Published in final edited form as:

J Struct Biol. 2010 May ; 170(2): 354–363. doi:10.1016/j.jsb.2010.01.002.

INSIGHTS INTO THE DOMAIN AND REPEAT ARCHITECTURE OF TARGET OF RAPAMYCIN

Bruce A. Knutson

Fred Hutchinson Cancer Research Center Basic Sciences Division 1100 Fairview Ave N PO Box 19024, Mailstop A1-162 Seattle, WA 98109, USA

Abstract

A simple and efficient protein sequence analysis strategy was developed to predict the number and location of structural repeats in the TOR protein. This strategy uses multiple HHpred alignments against proteins of known 3D structure to enable protein repeats referenced from the 3D structure to be traced back to the query protein sequence by using user-directed repeat assignments. The HHpred strategy performed with high sensitivity by predicting 100% of the repeat units within a test set of HEAT- and TPR-repeat containing proteins of known three dimensional structure. The HHpred strategy predicts that TOR contains 32 tandem HEAT repeats extending from the N-terminus to the FAT domain, which is itself comprised of 16 tandem TPR repeats. These findings were used to assemble a 3D atomic model for the TOR protein.

Keywords

Protein repeats; HEAT; TPR; TOR; HHpred; PIKK

1. Introduction

TOR (Target of rapamycin) is a serine/threonine protein kinase and signaling protein that is conserved in all eukaryotes. TOR was initially identified in *Saccharomyces cerevisiae* as a gene locus that confers drug resistance to the immunosuppressant rapamycin, an inhibitor of TOR kinase activity [1-5]. TOR is one of six members of the phosphoinositide-3-OH-related kinase (PIKK) protein family of eukaryotic signaling proteins that are essential for monitoring cellular nutrient supply, genome and transcript integrity, and regulating chromatin structure [6-8]. Importantly, TOR regulates both cell growth and cell cycle progression and is a major target for cancer therapeutics [9-11].

The overall domain structure of TOR is well conserved and is defined by a C-terminal PI3K (phosphatidylinositol-3-kinase) catalytic domain that is immediately flanked by four different alpha helical domains [8]. The amino terminus of the PI3K domain is flanked by a four-helix bundle called the FRB (FKBP12 rapamycin binding) domain and an extended

© 2009 Elsevier Inc. All rights reserved.

Corresponding author, Phone 206-667-5263, Fax 206-667-6497, bknutson@fhcrc.org.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Competing interests

The author declares that he has no competing interests.

helical region called the FAT (FRAP, ATM, and TRRAP) domain [2,12,13]. The FAT domain is conserved among PIKK family members and possibly folds into a combination of HEAT (Huntington, EF3A, ATM, TOR) and TPR (Tetratricopeptide repeat) repeats [2,12,14,15]. The C-terminus of TOR consists of the FATC (FRAP, ATM and TRRAP C-terminal) domain, composed of one short and one long helix joined by a disulfide bond [16]. The N-terminus of PIKKs, which is less conserved at the primary structure level, is composed almost entirely of tandem alpha helices that possibly fold into repeats [14,15].

Little is known about the structure and function of the repeat domains of PIKKs, but in other repeat-containing proteins these structures mediate protein-protein interactions. Sequence analysis of the PIKK family indicates that these proteins contain HEAT and TPR repeats in their HEAT and FAT domains [14,15]. However, the number and position of these repeats remains undetermined. Individual repeat units vary in length from 30 to 40 amino acid residues and number from as few as two repeats to more than dozens within a protein [17,18]. HEAT repeats consist of two antiparallel alpha helices, forming a helical hairpin ~30 amino acids in length. These repeats pack together in parallel via hydrophobic interactions between adjacent repeat helices and ultimately fold into an elegant superhelix [19]. TPR repeats also consist of ~34 amino acids that fold into two alpha helices in an anti-parallel arrangement that in turn pack into a helical superstructure similar to that described for HEAT repeat-containing proteins [20,21].

Understanding the repeat architectures of TOR is key to understanding how this protein ultimately functions in the cell and how TOR interacts with its targets. A major obstacle to detailed structural and functional analysis of TOR and other members of the PIKK protein family is their large size (~280-470 KDa), which renders experimental manipulation difficult. To date, only three structures of TOR have been solved: two high resolution structures of the FRB and FATC domains and a low resolution EM structure of the full-length protein [2,13,16,22]. Structural and functional studies of PIKKs have also been hindered by poor understanding of their repeat architectures, as traditional web-based repeat prediction methods detect few to none of the repeats in their HEAT and FAT domains. The main reason for this is that the primary amino acid structures of protein repeats exhibit too much variation for accurate prediction. Therefore, the HEAT and FAT domains of PIKKs may contain more divergent protein repeats that would not be identified by sequence similarity alone.

For this reason, I developed a simple method capable of predicting structural repeats in TOR that could be used to guide structural and functional analysis. The strategy uses HHpred, a powerful and fast sequence homology detection method that predicts a new and potentially more complete depiction of the domain and repeat architecture of the TOR protein family. The HHpred strategy presented here performed significantly better than traditional web-based methods by predicting 100% of the HEAT and TPR repeats within a test set of proteins of known three dimensional (3D) structure. Based on these results, I predict that the HEAT and FAT domains of TOR fold into a continuous superhelix of 48 tandem HEAT and TPR repeats. Comparative sequence analysis of TOR proteins from divergent eukaryotic species provides further supporting evidence that TOR proteins from fly, plant, yeast, and humans fold into remarkably similar structures and exhibit several intriguing structural differences in their HEAT and FAT domains that are examined in this report. These findings were used to assemble a 3D structural model for the TOR protein.

2. Methods

2.1. Computational methods

Protein domain searches were performed by Pfam, Prosite, SMART, REP, HHrepID, ARD, and TPRpred protein signature recognition tools [23-32]. Pfam (<http://pfam.sanger.ac.uk>), Prosite (<http://ca.expasy.org/prosite>), SMART (<http://smart.embl-heidelberg.de>), REP (<http://www.embl.de/~andrade/papers/rep/search.html>), TPRpred (<http://toolkit.tuebingen.mpg.de/tprpred>), ARD (<http://www.ogic.ca/projects/ard/client.pl>), HHrepID (<http://toolkit.tuebingen.mpg.de/hhrepid>). Pfam, Prosite, SMART, REP, and TPRpred are protein domain and repeat databases that can be used to determine domain and repeat architectures by searching a protein sequence of interest and comparing it to repeat and domain sequence profiles [23-30]. REP is specific for protein repeats, and TPRpred is specific for TPR repeats, while ARD and HHrepID are de novo repeat detection programs. ARD has been trained to detect HEAT repeats using a neural net [32]. HHrepID identifies protein repeats by searching for internal sequence symmetry by HMM-HMM comparisons [31]. All domain and repeat searches were conducted using each program's default settings and thresholds. The threshold for Pfam and SMART was set to E-values of 1.0 and 0.01, respectively. Prosite was set at a high confidence profile score cut off of 0, and REP and TPRpred searches were performed using the default selection thresholds. HHrepID searches were also conducted using program default settings that include 8 PSI-BLAST iterations, secondary structure scoring, repeat family E-value of 0.01, self alignment P-value of 0.1, and a MAC threshold of 0.5. A threshold of 0.8 was implemented for ARD searches. The Kippert et al. HEAT repeat prediction protocol was performed as described in [33] using HHpred with program default settings and parameters searched against Pfam database.

Structure similarity searches for homologous proteins were performed by HHpred (<http://toolkit.tuebingen.mpg.de/hhpred>) using the database of Hidden Markov Models (HMM) domains of known structure from the protein databank (PDB) database (<http://www.pdb.org/pdb/home/home.do>) [34]. All HHpred searches were conducted with the program's default settings and thresholds, and known structural repeat units were located by feature aligner available through the UniProtKB/Swiss-Protein database (<http://au.expasy.org/uniprot>) or from the 3D structures deposited in the PDB database [35]. Accession numbers for TOR protein sequences used for HHpred analysis are listed as follows: mTOR (NM_019906.1), hTOR (NP_004949.1), fTOR (NP_524891.1), pTOR (NP_001117459), yTOR1 (NP_012600.1), yTOR2 (NP_012719.2). 3D models were built using pymol (www.pymol.org) [36]. HHpred results used to assign HEAT and TPR repeats in TOR are listed in Supplementary Table 4. Protein secondary structure searches were performed by PSIPRED and DISOPRED2 [37,38]. PSIPRED (<http://bioinf.cs.ucl.ac.uk/psipred/psiform.html>), DISOPRED2 (<http://bioinf.cs.ucl.ac.uk/disopred/disopred.html>). HMM alignments between TOR proteins were generated by HHpred using the mTOR protein sequence as the query and the genomes of human, fly, plant, and yeast as the HMM database source.

2.2 Benchmarking

Prediction performance was measured by using a benchmark dataset of 37 proteins of known 3D structure. The dataset consisted of 13 HEAT-containing protein, 13 TPR-containing proteins, and 11 non-repeat-containing proteins. The dataset of non-repeat containing proteins consisted of a negative training set previously used by Palidwor et. al. with the addition of three alpha helical proteins selected from the PDB database at random [32]. Each of the 3D structures selected for the negative dataset consist almost entirely of alpha helices. The protein sequences from the following 3D structures from the PDB

database were used: 1egd A, 1bpy A, 1chu A, 1aua A, 1mhy D, 1fts A, 1cpo A, 1bmt A, 2wuz A, 3h37 A, and 3a4l A.

To access whether a repeat prediction by HHpred and web-based methods was correct, it is necessary that the repeat prediction overlap helix 1, the interlinker, and helix 2 of a HEAT or TPR repeat unit. To access correct repeat predictions by ARD, the highlighted amino acid must lie within the interlinker region of a HEAT repeat unit. For the HHpred strategy, two criteria were used to select the best templates: 1) probability, and 2) protein identity. Methods were evaluated based on their sensitivity, as each method predicted protein repeats with equal specificity, since no false positive protein repeats were detected in the negative dataset. Sensitivity was defined as: $(\text{true positives} / (\text{true positives} + \text{false negatives}) \times 100\%)$. A comprehensive list of HEAT and TPR repeat predictions is depicted in Supplementary Figure 1. HHpred results used to predict the number and location of HEAT and TPR repeats are listed in Supplementary Tables 1 and 2, respectively.

2.3. 3D modeling

Structural templates for each of the five domains in the TOR protein were identified by HHpred. PIR-alignments of the TOR protein sequence with the selected template were generated by HHpred using the “create model” option in the HHpred results page. Templates were selected by the highest probability score and highest protein identity. For repeat-containing domains, templates were selected based on the above criteria and also on the template alignment that matched the greatest number of consecutive repeat units. Next, PIR-alignments were manually converted to FASTA alignments. These alignments were used to build 3D models with the SWISS-MODEL workspace (<http://swissmodel.expasy.org/workspace>) using the alignment mode [39-42]. Protein model quality assessment was determined using the RMSD for each modeled region by PYMOL (www.pymol.org), and each of the modeled regions was manually fitted into a continuous structural model [36].

3. Results

3.1. HHpred repeat prediction strategy

The repeat prediction strategy begins by analyzing a target protein's primary amino acid structure using HHpred, which identifies homologous sequences from HMM-HMM comparison of the target protein with several protein family and structure databases [43,44]. HHpred is one of the most sensitive methods for remote homology detection as it can search a target protein against multiple protein databases. Because protein repeats vary considerably in their amino acid sequence, it is difficult to detect homology among repeat-containing proteins even though they ultimately form nearly identical 3D structures [45]. HHpred excels at discerning homology between divergent repeat-containing proteins, and the HHpred generated alignments between the target protein and a template of known 3D structure can be used to model the target protein's structure and to locate its individual protein repeats [44].

A schematic of the HHpred strategy is depicted in Figure 1A. If the target protein is searched against the PDB database, HHpred yields a list of pairwise query-template alignments of the target protein and potential protein matches of known 3D structure that are ranked from high to low probability. For example, HHpred analysis of Cand1, a HEAT repeat-containing protein, yields matches with itself and with several HEAT repeat proteins of known structure (Fig.1B). Cand1 is a 120 KDa protein comprised of 27 tandem HEAT repeats that form a unique U-shaped superhelix [46]. HHpred matched Cand1 with the

HEAT repeat-containing proteins importin $\beta 1$, protein phosphatase 2A, and karyopherin $\beta 2$ and it assigned probabilities between 99 and 100% for each template alignment (Fig.1B).

If the target protein is very large, the protein sequence is segmented into multiple overlapping pieces and each is searched separately to obtain optimal coverage of the entire protein sequence. For example, optimal coverage is not achieved upon searching the entire protein sequence of Cand1 in a single search because the next largest HEAT repeat-containing protein of known 3D structure contains only 19 HEAT repeats, and a perfect match between the two proteins would therefore miss 8 repeat units. For this reason, I performed two independent HHpred searches, one using the N-terminal portion and a second with the C-terminal portion of the protein to ensure complete coverage of the Cand1 protein sequence (Fig.1B, Search 1 and 2).

Alignment gaps occur often, as the amino acid identity between target and template proteins routinely fall below 17% (Fig.1B, Search1). To correct for this, multiple alignments are used to fill in the sequence gaps. In most cases, two template alignments are sufficient to precisely locate repeats in a target's protein sequence. If the repeat units within the template sequence are known, the position and sequence of the repeats can be extrapolated from the matching 3D structure (Fig.1B). An example of a primary structure alignment for Cand1 HEAT repeat 1 and matching template repeats is illustrated in Figure 1C. Each Cand1 HEAT repeat unit matched a minimum of two repeats in the template structures, and the HHpred strategy predicted all 27 Cand1 HEAT repeat units.

3.2. Comparison of HEAT repeat prediction methods

Repeat prediction methods predict vastly different repeat architectures for the TOR protein and it is unclear whether these methods lack the sensitivity required to predict the repeats in the TOR protein. I compared the user-directed HHpred strategy with several automated web-based methods to determine their respective sensitivities in predicting and locating repeats in proteins with known 3D structure. The user defined HHpred strategy uses expert user input and is expected to perform better than web-based methods, so the extent that it can be compared to the automated web-based methods is limited. The HEAT repeat family defined by Pfam contains ~27 unique proteins with known 3D structures. Related protein sequences were excluded from the analysis to avoid bias of to one another. Therefore, protein sequences exceeding 30% protein identity were excluded since these proteins are likely to share similar protein structures [47]. I manually compiled a set of 13 different HEAT-containing protein sequences using the following criteria: 1) protein structure of closest to the full-length protein, 2) highest resolution protein structure, 3) number and position of the repeat units within the protein have been annotated. For all proteins tested, HHpred matched the target proteins with their respective PDB entries and closely related homologs. These matches were excluded because alignment between two closely related homologs with protein identities exceeding 30% would bias HHpred prediction sensitivities. Six different web-based methods were used for prediction of HEAT repeats including REP, Pfam, Prosite, SMART, ARD, and HHrepID.

For all 13 HEAT repeat-containing proteins, the HHpred strategy correctly identified and located 100% of the HEAT repeat units. SMART, Prosite, and Pfam performed poorly, detecting less than 18% of the HEAT repeats searched. HHrepID, ARD and REP performed significantly better than the other web-based methods, detecting between 34-42% of the HEAT repeats, but failing to identify more than half of the repeats searched. The results of HEAT repeat prediction for all the detection methods is tabulated in Table 1, and repeats detected by HHpred and the web-based methods are shown in Figure 2.

3.3. Comparison of HEAT repeat prediction methods

I next tested HHpred's ability to detect TPR-containing protein sequences, the other type of repeat found in the TOR protein. A total of 13 out of ~40 protein structures with TPR-containing sequences available in the PDB database were analyzed by either HHpred or other web-based methods. Each of the 13 TPR-containing proteins chosen for analysis represents a unique structure belonging to the TPR protein family defined by Pfam. The same criteria to select HEAT-containing protein sequences were used for the selection of TPR repeat containing protein sequences. Six different web-based methods were used for prediction of TPR repeats, including REP, Pfam, Prosite, SMART, TPRpred, and HHrepID. TPRpred, a profile-based algorithm specifically designed to detect TPR-containing sequences, correctly identified all but two TPR repeats or 97% of the repeats among the proteins tested. The remaining five web-based methods performed similarly by detecting between 57-70% of the TPR repeats. Finally, the HHpred strategy detected 100% of the TPR repeats in the proteins tested, demonstrating that the user-directed HHpred approach performs with high sensitivity. The results of TPR repeat prediction by all detection methods are tabulated in Table 2, and repeats detected by HHpred and the web-based methods are shown in Figure 3.

3.4. Prediction of HEAT and TPR repeats in the TOR protein

The number and position of HEAT and TPR repeats in the TOR protein is not well characterized, as web-based methods predict very different repeat architectures for TOR. HHrepID, REP, and ARD performed the best among the automated web-based methods by detecting 11-16 HEAT repeats clustered within the HEAT domain while Pfam, Prosite, and SMART detected no repeats (Fig.4A). None of the web-based methods detected TPR repeats with the exception of TPRpred (Fig.4A, column 6). Even though several HEAT and TPR repeats are detected in the TOR protein sequence by the web-based methods, significant gaps remain, leaving more than 45% of the HEAT and FAT domains unassigned. It is unclear whether these unassigned gaps represent protein sequences that lack repeats or if they contain divergent repeats whose detection requires the increased sensitivity of the user-directed HHpred method.

To analyze the TOR protein by HHpred, I segmented the TOR protein sequence into four overlapping fragments and their primary amino acid structures were analyzed individually by HHpred. Each segment generated high probability matches with protein structures of HEAT- and TPR-containing proteins as well as with domain structures of FRB, FATC, and the catalytic domain of PI3K (Supplementary Table 3). From these HHpred results, I predict that the HEAT domain of TOR forms a continuous stretch of 32 HEAT repeats (Fig.4A, column 1, white boxes). Immediately following the HEAT domain, HHpred detected a tandem array of 16 TPR repeats that encompasses the entire FAT domain located between the HEAT and FRB domains of the TOR protein (Fig.4A, column 1, grey boxes). The location and sequence of HEAT and TPR repeats predicted by the HHpred method are shown in Supplementary Table 4.

Previous comparative modeling of the TOR protein sequence predicted that it contains as many as 41 contiguous HEAT repeats that extend to the PI3K domain (Fig.4A, column 2) [15]. Perry et al. employed a combination of sequence alignments, secondary structure predictions, and expanded consensus to predict the location of HEAT repeats in TOR. A second approach using structure homology recognition predicted that the TOR protein contains 36 discontinuous HEAT and TPR repeats. The Brewerton et al. method used a structure homology recognition program called FUGUE, which recognizes homologous proteins from sequence and structure comparisons [14, 48]. A quantitative comparison between the Brewerton and HHpred methods is not possible, as the location of Brewerton's

predicted repeats has not been published. However the HHpred prediction presented here supports the conclusion of Brewerton et al. that the region proximal to the PI3K domain is composed of TPR repeats. In contrast, HHpred predicted that TOR contains a continuous stretch of HEAT and TPR repeats rather than clusters of repeats separated by regions of non-repeat protein sequence suggested by the Brewerton et al. model.

Recently, Kippert et al. developed a repeat prediction protocol that detects HEAT repeats with high sensitivity [33]. Their protocol uses multiple sequence alignments that are searched by HHpred against the Pfam database to identify repeat-containing segments as well as individual repeat units within a protein of interest. I analyzed the HEAT domain of TOR by the Kippert et al. protocol and detected only 15 HEAT repeats with a pattern very similar to the repeats detected by ARD and REP (Fig.4A, column 3). Pairwise comparisons of repeat prediction methods revealed that the predictions accomplished by HHpred strategy showed 6-53% similarity to the repeat architectures predicted by other methods, indicating that HHpred predicts a much different structure for the TOR protein (Fig.4A).

3.5. Conservation of TOR repeat architecture

I used the computational model of TOR to compare the repeat architectures of TOR proteins from divergent species to gain insight into the structural conservation of these proteins. Pairwise HMM alignments between mouse TOR (mTOR) and TOR orthologs from human, fly, yeast and plants were generated by HHpred. Using the predicted mTOR repeat structure as a guide, I superimposed the repeat units predicted for mTOR over the other TOR homologs. Overall, the repeat architectures of the six TOR protein sequence analyzed are nearly identical, but there are some intriguing differences between their predicted structures (Fig.4B). The overall repeat structures of mTOR and human TOR (hTOR) are identical as are the repeat structures of the two yeast TOR paralogs, yTOR1 and yTOR2. The HEAT domain repeat structures of mTOR, hTOR and fly TOR (fTOR) are also identical (Fig.4B). However, the HEAT domains of plant TOR (pTOR), yTOR1 and yTOR2 are missing two tandem HEAT repeats corresponding to HEAT 7 and 8 of mTOR (Fig.4B, column 4-6). Detailed sequence analysis of the deletions in pTOR and yTOR1 and 2 revealed that they are positioned precisely at the boundaries of HEAT repeat 6 and 9 of mTOR, indicating that these repeats may have been gained or lost as a result of speciation (Fig.5A).

Another difference between the TOR protein structures is that yTOR1 and yTOR2 contain an ~80 amino acid extension at the N-terminus of sufficient length to accommodate two additional HEAT repeats units. I analyzed the protein sequences of the N-terminal extensions by PSIPRED and DISOPRED2, which predict protein secondary structure and intrinsically unstructured protein sequences, respectively [37,38]. The N-termini of both proteins are predicted to fold into alpha helical and beta sheet structures by PSIPRED, and do not resemble the consensus HEAT repeat structure (Fig.5B and C, second row). DISOPRED2 also predicts that the majority of the N-terminal extension are unstructured, and these disordered regions extend up to the first predicted HEAT repeat in yTOR 1 and 2, strongly suggesting that the N-termini are likely unstructured and do not fold into HEAT repeats (Fig.7B and C, third row). Moreover, HHpred and web-based detected no repeat-containing protein sequences within the N-terminal extensions of yTOR1 and 2, further supporting the conclusion (data not shown).

I also identified a nonconserved region (NCR) that varies in length between the eukaryotic species examined and that corresponds to the interlinker within the helices of TPR repeat 13 (Fig.5D). In mTOR and hTOR, the interlinker protein sequence is ~65 amino acids, but is progressively smaller among the other TOR proteins. In yTOR 1 and 2, the NCR is between 36 and 41 amino acids, while in fTOR the NCR is only 22 amino acids. Surprisingly, the NCR is completely absent from the pTOR protein sequence. Secondary structure analysis of

NCRs suggests that they are intrinsically disordered, a common characteristic of interlinkers within TPR repeats since they form loops between the repeat helices [20,21]. Repeat analysis failed to detect any TPR repeats within the NCR protein sequence, suggesting that the NCR may represent a unique interlinker sequence within TPR repeat 13. The large size of the interlinker is quite surprising, as TPR interlinkers tend to average ~4 amino acids in length [20,21].

3.6. 3D Model of the TOR protein

The HHpred method compared the entire length of the TOR protein sequence with proteins of known 3D structure. From these alignments, I assembled a 3D model of mTOR using representative matches for each of the HEAT and TPR repeat units, FRB domain, PI3K catalytic domain, and FATC domain (Supplementary Table 5). Importin β 1 and O-linked N-acetylglucosamine protein structures were used as templates for HEAT and TPR repeat units, respectively. A 3D structural representation of this model is shown in Figure 6A-D. As other template structures could be used to build the 3D model of TOR, the model presented here is meant to suggest one possible structural conformation assumed by TOR.

4. Discussion

In this paper, I used HHpred to predict and locate protein repeats in the TOR protein family from their primary amino acid structure. The user-directed HHpred strategy predicted two different types of protein repeats with high sensitivity, and revealed several important findings regarding TOR's structure. The HHpred method predicted a tandem arrangement of HEAT and TPR repeats rather than a set of repeat clusters. Based on the above results, TOR may ultimately fold into an expansive superhelix as illustrated in Figure 6. Alternate structural conformations of TOR are likely possible since it is unclear whether there are intramolecular contacts between TOR's domains. The pitch of TOR's HEAT and TPR superhelices is also presently unclear. It would be useful to find characteristics of HEAT and TPR repeats that determine the helicoid pitch of their respective superhelices so that these properties could be incorporated into the 3D model. HHpred also predicted that TOR's FAT domain is composed entirely of TPR repeats. Due to the sequence conservation between the FAT domains of other PIKK protein family members, the FAT domains of all PIKKs may fold into a similar repeat architecture. The role of the FAT domain for PIKKs is presently unclear, but it is possible that this domain has similar functions for all PIKKs such as binding substrate proteins.

Comparative sequence analysis of TOR proteins from different eukaryotes indicates that the HEAT and FAT domain structures are conserved by more than 95% between the species examined. The main difference between these structures lies within the HEAT domain, where an ~80 amino acid stretch of protein sequence is missing in pTOR and yTOR. Detailed sequence analysis revealed that this region corresponds to the boundaries of HEAT repeat 7 and 8 in mTOR and appears as though they have been precisely removed from the pTOR and yTOR protein sequences, indicating a unique site of repeat gain or loss as a result of speciation. It remains unclear how these two additional HEAT repeats contribute to metazoan TOR function, but one model is that they provide a platform for additional protein interactions.

The relatively large and variable interlinker within TPR repeat 13 may also provide additional contacts for TOR-interacting proteins. Interlinkers of HEAT- and TPR-containing proteins are often sites of protein-protein interaction [17,19-21]. Therefore, an expanded interlinker within TPR repeat 13 may be an important protein interaction interface that mediates a unique role specific for each TOR protein. Interestingly, yeast encode two copies of TOR, which preferentially form two separate complexes with distinct functions and

interact with different proteins [4,5]. The primary structures of the interlinker of TPR repeat 13 for yTOR1 and 2 are poorly conserved and likely highly disordered based on secondary structure predictions. This is analogous to the N-terminal extensions of yTOR1 and 2 that are also poorly conserved and likely disordered. These variable regions could help govern the assembly of the two complexes in yeast by providing unique protein interfaces that preferentially bind to one of the yTOR proteins.

Finally, the 3D model and repeat annotations predicted for the TOR protein by HHpred provides useful structural information that could be used as a guide to future work. The repeat architecture of TOR has been mapped and the repeat structures could be altered to test the importance of these features biochemically. The 3D model could be used in fitting the atomic model into EM densities, which may be close realization since an EM structure of yeast TOR has been determined, but atomic fitting of TOR's domain structures remains incomplete. Moreover, a similar strategy could be used to predict the repeat architectures of other PIKK protein family members as well as other large repeat-containing proteins that perform poorly when searched by automated web-based methods.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

I would like to thank members of the Hahn lab at the FHCRC for helpful discussions and Beth Moorefield, Steve Hahn, Ivanka Kamenova, Jorja Henikoff, and Steve Henikoff for critical reading of the manuscript. BAK was supported by the Chromosome Metabolism and Cancer Training grant NIH T32 CA09657 and grant RO1GM075114 to Steve Hahn.

References

1. Chen J, et al. Identification of an 11-kDa FKBP12-rapamycin-binding domain within the 289-kDa FKBP12-rapamycin-associated protein and characterization of a critical serine residue. *Proc Natl Acad Sci U S A* 1995;92(11):4947–51. [PubMed: 7539137]
2. Choi J, et al. Structure of the FKBP12-rapamycin complex interacting with the binding domain of human FRAP. *Science* 1996;273(5272):239–42. [PubMed: 8662507]
3. Heitman J, Movva NR, Hall MN. Targets for cell cycle arrest by the immunosuppressant rapamycin in yeast. *Science* 1991;253(5022):905–9. [PubMed: 1715094]
4. Helliwell SB, et al. TOR1 and TOR2 are structurally and functionally similar but not identical phosphatidylinositol kinase homologues in yeast. *Mol Biol Cell* 1994;5(1):105–18. [PubMed: 8186460]
5. Kunz J, et al. Target of rapamycin in yeast, TOR2, is an essential phosphatidylinositol kinase homolog required for G1 progression. *Cell* 1993;73(3):585–96. [PubMed: 8387896]
6. Abraham RT. Phosphatidylinositol 3-kinase related kinases. *Curr Opin Immunol* 1996;8(3):412–8. [PubMed: 8793996]
7. Abraham RT. PI 3-kinase related kinases: 'big' players in stress-induced signaling pathways. *DNA Repair (Amst)* 2004;3(8-9):883–7. [PubMed: 15279773]
8. Templeton GW, Moorhead GB. The phosphoinositide-3-OH-kinase-related kinases of *Arabidopsis thaliana*. *EMBO Rep* 2005;6(8):723–8. [PubMed: 16065066]
9. Abraham RT. TOR signaling: an odyssey from cellular stress to the cell growth machinery. *Curr Biol* 2005;15(4):R139–41. [PubMed: 15723787]
10. Guertin DA, Sabatini DM. An expanding role for mTOR in cancer. *Trends Mol Med* 2005;11(8):353–61. [PubMed: 16002336]
11. Sampson JR. Therapeutic targeting of mTOR in tuberous sclerosis. *Biochem Soc Trans* 2009;37(Pt 1):259–64. [PubMed: 19143643]

12. Bosotti R, Isacchi A, Sonnhammer EL. FAT: a novel domain in PIK-related kinases. *Trends Biochem Sci* 2000;25(5):225–7. [PubMed: 10782091]
13. Veverka V, et al. Structural characterization of the interaction of mTOR with phosphatidic acid and a novel class of inhibitor: compelling evidence for a central role of the FRB domain in small molecule-mediated regulation of mTOR. *Oncogene* 2008;27(5):585–95. [PubMed: 17684489]
14. Brewerton SC, et al. Structural analysis of DNA-PKcs: modelling of the repeat units and insights into the detailed molecular architecture. *J Struct Biol* 2004;145(3):295–306. [PubMed: 14960380]
15. Perry J, Kleckner N. The ATRs, ATMs, and TORs are giant HEAT repeat proteins. *Cell* 2003;112(2):151–5. [PubMed: 12553904]
16. Dames SA, et al. The solution structure of the FATC domain of the protein kinase target of rapamycin suggests a role for redox-dependent structural and cellular stability. *J Biol Chem* 2005;280(21):20558–64. [PubMed: 15772072]
17. Andrade MA, Perez-Iratxeta C, Ponting CP. Protein repeats: structures, functions, and evolution. *J Struct Biol* 2001;134(2-3):117–31. [PubMed: 11551174]
18. Kajava AV. Review: proteins with repeated sequence--structural prediction and modeling. *J Struct Biol* 2001;134(2-3):132–44. [PubMed: 11551175]
19. Andrade MA, et al. Comparison of ARM and HEAT protein repeats. *J Mol Biol* 2001;309(1):1–18. [PubMed: 11491282]
20. Blatch GL, Lassle M. The tetratricopeptide repeat: a structural motif mediating protein-protein interactions. *Bioessays* 1999;21(11):932–9. [PubMed: 10517866]
21. D'Andrea LD, Regan L. TPR proteins: the versatile helix. *Trends Biochem Sci* 2003;28(12):655–62. [PubMed: 14659697]
22. Adami A, et al. Structure of TOR and its complex with KOG1. *Mol Cell* 2007;27(3):509–16. [PubMed: 17679098]
23. Andrade MA, et al. Homology-based method for identification of protein repeats using statistical significance estimates. *J Mol Biol* 2000;298(3):521–37. [PubMed: 10772867]
24. Karpenahalli MR, Lupas AN, Soding J. TPRpred: a tool for prediction of TPR-, PPR- and SEL1-like repeats from protein sequences. *BMC Bioinformatics* 2007;8:2. [PubMed: 17199898]
25. Ponting CP, et al. SMART: identification and annotation of domains from signalling and extracellular protein sequences. *Nucleic Acids Res* 1999;27(1):229–32. [PubMed: 9847187]
26. Schultz J, et al. SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res* 2000;28(1):231–4. [PubMed: 10592234]
27. Schultz J, et al. SMART, a simple modular architecture research tool: identification of signaling domains. *Proc Natl Acad Sci U S A* 1998;95(11):5857–64. [PubMed: 9600884]
28. Sigrist CJ, et al. PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinform* 2002;3(3):265–74. [PubMed: 12230035]
29. Sonnhammer EL, et al. Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res* 1998;26(1):320–2. [PubMed: 9399864]
30. Sonnhammer EL, Eddy SR, Durbin R. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* 1997;28(3):405–20. [PubMed: 9223186]
31. Biegert A, Soding J. De novo identification of highly diverged protein repeats by probabilistic consistency. *Bioinformatics* 2008;24(6):807–14. [PubMed: 18245125]
32. Palidwor GA, et al. Detection of alpha-rod protein repeats using a neural network and application to huntingtin. *PLoS Comput Biol* 2009;5(3):e1000304. [PubMed: 19282972]
33. Kippert F, Gerloff DL. Highly sensitive detection of individual HEAT and ARM repeats with HHpred and COACH. *PLoS One* 2009;4(9):e7148. [PubMed: 19777061]
34. Berman HM, et al. The Protein Data Bank. *Nucleic Acids Res* 2000;28(1):235–42. [PubMed: 10592235]
35. Bairoch A, et al. The Universal Protein Resource (UniProt). *Nucleic Acids Res* 2005;33(Database issue):D154–9. [PubMed: 15608167]
36. DeLano, WL. DeLano Scientific LLC. Palo Alto; CA, USA: 2008. The PyMOL Molecular Graphics System.

37. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. *Bioinformatics* 2000;16(4):404–5. [PubMed: 10869041]
38. Ward JJ, et al. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol* 2004;337(3):635–45. [PubMed: 15019783]
39. Arnold K, et al. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* 2006;22(2):195–201. [PubMed: 16301204]
40. Guex N, Peitsch MC. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* 1997;18(15):2714–23. [PubMed: 9504803]
41. Kiefer F, et al. The SWISS-MODEL Repository and associated resources. *Nucleic Acids Res* 2009;37(Database issue):D387–92. [PubMed: 18931379]
42. Schwede T, et al. SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res* 2003;31(13):3381–5. [PubMed: 12824332]
43. Soding J. Protein homology detection by HMM-HMM comparison. *Bioinformatics* 2005;21(7):951–60. [PubMed: 15531603]
44. Soding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res* 2005;33(Web Server issue):W244–8. [PubMed: 15980461]
45. Kinch LN, Grishin NV. Evolution of protein structures and functions. *Curr Opin Struct Biol* 2002;12(3):400–8. [PubMed: 12127461]
46. Goldenberg SJ, et al. Structure of the Cand1-Cul1-Roc1 complex reveals regulatory mechanisms for the assembly of the multisubunit cullin-dependent ubiquitin ligases. *Cell* 2004;119(4):517–28. [PubMed: 15537541]
47. Peterson ME, et al. Evolutionary constraints on structural similarity in orthologs and paralogs. *Protein Sci* 2009;18(6):1306–15. [PubMed: 19472362]
48. Shi J, Blundell TL, Mizuguchi K. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J Mol Biol* 2001;310(1):243–57. [PubMed: 11419950]
49. Moreland JL, et al. The Molecular Biology Toolkit (MBT): a modular platform for developing molecular visualization applications. *BMC Bioinformatics* 2005;6:21. [PubMed: 15694009]

Figure 1. HHpred analysis scheme

A. Flow diagram of the HHpred repeat prediction strategy. B. Identification of individual repeat units from multiple HHpred alignments of the Cand1 target protein and templates of known 3D structure. Black boxes with white numbering represent the number and location of repeat units in the Cand1 protein sequence. Two independent HHpred searches were performed for the Cand1 protein, one with residues 1-750 (Search 1, results shown above) and a second with residues 730-1230 (Search 2, results shown below). HHpred matches the Cand1 protein with high probability to three different template structures that include the importin β 1 (IMB1), protein phosphatase 2A (PR65A), and karyopherin β 2 (KARY2). Each repeat unit was located in the template structures as shown by white boxes and the number of the repeat units within their structure is indicated in black numbering. C. Protein sequence alignments of Cand1 HEAT repeat 1 and template repeat units from IMB1, PR65/A, and KARY2. HEAT repeat structural features are indicated above the sequence alignment and include the following: L1, linker 1, H1, helix 1, IL, interlinker, H2, helix 2, L2, linker 2. The 3D structure of an individual repeat unit is shown below the sequence alignment with the corresponding structural components labeled. H1 and H2 of a single HEAT repeat unit are highlighted in grey and correspond to the grey portions of the ribbon structure while the linkers are drawn in black.

Figure 2. Detection of HEAT repeat units in proteins of known 3D structures

HEAT repeats identified only by HHpred only are colored in red. HEAT repeats identified by HHpred and one or more of the web-based methods are colored in yellow. Grey colored ribbons represent portions of the protein that do not contain HEAT repeat sequences.

Individual HEAT repeat units are numbered. Protein structure images were generated by MBT protein workshop available online (<http://mbt.sdsc.edu/software/applications>) [49]. A. Cand1 (PDB, 1u6g, C), B. CSE1 (PDB, 1wa5, C), C. TNPO1 (PDB, 2z5k, A), D. IMB1 (PDB, 1qgr, A), E. 2AAA (PDB, 1b3u, A), F. AP2 α 2 (PDB, 2vgl, A), G. CTNB1 (PDB, 1g3j, A), H. VATH (PDB, 1ho8, A), I. IMA1 (PDB, 1bk5, A), J. PKP1 (PDB, 1xm9, A), K. EF3A (PDB, 2iw3, A), L. HSPB1 (PDB, 1xqr, A), M. METH1 (PDB, 1te4, A).

Figure 3. Detection of TPR repeat units in proteins of known 3D structures

TPR repeats identified by HHpred and one or more of the web-based methods are colored in yellow. TPR repeats identified by HHpred and TPRpred only are colored in orange while TPR repeats identified only by HHpred are colored in red. Grey colored ribbons represent portions of the protein that do not contain HEAT repeat sequences. Individual TPR repeat units are numbered. Protein structure images were generated by MBT protein workshop available online (<http://mbt.sdsc.edu/software/applications>) [49]. A. OGT1 (PDB, 1w3b, A), B. MALT (PDB, 1hz4, A), C. PEX5 (PDB, 1fch, A), D. HCPC (PDB, 1ouv, A), E. HCPB (PDB, 1klx, A), F. PPP5 (PDB, 1wao, A), G. PPID (PDB, 1ihg, A), H. PLAF7 (PDB, 2fbn,

A), I. STIP1 (PDB, 1elr, A), J. STUB1 (PDB, 2c2l, A), K. FKBP4 (PDB, 1qz2, A), L. SYCD (PDB, 2vgx, A), M. FIS1 (PDB, 1nzn, A).

Figure 4. Prediction and comparison of TOR HEAT and TPR repeats

Rectangular boxes denote individual repeat units. A solid black line represents a gap in the structure corresponding to a missing repeat unit relative to the HHpred predicted structure. Individual repeat units in the mTOR structure are numbered within each rectangular box. A. The method used to predict the protein repeats is indicated at the top of each column. P.K., Perry and Klechner prediction. K.G., Kippert and Gerloff protocol. White and grey colored boxes indicate a HEAT and TPR repeats correctly identified by both HHpred and web-based method, respectively. Blue boxes indicate a mismatch between the HHpred and web-based method. Green boxes indicate a HEAT repeat identified by a web-based method that matches with a TPR repeat identified by HHpred. B. The HEAT and FAT/TPR domains from six different TOR proteins are shown. The name of each TOR protein is labeled at the top of each structure. Percent identity of each protein relative to the mTOR protein sequence is indicated at the bottom of the structures. White and grey rectangular boxes denote HEAT and TPR repeat units, respectively

Figure 5. Sequence analysis of divergent TOR repeats structures

Panel A shows a multiple sequence alignment of HEAT 6-9 for six TOR protein sequences from different eukaryotic species. The name of each TOR protein is indicated at the left of each sequence. Grey and black boxes indicate individual HEAT repeat units that are labeled at the bottom of each box. The numbers to the right of the sequence denote amino acid position. Bold lettering indicates an amino acid conserved for at least four of the sequences. Protein sequence analysis of yTOR1 (Panel B) and yTOR2 (Panel C) N-terminal extensions by PSIPRED (middle row) and DISOPRED2 (bottom row). H, helix, E, beta sheet, C, coil, D, disordered, =, ordered, -, gap. The black box and white lettering indicates predicted HEAT repeat 1 of yTOR1 and 2. Panel D shows a multiple sequence alignment and predicted secondary structure of TPR repeat 13. The first row labeled HHpred depicts the HMM alignments generated by HHpred that were combined for six different TOR protein sequences. Below the primary structure alignment, the secondary structures predicted by HHpred using PSIPRED are listed for each TOR protein amino acid position. The final row labeled DISOPRED2 shows disordered protein sequence predictions that are listed for each TOR protein amino acid position. D, disordered, =, ordered. Dark and light grey boxes indicate the two helices and interlinker of TPR repeat 13, respectively. H1, Helix 1, H2, Helix 2, IL, interlinker.

Figure 6. 3D model of mTOR based on HHpred results

Panels A-D show different views of the TOR 3D model built using the HHpred results listed in supplementary Table 3. Structures were corrected manually to minimize insertions of the template structure that are not found in the target protein. Panels A and B show side views while Panels C and D show top and bottom views, respectively. The 3D model was built in pymol. Orange and green colors correspond to HEAT and TPR/FAT domains, respectively. Purple and blue colors correspond to the FRB and FATC domains, respectively. Yellow and Red corresponds to the N- and C-lobe of the PI3K catalytic domain, respectively.

Table 1

Comparison of methods to detect HEAT repeats in protein of known structure

Name	PDB ^a	Actual Repeats	HHpred	ARD	HHrepID	REP	Pfam	Prosite	SMART
CAND1	1u6g, C	27	27	13	3	17	0	0	0
CSE1	1wa5, C	20	20	0	0	0	0	0	0
TNPO1	2z5k, A	19	19	11	14	11	2	0	0
IMB1	1qgr, A	19	19	5	4	9	1	1	0
2AAA	1b3u, A	15	15	11	1	13	4	11	0
AP2A2	2vgl, A	14	14	5	0	4	0	0	0
CTNB1	1g3j, A	12	12	5	11	10	6	9	12
VATH	1ho8, A	10	10	1	0	0	0	0	0
IMA1	1bk5, A	10	10	6	10	9	8	2	8
PKP1	1xm9, A	9	9	3	7	0	4	3	2
EF3A	2ix3, A	8	8	4	1	6	0	1	0
HSPB1	1xqr, A	6	6	3	0	0	1	0	0
METH1	1te4, A	3	3	2	3	0	3	0	3
Total Repeats		172	172	69	54	79	29	27	25
Sensitivity		N/A	100	40%	31%	46%	17%	16%	15%

^aFour letter PDB structure code, protein chain identifier.

Table 2

Comparison of methods to detect TPR repeats in proteins of known structure.

Name	PDB ^a	Actual Repeats	HHpred	TPRpred	HHrepID	REP	Pfam	Prosite	SMART
OGT1	1w3b, A	12	12	12	12	10	9	9	9
MALT	1hz4, A	8	8	8	0	0	0	0	0
PEX5	1fch, A	7	7	7	7	6	3	5	4
HCPB	1ouv, A	7	7	7	7	7	6	1	7
HCPB	1klx, A	4	4	3	4	3	3	0	3
PPP5	1wao, A	3	3	3	3	0	2	3	3
PPID	1ihg, A	3	3	3	0	3	2	3	3
PLAF7	2fhn, A	3	3	3	0	3	2	3	3
STIP1	1elr, A	3	3	3	0	3	3	3	3
STUB1	2c2l, A	3	3	3	3	3	2	3	3
FKBP4	1qz2, A	3	3	3	0	3	3	3	3
SYCD	2vgx, A	3	3	3	3	0	2	0	0
FIS1	1nzn, A	2	2	1	0	0	1	0	0
Total Repeats		61	61	59	39	41	38	33	41
Sensitivity		N/A	100%	97%	64%	67%	62%	54%	67%

^aFour letter PDB structure code, protein chain identifier.