doi:10.1006/jmbi.2001.4624 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2001) **309**, 1–18

JMB

REVIEW ARTICLE

Comparison of ARM and HEAT Protein Repeats

**Miguel A. Andrade[1,2], Carlo Petosa[3], Sean I. O'Donoghue[1]
Christoph W. Müller[3] and Peer Bork[1,2]***

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

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

[3]*Grenoble Outstation, European
Molecular Biology Laboratory
BP 156, 38042, Grenoble
Cedex 9, France*

**Corresponding author*

ARM and HEAT motifs are tandemly repeated sequences of approximately 50 amino acid residues that occur in a wide variety of eukaryotic proteins. An exhaustive search of sequence databases detected new family members and revealed that at least 1 in 500 eukaryotic protein sequences contain such repeats. It also rendered the similarity between ARM and HEAT repeats, believed to be evolutionarily related, readily apparent. All the proteins identified in the database searches could be clustered by sequence similarity into four groups: canonical ARM-repeat proteins and three groups of the more divergent HEAT-repeat proteins. This allowed us to build improved sequence profiles for the automatic detection of repeat motifs. Inspection of these profiles indicated that the individual repeat motifs of all four classes share a common set of seven highly conserved hydrophobic residues, which in proteins of known three-dimensional structure are buried within or between repeats. However, the motifs differ at several specific residue positions, suggesting important structural or functional differences among the classes. Our results illustrate that ARM and HEAT-repeat proteins, while having a common phylogenetic origin, have since diverged significantly. We discuss evolutionary scenarios that could account for the great diversity of repeats observed.

Keywords: sequence analysis; protein structure; evolution; HEAT repeats; armadillo repeats

Introduction

Duplication events in coding genes produce repeated fragments of protein sequence ranging from one amino acid residue to thousands (Heringa, 1994). Among the most common tandemly repeated motifs with regular three-dimensional structure are α-helical domains of roughly 50 residues which pack together to form elongated super-helices, or "solenoids" (Groves & Barford, 1999; Kobe & Kajava, 2000). Two such motifs with particularly prominent roles in the eukaryotic cell are the Armadillo (ARM) and HEAT motifs. ARM repeats were first discovered in the *Drosophila* segment polarity gene product Armadillo (Riggleman *et al.*, 1989) and later in several other proteins, including the junctional plaque protein plakoglobin (Franke *et al.*, 1989), the tumor suppresor adenomatous polyposis coli (APC) (Peifer *et al.*, 1994), and the nucleocytoplasmic transport factor importin (or karyopherin) α (Görlich *et al.*, 1994). The repeated HEAT motif was initially found in a diverse family of proteins, including the four from which it derives its name: huntingtin, elongation factor 3, the PR65/A subunit of protein phosphatase 2A (PP2A), and the lipid kinase TOR (target of rapamycin) (Andrade & Bork, 1995). Although ARM and HEAT-repeat proteins are involved in a great diversity of cellular processes, a function common to many is that of mediating important protein-protein interactions.

Crystal structures have been determined for five ARM and HEAT-repeat-containing proteins (Figure 1): β-catenin (Huber *et al.*, 1997), the PR65/A subunit of PP2A (Groves *et al.*, 1999), and importins α, β1, and β2 (transportin) (Conti *et al.*, 1998; Cingolani *et al.*, 1999; Chook & Blobel, 1999). Importin-α and β-catenin contain 10 and 12 tan-

Present address: S. I. O'Donoghue, LION Bioscience AG, Waldhoferstr. 98, 69123 Heidelberg, Germany

Abbreviations used: ARM, armadillo; PP2A, protein phosphatase 2A; RMS, root-mean-square deviation; NILS, nuclear localization signal.

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

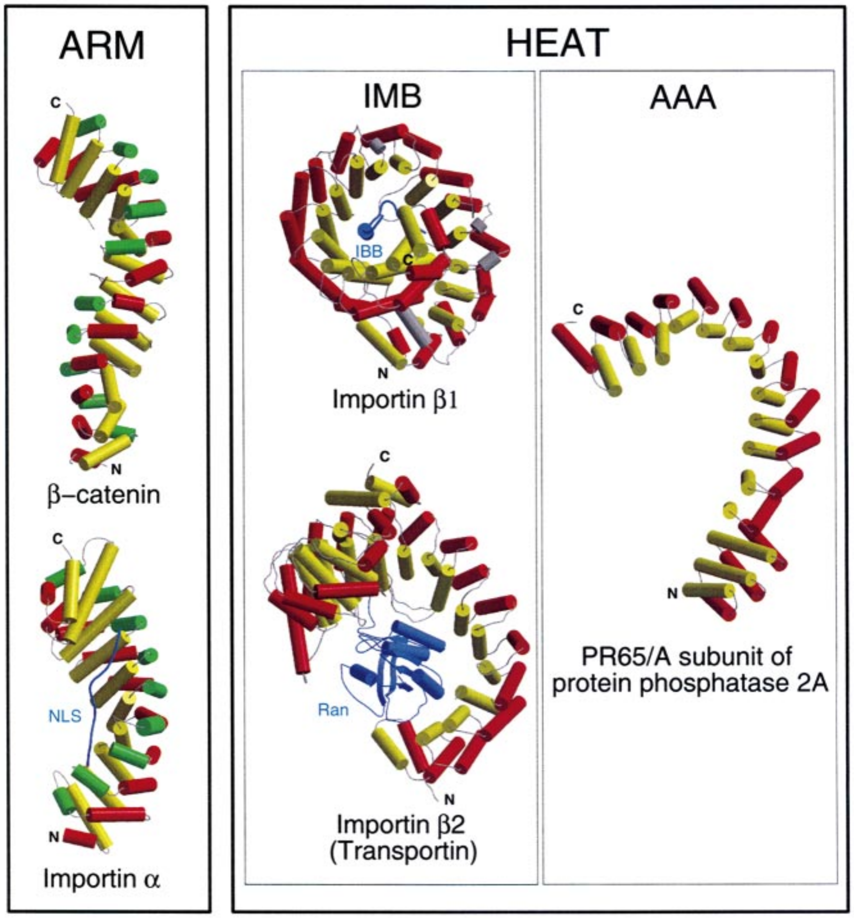

Figure 1. A gallery of HEAT and ARM repeat-containing proteins. H1, H2, and H3 helices of ARM proteins are shown in green, red, and yellow, respectively; A and B helices of HEAT proteins are in red and yellow, respectively; bound ligands are in blue. The structures shown are those of: mouse β-catenin (PDB 2bct; Huber *et al*., 1997); yeast importin-α bound to the bipartite NLS of nucleoplasmin (PDB 1ee5; Conti & Blobel, 2000); transportin (karyopherin β2) bound to Ran (PDB 1qbk, Chook & Blobel, 1999); human importin-β bound to the importin-α IBB domain (PDB 1qgk; Cingolani *et al*., 1999); and the PR65/A subunit of human protein phosphatase 2A (PDB 1b3u; Groves *et al*., 1999). The structures are oriented with their N termini at the bottom and C termini at the top.

dem ARM repeats, respectively, while the PR65/A subunit and importins β1 and β2 contain 15, 19, and 18 HEAT repeats, respectively. All these proteins participate in protein-protein interactions: importin-α recognizes cytosolic proteins bearing a basic nuclear localization signal (NLS), β-catenin binds to the conserved cytoplasmic domain of cadherins, the PR65/A subunit interacts with various regulatory B subunits of PP2A, and importins β1 and β2 bind to the small GTPase Ran and to various protein substrates destined for nuclear import. Except for β-catenin and the PR65/A subunit, details of these interactions have been revealed by three-dimensional structures of the protein-ligand complexes.

The canonical ARM repeat consists of three helices, denoted H1, H2, and H3. The H2 and H3 helices pack against each other in an antiparallel fashion and are roughly perpendicular to the shorter H1 helix, with a sharp bend between helices H1 and H2 mediated by a conserved glycine residue. The canonical HEAT repeat consists of only two helices, A and B, which form a helical hairpin. In both ARM and HEAT-repeat proteins, neighbouring repeats stack together into a single domain with a continuous hydrophobic core, forming an elongated super-helix. Despite having one less helix, the structure of the HEAT motif is in fact very similar to that of the ARM motif, with its strongly bent helix A corresponding to helices H1 and H2 of the ARM motif, and helix B corresponding to helix H3. This correspondence appears to extend to protein function as well. For example, in the ARM protein structures, the H3 helices form a highly conserved, concave surface implicated in ligand binding (Huber *et al*., 1997; Conti *et al*., 1998), while in the HEAT protein structures, the B helices also form a concave surface, which in importins β1 and β2 is a highly conserved, ligand-binding surface (Cingolani *et al*., 1999; Chook & Blobel, 1999). Indeed, a common phylogenetic origin has been proposed for the ARM repeats in α-importins and the HEAT repeats in the β-importins (Malik *et al*., 1997; Cingolani *et al*., 1999), indicating a common ancestor for much of the nuclear protein import/export machinery. This suggests that a single origin is common to all ARM/HEAT repeats, which subsequently diverged into different structural families (Cingolani *et al*., 1999; Kobe *et al*., 1999).

In contrast to ARM repeats which are relatively uniform, HEAT repeats are much more variable in length, amino acid sequence, and three-dimensional structure, rendering their identification by sequence comparison difficult, even using methods particularly tuned for repeat detection (Andrade *et al*., 2000; Bateman *et al*., 2000). It is clear that the proper identification and classification of repeats requires the combination of both structural and sequence information. Here, we present a survey of ARM and HEAT repeats attained *via* a combination of profile searches and the use of ARM and HEAT structural features. The analysis confirms that ARM and HEAT are separate repeat classes. Furthermore, the more divergent HEAT repeats could be subdivided into three clusters, which were used to generate an improved set of profiles. With these new profiles we could automatically identify a larger set of HEAT-repeat proteins and detect a greater number of repeats per protein. Our results strengthen the hypothesis of a common phylogenetic origin for ARM and HEAT repeats. We propose a mechanism to explain the evolutionary divergence of these repeats into at least four distinct classes in light of the relationship observed between sequences and structural features.

Classification of ARM and HEAT family members

As a first step towards understanding the relationships between ARM and HEAT repeats, we compared the sequences of a large number of repeat-containing protein fragments. We began by retrieving all the HEAT and ARM repeats in the SwissProt protein sequence database (Bairoch & Apweiler, 2000) which could be detected by the REP program, previously trained on sequence profiles for the two families (Andrade *et al*., 2000; for profile details, see the REP server at http://www.embl-heidelberg.de/ ~ andrade/papers/rep/search.html). The regions of protein sequence containing clusters of significantly identified repeats were extracted, and each was used as a query for PSI-BLAST searches (Altschul *et al*., 1997). Usually, in four to five iterations all the proteins containing ARM and HEAT repeats scored below the default *P*-value (probability of chance match) threshold of 0.001. Additional hits below this threshold but not detected by REP were individually inspected using reciprocal searches (Bork & Gibson, 1996), and were included in the repeat set if a known repeat-containing region scored lower than $P = 0.01$. Each case was checked against the domain databases Pfam (Bateman *et al*., 2000) and SMART (Schultz *et al*., 2000) to avoid spurious hits and to ensure that only repeat-containing regions of sequence were used for the subsequent analysis.

These repeat-containing protein fragments were then clustered using the BLAST *P*-values obtained during their sequence comparison as a measure of distance (Figure 2). (In case of multiple matches in runs with different queries the lowest values were used.) First, we grouped together all proteins with a high degree of similarity (BLAST $P\text{-value} < 10^{-40}$) to a leader sequence. These groups were then connected using a less stringent similarity criteria (BLAST $P\text{-value} < 10^{-5}$). Finally, sequences more weakly related (BLAST $P\text{-value} < 0.01$) to members of only one cluster were included. This led to four major clusters: one ARM class, and three groups of HEAT repeats which we designate as AAA, IMB and ADB classes (see Figure 2). At this similarity cut-off, only nine sequences overlap between these clusters: four related to human huntingtin (HD), four related to the target of rapamycin (TOR/FRAP), and the human TOGp protein (hCTOG). Profiles were built for each of the four clusters (classes) and used in REP for a new round of database searches. The newly identified members were added to the clusters if the similarity (by either sequence comparison or profile search) was conclusive. Only a few proteins could not be recognized by any of the profiles (see Table 1). For instance, the *Saccharonyces cerevisiae* Mot1p protein represents a special case, as it was connected by the PSI-BLAST search to the ARM class but was recognized by the AAA, not by the ARM, class profile.

In total, 37 ARM- and 71 HEAT-repeat containing sequences were detected in the SwissProt database by this approach (Table 1), accounting for 0.02-0.14 % and 0.06-0.32 %, respectively, of all proteins in yeast, *Caenorhabditis elegans*, *Drosophila* and humans (Table 2). Thus, we estimate that roughly 1 out of 500 eukaroytic proteins contains ARM or HEAT repeats. The number of repeats detected in each sequence using the four profiles is summarized in Table 1. These data were used to derive the final classification scheme shown in Figure 2. Additional information about the sequences and fragments used for the search can be found in Table 3.

The 37 ARM repeat sequences identified included proteins involved in nucleocytoplasmic transport (e.g. the family of importin α proteins), maintenance and control of the cellular cytoskeleton (e.g. β-catenin and plakoglobins), the tumor suppresor adenomatous polyposis coli and the guanine nucleotide exchange factor Gds1 (GDP dissociation stimulator) (Yamamoto *et al*., 1990). All 37 proteins cluster into a single class, including both proteins of known structure, importin α and β-catenin (Figure 2).

In contrast, the previously known HEAT repeats fell into two classes, the IMB and AAA classes. The 20 proteins of the IMB class are all nucleocytoplasmic shuttling factors with Ran-binding activity, including two, importins β1 and β2, of known structure (Cingolani *et al*., 1999; Chook & Blobel, 1999). Five additional sequences (all from *S. cerevisiae*) previously reported as HEAT proteins with a Ran-binding domain† were omitted from Figure 2, as we could neither detect them to contain repeats (with the conservative thresholds used here) nor relate them by PSI-BLAST to the sequences already clustered. Although these proteins indeed show features of HEAT repeats, they seem to have diverged rapidly. The 29 sequences of the AAA class are represented by one known 3D structure, that of the regulatory PR65/A subunit of the 2A protein phosphatase. However, they include proteins of widely differing activity, such as the yeast protein Vps15p, which forms a complex with Vps34p that is essential for vacuolar protein sorting (Stack *et al*., 1993); yeast elongation factor 3, an ATPase which stimulates the release of deacylated-tRNA from the ribosomal E-site; and the yeast GCN1 protein, which in complex with GCN20 activates the translational regulator GCN2 kinase (Garcia-Barrio *et al*., 2000).

Interestingly, we identified a novel fourth class, the ADB class (Figure 2). It comprises 22 protein sequences related to the clathrin-associated adaptor, a protein complex associated with the polyhedral clathrin lattice of clathrin-coated vesicles (Brodsky, 1997). An alignment of these protein repeats is shown in Figure 3. Two adaptors have been characterized in mammalian cells; both comprise four subunits, three of which are similar to one other (Traub, 1997). Other members of the class are the *Drosophila* Garnet protein (dGARN, Lloyd *et al*., 1999) and a hypothetical Yeast protein YBD7 (De Wergifosse *et al*., 1994). One member has previously been identified as containing ARM repeats (Küssel & Frasch, 1995), whereas others described the proteins of this class as candidate HEAT-repeat proteins (Andrade & Bork, 1995). Indeed, the class contains one sequence, yADB6, with significant similarity to both the ARM and AAA classes (Figure 2). However, the analysis of the repeat-containing regions by comparison to profiles and structural analysis indicates that the repeats in this class are more similar to HEAT than to ARM repeats (see below). It is unfortunate that no 3D structure for the repeat region of any member of this class has yet been determined, although a 3D structure of a fragment of α adaptin C (residues 701-938 of mADAC), which does not include the repeat region, is known (Traub *et al*., 1999).

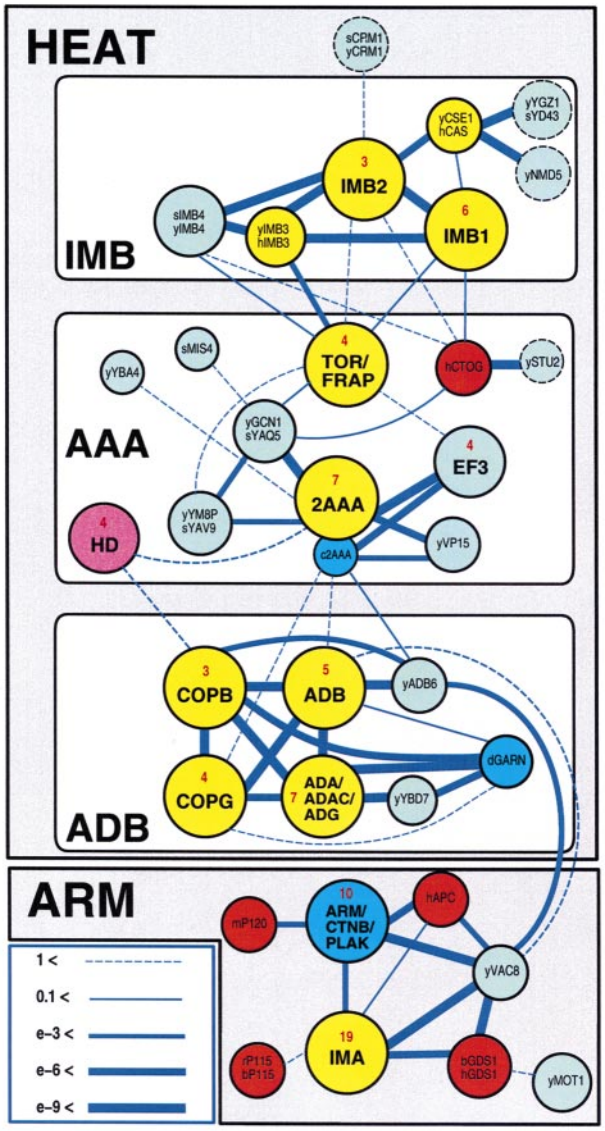

Figure 2. Similarity relationships between ARM or HEAT-repeat containing regions of proteins detected in the SwissProt database (Bairoch & Apweiler, 2000). The diagram represents the BLAST scores obtained in sequence similarity searches using regions of proteins identified as containing the repeat (therefore excluding other possible domains). Each bubble represents one sequence or a group of sequences, whose size is indicated by a red number inside the bubble. The colour of the bubble indicates the taxonomic span of the group: yellow, eukaryotic; green, fungi; cyan, metazoan; pink, vertebrate; orange, mammalian. The presence in a wide range of eukaryotes of all four classes suggests that these shared a common origin before the divergence of eukaryotes. For each group a representative ''leader'' sequence was chosen. The identifiers correspond to the SwissProt nomenclature with the species names abbreviated in front: s, *S. pombe;* y, yeast; h, human; m, mouse; b, bovine; r, rat; d, *D. melanogaster*. The connections between bubbles indicate the best *P*-value in a BLAST search from the representative of one group to any member of the other group (see legend in bottom left part of the Figure). Bubbles surrounded by a discontinuous line indicate a group for which no good repeat hits were identified, but which had a BLAST link to at least one member in a previously identified repeat subgroup.

† A profile of known members of the family was constructed (based on the alignment given by Gorlich *et al*., 1997). A Hidden Markov Model of the alignment was constructed and the SwissProt database scanned for hits above a score threshold ($E < 0.01$) using the hmmsearch algorithm (Eddy, 1998; http://hmmer.wustl.edu/).

Table 1. Number of repeats detected in SwissProt sequences with the profiles and thresholds described in Table 2

AAA		Standard annotation						
		sw	pfam	arm	aaa	imb	adb	H-old
2AAA	2AAA_CAEEL	15H			12	8		8
	2AAA_DROME	15H			14	10		10
	2AAA_HUMAN	15H			13	9		10
	2AAA_PEA	R			10	8		8
	2AAA_PIG	15H			13	9		10
	2AAA_YEAST	15H			12	10		12
	2AAB_HUMAN	15H			13	10		9
	2AAB_PIG	15H			13	9		10
EF3	EF3A_YEAST	10H			6			5
	EF3B_YEAST	10H			6			5
	EF3_CANAL	10H			6	5		5
	EF3_PNECA	10H						
	YAQ5_SCHPO				23	19	9	19
	GCN1_YEAST	36H			21	20	11	20
HD	HD_FUGRU	10H			7			5
	HD_HUMAN	10H			10			
	HD_MOUSE	10H			9		4	5
	HD_RAT	10H			8		4	5
	YBA4_YEAST	4H			10			
	VP15_YEAST	7H					4	
	YAV9_SCHPO				12	9	4	11
	YM8P_YEAST				12	9		9
	MIS4_SCHPO				6			5
TOR/	FRAP_HUMAN				9	9		8
FRAP	FRAP_RAT				9	9		8
	TOR1_YEAST	20H			8	8	8	8
	TOR2_YEAST				8	8	5	8
	CTOG_HUMAN				11	8	4	9
	STU2_YEAST							

ADB		Standard annotation						
		sw	pfam	arm	aaa	imb	adb	H-old
ADA/	ADA_DROME		ADA					
ADAC/	ADAA_MOUSE		ADA				4	
ADG	ADAC_MOUSE		ADA				4	
	ADAC_RAT		ADA				4	
	ADG_HUMAN		ADA					
	ADG_MOUSE		ADA					
	ADG_USTMA		ADA				5	
ADB	ADB1_HUMAN		ADA		5		7	
	ADB1_RAT		ADA				6	
	ADB1_YEAST		ADA					
	ADB2_YEAST		ADA				5	
	ADB_HUMAN		ADA		5		7	
	ADB6_YEAST							
COPB	COPB_DROME						5	
	COPB_RAT						5	
	COPB_YEAST						4	
COPG	COPG_BOVIN						4	
	COPG_CAEEL							
	COPG_SCHPO				5		4	
	COPG_YEAST				5		4	
	GARN_DROME		ADA					
	YBD7_YEAST		ADA					

IMB1		sw	pfam	arm	aaa	imb	adb	H-old
IMB1	IMB1_HUMAN	11H	1A		9	9		8
	IMB1_MOUSE	11H	1A		9	9		8
	IMB1_RAT	11H	1A		9	9		8
	IMB1_SCHPO	11H				9		
	IMB1_YEAST	11H				7		
	IMB_DROME	11H				8		
IMB2	IMB2_HUMAN	10H			9	10	4	9
	IMB2_SCHPO	10H			6	6		5
	IMB2_YEAST	10H			6	8	4	5
	IMB3_HUMAN	12H			7	8		6
	IMB3_YEAST	12H			11	10	4	9
	IMB4_SCHPO	12H			7	8		6
	IMB4_YEAST	12H			7	12		
	CSE1_YEAST							
	CAS_HUMAN							
	CRM1_SCHPO							
	CRM1_YEAST							
	NMD5_YEAST							
	YGZ1_YEAST							
	YD43_SCHPO							

ARM		Standard annotation						
		sw	pfam	arm	aaa	imb	adb	H-old
	APC_HUMAN		4A	4				
ARM/	ARM_DROME	12.5A	6A	10				5
CTNB/	ARM_MUSDO	R	6A	10				5
PLAK	CTNB_HUMAN		5A	10				
	CTNB_MOUSE		5A	10				
	CTNB_TRIGR		7A	10				
	CTNB_URECA		8A	7				
	CTNB_XENLA		5A	10				
	PLAK_HUMAN		6A	7	6			
	PLAK_MOUSE		6A	9	6			
	PLAK_XENLA		7A	10	5			
	VAC8_YEAST		8A	9	6	5	7	5
	GDS1_BOVIN			3				
	GDS1_HUMAN			3	5			
IMA	IMA1_ARATH	8A	8A	7	5	5		5
	IMA1_HUMAN	8A	8A	7				
	IMA1_MOUSE	8A	8A	7				
	IMA1_SCHPO	8A	8A	8	6	6		5
	IMA1_XENLA	8A	8A	8	5	6		5
	IMA1_YEAST	8A	8A	8	6	5		
	IMA2_ARATH	8A	8A	8	6			
	IMA2_HUMAN	8A	8A	8	6	6	4	6
	IMA2_MOUSE	8A	8A	8	6	6	4	5
	IMA2_XENLA	8A	8A	8	6	6		
	IMA3_HUMAN	8A	8A	7	6			5
	IMA3_MOUSE	8A	8A	8	6			5
	IMA4_HUMAN	8A	8A	7	6	6		5
	IMA4_MOUSE	8A	8A	7	6	6		5
	IMA6_HUMAN	8A	8A	8				4
	IMA6_MOUSE	8A	8A	8		6		
	IMA_CAEEL	8A	8A	8		5		
	IMA_DROME	10A	8A	8	5			
	IMA_LYCES	8A	8A	8				
	P120_MOUSE		4A	4				
	P115_RAT							
	P115_BOVIN							
	MOT1_YEAST	6T			6			6

Sequences are grouped as in the clusters and bubbles shown in Figure 2. The current annotation in SwissProt (standard) is compared to the results of this (arm, aaa, imb, adb) and previous work (H-old, Andrade *et al.*, 2000) in terms of protein numbers and repeats per protein. Sw column: repeats described in the SwissProt entry (H, HEAT, A, ARM, R, keyword "REPEATS" present but no other information, T, TPR). Pfam column: repeats detected by pfam as indicated in the corresponding SwissProt entry (A, ARM, ADA, adaptin). arm, aaa, imb, and adb columns: repeats detected by REP using the corresponding profiles. H-old column: repeats detected by REP using a unique profile for HEAT repeats (Andrade *et al.*, 2000).

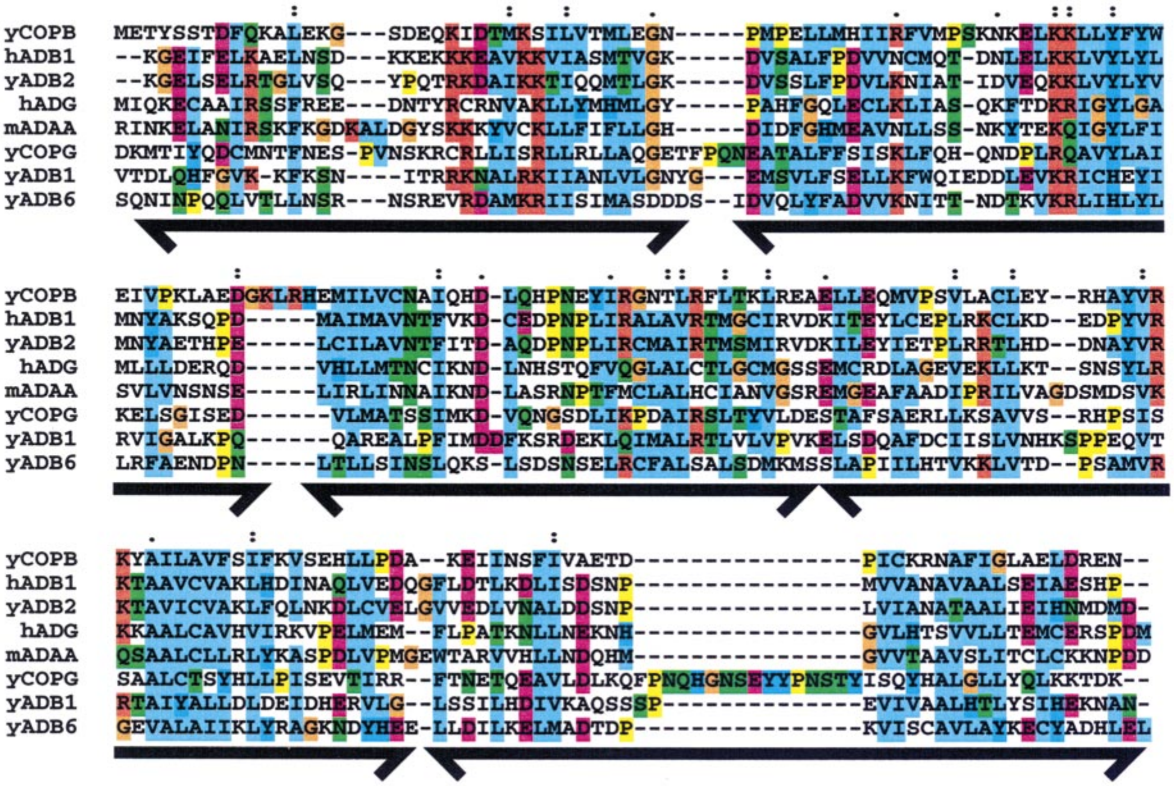
yCOPB METYSSTDFQKALEKG---SDEQKIDTMKSILVTMLEGN-----PMPELLMHIIRFVMPSKNKELKKLLYFYW
hADB1 --KGEIFELKAELNSD---KKEKKKEAVKKVIASMTVGK-----DVSALFPDVVNCMQT-DNLELKKLVYLYL
yADB2 --KGELSELRTGLVSQ---YPQTRKDAIKKTIQQMTLGK-----DVSSLFPDVLKNIAT-IDVEQKKLVYLYV
hADG MIQKECAAIRSSFREE---DNTYRCRNVAKLLYMHMLGY-----PAHFGQLECLKLIAS-QKFTDKRIGYLGA
mADAA RINKELANIRSKFKGDKALDGYSKKKYVCKLLFIFLLGH-----DIDFGHMEAVNLLSS-NKYTEKQIGYLFI
yCOPG DKMTIYQDCMNTFNES-PVNSKRCRLLISRLLRLLAQGETFPQNEATALFFSISKLFQH-QNDPLRQAVYLAI
yADB1 VTDLQHFGVK-KFKSN----ITRRKNALRKIIANLVLGNYG---EMSVLFSELLKFWQIEDDLEVKRICHEYI
yADB6 SQNINPQQLVTLLNSR---NSREVRDAMKRIISIMASDDDS--IDVQLYFADVVKNITT-NDTKVKRLIHLYL
yCOPB EIVPKLAEDGKLRHEMILVCNAIQHD-LQHPNEYIRGNTLRFLTKLREAELLEQMVPSVLACLEY--RHAYVR
hADB1 MNYAKSQPD-----MAIMAVNTFVKD-CEDPNPLIRALAVRTMGCIRVDKITEYLCEPLRKCLKD--EDPYVR
yADB2 MNYAETHPE-----LCILAVNTFITD-AQDPNPLIRCMAIRTMSMIRVDKILEYIETPLRRTLHD--DNAYVR
hADG MLLLDERQD-----VHLLMTNCIKND-LNHSTQFVQGLALCTLGCMGSSEMCRDLAGEVEKLLKT--SNSYLR
mADAA SVLVNSNSE-----LIRLINNAIKND-LASRNPTFMCLALHCIANVGSREMGEAFAADIPRILVAGDSMDSVK
yCOPG KELSGISED------VLMATSSIMKD-VQNGSDLIKPDAIRSLTYVLDESTAFSAERLLKSAVVS--RHPSIS
yADB1 RVIGALKPQ------QAREALPFIMDDFKSRDEKLQIMALRTLVLVPVKELSDQAFDCIISLVNHKSPPEQVT
yADB6 LRFAENDPN-----LTLLSINSLQKS-LSDSNSELRCFALSALSDMKMSSLAPIILHTVKKLVTD--PSAMVR
yCOPB KYAILAVFSIFKVSEHLLPDA--KEIINSFIVAETD----------------PICKRNAFIGLAELDREN--
hADB1 KTAAVCVAKLHDINAQLVEDQGFLDTLKDLISDSNP----------------MVVANAVAALSEIAESHP--
yADB2 KTAVICVAKLFQLNKDLCVELGVVEDLVNALDDSNP----------------LVIANATAALIEIHNMDMD-
hADG KKAALCAVHVIRKVPELMEM--FLPATKNLLNEKNH----------------GVLHTSVVLLTEMCERSPDM
mADAA QSAALCLLRLYKASPDLVPMGEWTARVVHLLNDQHM----------------GVVTAAVSLITCLCKKNPDD
yCOPG SAALCTSYHLLPISEVTIRR--FTNETQEAVLDLKQFPNQHGNSEYYPNSTYISQYHALGLLYQLKKTDK--
yADB1 RTAIYALLDLDEIDHERVLG--LSSILHDIVKAQSSSP--------------EVIVAALHTLYSIHEKNAN-
yADB6 GEVALAIIKLYRAGKNDYHEE-LLDILKELMADTDP----------------KVISCAVLAYKECYADHLEL

Figure 3. Alignment of selected members of the ADB class. The phasing of the ADB repeats identified is indicated with bold arrows. The similarity to HEAT and ARM repeats has been mentioned briefly by Andrade & Bork, (1995) and Küssel & Frasch (1995), respectively, however, the current analysis illustrates the significance of the repeats and their boundaries for the first time.

Specificity of profiles and relation between repeat classes

Use of the four profiles instead of only two (ARM plus old HEAT) increased both the sensitivity and selectivity of the identification of family members. For instance, the total number of HEAT repeats detected increased by roughly 50 % with respect to the previous profile used (Table 1). The new profiles have been included in the REP and SMART servers (http://www.embl-heidelberg.de/ ~ andrade/papers/rep, Andrade *et al*., 2000; http://www.embl-heidelberg.de/SMART/, Schultz *et al*., 2000). Table 1 shows that the ARM profile is 100 % specific for the ARM proteins, i.e. only ARM-containing proteins are recognized by this profile, but that the AAA and IMB profiles are not very specific and contain significant overlap. Comparatively, the ADB profile is quite specific, leading us to conclude that ADB is a separate class of repeats distinct from the ARM, IMB and AAA classes (Table 1, Figure 2).

The distribution of repeats detected using the sequence profiles is shown in Figure 4 for a representative sequence of each class. Of the four examples shown, the best results were obtained for yeast importin α (yIMA1), known from the crystal

Table 2. Number of proteins containing ARM or HEAT repeats detected in genomic sets

Protein set	ARM	AAA	ADB	IMB	HEAT-new[a]	HEAT-old[b]
S. cerevisiae	2 (0.3)	11	4	5	20 (3.2)	12
C. elegans	3 (0.2)	5	5	2	12 (0.6)	8
D. melanogaster	9 (0.6)	12	4	2	18 (1.3)	14
H. sapiens[c]	16 (1.4)	15	7	6	28 (2.4)	22

The number of proteins per 1000 (o/oo) of the total genomic set is shown between parentheses. REP thresholds used (*P*-value, minimum repeat number; Andrade *et al*., 2000) were: ARM 10^{-8}, 3; HEAT-old 10^{-6}, 4; AAA 10^{-5}, 5; ADB 10^{-8}, 4; IMB10^{-6}, 5. Average o/oo of ARM and HEAT repeats per organism were 0.6 and 1.5, respectively.

[a] Sum of AAA, ADB, and IMB.

[b] Unique profile defined previously (Andrade *et al*., 2000).

[c] Subset of the almost complete genome with sequences of reduced redundancy (no pair with more than 97 % of identical amino acid residues).

Table 3. Definition of the protein fragments used for compiling Figure 2

Class	Subclass	*P*-value	3D	Leader	Fragment
IMB	sIMB4/yIMB4	e-58		yIMB4	159-633
	yIMB3/hIMB3	e-59		hIMB3	212-979
	IMB2	e-105	+	hIMB2	9-700
	yCSE1/hCAS	e-134		hCAS	5-773
	yYGZ1/sYD43	0		sYD43	1-785
	IMB1	e-102	+	hIMB1	124-726
AAA	HD	0		mHD	183-920
	yYM8P/sYAV9	0		sYAV9	393-1102
	yGCN1/sYAQ5	0		sYAQ5	1062-2090
	TOR/FRAP	e-105		yTOR1	627-1147
	2AAA	e-95	+	h2AAA	5-552
	c2AAA	-		c2AAA	84-664
	hCTOG	-		hCTOG	160-1051
	EF3	e-49		yEF3A	85-278
	yVP15	-		yVP15	418-730
ADB	COPB	e-111		yCOPB	55-563
	COPG	e-100		yCOPG	61-564
	ADB	e-71		hADB1	61-564
	ADA/ADAC/ADG	e-52		hADG	55-563
	yADB6	-		yADB6	77-601
	yYBD7	-		yYBD7	91-657
	dGARN	-		dGARN	103-467
ARM	mP120	-		mP120	398-733
	ARM/CTNB/PLAK	0	+	hPLAK	131-611
	hAPC	-		hAPC	341-731
	rP115/bP115	0		bP115	1-651
	IMA	e-95	+	hIMA2	109-449
	yVAC8	-		yVAC8	81-450
	bGDS1/hGDS1	e-142		hGDS1	79-341
	yMOT1	-		yMOT1	356-979

P-value: minimum *P*-value of the sequence similarity search from the leader of a subclass to any other member in the subclass.

structure to contain ten ARM repeats. Our profile search successfully identified eight of these and detected two additional repeats, although these were classified as ADB and AAA repeats. The search performed less well for human importin β1 (hIMB1) correctly identifying only about half of the 19 HEAT repeats observed in the 3D structure. Several other repeats were identified in this sequence, but were mis-classified and/or out of phase with the observed structural motif, typically because the structural repeat was unusually long or had the insertion of an α-helix. Similarly, for the

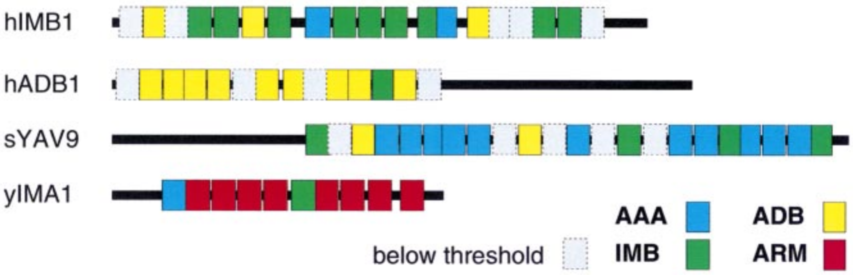

Figure 4. Distribution of repeats in four protein sequences and similarity to a repeat class. The protein sequences hIMB1, hADB1, sYAV9, and yIMA1 were clustered in the IMB, ADB, AAA, and ARM classes, respectively (see Figure 2). Repeats are represented as boxes whose colour indicates the most similar repeat profile: green, IMB; yellow, ADB; blue, AAA; red, ARM. Similarity of each repeat to a profile was evaluated with the alignment program SearchWise (Birney *et al*., 1996). The similarity score was normalized by a factor that accounted for the average score value obtained for repeats of each class: ADB, 1526; AAA, 1784; IMB, 1899; ARM, 2495. Each repeat was assigned to the class corresponding to the profile giving the highest normalized score and above a threshold of 0.6. Missing repeats were either known from 3D structures (hIMB1, yIMA1) or estimated by manual analysis (hADB1, sYAV9). There is an agreement between protein class and repeat class (e.g. most of the repeats in yIMA1, classified in the ARM class, are also assigned to ARM). However, some heterogeneity of repeat classes inside each sequence indicates the similarity between the four repeat classes. The cause are key residues that apparently can change within each protein sequence. Missing repeats (grey colour) reflect the limits of repeat detection by sequence similarity.

ADB and AAA examples, most of the repeats detected were specific for the corresponding class; however, not all the repeats detected belong to a single class, and regions of sequence occurring between two detected repeats, which likely also contain a repeat, fall below the detection threshold used.

Another fact that complicates repeat detection is that in a repeat ensemble, the terminal repeats have different constraints in packing and hydrophobicity. This results in larger divergence from the consensus and hampers detection. Examples are hIMB1 and hADB1 (Figure 4).

These examples illustrate some of the difficulties associated with detecting and classifying repeats. The fact that repeats of different classes are detected in the same sequence reflects the similarity of the repeat classes, but could also reflect independent evolution of repeats causing a switch of repeat class by the mutation of a few residues.

Relationship between repeats of known structure

In order to identify structural features that could support our sequence-based classification scheme, we examined the structures of individual ARM/HEAT repeats to see if these also clustered into distinct groups. We structurally aligned the individual repeats from importin α, β-catenin, importin β1 and the PR65/A subunit of PP2A, using only those residues for which the topological equivalence was unambiguous across all repeats. Thus, we aligned the ARM H2 helix with the upper part of HEAT helix A, and ARM helix H3 with HEAT helix B, but excluded a number of residues in the turns between helices, as well as the entire ARM helix H1, as its correspondence with residues of HEAT helix A was unclear. We then examined the structural differences between all repeats by principal component analysis. The matrix of RMS distances between all possible pairs of structures was computed, and the eigenvalues of the data distribution were calculated using the metric matrix distance geometry method implemented in X-PLOR (Kuszewski *et al.*, 1992). These data were projected onto the plane generated by the first two eigenvectors, which describe the directions of maximum spread of the data (Figure 5). The distribution indicates two trends; first, there is a well-defined cluster with members of all structures, indicating the existence of a canonical repeat structure common to all ARM/HEAT proteins; second, there are two other more loosely defined clusters: one with

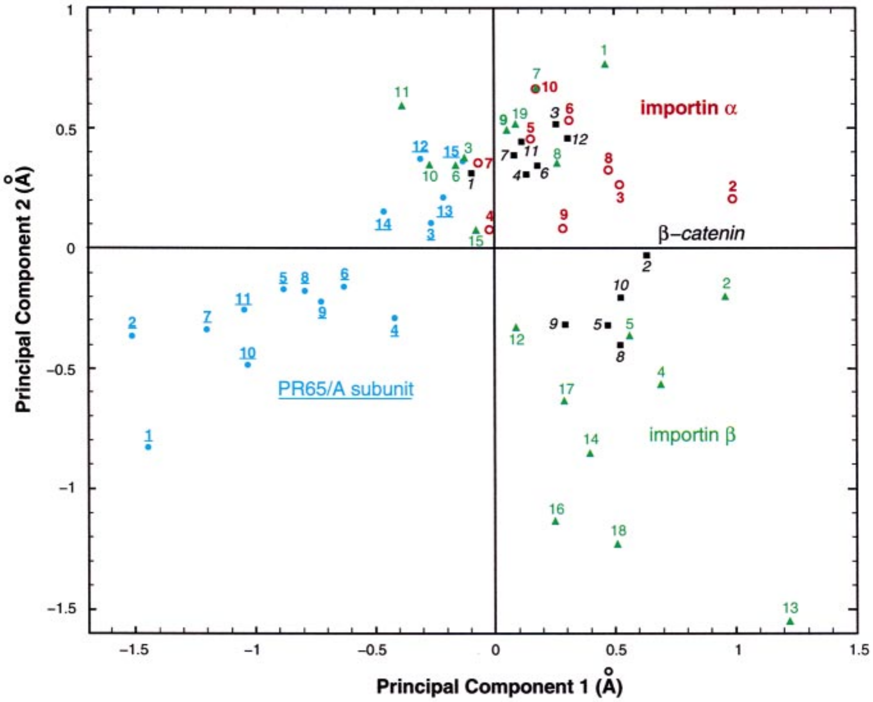

Figure 5. Principal component analysis of repeat structures. An "all against all" comparison of individual repeat structures was performed. From the resulting table of RMS distances, metric matrix distance geometry was used to find the plane of maximum separation. The Figure shows the data projection onto this plane, with each point representing the structure of one repeat. The color and shape of the symbol indicates protein of origin, and the numbers refer to the various structural repeats (some repeats are severely truncated and were not included in the analysis). Repeats were taken from the following structures: (●) PR65/A subunit (y2AAA, PDB:1b3u, Groves *et al.*, 1999); (▲) importin-β (yIMB1, PDB:1qgk, Cingolani *et al.*, 1999); (■) importin α (hIMA1, PDB:1bk5, Conti *et al.*, 1997); (○) β-catenin (mCTNB, PDB:2bct, Huber *et al.*, 1997). The Figure shows that there is a core structure found in all four structures with outliers from the PR65/A subunit and the importin-β falling in separable regions.

repeats from PP2A (left-lower quadrant), and another with repeats from IMPB and BCAT (right-lower quadrant). This result points out the variability of repeat structures within a protein structure. However, it does not reveal any obvious correspondence between the structural clusters and the sequence clusters. We therefore concentrate in the following section on conserved residues that might cause the sequence clustering.

Comparison of individual repeat sequences

In order to identify residues important for discriminating the four repeat classes, we compared the four sets of repeat alignments used to generate the sequence profiles for each class. The comparison was greatly facilitated by representing each alignment as a sequence logo, a convenient method to display the information content and relative frequency distribution of residues within a sequence alignment (Schneider & Stephens, 1990). Sequence alignments of 461 ARM, 94 IMB, 143 AAA, and 74 ADB repeats were used to generate the four sequence logos shown in Figure 6. These logos were then aligned to be consistent with the structural alignment of repeats, or in the case of the ADB class, to maximize similarity with the other sequence logos. This approach readily allowed us to identify features shared among the four classes as well as those distinct to a subset of classes.

Features common to ARM and HEAT repeats: hydrophobic core and Proline 11

The four sequence classes share seven highly conserved hydrophobic residues, located at positions 10, 13, 17, 24, 28, 32, and 35 (highlighted in yellow in Figure 6). The consensus residue at these positions is Leu, except at positions 24 and 28 where it is Val and Ala, respectively. In the known crystal structures, the seven residues are located on the buried face of either helix A (ARM helix H2) or helix B (ARM helix H3), and form the hydrophobic core of the repeat (Figure 7). Four of these residues have little or no contact with the preceding ($N-1$) or following ($N+1$) repeat: Ala28 on helix B (which is surrounded by residues 13, 16, and 17 from helix A) and residues Leu13, Leu32 and Leu35 (which interact with one another and with residues 10, 17 and 28). In contrast, the remaining three residues have significant inter-repeat interactions: Leu10 and Leu17 are both located on the same face of helix A and interact with several residues (4, 5, 9, 12, and 16) of repeat $N+1$, while Val24 on helix B faces the opposite direction and contacts residues 22 and 26 of repeat $N-1$. In general, the set of intra-repeat interactions for the seven residues is similar among the known repeat structures. A notable exception involves Leu13, which interacts primarily with Leu32 in ARM and IMB repeats, but with Leu35 in AAA repeats (compare green circles in Figure 7). In contrast, the set of inter-repeat interactions varies greatly among classes, and often even within a single class, reflecting different relative orientations between adjacent repeats.

Another highly conserved residue common to the ARM, AAA, and IMB alignments is Pro11 (highlighted in grey in Figure 6). In the ARM repeat, Pro11 acts as a helix breaker at the N terminus of helix H2, favouring the abrupt turn between the H1 and H2 helices (Figure 7). It plays a similar role in IMB and AAA-type HEAT structures, where it introduces a kink or a bend in helix A. Interestingly, a proline residue occurs rarely at position 11 in the ADB alignment (see below).

The four sequence logos also resemble one another at less strictly conserved positions. For example, residue 14 is usually hydrophobic, residues 31 and 36 are often Ala, and residues 15, 18, 21, 23 and 24, which are solvent-exposed on the external faces of helices A and B (H2 and H3), are almost always hydrophilic.

Specific features of the ARM repeat

In general, the ARM repeats are more highly conserved than the other three classes. This is particularly true at positions 4, 8, 16, and 34 (Figure 6). The higher degree of conservation at positions 4 and 8 reflects the presence of two helices (H1 and H2) in the ARM motif rather than one (helix A) in the HEAT motif. ARM residue 4 is a conserved Val, Ile, or Leu residue in the middle of helix H1 which fits into the ridge formed by residues 32, 35, and 36 on helix H3 of repeat $N-1$. It also contacts residue 35 within the same repeat, thus resulting in a ladder of van der Waals contacts involving residues 4 and 35 from successive ARM repeats. In contrast, residue 4 is a hydrophilic residue in the ADB repeats, and highly variable in the AAA and IMB repeats, consistent with its location at the bottom of helix A, either exposed to the solvent or in contact with diverse residues from the preceding repeat. Residue 8 is a strongly conserved Gly in the ARM repeats, located at the C terminus of helix H1. Gly8 has backbone dihedral angles in the α_L region of the Ramachandran plot ($\phi = 75^\circ$, $\psi = 25^\circ$), permitting the sharp bend between helices H1 and H2. In contrast, position 8 is highly variable in the HEAT repeats, playing no such specialized role.

Residue 16 is a highly conserved hydrophobic residue (usually Leu) on helix H2 of the ARM motif, but is less conserved among AAA repeats and least conserved among IMB repeats. Because this residue projects towards the preceding repeat, its local environment depends on the angle between repeats N and $N-1$. Among ARM repeats, this angle is nearly constant, and hence residue 16 interacts with the same residues (positions 17, 25, and 29 of repeat $N-1$) in most cases. In contrast, the inter-repeat angle, and hence the local environment of residue 16, is more variable among AAA repeats and most variable among

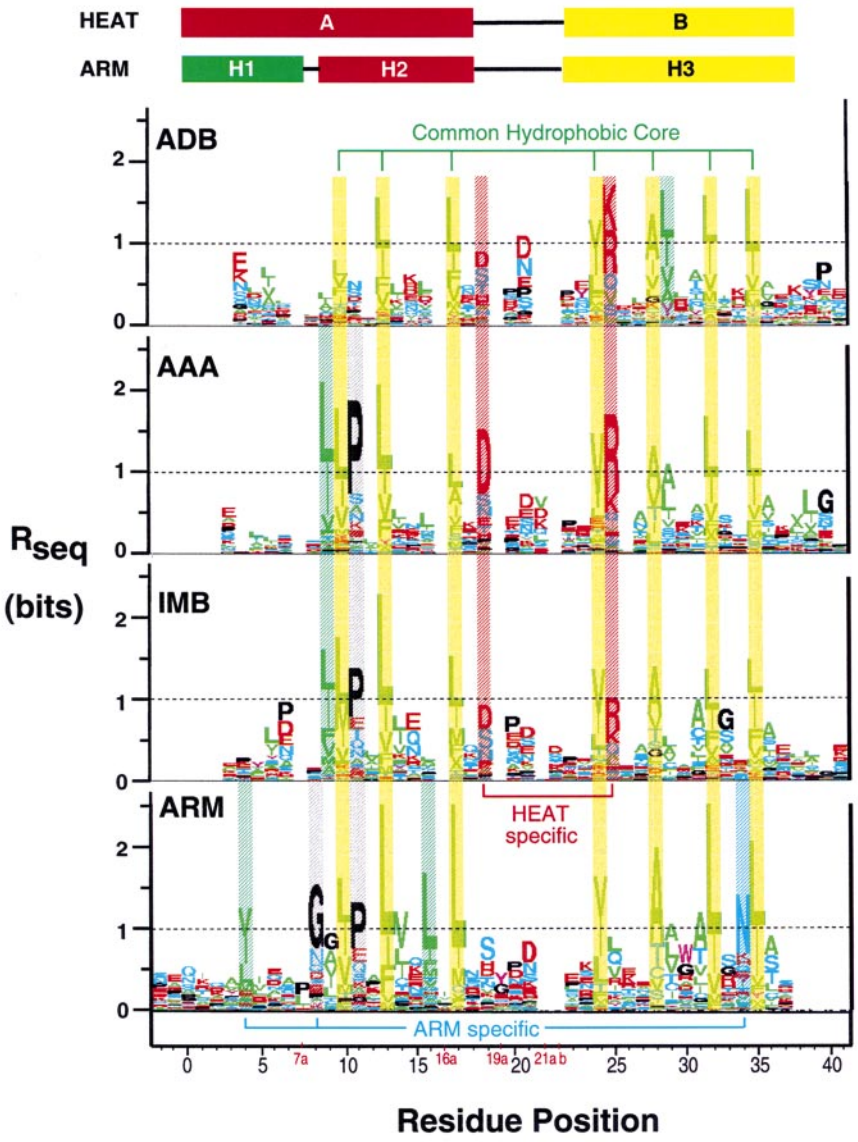

Figure 6. Sequence logos generated from alignments of ARM and HEAT repeats. Logos were generated using the WebLogo server (http://www.bio.cam.ac.uk/seqlogo/). The ARM, IMB, and AAA sequence logos were aligned according to a structural alignment of repeat units from importin-α, importin-β, and the PR65/A subunit of PP2A. The ADB sequence logo was aligned so as to maximize the similarity with the other profiles. Residues within a repeat are numbered according to the convention describing the consensus HEAT motif for both PR65/A and importin-β. Five positions occurring as infrequent insertions in the various alignments are labelled as 7a, 16a, 19a, 21a, and 21b. The letter plots represent amino acid conservation at each position. Residues occurring at each position are shown by their one-letter code, stacked from bottom to top in order of increasing frequency, with the size of each character proportional to its frequency at the position. The height of a column indicates the total information content of the aligned sequences at that position, R_{seq} (measured in bits). R_{seq} can range from 0, when all 20 residues are equally frequent (the alignment gives no information), to a maximum of 4.32 ($=\log_2 20$), when a single residue is invariant (no uncertainty at that position) (Schneider & Stephens, 1990). Positions with significant informational content ($R_{seq} > 1$ bit) have been highlighted. Colour code used was: green, hydrophobic amino-acids; red, charged; black, turn-like; magenta, aromatic; blue, polar.

IMB repeats, explaining the observed differences in sequence conservation.

Residue 34 is a strongly conserved hydrophilic residue among ARM repeats, most frequently an Asn, and is solvent exposed on the external face of helix H3. Because a variety of hydrophilic residues could be accomodated stereochemically at this location, the high degree of sequence conservation likely reflects an important functional role. Indeed, in importin-α, Asn34 residues are involved in substrate recognition, forming bidentate hydrogen bonds with the main-chain amide groups of the NLS peptide (Conti *et al*., 1998). Also, in β-catenin (functionally distinct from importin-α), five of the

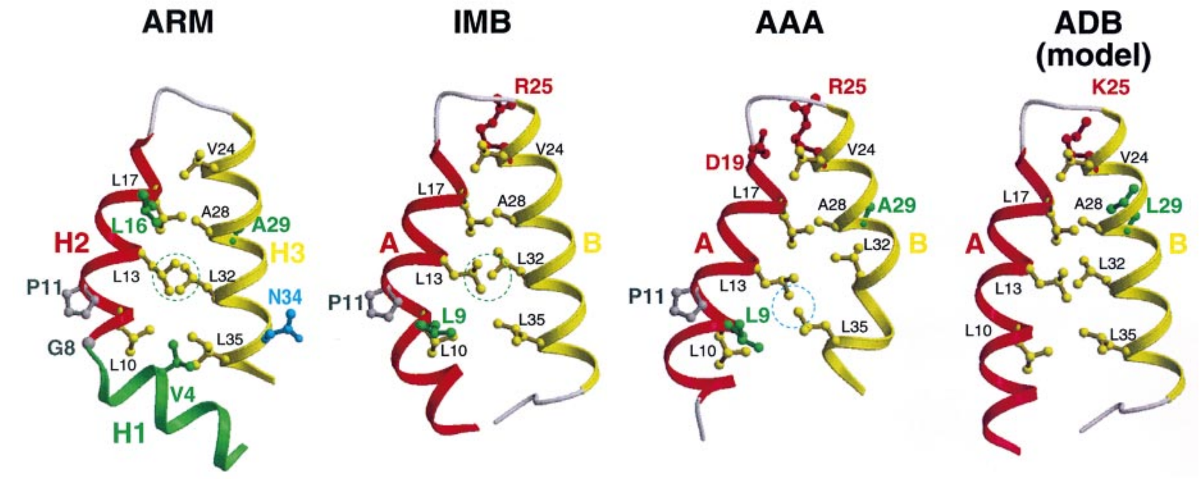

Figure 7. Representative ARM and HEAT repeat structures. The three structures on the left are those of repeats ARM-7, HEAT-10 and HEAT-8 of importin-α, importin-β, and the PR65/A subunit of phosphatase 2A, respectively. The ADB structure is a hypothetical model for illustrative purposes only. Residues shown in ball-and-stick representation are those highlighted in Figure 6, numbered and coloured according to the same scheme. The view is such that the preceding repeat, $N-1$, stacks in front of the page and the subsequent repeat, $N+1$, behind the page.

12 repeats have an Asn residue at position 34, and two have the similar residues His and Gln. In contrast, position 34 is highly variable (and rarely asparagine, <7%) in the AAA, IMB and ADB repeats, suggesting fewer functional interactions involving this residue.

Features of the IMB and AAA classes

A feature which distinguishes HEAT repeats from ARM repeats is the presence of consensus residues Asp19 and Arg/Lys25, which are most prevalent in the AAA alignment (highlighted in red in Figure 6). In the structure of the PP2A PR65/A subunit (a member of the AAA class), these residues form a ladder of hydrogen bonds extending across multiple repeats, involving the guanidinium group of Arg25 and the carboxylate and main-chain carbonyl groups of Asp19 (Groves *et al*., 1999). The two residues are somewhat less well conserved in the IMB alignment, consistent with the absence of the Asp-Arg ladder in the structures of IMB class members importins β1 and β2. Nevertheless, their relatively high frequency may reflect a recent divergence of the IMB class from AAA-type structures.

Another characteristic of the AAA and IMB-alignments is a strongly conserved hydrophobic residue at position 9 (Figure 6). This residue is on the bottom half of helix A, and typically interacts with conserved hydrophobic residues 10, 14, 32, and 36 of repeat $N-1$, and to a lesser extent with residues on helix B of repeat N. In contrast, this position is poorly conserved in both the ADB (see below) and ARM repeats. In ARM structures residue 9 is located at the junction of the H1 and H2 helices and is either small (Gly or Ala), as is typically the case in β-catenin, or a hydrophobic residue buried in the core, as is frequently observed in the repeats of importin-α.

Despite the clear separation of the IMB and AAA classes during the sequence comparison step shown in Figure 2, the corresponding sequence logos are remarkably similar and likely account for the high degree of overlap when the two profiles are used for repeat detection (Table 1). Thus, the IMB and AAA classes appear more closely related to each other than to either the ARM or ADB classes. Nonetheless, in addition to positions 19 and 25 being less well conserved in the IMB repeats than in the AAA repeats, other differences exist between the two sequence logos. For example, a proline at position 7 and a small residue (Gly, Ala or Ser) at position 33 are moderately conserved in the IMB repeats but not in the AAA repeats; while an aliphatic residue at position 39 and a Gly at position 40 occur frequently in the AAA repeats but not in the IMB repeats. However, these differences are evidently too weak to completely discriminate between the two classes.

Features of the ADB repeat

Although no 3D structure has been determined for the ADB class, the conservation of the seven core hydrophobic residues suggests that the structure of the individual ADB repeat closely resembles that of the ARM, AAA, and IMB repeats. Three features suggest a closer relationship to the AAA and IMB classes than to the ARM class. First, the absence of a consensus Gly8 residue and of a conserved hydrophobic residue at position 4 suggests that ADB-type repeats have a single A helix like the AAA and IMB-type repeats, rather than H1 and H2 helices as in the ARM repeats. Second, the consensus Asn34 residue characteristic

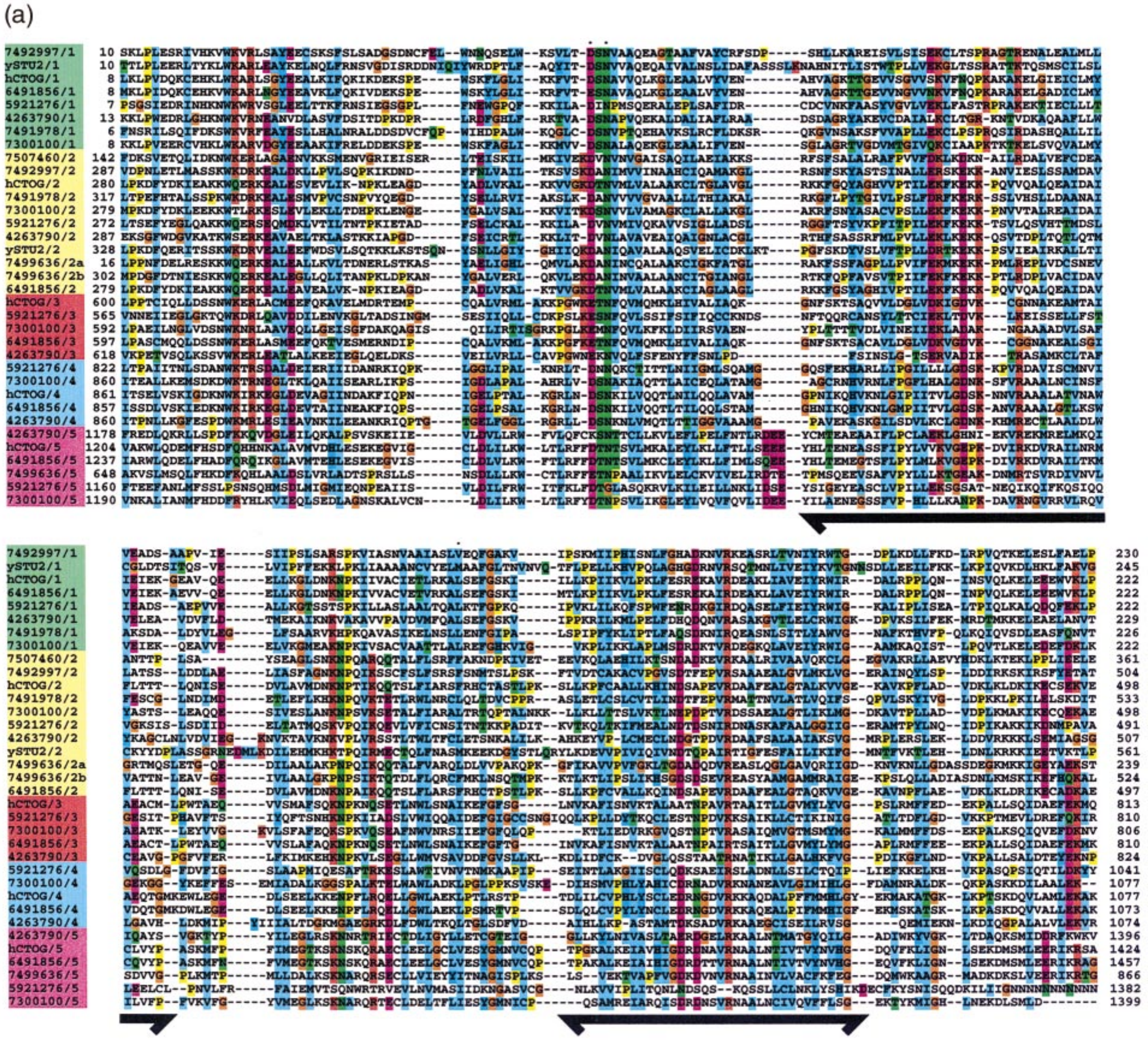

Figure 8 (*legend shown on page 14*)

of the ARM alignment is missing from the ADB alignment. Third, like the AAA and IMB repeats, ADB repeats have the signature Asp and Arg/Lys residues at positions 19 and 25. As in the IMB repeats, Asp19 is more poorly conserved than Arg/Lys25, suggesting the absence of an extended hydrogen-bonding network.

Interestingly, the absence of a consensus Pro residue at position 11 is unique to the ADB repeats, and suggests that helix A may be less distorted than in the AAA and IMB structures. Distortions in the A and B helices of the IMB and AAA structures (and the sharp turn between helices H1 and H2 in the ARM repeat) place the C terminus of one repeat close to the N terminus of the next repeat, reducing the number of linker residues required to connect the two. A straighter A-helix would tend to increase this distance, and might be compensated in an ADB protein either by the use of more linker residues, by a sharper change in chain direction within or at the end of helix B, or by a significantly different relative orientation of neighbouring repeats. An increased number of linker residues is unlikely because consecutive ADB repeats are spaced somewhat closer together in sequence than are AAA repeats (average spacings of 37.2 *versus* 40.6 residues). On the other hand, a sharp change in chain direction may occur in certain ADB repeats immediately after helix B, as there is a moderately conserved proline residue at position 39 (Figure 6). Also, certain features in the sequence logos suggest that interactions between neighbouring repeats in the ADB class differ from those of the AAA and IMB class, and thus may reflect a different relative orientation of adjacent repeats. For example, residues 9 and 10 (which interact with repeats $N-1$ and $N+1$, respectively) are highly conserved leucines in the IMB

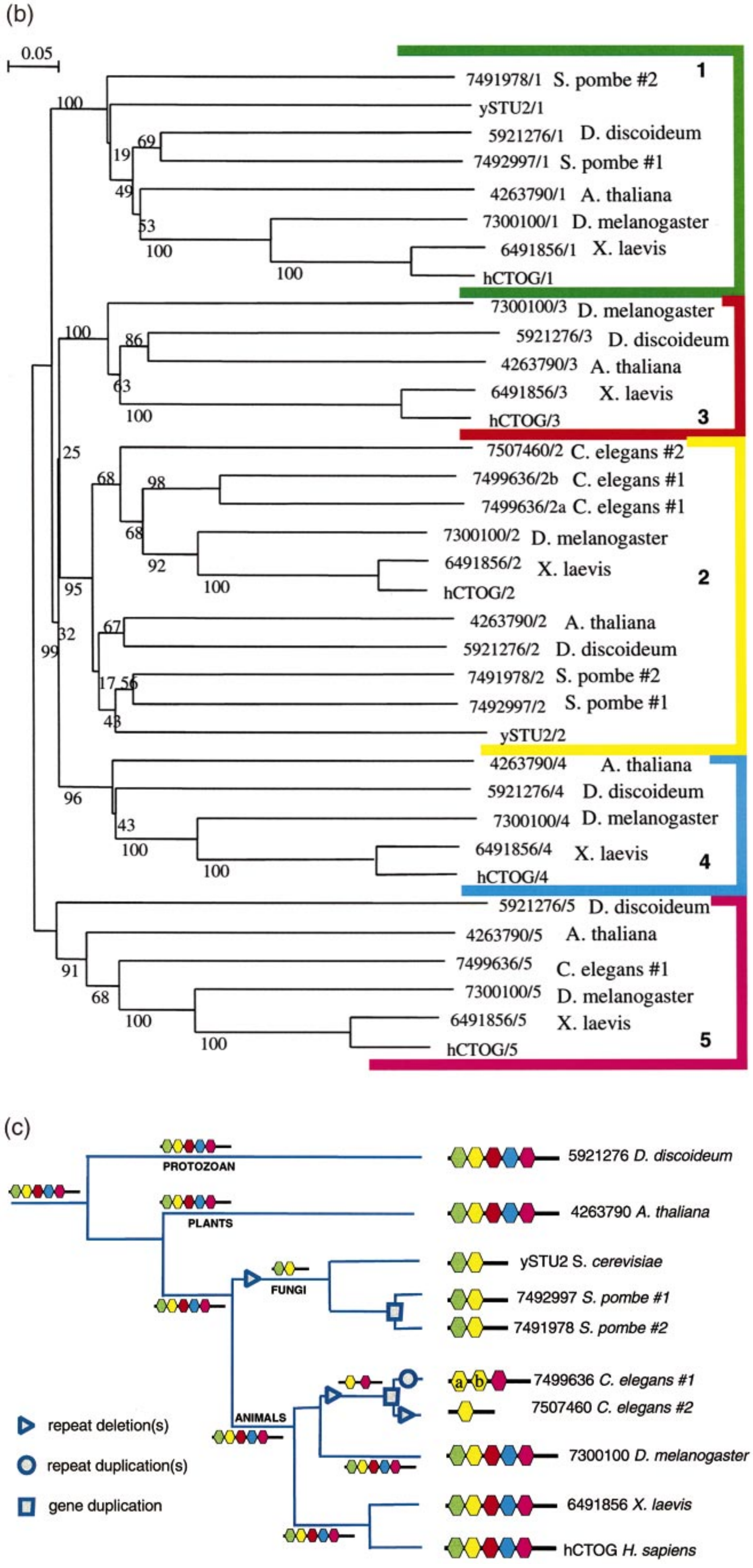

Figure 8 (*legend shown on page 14*)

and AAA repeats, but are poorly conserved among ADB repeats; whereas residue 29 (which projects from helix B toward the $N + 1$ repeat) is a highly conserved leucine in the ADB repeats but is poorly conserved among IMB and AAA repeats. Clearly, accurate knowledge of the structure of an individual ADB repeat and of the inter-repeat packing arrangement will require determination of the atomic structure of a representative protein of this class.

Evolution of repeats

Despite common features in sequence and structure, the evolution of ARM/HEAT repeats remains difficult to reconstruct. Repeats within one of the four classes described above are usually more similar to each other than to repeats of another class. This indicates an independent duplication of repeats after divergence of the four classes, possibly with the most recent common ancestor of the four protein classes consisting of only a single repeat. However, this idea raises two obvious problems. First, the role of modern ARM and HEAT proteins in mediating protein-protein interactions is carried out by an extended surface composed of a contiguous array of multiple repeats, and it is unlikely that such interactions could be encoded by an ancestral polypeptide composed of only one repeat. Second, all four classes of motif contain seven core hydrophobic residues which form the extensive surface mediating the interactions between neighbouring repeats (Figure 6). Exposed to the solvent, such a hydrophobic surface would likely render an ancestral, single-repeat polypeptide highly insoluble. In fact, both problems could be overcome if the ancestral protein formed homo-multimers, thereby burying all its core hydrophobic residues (except for those in terminal repeats) in the interface between monomers, and creating an extended surface for binding ligands. The ancestral protein could then have diverged (by gene duplication) into several single-repeat proteins, which subsequently became elongated (by internal replication of the motif), leading to the various classes of repeat proteins currently in existence. Indeed, such a homo-multimer hypothesis has been postulated for the origin of β-trefoil repeats (Ponting & Russell, 2000) and to explain the presence of proteins with only one leucine rich repeat in the receptor for von Willebrand factor (Kobe & Deisenhofer, 1994).

It is clear that the genesis of the entire repeat family is difficult to deduce from the sequence information presently available. However, to gain insight into the evolutionary process, we examined individual members of the four classes for evidence of more recent evolutionary events. We found one example, described below, which illustrates a complexity in the evolution of repeat motifs that involves not only duplications of one or several repeat units, but also various deletion and gene duplication events.

TOG family: an example of repeat evolution

One of the proteins in the AAA class identified as containing at least nine HEAT repeats is the human microtubule-binding, colonic and hepatic tumour related protein, CTOG (Charrasse *et al*., 1995, 1998). The protein has orthologous sequences of a similar length in *Xenopus laevis* (XMAP215; Tournebize *et al*., 2000), *Drosophila melanogaster* (Msps; Cullen *et al*., 1999), *Dictyostelium discoideum* (CP224; Graf *et al*., 2000) and *Arabidopsis thaliana*, but of much shorter length in *S. cerevisiae* (Stu2p; Wang & Huffaker, 1997), *Schizosaccharomyces pombe* (with two homologues, Alp14 and p93dis1; Radcliffe, 1998; Nabeshima *et al*., 1995) and *C. elegans* (with two putatively expressed homologues).

Comparison of the HEAT-repeat regions in the various sequences showed a surprising variety of repetitions of a ~250 residue cassette likely to contain six HEAT repeats (only two of which can be readily identified by a profile search). Five instances of the cassette appear to be present in CTOG and similarly sized orthologues, but only two in the shorter budding and fission yeast versions, and only either one or three instances in the *C. elegans* sequences. A series of events leading to

Figure 8. Duplication and deletion events in the Ctog/Stu2 proteins and homologues. (a) Alignment of a repeated cassette predicted to contain at least two (but more likely six as supported by secondary structure analysis) HEAT repeats. The identified repeats are denoted with bold arrows. Sequence identifiers are from GENBANK (only digits), or otherwise from SwissProt. Repetitions of the cassette are numbered from one to five according to their relative position in *Arabidopsis thaliana, Drosophila melanogaster, Dictyostelium discoideum, Xenopus laevis,* and *Homo sapiens* sequences. Other sequences from *Schizosaccharomyces pombe* (two), *Saccharomyces cereviasiae*, and *Caenorhabditis elegans* (two) show a lower number of cassette repetitions. (b) Phylogenetic tree corresponding to the alignment (bootstrapping values close to 100 indicate high branch stability). The five major branches correspond to the five repetitions (coloured and numbered accordingly). Note the duplication seen in *C. elegans* 7499636. Species names have been added after the species identifiers (with #1 or #2 to indicate multiple sequences in one species). (c) Proposed model for the evolution of cassettes in CTOG and orthologues. A likely sequence of events based on the phylogenetic tree that illustrates the complex evolution of the family from an original ensemble of five repetitions of the cassette (top left). The arrangement of repetitions seen in each sequence has been plotted in the tree of life of the corresponding species (following Philippe & Adoutte, 1998; Castresana, 2000): each cassette is represented as an hexagon using the previous colour coding. The lengths of the branches were chosen for ease of representation and have no phylogenetic meaning.

the early genesis of the cassettes are suggested by inspecting a dot plot of the larger sequences. For example, in the dot plot for human CTOG (data not shown) the sequence spanning cassettes 1 and 2 is highly similar to that spanning cassettes 3 and 4 (*E*-value of 2.8 e-20), suggesting that the four cassettes arose through the recent duplication of an earlier two-cassette sequence fragment. Cassettes 1 and 2 are moderately similar ($E = 0.0034$), as are 3 and 4 ($E = 0.17$), suggesting a somewhat earlier duplication event. Cassette 5 is least similar to the others ($E = 1.3$, 1.6, 220 with 1, 3 and 4) and thus likely arose through a still earlier duplication event.

The alignment and phylogeny of the cassette repetitions (Figure 8(a) and (b)) suggest how the ancestral 5-cassette protein subsequently evolved to its modern descendants in the various species. The shorter *S. cerevisiae* and *S. pombe* sequences contain only cassettes 1 and 2, and appear to have lost cassettes 3, 4 and 5 through one or more deletion events in a common ancestor (Figure 8(c)). Subsequently, after divergence of the two yeast species, a gene duplication event in *S. pombe* gave rise to the two sequences Alp14 and p93dis1. Similarly, *C. elegans* appears to have lost cassettes 1, 3 and 4 after its divergence from *Drosophila*; and following a gene duplication event, one of the genes lost cassette 5, while in the other cassette two were duplicated.

The model proposed in Figure 8(c) is a minimal sequence of events leading to the present day situation. We cannot exclude other events that increase the complexity of the evolution of this subfamily even further. This single example indicates that numerous duplication or deletion events have occurred independently in each of the four main repeat classes and probably even in each member of the classes. Differences in repeat evolution become also visible when comparing the divergence within different taxa. Given the current data set this makes it impossible to delineate a simple model for the evolution of the entire repeat super-family.

Evolution to the present day situation

The most parsimonious evolutionary model consistent with the data is that an ancestral repeat unit was duplicated, perhaps even functioning as a homo-multimer with a low copy number, and that further duplications formed intrinsically stable repeat-containing proteins, possibly the ancestors of the current four main classes. Alternatively, certain classes might have evolved later, separating merely because of functional constraints (e.g. specifically conserved residues due to Ran-binding or other functions).

The ability of different repeat ensembles to interact with other proteins might be ancient or might have been re-invented several times. In any case, the evolution of the ensemble was probably constrained at some point by the molecular partners. We note the following: (i) Elongation of the ensemble by repeat duplication is likely constrained by adaptation to the size of the partner. (ii) The overall shape of the multi-repeat super-helix may adapt to the shape of the partner by altering the angle between repeats, with a corresponding change in inter-repeat interactions. (iii) Special residues can be selected to match the physicochemical properties of the partner's surface to improve the specificity of the interaction. (iv) More complex changes could be selected to control conformational changes of the ensemble upon protein binding (this mechanism is proposed for importins and exportins where binding Ran·GTP provokes the release or the binding of the cargo protein in the nucleus, respectively; Mattaj & Englmeier, 1998). Steps (ii)-(iv) lead to the divergence of repeats within the same sequence (making it difficult to trace the evolution of the whole family). The only properties that must remain are the secondary structure and the hydrophobic core, already present in the ancestral unit. On top of these loose constraints there is an extreme divergence of the terminal repeats which alter their structure in order to shield the hydrophobic core of the internal repeats from the solvent. Those repeats are hardly detectable by sequence analysis and can diverge greatly in structure. The example of the ChTOG/Stu2p family illustrates that even within orthologous sequences a variety of duplication and deletion events can lead to further divergence and probably functional differences.

Conclusions

Our sequence analysis of protein fragments containing ARM and HEAT repeats suggests a separation into four main classes. Two of the classes correspond to well-defined protein families with common functional features: the clathrin-associated adaptor family and the Ran-binding family of nucleocytoplasmic transport proteins. The third class contains proteins with armadillo repeats, and the fourth is the most heterogeneous, with several subfamilies similar to the PR65/A subunit of PP2A. This classification is supported by the fact that profiles derived from each class are more sensitive and more selective than profiles derived from the entire super-family. Although the detection of ARM and HEAT repeats with the automatic method used here is still far from perfect (see, for example Figure 4), more proteins and repeat instances therein can now be identified (Table 3).

Structural analysis indicates that most of the repeats share detailed features related to similar super-secondary structures. The hydrophobic repeat core also appears to be highly conserved and is (probably together with repeat length) evidence for a common ancestry for all ARM and HEAT repeats. Each of the four repeat classes has key residues that are mainly responsible for the distinction at the sequence, but also at the structural level. This includes, for example, the absence in the ADB class of a Pro11 residue, which probably indicates a straighter helix A; the presence in the AAA class of charged residues at positions 19 and 25, forming a ladder of electrostatic inter-repeat interactions; the presence in α importins of an Asn34 residue involved in substrate recognition; and the conserved Gly8 in the ARM class facilitating the turn between helices H1 and H2.

One might argue that ARM and HEAT repeats have similar sequences simply because they share similar structures, arrived at by convergent evolution; however, the overall sequence identity between these repeats (about 13 %) is significantly higher than that expected for phylogenetically unrelated proteins with similar structures (8.5 % according to Rost, 1997). Based on our sequence and structural analysis, we postulate a common origin for ARM and HEAT repeats. However, due to the great divergence and the apparently complex evolution of repeats, we cannot resolve the order of early events which gave rise to the different classes of repeat proteins. In one group of orthologous sequences, the ChTOG/Stu2p family, we show by phylogenetic tree analysis that multiple gene duplication, repeat duplication and various deletion events can occur after a repeat ensemble is established. Such a complex series of events indicates that various functional adaptations of the proteins have occurred throughout their evolutionary history until recent times.

Although the methods for homology detection can certainly be further improved and although more sequence and structural data might give further insights into the evolution of the repeats, there are also conceptual limits in our current understanding of molecular evolution. For example, the term homology (like some other fundamental terms in molecular biology) is, on a closer look, not properly defined (see e.g. Doolittle, 2000; Fitch, 2000). We are proposing that the ancient repeats have a clear-cut homology relation, but since then repeats in each of the four families have undergone independent duplications leading to the present day architectures (Figure 4). When looking for protein similarities, we thus compared stretches of repeats (i.e. the first repeat in protein A with the first repeat in protein B, and so on), which is debatable. A proper analysis would have to concentrate on individual repeat units. These carry, however, too little signal to reveal significant relationships.

Despite all the limitations, we have given a systematic analysis of the expanding ARM/HEAT family, summarizing the common features and differences among its various members. This classification has allowed us to relate sequence features of these repeats to functional and structural properties. We can already predict the rough structure (an elongated super-helix of alpha-alpha repeats) and function (involvement in protein-protein interactions) for large parts of many proteins. Ultimately, this should be a first step towards a finer prediction of the structure and flexibility of these super-helical molecules from their sequence and of the concrete residues involved in the protein-protein contacts. This is very important since we cannot expect to obtain crystal structures of every ARM/HEAT repeat containing protein with all their different interacting factors. Understanding important biological processes such as nuclear transport, vacuolar transport, translation, and cytoskeleton organization, and diseases like Huntingtin, adenomatous polyposis, hepatomas and colonic tumours is at stake.

Acknowledgements

Thanks to Scott Kuersten and to the members of the Bork group for fruitful discussions. The work is supported by the BMBF and the DFG.

References

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

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

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

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

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

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.

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

Brodsky, F. M. (1997). New fashions in vesicle coats. *Trends Cell Biol.* **7**, 175-179.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**, 540-552.

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.

Charrasse, S., Schroeder, M., Gauthier-Rouviere, C., Ango, F., Cassimeris, L., Gard, D. L. & Larroque, C. (1998). The TOGp protein is a new human microtubule-associated protein homologous to the *Xenopus* XMAP215. *J. Cell Sci.* **111**, 1371-1383.

Chook, X. & Blobel, Y. (1999). Structure of the nuclear transport complex kanyopherin-beta2-Ran × GppNHp. *Nature,* **399**, 230-237.

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

Conti, E. & Kuniyan, J. (2000). Crystallographic analysis of the specific yet versatile recognition of distinct nuclear localization signals by Karyopherin alpha. *Structure Fold Des.* **8**, 329-338.

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 alpha. *Cell*, **94**, 193-204.

Cullen, C. F., Deak, P., Glover, D. M. & Ohkura, H. (1999). Mini spindles: a gene encoding a conserved microtubule-associated protein required for the integrity of the mitotic spindle in *Drosophila*. *J. Cell. Biol.* **146**, 1005-1018.

De Wergifosse, P., Jacques, B., Jonnuaux, J.-L., Purnelle, B., Skala, J. & Goffeau, A. (1994). The sequence of a 22.4 kb DNA fragment from the left arm of yeast chromosome II reveals homologues to bacterial proline synthetase and murine alpha-adaptin, as well as a new permease and a DNA-binding protein. *Yeast*, **10**, 1489-1496.

Doolittle, W. F. (2000). The nature of the universal ancestor and the evolution of the proteome. *Curr. Opin. Struct. Biol.* **10**, 355-358.

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

Fitch, W. M. (2000). Homology: a personal view on some of the problems. *Trends Genet.* **16**, 227-231.

Franke, W. W., Goldschmidt, M. D., Zimblemann, R., Mueller, H. M., Schiller, D. L. & Cowin, P. (1989). Molecular cloning and amino acid sequence of human plakoglobin, the common junctional plaque protein. *Proc. Natl Acad. Sci. USA*, **86**, 4027-4031.

Garcia-Barrio, M., Dong, J., Ufano, S. & Hinnebusch, A. G. (2000). Association of GCN1-GCN20 regulatory complex with the N terminus of eIF2alpha kinase GCN2 is required for GCN2 activation. *EMBO J.* **19**, 1887-1899.

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.

Görlich, D., Dabrowski, M., Bischoff, F. R., Kutay, U., Bork, P., Hartmann, E., Prehn, S. & Izaurralde, E. (1997). A novel class of RanGTP binding proteins. *J. Cell Biol.* **138**, 65-80.

Graf, R., Daunderer, C. & Schliwa, M. (2000). Dictyostelium DdCP224 is a microtubule-associated protein and a permanent centrosomal resident involved in centrosome duplication. *J. Cell. Sci.* **113**, 1747-1758.

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

Groves, M. R., Hanlon, N., Turowski, 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.

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

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

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

Kobe, B. & Kajava, A. V. (2000). When protein folding is simplified to protein coiling: the continuum of solenoid protein structures. *Trends Biochem. Sci.* **25**, 509-515.

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

Küssel, P. & Frasch, M. (1995). Pendulin, a *Drosophila* protein with cell cycle-dependent nuclear localization, is required for normal cell proliferation. *J. Cell Biol.* **129**, 1491-1507.

Kuszewski, J., Nilges, M. & Brünger, A. T. (1992). Sampling and efficiency of metric matrix distance geometry: a novel partial metrization algorithm. *J. Biomol. NMR*, **2**, 33-56.

Lloyd, V. K., Sinclair, D. A., Wennberg, R., Warner, T. S., Honda, B. M. & Grigliatti, T. A. (1999). A genetic and molecular characterization of the garnet gene of *Drosophila melanogaster*. *Genome*, **42**, 1183-1193.

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

Mattaj, I. W. & Englmeier, L. (1998). Nucleocytoplasmic transport: the soluble phase. *Annu. Rev. Biochem.* **67**, 265-306.

Nabeshima, K., Kurooka, H., Takeuchi, M., Kinoshita, K., Nakaseko, Y. & Yanagida, M. (1995). p93dis1, which is required for sister chromatid separation, is a novel microtubule and spindle pole body-associating protein phosphorylated at the Cdc2 target sites. *Genes Dev.* **9**, 1572-1585.

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

Philippe, H. & Adoutte, A. (1998). The molecular phylogeny of eukaryota: solid facts and uncertainties. In *Evolutionary Relationships Among Protozoa* (Coombs, G. H., Vickerman, K., Sleigh, M. A. & Warren, A., eds), Chapman & Hall, London, UK.

Ponting, C. P. & 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.

Radcliffe, P., Hirata, D., Childs, D., Vardy, L. & Toda, T. (1998). Identification of novel temperature-sensitive lethal alleles in essential beta-tubulin and nonessential alpha 2-tubulin genes as fission yeast polarity mutants. *Mol. Biol. Cell.* **9**, 1757-1571.

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.

Rost, B. (1997). Protein structures sustain evolutionary drift. *Fold. Design*, **2**, S19-S24.

Schneider, T. D. & Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucl. Acids Res.* **18**, 6097-6100.

Schultz, J., Copley, R. R., Doerks, T., Ponting, C. P. & Bork, P. (2000). SMART: a web-based tool for the study of genetically mobile domains. *Nucl. Acids Res.* **28**, 231-234.

Stack, J. H., Herman, P. K., Schu, P. V. & Emr, S. D. (1993). A membrane-associated complex containing the Vps15 protein kinase and the Vps34 PI 3-kinase is essential for protein sorting to the yeast lysosome-like vacuole. *EMBO J.* **12**, 2195-2204.

Tournebize, R., Popov, A., Kinoshita, K., Ashford, A. J., Rybina, S., Pozniakovsky, A., Mayer, T. U., Walczak, C. E., Karsenti, E. & Hyman, A. A. (2000). Control of microtubule dynamics by the antagonistic activities of XMAP215 and XKCM1 in *Xenopus* egg extracts. *Nature Cell Biol.* **2**, 13-19.

Traub, L. M. (1997). Clathrin-associated adaptor proteins - putting it all together. *Trends Cell Biol.* **7**, 43-46.

Traub, L. M., Downs, M. A., Westrich, J. L. & Fremont, D. H. (1999). Crystal structure of the alpha appendage of AP-2 reveals a recruitment platform for clathrin-coat assembly. *Proc. Natl Acad. Sci. USA*, **96**, 8907-8912.

Wang, P. J. & Huffaker, T. C. (0000). Stu2p: A microtubule-binding protein that is an essential component of the yeast spindle pole. *J. Cell. Biol.* **139**, 1271-1280.

Yamamoto, T., Kaibuchi, K., Mizuno, T., Hiroyoshi, M., Shirataki, H. J. & Takai, Y. (1990). Purification and characterization from bovine brain cytosol of proteins that regulate the GDP/GTP exchange reaction of smg p21 s, ras p21-like GTP-binding proteins. *Biol. Chem.* **265**, 16626-16634.

Edited by P. E. Wright

(Received 27 November 2000; received in revised form 5 March 2001; accepted 14 March 2001)