Journal of Structural Biology **134,** 117–131 (2001)
doi:10.1006/jsbi.2001.4392, available online at http://www.idealibrary.com on IDEAL®

Protein Repeats: Structures, Functions, and Evolution

Miguel A. Andrade,*,† Carolina Perez-Iratxeta,*,† and Chris P. Ponting‡

*European Molecular Biology Laboratory, Meyerhofstr. 1, Heidelberg 69012, Germany; †Department of Bioinformatics, Max Delbrück Center for Molecular Medicine, Berlin-Buch 13092, Germany; and ‡MRC Functional Genetics Unit, Department of Human Anatomy and Genetics, University of Oxford, South Parks Road, Oxford OX1 3QX, United Kingdom

Received March 13, 2001, and in revised form June 11, 2001; published online August 2, 2001

Internal repetition within proteins has been a successful strategem on multiple separate occasions throughout evolution. Such protein repeats possess regular secondary structures and form multirepeat assemblies in three dimensions of diverse sizes and functions. In general, however, internal repetition affords a protein enhanced evolutionary prospects due to an enlargement of its available binding surface area. Constraints on sequence conservation appear to be relatively lax, due to binding functions ensuing from multiple, rather than, single repeats. Considerable sequence divergence as well as the short lengths of sequence repeats mean that repeat detection can be a particularly arduous task. We also consider the conundrum of how multiple repeats, which show strong structural and functional interdependencies, ever evolved from a single repeat ancestor. In this review, we illustrate each of these points by referring to six prolific repeat types (repeats in β-propellers and β-trefoils and tetratricopeptide, ankyrin, armadillo/HEAT, and leucine-rich repeats) and in other less-prolific but nonetheless interesting repeats. © 2001 Academic Press

INTRODUCTION

Past innovation in protein functions and structures is due, for the most part, to gene duplication (Ohno, 1970). Duplication and recombination within a single gene have often given rise to non-overlapping regions of a protein sequence that share significant sequence similarity. Such repeats are relatively common, occurring in at least 14% of all proteins (Marcotte *et al.*, 1999). Repeats vary considerably from short amino acid repetitions, for example, the polyglutamine tracts of the Huntington disease gene product huntingtin, to large repetitions containing multiple domains, such as in the cytoskeletal protein titin.

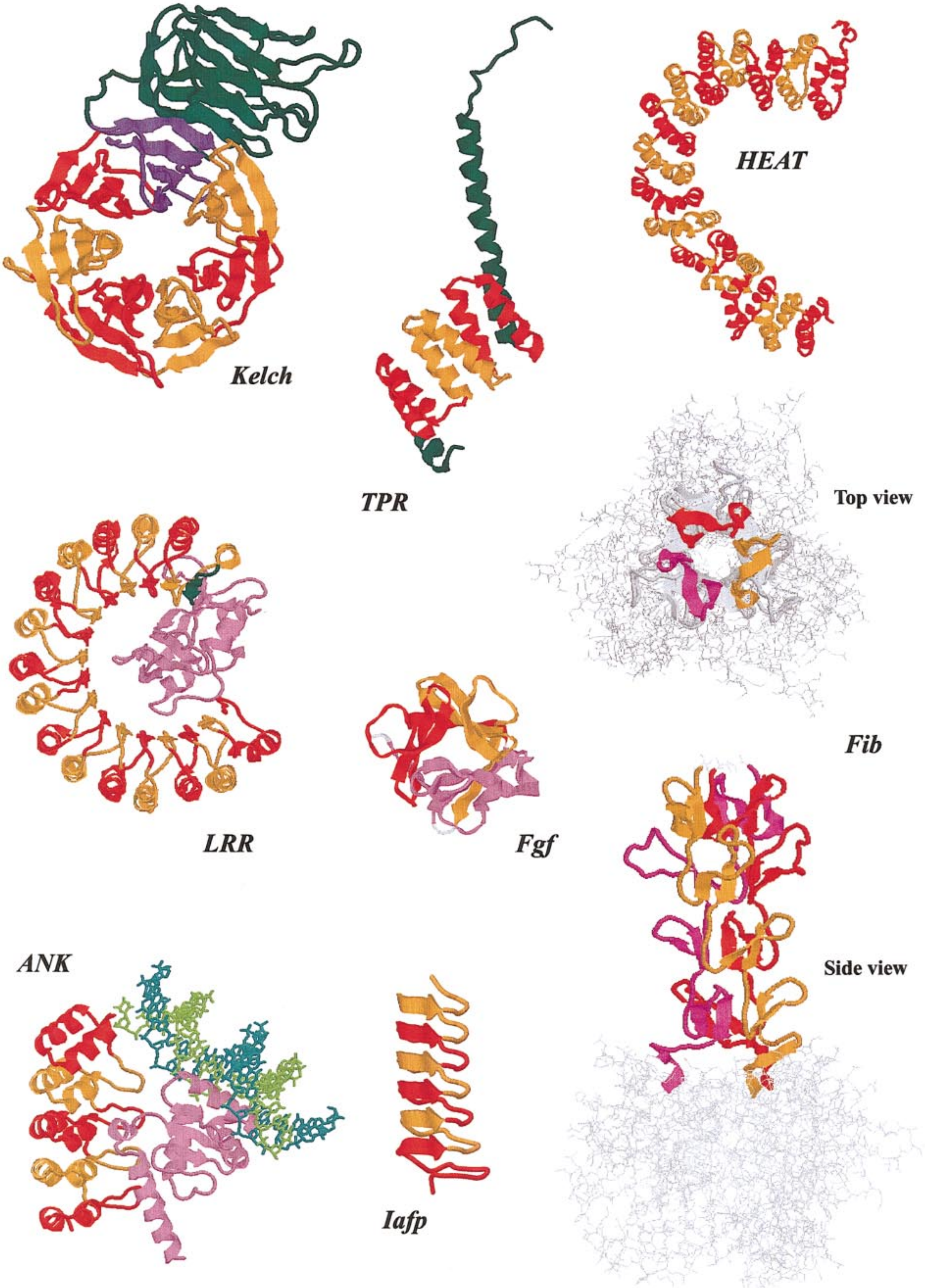
In this review, we concentrate on sequence repeats that occur tandemly in sequence and that form integrated assemblies when viewed as three-dimensional structures. Such repeats are essentially defined by their multiplicity and thus differ from both domains and motifs since these can occur singly. The importance of repeats in understanding biological function resides not only in their high frequency among known sequences, but also in their abilities to confer multiple binding and structural roles on proteins. This functional versatility is apparent not only among different repeat types, but also for similar repeats from the same family.

Our understanding of repeats, with respect to their structures, functions, and evolution, therefore represents a considerable challenge. How are we able to predict repeats within protein sequences? What are the relationships between repeats and their functions? In this review we describe six major repeat classes and their functions, structures, and possible evolutionary mechanisms. We attempt to describe how repeat identification can be linked to enhanced biological knowledge.

EVOLUTION OF REPEATS

Repeats are thought to arise via intragenic duplication and recombination events. Selective advantage of multiple repeats results in these mutations being fixed among populations. With the benefit of hindsight and the large numbers of sequences known, it is clear that repetitions of small structural units might confer several advantages on proteins, and thereon to their organisms, that are distinct from those of repeated domains. For example, tandemly repeated structures often occur in regular arrangements, either in linear arrays (e.g., see *Iafp* in Fig. 1) or as a superhelix with repeats arranged about a common axis (e.g., see *HEAT in* Fig. 1). For such "open" structures there is no theoretical limit on their repeat number, since incremental addition of repeats is not sterically impeded. These rod-like or superhelical structures present an extensive sol-

1047-8477/01 $35.00


Kelch
TPR
HEAT
Top view
Fib
LRR
Fgf
Side view
ANK
Iafp

vent-accessible surface that is well suited to binding large substrates such as proteins and nucleic acids.

By contrast, duplication of repeats in a superhelix with a small pitch results in a closed barrel-like structure, with a relatively small surface area available for ligand interactions with smaller ligands (e.g., see *Kelch* in Fig. 1). These assemblies are likely to present different advantages than the open structures of rods and superhelices. They are compact and stable, with opportunities for small ligands to be bound either along the internal axis of the barrel or on the axis at the barrel's periphery.

Following fixation of a repeat duplication, sequence similarities among repeats may erode quickly. Thus equivalent HEAT repeats in invertebrate and mammalian orthologues average only 13% sequence identity (Andrade *et al.*, 2001). These slight similarities imply that the functional constraints on individual repeats are relatively weak, when compared to the constraints imposed on the repeat assembly as a whole. By contrast, a function that is exacting on the structure of repeats, such as those in the ice-binding β-sheet domain of insect antifreeze proteins (Liou *et al.*, 2000), results in repeats being highly similar in sequence.

The numbers of repeats can vary even between orthologues, indicating that rapid loss and/or gain of repeats occurs frequently in evolution. This is neatly underscored by the demonstration that different alleles of a protein from the fungus *Podospora anserina* possess different numbers of WD40 repeats (Saupe *et al.*, 1995).

As we discuss below when describing major repeat classes, the most common function of repeat ensembles is that of binding to proteins. Such a property provides opportunities for the organism to expand its repertoire of cellular functions, such as protein transport, protein-complex assembly, and protein regulation using preexisting genetic material. Accordingly, even though the ability to generate repeats appears to be a general phenomenon of all phyla, repeats are more common in eukaryotic organisms than in prokaryotic ones (Marcotte *et al.*, 1999) and in metazoans more than in the rest of the eukaryotes (see Table I). This may be associated with the increasing complexity of cellular functions that are readily available from assemblies of repeats.

TABLE I

The Numbers and Percentages of Proteins That Are Annotated by the SwissProt Database (Bairoch and Apweiler, 2000) with the Feature "Repeat," Sorted by Taxon

Taxon	Number containing repeats/total	Percentage
Archaea	27/3428	0.79
Viruses	81/8048	1.00
Bacteria	299/28438	1.05
Fungi	232/8334	2.78
Viridiplantae	153/6963	2.20
Metazoa	1538/28948	5.31
Rest of Eukaryota	92/2434	3.78

DETECTION OF REPEATS

Identifying tandem repeats with high sequence similarities is relatively straightforward. Detecting homologous repeats whose similarities are low, however, represents a more considerable challenge. Compounding this is the issue of defining the boundaries of repeats. In some cases repeat boundaries may be assigned from the positions of flanking domains or repeats or from bona fide protein termini. Frequently the boundaries are predicted simply from an expectation that repeats occur in integer multiples and that homologues' repeat boundaries are always coincident.

Unfortunately repeats can occur in noninteger multiples and their boundaries often do not coincide. For example, arrays of bihelical repeats may consist of an integer number of helices 1-2, with a single additional flanking helix (helix 2 at the N-terminus or helix 1 at the C-terminus) representing a "half-repeat." Repeats in closed β-propeller barrel structures do occur only in integer multiples but often do not exactly correspond to the repeats seen in structure. This is due to the circular permutation of the sequence repeats with respect to the structure repeats.

FIG. 1. Tertiary structures of several proteins with structural repeats. Alternating repeats are shown in different colours. Kelch is the galactose oxidase from *D. dendroides* (Ito *et al.*, 1991) and Fgf is the acidic fibroblast growth factor from *H. sapiens* (Eriksson *et al.*, 1993); these are examples of different closed structures, a β-barrel and a β-trefoil, repectively. TPR is a fragment of the human protein phosphatase 5 (Das *et al.*, 1998). HEAT is the protein phosphatase 2A PR65/A from *H. sapiens*, which is an open solenoid-like structure (see text) (Groves *et al.*, 1999). LRR is the porcine ribonuclease inhibitor complexed with the ribonuclease (Kobe and Deisenhofer, 1995). Fib corresponds to the adenovirus fibre protein from the human adenovirus type 2 (van Raiij *et al.*, 1999); the two views of the structure show a triple β spiral (Table III). Iafp is the insect antifreeze protein from *Tenebrio molitor* (Liou *et al.*, 2000), a small all β protein (Table III). ANK is a fragment of the of the β-subunit of the of the GA-binding protein from mouse (Batchelor *et al.*, 1998) complexed with the α-subunit and 21 bp of DNA. The corresponding PDB identifiers are Kelch, 1gof. Fgf, chain A from 2afg. TPR, 1a17. HEAT, chain A from1b3u. LRR, chain I from 1dfj. Fib, chain A from 1qiu, Iafp, chain A from 1ezg. ANK, chain B from 1awc.

Nevertheless, repeat detection has become considerably easier in recent years due to the advent of Web-based resources, such as SMART (smart.embl-heidelberg.de; Schultz *et al.*, 1998) and Pfam (www.sanger.ac.uk/Pfam; Bateman *et al.*, 2000), both of which perform well in predicting frequently occurring repeats. A new server, REP (www.embl-heidelberg.de/~andrade/papers/rep/search.html, Andrade *et al.*, 2000), also is proficient in detecting common repeats. It is emphasised that, due to the problems outlined above, these and other methods are unable to predict all repeats with complete accuracy.

Identifying repetitive regions of single protein sequences invariably involves the analysis of suboptimal alignments. An optimal alignment of a sequence (with i amino acids) is the path with the highest associated alignment score taken through the $i \times i$ trace matrix. The first and subsequent suboptimal alignments are given by the next highest scoring paths. High-scoring paths can be visualized using Dotter (www.cgr.ki.se/cgr/groups/sonnhammer/Dotter.html; Sonnhammer and Durbin, 1995). Estimating whether such alignments represent past evolutionary duplication events or whether the internal sequence similarity arose simply by chance has, until recently, been a thorny issue. A classic approach to estimating the significance of sequence similarity has been to compare the alignment score to those generated by randomly shuffling the aligned sequences (McLachlan, 1983). Useful implementations of this have recently been described (Pellegrini *et al.*, 1999; Heger and Holm, 2000).

MACAW (Schuler *et al.*, 1991) can also be used to assess sequence similarity significance. By contrast to the aforementioned methods, MACAW provides probabilities P that the repeats have not arisen through chance alone. Here the sequence must be compared against itself and a search space used that is the square of the sequence length in amino acids. This method is not entirely satisfactory since it is not amenable to large-scale studies looking for internal repeats in more than one protein, and it considers only ungapped alignments.

One further elegant and statistically robust approach, which generates P values for suboptimal alignments, is provided in the Prospero/Ariadne suite (www.well.ox.ac.uk/~rmott/ariadne.html; Mott and Tribe, 1999; Mott, 2000). This method accounts for variations in sequence composition and length in its derivation of P for gapped alignments and thus should be the method of choice in assessing the significance of internal sequence similarities.

The popular BLAST suite of programs (Altschul *et al.*, 1997), and in particular PSI-BLAST, may also be used to detect repeats. It is emphasized, however, that BLAST's statistics are provided on the basis of optimal, rather than suboptimal, alignments. Consequently, these statistics are not able to provide good estimates of either P or E, the number of proteins with associated (optimal) alignment scores greater than, or equal to, a score x expected purely by chance. The presence of repeats in a sequence used as a query in PSI-BLAST runs is indicated usually by: (1) the *same* region of the query being aligned against two distinct regions of a second protein with an associated E value less than about 10 or (2) *different* regions of the query being aligned against the *same* region of a second protein, again with $E < 10$.

Once the presence of repeats with statistically significant similarities in a protein has been established, it is appropriate to construct their multiple alignment. Further repeat homologues, identified by (PSI-) BLAST searches of databases (with an E value inclusion threshold $E_T = 0.002$, for inclusion in the profile used in the subsequent search iteration) using the original repeats as queries should be added to this alignment. The multiple alignment should be optimized by hand editing following guidelines given elsewhere (Bork and Gibson, 1996; Ponting and Birney, 2000). From this alignment, a hidden Markov Model (HMM) may be constructed and compared with protein sequence databases using, for example, the HMMER suite (hmmer.wustl.edu; Eddy, 1998). HMMER is appropriate for collating protein repeats since it successfully applies a heuristic strategy to detect bona fide repeats whose individual E values (for optimal alignment statistics) appear to be insignificant, but are deemed significant by combining the highest scores of other repeats in the protein. Repeats should be considered significant if their (per-sequence, rather than per-repeat) E values are less than 0.1.

Detection Example: New Repeats in Spindlin

As an example of detecting repeats, we describe an analysis of spindlin, a spindle-associated protein with roles in early mouse embryo development (Oh *et al.*, 1997). Repeats were detectable within spindlin using one or more of four methods. First, comparison of this sequence with itself using Dotter (Sonnhammer and Durbin, 1995) showed similarity not only along the diagonal (which represents an exact match of the sequence with itself) but also in off-diagonal positions (which represent similar, but nonidentical, regions) (Fig. 2a). This suggests, but does not provide statistical evidence for, internal repeats within spindlin. Second, a gapped BLAST (Altschul *et al.*, 1997) search of NCBI's nonredundant database using the *Mus musculus* spindlin sequence as a query revealed significant similarity to,

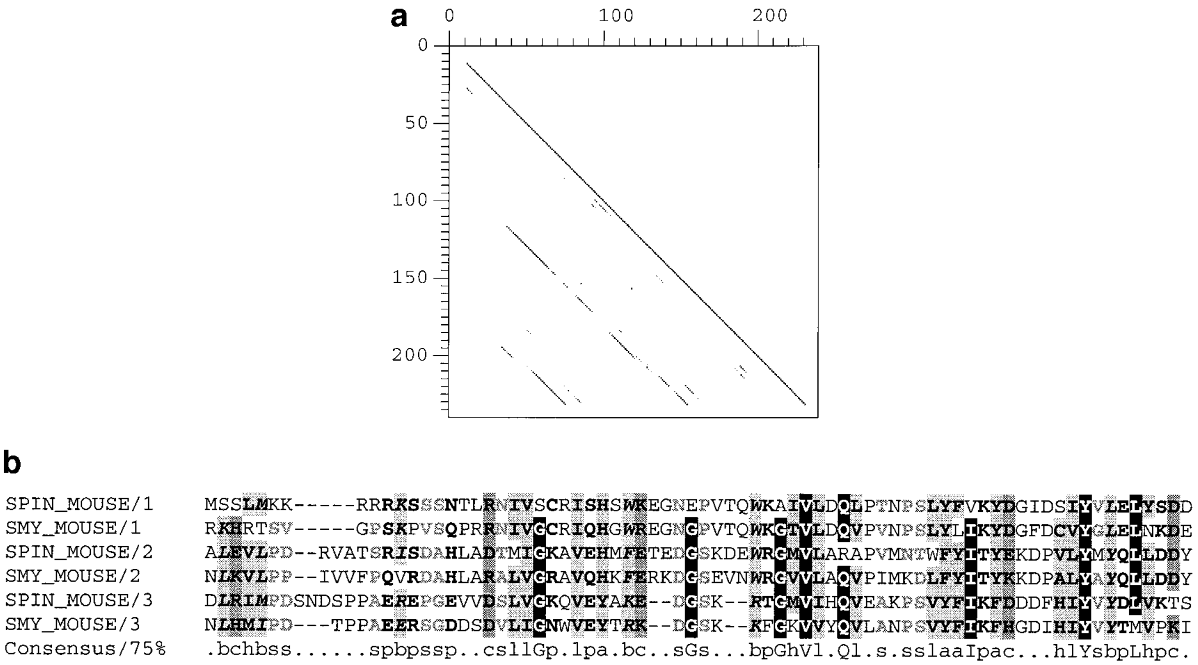


FIG. 2. Detection of repeats in spindlin. (a) Dot plot of spindlin (SPIN_MOUSE (horizontal) vs SPIN_MOUSE (vertical). (b) Multiple alignment of repeats in spindlin.

among others, its orthologue in *Mus spicilegus*. The significant similarity again resided not only along the diagonal ($E = 6 \times 10^{-63}$) but also in an off-diagonal second alignment ($E = 1 \times 10^{-5}$). Third, self-comparison of the spindlin sequence using Prospero (Mott and Tribe, 1999) showed two off-diagonal regions of significant similarity ($P = 1.1 \times 10^{-13}$ and 6.0×10^{-5}). Last, a self-comparison of spindlin using MACAW (Schuler *et al.*, 1991) revealed three pairs of ungapped alignment blocks with significant (6.9×10^{-9}, 5.2×10^{-5} and 6.2×10^{-3}) similarities (here, the relevant search space is the square of the number of amino acids in spindlin, 240^2).

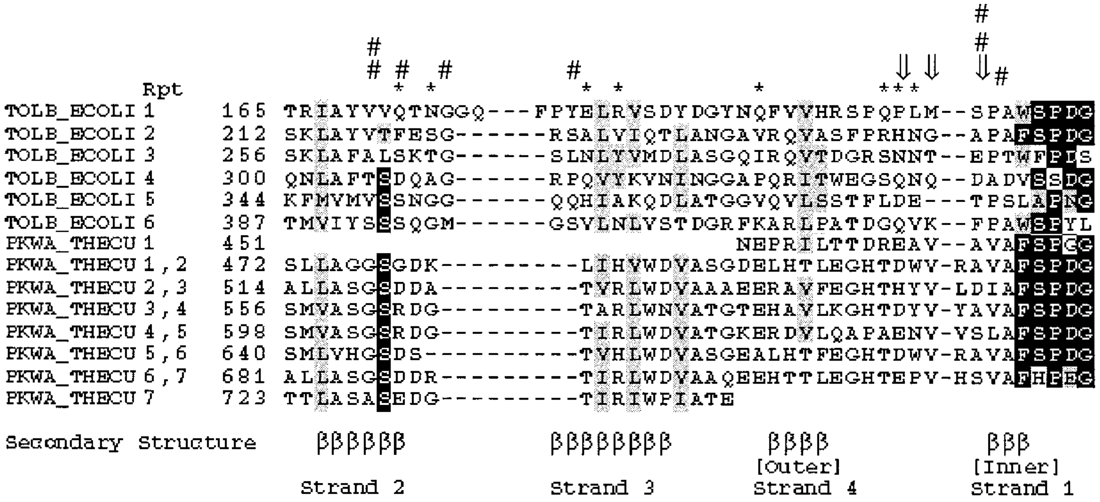
Once statistical significance of repeats was assured their sequences were multiply aligned (Fig. 2b). For this, the boundaries of repeats needed to be assigned. In the case of the three spindlin repeats this was not particularly problematic since these together span the complete protein sequence. Thus, the N-terminal repeat boundary coincides with the protein's N-terminus and the C-terminal boundary coincides with the protein's C-terminus. For the sake of completeness, a HMM constructed from the spindlin repeats' multiple alignment was compared with current protein sequence databases using HMMER (Eddy, 1998), but no further homologues were detected. The spindlin repeats appear to be all β-strand structures, but their functions remain unknown.

SIX MAJOR REPEAT FAMILIES

Many protein repeat families are known, each with different structures, functions, and phylogenetic distributions. For the purpose of this review, we have chosen to classify families according to their tertiary structures, although other ways of classification are of equal merit. The six repeat families we shall discuss (Table II) include two families each of the three major structural types: all-β (β-propellers and β-trefoils), all α structure (armadillo/HEAT and TPR-like repeats), and mixed α/β (leucine-rich and ankyrin repeats). These examples provide ample evidence for the evolutionary mechanisms of their propagation.

β-Propellers

The WD40 repeat (Neer *et al.*, 1994) is the most common repeat detected among known human proteins. These contain approximately 40 amino acids and include well-conserved Trp (W) and Asp (D) amino acids. The crystal structure of an assembly of seven WD40 repeats (e.g., Sondek *et al.*, 1996) revealed that each repeat represents a four-stranded antiparallel β-sheet (a "blade") arranged radially in a "propeller" arrangement about a central axis. Such β-propeller structures are also seen in methylamine dehydrogenase heavy chain (PQQ repeats), regulator of chromosome condensation 1 (RCC1 repeats),



```
                                       #                                                     #
                                       #
                                       #  #   #         #                          ⇓  ⇓     ⇓#
            Rpt                           *   *           *  *            *          ***
TOLB_ECOLI  1     165  TRIAYVVQTNGGQ---FPYELRVSDYDGYNQFVVHRSPQPLM--SPAWSPDG
TOLB_ECOLI  2     212  SKLAYVTFESG------RSALVIQTLANGAVRQVASFPRHNG--APAFSPDG
TOLB_ECOLI  3     256  SKLAFALSKTG------SLNLYVMDLASGQIRQVTDGRSNNT--EPTWFPIS
TOLB_ECOLI  4     300  QNLAFTSDQAG------RPQVYKVNINGGAPQRITWEGSQNQ--DADVSSDG
TOLB_ECOLI  5     344  KFMVMVSSNGG------QQHIAKQDLATGGVQVLSSTFLDE---TPSLAPNG
TOLB_ECOLI  6     387  TMVIYSSSQGM------GSVLNLVSTDGRFKARLPATDGQVK--FPAWSPYL
PKWA_THECU  1     451                               NEPRILTTDREAV--AVAFSPGG
PKWA_THECU  1,2   472  SLLAGGSGDK---------LIHVWDVASGDELHTLEGHTDWV-RAVAFSPDG
PKWA_THECU  2,3   514  ALLASGSDDA---------TVRLWDVAAAEERAVFEGHTHYV-LDIAFSPDG
PKWA_THECU  3,4   556  SMVASGSRDG---------TARLWNVATGTEHAVLKGHTDYV-YAVAFSPDG
PKWA_THECU  4,5   598  SMVASGSRDG---------TIRLWDVATGKERDVLQAPAENV-VSLAFSPDG
PKWA_THECU  5,6   640  SMLVHGSDS----------TVHLWDVASGEALHTFEGHTDWV-RAVAFSPDG
PKWA_THECU  6,7   681  ALLASGSDDR---------TIRLWDVAAQEEHTTLEGHTEPV-HSVAFHPEG
PKWA_THECU  7     723  TTLASASEDG---------TIRIWPIATE

Secondary Structure    ββββββ          ββββββββ      ββββ          βββ
                                                    [Outer]       [Inner]
                       Strand 2        Strand 3     Strand 4      Strand 1
```

FIG. 3. Prediction of a supersite in TolB. Sequence analysis indicated the presence of a β-propeller domain in TolB (Ponting and Pallen, 1999a). On the basis of supersite information (Russell *et al.*, 1998), the binding site of TolB was mapped from other β-propeller heterodimer structures onto the multiple alignment (alignment positions marked with asterisks). This prediction corresponds well with several amino acids involved in suppressor mutations of *pal* A88V (Ray *et al.*, 2000).

and galactose oxidase (Kelch repeats) (each containing seven blades) and in neuraminidase (containing six blades) (reviewed in Murzin, 1992).

In recent years several families of domains have been shown to adopt β-propeller structures with four, five, six, seven, or eight blades. These structures may be browsed using the SCOP resource (scop.mrc-lmb.cam.ac.uk/scop/; Lo Conte *et al.*, 2000). Several other families of repeats have also been predicted to adopt a propeller-like structure, for example, YWTD (Springer, 1998) and integrin α subunits (Springer, 1997).

β-Propeller structures are closed structures with interactions between the N- and C-terminal repeats. As described previously, the periodicities of some β-propeller repeats do not exactly match the periodicities of their repeats structures. In these cases the sequence repeat is circularly permuted with respect to the structural repeat. A "Velcro" model of closure of propellers has been proposed (Neer and Smith, 1996), with one of the blades being formed from β-strands from both the most N-terminal and the most C-terminal of sequence repeats.

Repeat families commonly represent either enzymes or nonenzymes, but rarely both. It is unusual therefore that some β-propellers are enzymes, whereas others are not. Whether catalytic or not, β-propellers have a significant preference for binding proteins and other ligands along the propeller axis at the surface formed by the N-termini of interior β-strands (Russell *et al.*, 1998). This observation of a ligand-binding "supersite" in β-propellers was recently used to predict residues that contribute to the ligand-binding site of TolB (Ponting and Pallen, 1999a). Not only was the prediction (Ponting and Pallen, 1999a) of a β-propeller domain in TolB correct (Abergel *et al.*, 1999) but also the predicted ligand-binding residues (in the loops between β-strands 2 and 3, and 4 and 1, Fig. 3) were found to correlate with experimentally derived functional residues (Ray *et al.*, 2000) (Fig. 3). This demonstrates that supersite information can be used to predict binding-sites even in the absence of tertiary structure data.

Recent studies indicate that the β-propellers of multidomain proteases may directly select substrates by size exclusion. The crystal structure of the prolyl oligopeptidase β-propeller domain shows that it lacks the usual "Velcro" of a blade formed by N- and C-terminal β-strands (Fülöp *et al.*, 1998). Instead, the terminal blades associate only via hydrophobic interactions. The enzyme's active site, which cleaves substrates no longer than 30 amino acids, faces the narrow (~4 Å) entrance of the propeller. It is proposed that this entrance is enlarged by the "breathing" of the propeller between the first and last blades. The size of the enlarged entrance is thought to act to exclude large substrates, thereby preferentially specifying the small (<30 amino acid) polypeptide substrates. By analogy, a similar mechanism has been proposed for the β-propeller domain of the tricorn protease (Ponting and Pallen, 1999b).

β-Trefoils

Another all β-sheet "closed" structure with internal repeats is the β-trefoil. This fold is found in known tertiary structures of fibroblast growth factors (FGFs), interleukin-1s, Kunitz soybean trypsin inhibitors, ricin-like toxins, plant agglutinins, and hisactophilin-like actin-bundling proteins (Murzin *et al.*, 1992; Ponting and Russell, 2000). By contrast to the β-propellers, however, β-trefoils do not appear to possess a "supersite" since members of the fold family often bind their respective protein ligands in different topological locations (Russell *et al.*, 1998). Consequently, predictions of binding sites, such as those described above for β-propellers, are not plausible.

A recent study of β-trefoil structures and sequences (Ponting and Russell, 2000) provides insights into the evolution of closed repeat assemblies. The β-trefoil fold consists of six two-stranded β-hairpins, three of which form a barrel structure, while the remaining three form a triangular cap on the barrel (Murzin *et al.*, 1992). Three pairs of these two-stranded β-hairpins can be seen as repeats in the crystal structures, but are not immediately apparent from their sequences. The recent more detailed analysis, however, demonstrated the presence of four β-trefoils in the actin-binding proteins fascins and showed that the internal triplications within each of the β-trefoils are significantly similar in sequence.

This indicates that the three internal repeats in fascin β-trefoils arose not via convergent evolution but instead by divergence from a single repeat common ancestor. As a protein possessing only a single repeat is unlikely to be stable as a monomer, perhaps the most parsimonious explanation for the evolution of the β-trefoil triplicated repeat is that a homotrimer-forming progenitor repeat underwent successive gene duplication events giving rise to a three-repeat-containing monomer. We return to this issue at the end of this review.

TPR-Like

Tetratricopeptide repeats contain approximately 34 amino acids arranged in two α-helices that are packed together in a knobs-in-holes manner (Sikorski *et al.*, 1990; Lamb *et al.*, 1995). Convergent evolution of TPRs is unlikely given its relatively strong conservation of sequence. The TPR is likely to be an ancient repeat since it is found in eukarya, bacteria, and archaea (Ponting *et al.*, 1999). Multiple TPRs form a right-handed superhelix (Das *et al.*, 1998) with a groove of large surface area available for ligand binding. This groove is employed in the binding of molecular chaperone Hsp70's C-terminal tail (Scheufler *et al.*, 2000). By contrast the groove is not used for molecular recognition by the TPRs of p67$^{\text{phox}}$ (Lapouge *et al.*, 2000). Thus TPR assemblies show multiple modes of ligand binding and do not appear to possess a single supersite.

TPRs come in many different flavours that form distinct sequence subfamilies. These include repeats in: kinesin light chains (Ginhart and Goldstein, 1996), SNAP secretory proteins (Ordway *et al.*, 1994), clathrin heavy chains, and bacterial aspartyl-phosphate phosphatases (Andrade *et al.*, 2000). In-depth studies of helical repeats (Andrade *et al.*, 2000; Ponting, 2000) also show that repeat families, such as HAT repeats (Preker and Keller, 1998), protein farnesyl transferase α-subunit repeats (Boguski *et al.*, 1992), and Sel-1 repeats are distant homologues of TPRs. These sequence-based studies indicate that the characteristic bihelical TPR has proliferated as a result of its ability to acquire multiple functional roles. However, the prediction of these different roles solely on the basis of sequence currently remains elusive.

Ankyrin

These repeats take their name from one of the proteins in which they were first found, the human erythrocyte protein ankyrin (Lux *et al.*, 1990). Each repeat contains approximately 33 residues and forms an L-shaped structure consisting of two anti-parallel α-helices followed by a β hairpin (Gorina and Pavletich, 1996). The hairpins of different repeats pack tightly together forming an anti-parallel β-sheet. Hydrophobic residues in the α-helices form complementary nonpolar surfaces that assemble forming an extended helical bundle. Additional hydrogen bonds between residues of adjacent repeats contribute to further stabilization of the ensemble. The smaller sizes of the side chains lining the inner α-helices, and the left-handed twist of the stacking, produce a characteristic solvent-accessible groove (Sedgwick and Smerdon, 1999).

The function of the ankyrin repeats is to bind other proteins but they do not bind a single class of proteins. For example, several structures show ankyrin repeats complexed with another proteins (reviewed in Swedgwick and Smerdon, 1999), such as p53 (a nuclear tumour suppressor), CDK6 (cell division protein kinase), and p65 (a transcriptional regulator). Other known cases are the interaction between the development protein Notch and deltex (a cytoplasmic protein) (Diederich *et al.*, 1994) and the interaction between the noncatalytic subunit M130 and the catalytic subunit PP1c of the smooth muscle myosin phosphatase (Hirano *et al.*, 1997).

These tertiary structures of complexes show that although there is considerable sequence variation at the heterodimer interface, the interactions involve the extended groove formed by the anti-parallel β-sheet (Sedgwick and Smerdon, 1999). This mechanism is similar to that observed in armadillo and HEAT repeats.

Ankyrin repeats are present in a large number of protein families, including transcription factors, development regulators, cytoskeletal proteins, and toxins. Sequence and taxonomic analysis of these repeats suggests that their phyletic propagation between eukaryotes, bacteria, and viruses has involved multiple events of horizontal gene transfer (Bork, 1993). For example, the only archaeal sequence currently known to have these repeats (possibly five copies) is a *Thermoplasma acidophilum* hypothetical sequence (SPTREMBL code Q9HLN1) that is more similar to other eukaryotic sequences than to any archaeal sequence.

Armadillo/HEAT

Armadillo repeats (Peifer *et al.,* 1994) were first identified in the product of the eponymous *D. melanogaster* segment polarity gene (Riggleman *et al.,* 1989). They were later found in several eukaryotic proteins, including the junctional plaque protein plakoglobin, β-catenin, the tumour suppressor adenomatous polyposis coli, and the nuclear transport factor importin-α, among others.

HEAT repeats derive their name from four diverse eukaryotic proteins in which they were first identified: *h*untingtin (involved in Huntington's disease), *e*longation factor 3, PR65/*A* subunit of protein phosphatase A, and the *T*OR (target of rapamycin) (Andrade and Bork, 1995). It is also present in importins β1 and β2 (with a Ran-binding function), in proteins related to the clathrin-associated adaptor complex (Andrade and Bork, 1995), in the microtubule-binding colonic and hepatic tumor-related protein (CTOG) family (Andrade *et al.,* 2000) and in many other proteins related to chromosome dynamics (Neuwald and Hirano, 2000).

Armadillo repeats consist of three $\alpha-$helices. The first of these is short (about eight amino acids long) and lies perpendicular to the other two, longer, α-helices that pack against one another. HEAT repeats have two anti-parallel $\alpha-$helices. The first HEAT helix has a kink (of variable extent) that makes it equivalent to both the first and the second helices of armadillo repeats. The C-terminal helices of both armadillo and HEAT repeats are also superimposable. The parallel stacking of repeat units forms a solenoid. Depending on the structure, these solenoids may have different degrees of curvature but all exhibit a groove formed by the last helix of each repeat. As in ankyrins, protein–protein interactions have been seen to occur within this groove. The binding of importin-α by importin-β (Cingolani *et al.,* 1999), Ran^{GTP} by transportin (Chook *et al.,* 1999), and nuclear localization signal peptides by importin-α (Conti *et al.,* 1998) all exhibit binding sites within this groove. However, protein recognition can also occur on the opposite end of the solenoid, as with the binding of FxFG nucleoporin repeats by importin-β (Bayliss *et al.,* 2000). Further similarities between Armadillo and HEAT repeat families include a series of conserved residues that form the repeats' hydrophobic cores (Andrade *et al.,* 2001).

In some cases sequence and structural features can distinguish between different variants of these repeats (discussed in Andrade *et al.,* 2001). For example, for the HEAT repeats of the PR65/A subunit of protein phosphatase A, charged residues in the loop linking the repeats' $\alpha-$helices were shown to form a ladder of electrostatic interactions between adjacent repeats (Groves *et al.,* 1999). These are also present in the HEAT repeats of elongation factor 3, but not in those of importin-β. A conserved asparagine in the last helix of armadillo repeats is involved in protein–protein contacts, such as recognition of the nuclear localization signal by importin-α (Conti *et al.,* 1998). This conserved asparagine is absent in HEAT repeats.

A common phylogenetic origin (homology) for the armadillo and HEAT repeats present in the nuclear protein transport complex has been proposed (Malik *et al.,* 1997; Cingolani *et al.,* 1999). Other repeat families are known which exhibit considerable structural similarity to armadillo/HEAT repeats but show no detectable sequence similarity. These include the all helical structures of VHS domains (Lohi and Lehto, 1998; Mao *et al.,* 2000) and regions of phosphoinositide 3-kinase γ (Walker *et al.,* 1999) and eukaryotic initiation factor 4G (Marcotrigiano *et al.,* 2001). Without additional evidence, divergent and convergent evolution of HEAT/Armadillo repeats and these structures appear equally plausible.

Leucine-Rich Repeats

Leucine-rich repeats (Kobe and Deisenhofer, 1994) (LRRs) are relatively short in comparison to other repeat families, with lengths of about 20 amino acids. They are associated with an astonishing variety of functions, including signal transduction, transmembrane receptors, DNA repair, cell adhesion, and extracellular matrix proteins. They are also not restricted to eukaryotes, since bacterial and viral versions are known. The common function among LRRs is that they form complexes with other proteins. For example, the LRRs of ribonuclease A inhibitor bind to ribonuclease A (Kobe and Deisenhofer, 1995), LRRs of the extracellular matrix leucine-rich repeat glycoprotein/proteoglycan family (Iozzo, 1998) interact with transforming growth factor β (Hildebrand *et al.,* 1994) and collagen (Svenson *et al.,* 2000), LRRs of platelet glycoproteins associate with thrombin and von Willebrand factor (Shen *et*

TABLE II
Examples of Frequently Occurring Repeat Families

Repeat	Ref1	L	3D	PDB	Ref2	Distribution	Function	Pfam
Kelch	Neer *et al.* (1994)	40	β-Barrel	1gof	Ito *et al.* (1991)	Eukaryotic	Enzyme. Protein processing	PF01344
Fibroblast growth factor	Murzin *et al.* (1992)	40	β-Trefoil	2afg_A	Eriksson *et al.* (1993)	Eukaryotic–viral	Development	PF00167
Tetratrico-peptide repeats	Zhang *et al.* (1991)	34	α-α	1a17	Das *et al.* (1998)	Eukariotic–bacterial–archaeal	PPI	PF00515
Ankyrin	Lux *et al.* (1990)	33	α-α-β-Hairpin	1awc_B	Batchelor *et al.* (1998)	Eukaryotic–bacterial–viral	PPI	PF00023
HEAT	Andrade and Bork (1995)	47	α-α	1b3u_A	Groves *et al.* (1999)	Eukaryotic	PPI	None
Leucine-rich repeats	Kobe and Deisenhofer (1994)	20	α-β	1dfj_I	Kobe and Deisenhofer (1995)	Eukaryotic–bacterial	PPI	PF00560

Note. Abbreviations used: Repeat, name of the repeat; Ref1, the original description and/or characterization of the repeat in the literature; L (length), average length of the repeat in amino acids; 3D, fold category; PDB, the PDB identifier of the structure shown in Ref2; Distribution, phyletic distribution of the repeat family; Function, summary of the function of the family (PPI, protein–protein interaction); Pfam, Identifier of the corresponding entry in the Pfam database.

al., 2000), and LRRs of plant disease resistance gene products form a pathogen-recognition domain (Van Der Biezen and Jones, 1998).

The first crystal structures of LRRs showed each repeat to contain a β-strand and an α-helix that are oriented in an antiparallel manner (Kobe and Deisenhofer, 1995; Price *et al.*, 1998). The side-by-side association of repeats builds an arch, with the β-strands forming the arch's interior harboring an extended protein-binding surface.

Somewhat surprisingly, later structures were found to be rather different. In particular, the structure of the Internalin B protein from *Listeria monocytogenes* also shows an array of β-strands, forming the inside surface of the arch, but its outside surface is composed of 3_{10}, rather than α-, helices (Marino *et al.*, 1999).

The so-called leucine-rich-variant repeats of a hypothetical protein from *Azotobacter vinelandii* also assemble as an arch, but with an α-helix on its inside and a 3_{10} helix on its outside (Peters *et al.*, 1996). Furthermore, there is only slight sequence similarity to leucine-rich repeats in their patterns of conserved hydrophobic residues. Therefore, these repeats are unlikely to be homologues of leucine-rich repeats.

OTHER REPEAT FAMILIES

Other protein families are too numerous to describe here. Instead, in this section we shall discuss families that demonstrate important differences in structure, function, and evolution, when compared to β-propellers, β-trefoils, and TPRs, and ankyrin, ARM/HEAT and leucine-rich repeats (see Table III). Since these six repeat families form regular nonfibrous and monomeric structures, other repeat families that lack structure, that form rod-like structures or that form oligomers, will be discussed.

The fibronectin-binding repeats of staphylococcal proteins are known not to form a regular tertiary structure in solution (Penkett *et al.*, 2000). These are unusual in that they appear to only adopt a regular tertiary structure when bound to their ligand, the mammalian extracellular protein fibronectin. Their unfolded conformations may be linked to the bacterial proteins' abilities to evade both proteolytic and immune defenses of the mammalian hosts.

Many repeats form rigid linear arrays, or rods. A great number of these are oligomeric coiled-coil proteins containing between two and five amphipathic α-helices (Burkhard *et al.*, 2001). These long helices often wind about one another forming parallel left-handed coiled coils. These structures contain characteristic seven-residue (heptad) repeats and may extend up to several tens of nanometers long.

By contrast to these fibrous proteins of α-structure, long filaments can be composed of repeated β-structures, such as in the adenovirus fiber protein (van Raaij *et al.*, 1999). The crystal structure of the shaft region of this protein shows that it forms homotrimers with close association of three two-strand repeating units (the "triple β-spiral fold") in the shaft. Consequently, beyond its construction from β-structure rather than from α-helices, it is similar to three-chain coiled-coil filaments.

Other repeat families provide additional insights into the evolution of repeated structures.

Filaments may also be built from short, few resi-

TABLE III
Other Less Frequently Occurring Repeat Families

Repeat	Ref1	L	3D	PDB	Ref2	Distribution	Function	Pfam
β-Farnesyl transferase	Park *et al.* (1997)	42	α-Barrel	1ft2b	Park *et al.* (1997)	Eukaryotic	Enzyme. Protein processing	None
Adenovirus fiber protein	Green *et al.* (1983)	15	Triple β spiral	1qiu	van Raiij *et al.* (1999)	Viral	PPI. Binds to host receptor	None
Zein	Argos *et al.* (1982)	20	α-Helix (proposed)	Model	Matsushima *et al.* (1997)	Plants	Plant seed storage protein	PF01559
Bacterial glycosyl transferase	Wren (1991)	35	Unknown	None		Bacterial	Enzyme. Small molecules binding	None
Insect antifreeze protein	Graham *et al.* (1997)	12	β-sheet	1ezg_A	Liou *et al.* (2000)	Metazoa	Ice binding. Antifreeze	None
Ice nucleation protein	Gurian-Sherman and Lindow (1993)	16	Hairpin-loop	1ina	Tsuda *et al.* (1997)	Bacterial	Catalyst of ice formation	PF00818
Nebulin	Pfuhl *et al.* (1996)	35	α-Helix (proposed)	None		Metazoa	PPI. Binds to F-actin	PF00880
Notch/lin-12	Wharton *et al.* (1985)	31	Unknown	None		Metazoa	PPI. Lateral inhibition of development processes	PF00066
Plectin	Wiche *et al.* (1991)	38	Unknown	None		Metazoa	PPI. Cytoskeleton. Cell adhesion. Antigens	PF00681
Spectrin	Speicher and Marchesi (1984)	106	Three-helix bundle	1cun	Pascual *et al.* (1997)	Metazoa	PPI. Cell shape. Cytoskeleton	PF00435
Annexin	Barton *et al.* (1991)	60	Five-helix bundle	1ain	Weng *et al.* (1993)	Eukaryotic	Regulatory. Membrane fusion. Exocytosis	PF00191
Flocculin	Watari *et al.* (1994)	45	Unknown	None		*S. cerevisiae*	Regulatory of flocculation	PF00624
Major vault protein	Vasu *et al.* (1993)	52	Unknown	None		Eukaryotic	Multidrug resistance	PF01505

Note. The columns are defined as in Table II. Here the Notch repeat is also called lin12.

due, repeats. Spider silk proteins contain numerous glycine-rich repeats: GPGG$(X)_n$ β-turn spiral and GGX 3_{10} helix repeats; here X, denotes any residue. Interestingly only a subset of silk protein genes contain introns, but these introns show even greater average sequence identity among themselves (87%) than do the exons (73%) (Hayashi and Lewis, 2000). One explanation for this is that the coding regions have undergone accelerated evolution (Hill and Hastie, 1987), due to extreme selective pressures arising from the importance of these genes to the spider's survival. Meanwhile, the conservation of introns is associated with rapid internal duplication of gene portions, due in part to slippage during replication. Thus, rapid internal gene duplications and mutation might also account, although to lesser extents, for many other repetitive proteins, including each of those discussed previously.

Flocculation in yeast is mediated, in part, by flocculins which, in *Saccharomyces cerevisiae,* contain at least four flocculin repeats. The only exception to this is YHR213w, whose hypothetical translation product contains a single flocculin repeat. Examination of the genomic sequence of yeast chromosome VIII around YHR213w indicates that the similarity to a neighboring flocculin gene (Flo5) extends beyond both the N- and C-terminal ends of the open reading frame over a number of stop codons. This is a clear indication of a pseudogene, and it is identified as such in a *S. cerevisiae* database (the Munich Information Centre for Protein Sequences, www.mips.biochem.mpg.de/proj/yeast/). This is an example where a possible error in the predicted gene structure may be highlighted when a conceptual translation of a genomic sequence presents an unusual domain architecture (defined as the sequential arrangement of domains, repeats, and motifs).

The more pervasive functions displayed by repeat ensembles are catalysis and protein–protein recognition. However, a repetitive structure can be used for other different tasks. The multiplicity of repeats that mimick water structure is a good example of the functional flexibility that can be acquired via protein repeat evolution. On one hand, insect and plant proteins protect themselves from freezing using repeats that impede ice formation (Liou *et al.*, 2000; Worrall *et al.*, 1998). On the other hand, bacterial proteins use different repeat types to favor the formation of ice as a mechanism of weakening an infected plant (Gurian-Sherman and Lindow, 1993).

A more passive function is played by the repeats of the plant storage proteins, α-prolamins, First identified in maize zein proteins (Argos *et al.*, 1982) these repeats are likely to form a layer of helices packed in an hexagonal arrangement (Matsushima *et al.*, 1997). In this case, the structure of the repeat bears little relation to its organismal function, since it is the unusual composition of nitrogen-rich amino acids that is required for its seed germination properties.

The vault is a ribonuclear particle observed in higher and lower eukaryotes. Its function remains unclear, but its elevated expression in cancer lines seems to be related to multidrug resistance (Kickhoefer *et al.*, 1998). The whole molecule is hollow and this suggested that drugs may be sequestered from their targets inside the particle (Kong *et al.*, 1999); 78% of the total mass of the particle is composed of 96 copies of the MVP (major vault protein, Vasu *et al.*, 1993). MVP homologues display seven copies of a 52-amino-acid repeat. These numbers resemble repeats present in β−propellers, in particular RCC1 repeats (Renault *et al.*, 1998), suggesting that MVP repeats may also form a similar closed structure.

CONCLUSIONS

Our survey of protein repeats has highlighted the multifunctionality of repeat types, their structural differences, and their proliferations in different evolutionary lineages. One likely reason for their evolutionary success is that repeat-containing proteins are relatively "cheap" to evolve. By this we mean that large and thermodynamically stable proteins may arise by the simple expedient of intragenic duplications, rather than the more complex processes of *de novo* α-helix and β-sheet creation. This is supported by the larger sizes of most repeat-containing structures relative to compact domains (Fig. 4).

This does not, of course, present a complete answer to their success since it addresses the question of how repeat-containing proteins arose, rather than why they have been selected for and fixed in evolutionary lineages on so many separate occasions. As suggested throughout this review, the reasons for the *functional* successes of repeat classes may be a proclivity of repeat assemblies to acquire different molecular functions, namely, the association with different protein ligands. This, in turn, might be associated with the large solvent-accessible surface areas, presented by extended "open" assemblies, that are available for interactions with ligands. This is because burial of nonpolar residues at protein–protein interfaces is thought to be an important contributor to heterodimer stability (Tsai *et al.*, 1997).

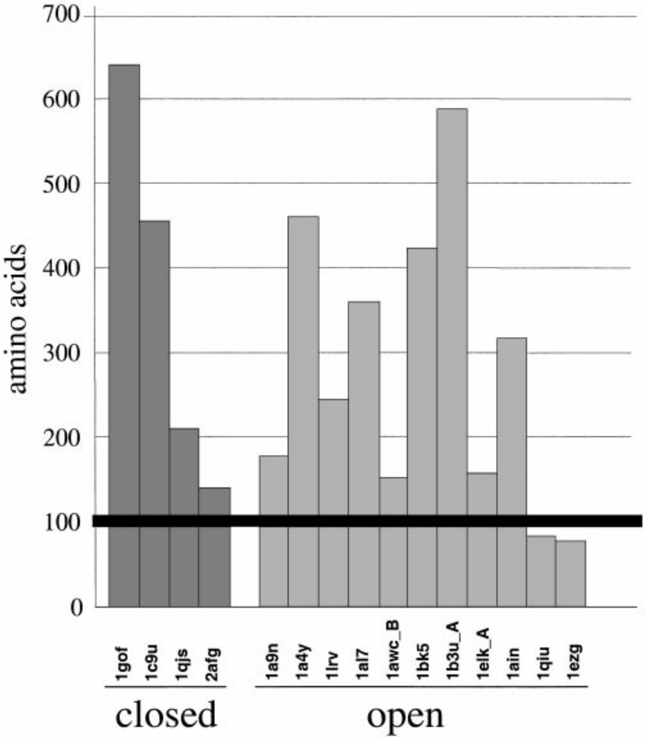


FIG. 4. Distribution of domain size in known structures. The bold line indicates the average size of domains, of approximately 100 amino acids (Wheelan *et al.*, 2000). Repeats and their corresponding PDB codes are shown (from left to right). Closed structures: kelch, 1gof; glucose dehydrogenase-B, 1c9u; hemopexin, 1qjs; fibroblast growth factor, 2afg; open structures: LRR/typical, 1a9n; LRR/ribose inhibitor, 1a4y; LRV, 1lrv; TPR, 1al7; ankyrin, 1awc_B; armadillo, 1bk5; HEAT, 1b3u_A; VHS, 1elk_A; annexin, 1ain; adenovirus fibrous protein, 1qiu; IAFP, 1ezg.

In understanding the evolution of repeats, one major problem remains. Repeats are defined as occurring multiply, and all repeats in a family are homologous. This means that these repeats all evolved from a common ancestor, which necessarily must have contained only a single repeat. This is apparently contradictory, since it is not expected that a single repeat could exist in isolation, as a single folded functional unit. Rescue is at hand if one suggests that the family's common ancestor indeed represented a single repeat, but one that formed homooligomers. The homooligomeric structure of the ancestor might mirror that of the intrachain repetitive structure of its modern homologue, except in its multichain character. This scenario has recently been suggested for the evolution of the β-trefoil fold (Ponting and Russell, 2000).

A problem with this proposal is that there are few, if any, known examples where homologous multirepeat assemblies are formed *both* from oligomers of single repeats *and* from a single chain of multiple repeats. However, this might not be too surprising since the highly cooperative process of folding a multirepeat protein must be significantly more favorable than folding a homooligomeric protein from its constituent monomers. This is because the kinetic folding pathways of multirepeat protein structures may be nucleated at many positions. In this way ancient oligomeric single repeat proteins might have been driven to extinction by their monomeric multiple repeat-containing homologues.

REFERENCES

Abergel, C., Bouveret, E., Claverie, J. M., Brown, K., Rigal, A., Lazdunski, C., and Benedetti, H. (1999) Structure of the *Escherichia coli* TolB protein determined by MAD methods at 1.95 A resolution, *Struct. Fold Des.* **7,** 1291.

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

Andrade, M. A., and Bork, P. (1995) HEAT repeats in the Huntington's disease protein, *Nature Genet.* **11,** 115–116.

Andrade, M. A., Petosa, C., O'Donoghue, S. I., Müller, C. W., and Bork, P. (2001) Comparison of ARM and HEAT protein-repeats, *J. Mol. Biol.* **309,** 1–18.

Andrade, M. A., Ponting, C., Gibson, T., and Bork, P. (2000) Identification of protein repeats and statistical significance of sequence comparisons, *J. Mol. Biol.* **298,** 521–537.

Argos, P., Pedersen, K., Marks, M. D., and Larkins, B. A. (1982) A structural model for maize zein proteins, *J. Biol. Chem.* **257,** 9984–9990.

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

Barton, G. J., Newman, R. H., Freemont, P. S., and Crumpton, M. J. (1991) Amino acid sequence analysis of the annexin super-gene family of proteins, *Eur. J. Biochem.* **198,** 749–760.

Batchelor, A. H., Piper, D. E., de la Brousse, F. C., McKnight, S. L., and Wolberger, C. (1998) The structure of GABPα/β: An ETS domain-ankyrin repeat heterodimer bound to DNA, *Science* **279,** 1037–1041.

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

Bayliss, R., Littlewood, T., and Stewart, M. (2000) Structural basis for the interaction between FxFG nucleoporin repeats and importin-beta in nuclear trafficking, *Cell* **102,** 99–108.

Boguski, M. S., Murray, A. W., and Powers, S. (1992) Novel repetitive sequence motifs in the alpha and beta subunits of prenyl-protein transferases and homology of the subunit to the MAD2 gene product of yeast, *New Biol.* **4,** 408–411.

Bork, P., and Gibson, T. J. (1996) Applying motif and profile searches, *Methods Enzymol.* **266,** 162–184.

Burkhard, P., Stetefeld, J., and Strelkov, S. V. (2001) Coiled coils: A highly versatile protein folding motif, *Trends Cell Biol.* **11,** 82–88.

Cingolani, G., Petosa C., Weis K., and Müller, C. W. (1999) Structure of importin-β bound to the IBB domain of importin-α, *Nature* **399,** 221–229.

Chook, Y. M., and Blobel, G. (1999) Structure of the nuclear transport complex karyopherin-beta2-Ran × GppNHp, *Nature* **399,** 230–237.

Conti, E., Uy, M., Leighton, L., Blobel, G., and Kuriyan, J. (1998) Crystallographic analysis of the recognition of a nuclear localization signal by the nuclear import factor karyopherin alpha, *Cell* **94,** 193–204.

Das, A. K., Cohen, P. W., and 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.

Diederich, R. J., Matsuno, K., Hing, H., and Artavanis-Tsakonas, S. (1994) Cytosolic interaction between deltex and Notch ankyrin repeats implicates deltex in the Notch signaling pathway, *Development* **120,** 473–481.

Eddy, S. (1998) Profile hidden Markov models, *Bioinformatics* **14,** 755–763.

Eriksson, A. E., Cousens, L. S., and Matthews, B. W. (1993) Refinement of the structure of human basic fibroblast growth factor at 1.6 A resolution and analysis of presumed heparin binding sites by selenate substitution, *Protein Sci.* **2,** 1274–1284.

Fülöp, V., Böcskei, Z., and Polgár, L. (1998) Prolyl oligopeptidase: An unusual β-propeller domain regulates proteolysis, *Cell* **94,** 161–170.

Gindhart, J. G., Jr., and Goldstein, L. S. (1996) Tetratrico peptide repeats are present in the kinesin light chain, *Trends Biochem Sci.* **21,** 52–53.

Gorina, S., and Pavletich, N. P. (1996) Structure of the p53 tumor suppressor bound to the ankyrin and SH3 domains of 53BP2, *Science* **274,** 1001–1005.

Graham, L. A., Liou, Y. C., Walker, V. K., and Davies, P. L. (1997) Hyperactive antifreeze protein from beetles, *Nature* **388,** 727–728.

Green, N. M., Wrigley, N. G., Russell, W. C., Martin, S. R., and McLachlan, A. D. (1983) Evidence for a repeating cross-beta sheet structure in the adenovirus fibre, *EMBO J.* **2,** 1357–1365.

Groves, M. R., Hanlon, N., Turowski, P., Hemmings, B. A., and Bartford, D. (1999) The structure of the protein phosphatase 2A PR65/A subunit reveals the conformation of its 15 tandemly repeated HEAT motifs, *Cell* **96,** 99–110.

Gurian-Sherman, D., and Lindow, S. E. (1993) Bacterial ice nucleation: Significance and molecular basis, *FASEB J.* **7,** 1338–1343.

Hayashi, C. Y., and Lewis, R. V. (2000) Molecular architecture and evolution of a modular spider silk protein gene, *Science* **287,** 1477–1479.

Heger, A., and Holm, L. (2000) Rapid automatic detection and alignment of repeats in protein sequences, *Proteins* **41,** 224–237.

Hildebrand, A., Romaris, M., Rasmussen, L. M., Heinegard, D., Twardzik, D. R., Border, W. A., and Ruoslahti, E. (1994) Interaction of the small interstitial proteoglycans biglycan, decorin and fibromodulin with transforming growth factor beta, *Biochem. J.* **302,** 527–534.

Hill, R. E., and Hastie, N. D. (1987) Accelerated evolution in the reactive centre regions of serine protease inhibitors, *Nature* **326,** 96–99.

Hirano, K., Phan, B. C., and Hartshorne, D. J. (1994) Interactions of the subunits of smooth muscle myosin phosphatase, *J. Biol. Chem.* **272,** 3683–3688.

Iozzo, R. V. (1998) Matrix proteoglycans: From molecular design to cellular function, *Annu. Rev. Biochem.* **67,** 609–652.

Ito, N., Phillips, S. E., Stevens, C., Ogel, Z. B., McPherson, M. J., Keen, J. N., Yadav, K. D., Ju, B. G., Jeong, S., Bae, E., Hyun, S., Carroll, S. B., Yim, J., and Kim, J. (2000) Fringe forms a complex with Notch, *Nature* **405,** 191–195.

Kickhoefer, V. A., Rajavel, K. S., Scheffer, G. L., Dalton, W. S., Scheper, R. J., and Rome, L. H. (1998) Vaults are up-regulated in multidrug-resistant cancer cell lines, *J. Biol. Chem.* **273,** 8971–8974.

Kobe, B., and Deisenhofer, J. (1994) The leucine-rich repeat: A versatile binding motif, *Trends Biochem. Sci.* **19,** 415–421.

Kobe, B., and Deisenhofer, J. (1995) A structural basis of the interactions between leucine-rich repeats and protein ligands, *Nature* **374,** 183–186.

Kong, L. B., Siva, A. C., Rome, L. H., and Stewart, P. L. (1999) Structure of the vault, a ubiquitous cellular component, *Structure* **7,** 371–379.

Lamb, J. R., Tugendreich, S., and Hieter, P. (1995) Tetratrico peptide repeat interactions: To TPR or not to TPR? *Trends Biochem. Sci.* **20,** 257–259.

Lapouge, K., Smith, S. J. M., Walker, P. A., Gamblin, S. J., Smerdon, S. J., and Rittinger, K. (2000) Structure of the TPR domain of p67^{phox} in complex with Rac-GTP, *Mol. Cell* **6,** 899–907.

Liou, Y. C., Tocilj, A., Davies. P. L., and Jia, Z. (2000) Mimicry of ice structure by surface hydroxyls and water of a beta-helix antifreeze protein, *Nature* **406,** 322–324.

Lo Conte, L., Ailey, B., Hubbard, T. J., Brenner, S. E., Murzin, A. G., and Chothia, C. (2000) SCOP: A structural classification of proteins database, *Nucleic Acids Res.* **28,** 257–259.

Lohi, O., and Lehto, V. P. (1998) VHS domain marks a group of proteins involved in endocytosis and vesicular trafficking, *FEBS Lett.* **440,** 255–257.

Lux, S. E., John, K. M., and Bennett, V. (1990) Analysis of cDNA for human erythrocyte ankyrin indicates a repeated structure with homology to tissue-differentiation and cell-cycle control proteins, *Nature* **344,** 36–42.

Malik, H. S., Eickbush, T. H., and Goldfarb, D. S. (1997) Evolutionary specialization of the nuclear targeting apparatus, *Proc. Natl. Acad. Sci. USA* **94,** 13738–13742.

Mao, Y., Nickitenko, A., Duan, X., Lloyd, T. E., Wu, M. N., Bellen, H., and Quiocho, F. A. (2000) Crystal structure of the VHS and FYVE tandem domains of Hrs, a protein involved in membrane trafficking and signal transduction, *Cell* **100,** 447–456.

Marcotrigiano, J., Lomakin, I. B., Sonenberg, N., Pestova, T. V., Hellen, C. U. T., and Burley, S. K. (2001) A Conserved HEAT Domain within eIF4G Directs Assembly of the Translation Initiation Machinery, *Mol. Cell* **7,** 193–203.

Marcotte, E. M., Pellegrini, M., Yeates, T. O., and Eisenberg, D. (1999) A census of protein repeats, *J. Mol. Biol.* **293,** 151–160.

Marino, M., Braun, L., Cossart, P., and Ghosh, P. (1999) Structure of the lnlB leucine-rich repeats, a domain that triggers host cell invasion by the bacterial pathogen L. monocytogenes, *Mol. Cell* **4,** 1063–1072.

Matsushima, N., Danno, G., Takezawa, H., and Izumi, Y. (1997) Three-dimensional structure of maize alpha-zein proteins studied by small-angle X-ray scattering, *Biochem. Biophys. Acta* **1339,** 14–22.

McLachlan, A. D. (1983) Analysis of gene duplication repeats in the myosin rod, *J. Mol. Biol.* **169,** 15–30.

Mott, R., and Tribe, R. (1999) Approximate statistics of gapped alignments, *J. Comput. Biol.* **6,** 91–112.

Mott, R. (2000) Accurate Formula for P-values of gapped local sequence and profile alignments, *J. Mol Biol.* **300,** 649–659.

Murzin, A. G. (1999) Structural principles for the propeller assembly of beta-sheets: The preference for seven-fold symmetry, *Proteins* **14,** 191–201.

Murzin, A. G., Lesk, A. M., and Chothia, C. (1992) β-Trefoil fold. Patterns of structure and sequence in the Kunitz inhibitors interleukins-1 beta and 1 alpha and fibroblast growth factors, *J. Mol. Biol.* **223,** 531–543.

Neer, E. J., Schmidt, C. J., Nambudripad, R., and Smith, T. F. (1994) The ancient regulatory-protein family of WD-repeat proteins, *Nature* **371,** 297–300.

Neer, E. J., and Smith, T. F. (1996) G protein heterodimers: New structures propel new questions, *Cell* **84,** 175–178.

Neuwald, A. F., and Hirano, T. (2000) HEAT repeats associated with condensins, cohesins, and other complexes involved in chromosome-related functions, *Genome Res.* **10,** 1445–1452.

Oh, B., Hwang, S. Y., Solter, D., and Knowles, B. B. (1997) Spindlin, a major maternal transcript expressed in the mouse during the transition from oocyte to embryo, *Development* **124,** 493–503.

Ohno, S. (1970) Evolution by Gene Duplication, Springer-Verlag, Berlin.

Ordway, R. W., Pallanck, L., and Ganetzky, B. (1994) A TPR domain in the SNAP secretory proteins, *Trends Biochem Sci.* **19,** 530–531.

Pascual, J., Pfuhl, M., Walther, D., Saraste, M., and Nilges, M. (1997) Solution structure of the spectrin repeat: A left-handed antiparallel triple-helical coiled-coil, *J. Mol. Biol.* **273,** 740–751.

Peifer, M., Berg, S., and Reynolds, B. (1994) A repeating amino acid motif shared by proteins with diverse cellular roles, *Cell* **76,** 789–791.

Pellegrini, M., Marcotte, E. M., and Yeates, T. O. (1999) A fast algorithm for genome-wide analysis of proteins with repeated structures, *Proteins* **35,** 440–446.

Penkett, C. J., Dobson, C. M., Smith, L. J., Bright, J. R., Pickford, A. R., Campbell, I. D., and Potts, J. R. (2000) Identification of residues involved in the interaction of *Staphylococcus aureus* fibronectin-binding protein with the (4)F1(5)F1 module pair of human fibronectin using heteronuclear NMR spectroscopy, *Biochemistry* **39,** 2887–2893.

Peters, J. W., Stowell, M. H., and Rees, D. C. (1996) A leucine-rich repeat variant with a novel repetitive protein structural motif, *Nat. Struct. Biol.* **3,** 991–994.

Pfuhl, M., Winder, S. J., Castiglione Morelli, M. A., Labeit, S., and Pastore, A. (1996) Correlation between conformational and binding properties of nebulin repeats, *J. Mol. Biol.* **257,** 367–384.

Ponting, C. P. (2000) Proteins of the endoplasmic-reticulum-associated degradation pathway: Domain detection and function prediction, *Biochem J.* **351,** 527–535.

Ponting, C. P., and Pallen, M. J. (1999a) A beta-propeller domain within TolB, *Mol. Microbiol.* **31,** 739–740.

Ponting, C. P., and Pallen, M. J. (1999b) β-propeller repeats and a PDZ domain in the tricorn protease: Predicted self-compartmentalisation and C-terminal polypeptide-binding strategies of substrate selection, *FEMS Microbiol. Lett.* **179,** 447–451.

Ponting, C. P., Aravind, L., Schultz, J., Bork, P., and Koonin, E. V. (1999) Eukaryotic signalling domain homologues in archaea and bacteria: Ancient ancestry and horizontal gene transfer, *J. Mol. Biol.* **289,** 729–745.

Ponting, C. P., and Birney, E. (2000) Identification of domains from protein sequences, *in* Webster, D. (Ed.), Protein Structure Prediction, Humana Press, Clifton, NJ.

Ponting, C. P., and Russell, R. B. (2000) Identification of distant homologues of fibroblast growth factors suggests a common ancestor for all β-trefoil proteins, *J. Mol. Biol.* **302,** 1041–1047.

Preker, P. J., and Keller, W. (1998) The HAT helix, a repetitive motif implicated in RNA processing, *Trends Biochem. Sci.* **23,** 15–16.

Price, S. R., Evans, P. R., and Nagai, K. (1998) Crystal structure of the spliceosomal U2B"-U2A′ protein complex bound to a fragment of U2 small nuclear RNA, *Nature* **394,** 645–650.

Ray, M.-C., Germon, P., Vianney, A., Portalier, R., and Lazzaroni, J. C. (2000) Identification by genetic suppression of *Escherichia coli* TolB residues important for TolB-Pal interaction, *J. Bacteriol.* **182,** 821–824.

Renault, L., Nassar, N., Vetter, I., Becker, J., Klebe, C., Roth, M., and Wittinghofer, A. (1998) The 1.7 A crystal structure of the regulator of the chromosome condensation (RCC1) reveals a seven-bladed propeller, *Nature* **392,** 97–101.

Riggleman, B., Wieschaus, E., and Schedl, P. (1989) Molecular analysis of the armadillo locus: Uniformly distributed transcripts and a protein with novel internal repeats are associated with a Drosophila segment polarity gene, *Genes Dev.* **3,** 96–113.

Russell, R. B., Sasieni, P. D., and Sternberg, M. J. E. (1998) Supersites within superfolds. Binding site similarity in the absence of homology, *J. Mol. Biol.* **282,** 903–918.

Saupe, S., Turcq, B., and Begueret, J. (1995) A gene responsible for vegetative incompatibility in the fungus *Podospora anserina* encodes a protein with a GTP-binding motif and G beta homologous domain, *Gene* **162,** 135–139.

Scheufler, C., Brinker, A., Bourenkov, G., Pegoraro, S., Moroder, L., Bartunik, H., Hartl, F. U., and Moarefi, I. (2000) Structure of TPR domain-peptide complexes: Critical elements in the assembly of the Hsp70-Hsp90 multichaperone machine, *Cell* **101,** 199–210.

Sedgwick, S. G., and Smerdon, S. J. (1999) The ankyrin repeat: A diversity of interactions on a common structural framework, *Trends Biochem. Sci.* **24,** 311–316.

Schuler, G. D., Altschul, S. F., and Lipman, D. J. (1991) A workbench for multiple alignment construction and analysis, *Proteins* **9,** 180–190.

Schultz, J., Doerks, T., Ponting, C. P., Copley, R. R., and Bork, P. (2000) More than 1000 putative novel human signalling proteins revealed by EST data mining, *Nature Genet.* **25,** 201–204.

Schultz, J., Milpetz, F., Bork, P., and Ponting, C. P. (1998) SMART, a simple modular architecture research tool: Identification of signaling domains, *Proc. Natl. Acad. Sci. USA* **95,** 5857–5864.

Shen, Y., Romo, G. M., Dong, J., Schade A., McIntire, L. V., Kenny, D., Whisstock, J. C., Berndt, M. C., López, J. A., and Andrews, R. K. (2000) Requirement of leucine-rich repeats of glycoprotein (GP) Ib for shear-dependent and static binding of von Willebrand factor to the platelet membrane GP Ib-IX-V complex, *Blood* **95,** 903–910.

Sikorski, R. S., Boguski, M. S., Goebl, M., and Hieter, P. (1990) A repeating amino acid motif in CDC23 defines a family of proteins and a new relationship among genes required for mitosis and RNA synthesis, *Cell* **60,** 307–317.

Sondek, J., Bohm, A., Lambright, D. G., Hamm, H. E., and Sigler, P. B. (1996) Crystal structure of a G_A protein $\beta\gamma$ dimer at 2.1Å resolution, *Nature* **379,** 369–374.

Sonnhammer, E. L., and Durbin, R. (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis, *Gene* **167,** GC1–10.

Speicher, D. W., and Marchesi, V. T. (1984) Erythrocyte spectrin is comprised of many homologous triple helical segments, *Nature* **311,** 177–180.

Springer, T. A. (1997) Folding of the N-terminal, ligand-binding region of integrin alpha-subunits into a beta-propeller domain, *Proc. Natl. Acad. Sci. USA* **94,** 65–72.

Springer, T. A. (1998) An extracellular beta-propeller module predicted in lipoprotein and scavenger receptors, tyrosine kinases, epidermal growth factor precursor, and extracellular matrix components, *J. Mol. Biol.* **283,** 837–862.

Svensson, L., Narlid, I., and Oldberg, A. (2000) Fibromodulin and lumican bind to the same region on collagen type I fibrils, *FEBS Lett.* **470,** 178–182.

Tsai, C. J., Lin, S. L, Wolfson, H. J., and Nussinov, R. (1997) Studies of protein-protein interfaces: A statistical analysis of the hydrophobic effect, *Protein Sci.* **6,** 53–64.

Tsuda, S., Ito, A., and Matsushima, N. (1997) A hairpin-loop conformation in tandem repeat sequence of the ice nucleation protein revealed by NMR spectroscopy, *FEBS Lett.* **409,** 227–231.

Van Der Biezen, E. A., and Jones, J. D. G. (1998) Plant disease-resistance proteins and the gene-for-gene concept, *Trends Biochem. Sci.* **23,** 454–456.

van Raaij, M. J., Mitraki, A., Lavigne, G., and Cusack, S. (1999) A triple beta-spiral in the adenovirus fibre shaft reveals a new structural motif for a fibrous protein, *Nature* **401,** 935–938.

Vasu, S. K., Kedersha, N. L., and Rome, L. H. (1993) cDNA cloning and disruption of the major vault protein alpha gene (mvpA) in Dictyostelium discoideum, *J. Biol. Chem.* **268,** 15356–153560.

Walker, E. H., Perisic, O., Ried, C., Stephens, L., and Williams, R. L. (1999) Structural insights into phosphoinositide 3-kinase catalysis and signalling, *Nature* **402,** 313–320.

Watari, J., Takata, Y., Ogawa, M., Sahara, H., Koshino, S., Onnela, M. L., Airaksinen, U., Jaatinen, R., Penttila, M., and Keranen, S. (1994) Molecular cloning and analysis of the yeast flocculation gene FLO1, *Yeast* **10,** 211–225.

Weng, X., Luecke, H., Song, I. S., Kang, D. S., Kim, S. H., and Huber, R. (1993) Crystal structure of human annexin I at 2.5 A resolution, *Protein Sci.* **2,** 448–458.

Wharton, K. A., Johansen, K. M., Xu, T., and Artavanis-Tsakonas, S. (1985) Nucleotide sequence from the neurogenic locus notch implies a gene product that shares homology with proteins containing EGF-like repeats, *Cell* **43,** 567–581.

Wheelan, S., Marchler-Bauer, A., and Bryant, S. H. (2000) Domain size distribution can predict domain boundaries, *Bioinformatics* **16,** 613–619.

Wiche, G., Becker, B., Luber, K., Weitzer, G., Castanon, M. J., Hauptmann, R, Stratowa, C., and Stewart, M. (1991) Cloning and sequencing of rat plectin indicates a 466-kD polypeptide chain with a three-domain structure based on a central alpha-helical coiled coil, *J. Cell. Biol.* **114,** 83–99.

Worrall, D., Elias, L., Ashford, D., Smallwood, M., Sidebottom, C., Lillford, P., Telford, J., Holt, C., and Bowles, D. (1998) A carrot leucine-rich-repeat protein that inhibits ice recrystallization, *Science* **282,** 115–117.

Wren, B. W. (1991) A family of clostridial and streptococcal ligand-binding proteins with conserved C-terminal repeat sequences, *Mol. Microbiol.* **5,** 797–803.

Zhang, K., Smouse, D., and Perrimon, N. (1991) The crooked neck gene of *Drosophila* contains a motif found in a family of yeast cycle genes, *Genes Dev.* **5,** 1080–1091.