedral angles of two successive amino-acid residues: φ_i, ψ_i and φ_{i+1}, ψ_{i+1} of approximately -90, 0° and 90, 0° or 90, 0° and -90, 0°. Negative φ values are considered to be right-handed (R) and positive φ values left-handed (L). The two types of conformation are therefore described as RL (-90, 0°; 90, 0°) or LR (90, 0°; -90, 0°). Both occur in native proteins.

In proteins the majority of eggs (atoms bound in the nest) are O atoms, typically main-chain O atoms from residues preceding the nest residues in sequence by two or three residues. There are often two eggs and there can be more. Typically, one egg, which can be considered to be the main one, forms strong hydrogen bonds to the NH groups of residues i and $i + 2$ of the nest. The $i + 1$ NH group may also form a hydrogen bond to the main egg, but is frequently angled such that it is available to hydrogen bond to a second egg atom.

Sometimes nests occur with more than two residues with alternating L and R conformations. This gives rise to overlapping nests whose main-chain NH groups all point roughly inwards towards the centre of the nest such that a wider anion-binding site is formed. This arrangement is called a compound nest and is described as RLR, LRL and so on, depending on the dihedral angles. Compound nests, being wider, often form binding sites for anionic groups rather than single atoms. One example is in the P-loop, the conserved part of the commonest nucleotide triphosphate-binding motif in proteins; its LRLR nest binds the β-phosphate of GTP and ATP. Compound nests often occur surrounding the square-planar $[Fe_2S_2][RS_4]$ or cuboid $[Fe_4S_4][RS_4]$ iron–sulfur centres in proteins. These groups of atoms have a net negative charge if all the S atoms are included. It has been suggested (Milner-White & Russell, 2005) that such nests may have had a role in very early evolution.

The residue with left-handed conformation in RL or LR nests is usually glycine in native proteins. This is because the side chains of L-amino acids favour the R rather than the L conformation but, as glycine has no side chain, it is indifferent. D-Amino acids, on the other hand, favour the L rather than the R conformation and since the CSD, unlike the PDB, contains peptides incorporating D-amino acids, examining it should allow an idea of the conformations adopted by such sequences.

A majority of the nests in proteins occur in association with certain hydrogen-bonded motifs (Watson & Milner-White, 2002*a*) of up to six residues. The commonest of these is the Schellman loop (Schellmann, 1980; Milner-White, 1988) which always incorporates a nest. Most Schellmann loops, though not all, occur at α-helical C-termini. Other common motifs sometimes associated with nests are three types of β-turn (Watson & Milner-White, 2002*a*). In the same work, a new protein motif was described called an Asx-nest or ST-nest in which a side-chain atom from the first (aspartate, asparagine, serine or threonine) nest residue binds in the concavity of the nest. It is of interest to see to what extent nests in the CSD occur in association with these or other motifs.

3. Methods

Searching for peptides among the 270 000 entries in the CSD was performed with version 1.3 of the program *CONQUEST* (Cusack *et al.*, 2000; Allen, 2002;, Bruno *et al.*, 2002). To find nests, we searched the entire database for all atom arrangements (not just those labelled as peptide) of the form N—C—CO—N—C—CO—N where the two peptide bonds are *trans* ($\omega = 180 \pm 90^\circ$). Only amino-acid residues potentially able to form anion-binding nests were included; those with, for example, prolines or N-methylated amino acids at the NH groups of putative nests were eliminated. For RL nests all four dihedral angles (for residues i and $i + 1$), φ_i, ψ_i, φ_{i+1} and ψ_{i+1} have to be in the ranges -140 to -20°, -100 to 40°, 20 to 140° and -40 to 100°, respectively, while for LR nests the values for residues i and $i + 1$ are reversed. The angles encompass the α_R and α_L as well as the γ_R and γ_L and other regions. Hydrogen bonds between the relevant N and O atoms were defined by a distance between them of less than 3.8 Å. Short peptides within the PDB were also searched for the same set of angles, by examining the peptides section of the SCOP database, but only within crystal structures of naturally occurring polypeptides (proteolytic fragments were not considered) of less than 20 amino-acid residues.

4. Results

Searches of the CSD for nests found 37 unique examples, of which 19 were RL and 18 LR. There were two compound nests of the form LRLR consisting of three overlapping nests; hence, the 37 unique nests are situated in the 33 polypeptide crystal structures listed in Table 1. In the full name of each peptide given on the right-hand side of Table 1 the first two nest residues are highlighted in bold. In Table 1 nests are divided into three categories depending on the chirality of the nest residues and the relevant features of each peptide and its nest are presented.

Examination of Fig. 1 shows that the main chain (N—C^α—C—N—C^α—C—N) of an LR nest traces an S shape when viewed from the egg, *i.e.* from the hydrogen side of the NH groups. This can be seen in Figs. 1(*b*), 1(*d*), 1(*h*) and 1(*j*). On the other hand, the main chain of an RL nest viewed from the same direction looks like the mirror image of an S shape. One is seen in Fig. 1(*c*). These shapes provide a visual means of finding RL and LR nests when examining polypeptides. They are only seen when viewed from the right direction, which is why the shapes are not visible in the other peptides in Fig. 1.

5. Amino-acid sequences of nests

In proteins 65–75% of nests are of the RL type and 35–25% are LR, the RL/LR ratio being at the higher end if single nests are considered. Possible reasons for this ratio have been discussed (Watson & Milner-White, 2002*a*; Pal *et al.*, 2002) and no incontrovertible explanations have emerged. However, it is conceivable that, just as glycines in α-helices adopt the α_R rather than the α_L conformation, so RL nests are favoured in proteins for the same sorts of reason, *i.e.* to fit in with features preferred by L-amino acids. It is of interest to compare the situation in short peptides. In the admittedly small sample, the RL/LR nest distribution is near to half and half.

The middle part of Table 1 shows that a number of achiral amino acids can take the place of glycine (or the residue with positive φ value) in nests. Most of these amino acids [α-aminoisobutyrate, dipropylglycine, 1-aminocyclopropane-1-carboxy, α-(1,1′-biphenyl-2,2′-dimethylene)glycyl and those containing piperidine] are α-tetrasubstituted (Toniolo *et al.*, 2001). Another α-tetrasubstituted amino acid occurring in the same situation is *S*-isovaline; although chiral, its having two β C atoms makes it appropriate there. A different type of achiral amino acid found at the L residue in a nest is α,β-dehydrophenylalanine, which has a double bond between its α and β C atoms.

From the CSD we learn whether sequences with adjacent (or alternating) D- and L-amino acids might form nests. In the whole database there are 29 candidate peptides exhibiting a pair of such adjacent amino acids and of these 14 incorporate nests (listed at the top of Table 1). These results show that adjacent or alternating D- and L-amino acids are compatible with nests. This is confirmed by model building of nests. Furthermore, although the sample is small and the peptides in it far from randomly selected, the evidence is consistent with such sequences favouring the nest conformation even more than alternating chiral and achiral residues do.

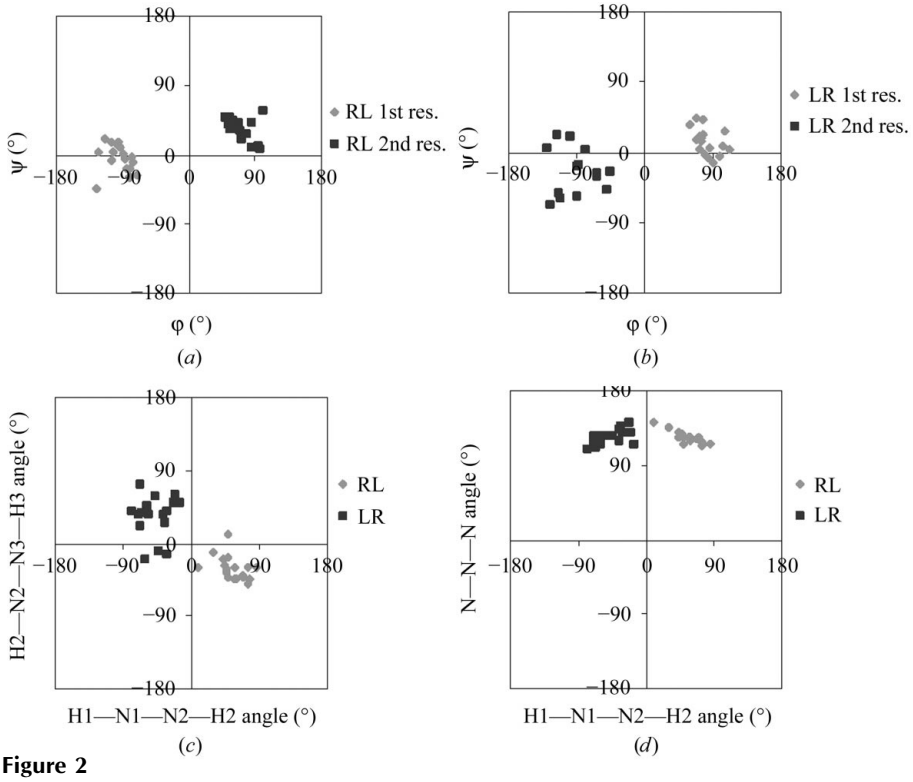

Figure 2
Nest geometry. (*a*) For RL nests in the CSD, angle φ_i is plotted against ψ_i and angle φ_{i+1} is plotted against ψ_{i+1}. (*b*) For LR nests in the CSD, angle φ_i is plotted against angle ψ_i and angle φ_{i+1} is plotted against angle ψ_{i+1}. (*c*) For both RL and LR nests in the CSD the torsion angles H1—N1—N2—H2 and H2—N2—N3—H3 are plotted against each other. The numbers 1 and 2 and 3 represent nest residues *i*, *i* + 1 and *i* + 2. N and H are main-chain atoms. (*d*) For RL and LR nests in the CSD the N—N—N angle is plotted against the H1—N1—N2—H2 angle.

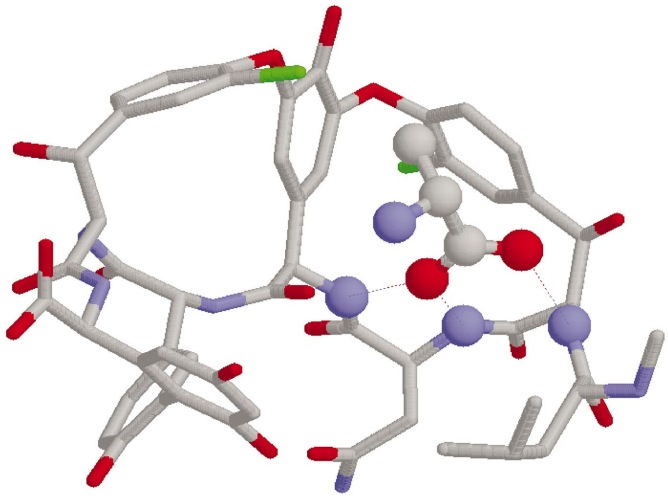

Figure 3
The C-terminal D-alanine binding at the nest in vancomycin (taken from a model of vancomycin with a D-Ala-D-Ala peptide; PDB code 1van; Kelly *et al.*, 1989). The atoms of the C-terminal D-alanine and the main-chain N atoms of the nest in vancomycin are portrayed in ball-and-stick representation. C atoms are grey, O atoms red, N atoms blue and Cl atoms green. Hydrogen bonds are shown as thin dashed lines. The sugar moieties of vancomycin are not shown.

6. Nest geometry

The shapes of the 37 nests have been investigated by analyzing various geometrical parameters. Their φ, ψ angles are shown in Figs. 2(*a*) and 2(*b*) and average values are given in Table 2. The results resemble those for protein nests. The approximate enantiomeric relationship between the two pairs of angles of RL and LR nests is evident in the two figures and the more subtle variations in these angles observed in proteins are also evident in peptides. The torsion angles H1—N1—N2—H2 and H2—N2—N3—H3 are parameters (Pal *et al.*, 2002) that measure the relative directionality of adjacent NH groups.

Table 2
Averages of nest parameters (°).

The average values for the geometrical nest parameters of Fig. 2 are given.

	φ_i	ψ_i	φ_{i+1}	ψ_{i+1}	H1—N1—N2—H2	H2—N2—N3—H3	N—N—N
RL	−94	−5	71	34	54	−31	123
LR	81	15	−84	−25	−48	34	124

Fig. 2(*c*) shows them plotted against each other and Fig. 2(*d*) shows the H1—N1—N2—H2 angle plotted against the NNN angle. The average values are given in Table 2. The results confirm that the geometries of peptide nests are broadly similar to those of protein nests. The corresponding data for nests with alternating D/L- or L/D-amino acids and also for those occurring within cyclic peptides (results not shown) were examined; both sets had similar geometrical parameters to other nests.

7. Nests as part of other motifs

Table 3 compares the situations of nests in the CSD to those in the PDB. A high proportion of RL nests occur in Schellmann loops (Watson & Milner-White, 2002*a*), as seen in Figs. 1(*e*), 1(*f*), 1(*g*), 1(*k*), 1(*p*) and 1(*s*), and also in type I β-turns, as in Figs. 1(*n*), 1(*q*) and 1(*w*). Since Schellmann loops by definition incorporate a type I β-turn, there are similarities between these two motif types. LR nests, on the other hand, often occur in association with type I′ β-turns, as seen in Fig. 1(*v*), or type II β-turns, as in Fig. 1(*x*).

An obvious difference with proteins is that seven of the eggs bound to small peptide nests are C-terminal carboxylate O atoms from another peptide altogether. In previous work such a feature was not found in proteins, probably because an entire protein chain only has one C-terminus. However, side-chain carboxylates, usually from the same polypeptide chain, sometimes bind to nests in proteins.

Many nests, in short peptides as in proteins, occur at the ends of α-helices. RL nests often occur at right-handed α-helical C-termini (Figs. 1*e*–1*g*), while LR nests can occur at the N-termini of right-handed α-helices (Pal *et al.*, 2002; Fig. 1*t*). For each nest this is listed, as N or C, in Table 1. In NAHTEV in Fig. 1(*t*) a nest lies between two helices, one right- and one left-handed, and is thus labelled NC. The RL nests at right-handed helical C-termini occur in association with two alternative motifs, the Schellman loop or, less commonly, the type I β-turn (in proteins 95% of Schellmann loops and 40% of nest-associated β-turns are at α-helical C-termini). In both cases the nests bind the free carbonyl O atoms at the helical C-terminus. In both DALSAK and NAHTEV there happens to be a left-handed α-helix with, at its C-terminus, an LR nest associated with a type I′ β-turn; these features are all of opposite hand compared to those usual in proteins.

Table 3
Occurrence of nests in association with different motifs.

Nests are divided according to the hydrogen-bonded motif category of which they are part. When the main nest atoms are O atoms from the other peptide in the crystal, the motif type is listed as 'Other Peptide'. The relationships between motifs and nests have been described elsewhere (Watson & Milner-White, 2002) but, for nest-associated β-turns (Schellmann loops incorporate such a β-turn too), the defining β-turn hydrogen bond is between the CO of nest residue $i-2$ and the NH of nest residue $i+1$. At the right-hand side is a figure for the percentage of all nests (taking RL and LR together) occurring as the various motifs in native proteins.

	RL, No. in CSD	LR, No. in CSD	Occurrence in proteins (%)
Schellmann/paperclip	9	—	22
Type I β-turn	4	—	8
Type II β-turn	—	5	8
Type I′ β-turn	—	4	4
Other peptides	3	4	0
Miscellaneous	3	5	NA
Totals	19	18	

8. Nest-egg atoms

All 33 nests or compound nests from the CSD have at least one egg atom. This differs from the situation in proteins, where 23% of RL nests and 48% of LR nests appear to be unoccupied (Pal *et al.*, 2002). These observations lead to the suggestion that the empty nests in the PDB result from the low resolution of protein crystals and the presence of many nests at protein surfaces adjacent to motile solvent.

As in proteins, the majority of egg atoms within the CSD nests are O atoms. The commonest are main-chain carbonyl O atoms from preceding amino acids. These are summarized in the nest-egg column in Table 1, where main-chain carbonyl O atoms are listed by their residue number in relation to the nest; for example, $i-2$ means the carbonyl O atom of the residue two residues behind the first nest residue. Apart from main-chain carbonyls, Table 1 shows that carboxylate (five) and hydroxyl (two, excluding HOH, EtOH) O atoms are frequently observed as egg atoms in peptide nests, as in proteins. In the CSD peptides, there are also examples of ethanol (one), water (four) and acetate (one) O atoms binding to nests. There is also a peptide (WEHDES; Fig. 1*v*) with a chloride ion binding in a nest. This, combined with the observation that phosphate groups and iron–sulfur centres often bind nests in proteins, indicates that the egg's negative charge, rather than its hydrogen-bonding ability, governs nest binding.

9. Naturally occurring peptides

The peptides containing nests in Table 1 include several synthesized within organisms rather than in the chemistry laboratory. They include the glycopeptide antibiotic vancomycin from the bacterium *Amycolatopsis orientalis*, the siderophore pseudobactin from the bacterium *Nocardia orientalis* and the possible antifungal agents tensin from the bacterium *Pseudomonas fluorescens*, cyclinopeptide A from linseed and the putative anti-cancer agent stylopeptide from a

marine sponge. There are also three synthetic peptides designed to mimic enkephalins.

The compound LRLR nest in tensin is seen in Fig. 1(*k*). Five main-chain N atoms form a wide nest whose hydrogen-bonding potential is fulfilled by means of a type I β-turn, a serine side-chain hydroxyl group (forming an ST nest) and a water molecule. Ferric pseudobactin is a peptidic siderophore. It has a compound LRLR nest, which binds an O atom of the FeO_6 moiety of the pseudobactin and an OH group on the peptide.

It is intriguing that a nest forms a key functional part of vancomycin. A model of its three-dimensional structure omitting the sugars is shown in Fig. 3, with the three N atoms of the nest displayed as blue spheres. From this it can be seen that the anion-binding site of the nest is at the bottom of a peptide-sized cleft and it is here that the carboxylate group of its ligand binds. Vancomycin inhibits bacterial cell-wall synthesis by binding D-alanyl-D-alanine at the end of a piece of the cell-wall peptide chain, the cross-linking of which a final stage in bacterial cell-wall synthesis. It is the C-terminal carboxylate group that binds to the nest. In Fig. 3 the terminal D-alanine is shown in ball-and-stick mode. The three-dimensional structure of a vancomycin–acetyl-D-alanyl-D-alanine complex was first revealed by NMR (Williams *et al.*, 1983) and the interaction has been much investigated since (Williams & Bardsley, 1999; Knox & Pratt, 1990). In the crystal structure (Loll *et al.*, 1997) listed in Table 1 acetate binds to the antibiotic in place of the C-terminal end of the peptide. In vancomycin-resistant enterobacteria the C-terminal D-alanine is replaced by a D-lactyl residue (Bugg *et al.*, 1991). This still has a carboxylate group that binds to the nest, although its overall affinity is lower.

10. Nest flexibility

One aspect of nests not hitherto considered is their flexibility. Nests have been described as structures with an affinity for anions. However, the converse is also likely, that the presence of appropriate anions encourages formation of the nest structure. Thus, if the anion is lost, the nest conformation may not be retained. One example of this is seen in the P-loop proteins where several (F_1-ATPase, myosin, guanylate binding protein and nitrogenase) have been shown (Abrahams *et al.*, 1994; Ramakrishnan *et al.*, 2002; Ramasarma & Ramakrishnan, 2002) to lose part of the characteristic P-loop nest structure in the absence of a correctly positioned ligand phosphate ion in the nest; the main-chain parts of the first two residues flip such that the LRLR nest becomes a simple LR nest.

The vancomycin structure is of special interest here because it is evident from Fig. 3 that the atoms occurring in the place of the tyrosine-like amino-acid side chains are covalently cross-linked such that the nest is more rigid than in a typical polypeptide. This should cause it to have higher affinity for the carboxylate group of its ligand. This is in addition to the favourable effect already suggested arising from the nest deriving from alternating D- and L-amino acids. It seems evolutionary pressure may have generated a specially strong binding site in vancomycin for a carboxylate group.

Several nests in the CSD occur within cyclic peptides. There are many cyclic di-, tri-, tetra-, penta- and hexapeptides in the CSD, all apparently too constrained to allow nest formation, as none were found. However, larger cyclic peptides can accommodate nests. In our sample there are two heptapeptides (*e.g.* Fig. 1*x*), four octapeptides (see Figs. 1*k* and 1*r*) and two nonapeptides (see Figs. 1*p* and 1*s*). The octapeptide in Fig. 1(*q*) has two symmetry-related nests within it. Five other nests (including vancomycin and pseudobactin) occur within peptides cyclized in various other ways.

11. Short peptides in the Protein Data Bank

To find out how many small peptides are stored in the PDB, we searched there for crystal structures of peptides of less than 20 amino-acid residues. Only naturally occurring peptides and not protein fragments were included. 12 unique peptides were found. One of these, a heat-stable enterotoxin (PDB code 1etn; Ozaki *et al.*, 1991) Cys-Glu-**Leu**-**Cys**-**Cys**-Asn-Pro-Ala-Cys-Ala-Gly-Cys, incorporated a nest and the others had none. The nest residues are in bold. The nest is a compound one of type RLR and the first nest forms part of a paperclip/Schellmann loop.

12. Conclusions

Nests, anion-binding three-residue main-chain motifs, occur frequently in crystals of short peptides, as they do in proteins. All 37 observed in the CSD are bound to an anionic egg atom or group of atoms, whereas in proteins a number of nests (about 30%) are apparently empty. In about half of small polypeptide nests the egg consists of one or two carbonyl O atoms from an amino acid in the same piece of polypeptide and less than three residues away in sequence; in proteins the proportion of occupied nests with eggs of this type is higher, at around 90%. Motifs with which such nests are associated include the Schellman loop, a number of β-turns (types I, II and I′) and Asx- or Ser/Thr-turns; their distribution is broadly similar in small polypeptides and proteins.

In the remaining small peptides with nests the anionic atoms in the nest cavity are main-chain carbonyl O atoms from a separate molecule within the crystal. Examples with O atoms (from water and methanol) and a chloride ion as the anionic atom or group are observed. In seven peptide crystals, the C-terminal carboxylate of one peptide binds to the nest of another peptide. Not surprisingly, such binding has not been seen in proteins.

A striking example is the glycopeptide antibiotic vancomycin and its relatives, which incorporate a rigid nest at the bottom of a cleft with the function of binding the carboxylate group of the C-terminal D-alanine of its target, the bacterial cell-wall precursor peptide. As well as the mode of action of this antibiotic being important in itself, it is interesting because the side chains of the nest amino acids are covalently cross-

linked, making the nest more rigid than usual and therefore presumably more effective at anion binding.

In short polypeptides there are about equal numbers of RL and LR nests, the two mirror-image forms of nests, whereas in proteins there are more RL nests. There are two compound nests in the CSD, both of the LRLR type and from naturally occurring peptides: tensin and pseudobactin.

In both proteins and peptides, sequences with alternating glycine/L-amino-acid sequences favour nests, whereas in the CSD, which includes synthetic as well as genetically encoded amino acids, other achiral amino acids such as α-aminoisobutyrate can take the place of glycine. Alternating D-amino-acid/L-amino-acid sequences also appear to favour nests, perhaps even more strongly than glycine/L-amino-acid sequences. Vancomycin also has adjacent L- and D-amino acids helping to form its nest.

References

Abrahams, J. P., Leslie, A. G. W., Lutter, R. & Walker, J. E. (1994). *Nature (London)*, **370**, 621–628.

Allen, F. H. (2002). *Acta Cryst.* B**58**, 380–388.

Aravinda, S., Shamala, N., Pramanik, A., Das, C. & Balaram, P. (2000). *Biochem. Biophys. Res. Commun.* **273**, 933–936.

Banerjee, A., Ragothama, S. R., Karle, I. L. & Balaram, P. (1996). *Biopolymers*, **39**, 279–285.

Benedetti, E., Di Blasio, B., Pavone, P., Pedone, C., Santini, A., Bavoso, A., Toniolo, C., Crisma, C. & Sartore, L. (1990). *J. Chem. Soc. Perkin Trans. II*, pp. 1829–1837.

Bhandary, K. K. & Kopple, K. D. (1991). *Acta Cryst.* C**47**, 1483–1487.

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.

Bugg, T. D. H., Wright, G. D., Dutka-Malen, S., Arthur, M., Courvalin, P. & Walsh, C. T. (1991). *Biochemistry*, **30**, 10408–10415.

Collins, N., Flippen-Anderson, J. L., Haaseth, R. C., Deschamps, J. R., George, C., Kover, K. & Hruby, V. J. (1996). *J. Am. Chem. Soc.* **118**, 2143–2152.

Cusack, R. M., Grondahl, L., Abbenante, G., Fairlie, D. P., Gahan, L. R., Hanson, G. R. & Hambley, T. W. (2000). *J. Chem. Soc. Perkin Trans. II*, pp. 323–331.

Datta, S., Shamala, N., Banerjee, A., Pramanik, A., Bhattacharjya, S. & Balaram, P. (1997). *J. Am. Chem. Soc.* **119**, 9246–9251.

Deschamps, J. R., George, C. & Flippen-Anderson, J. L. (1996). *Acta Cryst.* C**52**, 1583–1585.

Dey, S., Mitra, S. N. & Singh, T. P. (1996). *Int. J. Pept. Protein Res.* **48**, 123–129.

Dhanasekharan, M., Fabiola, F., Pattabhhi, V. & Durani, S. (1999). *J. Am. Chem. Soc.* **121**, 5575–5576.

DiBlasio, B., Pavone, V., Saviano, M., Fattorusso, R., Pedone, C., Benedetti, E., Crisma, M. & Toniolo, C. (1994). *Pept. Res.* **7**, 55–68.

DiBlasio, B., Rossi, F., Benedetti, E., Pavone, V., Pedone, C., Temussi, P. A., Zanotti, G. & Tancredi, T. (1989). *J. Am. Chem. Soc.* **111**, 9089–9098.

DiBlasio, B., Rossi, F., Benedetti, E., Pavone, V., Saviano, M., Pedone, C., Zanotti, G. & Tancredi, T. (1992). *J. Am. Chem. Soc.* **114**, 8277–8283.

Doi, M., In,Y., Fanaka, M., Inoue, M. & Ishida, T. (1993). *Acta Cryst.* C**49**, 1530–1532.

Fabiano, N., Valle, G., Crisma, M., Toniolo, C., Lombardi, A., Isemia, C., Pavone, V., DiBlasio, B., Pedone, C. & Benedetti, C. (1993). *Int. J. Pept. Protein Res.* **42**, 459–465.

Fabiola, F., Pattabhi, G. V., Raju, E. B. & Durani, S. (1997). *J. Pept. Res.* **50**, 352–356.

Formaggio, F., Crisma, M., Toniolo, C., Tchertanov, L., Gwilken, J., Mazaleyrat, J. P., Gaucher, A. & Wakelsman, M. (2000). *Tetrahedron*, **56**, 8721–8734.

Henriksen, A., Anthoni, U., Nielsen, T. H., Sorensen, J., Christopherson, C. & Gajhede, M. (2000). *Acta Cryst.* C**56**, 113–115.

Karle, I., Gurunath, R., Prasad, S., Kaul, R., Rao, R. B. & Balaram, P. (1995). *J. Am. Chem. Soc.* **117**, 9632–9637.

Karle, I., Kaul, R., Rao, R. B. & Ragonathan, S. (1997). *J. Am. Chem. Soc.* **119**, 12048–12054.

Karle, I., Pramanik, A., Banerjee, A., Bhattacharijya, S. & Balaram, P. (1997). *J. Am. Chem. Soc.* **119**, 9087–9095.

Karle, I. L., Das, C. & Balaram, P. (2000). *Proc. Natl Acad. Sci. USA*, **119**, 12048–12061.

Kelly, J. A., Knox, J. R., Zhao, H., Frere, J. M. & Ghuysen, J. M. (1989). *J. Mol. Biol.* **209**, 281–295.

Knox, J. R. & Pratt, R. F. (1990). *Antimicrob. Agents Chemother.* **34**, 1342–1347.

Kopple, K. D., Bhandry, K. K., Kartha, G., Wang, Y. S. & Parameswaran, K. N. (1986). *J. Am. Chem. Soc.* **108**, 4636–4642.

Loll, P. J., Bevivno, A. E., Korty, B. N. & Axelsen, P. H. (1997). *J. Am. Chem. Soc.* **119**, 1516–1522.

Lomize, A. L., Flippen-Anderson, J. L., George, C. & Mosberg, H. I. (1994). *J. Am. Chem. Soc.* **116**, 429–436.

Milner-White, E. J. (1988). *J. Mol. Biol.* **199**, 503–511.

Milner-White, E. J. & Russell, M. J. (2005). *Origins of Life and Evolution of the Biosphere.* In the press.

Morita, M., Kayashita, T., Takeya, K., Itokawa, H. & Shiro, M. (1995). *Tetrahedron*, **51**, 12539–12548.

Nebel, K., Altmann, E., Mutter, M., Bardi, R., Piazzesi, A. M., Crisma, M., Bonora, G. M. & Toniolo, C. (1991). *Biopolymers*, **31**, 1135–1148.

Ozaki, H., Sato, T., Kubota, H., Hata, Y., Katsube, Y. & Shimonishi, Y. (1991). *J. Biol. Chem.* **266**, 5934–5941.

Padmanabhan, B. & Singh, T. P. (1993). *Biopolymers*, **33**, 613–619.

Pal, D., Suhnel, J. & Weiss, M. (2002). *Angew. Chem. Int. Ed.* **41**, 4663–4665.

Peersen, O. B., Yoshimura, S., Hojo, H., Aimoto, S. & Smith, S. O. (1992). *J. Am. Chem. Soc.* **114**, 4332–4335.

Pettit, G. R., Svirangam, J. K., Herald, D. L., Xu, J.-P., Boyd, M. R., Cichacz, Z. & Erickson, K. L. (1995). *J. Org. Chem.* **60**, 8257–8261.

Rajashankar, K. R., Ramakumar, S., Jain, R. M. & Chauhan, V. S. (1996). *J. Biomol. Struct. Dyn.* **13**, 641–647.

Ramakrishnan, C., Dani, S. & Ramasarma, T. R. (2002). *Protein Eng.* **15**, 783–79.

Ramasarma, T. R. & Ramakrishnan, C. (2002). *Ind. J. Biochem. Biophys.* **39**, 5–15.

Saviano, M., Isemia, C., Rossi, F., DiBlasio, B., Iacovino, R., Mazzeo, M., Pedone, C. & Benedetti, E. (2000). *Biopolymers*, **53**, 189–204.

Schellmann, J. A. (1980). *Protein Folding*, edited by R. Jaenicke, pp. 53–61. Amsterdam: Elsevier.

Teintze, M., Hussain, M. B., Barnes, C. L., Leong, J. & van der Helm, D. (1981). *Biochemistry*, **20**, 6446–6457.

Toniolo, C., Crisma, M., Formaggio, F. & Peggion, C. (2001). *Biopolymers*, **60**, 396–419.

Toniolo, C., Formaggio, F., Crisma, M., Mazalayrat, J.-P., Wakselman, C., George, C., Deschamps, J. H., Flippen-Anderson, J. L., Pispisa, B., Venanzi, M. & Palleschi, A. (1999). *Chem. Eur. J.* **5**, 2254–2264.

Watson, J. D. & Milner-White, E. J. (2002*a*). *J. Mol. Biol.* **315**, 199–207.

Watson, J. D. & Milner-White, E. J. (2002*b*). *J. Mol. Biol.* **315**, 183–191.

Williams, D. H. & Bardsley, B. (1999). *Angew. Chem.* **38**, 1172–1193.

Williams, D. H., Williamson, M. P., Butcher, D. W. & Hammond, S. J. (1983). *J. Am. Chem. Soc.* **105**, 1332–1339.