doi:10.1006/jmbi.2000.3684 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2000) 298, 521–537

JMB

Homology-based Method for Identification of Protein Repeats Using Statistical Significance Estimates

Miguel A. Andrade[1,2], Chris P. Ponting[3], Toby J. Gibson[1] and Peer Bork[1,2]*

[1]*European Molecular Biology Laboratory, Meyerhofstr. 1 Heidelberg 69012, Germany*

[2]*Max Delbrück Center for Molecular Medicine Department of Bioinformatics PO Box 740238 D-13092, Berlin-Buch Germany*

[3]*MRC Functional Genetics Unit, University of Oxford Department of Human Anatomy and Genetics, South Parks Road, Oxford, OX1 3QX UK*

**Corresponding author*

Short protein repeats, frequently with a length between 20 and 40 residues, represent a significant fraction of known proteins. Many repeats appear to possess high amino acid substitution rates and thus recognition of repeat homologues is highly problematic. Even if the presence of a certain repeat family is known, the exact locations and the number of repetitive units often cannot be determined using current methods. We have devised an iterative algorithm based on optimal and sub-optimal score distributions from profile analysis that estimates the significance of all repeats that are detected in a single sequence. This procedure allows the identification of homologues at alignment scores lower than the highest optimal alignment score for non-homologous sequences. The method has been used to investigate the occurrence of eleven families of repeats in *Saccharomyces cerevisiae, Caenorhabditis elegans* and *Homo sapiens* accounting for 1055, 2205 and 2320 repeats, respectively. For these examples, the method is both more sensitive and more selective than conventional homology search procedures. The method allowed the detection in the SwissProt database of more than 2000 previously unrecognised repeats belonging to the 11 families. In addition, the method was used to merge several repeat families that previously were supposed to be distinct, indicating common phylogenetic origins for these families.

Keywords: protein repeats; homology; sub-optimal alignment; extreme value distribution; sequence analysis

Introduction

Prediction of homologous proteins by sequence analysis greatly aids the experimental determination of molecular structure and function. However, many protein molecules possess more than one compact structural and functional unit (''domain'', also called ''module'') (Baron *et al.*, 1991; Bork, 1992) that have been independently propagated during evolution (Doolittle, 1989; Heringa & Taylor, 1997). Consequently, homologue detection necessitates detailed consideration of the domain architectures of proteins.

Although it is not unusual for domains to be repeated within a single polypeptide, a distinction is made here between autonomous domains, that may be found as single copies in proteins, and repeats that are invariably found as two or more copies in proteins.

Many families of repeats have been identified first by sequence analysis and subsequently shown by structure determination to represent repeated secondary structural elements. These are usually arranged in close-packing arrangements either as an ''open'' structure with repeats forming an elongated super-helix, or else as a ''closed'' structure with repeats arranged radially about a common axis and with associations between N and C-terminal repeats. Some collections of repeats arranged as closed structures, such as β-propellers (Murzin, 1992), may be formally designated as domains since they are compact structural and

Abbreviations used: ANK, ankyrin; ARM, armadillo; CE, *Caenorhabditis elegans*; EVD, extreme value distribution; HAT, half-a-TPR; HS, *Homo sapiens*; HSP, high-scoring segment pairs; HMM, hidden Markov model; LRR, leucine-rich repeat; PFTA, protein farnesyl transferase α sub-unit; PFTB, protein farnesyl transferase β sub-unit; RCC1, regulator of chromosome condensation 1; REP, REPeat finding method; SC, *Saccharomyces cerevisiae*; SW, SwissProt protein database; TPR, tetratrico peptide repeat.

E-mail address of the corresponding author: bork@embl-heidelberg.de

0022-2836/00/030521–17 $35.00/0

functional units that appear to have been propagated *in toto* by gene duplication.

Detection of such repeats in sequence databases differs from detection of domains in three important respects. Firstly, most repeats are considerably shorter in length than domains and are often highly divergent in sequence. This has the consequence that database searches for repeats usually identify a lesser percentage of true homologues than do domain searches. Secondly, the numbers of repeats in individual proteins can be extremely variable. This is true even for some repeats that form β-propeller closed structures that might otherwise have been thought to possess a constant number of repeats (Neer *et al.*, 1994; Saupe *et al.*, 1995). Consequently, attempts to improve detection of repeats using multiple alignments of N tandem repeats will not detect all repeats for those proteins that have numbers of repeats that are not integer multiples of N. Finally, defining the first and last residues of a repeat is more contentious than for a domain, since repeats are more prone to circular permutation than are domains, particularly within closed structures (Russell & Ponting, 1998), and are also prone to partial truncation resulting in non-integer repeat numbers.

The tandem arrangements of repeats impose constraints on amino acid residue conservation that are characteristic of the repeat family. These characteristics, once detected, may be used to identify additional repeats in the sequence. For example, a positively and a negatively charged residue at defined positions of HEAT repeats (Andrade & Bork, 1995) have been shown recently to form ladder of hydrogen bonds between repeats in the crystal structure (Groves *et al.*, 1999).

These constraints are not well reflected in the existing methods of *de novo* repeat detection in protein sequences, such as dot-plot and other similarity-based methods (Heringa & Argos, 1993; Heringa, 1994) and Fourier analysis techniques (McLachlan, 1977, 1978; Pasquier *et al.*, 1990). Furthermore, current methods of repeat identification do not use a robust probabilistic model to determine the significance of an individual repeat's alignment. To overcome these limitations, we have developed an iterative, homology-based REPeat finding method (REP) that complements first order approaches for repeat identification. If a repeat has already been identified with approximate borders, the method can detect new repeat units based on the probabilities of finding matches of different suboptimal alignments when compared to random sequences (see Methods). This is possible because the scores of non-overlapping sub-optimal, as well as optimal, local alignments taken from a search with the profile on a randomised database, are found to be commonly described by extreme value distributions (EVDs). Consequently, from the scores of optimal and non-overlapping sub-optimal alignments (x_i, for $i = 1, 2, \ldots, n$) we have been able to iteratively estimate P-values (P_i, for $i = 1, 2, \ldots, n$) that represent, for the ith highest scoring local alignment, the probability of finding a hit scoring at least x_i in a random sequence database. Repeat units are identified on the basis of these P_i-values instead of the original scores x_i. This allows an iterative approach to the detection of repeats since newly identified repeats can be included in an alignment that is used for a subsequent search.

This approach is demonstrated by the identification of numerous previously unrecognised representatives of 11 repeat families that are widely represented among eukaryotes.

Results

A new repeat-detection method (REP) has been applied to 11 families of repeats that have representatives in three divergent eukaryotic species (*Saccharomyces cerevisiae* (SC), *Caenorhabditis elegans* (CE), *Homo sapiens* (HS)) and hence are expected to be widespread among all eukaryotic organisms. Fortunately, the three-dimensional structure is known for at least one family member for all but one of these families. This allows detailed comparisons to be made between sequence alignments and three dimensional structures in order to ascertain whether conserved positions reflect preservation of intra-, or inter-, repeat interactions, or other structural or functional constraints.

The selected repeat families vary in length and contain a variety of secondary structure arrangements: a summary of the repeats' properties is given in Table 1. Differences in these properties are not reflected in the method's results (see below).

From known protein structures containing those repeats it can be seen that the variety of folds is restricted to two major types of super-structure (Figure 1): (i) open structures formed by the assembly of repeats composed of either two anti-parallel helices (Armadillo, HEAT, ankyrin, TPR and PFTA repeats; reviewed by Groves & Bartford (1999)) or else one helix and one β-strand (LRR); and (ii) propellers of six or seven secondary structural elements formed either by pairs of antiparallel helices (PFTB) or by ''blades'' of four antiparallel β-strands (RCC1, kelch and WD40 repeats; see Murzin, 1992). The three-dimensional structure of HAT repeats is currently unknown.

Fitting of extreme value distributions to non-overlapping sub-optimal alignment scores

Optimal and sub-optimal non-overlapping alignment scores were obtained from the comparison of profiles, calculated from these repeats' alignments, with an artificial sequence database (see Methods). The distributions of these scores were found to be well-described by EVDs (Figure 2). This finding enabled the estimation of the probability of obtaining a similarity score S larger than x_i by chance, $P(S > x_i)$ ($P(x_i)$-value), which is specific for the ith-highest scoring alignment. In a database search, the number of unrelated hits in repeat order i expected to score above x may be estimated as the

Table 1. Properties of repeat types studied in this work

Repeat	Ref1	Length	2D	3D PDB	3D Frag.	3D SW	3D Ref2
ANK	Lux *et al.* (1990)	30	α/β	1awc_B	5-157	GABB_MOUSE	Batchelor *et al.* (1998)
ARM	Peifer *et al.* (1994)	40	α	1bk5	46-530	IMA1_YEAST	Conti *et al.* (1990)
HAT	Preker & Keller (1998)	33	α?	none			
HEAT	Andrade & Bork (1995)	38	α	1b3u		2AAA_HUMAN	Groves *et al.* (1999)
KELCH	Bork & Doolittle (1994)	47	β	1gof		GAOA_DACDE	Ito *et al.* (1991)
LRR	Kajava (1998)	23	α/β	1dfj_I		RINI_PIG	Kobe & Deisenhofer (1995)
PFTA	Boguski *et al.* (1992)	34	α	1ft2_A		PFTA_RAT	Long *et al.* (1998)
PFTB	Boguski *et al.* (1992)	42	α	1ft2_B		PFTB_RAT	Long *et al.* (1998)
RCC1	Ohtsubo *et al.* (1987)	51	β	1a12		RCC_HUMAN	Renault *et al.* (1998)
TPR	Zhang *et al.* (1991)	34	α	1a17	19-177	PPP5_HUMAN	Das *et al.* (1998)
WD40	Neer *et al.* (1994)	40	β	1gp2_B		GBB1_HUMAN	Wall *et al.* (1995)

ANK, ankyrin; ARM, armadillo, HAT, HEAT, KELCH, LRR, leucine-rich repeats; PFTB, protein farnesyl transferase α-subunit repeats; PFTB, protein farnesyl transferase β-subunit repeats; RCC1, TPR, tetratricopeptide repeats; ref1, is a reference to the discovery or a review of the corresponding repeat family; length, is the length of the profile used (this excludes positions in alignments containing more than 66% gaps); 2D, secondary structure content; PDB, pointer (PDB database code) to a representative 3D structure containing the repeat (if necessary, the chain is indicated following the PDB code). These structures are shown in Figure 1; frag, a fragment of the original protein was used; SW, SwissProt code of the corresponding protein sequence; ref2, reference to the structure determination.

product of the database size used for the search (in sequences) and the $P_j(x_i)$ value (Pearson, 1998). For the selection of true positive repeats, a single *P*-value threshold P_0 per repeat family can be applied to all repeats, irrespective of their order. Furthermore, to first approximation the total number of false positive repeats expected in a database search may be estimated as the triple product of P_0 with the database size and the average number of repeats found in true positive sequences.

Single *P*-value thresholds P_0 were assigned for each repeat family that discriminated between true positives and the top scoring false positive sequence. True positive sequences were assigned using exhaustive PSI-BLAST analysis (Altschul *et al.*, 1997). Values of P_0 represent P_i values that are bettered by at least one repeat in a given sequence (see Methods for a complete description of the application of repeat thresholds). Thus they are typically larger than $1/N$ where N is the size of a typical database ($\approx 10^4$-10^5).

Analysis of SwissProt

In order to calibrate our method, and to compare it with existing sets of identified repeats, the SwissProt database (Bairoch & Apweiler, 1999) was searched for all detectable members of the 11 repeat families. The SwissProt release used (37.0 Dec 1998) contains 77,977 protein sequences. Many of the repeats that are defined in this database have been annotated using literature sources. However, other repeats have been defined using the protein family database Pfam (Bateman *et al.*, 1999).

Pfam is a collection of protein domains and motif families each represented by a multiple alignment. These families may be searched using a suite of algorithms employing hidden Markov models (HMMs) (HMMER, Eddy, S., unpublished data). The latest version of this, HMMER2, derives a single *E*-value for each sequence from a single calibrated EVD. This EVD is a fit to the distribution of optimal scores calculated from alignments between an HMM and a randomised sequence database. This approach is successful in annotating numerous repeats in SwissProt that have previously escaped attention. Consequently, it is considered by us to be the most sensitive approach currently in use for repeat identification given an initial multiple alignment.

Here we detect representatives from each of the 11 repeat families using REP, and compare these with the results available from Pfam (using the links provided by SwissProt 37.0 from sequences to Pfam domains). In most of the 11 cases, the majority of detected repeats are found by both methods, but some repeats are found by only one method. In order to distinguish true positive homologues from false positives, doubtful sequences were subjected to reciprocal searches (Bork & Gibson, 1996) using the position-specific, iterative version of BLAST (PSI-BLAST (Altschul *et al.*, 1997)) and an *E* value inclusion threshold of 0.001. PSI-BLAST is currently a method of choice in detecting homologues, although studies have shown its generated alignments to be typically sub-optimal (unpublished results).

Ankyrin repeats

Ankyrin repeats (Lux *et al.*, 1990) were first detected in human erythrocyte ankyrin. They were found subsequently in a large number of protein families (reviewed by Bork, 1993), including GA-binding Protein β subunit (a transcription factor) (see its structure in Figure 1), *Drosophila melanogaster* Cactus and Notch, human p53, several nuclear factors, 2-5α-dependent RNase and myotrophin. The structure of each independent repeat contains

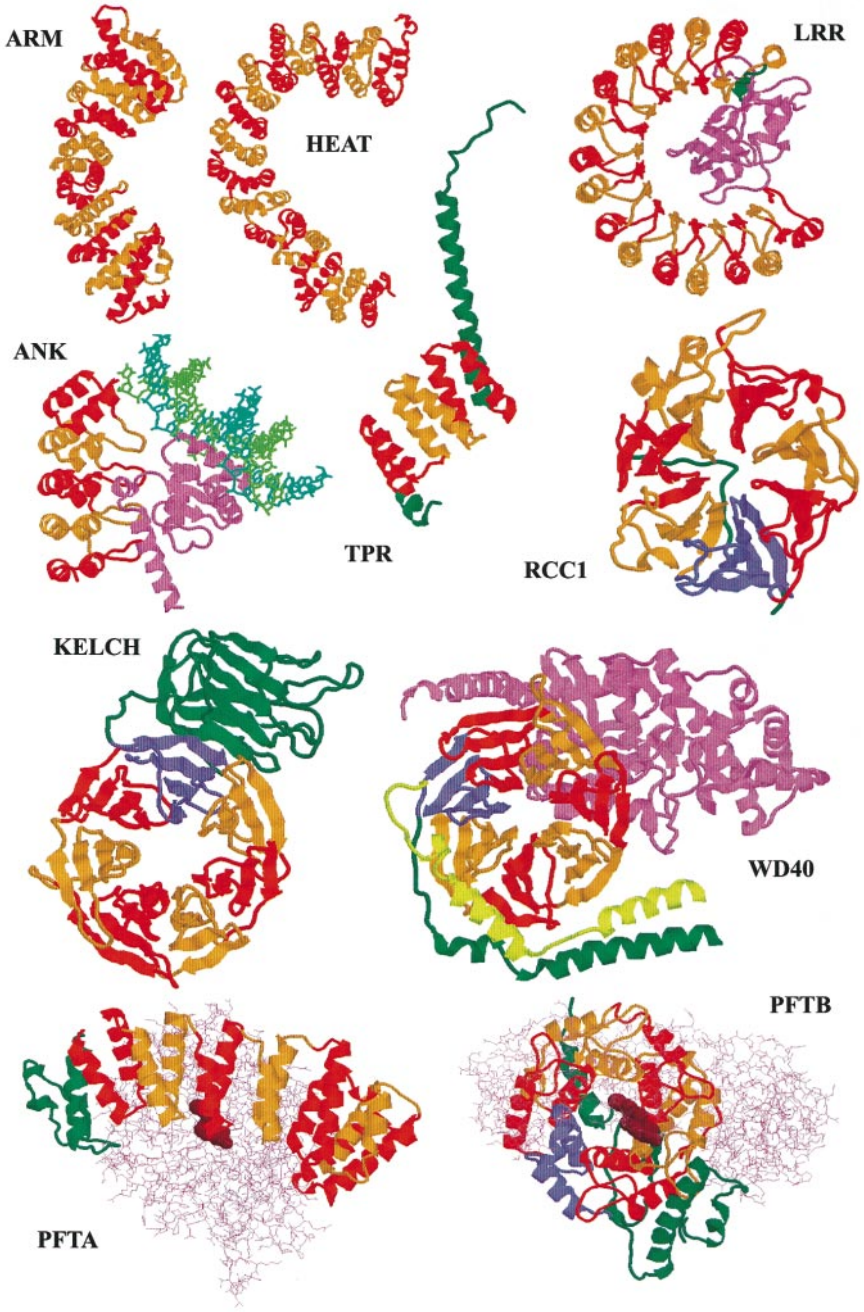

Figure 1 (*legend opposite*)

two antiparallel α-helices followed by a loop which contains a type I β turn. The repeats assemble into a rod of highly packed helices (Figure 1).

REP and Pfam identified in common 107 ankyrin repeat-containing proteins in SwissProt. In addition, Pfam uniquely detects seven proteins with nine repeats and REP three proteins with 13 repeats (Table 2). All three proteins not identified by Pfam, have either already been reported in the literature to contain ANK repeats (e.g. TRP-1 proteins (Phillips *et al.*, 1992)) or else were manually annotated in SwissProt entries (e.g. SW:CDN5_HUMAN, cyclin-dependent kinase 4 inhibitor B, a tumor suppressor, for which two repeats were found in addition to two that were annotated previously).

Armadillo repeats

Armadillo repeats (Peifer *et al.*, 1994) are eukaryotic-specific repeats implicated in protein-protein interaction. First identified in the *D. melanogaster*

Table 2. Comparison of the results on the analysis of the set of repeats in SwissProt as with the Pfam analysis results as presented in release 37.0 (field 'DR', with corresponding Pfam identifier and number of hits detected)

		PFAM	Only	Only	Rep/Prot		REP thr	
Repeat	PFAM id	+REP	REP	PFAM	REP	PFAM	P_θ	n_{min}
ANK	PF00023	107	3	7	5.3	4.3	1e-5	1
ARM	PF00514	34	0	8	7.7	6.9	1e-8	3
HAT			7		6.3		1e-4	3
HEAT			47		7.4		1e-6	4
KELCH	PF01344	21	10	0	5.9	4.5	1e-4	3
LRR[a]	PF00560	118	2	28	9.0	9.1	1e-4	2
LRRd[b]	PF00560	135	6	11	8.8	9.0	1e-5	1
PFTA[c]	PF01239	8	0	0	5.1		1e-4	4
PFTB	PF00432	25	2	0	4.5	4.1	1e-5	2
RCC1	PF00415	11	0	0	6.3	5.2	1e-5	3
TPR	PF00515	50	70	4	6.6	3.0	1e-4	3
WD40	PF00400	253	55	0	6.5	4.4	1e-4	3

Pfam + REP, Pfam and our method agree in assigning repeat(s) to a protein sequence; only REP, our method assigns repeats to a protein sequence and Pfam does not; only Pfam, Pfam assigns repeats to a protein and our method does not; rep/prot how many repeats per protein are assigned by each method (in the common set of assigned proteins); Pfam id indicates the Pfam identifier; REP thr columns indicate the thresholds used by our method: P_θ threshold in P-value to accept a hit as repeat, n_{min} minimum number of repeats required to validate the assignment. Note that HAT and HEAT repeats could not be used for this comparison because their profiles were not yet included in the Pfam database at the moment of the release of SwissProt 37.0 (December 1998).

Total number of repeats detected by each method is 5,060 and 3,497 for REP and Pfam, respectively, for those cases where the comparison is possible, namely: ANK, ARM, KELCH, LRRd, PFTB, RCC1, TPR, and WD40. Details of this analysis can be accessed through the internet at the web address http://www.embl-heidelberg.de/ ~ andrade/papers/rep.

[a] The Pfam profile of LRRs, PF00560 consists of two LRRs. Accordingly, the number in the table is the number of hits multiplied by two.

[b] LRRd is a profile containing two consecutive LRRs. The results obtained with this profile are shown for comparison.

[c] The Pfam profile for PFTA corresponds to a domain rather than to a single repeat.

Armadillo protein (Riggleman *et al.*, 1989), similar repeats were later found in importin α-subunit, plakoglobin, vacuolar yeast protein 8, β-catenin and Rap1 (GTPase-GDP dissociation stimulator 1).

Each armadillo repeat (ARM) contains two antiparallel helices, with the first often kinked due to the presence of a proline residue, or a series of small amino acid residues. The repeats pack in a super-helical assembly (Conti *et al.*, 1998). One example of an ARM repeat-containing protein is found in importin, a hetero-dimeric protein complex which takes proteins into the nucleus probably by recognition of nuclear localisation signals (NLSs) (Görlich *et al.*, 1994). The importin α-sub-

Figure 1. 3D structures of proteins and macromolecular complexes containing repeats relevant to this work. Series of repeats are coloured in alternating red and orange, with the remainder of the protein in olive green and other protein chains in violet. ANK, four and a half ankyrin repeats in a fragment of the β-subunit of the GA-binding protein (Batchelor *et al.*, 1998) complexed with a fragment of the α-subunit, and 21 bp of DNA (green). Note that the long inter-repeat loops mediate the interaction between the α-subunit and the DNA. ARM, fragment of the importin α-subunit containing ten armadillo repeats (Conti *et al.*, 1998). Amino acid analysis shows high residue conservation in a groove in the concave region of the arch indicating its possible role in protein recognition. Kelch, the structure of galactose oxidase from *Dactylium dendroides* (Ito *et al.*, 1991) showing only the N-terminal and the β-propeller domains (positions 1-532) with the latter formed by seven kelch repeats. The void within the propeller domain is filled in the crystal structure by two β-strands from the C-terminal domain and water (not shown). Note that the external β-strand of the last blade (purple) is not consecutive in sequence with the internal β-strands due to the circular permutation of sequence repeats relative to the structural repeats. This closing ''clasp'' mechanism has also been observed for WD40 repeats and has been proposed to contribute to the propeller's structural stability. LRR, structure of porcine ribonuclease inhibitor (Kobe & Deisenhofer, 1995). The inhibitor surrounds the ribonuclease in a horseshoe-like structure that is composed entirely of LRRs. TPR, structure of TPRs in protein phosphatase 5 (Das *et al.*, 1998). RCC1, Structure of the regulator of chromosome condensation (Renault *et al.*, 1998), a seven-bladed propeller. Note that although there is some structural resemblance to the WD40 and kelch structures, the twist angles between the blades' β-sheets are larger. PFTA/PFTB, structure of the heterodimeric protein farnesyltransferase (Long *et al.*, 1998). The complex has been represented in two pictures that highlight the α and β-sub units. Left, the α-subunit contains five PFTA repeats (the β-subunit, in the back, is represented by thin sticks). Right, the β-subunit contains a barrel containing six PFTB repeats (the α-subunit, in the front, is represented by thin sticks). The substrate (farnesyl diphosphate, in brown) is located between the two subunits. WD40, structure of the G-protein heterotrimer (Wall *et al.*, 1995). α-subunit in violet and γ-subunit in yellow. The β-subunit contains a barrel formed by seven repeats that are similar in structure and in the clasp closing mechanism to kelch repeats of galactose oxidase. In comparison to kelch repeats, the WD40 blades are arranged more compactly resulting in a smaller central void.

unit contains a rod formed by ARM repeats (see the structure in Figure 1) (Conti *et al*., 1998). A similar structure is observed in β-catenin (Huber *et al*., 1997).

Identification of ARM repeat containing proteins in SwissProt using REP resulted in a subset of those found by Pfam. However, the eight proteins that are predicted by Pfam analysis to contain ARM repeats include six that are importin β-subunits, which were identified by REP as HEAT repeats (Andrade & Bork, 1995). This supports the similarity between ARM and HEAT repeats, which is also observable at the structural level (see Figure 1 and see the Discussion). The remaining two true positive cases detected by Pfam but not by REP, were mammalian general vesicular transport factors P115.

HEAT repeats

HEAT repeats (Andrade & Bork, 1995) are present in eukaryotic proteins including human Huntingtin, elongation factor III, protein phosphatase PP2A α-subunit and the TOR/FRAP family of phosphatidyl inositol kinases. As discussed above, they resemble ARM repeats in sequence and in structure. They are distinguishable, however, from ARM repeats by: (a) the presence of two charged amino acid residues per repeat that has been observed to compose a ladder of electrostatic interactions within the repeat super-structure (Groves *et al*., 1999); and (b) the lack of a glycine residue, that is conserved in helix 1 of ARM repeats. The latter difference has the effect that helix 1 of ARM repeats is often more kinked than the corresponding helix of HEAT repeats, causing disruption of the ARM repeat helix into two parts. Recently-determined tertiary structures (Cingolani *et al*., 1999; Chook & Blobel, 1999; Vetter *et al*., 1999; Kobe *et al*., 1999) demonstrate their differences in structure when compared with ARM repeats, in particular, their lower rotation angles of adjacent repeats (15° in HEAT *versus* 30° in ARM). These structural differences necessitated the detection of HEAT and ARM repeats using two independent searches with high discriminatory thresholds ($P_0 = 10^{-6}$ and 10^{-8}). Nevertheless, as might be expected application of this high threshold was insufficient to completely discriminate between these two related repeat families.

REP identified 47 proteins containing 348 HEAT repeats. Many of these had already been described as containing HEAT repeats, but a number of new findings were revealed. For example, five HEAT repeats, in two clusters, were predicted in yeast Mot1p, a member of the SNF2/SWI2 family of ATPases (Davis *et al*., 1992). Corroboration of three of these repeats was provided by PSI-BLAST analysis of the N-terminal group of Mot1p HEATs (amino acid residues 280-590) that detected significant similarity with previously identified HEAT-containing proteins, including β-importins, within six iterations and using a threshold of $E < 0.001$. Previously predicted TPRs within this region (Davis *et al*., 1992) could not be verified using a variety of methods (results not shown). A sixth HEAT repeat predicted by REP (amino acid residues 1495-1537) is a likely false positive, given that this lies within the ATPase homology domain. The predicted HEAT repeats in Mot1p are contained within the region known to bind the TATA-binding protein (Auble *et al*., 1997).

S. cerevisiae and *Schizosaccharomyces pombe* orthologues (YMR288w and C27F1.09C, respectively) of human SAP155 (Wang *et al*., 1998) were identified by REP as containing HEATs. Indeed, SAP155 has been noted as containing a repeated structure similar to that of the regulatory subunit A of protein phosphatase PP2A (Wang *et al*., 1998) and these, in turn, are known (Andrade & Bork, 1995) to represent HEATs. REP also identified nine HEATs in the colonic and hepatic tumor over-expressed protein (CH-TOG) (Charrasse *et al*., 1995). This is a member of a family of microtubule-associated components of the meiotic and mitotic spindle poles (Matthews *et al*., 1998; Wang & Huffaker, 1997). PSI-BLAST analysis of the *C. elegans* CH-TOG orthologue, ZYG-9, confirmed the significance of sequence similarities between this family and HEAT repeats (data not shown).

HEAT repeats were also detected in *S. pombe* hypothetical protein C31A2.05c (SW:YA45_SCHPO). This appears to be an orthologue of *Coprinus cinereus* Rad9 (Seitz *et al*., 1996), *S. cerevisiae* YDR18Ow, human IDN3, and *Drosophila* Nipped-B (Rollins *et al*., 1999). This family of proteins is predicted to possess roles in sister chromatid cohesion, chromosome condensation and DNA repair (Seitz *et al*., 1996; Rollins *et al*., 1999).

TPR repeats

Tetratrico peptide repeats (TPRs) (Zhang *et al*., 1991) are widespread among organisms drawn from the three kingdoms of cellular life and occur in proteins possessing a wide variety of functions. A single TPR contains two antiparallel α-helices which pack into an open structure (Das *et al*., 1998). REP reports a total of 120 proteins with TPR repeats. Pfam predicts four proteins containing a total of eight repeats that were unable to be identified by REP. Of the 70 TPR-containing proteins with a total of 330 individual repeats detected by REP, but not by Pfam, some such as kinesin light-chains (Ginhart & Goldstein, 1996) and SNAP secretory proteins (Ordway *et al*., 1994) were described to contain these repeats. REP also detected TPRs in the tandemly repeated superhelix of clathrin heavy chains. The resemblance of the clathrin superhelix to TPRs had been previously noted (Ybe *et al*., 1999), but these repeats had been suggested to be shorter and to contain fewer periodically spaced hydrophobic residues than TPRs. However, comparison of an HMM calculated from a multiple alignment of clathrin heavy chain sequences with current sequence databases

using HMMER2 revealed significant similarities ($10^{-3} < E < 10^{-15}$) with sequences that are annotated as containing TPRs. This suggests that the clathrin heavy chain is composed of divergent yet bone fide TPRs.

REP detected TPRs in other proteins that, to our knowledge, had not been suggested as containing these repeats. Of these, a 72 kDa signal recognition particle protein (Lutcke *et al*., 1993) that is cleaved during apoptosis (Utz *et al*., 1998) (SW:SR72_CANFA), several *B. subtilis* aspartyl-phosphate phosphatases that function in regulating sporulation (Perego & Hoch, 1996) (e.g. SW:RAPE_BACSU), a histone-binding protein (Kleinschmidt *et al*., 1986) (SW:HIBN_XENLA), and a gene required for cytochrome *c* maturation in *E. coli* (Thony-Meyer *et al*., 1995) (SW:CCMH_ECOLI) are perhaps the best experimentally characterised.

TPRs were also predicted in the RRP5 family of rRNA biogenesis proteins. As discussed below, these are likely to represent divergent TPRs that have been previously reported as HAT repeats.

HAT repeats

HAT (''half-a-TPR'') helices were recently identified in several eukaryotic proteins involved in RNA metabolism (Preker & Keller, 1998). These repeats are predicted to contain two α-helices that are tightly packed to form large super-helical structures. Preker & Keller (1998) noted sequence similarity between HATs and TPRs, in particular those previously predicted in *D. melanogaster* Crn, the *crooked neck* gene product (Zhang *et al*., 1991). They suggested, however, that HATs are distinct from TPRs in lacking residues that contribute to the TPR ''holes'' (Sikorski *et al*., 1990). Using a previously published alignment of HATs (Preker & Keller, 1998) and thresholds of $P_0 = 10^{-4}$ and $n_{\min} = 3$ (repeat number threshold, see Methods), REP identified the HAT repeats included in the alignment as well as another 11 repeats in *D. melanogaster* Crn that have previously been described as TPRs (Zhang *et al*., 1991). REP also identified four similar repeats in yeast RRP5p (Venema & Tollervey, 1996).

The RRP5p repeats had been described, on the basis of similarity to *D. melanogaster* Crn, as TPRs (Torchet *et al*, 1998). However, the repeats in Crn and RRP5p are not identified by Pfam or SMART (Ponting *et al*., 1999) domain identification systems as such. However, a PSI-BLAST database search using human CstF-77 (Genbank identifier (gi) 632498) as query, identified regions of Crn and human RRP5p as significantly similar to HATs ($E = 4 \times 10^{-12}$ and $E = 1 \times 10^{-5}$, in round 1). Subsequent iterations of the PSI-BLAST search, however, revealed significant similarity ($E < 10^{-8}$) to TPRs in prokaryotic proteins that are predicted by SMART and PFAM (results not shown). As suggested (Preker & Keller, 1998), HAT repeats appear to represent a distinct subfamily within the large and diverse family of TPRs. It would appear that HAT repeats are unusual TPRs in containing insertions and deletions; this distinction is one that they share with the SNAP subfamily of TPRs (Ordway *et al*., 1994).

Identification of HAT repeats in RRP5p is of particular interest given that over-expression of human *RRP5* (also known as *ALG-4*) induces transcription of the Fas ligand, and consequently, apoptosis (Lacana' & D'Adamio, 1999). An in-frame deletion of two amino acid residues within the first HAT repeat in yeast RRPSP has been shown to inhibit the synthesis of 18 S rRNA from its 35 S precursor (Torchet *et al*., 1998). Thus, the effect of *RRP5* over-expression might have a more general effect than simply increasing transcription of the Fas ligand.

Kelch repeats

Kelch repeats are found in eukaryotes and bacteria in protein families such as *D. melanogaster* kelch (ring canal protein), α and β-scruin, calicin and galactose oxidase (Bork & Doolittle, 1994). The repeats are arranged in a β-propeller fold formed by seven blades of four anti-parallel β-strands. The structural similarity of kelch and WD40 repeat-containing folds is readily apparent (see Figure 1) although this is not reflected at the sequence level. These and other examples of repeat-containing β-propeller structures (for example, hemopexin, sialidases and RCC1) suggest that these structures have arrived at a common structure independently.

REP identified ten kelch proteins with 47 repeats that were not annotated in SwissProt; of these, some have been described. For example, REP identified kelch repeats in *S. pombe* Ral2, a guanine nucleotide releasing factor for Ras1 (SW:RAL2_SCHPO), as discussed by Bork & Doolittle (1994), and in *S. pombe* C15A10.10, an orthologue of mouse muskelin (Adams *et al*., 1998).

Three kelch repeats were predicted in yeast Mds3p (SW:MDS3_YEAST), a negative regulator of the sporulation pathway (Benni & Neigeborn, 1997). REP reported two repeats in the Mds3p paralogue Pmd1p (SW: YEW2_YEAST) but these were not automatically reported due a repeat number threshold of $n_{\min} = 3$. However, closer examination of these results indicated the existence of several other kelch repeats in this sequence. These proteins' kelch repeats are located in regions that are distinct from their binding sites for protein kinases. Given the propensity of eukaryotic intracellular kelch repeat-containing proteins to bind actin (Way *et al*., 1995; Hernandez *et al*., 1997; Robinson & Cooley, 1997; Kim *et al*., 1999), it is likely that the Mds3p and Pmd1p kelch repeats possess similar actin-binding functions.

Leucine-rich repeats

Leucine-rich repeats (LRRs) (Kajava *et al*., 1998) each contain a β-strand and an α-helix. LRR-

containing proteins have been detected in bacteria, eukaryotes and viruses, but not, to date, in archaea. Notable examples of these proteins are fungal adenylate cyclases, carboxypeptidase N-regulatory subunit, proteoglycan II (decorin), Ras suppressor protein 1, *D. melanogaster* toll and flightless proteins, and fibromodulin.

Assignment of LRRs has been problematic. The SwissProt entry SW:RINI_PIG, representing a sequence whose structure is known (Kobe & Deisenhofer, 1995), lists mis-assigned structural repeats that begin and end in the middle of α-helices. Our assignment of these repeats corresponded to the repeats apparent from the structure (Kobe & Deisenhofer, 1995). The majority of LRRs were detected by both REP and Pfam (Table 2), although 28 LRR-containing proteins were identified only by Pfam. A likely cause of Pfam's success in this case is that the Pfam LRR alignment used represents two consecutive LRRs (however, we note that the most recent version of the LRR Pfam alignment represents only a single repeat). Although this increases the sensitivity of detecting even numbers of repeats, we chose not to employ this approach as it is unable to detect all repeats in proteins containing odd numbers of repeats. For comparison, we used a profile including two repeats (LRRd in Table 2). The results were more sensitive (with an increase from 120 to 141 proteins identified) keeping selectivity (all new cases were true positives) but, as we expected, the total number of repeats detected per protein decreased (from 9.0 to 8.8) due to missing single repeats.

PFTA repeats

The protein farnesyl transferase (PFT) heterodimeric complex is unusual in containing two families of repeats, one in each subunit (Boguski *et al.*, 1992) (Figure 1). The α-subunit contains an open structure of repeats (PFTA) composed of two anti-parallel α helices arranged in a similar way to Ankyrin repeats or TPRs.

REP and Pfam performed identically in the detection of PFTA repeats. Although not detected above threshold by REP, four putative PFTA repeats in *Methanococcus jannaschii* hypothetical protein MJ1345 (SW:YD45_METJA) were detected with low P values ($10^{-6} < P < 10^{-3}$). These were coincident with nine TPRs predicted by REP ($10^{-16} < P < 10^{-6}$) and by Pfam. In order to investigate whether TPRs and PFTA repeats are structurally, as well as sequentially, similar, their known tertiary structures (PDB codes 1FT2 chain A and 1AL7) were compared using Dali (Holm & Sander, 1993). This resulted in a Z-score of 5.0, which is above the threshold of 2.0 that is considered significant. These results suggest that PFTA repeats are divergent TPRs.

PFTB repeats

The β-subunits of PFT contains repeats that have also been observed in other protein prenyltransferases (Boguski *et al.*, 1992). PFTB repeats are formed by two antiparallel α-helices arranged in a barrel of six repeats (Figure 1). REP and Pfam identified the same PFTB repeats in SwissProt proteins with the exceptions of the close homologues SW:Y4KT_RHISN and SW:YCP7_BRAJA that were only detected by REP (two repeats at 45-86 and 401-443 with P-values of 10^{-8} and 10^{-7}, respectively). These bacterial proteins are of unknown function, but are homologues ($E < 10^{-6}$ in a BLAST search) of plant an fungal diterpene cyclases. These function in the cyclisation of geranyl geranyl pyrophosphate into copalyl pyrophosphate (Kawaide *et al.*, 1997). The substrate and sequence similarities of these enzymes to those of PFT suggest that SW:Y4KT_RHISN and SW:YCP7_BRAJA are PFT homologues with PFTB repeats.

The structure of squalene-hopene synthase has been shown to be similar to that of PFT (Wendt *et al.*, 1998). REP detected seven PFTB repeats in domains 1 and 2 of this enzyme. Each of these repeats contains an external and an internal α-helix with an intervening QW motif (Poralla *et al.*, 1994). The N-terminal repeat predicted by REP contains sequences from both domains 1 and 2 due to the insertion of domain 2 into domain 1. These results indicate that PFTB repeats are divergent representatives of QW motif-containing repeats that are apparent in the squalene-hopene synthase crystal structure.

RCC1 repeats

Regulator of chromosome condensation 1 (RCC1) repeats (Ohtsubo *et al.*, 1987) form a seven-bladed β-barrel, with each blade containing four antiparallel β-strands, as in kelch and WD40 structures. The twist of the blades with respect to the barrel's axis is more pronounced for the RCC1 structure than for these other β-barrel structures (Renault *et al.*, 1998). REP and Pfam performed identically in detecting RCC1 repeat-containing proteins, although on average REP detected more repeats per protein than Pfam. No sequence similarities were detected to other β-barrel structures such as kelch or WD40.

WD40 repeats

Identifying WD40 repeats is problematic since the structural repeat, represented by each of the β-propeller blades in the Gβ transducin structure (Wal *et al.*, 1995; Sondek *et al.*, 1996), is permuted with respect to the sequence repeat by a single β-strand. Consequently, an alignment of WD40 structural repeats is unable to be used to detect the complete set of seven complete repeats in this Gβ transducin structure. Furthermore, an alignment of WD40 sequence repeats would be unable to detect

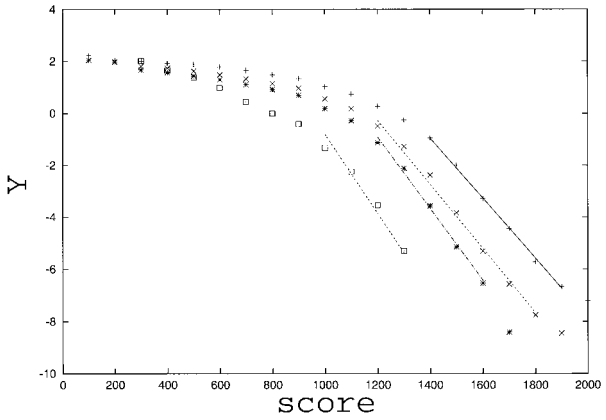

Figure 2. Fitting of sub-optimal non-overlapping alignment scores to EVDs. Symbols correspond to the score distributions of non-overlapping sub-optimal hits of the WD40 repeat profile against a database of 20,000 randomised sequences. Shown from right to left: score distributions of the 3rd, 4th, 5th and 15th repeat orders. The *X*-axis represents scores, whereas the *Y*-axis represents $Y = \log_e(-\log_e(1 - c_i))$ with c_i being the cumulative fraction of hits above scores x for the ith bin. Continuous lines correspond to the fit functions derived by SearchWise from the upper regions of the distributions. For example, the rightmost distribution was fitted between scores 1400 to 1900. The χ^2 values for the fits were 0.0013, 0.03, 0.013 and 0.08.

the complete set of repeats in a WD40 repeat-containing protein for which there is no permutation. Unfortunately, the relative populations of permuted *versus* unpermuted versions of these WD40 β-propeller structures are unknown. This has the result that no single multiple alignment will be able to detect the complete set of WD40 repeats.

We have chosen to construct a multiple alignment that contains the WD40 sequence repeats that are permuted with respect to the structural repeats, in order to be consistent with the only known structures of WD40 repeat-containing proteins (Wall *et al*., 1995; Sondek *et al*., 1996). Consequently, assignments of WD40 repeats in this study differ from those in previous studies (Neer *et al*., 1994; Garcia-Higuera *et al*., 1996) and those shown in the SwissProt database (Bairoch & Apweiler, 1999).

REP identified 308 WD40 repeat-containing proteins, as compared with Pfam's 253 and a total of 188 that were identified in an independent study (Smith *et al*. (1999); http://BMERC-www.bu.edu/wdrepeat; last updated in February 1998). All the repeats identified by REP were validated by PSI-BLAST searches. Many WD40 repeat-containing proteins that were identified only by REP represent hypothetical and experimentally uncharacterised proteins, or else close homologues of known WD40 repeat-containing proteins. Homologues of yeast Prtlp (SW:IF3X_YEAST), however, represent previously unrecognised WD40 repeat-containing proteins that were detected by REP (for example, REP detected repeats in the eukaryotic translation initiation factor 3 β-subunit, as in SW:IF3X_HUMAN from 332-370, 372-417 and 649-694 with *P*-values ranging from $10^{-3}$to 7×10^{-5}). Yeast Prt1p was originally identified by a screen for cell division cycle mutants that affect cell proliferation (Hanic-Joyce *et al*., 1987). Subsequently, it was found to be a subunit of the eukaryotic initiation factor 3 (eIF3) complex that is conserved in yeast (Naranda *et al*., 1994) and in mammals (Asano *et al*., 1997; Methot *et al*, 1997).

There are four strongly predicted ($P < 10^{-4}$) WD40 repeats in yeast Prt1p: amino acid residues 217-257, 260-297, 537-580 and 596-641. The C-terminal pair of Prt1p WD40 repeats are included in a region that is important for interaction with TIF34 and for thermostability (Evans *et al*., 1995; Asano *et al*., 1998). In addition, missense mutations that occur in temperature-sensitive *prt1* mutants (Evans *et al*., 1995) are predicted to occur in WD40-containing regions, after strands a and c (Smith *et al*., 1999). These regions of several β-propeller structures have been noted to participate in ligand-binding (Li *et al*., 1995). It is possible, therefore, that these amino acid substitutions in temperature-sensitive mutants also result in defective complex formation of Prt1p within eIF3 (Asano *et al*., 1998).

Repeat detection performance within the proteins identified

One of the main criteria for the success of the REP procedure is the selectivity and sensitivity for the identification of the repeat-containing proteins (see above). Due to the divergence of repeats, another major issue is the correct identification of the repeats therein. The sixth and seventh column in Table 2 compare the number of repeats detected by REP and Pfam in those proteins that are judged to contain repeats by both methods. This indicates that REP is superior to Pfam in terms of sensitivity (5060 *versus* 3497 identified repeats in these proteins, see Table 2). There is no easy way, however, to compare these methods' levels of selectivity. Most of those additional repeats could be confirmed using reciprocal PSI-BLAST searches and only very few likely false positives have been revealed. This is indicative of a high selectivity.

Furthermore, we identified numerous twilight zone candidates, both in terms of proteins and in terms of repeats contained therein. For example, protein B19 from Vaccinia virus (Smith *et al*., 1991), SW:V19R_VACCV, scored close to the threshold for ankyrin repeats. This was confirmed after PSI-BLAST iteration 3 of the database search revealed significant similarity to the ankyrin repeats of tankyrase. This small protein of only 176 amino acid residues displays two repeats above the thresholds (positions 13-44 and 110-142). Further analysis confirmed these repeats and produced an additional C-terminal repeat (amino acid residues 148 to 176). It appears plausible that this protein is entirely

formed by ankyrin repeats as this has been shown for similar proteins in Orthopoxviruses (e.g. Bork, 1993). The remaining repeats are too divergent to be detected by REP. Such examples indicate that REP uses rather stringent thresholds and that many more repeats of these families remain undiscovered in public databases.

Analysis using Pfam and SMART alignments

Differences in the results of the two methods could be due, at least in part, to the use of different multiple alignments. In order to investigate this we performed searches of SwissProt with REP using the kelch and TPR alignments taken from Pfam, and searches of Swissprot with HMMER2 using kelch and TPR alignments taken from SMART.

When searching for kelch repeats using the Pfam kelch alignment (PF01344) for the profile/HMM, 12 more true positive kelch repeat-containing proteins were identified with REP than were detected using HMMER2. Similarly, when searching for TPRs using the Pfam TPR alignment (PF00515), 45 more TPR-containing proteins were identified only by REP than were identified only by HMMER2. These are slightly fewer than the corresponding numbers (19 and 66) of kelch or TPR-containing proteins detected only by REP, using the SMART alignment, over-and-above the proteins detected only by HMMER2, using the Pfam alignment.

Similarly, querying the SwissProt database using the SMART-derived kelch and TPR alignments resulted in 19 and 3, respectively, more repeat-containing proteins detected only by REP than detected only by HMMER2. The results for HMMER2 searches were compiled using a relatively liberal *E*-value acceptance threshold of 0.01.

These results demonstrate that although improvements can be made to the detection of repeats using more optimal alignments, more repeat-containing proteins were detected using REP than they were using HMMER2. This was found to be the case even when identical multiple alignments were used for the searches.

Genome-wide analysis

After benchmarking by comparing REP with SwissProt annotations and HMMER2 searches, we sought to quantify the spread and the evolution of these repeats using genomic data. Three sets of proteins were chosen: 6218 proteins from *S. cerevisiae* (SC), 19,351 proteins from *C. elegans* (CE), and 11,827 proteins from *H. sapiens* with less than a 97% sequence identity to each other (HS) (Table 3). SC and CE data represent completed genomes, whereas the HS set represents approximately 10-15% of the complete human proteome.

The 11 repeat families studied were found to be contained in 2-3% of the proteins for each of the three genomes. Given the conservative detection thresholds used and given the existence of many more repeat families, this figure represents a considerable underestimate of the fraction of eukaryotic repeat-containing proteins.

Table 3 shows that *C. elegans* contains fewer proteins with these repeats than either *S. cerevisiae* or *H. sapiens*. *H. sapiens* appears to be more enriched in ANK, ARM and LRR repeats, whereas *S. cerevisiae* is enriched in WD40 repeat-containing proteins. The numbers of repeats per protein (*R*) is relatively constant between these three species. As expected given the choice of the 11 repeat families, these repeats occur primarily in eukaryotes, although six families are also observed in bacteria, one in archaea and three in viruses.

A more complete analysis of the distributions of repeat numbers per protein (Figure 3) shows that closed β-barrel structures display maxima around the observed barrel repeat number of seven in

Table 3. Analysis of repeats represented in three eukaryotic proteomes, those of SC, *S. cerevisiae*, CE, *C. elegans* and HS, *H. sapiens*

	SC			CE			HS (97% n.r.)			SW				
Repeat	*P*	o/oo	*R*	*P*	o/oo	*R*	*P*	o/oo	*R*	Bact	Euk	Arch	Vir	Total
ANK	18	2.9	3.56	78	4.0	5.59	64	5.4	5.17	4	86	0	20	110
ARM	2	0.3	8.50	3	0.2	6.00	16	1.4	6.06	0	34	0	0	34
HAT	7	1.1	6.43	5	0.3	9.60	4	0.3	6.25	0	7	0	0	7
HEAT	12	1.9	8.08	8	0.4	5.75	18	1.5	8.61	0	47	0	0	47
KELCH	5	0.8	4.00	17	0.9	6.41	11	0.9	5.27	2	20	0	9	31
LRR	10	1.6	6.20	52	2.7	10.15	59	5.0	10.92	7	112	0	1	120
PFTA	2	0.3	5.50	2	0.1	5.50	3	0.3	5.00	0	8	0	0	8
PFTB	4	0.6	4.25	3	0.2	5.00	4	0.3	4.75	6	21	0	0	27
RCC1	3	0.5	5.00	6	0.3	4.50	7	0.6	7.71	0	11	0	0	11
TPR	24	3.9	6.42	49	2.5	5.49	49	4.1	6.22	24	89	7	0	120
WD40	93	15.0	5.95	117	6.0	5.97	91	7.7	6.78	7	301	0	0	308

SC and CE represent completely sequenced genomes with 6,218 and 19,351 protein sequences. For human sequences 11,827 sequences were obtained after removing sequences more than 97% identical from the set of human proteins deposited in several sequence databases (as of December 1998). *P*, Number of proteins with the repeat; o/oo, proteins with the repeat per 1000 proteins in the set; *R*, repeats per protein. Some of the sequences identified as containing HEAT repeats were also identified as containing ARM repeats (one for SC, none for CE, six for HS). The final columns show repeat numbers in SwissProt listed according to the separate phyla of Bacteria, Eukaryotes, Archaea and Viruses. Details of this analysis can be accessed through the internet at the web address http://www.embl-heidelberg.de/~andrade/papers/rep.

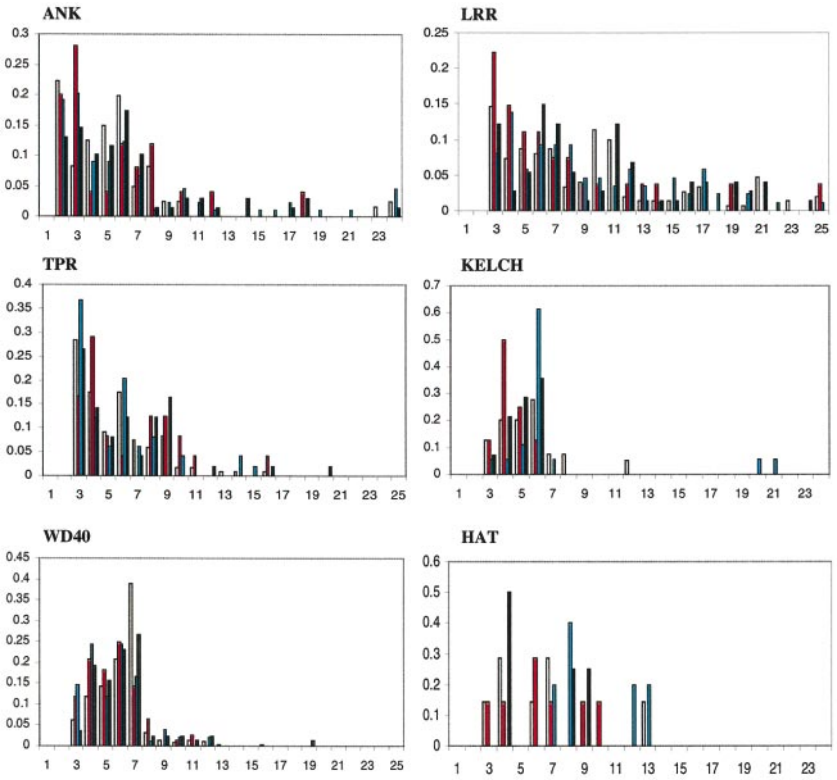

Figure 3. Length distributions of repeats in four sets of sequences: SW, SC, CE, HS as described in Table 3 caption, represented as white, red, blue and black columns, respectively. *X*-axis represents the number of repeats found in a protein. *Y*-axis represents the fraction of proteins of a given set containing the repeat. Note how repeat numbers in propellers (kelch, WD40) tend to cluster around specific values (six or seven or multiples thereof). By contrast open structures (ANK, LRR, TPR) show more variation. According to this observation, HAT repeats are predicted to form rods rather than propellers.

kelch and WD40 structures. Lower values are predicted to be due to the lack of detection of all repeats, whereas higher numbers appear to represent higher order harmonics due to the presence of multiple barrels. By contrast open structures containing ANK, LRR, and TPR repeats show more gradually decreasing distributions indicating lesser constraints on repeat numbers.

These findings present implications for the evolution of repeated structures that differ depending on whether the repeats occur in either open or close structures. Due to geometrical considerations there are likely to be few numbers of repeats possible in β-propeller structures (Murzin, 1992). However, the fact that these structures have been generated by (stepwise) duplication events indicates that open structures with smaller number of repeats have existed in the evolutionary past. By contrast, there appear to be few constraints on repeat numbers in open structures.

This analysis was used to predict the structure of HAT repeats, the only one of the eleven families without a known structure. The repeat numbers distribution is more similar to those of open structures than it is to closed structures (Figure 3). This suggests that HAT repeats form open structures. This is consistent with their sequence similarity to TPRs that form open structures (Figure 1).

For *C. elegans* proteins, an additional analysis of all annotated alternative splicing variants was performed. It was observed that alternative splicing results in clusters of repeats remaining either intact or else deleted in their entireties (data not shown). This implies that clusters of repeats mostly form super-structures that function and fold only when fully intact.

Discussion

The detection of repeats in database searches is problematic due in part to their low levels of sequence similarity and their shortness in length. To date, database search methods that have estimated *P*-values for repeats' alignment scores have not explicitly used the distributions of scores for suboptimal alignments. Here, we have demonstrated for a variety of repeat families, that the distributions of sub-optimal non-overlapping alignment scores for searches against a randomised sequence database are well-described by EVDs. This enables the estimation of *P*-values for a single repeat, whether it corresponds to the optimal alignment or a sub-optimal alignment.

These repeat-specific *P*-values have been used for the detection of 11 repeat types in the SwissProt database and in complete genomes. This approach has been encapsulated in a program called REP. The performance of REP has been compared with HMMER2, a method that employs a different methodology for the detection of repeats. HMMER2 is the search method used by the Pfam domain database which, in turn, is now used to annotate repeats in SwissProt. Although the primary goal of HMMER2 is to identify domains, its capability to detect repeats using empiric thresholds has drastically improved the annotation

quality of repeats in SwissProt. Yet, the comparison to REP demonstrated that although the quality of alignments used for searches is a factor in repeat detection, REP is superior at least in terms of sensitivity for the detection of both repeat-containing proteins and the repeats therein. However, one has to consider that the comparison was done with the Pfam annotation of SwissProt. The latter implies a manual control and, if required, adjustment of thresholds for each domain. Although this is also true for REP (Table 2), the computation of the *P*-values for non-overlapping sub-optimal alignments appears to be a better statistical model and thus the major reason for the improvement.

As a result, previously unrecognised repeats were identified in a number of proteins. In addition, REP indicated homology of different repeat types studied. This implies that there are divergent families of larger repeat superfamilies. Our analysis supports that of Cingolani *et al*. (1999) in suggesting a common evolutionary origin of ARM and HEAT repeats. ARM and HEAT repeats possess structural similarities with differences in the kinking of the first α-helix in ARM repeats (Groves & Bartford, 1999; Kobe *et al*., 1999). In addition, the consensus sequence of HEAT repeats contains several conserved charged amino acid residues that are absent in ARM repeats.

The TPR family of repeats is represented in extremely diverse phyla and is likely to have originated prior to the last common ancestor of archaea, eukarya and bacteria. TPR sequences are extremely diverse. One consequence of this is that TPRs have only recently been recognised in proteins such as kinesin light chains (Rollins *et al*., 1999) and SNAP secretory proteins (Ordway *et al*., 1994). Here, we have provided evidence that TPRs occur within clathrin heavy chain repeats and that HAT and PFTA repeats represent divergent TPR subclasses.

Kajava (1998) suggested separate phylogenetic origins for several different classes of LRRs based on the high levels of conservation within each LRR subfamily. Although our analysis using REP showed distinctions in the sequences of each LRR class, it was also found that searches could not absolutely partition LRRs into these separate classes (unpublished results). This suggests a common phylogenetic origin for these repeats, rather than separate origins as proposed by Kajava (1998).

Improved detection of repeats has allowed greater insights into the differences in repeat numbers between structurally-distinct repeat assemblies. Results for closed structures (WD40, PFTB, RCC1 and kelch repeats) indicate upper bounds for the numbers of repeats in each structure, whereas there appear to be no such limits for open structures (ARM, HEAT, ANK, TPR and PFTA repeats). It is suggested that this type of repeat number analysis can be used to predict the type of structural assembly formed by repeats of unknown structure.

It is foreseen that the approach used here to assign *P*-values to single repeats might be applicable for additional purposes. These might include the estimation of a single *P*-value per protein for the detection of its *n* repeats, and investigations of inter-repeat distances. The latter distributions could be used to distinguish further between true positive repeat scores from false positive scores since, in general, repeats appear to cluster in sequence as well as in structure. Finally, a similar approach to that employed here might improve the detection of domain pairs, such as Dbl and pleckstrin homology domains, whose occurrences in proteins are positively correlated.

Methods

Background

The question of detecting divergent repeats is associated with the question of assigning appropriate statistical estimates of whether sequence similarities result from divergence from a common ancestor (''homology''), and consequently represent true similarities in protein structure and in function. Many sequence database search algorithms use local alignment statistics reviewed by Altschul *et al*. (1994) to estimate the number of sequences with scores equal to a value x, or greater, that are expected purely by chance for a particular database size. Calculation of this ''Expectation value'' is based on the analytically derived statistics (Karlin & Altschul, 1990). This predicts that, when searching a database of uniform length random sequences, the scores of high-scoring segment pairs (HSPs) obey an extreme value distribution (EVD). It is noteworthy that this theory relates only to the distribution of highest scoring (optimal) local alignments, and is not formally applicable to the score distributions of the second highest, and of subsequent, sub-optimal alignments.

Database searches that are specific for protein sequences that contain repeats require an explicit correlation between the scores of optimal and sub-optimal alignments. This might involve setting lower acceptance score thresholds for suboptimal alignments than for optimal alignments, as initially implemented in the SMART server (Schultz *et al*., 1998). Alternatively, the bit scores of repeats that score above a low threshold might be summed. This approach is adopted in HMMER2 (Eddy, S., unpublished; http://hmmer.wustl.edu/) which is the underlying search method currently used in the domain detection systems Pfam (Bateman *et al*., 1999) and SMART (Ponting *et al*., 1999). A third approach would be to use the statistics of sub-optimal alignment scores (Karlin & Altschul, 1993). However, this approach is best suited to long sequences and/or profiles since ''edge-effects'' are expected to significantly distort score distributions for the short repeats under investigation here.

As an alternative to these approaches, REP calibrates a profile of a repeat by fitting the EVDs of optimal and sub-optimal scores of non-overlapping alignments to a database of random sequences.

Calibration of a profile

The probability of observing a score greater than x by chance in a pairwise comparison (*P*-value) is described by an extreme value distribution:

$$P(x) = 1 - e^{-e^{-\lambda(x-u)}} \quad (1)$$

where λ is the decay constant and u the characteristic value. For the ungapped comparison of two sequences of length m and n (in the limit where m and n are sufficiently large):

$$u = \frac{\ln Kmn}{\lambda} \quad (2)$$

where K and λ are constants that depend on the substitution scores and sequences (Altschul *et al.*, 1994). Equation (1) can be rewritten as:

$$P(x) = 1 - e^{-Kmne^{-\lambda x}} \quad (3)$$

This pairwise, ungapped comparison P-value is multiplied by the number of database sequences to produce the number of HSPs with scores at least x expected in a given database search (E-value). Although there is no similar formalism for alignments containing gaps, computational experiments suggest that this theory remains valid (Smith *et al.*, 1985; Collins *et al.*, 1988; Mott, 1992; Altschul & Gish, 1996).

Furthermore, the formulism is not applicable to global alignments. Consequently, in the method described below the statistics of sequence similarities are derived from local alignments, although alignments are subsequently extended to global alignments by HSP extensions.

Here, we have used a slightly modified description of equation (3) for the computation of P-values (as implemented in SWise (Birney *et al.*, 1996)):

$$P(x) = 1 - e^{a/e^{bx}} \quad (4)$$

Implementations of the Smith-Waterman algorithm (Smith & Waterman, 1981) estimate P-values by fitting the distribution of local alignment scores, generated from a database search, to equation (4). Equation (4) can be rewritten as:

$$\log_e(-\log_e(1 - P(x))) = \log_e(a) + bx \quad (5)$$

A histogram was constructed from the list of scores for all sequences in the database search. Cumulative frequencies c_i were calculated for each bin i. Parameters a and b were estimated from a linear fit to the function $\log_e(-\log_e(1 - c_i))$ using linear regression. The region of the histogram used for the fit was defined as between the bin with the highest frequency value (the modal bin) and the highest bin i with $c_i > 0.01$ (see Figure 2).

However, in cases where the number of true homologues is large, this procedure results in inaccurate P-value estimates due to the presence of true homologues with scores in the range over which the EVD is fitted. Here we follow the method described (Hoffmann & Bucher, 1995) by generating an artificial amino acid sequence database. This was created by extracting at random N sequences from a parent database and subsequently randomly reordering the sequences within windows of length w. This has the advantages that the created ''randomised'' database and the parent database have effectively equivalent amino acid residue compositions, and that individual sequences within the randomised database are exactly equivalent in local composition to their parent sequences. In this application $N = 20{,}000$ sequences were extracted at random from the SwissProt database (Bairoch & Apweiler, 1999). These were shuffled using a window length of $w = 20$ amino acid residues following a protocol described elsewhere (Hoffmann & Bucher, 1995). The chosen value of w is of the order of the size of the repeats studied here. This ensures that the randomised sequence database retains local compositional bias without conservation of repeat families' sequence motifs.

Local alignment scores were derived from SWise (Thompson *et al.*, 1994; Birney *et al.*, 1996) database searches using log-odds (''negative'') profiles calculated from multiple alignments (Birney *et al.*, 1996), using the Gonnet250 substitution matrix. ''Positive profiles'', used for producing global alignments, were calculated similarly (Birney *et al.*, 1996). Multiple alignments were constructed using sequences identified as homologues in PSI-BLAST (Altschul *et al.*, 1997) ($E < 0.001$) and/or HMMER2 (http://hmmer.wustl.edu/) ($E < 0.01$) database searches, and by reference to determined tertiary structures if available. Alignments with more than 400 sequences were purged of one of each pair of sequences with greater identity than a certain threshold to reduce the size of the alignment and the time for making the profile, losing a minimum of sequence variability. Fitting an EVD to a plot of cumulative frequency *versus* optimal local alignment scores provided the constants a_1 and b_1 of equation (4) that allowed calculation of $P_1(x_1)$, the probability of obtaining a score greater than, or equal to, the optimal local alignment score by chance alone. Note that P-values are independent of the database size. This allows the establishment of a single threshold P_0 for each repeat family that is independent of both database size and repeat order. The corresponding expectation value is the number of unrelated sequences of repeat order i scoring above x_i in a given database of size N is simply the product of $E_i(x_j) = P_i(x_i) \times N$.

Constants were similarly calculated for sub-optimal non-overlapping alignment scores thereby allowing estimations of $P_i(x_i)$ $(i > 1)$. The highest sub-optimal non-overlapping alignment score distribution for $i = 2$ was generated by comparison of a negative profile with a derivative randomised database. This was assembled from all sequences giving any detectable hit and contained substitutions of ''X'' for all amino acid residues that were identified previously as being present within optimal global alignments. Global alignments were generated from local alignments by comparison of the positive profile of length l with the region identified by the local alignment method extended by $l/2$ at both N and C-terminal ends.

This procedure was iterated for further orders $(i > 2)$ of sub-optimal alignment score distributions until the derivative randomised database size was less than 500 sequences (2.5% of the original database size). Fitting EVDs to plots of cumulative frequency *versus* sub-optimal local alignment scores provided the constants a_i and b_i of equation (4) that allowed calculation of $P_i(x_i)$, the P-value of the $(i - 1)$th non-overlapping sub-optimal local alignment score for a given database.

Recognition of repeats in one sequence

Comparison of a profile with a single sequence that has been masked over its optimal alignment sequence yields the first sub-optimal alignment sequence. Subsequent masking of this sequence and profile comparison reveals the second sub-optimal alignment. Further iterations of this procedure yield a series of non-overlapping alignments that are ranked according to their decreasing scores. Each of these scores is converted into corresponding $P_i(x_i)$ values using equation (4). In order to delineate

true from false positive repeats we required the choice of two thresholds. The first of these is the minimum number of repeats reported per sequence (n_{min}) Increasing this number improves the selectivity of the search, at the expense of sensitivity. The second is an *P*-value threshold, P_θ, that is applicable to all repeat orders *i*. This threshold is applied in the following manner. The top scoring $i = 1, 2, \ldots, n$ repeats are considered as true positives if all repeats possess $P_{i+1}(x_i) < P_\theta$, including $P_n(x_n) < P_\theta$. Thus, repeats *i* that obey $P_{i+1}(x_i) < P_\theta$ are held as ''pending'' until a higher order repeat *j* satisfies condition $P_j(x_j) < P_\theta$. At this point they are assigned as true positives. Pending repeats *i* that have no higher order *j* with $P_j(x_j) < P_\theta$ are assigned as false positives. This manner of applying a single threshold P_θ was found to be more sensitive than requiring that all repeats $i = 1, 2, \ldots, n$ have $P_i(x_i)$-values less than a threshold P_θ (unpublished results).

Implementation details

Calculations were computed on a 440 MHz DEC Alpha. Generation of the 20,000 sequence randomised database (occupying 7.7 Mb of disk space) using $w = 20$ required approximately 44 minutes. Calculation of the a_i, b_i coefficients of equation (4) for a typical repeat family of length 30 required approximately 30 minutes of CPU In comparison, the identification of putative homologues from databases is rapid: for example, a typical search of the effectively complete collection of 19,351 *C. elegans* sequences required five minutes of CPU. REP can be used via a public web server accessible from http://www.embl-heidelberg.de/~andrade/papers/rep/ together with detailed results of the analysis presented in Tables 2 and 3.

Acknowledgments

We thank Julie Thompson for her help in using WiseTools, Ewan Birney and Richard Mott for fruitful discussions and Joerg Schultz and Richard Copley for discussions relating to the SMART database. The work is supported by the BMBF and the DFG.

References

Adams, J. C., Seed, B. & Lawler, J. (1998). Muskelin, a novel intracellular mediator of cell adhesive and cytoskeletal responses to thrombospondin-1. *EMBO J.* **17**, 4964-4974.

Altschul, S. F. & Gish, W. (1996). Local alignment statistics. *Methods Enzymol.* **266**, 460-480.

Altschul, S. F., Boguski, M. S., Gish, W. & Wootton, J. C. (1994). Issues in searching molecular sequence databases. *Nature Genet.* **6**, 119-129.

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. *Nucleic Acids Res.* **25**, 3389-3402.

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

Asano, K., Kinzy, T. G., Merrick, W. C. & Hershey, J. W. (1997). Conservation and diversity of eukaryotic translation initiation factor eIF3. *J. Biol. Chem.* **272**, 1101-1109.

Asano, K., Phan, L., Anderson, J. & Hinnebusch, A. G. (1998). Complex formation by all five homologues of mammalian translation initiation factor 3 subunits from yeast Saccharomyces cerevisiae. *J. Biol. Chem.* **273**, 18573-18585.

Auble, D. T., Wang, D., Post, K. W. & Hahn, S. (1997). Molecular analysis of the SNF2/SWI2 protein family member MOT1, an ATP-driven enzyme that dissociates TATA-binding protein from DNA. *Mol. Cell. Biol.* **17**, 4842-4851.

Bairoch, A. & Apweiler, R. (1999). The SWISS-PROT protein sequence data bank and its new supplement TrEMBL in 1999. *Nucl. Acids Res.* **27**, 49-54.

Baron, M., Norman, D. G. & Campbell, I. D. (1991). Protein modules. *Trends Biochem. Sci.* **16**, 13-17.

Batchelor, A. H., Piper, D. E., de la Brousse, F. C., McKnight, S. L. & 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., Finn, R. D. & Sonnhammer, E. L. (1999). Pfam 3.1: 1313 multiple alignments match the majority of proteins. *Nucl. Acids Res.* **27**, 260-262.

Benni, M. L. & Neigeborn, L. (1997). Identification of a new class of negative regulators affecting sporulation-specific gene expression in yeast. *Genetics,* **147**, 1351-1366.

Birney, E., Thompson, J. D. & Gibson, T. J. (1996). PairWise and SearchWise: finding the optimal alignment in a simultaneous comparison of a protein profile against all DNA translation frames. *Nucl. Acids Res.* **24**, 2730-2739.

Boguski, M. S., Murray, A. W. & Powers, S. (1992). Novel repetitive sequence motifs in the α and β subunits of prenyl-protein transferases and homology of the α subunit to the *MAD2* gene product of Yeast. *New Biologist,* **4**, 408-411.

Bork, P. & Doolittle, R. F. (1994). Drosophila kelch motif is derived from a common enzyme fold. *J. Mol. Biol.* **236**, 1277-1282.

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

Bork, P. (1992). Mobile modules and motifs. *Curr. Opin. Struct. Biol.* **2**, 413-421.

Bork, P. (1993). Hundreds of ankyrin-like repeats in functionally diverse proteins: Mobile modules that cross phyla horizontally? *Proteins: Struct. Funct. Genet.* **17**, 363-374.

Charrasse, S., Mazel, M., Taviaux, S., Berta, P., Chow, T. & Larroque, C. (1995). Characterization of the cDNA and pattern of expression of a new gene over-expressed in human hepatomas and colonic tumors. *Eur. J. Biochem.* **234**, 406-413.

Chook, Y. M. & Blobel, G. (1999). Structure of the nuclear transport complex karyopherin-β2-Ran × GppNHp. *Nature,* **399**, 230-237.

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

Collins, J. F., Coulson, A. F. & Lyall, A. (1988). The significance of protein sequence similarities. *Comput. Appl. Biosci.* **4**, 67-71.

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

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.

Davis, J. L., Kunisawa, R. & Thorner, J. (1992). A presumptive helicase (*MOT1* gene product) affects gene expression and is required for viability in the yeast *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **12**, 1879-1892.

Doolittle, R. F. (1989). Similar amino acid sequences revisited. *Trends Biochem. Sci.* **14**, 244-245.

Evans, D. R., Rasmussen, C., Hanic-Joyce, P. J., Johnston, G. C., Singer, R. A. & Barnes, C. A. (1995). Mutational analysis of the Prt1 protein subunit of yeast translation initiation factor 3. *Mol. Cell. Biol.* **15**, 4525-4535.

Garcia-Higuera, I., Fenoglio, J., Li, Y., Lewis, C., Panchenko, M. P., Reiner, O., Smith, T. F. & Neer, E. J. (1996). Folding of proteins with WD-repeats: comparison of six members of the WD-repeat superfamily to the G protein β-subunit. *Biochemistry,* **35**, 13985-13994.

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

Görlich, D., Prehn, S., Laskey, R. A. & Hartmann, E. (1994). Isolation of a protein that is essential for the first step of nuclear protein import. *Cell,* **79**, 767-778.

Groves, M. R. & Bartford, D. (1999). Topological characteristics of helical repeat proteins. *Curr. Opin. Struct. Biol.* **9**, 383-389.

Groves, M. R., Hanlon, N., Throwski, P., Hemmings, B. A. & 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.

Hanic-Joyce, P. J., Singer, R. A. & Johnston, G. C. (1987). Molecular characterization of the yeast PRT1 gene in which mutations affect translation initiation and regulation of cell proliferation. *J. Biol. Chem.* **262**, 2845-2851.

Heringa, J. & Argos, P. (1993). A method to recognize distant repeats in protein sequences. *Proteins: Struct. Funct. Genet.* **17**, 391-441.

Heringa, J. & Taylor, W. R. (1997). Three-dimensional domain duplication, swapping and stealing. *Curr. Opin. Struct. Biol.* **7**, 416-421.

Heringa, J. (1994). The evolution and recognition of protein sequence repeats. *Comput. Chem.* **18**, 233-243.

Hernandez, M. C., Andres-Barquin, P. J., Martinez, S., Bulfone, A., Rubenstein, J. L. & Israel, M. A. (1997). ENC-1: a novel mammalian kelch-related gene specifically expressed in the nervous system encodes an actin-binding protein. *J. Neurosci.* **17**, 3038-3051.

Hoffmann, K. & Bucher, P. (1995). The FHA domain: a putative nuclear signalling domain found in protein kinases and transcription factors. *Trends Biochem. Sci.* **20**, 347-349.

Holm, L. & Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**, 123-138.

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

Ito, N., Phillips, S. E., Stevens, C., Ogel, Z. B., McPherson, M. J., Keen, J. N., Yadav, K. D & Knowles, P. F. (1991). Novel thioether bond revealed by a 1.7 Å crystal structure of galactose oxidase. *Nature,* 87-90.

Kajava, A. V. (1998). Structural diversity of leucine-rich repeat proteins. *J. Mol. Biol.* **277**, 519-527.

Karlin, S. & Altschul, S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl Acad. Sci. USA,* **87**, 2264-2268.

Karlin, S. & Altschul, S. F. (1993). Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc. Natl Acad. Sci. USA,* **90**, 5873-5877.

Kawaide, H., Imai, R., Sassa, T. & Kamiya, Y. (1997). Ent-kaurene synthase from the fungus Phaeosphaeria sp. L487. cDNA isolation, characterization, and bacterial expression of a bifunctional diterpene cyclase in fungal gibberellin biosynthesis. *J. Biol. Chem.* **272**, 21706-21712.

Kim, I. F., Mohammadi, E. & Huang, R. C. (1999). Isolation and characterization of IPP, a novel human gene encoding an actin-binding, kelch-like protein. *Gene,* **228**, 73-83.

Kleinschmidt, J. A., Dingwall, C., Maier, G. & Franke, W. W. (1986). Molecular characterization of a karyophilic, histone-binding protein: cDNA cloning, amino acid sequence and expression of nuclear protein N1/N2 of *Xenopus laevis*. *EMBO J.* **5**, 3547-3552.

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

Kobe, B., Gleichmann, T., Horne, J., Jennings, I. G., Scotney, P. D. & Teh, T. (1999). Turn up the HEAT. *Structure,* **7**, R91-R97.

Lacana′, E. & D′Adamio, L. (1999). Regulation of Fas ligand expression and cell death by apoptosis-linked gene 4. *Nature Med.* **5**, 542-547.

Li, J., Brick, P., O′Hare, M. C., Skarzynski, T., Curry, L. F. L. V. A., Clark, I. M., Bigg, H. F., Hazleman, B. L., Cawston, T. E., *et al.* (1995). Structure of full-length porcine synovial collagenase reveals a C-terminal domain containing a calcium-linked, four-bladed β-propeller. *Structure,* **3**, 541-549.

Long, S. B., Casey, P. J. & Beese, L. S. (1998). Cocrystal structure of protein farnesyltransferase complexed with a farnesyl diphosphate substrate. *Biochemistry,* **37**, 9612-9618.

Lutcke, H., Prehn, S., Ashford, A. J., Remus, M., Frank, R. & Dobberstein, B. (1993). Assembly of the 68 and 72-kD proteins of signal recognition particle with 75 RNA. *J. Cell. Biol.* **121**, 977-985.

Lux, S. E., John, K. M. & 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.

Matthews, L. R., Carter, P., Thierry-Mieg, D. & Kemphues, K. (1998). ZYG-9, a *Caenorhabditis elegans* protein required for microtubule organization and function, is a component of meiotic and mitotic spindle poles. *J. Cell. Biol.* **141**, 1159-1168.

McLachlan, A. D. (1977). Analysis of periodic patterns in amino acid sequences: collagen. *Biopolymers,* **16**, 1271-1297.

McLachlan, A. D. (1978). Coiled coil formation and sequence regularities in the helical regions of α-keratin. *J. Mol. Biol.* **124**, 297-304.

Methot, N., Rom, E., Olsen, H. & Sonenberg, N. (1997). The human homologue of the yeast Prt1 protein is an integral part of the eukaryotic initiation factor 3

complex and interacts with p170. *J. Biol. Chem.* **272**, 1110-1116.

Mott, R. (1992). Maximum likelihood estimation of the statistical distribution of Smith-Waterman local sequence similarity scores. *Bull. Math. Biol.* **54**, 59-75.

Murzin, A. G. (1992). Structural principles for the propeller assembly of β-sheets: the preference for seven-fold symmetry. *Proteins: Struct. Funct. Genet.* **14**, 191-201.

Naranda, T., MacMillan, S. E. & Hershey, J. W. (1994). Purified yeast translational initiation factor eIF-3 is an RNA-binding protein complex that contains the PRT1 protein. *J. Biol. Chem.* **269**, 32286-32292.

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

Ohtsubo, M., Kai, R., Furuno, N., Sekiguchi, T., Sekiguchi, M., Hayashida, H., Kuma, K., Miyata, T., Fukushige, S., Murotsu, T., Matsubara, K. & Nishimoto, T. (1987). Isolation and characterization of the active cDNA of the human cell cycle gene (RCC1) involved in the regulation of onset of chromosome condensation. *Genes Dev.* **1**, 585-593.

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

Pasquier, G. M., Promponas, V. I., Varvayannis, N. J. & Hamodrakas, S. J. (1990). A Web server to locate periodicities in a sequence. *Bioinformatics,* **215**, 403-410.

Pearson, W. R. (1998). Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol.* **276**, 71-84.

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

Perego, M. & Hoch, J. A. (1996). Protein aspartate phosphatases control the output of two-component signal transduction systems. *Trends Genet.* **12**, 97-101.

Phillips, A. M., Bull, A. & Kelly, L. E. (1992). Identification of a *Drosophila* gene encoding a calmodulin-binding protein with homology to the trp phototransduction gene. *Neuron,* **8**, 631-642.

Ponting, C. P., Schultz, J., Milpetz, F. & Bork, P. (1999). SMART: identification and annotation of domains from signalling and extracellular protein sequences. *Nucl. Acids Res.* **27**, 229-232.

Poralla, K., Hewelt, A., Prestwich, G. D., Abe, I., Reipen, I. & Sprenger, G. (1994). A specific amino acid repeat in squalene and oxidosqualene cyclases. *Trends Biochem. Sci.* **19**, 157-158.

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

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

Riggleman, B., Wieschaus, E. & 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.

Robinson, D. N. & Cooley, L. (1997). *Drosophila* kelch is an oligomeric ring canal actin organizer. *J. Cell. Biol.* **138**, 799-810.

Rollins, R. A., Morcillo, P. & Dorsett, D. (1999). Nipped-B, a *Drosophila* homologue of chromosomal adherins, participates in activation by remote enhancers in the cut and Ultrabithorax genes. *Genetics,* **152**, 577-593.

Russell, R. B. & Ponting, C. P. (1998). Protein fold irregularities that hinder sequence analysis. *Curr. Opin. Struct. Biol.* **8**, 364-371.

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

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

Seitz, L. C., Tang, K., Cummings, W. J. & Zolan, M. E. (1996). The *rad9* gene of *Coprinus cinereus* encodes a proline-rich protein required for meiotic chromosome condensation and synapsis. *Genetics,* **142**, 1105-1117.

Sikorski, R. S., Boguski, M. S., Goebl, M. & 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.

Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195-197.

Smith, T. F., Waterman, M. S. & Burks, C. (1985). The statistical distribution of nucleic acid similarities. *Nucl Acids Res.* **13**, 645-656.

Smith, G. L., Chan, Y. S. & Howard, S. T. (1991). Nucleotide sequence of 42 kbp of vaccinia virus strain WR from near the right inverted terminal repeat. *J. Genet. Virol.* **72**, 1349-1376.

Smith, T. F., Gaitatzes, C., Saxena, K. & Neer, E. J. (1999). The WD repeat: a common architecture for diverse functions. *Trends Biochem. Sci.* **24**, 181-185.

Sondek, J., Bohm, A., Lambright, D. G., Hamm, H. E. & Sigler, P. B. (1996). Crystal structure of a G-protein βγ dimer at 2.1 Å resolution. *Nature,* **379**, 369-374.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). Improved sensitivity of profile searches through the use of sequence weights and gap excision. *Comput. Appl. Biosci.* **10**, 19-30.

Thony-Meyer, L., Fischer, F., Kunzier, P., Ritz, D. & Hennecke, H. (1995). *Escherichia coli* genes required for cytochrome c maturation. *J. Bacteriol.* **177**, 4321-4326.

Torchet, C., Jacq, C. & Hermaun-Le Denmat, S. (1998). Two mutant forms of the S1/TPR-containing protein Rrp5p affect the 18S rRNA synthesis in *Saccharomyces cerevisiae*. *RNA,* **4**, 1636-1652.

Utz, P. J., Hottelet, M., Le, T. M., Kim, S. J., Geiger, M. E., van Venrooij, W. J. & Anderson, P. (1998). The 72-kDa component of signal recognition particle is cleaved during apoptosis. *J. Biol. Chem.* **273**, 35362-35370.

Venema, J. & Tollervey, D. (1996). RRP5 is required for formation of both 18S and 5.8S rRNA in yeast. *EMBO J.* **15**, 5701-5714.

Vetter, I. R., Arndt, A., Kutay, U., Görlich, D. & Wittinghofer, A. (1999). Structural view of the ran-importin β interaction at 2.3 Å resolution. *Cell,* **97**, 635-646.

Wall, M. A., Coleman, D. E., Lee, E., Iñiguez-Lluhi, J. A., Posner, B. A., Gilman, A. G. & Sprang, S. R. (1995). The structure of the G protein heterotrimer $G_i\alpha_1\beta_1\gamma_2$. *Cell,* **83**, 1047-1058.

Wang, P. J. & Huffaker, T. C. (1997). Stu2p: a microtubule-binding protein that is an essential component of the yeast spindle pole body. *J. Cell Biol.* **139**, 1271-1280.

Wang, C., Chua, K., Seghezzi, W., Lees, E., Gozani, O. & Reed, R. (1998). Phosphorylation of spliceosomal protein SAP 155 coupled with splicing catalysis. *Genes Dev.* **12**, 1409-1414.

Way, M., Sanders Garcia, C., Sakai, J. & Matsudaira, P. (1995). Sequence and domain organization of scruin, an actin-cross-linking protein in the acrosomal process of Limulus sperm. *J. Cell. Biol.* **128**, 51-60.

Wendt, K. U., Lenhart, A. & Schulz, G. E. (1998). The structure of the membrane protein squalene-hopene cyclase at 2.0 Å resolution. *J. Mol. Biol.* **286**, 175-187.

Ybe, J. A, Brodsky, F. M., Hofmann, K., Lin, K., Liu, S. H., Chen, L., Earnest, T. N., Fletterick, R. J. & Hwang, P. (1999). Clathrin self-assembly is mediated by a tandemly repeated superhelix. *Nature,* **399**, 371-375.

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

Edited by J. Thornton

(Received 3 August 1999; received in revised form 6 December 1999; accepted 14 January 2000)