doi:10.1006/jmbi.2001.4545 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2001) 307, 1427–1450

JMB

# SOCKET: A Program for Identifying and Analysing Coiled-coil Motifs Within Protein Structures

**John Walshaw and Derek N. Woolfson***

*Centre for Biomolecular Design and Drug Development, School of Biological Sciences University of Sussex, Falmer East Sussex, BN1 9QG, UK*

*Corresponding author

The coiled coil is arguably the simplest protein-structure motif and probably the most ubiquitous facilitator of protein-protein interactions. Coiled coils comprise two or more α-helices that wind around each other to form ''supercoils''. The hallmark of most coiled coils is a regular sequence pattern known as the heptad repeat. Despite this apparent simplicity and relatedness at the sequence level, coiled coils display a considerable degree of structural diversity: the helices may be arranged parallel or anti-parallel and may form a variety of oligomer states. To aid studies of coiled coils, we developed SOCKET, a computer program to identify these motifs automatically in protein structures. We used SOCKET to gather a set of unambiguous coiled-coil structures from the RCSB Protein Data Bank. Rather than searching for sequence features, the algorithm recognises the characteristic knobs-into-holes side-chain packing of coiled coils; this proved to be straightforward to implement and was able to distinguish coiled coils from the great majority of helix-helix packing arrangements observed in globular domains. SOCKET unambiguously defines coiled-coil helix boundaries, oligomerisation states and helix orientations, and also assigns heptad registers. Structures retrieved from the Protein Data Bank included parallel and anti-parallel variants of two, three and four-stranded coiled coils, one example of a parallel pentamer and a small number of structures that extend the classical description of a coiled coil. We anticipate that our structural database and the associated sequence data that we have gathered will be of use in identifying principles for coiled-coil assembly, prediction and design. To illustrate this we give examples of sequence and structural analyses of the structures that are possible using the new data bases, and we present amino acid profiles for the heptad repeats of different motifs.

© 2001 Academic Press

*Keywords:* coiled coil; helix packing; hydrophobic core packing; trigger motif; sequence-structure relationships

## Introduction

The coiled coil is a ubiquitous protein-folding motif (Lupas, 1996); current estimates indicate that approximately 5-10% of the sequences emerging from the various genome projects encode coiled-coil regions (Walshaw & Woolfson, unpublished data; (Mewes *et al.*, 2000)). As coiled coils facilitate and cement protein-protein interfaces the possibilities for both cognate and potentially promiscuous protein-protein interactions in any one genome or cell are considerable. Therefore, in this post-genome era, a better understanding of coiled-coil interactions would be useful. For instance, confident recognition of coiled-coil sequences would facilitate protein prediction and design studies, including: the definition of protein domain boundaries; the prediction of potential protein partners; the highlighting of sites for the action of novel diagnostics and therapeutics and the design of peptides targeted to these (Chan *et al.*, 1998; Sharma *et al.*, 1998). Reliable methods for identifying coiled coils in protein structures and sequences would expedite such studies; new prediction methods would lead to relational databases for coiled coils from which sequence-to-structure rules could be gleaned. Here, we describe how we have exploited unique structural features of the coiled

Abbreviations used: PDB, Protein Data Bank; PRC, photosynthetic reaction centre.

E-mail address of the corresponding author: dek@biols.susx.ac.uk

0022-2836/01/051427–24 $35.00/0 © 2001 Academic Press

coil to gather related structural and sequence data for this common motif.

Coiled coils comprise two or more α-helices wound around each other in regular, symmetrical fashions to produce rope-like structures (Figures 1(a) and (b)) (Crick, 1953). The sequence basis of these arrangements are repeating patterns of seven residues, which are often referred to as heptads and labelled *a* to *g* (Figure 1(c)). Usually, there is a consensus of hydrophobic residues at *a* and *d* positions, which form an apolar ''stripe'' on each helix. However, because the heptad repeat falls short of two complete turns of a regular α-helix, successive *a* (and *d*) positions wind around the α-helix surface in the opposite sense to the twist of the helix. Therefore, supercoiling of the helices is required for continuous interfacing of the hydrophobic stripes, and to form the core of the structure (Figure 1).

Based on the above, coiled coils would appear to be one of the most tractable targets for protein-structure prediction and design studies. Indeed in testament to this, reasonable predictors of coiled-coil motifs are available (Berger *et al.*, 1995; Lupas *et al.*, 1991; Wolf *et al.*, 1997; Woolfson & Alber, 1995); and a number of successful, coiled-coil-based designs have been reported (Harbury *et al.*, 1998; Lovejoy *et al.*, 1993; Nautiyal & Alber, 1999; Ogihara *et al.*, 1997). Nonetheless, a variety of coiled-coil types appear to be based on similar heptad patterns (Lupas, 1996): most coiled coils form homo-oligomers, although these may be two, three, four or five-stranded structures. In addition, there are examples of heterotypic coiled coils with two, three and four helices. Furthermore, in some coiled coils the helices are parallel, while in others they are anti-parallel. Finally, intra-chain coiled-coil interactions also occur, most commonly in

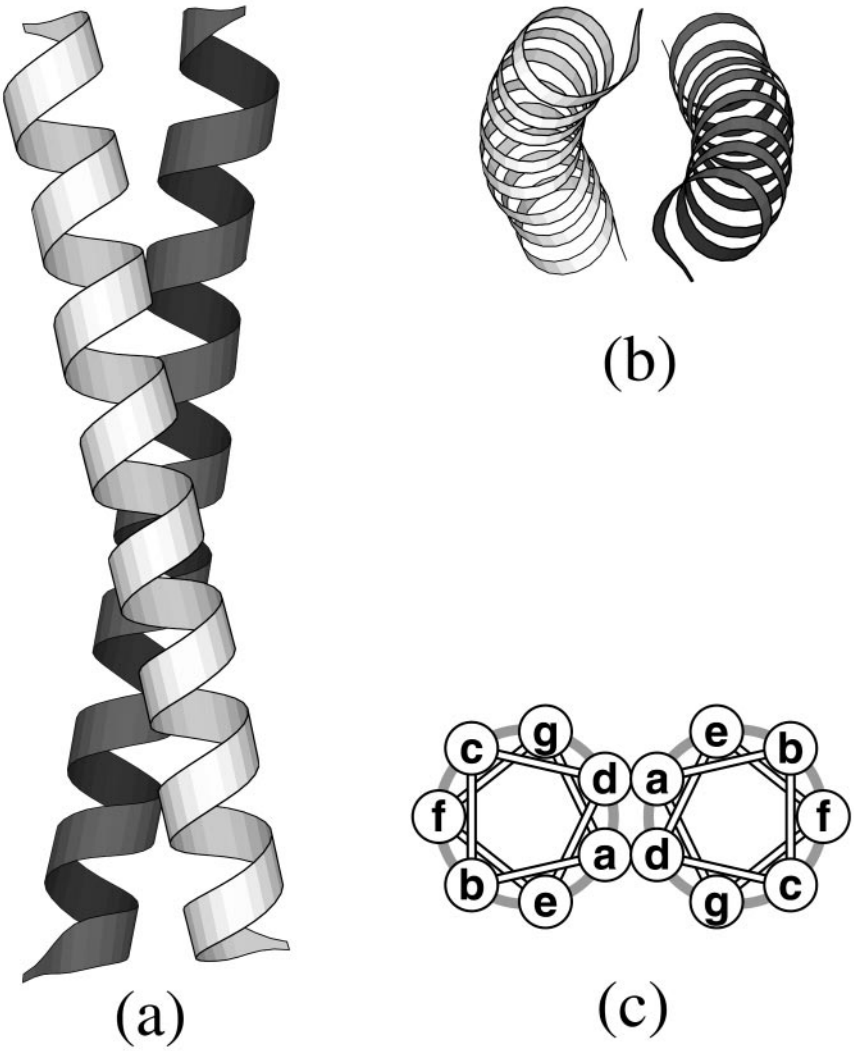


**Figure 1.** The leucine-zipper region from GCN4 (O'Shea *et al.*, 1991), a typical parallel dimeric coiled coil. (a) and (b) Orthogonal views of the backbone structure. (c) Helical-wheel representation showing the orientation and interaction between the two helices and indicating the heptad positions. This diagram assumes supercoiling of the two helices and uses 3.5 residues per turn.

anti-parallel helix-loop-helix motifs, though longer loops separating parallel strands are observed. Therefore, the problem of coiled-coil recognition and prediction is not limited to spotting heptad repeats.

To date, several directors of coiled-coil oligomerisation state and helix orientation have been elucidated. For instance, amino acid selection at the *a* and *d* positions strongly influences oligomer-state selection (Harbury *et al*., 1993, 1994; Woolfson & Alber, 1995), and electrostatic interactions between side-chains at *e* and *g* sites from neighbouring helices help specify binding partners (Kohn *et al*., 1998; O'Shea *et al*., 1992, 1993; Vinson *et al*., 1993). Nevertheless, the determinants of the number and identity of partner helices are not fully understood.

In our view, major remaining problems in coiled-coil research include: (i) What are the limitations on coiled coil topology, i.e. what helical arrangements are possible? (ii) What are the sequence-to-structure rules that link the heptad repeats to these arrangements? (iii) What guides partner selection in coiled coils? (iv) What features lead to high-order assemblies of coiled-coils, for example, as are observed in intermediate filaments (Steinert, 1993), SNARE complexes (Sutton *et al*., 1998) and spindle pole bodies (Wigge *et al*., 1998)?

The availability of reliable databases of positive coiled-coil structures and their associated sequences would provide a platform for addressing these issues. However, to achieve this, reliable coiled-coil recognition algorithms are required.

Several coiled-coil prediction schemes have been proposed. However, these have met with varying degrees of success. The methods that predict coiled-coil regions include COILS (Lupas *et al*., 1991), COILER (Woolfson & Alber, 1995), PAIRCOIL (Berger *et al*., 1995) and MULTICOIL (Wolf *et al*., 1997), with COILS being the most widely used. To our knowledge, no algorithms exist to predict partner selection in coiled-coil systems, i.e. predicting the preference for making homotypic or heterotypic interactions. A major concern for prediction, however, is that even when predicting "simply" where coiled-coil motifs occur in linear sequences, there is significant disagreement between results from the above, most commonly used methods, which goes beyond a simple degree of conservatism (Walshaw & Woolfson, unpublished data).

Here, we focus on problem (i) outlined above. Our aim was to collate sets of known coiled coils to determine the structural limits for these motifs and, where possible, to provide relational data bases of sequences and coiled-coil parameters for the different structural types. This work should aid studies geared at tackling the other questions, although problem (iv) is beyond the scope of the present work. To effect the study, we required a new means of identifying coiled-coil structures in protein structural databases.

Our approach focused on the packing arrangement particular to coiled-coil structures; namely, the interaction between the α-helices termed "knobs-into-holes" packing, which was first postulated almost 50 years ago by Crick (1953). Crick considered the dimensions of the α-helix and the positions of the α-carbon atoms within it. In short, for two helices with heptad repeats and an appropriate supercoil twist, the $C^{\alpha}$ atoms are able to interlace over an indefinite stretch (Figure 2). In Crick's (1953) packing scheme, every first and fourth residue of each heptad is a "knob", which fits into a diamond-shaped "hole" formed by four residues on another α-helix. Three of the four residues of this diamond are themselves knobs, so that a complementary interlocking structure results for two-stranded coiled coils. This basic model was first confirmed at atomic resolution for the dimeric, leucine-zipper motif (O'Shea *et al*., 1991). Crick also proposed trimeric coiled coils. Trimers, tetramers and a pentamer have all since been observed experimentally. In these structures a cyclic pattern of knobs-into-holes packing was predicted (Crick, 1953), which has also been confirmed by others (Harbury *et al*., 1993, 1994).

Since Crick's (1953) proposal, it has been shown that, with some caveats, the interlacing of α-carbon positions is a general feature of helix-helix packing, and is not restricted to heptad repeats and coiled coils (Walther *et al*., 1996). However, as emphasised elsewhere (Efimov, 1999), in knobs-into-holes packing it is the side-chains, not the α-carbons, which form the interdigitating interface. The nature of this interface is quite distinct from, and, as we show, may be distinguished from, the general model of α-helix packing in globular domains called "ridges into grooves" (Chothia *et al*., 1981). However, knobs-into-holes packing may be considered as one extreme of this scheme.

For these reasons, we did not explicitly search for classic coiled-coil attributes such as sequence repeats alone, or, at the structural level for features such as pitch, ideal symmetry, interhelical angles and distances. Rather, we were concerned only with the knob-and-hole interaction. This proved straightforward to describe in terms of the relative spatial arrangement of side-chains (Methods and Definitions). On this basis, we developed the algorithm SOCKET and applied it to cull a complete set of unambiguous coiled-coil structures from the RCSB Protein Data Bank (Berman *et al*., 2000). We illustrate the possibilities for such a database with a variety of sequence and structural analyses.

## Results and Discussion

### The design and development of SOCKET

For a full set of definitions of the terms used below, which are highlighted in bold, the reader is referred to Methods and Definitions.

We sought to identify coiled-coil motifs in protein structures based on structural criteria alone; that is, without the need to turn to sequence analysis. We focused on the packing interaction pro-

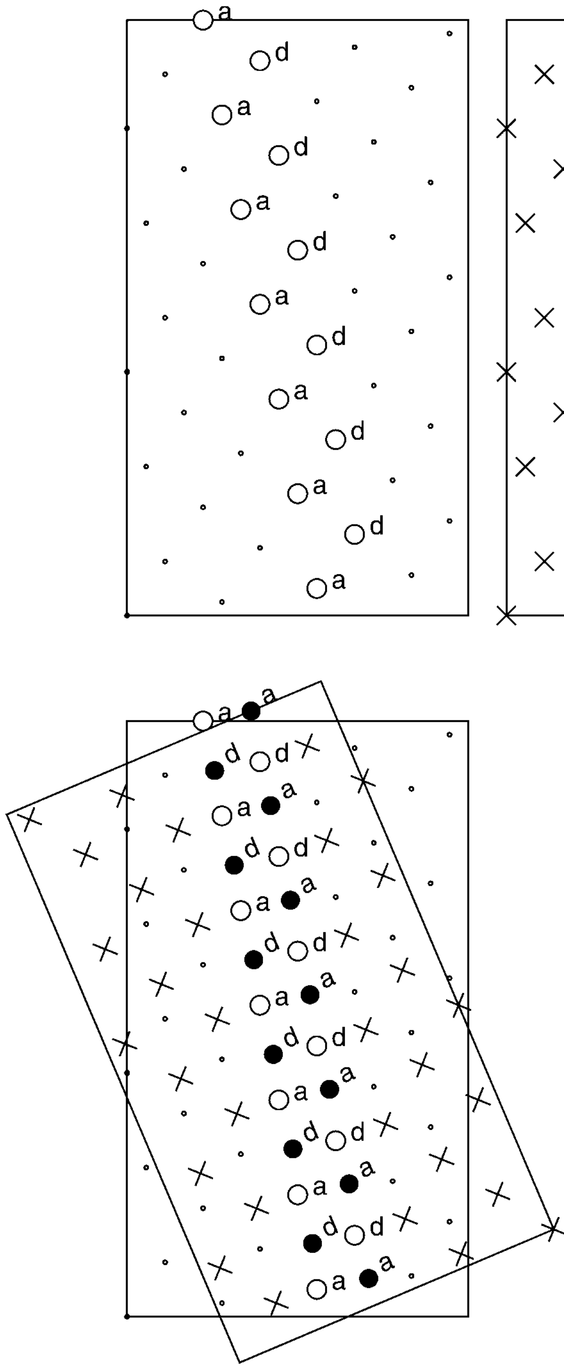


**Figure 2.** Helical-net representation of a coiled-coil interface (Crick, 1953). The uppermost figures show the external surfaces of two, identical α-helices viewed with the N termini at the top, and indicate the relative positions of the $C^{\alpha}$ atoms. The core *a* and *d* residues are highlighted and distinguished for each helix by hollow and filled circles. The remaining positions are shown and distinguished as dots and crosses. The lower diagram shows the interlacing of core (and other) positions when the two surfaces pack against each other; n.b. in this view, the slanted surface is now effectively being viewed from the inside of the helix.

posed by Crick in which a knob side-chain from one α-helix fits into a hole comprising four side-chains from *one other* α-helix (Crick, 1953), Figure 3. To achieve this, all residues were represented by a centre of mass. A side-chain was classed as a knob if it contacted four or more side-chain centres within a specified packing cutoff; the nearest four side-chains were taken as the corresponding hole. The lower the packing cutoff at which a structure exhibits knobs-into-holes, the closer-knit the side-chain intercalation. Packing cutoffs were determined empirically as described below.

We designate isolated cases of a knob in a hole as Type 1 and Type 2; complementary interactions,

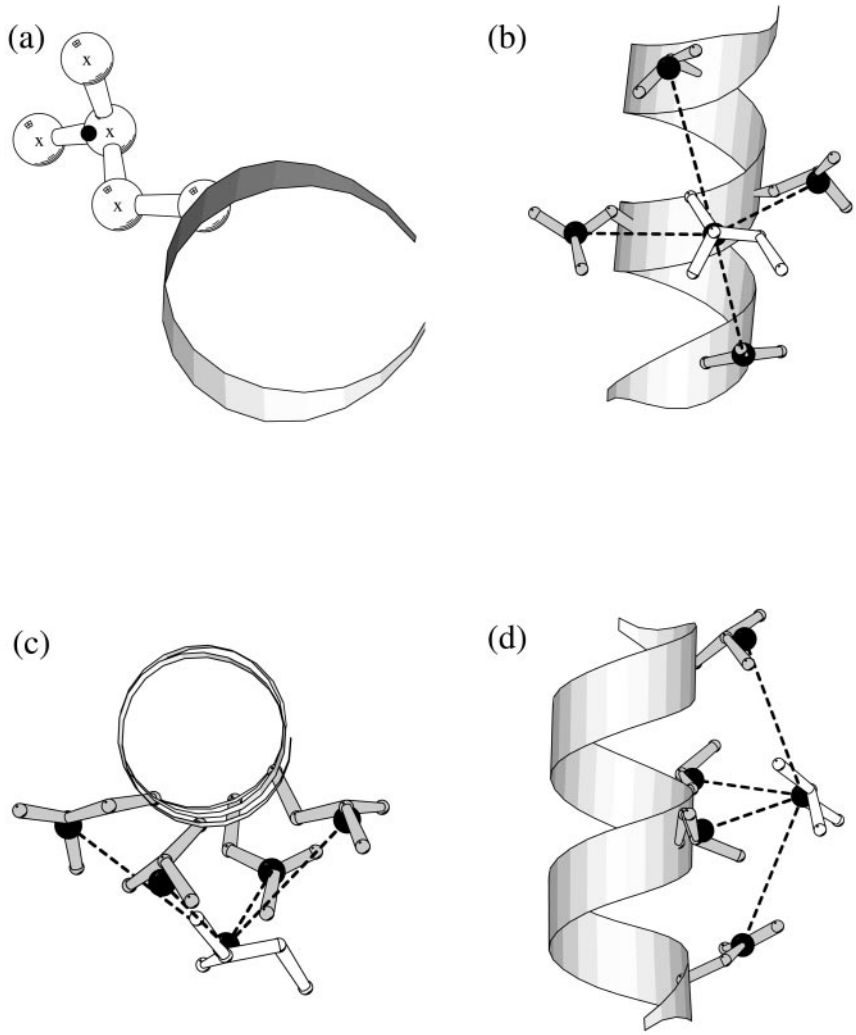


**Figure 3.** Knobs, holes and knobs-into-holes packing. (a) The centre of a side-chain, indicated here by a black dot, is represented by the mean of the co-ordinates of the side-chain; ''X's'' mark the co-ordinates of the side-chain atoms used. (b), (c) and (d) Orthogonal views of a knob in a hole. Here, distances between the centres of the knob and those of the four hole side-chains are all below a specified packing cutoff.

in which each knob is itself part of a hole, are Types 3 and 4 (Figure 4). Type 1 and 3 knobs are positioned across the hole rather than strictly inside it (as determined by an insertion cutoff), but still meet the packing-cutoff criteria. Such conformations of long side-chains are observed in some classic coiled coils (see Methods and Definitions).

True knobs-into-holes helix-helix packing requires complementary interactions (Crick, 1953). These come in two forms (though the first is the simplest extreme of the other). True two-stranded coiled coils exhibit pairwise-complementary knobs-into-holes interactions. This means that when a knob from helix X fits into a hole formed by four side-chains of helix Y, one of these side-chains on Y is itself a knob, which fits into a hole comprising four side-chains of helix X (Figure 5). The arrangement is complementary because one of the hole residues on X is also the first mentioned knob. In the above terminology, coiled-coil helical interfaces have Type 3 or Type 4 knobs-into-holes. Higher-order coiled coils have cyclically complementary knobs-into-holes interactions. In this case, the knob from X again fits into a hole on Y as described above. Although one of these hole residues is a knob it does not fit back into a hole on X. Rather, the knob on Y interacts with a hole on a third helix Z. In a three-order coiled coil, one of the side-chains that forms the hole on Z is a knob that fits into a hole on helix X to complete the cycle (Figure 6).

In order to identify all possible orders of coiled-coil structure, SOCKET was written to locate knobs-into-holes interactions and to find cycles of these. Early applications of the algorithm to the RCSB Protein Data Bank (PDB) returned a number of perpendicular pairs of neighbouring helices, which, nonetheless, had a single pair of pairwise-complementary knobs-in-holes. Therefore, to define a two-stranded ''coiled coil'', we set an additional

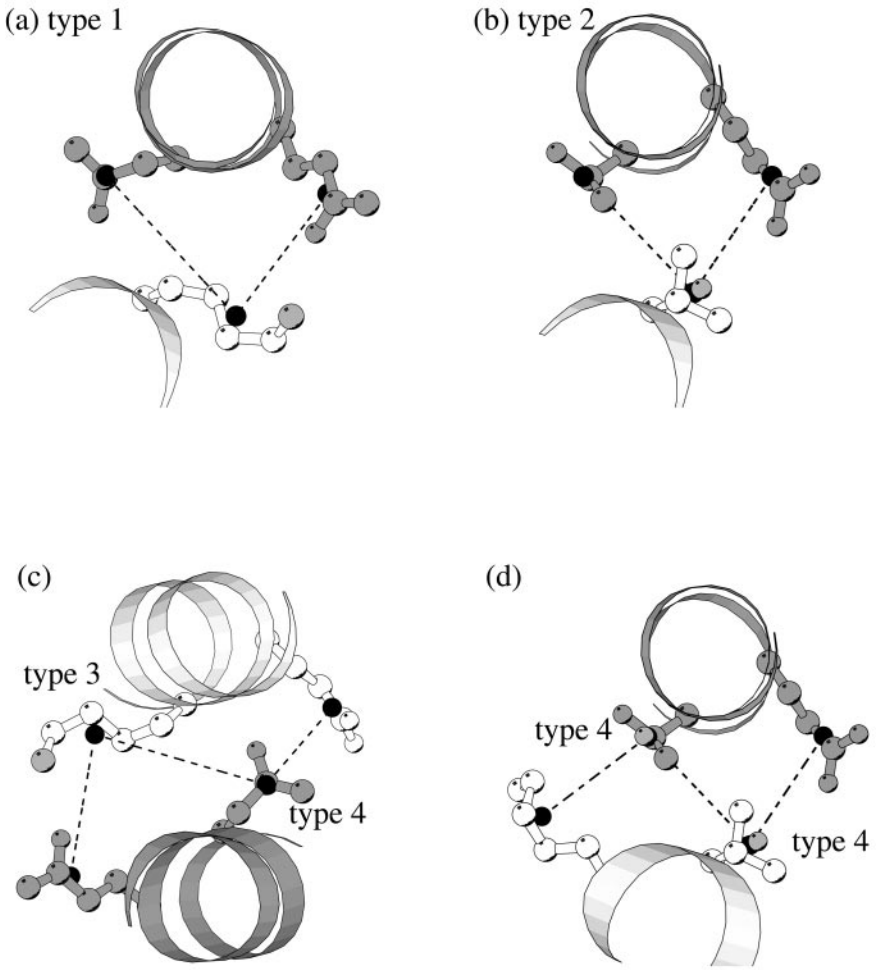


**Figure 4.** Ball-and-stick representations of different types of knobs-into-holes. (a) Type 1, a ''knob-across-a-hole''. Only the two sides of the hole are shown. The knob is in white. The distances between side-chain centres (marked as black discs) are below the packing cutoff, but the end of the knob side-chain (the grey atom) is too far from the left-most hole side-chain to be described as lying in the hole. (b) Type 2, as in (a), but the end of the knob side-chain (the grey pseudo-atom) is within the insertion-cutoff to all hole-side-chain centres. (c) and (d) Complementary pairwise (order-2) knob-into-hole interactions. (c) Shows an example of a Type 3 combined with a Type 4 knob. The packing-cutoff is satisfied by both knobs (white, top left; and grey, bottom right), but the white knob lies across its hole (Type 3). (d) Both knobs are Type 4, in their respective holes.

requirement for *at least* two pairwise-complementary knobs-into-holes. This meant that a two-stranded coiled coil could effectively be as short as a single heptad. By contrast, for higher ($N$) order coiled coils, we defined the presence of even a single, complete $N$-order cyclically complementary knobs-into-holes as sufficient to designate $N$ helices as belonging to an $N$-stranded coiled coil. By this definition coiled coils above dimer may comprise only a single layer of knobs, i.e. effectively ''half a heptad''. A final noteworthy point is that in higher-order coiled coils SOCKET located knob residues and corresponding holes beyond the classical core ($a$ and $d$) positions; side-chains at $e$ and $g$ sites of the heptad repeat acted as knobs increasingly in three, four and five-stranded coiled coils. We refer to these interactions as peripheral knobs and holes.

## Determination of inter-residue distance thresholds to specify knobs-into-holes packing

SOCKET was tested on several classical coiled coils of different oligomer states and orientations, and also on some non-coiled-coil α-helical domains. The minimum packing cutoffs at which any knobs-into-holes interactions were identified in these structures are shown in Table 1. This indicated that a packing cutoff of 7.0 Å was sufficiently large to observe all the expected core packing interactions in the classical coiled coils while excluding other types of packing in globular domains. While several type 1 and type 2 knobs appeared between neighbouring helices in the latter (Figure 7), layers of complementary knobs (Type 3 and 4 knobs) were only found in the coiled coils (Figure 8). In addition, using this cutoff numerous cases of the

**Table 1.** Knobs-into-holes packing in classic coiled coils and control structures

| Protein structure | PDB | Resolution (Å) | Minimum packing cutoff (Å) | Packing cutoff=7.0 Å: Non-complementary knobs: Type 1 | Non-complementary knobs: Type 2 | Complementary knobs: Type 3 | Complementary knobs: Type 4 | Complete layers of knobs |
|---|---|---|---|---|---|---|---|---|
| GCN4 leucine zipper | 2zta | 1.8 | 5.8 | 0 | 0 | 0 | 14 | 7 |
| C-Myc-Max leucine zipper | 1a93 | n/a | 6.4 | 0 | 1 | 1 | 9 | 5 |
| Seryl-tRNA synthetase (2) | 1ses | 2.5 | 6.1 | 1 | 15 | 1 | 15 | 8 |
| F1-ATPase | 1bmf | 2.85 | 6.3 | 2 | 37 | 2 | 12 | 7 |
| Replication terminator protein | 1ecr | 2.7 | 6.8 | 0 | 2 | 1 | 9 | 5 |
| Influenza virus hemagglutinin | 1htm | 2.5 | 6.2 | 0 | 12 | 4 | 34 | 10 |
| Mannose-binding protein-A | 1rtm | 1.8 | 5.8 | 1 | 1 | 0 | 15 | 5 |
| SNARE complex (3) | 1sfc | 2.40 | 6.3 | 10 | 72 | 7 | 117 | 3 |
| Repressor of primer | 1rpr | n/a | 6.5 | 0 | 4 | 2 | 34 | 3 |
| Haemoglobin (4) | 2hhb | 1.74 | 9.0 | 1 | 34 | 0 | 0 | 0 |
| Farnesyl diphosphate synthase | 1fps | 2.6 | 7.2 | 3 | 12 | 0 | 0 | 0 |
| Engrailed homeodomain (2) | 1hdd | 2.8 | 9.3 | 0 | 2 | 0 | 0 | 0 |
| Antitermination factor nusb | 1baq | n/a | 9.4 | 0 | 2 | 0 | 0 | 0 |
| Subtilisin novo | 1cse | 2.8 | 12.0 | 0 | 0 | 0 | 0 | 0 |
| p18-ink4c(ink6) (2) | 1ihb | 1.95 | 9.6 | 0 | 5 | 0 | 0 | 0 |
| Acyl-coenzyme a binding protein | 2abd | n/a | 8.7 | 1 | 2 | 0 | 0 | 0 |

Knobs-into-holes packing interactions identified in classic coiled-coil structures (top) and control structures (bottom). Where there was more than one monomer in the PDB file, the oligo number is shown in brackets (column 1). The minimum packing-cutoff at which complementary knobs-into-holes appear is shown in column 4. The number of non-complementary and complementary knobs (side-chains inserted in a group of four side-chains of a neighbouring helix) is shown in columns 5-8. The number of layers of complementary knobs (pairwise for two-stranded coiled coils; cyclic for others) is shown in the last column.

aforementioned peripheral knobs were identified in the four-stranded coiled-coil structures, and a small number in the three-stranded. 7.0 Å was therefore used as the default cutoff for the evaluation of the PDB. A more liberal cutoff of 7.4 Å was also used to identify ''marginal coiled coils''. An insertion cutoff of 7.0 Å proved to differentiate between knobs-into-holes and knobs-across-holes interactions (data not shown).

Our analysis of the whole PDB revealed that there was not a great difference between the results using the default (7.0 Å) and liberal (7.4 Å) packing cutoffs. This was true with the exception of the number of two-stranded coiled coils that were returned, which, as discussed below, dropped rapidly with cutoff. For this class of motif, the distinction between knobs-into-holes and other modes of helix-helix packing, therefore, appeared to be the most ''blurred''. For the remainder of this paper results refer to structures retrieved using a packing-cutoff of 7.0 Å except where stated.

## Knobs-into-holes based structures in the complete PDB

The numbers of positive coiled-coil structures at each cutoff are listed in Table 2. As indicated, these were grouped into a number of sequence-based homologous families. We found classical examples of parallel and anti-parallel two, three and four-stranded structures and a single, previously reported example of a parallel five-stranded coiled coil (Malashkevich *et al.*, 1996). However, it should be noted that some structures had more than one coiled-coil motif, and some included coiled-coil motifs of different order *N* and/or orientation. Thus, any one intact protein structure need not necessarily be exclusively classed as *N*-stranded. The numbers of coiled coils present in a set of non-homologous representatives from each is shown in Table 2. Full lists of the positive structures and details of their knob-into-holes packing motifs are available on the World Wide Web (http://www.biols.susx.ac.uk/coiledcoils). These pages are linked to sequence data giving all the homologous examples of the identified coiled-coil motifs.

The structural types mentioned above, together with the associated sequence data and some knobs-into-holes interactions observed in certain transmembrane domains provide the focus for this work. However, a number of additional structures tested positive for knobs-into-holes interactions and coiled-coil motifs. For instance, several variations on the four-helix-bundle motif with and without cyclically complementary knobs-into-holes interactions were highlighted. SOCKET was also able to locate non-canonical coiled-coil motifs, i.e. structures based on sequence patterns deviating from the hallmark heptad repeat (Brown *et al.*, 1996; Hicks *et al.*, 1997), for instance, those previously noted for hemagglutinin (1htm, Bullough *et al.*, 1994). Because we define the span of a coiled

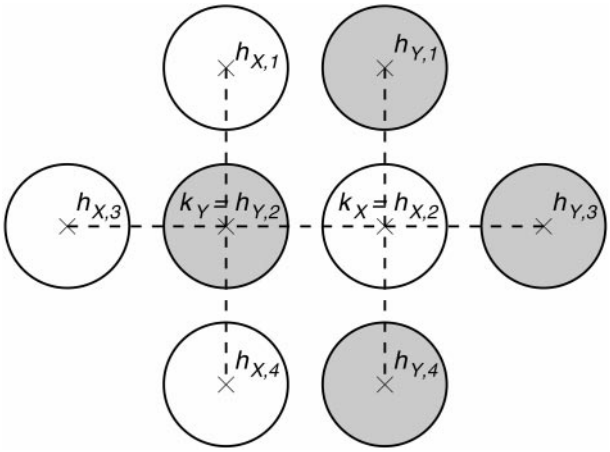


**Figure 5.** Schematic view of a complementary knob-into-hole interaction. Each circle represents a side-chain. The light circles are side-chains of helix X and face out of the page. The dark circles face into the page and are for helix Y. Crosses mark the centres of each side-chain, and the dotted lines the distances between them. The latter must be within the packing-cutoff to observe a knob-in-a-hole interaction. The four residues constituting each hole (*h*) are numbered in order as they appear in each helix. With the exception of holes in helices with distorted coiled coils caused by inserted residues, the four residues of a hole correspond respectively to residues $i$, $i+3$, $i+4$, $i+7$ with respect to amino acid sequence. This is whether the helices are parallel or anti-parallel, or whether the knob is at *a* or *d*. In the case shown, the two knob residues *k* are also equivalent to $h_2$ for X and Y, which indicates that this is a *d* layer. In an *a* layer in a parallel coiled coil, $k = h_3$. In an anti-parallel orientation, all layers are mixed *a*/*d*, with $k = h_2$ and $k = h_3$, respectively. In the core of a coiled coil with more than two helices, there would be no pairwise complementarity; that is, knob $k_Y$ would fit into hole $h_X$, but $k_X$ would not fit into hole $h_Y$ (or *vice versa*). Side-chains $h_{X,2}$ and $h_{Y,2}$ would however both be part of the same cyclic complementary *d* layer.

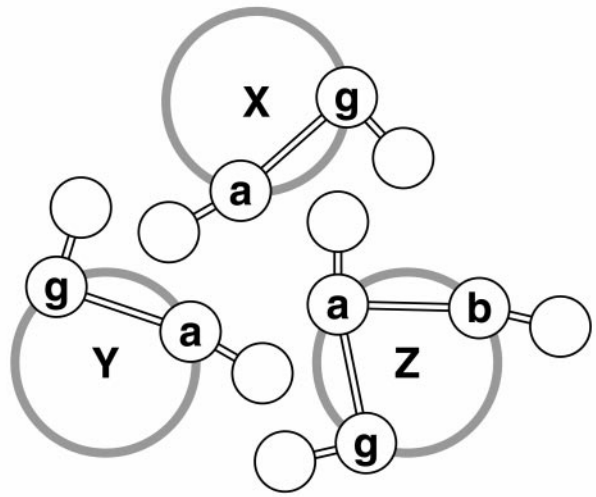


**Figure 6.** Partial helical wheel showing cyclically complementary knobs-into-holes in an *a*-layer of a three-stranded coiled coil. An *a* knob on helix X (X*a*) fits into a hole formed partly by *g* and *a* side-chains on helix Y. However, neither Y*g* nor Y*a* forms a knob to fit into a hole on helix X. Instead, Y*a* interacts with a hole on helix Z formed by Z*g* and Z*a*. Z*a* completes the cycle by fitting into a hole formed by X*g* and X*a*. This is cyclic complementarity of order 3. In addition, if the X*g* side-chain is long, it may act as a peripheral (non-core) knob and fit into a Z*a*/Z*b* hole. This would be an example of an X*g*-Z*a* pairwise complementarity (i.e. order = 2).

coil as the region between the extreme knobs-into-holes layers, we found some structures where the intervening helical segments did not display contiguous runs of knobs-into-holes interactions. For example, an anti-parallel region found in colicin Ia (1cii (Wiener *et al.*, 1997)) spanned 36 residues (236 to 271 and 397 to 432), therefore 11 complete layers were possible, but only five were highlighted by SOCKET.

In addition, more-unusual structures were found in which one or more helices participated in two different coiled-coil units. These included three-stranded ''α-layer'' structures from the variant surface glycoproteins of the trypanosome (Blum *et al.*, 1993; Freymann *et al.*, 1990) and colicin Ia from

**Table 2.** Numbers of types of structures that tested positive for coiled-coil motifs in the SOCKET analysis of the PDB

| | | | | Structures and (families) with *N*-stranded coiled-coils | | | |
|---|---|---|---|---|---|---|---|
| | Structures | Families | Coiled coils | $N = 2$ | $N = 3$ | $N = 4$ | $N = 5$ |
| Packing cutoff <=7.4 Å | 660 | 134 | 197 | 602 (114) | 53 (14) | 13 (7) | 1 (1) |
| Packing cutoff <=7.0 Å | 561 | 92 | 148 | 509 (74) | 47 (12) | 13 (7) | 1 (1) |
| Packing cutoff <=7.0 Å length > = 15 residues | 198 | 42 | 62 | 150 (27) | 46 (11) | 8 (4) | 1 (1) |

Numbers of types of structures that tested positive for coiled-coil motifs in the SOCKET analysis of the PDB. PDB Release #87, which had a total of 9255 structures, was examined. The number of sequence families to which the positive structures belong, and the numbers of coiled-coil motifs in the families (each represented by a single protein structure) are also indicated. The bottom row indicates the number of structures, families and coiled coils that remained after removing the short (<15 residues long) motifs. Some structures had more than one coiled coil, and some of these had different topologies (orders (N) and/or orientations). This means that the right-hand side of the table reveals more detail and the numbers here need not necessarily add up to those collated on the left-hand side.

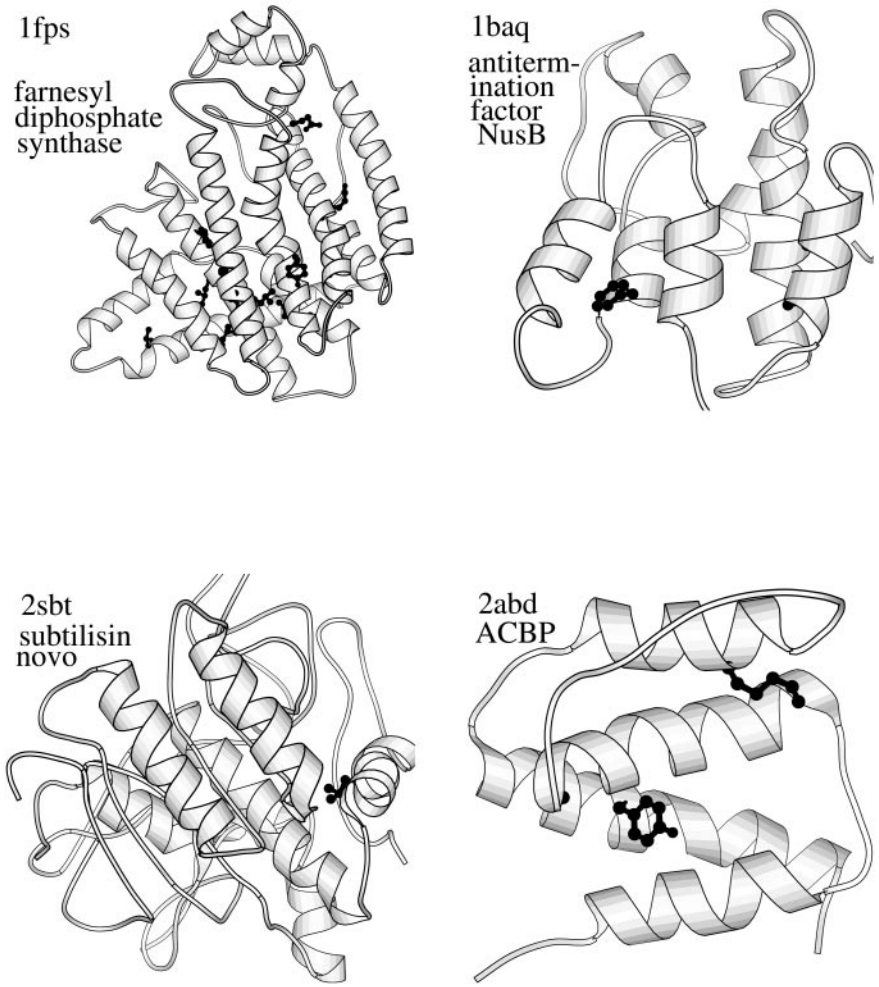


**Figure 7.** Isolated knobs-into-holes interactions in non-coiled-coil, globular, α-helical domains. Type 1 and 2 (non-complementary) knobs are shown as balls and sticks; significantly, no type 3 or 4 knobs were found in these examples.

*Escherichia coli* (Wiener *et al.*, 1997) and helix clusters, for instance in the core of the gp41 protein from HIV (Chan *et al.*, 1997). These structures are based on what we term multi-faceted helices, which effectively have two or more overlapping heptad repeats that facilitate association into multiple coiled coils. Full descriptions and theoretical analyses of these unusual structures and of the four-helix bundles will be presented elsewhere.

## Distinguishing different forms of two-stranded structures and the extent of knobs-into-holes interactions in non-fibrous, globular domains

A striking result from the analysis of the PDB was the large number (74) of sequence families that exhibited two-stranded coiled-coil motifs. After eliminating those that were shorter than 15 residues, only 28 families (nine with parallel and 19 with anti-parallel coiled coils) remained. The majority (46) of the two-stranded families, therefore, were for very short regions. Of these, ten had only parallel motifs, 32 had anti-parallel and four had both. Twenty-two of the anti-parallel arrangements in these families were the helix-turn-helix motifs and mostly occurred within globular domains. Examples included helices 5 and 6 of the endoglucanase catalytic core (1cem, Alzari *et al.*, 1996), helices 12 and 13 of *Streptomyces* N174 chitosananse (1chk, Marcotte *et al.*, 1996), helices 21 and 22 of methylmalonyl-coA mutase (3req, Mancia *et al.*, 1996), helices 1 and 2 of the tetratricopeptide repeats of the serine/threonine protein phosphatase domain (1a17, Das *et al.*, 1998) and helix pairs H2-H3 and J2-J3 of the beta-catenin armadillo repeat (3bct, Huber *et al.*, 1997). This class also showed the largest proportion of ''borderline'' motifs, which exhibit knobs-into-holes packing only when the packing-cutoff was increased to 7.4 Å. Therefore, these results appear to suggest that the short motifs do not share the signature of longer, ''true'' coiled coils, and possibly that they are inappropriately described by this term in its classic sense.

However, closer examination of the short two-stranded motifs revealed that, in many respects,

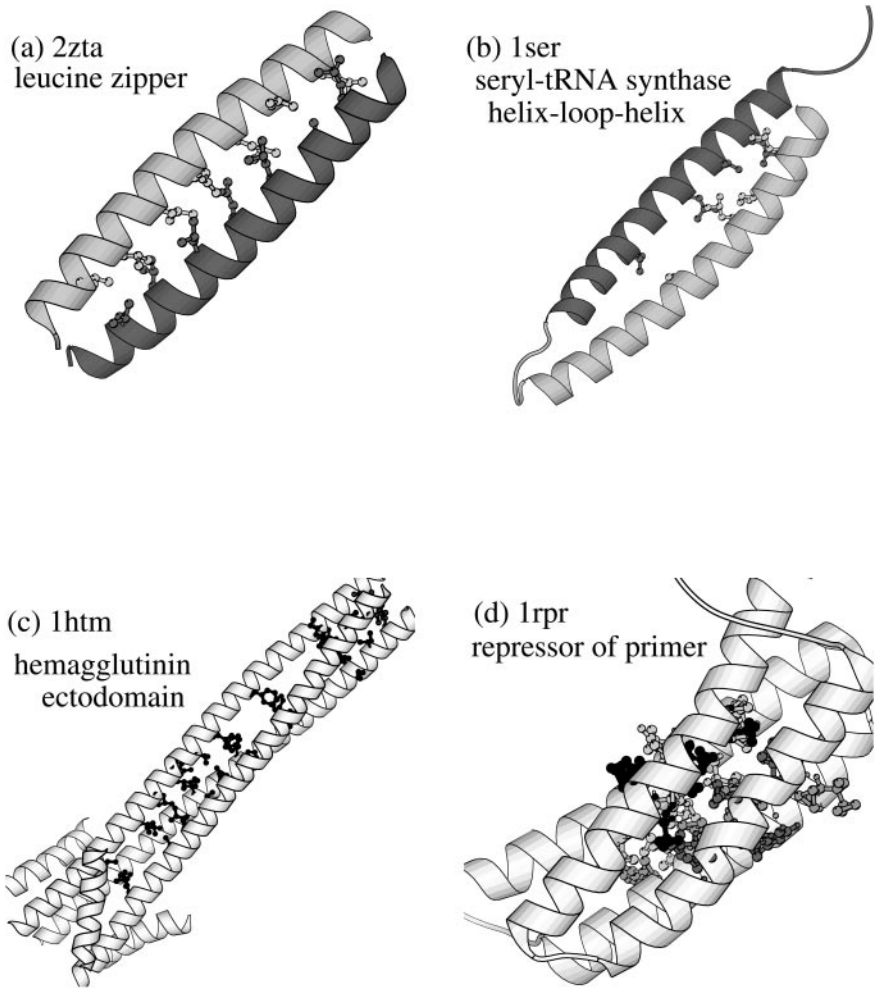


**Figure 8.** Examples of classic coiled coils displaying type 3 and 4 (complementary) knobs-into-holes. The knobs are highlighted as balls and sticks in (a) parallel two-stranded, (b) anti-parallel two-stranded, (c) parallel three-stranded and (d) anti-parallel four-stranded coiled-coil domains. In (c) and (d), *e* and *g* (non-core) knobs make complementary interactions with *gcdg* and *eabe* holes, respectively.

their packing mode was identical to the classic structures; it was simply exhibited over a shorter length. For instance, by taking one representative structure from each sequence family, we found that the interhelical angles and distances (Figure 9) and core-packing angles (Figure 10, see below) were indistinguishable from the longer anti-parallel two-stranded motifs. On the other hand, the packing appeared slightly ''looser'' in the short motifs, as indicated by the minimum packing cutoff required to identify knobs-into-holes interactions (Figure 11). All the ''long'' parallel motifs were found by a cutoff of no more than 6.8 Å, and the majority of these had still tighter packing, <6.4 Å (Figure 11(a)). However, the opposite was observed in the distribution for the short parallel motifs, which peaked at a cutoff of ≈6.5 Å (Figure 11(b)). Similar trends were clear in the data for the anti-parallel two-stranded motifs, although these distributions were slightly broader and their peaks moved to slightly higher cutoffs (Figure 11(c) and (d)). It was also apparent that all types of two-stranded coiled coil were typically looser-packed than three-stranded motifs, e.g. compare Figure 10(a)-(d) with 10(e) and (f).

## Trigger motifs

We envisage a variety of applications for the databases of positive coiled-coil structures that we have culled. For instance, one would be to test and develop theories for coiled-coil structure and assembly and establish sequence-to-structure rules for coiled-coil folding, stability, olgomer-state preference and partner selection.

An interesting example is the ''trigger motif'', which is a recently proposed sequence pattern stated to be obligatory for the folding of many if not all coiled-coil motifs (Steinmetz *et al.*, 1998). This notion is based on the following evidence. First, analysis of deletion mutants of the oligomerization domain (Ir) of *Dictyostelium discoideum* cortexillin I, a dimer, shows that monomers associate only when they include a motif corresponding to two

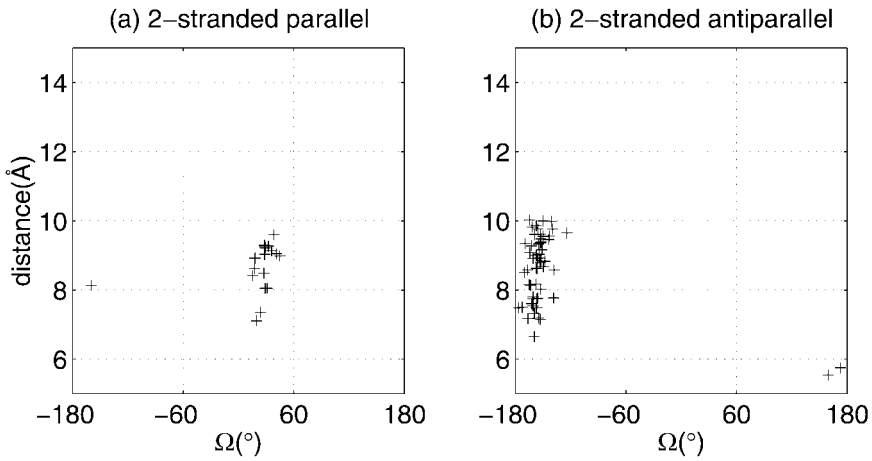


**Figure 9.** Helix-helix geometry in coiled-coil structures. The relative orientations of pairs of adjacent helices in (a) parallel and (b) anti-parallel two-stranded structures are gauged by inter-helical distance (*y*-axis) and inter-helical angle, Ω, (*x*-axis), which were calculated using XHELIX (Walther *et al.*, 1996). To avoid bias, data were taken from one representative structure from each family. Data from short and long examples were combined for this Figure.

particular consecutive heptad repeats from the wild-type sequence (Burkhard *et al.*, 2000). Second, a peptide corresponding to these 14 residues folds into a monomeric helix putatively stabilised by three intra-chain electrostatic interactions between side-chains spaced $i$ to $i+4$. It is proposed that this monomer mediates Ir dimerisation by forming local secondary structure prior to formation of the dimer, which is then stabilised by the pairing of other heptads from the sequence. Third, the same workers show that a similar motif is required for dimerisation of GCN4-leucine-zipper mutants (Kammerer *et al.*, 1998), and, by sequence comparisons, identify other 13-residue examples in a number of known dimeric coiled coils. Consensus analysis of these sequences result in two similar patterns, which may be represented by the PROSITE (Hofmann *et al.*, 1999) syntax, where residue positions are separated by hyphens and aligned against the heptad register:

```
d-e-f- g  -  a   -b-  c   -d-e

L-E-x-[RK]-[LVIA]-x-[REKD]-x-E

L-E-x- E  -[LVIA]-x-[REKD]-x-[RK]
```

.

We searched a recent issue of the PDB (#93) for such sequence patterns in the structures that displayed knobs-into-holes packing. In addition to the expected hits for the cortexillin I dimerisation domain and the GCN4-based mutants noted above, only one other family of coiled-coil sequences contained a trigger motif; namely, the tumour-necrosis receptor associated factor 2 structures (e.g. 1ca4, Park *et al.*, 1999). One of the family (PDB code 1czz, Ye *et al.*, 1999), however, did not have a trigger motif because of a point mutation. We also note that, although the 15 Å-resolution structure of rabbit α-chain tropomyosin (Phillips, 1986) contains only α-carbon atoms and was omitted from our initial analysis, none of its sequence matches either of the above proposed trigger patterns. Therefore, we find little corroborating evidence for the particular proposed trigger motifs being a general feature of true coiled-coil structures present in the PDB.

## Core-packing angles

Harbury *et al.* (1993) introduce core-packing angles to describe the orientation of a knob side-chain with respect to the hole into which it fits. They show that even disregarding the side-chain orientation, there is an inherent difference between the positions of knobs in different layers (*a* and *d*) in dimers and tetramers (Harbury *et al.*, 1993), and also in trimers (Harbury *et al.*, 1994). This is indicated by the angle between the $C^{\alpha}$-$C^{\beta}$ bond of the knob and the vector between the $C^{\alpha}$ atoms of the two residues constituting the sides of the hole (Methods and Definitions and Figure 12). These angles effectively describe the relative helix orientation, because the positions of $C^{\alpha}$ and $C^{\beta}$ atoms are basically invariant with respect to the α-helical axis. In *d* layers in parallel dimeric coiled coils, the $C^{\alpha}$-$C^{\beta}$ bond of the knob points straight into a *d,e* hole, and is described as *perpendicular*. In the *a* layers, the $C^{\alpha}$-$C^{\beta}$ bond is parallel to the vector between the *g* and *a* residue α-carbons. (Whether or not such a knob lies ''in'' or ''across'' the hole, i.e. whether it is a type 4 or 3 interaction, respectively, depends on the orientation of the side-chain beyond the β-carbon; which SOCKET determines using the insertion cutoff.) In parallel tetramers, the converse applies; that is, *a* knobs pack perpendicular and *d* knobs parallel. In trimers, both layers

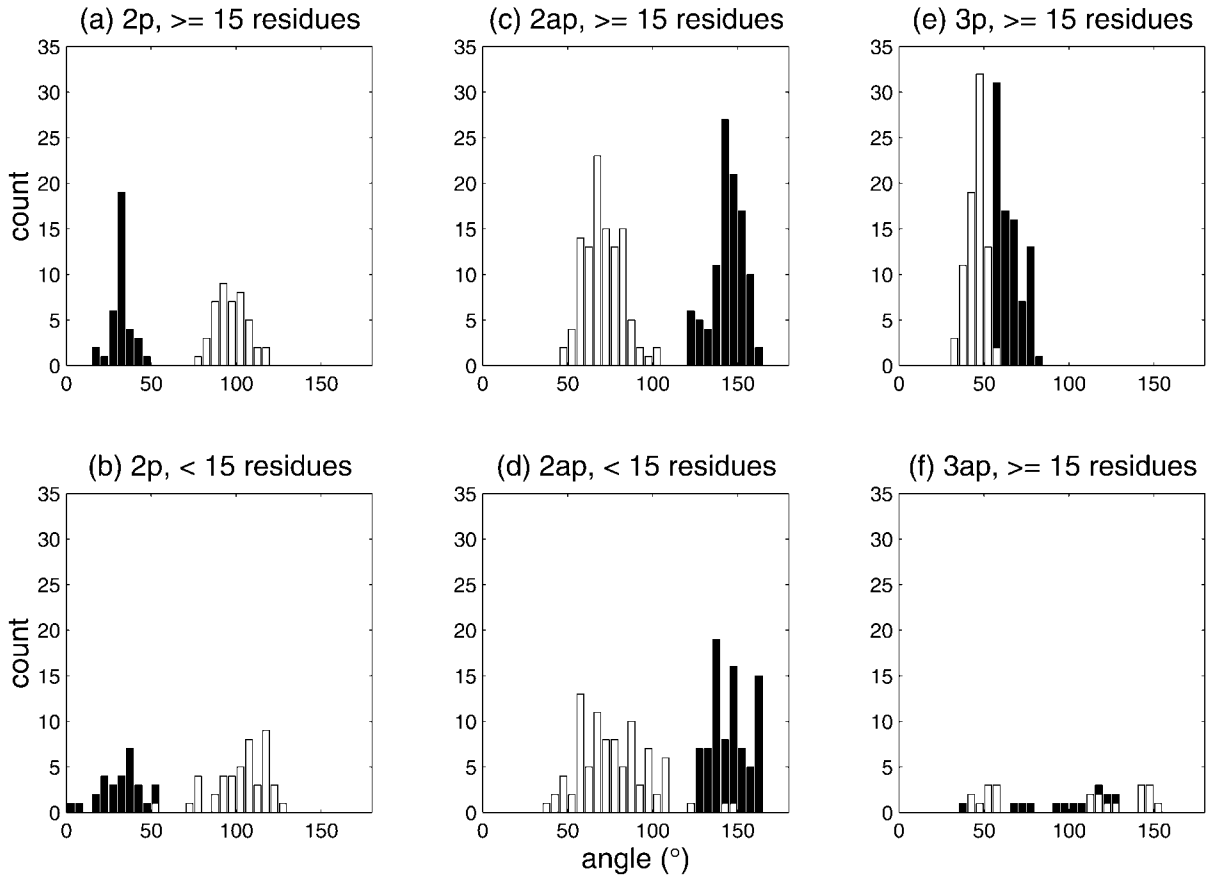


**Figure 10.** Core-packing angles. Angles were calculated for core (*a* and *d*) knobs packing into the base of their holes (*g*/*a* and *d*/*e*, respectively) as defined by Harbury and depicted in Figure 12 (Harbury *et al.*, 1993). Data for the *a* knobs are repesented by bold bars and those for the *d* knobs as hollow bars. The data were split according to topology and length of the retrieved coiled-coil motifs as follows: (a) and (b) long (⩾15 residues) and short two-stranded parallel motifs. (c) and (d) long and short two-stranded anti-parallel structures. (e) and (f) long (⩾15 residues) three-stranded parallel and anti-parallel coiled coils; Zp refers to two-stranded parallel and 2ap refers to two-stranded anti-parallel, etc.

exhibit angles between the two extremes that are described as acute.

An important concept developed by Harbury and colleagues is that the different packing geometries lead to differences in amino acid selection at the knob positions; alternatively, different amino acid choices at the *a* and *d* knobs dictate core-packing angles and, in turn, control oligomer-state during coiled-coil assembly. For example, the perpendicular orientation is thought to restrict the choice largely to side-chains that are not branched at the β-carbon. These principles have been used to considerable effect in protein-structure prediction (Berger *et al.*, 1995; Wolf *et al.*, 1997; Woolfson & Alber, 1995) and design (Harbury *et al.*, 1998; Nautiyal *et al.*, 1995; Pandya *et al.*, 2000). The sets of true coiled-coil structures that we have gathered provide an opportunity to test and further develop ideas on relationships between core-packing angle and amino acid profiles.

To illustrate the possibilities, we begin here by presenting and describing data on the core-packing angles in two-stranded parallel and anti-parallel coiled coils; we have also calculated the distributions of core-packing angles for all the other examples, which can be accessed at the aforementioned web site.

We considered the core-packing angles for long (⩾15 residues in length) and short parallel two-stranded motifs separately (Figure 10(a) and (b)). In the long structures the packing angles peaked at 30-35° for *a* layers and 90-95° in *d* layers (Figure 10(a)), which correspond to parallel and perpendicular packing orientations, respectively. The short structures showed a larger spread of core-packing angles (Figure 10(a) and (b)). Again, this suggests that the shorter structures are perhaps less ideal, or simply less constrained than the longer examples. Nonetheless, the similarity in the two distributions was clear, which provides further evidence that knobs-into-holes interactions made over short stretches could be classed as coiled-coil motifs even if they are not part of a classic fibrous structure.

In anti-parallel structures, the sense of the $C^{\alpha}$-$C^{\alpha}$ hole vector is reversed, but the core-packing angles are not simply a mirror of their parallel counterparts (Figure 9(c) and (d)). In long, anti-parallel

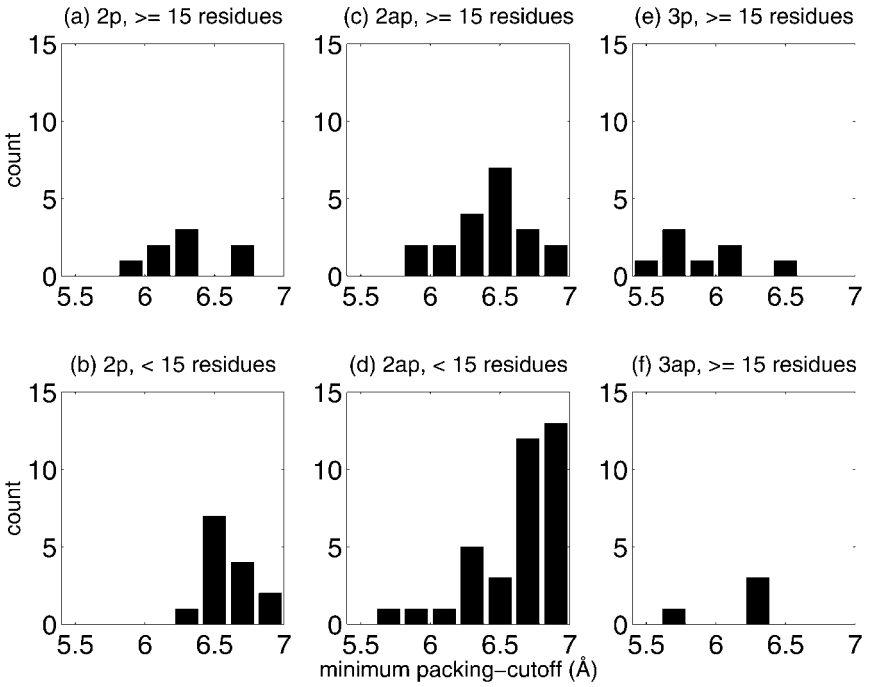


**Figure 11.** Minimum packing-cutoff required to observe complementary knobs-into-holes interactions in coiled-coil motifs. The plots show the number of retrieved structures for various packing-cutoffs below 7.0 Å. PDB entries were initially screened with one representative structure from each family that tested positive with a packing-cutoff of 7.0 Å taken forward for the analysis. (a) and (b) Long (⩾15 residues) and short two-stranded parallel structures. (c) and (d) Long and short two-stranded anti-parallel motifs. (e) and (f) long parallel and anti-parallel three-stranded coiled coils.

motifs, the angles between *a* knobs and *d*, *e* holes peaked at 140-145°; the orientation of the *d* knobs into *g*, *a* holes peaked at 65-70° (Figure 10(c)). Compared with the parallel structures, both ranges were shifted closer to the acute geometry described by Harbury (1993) for three-stranded coiled coils (Figure 10(e), see below). The broader distributions observed in the short, parallel motifs as compared with the longer counterparts were also apparent for the anti-parallel two-stranded structures.

Our analysis of higher-order structures also confirmed the Harbury (1993) theory: For three-stranded structures the distributions of core-packing angles for the parallel examples peaked at 45-50° and 55-60°, respectively (Figure 10(e) and (f)), which corresponded to acute packing (Harbury *et al.*, 1993, 1994). In a similar way, data (also not shown here) for a variety of natural four-stranded structures fitted the theory, which was developed from the crystal structure of the pLI mutant of the GCN4 leucine-zipper peptide (Harbury *et al.*, 1993). With some modifications, the structures of two and four-stranded parallel coiled coils are related, but the core-packing geometries at *a* and *d* layers are reversed. Accordingly, we found that for the *d* layers of parallel four-stranded arrangements the distribution of core-packing angles peaked at 30-35°, i.e. parallel, although the distribution was broader than for the related *a* sites of two-stranded structures. The distribution of angles made by the *a* knobs of four-stranded structures was also broader and had a slightly shifted peak (80-85°) relative to the d layers of two-stranded motifs, but, nonetheless, the distribution clearly indicated perpendicular packing.

## Amino-acid profiles for the heptad repeats of different coiled-coil topologies

The topological categories and the analysis of core-packing angles described above provide a basis for determining sequence-to-structure rules that may be used to discriminate different topologies for protein-structure prediction and design (Conway & Parry, 1990, 1991; Woolfson & Alber, 1995). The question is how, if at all, do the amino acid profiles for the various motifs differ (Woolfson & Alber, 1995)? For canonical coiled coils the term amino acid profile refers to 20 × 7 tables that give normalised rates of occurrence for each of the 20 amino acid residues at each of the seven heptad positions. Furthermore, do differences in the amino acid usage correlate with, and can they explain or be explained by, the different core-packing geometries in the different topologies as has been suggested elsewhere (Harbury *et al.*, 1993; Woolfson & Alber, 1995)?

For each coiled-coil topology, we compiled tables giving normalised frequencies of occurrence of amino acid residues at each heptad position (http://www.biols.susx.ac.uk/coiledcoils). To minimise bias, each family within a given topology

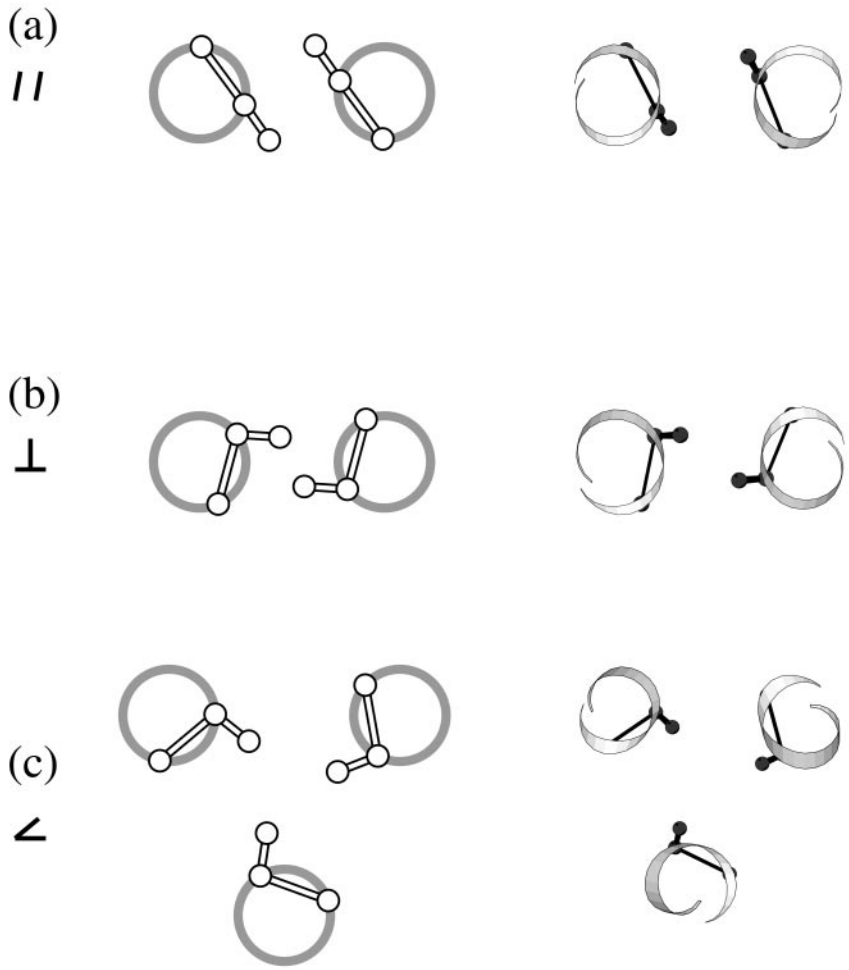


**Figure 12.** Schematic diagrams and experimental examples for different core-packing geometries. (a) Parallel packing, showing an *a*-layer from the GCN4 leucine zipper (O'Shea *et al*., 1991). (b) Perpendicular packing at a *d*-layer in the same structure. (c) Acute packing at a *d*-layer in a structure of a trimeric GCN4 mutant (Gonzalez *et al*., 1996).

contributed equally to the statistics, and multiple instances of the same protein were not counted more than once. We do not present statistical comparisons of the results for all of the topologies here because for some classes the sample sizes were too small. However, the tables for the two-stranded parallel and anti-parallel topologies (Tables 3 and 4) were reasonably populated even after excluding the short motifs with less than 15 residues. To assess potentially important differences that may help distinguish parallel and antiparallel dimers, we focused on the core-forming (*a*, *d*, *e* and *g*) sites. We asked which amino acid frequencies differed more than twofold between two data sets, namely, parallel *versus* anti-parallel two-stranded structures, and parallel *versus* a control set of frequencies derived from SWISSPROT (Bairoch & Apweiler, 2000).

First, it is noteworthy that the anti-parallel structures showed a broader spectrum of amino acid usage over all heptad positions; although there was approximately four times as much data for the anti-parallel structures (1089 residues, compared with 252 residues in long, parallel two-stranded structures). Regarding specific placements of particular residues, however, there were more examples that occurred with high rates in the parallel structures. For example, the following residue placements occurred in the parallel table at least twice as often as in either of the comparison data sets: (i) The β-branched, hydrophobic residue Val and the polar side-chain of Asn were favoured at *a* positions. (ii) Leu and Met were favoured at *d* sites. (iii) Half of the possible examples of charged residues were favoured at the core-flanking *e* and *g* sites, namely Glu at *e* and Glu, Lys and Arg at *g*. It is interesting that the occurrences of some hydrophobic residues showed the inverse correlation at these two sites, i.e. hydrophobic side-chains occurred more frequently at *e* and *g* in the anti-parallel structures, although not to the extent of a twofold increase.

As described above, for two-stranded structures the core-packing angles made by knobs at *a* and at *d* positions are different, parallel and perpendicular, respectively. Although this is true for both

**Table 3.** Amino acid profile for long parallel two-stranded coiled coils

| Amino acid | Normalised frequencies of occurrence at each heptad position | | | | | | |
|---|---|---|---|---|---|---|---|
| | *a* | *b* | *c* | *d* | *e* | *f* | *g* |
| A | 1.35 | 2.71 | 0.31 | 0.29 | 1.94 | 1.55 | 0.31 |
| C | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| D | 0 | 1.15 | 0.58 | 0 | 0 | 1.67 | 1.15 |
| E | 0.40 | 1.91 | 2.86 | 0 | 4.16 | 2.77 | 4.76 |
| F | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| G | 0 | 0 | 0 | 0 | 0 | 0 | 0.89 |
| H | 0 | 0 | 2.71 | 0 | 0 | 0 | 0 |
| I | 1.77 | 0.52 | 0 | 0 | 0.51 | 0 | 0 |
| K | 0.86 | 1.53 | 0.51 | 0.37 | 1.49 | 1.49 | 3.57 |
| L | 2.18 | 0.32 | 0.64 | 7.61 | 0.62 | 0.31 | 0 |
| M | 0 | 0 | 3.84 | 4.59 | 0 | 0 | 0 |
| N | 4.04 | 2.73 | 0.68 | 0 | 0 | 2.65 | 0 |
| P | 0 | 0 | 0.62 | 0.44 | 0 | 0 | 0 |
| Q | 0.65 | 0.76 | 4.58 | 0 | 2.96 | 0.74 | 2.29 |
| R | 0.41 | 1.17 | 1.17 | 0.42 | 2.85 | 1.71 | 2.94 |
| S | 0.36 | 1.70 | 0 | 0.30 | 0.41 | 2.06 | 0 |
| T | 0.45 | 0.53 | 2.14 | 0.38 | 1.04 | 1.04 | 1.07 |
| V | 3.12 | 0.46 | 1.38 | 0.33 | 0.89 | 0.89 | 0.46 |
| W | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Y | 0.80 | 1.81 | 0 | 0.68 | 0 | 0 | 0 |

Normalised frequencies of occurrence of amino acid residues at the different positions of the heptad pattern. These data were compiled for structured heptads in a selected set of non-redundant, long (⩾15 residues), parallel, two-stranded coiled coils; nine families with a total of 252 residues used.

parallel and anti-parallel two-stranded motifs, in the latter the angles are shifted slightly to the acute. To a first approximation, one might expect this to influence amino acid selection at these core positions; indeed this is the basis of the Harbury theory; this is not clear cut, however, because the nature of residues at the hole positions might also influence side-chain selection. Nonetheless, as described elsewhere for parallel structures amino acid selection does occur at the core sites in two-stranded motifs, but it is more even at the same sites in trimers where the core-packing angles are similar and acute (Harbury *et al.*, 1993, 1994; Woolfson & Alber, 1995). A pertinent question in the context of two-stranded structures therefore is: how does the attenuation of the core-packing angles in anti-parallel structures influence amino acid selection compared with that seen in the parallel motifs?

To address this question, we compared amino acid occupancies at the *a* and *d* positions for the two structures. Ideally, we would have calculated

**Table 4.** Amino acid profile for long anti-parallel two-stranded coiled coils

| Amino acid | Normalised frequencies of occurrence at each heptad position | | | | | | |
|---|---|---|---|---|---|---|---|
| | *a* | *b* | *c* | *d* | *e* | *f* | *g* |
| A | 1.42 | 1.48 | 1.28 | 1.55 | 0.69 | 1.04 | 1.98 |
| C | 0 | 0 | 0 | 0.59 | 0 | 1.19 | 0 |
| D | 0.09 | 1.06 | 1.71 | 0.56 | 1.37 | 1.75 | 0.25 |
| E | 0.54 | 1.32 | 1.75 | 1.31 | 1.45 | 1.76 | 1.23 |
| F | 0.84 | 0.51 | 0.85 | 0.96 | 0.80 | 0.96 | 0.64 |
| G | 0 | 0.72 | 0.81 | 0.07 | 0.48 | 0.38 | 0.19 |
| H | 1.09 | 1.56 | 0.31 | 0.44 | 0.88 | 1.47 | 3.79 |
| I | 2.87 | 0.48 | 0.24 | 1.60 | 1.59 | 0.79 | 1.12 |
| K | 0.33 | 2.24 | 1.52 | 0.41 | 1.22 | 1.55 | 0.66 |
| L | 2.96 | 1.04 | 1.33 | 3.64 | 1.31 | 0.35 | 1.11 |
| M | 1.45 | 1.18 | 0.59 | 1.65 | 0.83 | 0.83 | 1.93 |
| N | 0.88 | 1.58 | 0.94 | 0.22 | 0.44 | 1.78 | 0.59 |
| P | 0 | 0 | 0.28 | 0 | 0.27 | 0 | 0 |
| Q | 0.86 | 1.23 | 1.75 | 0.74 | 1.99 | 1.16 | 2.30 |
| R | 0.66 | 1.36 | 1.48 | 0.47 | 2.55 | 1.78 | 1.14 |
| S | 1.03 | 0.78 | 0.58 | 0.28 | 0.37 | 0.83 | 0.83 |
| T | 0.61 | 1.23 | 0.73 | 0.61 | 0.58 | 1.39 | 1.15 |
| V | 0.51 | 0.21 | 0.53 | 0.37 | 0.61 | 0.41 | 0.99 |
| W | 0 | 0.56 | 2.24 | 0.31 | 0.53 | 0.53 | 0.53 |
| Y | 1.23 | 0.66 | 0.44 | 1.69 | 0.82 | 0.41 | 0.20 |

Normalised frequencies of occurrence of amino acid residues at the different positions of the heptad pattern. These data were compiled for structured heptads in a selected set of non-redundant, long (⩾15 residues), anti-parallel, two-stranded coiled coils, 19 families with a total of 1089 residues used.

*a/d* ratios for every amino acid in each structure, as this carries the advantage of being self-normalising. However, for the representative (non-redundant) two-stranded structures not all of the amino acids were found at the *a* and *d* sites. Nonetheless from the ratios that could be evaluated, it was clear that the parallel structures showed greater discrimination of amino acid residues between the two sites. These were in line with the differences pointed out above: in parallel structures, a Leu residue was favoured at *d* by 3.5-times; whereas, the residue was more evenly spread between the *a* and *d* sites of anti-parallel motifs ($a/d = 0.84$). This is a particularly telling correlation because perpendicular packing is believed to favour strongly the non-β-branched hydrophobic residue (Harbury *et al*., 1993). By contrast, for parallel packing at *a* sites of parallel dimers β-branched hydrophobic residues are favoured (Harbury *et al*., 1993). In accordance, the *a/d* ratio for the Ile and Val residues combined was 14.2 for parallel structures, but reduced to 1.8 for the anti-parallels. Therefore, the use of certain hydrophobic residues clearly changes between the *a* and *d* sites of parallel and anti-parallel two-stranded structures and the differences are in line with expectations based on differences in core-packing angles. Interestingly, however, the proportion of all hydrophobic residues (Ala, Phe, Ile, Leu, Met, Val, Trp and Tyr) found at the $a + d$ sites was virtually the same (≈0.71) in the two structural types. Finally, the *a* sites of parallel dimers tolerate certain polar residues (Gonzalez *et al*., 1996; Woolfson & Alber, 1995). Asn side-chains appear to be particularly suited to this environment; indeed, this particular residue placement provides a key specifying interaction in the leucine zipper (Gonzalez *et al*., 1996; Lumb & Kim, 1995). It is interesting, therefore, that an Asn residue occurs four times more often at the *a* positions of parallel two-stranded structures, which include both leucine zipper and other structures, as compared with their anti-parallel counterparts.

### Knobs-into-holes packing between transmembrane helices

One analysis that uses nearest-neighbour measurements indicates that left-handed helix-pairs in the transmembrane regions of the photosynthetic reaction centre (PRC), cytochrome C oxidase and bacteriorhodopsin interact by knobs-into-holes packing characteristic of coiled coils (Langosch & Heringa, 1998). In accordance, the SOCKET analysis of the PDB gave positive results for the PRC and cytochrome C oxidase, as well as for the cytochrome bc1 transmembrane subunits. On the other hand, no bacteriorhodopsin structure showed more than one pairwise knobs-into-holes layer between any pair of helices, and was classed as negative. With a packing cutoff of 7.4 Å, however, anti-parallel two-stranded coiled coils with two consecutive layers were highlighted, and in the highest resolution crystal structure (Luecke *et al*., 1998) two pairs of helices interacted in this manner.

Returning to the photosynthetic reaction centre, there are 28 examples of the homologous L and M chains for the PRC listed in Version 1.6 of CATH (Orengo *et al*., 1997); these examples are from 12 structures, two of which have two copies of each chain. 18 of these chains (plus another not listed in CATH) tested positive for knobs-into-holes packing using SOCKET with a 7.0 Å packing cutoff. In the four structures from *Rhodopseudomonas viridis* (Deisenhofer *et al*., 1995; Lancaster & Michel, 1997), only one interaction was consistently reported, which was for an anti-parallel pair of helices in the M-chain. The interaction spanned three layers, but the middle layer was not a complementary knobs-into-holes interaction. In the remaining structures, which are from *Rhodobacter sphaeroides*, (Arnoux *et al*., 1989; Chang *et al*., 1991; Chirino *et al*., 1994; Ermler *et al*., 1994; McAuley-Hecht *et al*., 1998; Stowell *et al*., 1997; Yeates *et al*., 1988), shorter anti-parallel motifs (with two layers of knobs-into-holes) occurred between a different pair of helices. Depending on the structure, this was either in the L-chain, the M-chain, or both; in the structures with two copies of each chain these motifs were seen in both L-chains but neither M-chain. The apparent inconsistency of the packing interactions between different crystal structures and between different species suggests that the coiled-coil type interactions are marginal within this domain. This is consistent with the previous findings (Langosch & Heringa, 1998), which noted that the helix-helix interactions are less compact and regular than coiled coils from water-soluble proteins. It is possible that these findings reflect difficulties in achieving multiple coiled-coil interactions in multi-helix arrangements; we will explore this theme elsewhere.

## Conclusion

We introduce SOCKET, a program for identifying and analysing coiled-coil motifs in protein structures. Automated methods for analysing coiled coils have been sequence oriented and therefore predictive. By contrast, our method searches for the key structural features of coiled coils; namely, the knobs-into-holes interactions. SOCKET highlights complementary and cyclic arrangements for the knobs-into-holes, which is important in distinguishing isolated knobs-into-holes from networks of these that constitute *bona fide* coiled-coil structures. SOCKET uses this information to assign oligomer order (number of helices), orientation (parallel, anti-parallel and mixed) and heptad register for the identified coiled coils. The program also calculates ''core-packing angles'', which describe how each knob interacts with its corresponding hole.

By applying SOCKET to the PDB, we have shown that knobs-into-holes packing as observed in classic coiled coils is distinguishable from the majority of helix-helix packing in globular domains. Nonetheless, there was a low but steady frequency of the interaction between short, usually anti-parallel α-helices in globular contexts. These short motifs appeared only sparsely in isolated domains, and were not a characteristic of any particular α-helical globular fold. Whilst such helical pairs might not traditionally be called coiled coils, closer examination revealed that many of them had interhelical geometry and packing characteristic of the more-classical assemblies, although they were less symmetrical. In addition, the motifs tended to use similar amino acid residues at the interface positions (data not shown). Thus, the shorter structures appeared to be based on at least some of the structural principles of classic coiled coils. Therefore, these should not be excluded by the imposition of symmetry constraints for example. It is important that such examples should not be considered as ''false positives'' when predicted as coiled coils, for instance in algorithm-benchmarking exercises. For example, we found that some helix-bundle domains showed partial coiled-coil character and clearly had heptad repeats; accordingly, these sequences often gave strong positives in the commonly used coiled-coil prediction programs (data not shown).

In addition, to the non-classical short motifs and a variety of four-helix bundles that displayed partial (incomplete cycles) knobs-into-holes interactions, as expected our analysis of the PDB highlighted classic coiled coils. These included examples of two, three and four-stranded coiled coils with all possible orientations. SOCKET also retrieved the single known example of a five-stranded coiled coil, which is a parallel structure. In accord with previous reports (Langosch & Heringa, 1998), we also found examples of knobs-into-holes interactions in certain membrane-spanning structures. However, we conclude that many of these examples are borderline cases of coiled coils because SOCKET did not always report the same interactions for different structures of the same protein. We note also that in certain structures, particularly longer examples, contiguous layers of knobs-into-holes were not always evident. One example was found in a region of colicin Ia (1cii, Wiener *et al*., 1997) where 11 complete knobs-into-holes layers were possible, but only five were present. Structures of this type potentially pose problems for analysis and prediction. Such interruptions may have a biological role, for instance in modulating stability and dynamics of the assemblies.

It is intriguing that a number of more-complex structures, which nevertheless had regular arrangements of helices, also tested positive for knobs-into-holes interactions. Examples included clusters, sheets and cylinders of α-helices. We term these ''multi-faceted coiled coils'' because they are characterised by more than one heptad repeat superimposed on the same sequence. The offset between the heptad repeats determined the structural properties. For instance, we found that classic trimeric, tetrameric and pentameric coiled coils effectively exhibited two heptad repeats offset by 3, 1 and 1 residues, respectively; whereas, the cylinder and sheet structures were variations on multi-faceted coiled coils with a two-residue offset. These assemblies and a theoretical basis for them will be described elsewhere (Walshaw & Woolfson, 2001).

The relative tilt of helices in coiled coils, as indicated by the core-packing angles, was relatively uniform in each oligomer class. Moreover, the orientations of the wild-type and mutant peptides based on the GCN4 leucine zipper (Harbury *et al*., 1993, 1994) were characteristic of parallel two, three and four-stranded assemblies in general. The distributions of core-packing angles in the short two-stranded motifs indicated only slightly more variation than in the longer cases. Because the core-packing angles of anti-parallel helices are not simply a mirror image of their parallel counterparts, it would appear that a slightly different interhelical tilt is characteristic of the former. The angle distributions of three-stranded coiled coils also confirmed an almost identical orientation of *a* and *d* side-chains (with respect to each other) relative to the neighbouring helices with which they interact.

Finally, we derived amino acid profiles for all of the coiled-coil structures for which we identified multiple examples in the PDB. Here, we have compared the profiles for the long, parallel and anti-parallel two-stranded structures; the other tables are available at our Web site (http://www.biols.susx.ac.uk/coiledcoils). The comparison of the two-stranded motifs showed differences that could be rationalised to some considerable extent by differences in core-packing angles bewteen the two structural classes. We anticipate that similar analyses for the other coiled-coil topologies will help define sequence-to-structure rules for coiled-coil orientation, oligomerisation state and homo/hetero-specificity. In turn, such an understanding will aid the recognition of these motifs in protein sequences and improve protein designs.

## Methods and Definitions

### The SOCKET algorithm

SOCKET requires two data input files: a PDB-format (Berman *et al*., 2000) file containing three-dimensional atomic coordinates (including side-chains) and the corresponding DSSP-format file (Kabsch & Sander, 1983), which details the secondary structure of each residue.

### Knobs and holes

The basic packing interaction recognized by the program is a knob side-chain of one α-helix that fits into a

hole comprising four side-chains of a different, single α-helix.

All residues are represented by: (i) a centre (Figure 3(a)), which is the $C^{\alpha}$ atom for a glycine residue, and otherwise the mean co-ordinate of all the side-chain atoms (excluding hydrogens) from $C^{\beta}$ onwards; (ii) an end (Figure 3(a)), which is the terminal atom of the side-chain, or a mean co-ordinate where there are two termini. This representation of a side-chain is relatively insensitive to the position of individual atoms and carries advantages for measuring low-resolution structures. Because there are currently relatively few solved structures of coiled coils, poorer-resolution structures were included in the analysis described herein.

Contacts between side-chains of different helices are evaluated as the distance between their centres. All side-chain-side-chain contacts between all pairs of α-helices in the PDB structure are measured.

A side-chain is classed as a knob if it contacts four or more side-chains on another single helix within a specified packing cutoff; the nearest four side-chains constitute the corresponding hole (Figure 3(b), (c) and (d)). With a sensibly low packing cutoff (see below) the number of hole side-chains contacting a knob is rarely more than four. In an undistorted α-helix a hole will be composed of the four sequence-related residues $i$, $i+3$, $i+4$, $i+7$. However, this constraint is not imposed when compiling the components of each hole, because some α-helices exhibit local distortions due to an extra inserted residue, which does not grossly alter the direction of the helix. For example, the structure of GreA from *E. coli* (1grj (Stebbins *et al.*, 1995)) has a knob side-chain (Ile62) in a hole formed by residues $i$, $i+3$, $i+4$, $i+8$ (Leu18, Leu21, Lys22, Arg26) due to an extra residue deforming the helix backbone. Holes consisting of such unexpected patterns are reported by SOCKET.

In classic coiled-coil structures there are several examples of large side-chains from the hydrophobic core lying across a hole rather than inside it; for example, Lys176 of C-Fos in the C-Fos-C-Jun heterodimer (Glover & Harrison, 1995) lies across the hole formed by Leu296, Gln299, Asn300 and Leu303. Using a distance cutoff has the advantage that such instances are still observed (Figure 4). These two types of knob-hole interaction are differentiated by measuring the distance between the end of the knob side-chain and the four hole side-chains' centres. If this is within an insertion cutoff, then the knob is designated type-2, and is inside its hole (Figure 4(b)); otherwise it is a type-1 knob, lying across its hole (Figure 4(a)). The packing and insertion cutoffs were determined empirically as described in Results and Discussion.

## Complementary knobs

For all coiled coils, the true knobs-into-holes mode of packing advanced by Crick involves complementarity (Crick, 1953). In the two-stranded parallel case, a knob $k_X$ of helix X fits into a hole $h_Y$ of four side-chains $h_{Y,1}$, $h_{Y,2}$, $h_{Y,3}$, $h_{Y,4}$ in helix Y, which are usually the related residues $i$, $i+3$, $i+4$, $i+7$. One of these four hole side-chains is itself a knob ($k_Y$) which fits into a hole $h_X$ comprising four side-chains of helix X one of which is the aforementioned knob $k_X$ (Figure 5). The two knobs, $k_X$ and $k_Y$, which will be either both *a* or both *d* residues have pairwise-complementary and both have an order of 2 (see below). More precisely, they form equivalent side residues of the hole into which the other fits; i.e. $k_X \equiv h_{X,n}$ and $k_Y \equiv h_{Y,n}$, where $n$ is either 2 or 3. Following this, with the helix positioned vertically with the N terminus at the top (Figure 3(b)), hole residues $h_{Y,1}$ and $h_{Y,4}$ can be described as the top and bottom of the hole, respectively. The sides of each hole, $h_{Y,2}$, $h_{Y,3}$ and $h_{X,2}$, $h_{X,3}$ together form a layer, which can be an *a*-layer or a *d* layer; Figure 5 depicts a *d* layer. In two-stranded anti-parallel coiled coils, layers consist of one *a* and one *d*, with $n=2$ in one case and $n=3$ in the other.

This complementarity is expanded for higher-order coiled coils. Consider the X and Y helices above, this time in a three, four or five-stranded coiled coil: once more, knob $k_X$ of helix X fits into a hole, residues $h_Y$, one of which is itself a knob. However, $h_{Y,n}$ does *not* fit into a hole in X, but into a hole, $h_Z$, on the third helix Z. In a three-order coiled coil, $h_{Z,n}$ is a knob which fits into $h_X$, completing a cyclic arrangement of 3 knobs and 3 holes. Again, in a three-stranded parallel assembly, the knobs will all have the same sequence register, i.e. they are all *a* or all *d* residues, and the sides of the three holes form an *a*-layer or *d* layer. Larger rings of four or five knobs are present in higher-order four and five-stranded coiled coils, respectively. These arrangements in the *core* of coiled coils with more than two helices have cyclic-complementarity, i.e. like a daisy chain, and *not* pairwise-complementary; for example, see the three *a* side-chains of Figure 6. SOCKET finds such cycles by a recursive procedure that searches for closed loops amongst an initial list of knobs-into-holes interactions calculated for a structure. The order of a knob, or layer of knobs, is the number of knobs in the complementary arrangement, which is usually the same as the number of helices in the coiled coil (the latter is referred to as the order of the coiled coil).

One reason that this distinction must be observed is that coiled coils with more than two strands have additional, ''peripheral'' knobs and holes. For example, in the parallel tetrameric GCN4 mutant p-LI *e* side-chains fit into *gcdg* holes (Harbury *et al.*, 1993), and *g* side-chains fit into *eabe* holes. In this case, the result is that neighbouring pairs of helices in the four-helix ring pack *via* knobs-in-holes that have pairwise-complementarity. Nonetheless, core, *a* and *d* residues do set up a four-membered cyclic interaction indicating that each helix pair is part of a four-stranded, and not a simple two-stranded coiled coil. Pairwise interactions can also occur between helices in three-stranded coiled coils, e.g. *g* of helix X and *a* of Z in Figure 6.

Another problem is that it is possible for neighbouring helices that are perpendicular to have a single pair of pairwise-complementary knobs-in-holes; we found a number of short helices in globular domains that exhibited such interactions. Therefore, in SOCKET a requirement for at least two pairwise complementary knobs-into-holes is set to define a two-stranded coiled coil. This means that a two-stranded coiled coil can effectively be as short as a single heptad. On the other hand, the presence of even a single *N*-order cyclic-complementary knobs-into-holes interaction is considered enough to designate the *N* helices involved as belonging to an *N*-stranded coiled coil. By this definition a coiled coil can consist only of a single layer, i.e. half a heptad.

We term knob residues that exhibit complementarity as type-3 and type-4, which are extensions of type-1 and type-2, respectively. Note that any group (pair or cycle) of complementary knobs can have a mixture of the two types. For example, in the Fos-Jun structure (1fos (Glover & Harrison, 1995)), Lys176 in chain E is type-3, while its

partner residue Asn300 in chain F is type 4, Figure 4(c). A layer consisting of only type-3 knobs represents a relatively poor knobs-into-holes pattern. All pairwise or cyclic complementary layers will consist only of knobs of type 3 or greater.

It should be noted that it is possible for a side-chain to be a knob in more than one cyclic-complementary ring of knobs-in-holes, this occurs in some unusual arrangements of helices that we discuss elsewhere (J.W. & D.N.W., unpublished results). Therefore, it is important that the order, list of complementary knobs and register of each knob be compiled separately for each cyclic interaction.

### Determination of packing-cutoff and insertion-cutoff

To establish how well such domains can be distinguished from α-helical domains which lack coiled coils, a number of classic coiled-coil protein structures were assessed using SOCKET. These structures and the controls are listed below. In each case, an initial packing cutoff of 5.0 Å was used and successively incremented by 0.1 Å, until the result was positive for knobs-into-holes; any structure with at least two helices will be positive if the packing cutoff is sufficiently large. This effectively scored each coiled-coil and control structure and allowed the evaluation of a sensible standard cutoff value (minimum packing cutoff) that optimally distinguished coiled coils from other helical domains (Table 1). The same value was used as the insertion cutoff.

The following PDB structures were used as coiled-coil positive: parallel homodimers 2zta (O'Shea *et al.*, 1991) and 1a93 (Lavigne *et al.*, 1998); anti-parallel helix-turn-helix motifs 1ses (Belrhali *et al.*, 1994) and 1bmf (Abrahams *et al.*, 1994); the anti-parallel homodimer 1ecr (Kamada *et al.*, 1996); parallel homotrimers 1htm (Bullough *et al.*, 1994) and 1rtm (Weis & Drickamer, 1994); the parallel heterotypic four-helix coiled coil 1sfc (Sutton *et al.*, 1998); and the anti-parallel four-helix structure 1rpr (Eberle *et al.*, 1991). The control α-domain structures were 2hhb (Fermi *et al.*, 1984), 1fps (Tarshis *et al.*, 1994), 1hdd (Kissinger *et al.*, 1990), 1baq (Huenges *et al.*, 1998), 1ihb (Venkataramani *et al.*, 1998) and 2abd (Andersen & Poulsen, 1993). In addition, the α/β structure 1cse (Bode *et al.*, 1987) was used, which has helices that pack side-by-side in a Rossman fold.

### Processing of residues beyond the ends of α-helices

Using the DSSP definitions of secondary structure has the drawback that hole side-chains at the extremities of a coiled coil may be missed. For example, the most N-terminal knob residue on one α-helix might fit into a hole whose top residue is not α-helical. For this reason, an optional helix-extension parameter may be used in SOCKET to extend each DSSP-defined helix by a specified number of residues. This means that helices separated by short local distortions become merged. In some cases this is desirable: for example, the anti-parallel coiled coil of GreA has one helix effectively split into two by a ''kink'' caused by an extra residue inserted in the heptad pattern (Stebbins *et al.*, 1995). Because one half of the motif has only one pairwise complementary knob-in-hole interaction, it would not be considered as a coiled coil based on the above definitions. Joining the two halves, however, results a single coiled coil that is recognised by SOCKET. On the other hand, two anti-parallel helices joined by a four-residue hairpin become merged if a helix-extension of only two residues is used; the two helices are then incorrectly evaluated as constituting a single continuous helix. In the worst cases, erroneous knob-in-hole interactions will be found in which a hole might comprise three side-chains from one true helix and the fourth from another. Therefore, the maximum permissible helix extension is two, which should be used with care and the results inspected by eye. It is unfortunate that there are problems in identifying such instances automatically; large gaps in the serial numbers of the hole residues in the PDB file cannot be used as an indicator, because some PDB structures have discontinuities in the sequence numbering for the sake of compatibility with the numbering scheme of homologues.

### Assignment of heptad register, abcdefg

The register of complementary knob residues, including peripheral knobs, can be unambiguously assigned in complete layers of the same order as the coiled coil. Consider a knob $k_X$ of helix $X$ with a complementary partner $k_Y$ of helix $Y$. $k_Y$ is either the second or third of the four hole side-chains ($h_{Y,n}$) into which $k_X$ fits. In two-stranded coiled coils, whether $k_X$ is an *a* or *d* knob is determined simply by whether $n = 2$ or 3 and helix orientation. For example, for a parallel two-stranded coiled coil each *a* knob fits into a $d'g'a'd'$, the complementary knob of which is a′ and $n = 3$; whereas, d knobs fit into $a'd'e'a'$ holes and $n = 2$. The opposite is true for the anti-parallel case.

In coiled coils with more than two helices, the issue may be complicated by the presence of non-core knobs in holes at the periphery of the helix-helix interfaces. In assemblies with more than two strands, the register of $k_X$ can be assigned only if the order of $k_Y$ is the same as the order $N$ of the coiled coil. If it is less than this value, then $k_X$ and $k_Y$ are part of an incomplete layer, and no attempt is made to assign the register. For example, in the structure of the synaptic fusion (a.k.a. SNARE) complex (PDB: *1sfc*, Sutton *et al.*, 1998) the four-stranded coiled coils ''fray'' towards their ends, exhibiting only two-order complementary knobs-into-holes packing between pairs of adjacent helices. These could be considered as two-stranded coiled coils, but this would be inappropriate since such missing layers might occur in the middle of a four-stranded motif with otherwise complete cyclic layers. The presence of a single $N$-order layer of cyclic complementary knobs is taken to mean that the order of the coiled coil is $N$.

Assuming that the order of $k_Y$=$N$, then the register of $k_X$ is determined by its own order. If the order of $k_X = N$, then the same rules as for the two-stranded interactions are followed: for example, in Figure 6, where the N terminus of the helices are nearest the viewer, the *a* residues ($k_X$,$k_Y$,$k_Z$) are determined as such because their complementary partner with an order of 3 ($k_Y$,$k_Z$,$k_X$) are $h_{Y,3}$, $h_{Z,3}$, $h_{X,3}$ not $h_{Y,2}$ *etc.*, and similarly for the *d* residues. If, on the other hand, the order of $k_X$ is less than $N$ (specifically, $k_X = 2$), then $k_X$ is a peripheral knob; a register of *e* or *g* is once more determined by a combination of $n = 2$ or 3 and the relative orientation of the helices. The *g* residue ($k_X$) in Figure 6 has a complementary partner with an order of 2 ($k_Z$) which is $h_{Z,2}$. Note that helices $X$ and $Z$ form a pairwise complementary interaction between *a* and *g* side-chains. All of these rules are summarised in Table 5.

The register of non-knob residues is determined implicitly by their sequential position relative to the

**Table 5.** Assignment of heptad register

| $n$ | Orientation | $K = N$ | $K = 2, N > 2$ |
|---|---|---|---|
| 2 | P | *d* | *g* |
| | A | *a* | *e* |
| 3 | P | *a* | *e* |
| | A | *d* | *g* |

Assignment of heptad register. The heptad register of a given knob $k_X$ is determined by: (i) which side residue $n$ of its hole is its complementary knob $k_Y$ ($n \in \{1,2,3,4\}$, and is the set of four hole residues numbered in order of sequence; $k_Y$ will always be one of the side residues, namely 2 or 3 (Figure 5); (ii) the orientation of the knob helix $X$ relative to the hole helix $Y$ (P is parallel, A is anti-parallel); (iii) the order $K$ of the knob $k_X$ and (iv) the order $N$ of the coiled coil. For example, in two-stranded motifs ($N = 2$) complementary knobs (*a* and *d*) always make pairwise ($K = 2$) interactions. In four-stranded coiled coils ($N = 4$), *a* and *d* knobs form 4-membered cyclic arrangements ($K = 4$; column 3), while peripheral (*e* and *g*) knobs, where they occur, are involved in pairwise interactions ($K = 2$; column 4). Peripheral knobs are less common in three-stranded coiled coils (and more frequent in five-stranded structures), but the same rules apply. Only *a*, *d*, *e* and *g* knobs are explicitly assigned a register; the others are deduced from their sequential separation from these residues.

knob residues. Considering non-knob residues in turn as they appear in the sequence, registers are assigned in order following the previously determined knob. Discontinuities in the heptad repeat, such as in the influenza hemagglutinin ectodomain (Bullough *et al.*, 1994), will appear as an interruption between the first non-canonical knob and its preceding residue. The span of a knobs-into-holes packing region of an α-helix is defined as from the most N-terminal to the most C-terminal hole residues. Therefore, in some cases certain intervening residues may be assigned as *a* and *d* sites even if they were not identified as knobs by SOCKET. Residues N-terminal to the first explicitly determined knob are assigned a register that will be continuous with it.

The orientation of two helices can be determined either by a consideration of the relative sequence order of the most extreme complementary knobs in each helix, or by assessing the vectors describing the helical axes (only a pseudo-axis is calculated, based on the $C^\alpha$ positions of the terminal residues). The orientation of a coiled coil is parallel only if all of its helices are parallel.

### Core-packing angles

SOCKET measures the packing-angles between each knob side-chain and the holes into which it fits. The angle used here is as defined by Harbury (1993); i.e. the angle between two vectors: (i) the $C^\alpha$-$C^\beta$ bond vector of the knob residue and (ii) the $C^\alpha$-$C^\alpha$ vector between the two residues that form the sides of the hole. In SOCKET, the latter is always calculated from $C^\alpha$ of $h_{Y,2}$ to $C^\alpha$ of $h_{Y,3}$, which means that the angles for anti-parallel pairs equate to 180° minus the angles in parallel cases. The angles between *e* and *g* knobs and their holes are also measured. Since $C^\alpha$ and $C^\beta$ positions are basically invariant with respect to the helix axis, the packing angles are effectively an index of the relative tilt of neighbouring helices.

### Application of SOCKET to the PDB

In some crystal structures of coiled coils the assembly is not manifested in the asymmetric unit. ''Biological units'' of solved protein structures are made available by the European Bioinformatics Institute in the Macromolecular Structure Database (EBI-MSD (Henrick & Thornton, 1998)). For efficiency, MSD files were used only for the structure in the PDB_SELECT (Hobohm & Sander, 1994) list of non-homologous protein chains representative of all known folds. However, this list excludes short chains of <30 residues and, so, might omit some coiled coils. Therefore, where necessary, MSD files were used for any additional coiled coil structures found in the literature, particularly from reviews on this motif (Kohn *et al.*, 1997; Lupas, 1996). For all remaining structures, the plain PDB (Berman *et al.*, 2000) files were used. Release #87 of the PDB was processed, which contained a total of 9255 structures.

All the structures were analysed with SOCKET using the standard packing cutoff of 7.0 Å and a helix-extension of zero. Structures exhibiting no knobs-into-holes packing were reanalysed by incrementing the helix extension, up to a maximum of two. In order to highlight structures with only marginal knobs-into-holes packing, a similar analysis was performed using the more-liberal packing cutoff of 7.4 Å. For structures that tested positive for knobs-in-holes, we determined all the minimum packing-cutoffs and the associated minimum helix extensions required to find knobs-into-holes. In all cases an insertion-cutoff of 7.0 Å was used. In both analyses, structures that were positive only with the helix-extension parameter >0 residues were checked to ensure that no holes were split over two distinct helices; theoretical model structures were also removed.

Chains shown to pack *via* knobs-into-holes were grouped into families to reduce bias resulting from the duplication of proteins or their homologues in the PDB. The structure classification schemes SCOP (Murzin *et al.*, 1995) and CATH (Orengo *et al.*, 1997) could not always be used because not all the protein structures had been entered into these databases. Instead, the set of sequences of all the chains was self-compared using the BLAST pairwise-alignment program (Altschul *et al.*, 1997). A strict P-value of $<10^{-10}$ was taken to indicate that two sequences are homologues. There are dangers associated with searching a relatively small database, but a check indicated that the results for a given pair of sequences were similar whether searching versus the set of structures that tested positive for coiled-coil motifs, or versus the entire PDB. Inconsistent or dubious groupings were corrected manually using SCOP, CATH and the Pfam database of sequence families (Bateman *et al.*, 2000), though the taxonomy of these schemes differ in some cases; it is problematic to demonstrate homology for short sequences of relatively low complexity such as some coiled coils. The structure with the best resolution and/or longest knobs-into-holes packing motif was chosen as the representative for each family, and interhelical angles and distances were measured for these using XHELIX (Walther *et al.*, 1996).

### Compilation of amino acid profiles

Profiles, that is, normalised frequency tables giving the relative occurrence of each amino acid at each heptad position (Gribskov *et al*., 1987; Woolfson & Alber, 1995), were compiled from the heptads identified and assigned by SOCKET. This was done independently for each type of topology (e.g. parallel two-stranded, anti-parallel two-stranded, etc). Furthermore, separate tables were compiled for short (14 amino acid residues or fewer in length) and long two-stranded motifs of each orientation. The complete data set for any given topology comprised heptads from many families, which themselves contained widely different numbers of sequences some of which were identical or homologous. Therefore, we have not quoted summed frequencies across all the data for each topology, as these raw data would be very biased. Instead, amino acid frequencies were tallied separately for each family, and the profile of each topology was taken as the mean of the profiles of all families; in this way families with a large number of coiled-coil structures did not bias the data. To further reduce bias, where more than one structure had been solved, only one of each protein was used. Where structures possessed multiple motifs with different topologies, the statistics from different parts of the structure were summed separately such that they contributed to the appropriate topology profile. In addition, because coiled-coil motifs rarely comprise integral numbers of complete heptads, the numbers of residues at each position *a* to *g* are usually different. Therefore to allow comparison between positions, the profile values were scaled so that the total for each column (heptad position) was unity. Finally, the summed, averaged and scaled profiles of each topology were normalised by dividing by the relative frequency of each amino acid in SWISSPROT (Bairoch & Apweiler, 2000). To reiterate a point made above, not all residues assigned as *a* and *d* will necessarily have been identified by SOCKET as knob residues.

### Illustrations

All or parts of Figures 1, 3, 4, 7, 8 and 12 were created using MOLSCRIPT (Kraulis, 1991).

### World Wide Web resource

We have created a website (http://www.biols.susx.ac.uk/coiledcoils) from which the SOCKET program and a user manual may be downloaded. Alternatively, the program may be run interactively from this site. In addition, all of the coiled-coil regions highlighted from the PDB and analysed by SOCKET are tabulated and may be viewed through this site. Links are provided from each structure to related sequence data; namely, the corresponding amino acid profiles and the sequences of related proteins and protein families. The site will be updated periodically.

## Acknowledgements

We thank the MRC for financial support and the Sussex High Performance Computing Initiative for access to the Onyx2 parallel processor and O2 workstations.

## References

Abrahams, J. P., Leslie, A. G., Lutter, R. & Walker, J. E. (1994). Structure at 2.8 Å resolution of F1-ATPase from bovine heart mitochondria. *Nature,* **370**, 621-628.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389-3402.

Alzari, P. M., Souchon, H. & Dominguez, R. (1996). The crystal structure of endoglucanase CelA, a family 8 glycosyl hydrolase from Clostridium thermocellum. *Structure*, **4**, 265-275, Published erratum appears in *Structure* 1996**15**-633.

Andersen, K. V. & Poulsen, F. M. (1993). The three-dimensional structure of acyl-coenzyme A binding protein from bovine liver: structural refinement using heteronuclear multidimensional NMR spectroscopy. *J. Biomol. NMR,* **3**, 271-284.

Arnoux, B., Ducruix, A., Reiss-Husson, F., Lutz, M., Norris, J., Schiffer, M. & Chang, C. H. (1989). Structure of spheroidene in the photosynthetic reaction center from Y Rhodobacter sphaeroides. *FEBS Letters,* **258**, 47-50.

Bairoch, A. & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucl. Acids Res.* **28**, 45-48.

Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L. & Sonnhammer, E. L. (2000). The Pfam protein families database. *Nucl. Acids Res.* **28**, 263-266.

Belrhali, H., Yaremchuk, A., Tukalo, M., Larsen, K., Berthet-Colominas, C. & Leberman, R. *et al*. (1994). Crystal structures at 2.5 angstrom resolution of seryl-tRNA synthetase complexed with two analogs of seryl adenylate. *Science,* **263**, 1432-1436.

Berger, B., Wilson, D. B., Wolf, E., Tonchev, T., Milla, M. & Kim, P. S. (1995). Predicting coiled coils by use of pairwise residue correlations. *Proc. Natl Acad. Sci. USA,* **92**, 8259-8263.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N. & Weissig, H. *et al*. (2000). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235-242.

Blum, M. L., Down, J. A., Gurnett, A. M., Carrington, M., Turner, M. J. & Wiley, D. C. (1993). A structural motif in the variant surface glycoproteins of Trypanosoma-Brucei. *Nature,* **362**, 603-609.

Bode, W., Papamokos, E. & Musil, D. (1987). The high-resolution X-ray crystal structure of the complex formed between subtilisin Carlsberg and eglin c, an elastase inhibitor from the leech Hirudo medicinalis. Structural analysis, subtilisin structure and interface geometry. *Eur. J. Biochem.* **166**, 673-692.

Brown, J. H., Cohen, C. & Parry, D. A. D. (1996). Heptad breaks in alpha-helical coiled coils: stutters and stammers. *Proteins: Struct. Funct. Genet.* **26**, 134-145.

Bucher, P. & Bairoch, A. (1994). A generalized profile syntax for biomolecular sequence motifs and its function in automatic sequence interpretation. *ISMB-94,* **2**, 53-61.

Bullough, P. A., Hughson, F. M., Treharne, A. C., Ruigrok, R. W., Skehel, J. J. & Wiley, D. C. (1994). Crystals of a fragment of influenza haemagglutinin in the low pH induced conformation. *J. Mol. Biol.* **236**, 1262-1265.

Burkhard, P., Kammerer, R. A., Steinmetz, M. O., Bourenkov, G. P. & Aebi, U. (2000). The coiled-coil trigger site of the rod domain of cortexillin I unveils

a distinct network of interhelical and intrahelical salt bridges. *Structure Fold. Des.* **8**, 223-230.

Chan, D. C., Fass, D., Berger, J. M. & Kim, P. S. (1997). Core structure of gp41 from the HIV envelope glycoprotein. *Cell,* **89**, 263-273.

Chan, D. C., Chutkowski, C. T. & Kim, P. S. (1998). Evidence that a prominent cavity in the coiled coil of HIV type 1 gp41 is an attractive drug target. *Proc. Natl Acad. Sci. USA,* **95**, 15613-15617.

Chang, C. H., el-Kabbani, O., Tiede, D., Norris, J. & Schiffer, M. (1991). Structure of the membrane-bound protein photosynthetic reaction center from Rhodobacter sphaeroides. *Biochemistry,* **30**, 5352-5360.

Chirino, A. J., Lous, E. J., Huber, M., Allen, J. P., Schenck, C. C., Paddock, M. L., Feher, G. & Rees, D. C. (1994). Crystallographic analyses of site-directed mutants of the photosynthetic reaction center from Rhodobacter sphaeroides. *Biochemistry,* **33**, 4584-4593.

Chothia, C., Levitt, M. & Richardson, D. (1981). Helix to helix packing in proteins. *J. Mol. Biol.* **145**, 215-250.

Conway, J. F. & Parry, D. A. D. (1990). Structural features in the heptad substructures and longer range repeats of two-stranded alpha-fibrous proteins. *Int. J. Biol. Macromol.* **12**, 328-334.

Conway, J. F. & Parry, D. A. D. (1991). Three-stranded alpha-fibrous proteins: the heptad repeat and its implications for structure. *Int. J. Biol. Macromol.* **13**, 14-16.

Crick, F. H. C. (1953). The packing of alpha-helices: simple coiled-coils. *Acta. Crystallog.* **6**, 689-697.

Das, A. K., Cohen, P. W. & Barford, D. (1998). The structure of the tetratricopeptide repeats of protein phosphatase 5: implications for TPR-mediated protein-protein interactions. *EMBO J.* **17**, 1192-1199.

Deisenhofer, J., Epp, O., Sinning, I. & Michel, H. (1995). Crystallographic refinement at 2.3 Å resolution and refined model of the photosynthetic reaction centre from *Rhodopseudomonas viridis*. *J. Mol. Biol.* **246**, 429-457.

Eberle, W., Pastore, A., Sander, C. & Rosch, P. (1991). The structure of ColE1 rop in solution. *J. Biomol. NMR,* **1**, 71-82.

Efimov, A. V. (1999). Complementary packing of alpha-helices in proteins. *FEBS Letters,* **463**, 3-6.

Ermler, U., Fritzsch, G., Buchanan, S. K. & Michel, H. (1994). Structure of the photosynthetic reaction centre from Rhodobacter sphaeroides at 2.65 Å resolution: cofactors and protein-cofactor interactions. *Structure,* **2**, 925-936.

Fermi, G., Perutz, M. F., Shaanan, B. & Fourme, R. (1984). The crystal structure of human deoxyhaemoglobin at 1.74 Å resolution. *J. Mol. Biol.* **175**, 159-174.

Freymann, D., Down, J., Carrington, M., Roditi, I., Turner, M. & Wiley, D. (1990). 2.9 Å resolution structure of the N-terminal domain of a variant surface glycoprotein from *Trypanosoma-Brucei*. *J. Mol. Biol.* **216**, 141-160.

Glover, J. N. & Harrison, S. C. (1995). Crystal structure of the heterodimeric bZIP transcription factor c-Fos-c-Jun bound to DNA. *Nature,* **373**, 257-261.

Gonzalez, L. J., Woolfson, D. N. & Alber, T. (1996). Buried polar residues and structural specificity in the GCN4 leucine-zipper. *Nature Struct. Biol.* **3**, 1011-1018.

Gribskov, M., McLachlan, A. D. & Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. USA,* **84**, 4355-4358.

Harbury, P. B., Zhang, T., Kim, P. S. & Alber, T. (1993). A switch between two, three, and four-stranded coiled coils in GCN4 leucine zipper mutants. *Science,* **262**, 1401-1407.

Harbury, P. B., Kim, P. S. & Alber, T. (1994). Crystal structure of an isoleucine-zipper trimer. *Nature,* **371**, 80-83.

Harbury, P. B., Plecs, J. J., Tidor, B., Alber, T. & Kim, P. S. (1998). High-resolution protein design with backbone freedom. *Science,* **282**, 1462-1467.

Henrick, K. & Thornton, J. M. (1998). PQS: a protein quaternary structure file server. *Trends Biochem. Sci.* **23**, 358-361.

Hicks, M. R., Holberton, D. V., Kowalczyk, C. & Woolfson, D. N. (1997). Coiled-coil assembly by peptides with non-heptad sequence motifs. *Fold. Des.* **2**, 149-158.

Hobohm, U. & Sander, C. (1994). Enlarged representative set of protein structures. *Protein Sci.* **3**, 522-524.

Hofmann, K., Bucher, P., Falquet, L. & Bairoch, A. (1999). The PROSITE database, its status in 1999. *Nucl. Acids Res.* **27**, 215-219.

Huber, A. H., Nelson, W. J. & Weis, W. I. (1997). Three-dimensional structure of the armadillo repeat region of beta-catenin. *Cell,* **90**, 871-882.

Huenges, M., Rolz, C., Gschwind, R., Peteranderl, R., Berglechner, F. & Richter, G., *et al*. (1998). Solution structure of the antitermination protein NusB of *Escherichia coli*: a novel all-helical fold for an RNA-binding protein. *EMBO J.* **17**, 4092-4100.

Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers,* **22**, 2577-2637.

Kamada, K., Horiuchi, T., Ohsumi, K., Shimamoto, N. & Morikawa, K. (1996). Structure of a replication-terminator protein complexed with DNA. *Nature,* **383**, 598-603.

Kammerer, R. A., Schulthess, T., Landwehr, R., Lustig, A., Engel, J., Aebi, U. & Steinmetz, M. O. (1998). An autonomous folding unit mediates the assembly of two-stranded coiled coils. *Proc. Natl Acad. Sci. USA,* **95**, 13419-13424.

Kissinger, C. R., Liu, B. S., Martin-Blanco, E., Kornberg, T. B. & Pabo, C. O. (1990). Crystal structure of an engrailed homeodomain-DNA complex at 2.8 Å resolution: a framework for understanding homeodomain-DNA interactions. *Cell,* **63**, 579-590.

Kohn, W. D., Mant, C. T. & Hodges, R. S. (1997). alpha-helical protein assembly motifs. *J. Biol. Chem.* **272**, 2583-2586.

Kohn, W. D., Kay, C. M. & Hodges, R. S. (1998). Orientation, positional, additivity, and oligomerisation-state effects of interhelical ion pairs in alpha-helical coiled- coils. *J. Mol. Biol.* **283**, 993-1012.

Kraulis, P. J. (1991). Molscript: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallog.* **24**, 946-950.

Lancaster, C. R. & Michel, H. (1997). The coupling of light-induced electron transfer and proton uptake as derived from crystal structures of reaction centres from Rhodopseudomonas viridis modified at the binding site of the secondary quinone, QB. *Structure,* **5**, 1339-1359.

Langosch, D. & Heringa, J. (1998). Interaction of transmembrane helices by a knobs-into-holes packing

characteristic of soluble coiled coils. *Proteins: Struct. Funct. Genet.* **31**, 150-159.

Lavigne, P., Crump, M. P., Gagne, S. M., Hodges, R. S., Kay, C. M. & Sykes, B. D. (1998). Insights into the mechanism of heterodimerization from the 1H-NMR solution structure of the c-Myc-Max heterodimeric leucine zipper. *J. Mol. Biol.* **281**, 165-181.

Lovejoy, B., Choe, S., Cascio, D., McRorie, D. K., DeGrado, W. F. & Eisenberg, D. (1993). Crystal structure of a synthetic triple-stranded alpha-helical bundle. *Science,* **259**, 1288-1293.

Luecke, H., Richter, H. T. & Lanyi, J. K. (1998). Proton transfer pathways in bacteriorhodopsin at 2.3 angstrom resolution. *Science,* **280**, 1934-1937.

Lumb, K. J. & Kim, P. S. (1995). A buried polar interaction imparts structural uniqueness in a designed heterodimeric coiled-coil. *Biochemistry,* **34**, 8642-8648.

Lupas, A. (1996). Coiled coils: new structures and new functions. *Trends Biochem. Sci.* **21**, 375-382.

Lupas, A., Van Dyke, M. & Stock, J. (1991). Predicting coiled coils from protein sequences. *Science,* **252**, 1162-1164.

Malashkevich, V. N., Kammerer, R. A., Efimov, V. P., Schulthess, T. & Engel, J. (1996). The crystal structure of a five-stranded coiled coil in COMP: A prototype ion channel? *Science,* **274**, 761-765.

Mancia, F., Keep, N. H., Nakagawa, A., Leadlay, P. F., McSweeney, S. & Rasmussen, B. *et al.* (1996). How coenzyme B12 radicals are generated: the crystal structure of methylmalonyl-coenzyme A mutase at 2 Å resolution. *Structure,* **4**, 339-350.

Marcotte, E. M., Monzingo, A. F., Ernst, S. R., Brzezinski, R. & Robertus, J. D. (1996). X-ray structure of an anti-fungal chitosanase from streptomyces N174. *Nature Struct. Biol.* **3**, 155-162.

McAuley-Hecht, K. E., Fyfe, P. K., Ridge, J. P., Prince, S. M., Hunter, C. N. & Isaacs, N. W., *et al.* (1998). Structural studies of wild-type and mutant reaction centers from an antenna-deficient strain of *Rhodobacter sphaeroides*: monitoring the optical properties of the complex from bacterial cell to crystal. *Biochemistry,* **37**, 4740-4750.

Mewes, H. W., Frishman, D., Gruber, C., Geier, B., Haase, D. & Kaps, A. *et al.* (2000). MIPS: a database for genomes and protein sequences. *Nucl. Acids Res.* **28**, 37-40.

Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536-540.

Nautiyal, S. & Alber, T. (1999). Crystal structure of a designed, thermostable, heterotrimeric coiled coil. *Protein Sci.* **8**, 84-90.

Nautiyal, S., Woolfson, D. N., King, D. S. & Alber, T. (1995). A designed heterotrimeric coiled-coil. *Biochemistry,* **34**, 11645-11651.

O'Shea, E. K., Klemm, J. D., Kim, P. S. & Alber, T. (1991). X-ray structure of the GCN4 leucine zipper, a two-stranded, parallel coiled coil. *Science,* **254**, 539-544.

O'Shea, E. K., Rutkowski, R. & Kim, P. S. (1992). Mechanism of specificity in the Fos-Jun oncoprotein heterodimer. *Cell,* **68**, 699-708.

O'Shea, E. K., Lumb, K. J. & Kim, P. S. (1993). Peptide velcro: design of a heterodimeric coiled-coil. *Curr. Biol.* **3**, 658-667.

Ogihara, N. L., Weiss, M. S., Degrado, W. F. & Eisenberg, D. (1997). The crystal structure of the designed trimeric coiled coil coil-VaLd: implications for engineering crystals and supramolecular assemblies. *Protein Sci.* **6**, 80-88.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). CATH: a hierarchic classification of protein domain structures. *Structure,* **5**, 1093-1108.

Pandya, M. J., Spooner, G. M., Sunde, M., Thorpe, J. R., Rodger, A. & Woolfson, D. N. (2000). Sticky-end assembly of a designed peptide fibre provides insight into protein fibrillogenesis. *Biochemistry,* **39**, 8728-8734.

Park, Y. C., Burkitt, V., Villa, A. R., Tong, L. & Wu, H. (1999). Structural basis for self-association and receptor recognition of human TRAF2. *Nature,* **398**, 533-538.

Phillips, G. N., Jr. (1986). Construction of an atomic model for tropomyosin and implications for interactions with actin. *J. Mol. Biol.* **192**, 128-131.

Sharma, V. A., Logan, J., King, D. S., White, R. & Alber, T. (1998). Sequence-based design of a peptide probe for the APC tumor suppressor protein. *Curr. Biol.* **8**, 823-830.

Stebbins, C. E., Borukhov, S., Orlova, M., Polyakov, A., Goldfarb, A. & Darst, S. A. (1995). Crystal structure of the GreA transcript cleavage factor from *Escherichia coli*. *Nature,* **373**, 636-640.

Steinert, P. M. (1993). Structure, function, and dynamics of keratin intermediate filaments. *J. Invest Dermatol.* **100**, 729-734.

Steinmetz, M. O., Stock, A., Schulthess, T., Landwehr, R., Lustig, A. & Faix, J. *et al.* (1998). A distinct 14 residue site triggers coiled-coil formation in cortexillin I. *EMBO J.* **17**, 1883-1891.

Stowell, M. H., McPhillips, T. M., Rees, D. C., Soltis, S. M., Abresch, E. & Feher, G. (1997). Light-induced structural changes in photosynthetic reaction center: implications for mechanism of electron-proton transfer. *Science,* **276**, 812-816.

Sutton, R. B., Fasshauer, D., Jahn, R. & Brünger, A. T. (1998). Crystal structure of a SNARE complex involved in synaptic exocytosis at 2.4 Å resolution. *Nature,* **395**, 347-353.

Tarshis, L. C., Yan, M., Poulter, C. D. & Sacchettini, J. C. (1994). Crystal structure of recombinant farnesyl diphosphate synthase at 2.6-Å resolution. *Biochemistry,* **33**, 10871-10877.

Venkataramani, R., Swaminathan, K. & Marmorstein, R. (1998). Crystal structure of the CDK4/6 inhibitory protein p18INK4c provides insights into ankyrin-like repeat structure/function and tumor-derived p16INK4 mutations. *Nature Struct. Biol.* **5**, 74-81.

Vinson, C. R., Hai, T. & Boyd, S. M. (1993). Dimerization specificity of the leucine zipper-containing bZIP motif on DNA binding: prediction and rational design. *Genes Dev.* **7**, 1047-1058.

Walshaw, J. & Woolfson, D. N. (2001). Open-and-shut cases in coiled-coil assembly: α-sheets and α-cylinders. *Protein Sci.* **10**, 668-673.

Walther, D., Eisenhaber, F. & Argos, P. (1996). Principles of helix-helix packing in proteins: the helical lattice superposition model. *J. Mol. Biol.* **255**, 536-553.

Weis, W. I. & Drickamer, K. (1994). Trimeric structure of a C-type mannose-binding protein. *Structure,* **2**, 1227-1240.

Wiener, M., Freymann, D., Ghosh, P. & Stroud, R. M. (1997). Crystal structure of colicin Ia. *Nature,* **385**, 461-464.

Wigge, P. A., Jensen, O. N., Holmes, S., Soues, S., Mann, M. & Kilmartin, J. V. (1998). Analysis of the Saccharomyces spindle pole by matrix-assisted laser desorption/ionization (MALDI) mass spectrometry. *J. Cell Biol.* **141**, 967-977.

Wolf, E., Kim, P. S. & Berger, B. (1997). MultiCoil: a program for predicting two- and three-stranded coiled coils. *Protein Sci.* **6**, 1179-1189.

Woolfson, D. N. & Alber, T. (1995). Predicting oligomerization states of coiled coils. *Protein Sci.* **4**, 1596-1607.

Ye, H., Park, Y. C., Kreishman, M., Kieff, E. & Wu, H. (1999). The structural basis for the recognition of diverse receptor sequences by TRAF2. *Mol. Cell,* **4**, 321-330.

Yeates, T. O., Komiya, H., Chirino, A., Rees, D. C., Allen, J. P. & Feher, G. (1988). Structure of the reaction center from Rhodobacter sphaeroides R-26 and 2.4.1: protein-cofactor (bacteriochlorophyll, bacteriopheophytin, and carotenoid) interactions. *Proc. Natl Acad. Sci. USA,* **85**, 7993-7997.

***Edited by J. Thornton***

*(Received 23 October 2000; received in revised form 4 February 2001; accepted 5 February 2001)*