Published in final edited form as:

J Mol Biol. 2013 May 27; 425(10): 1826–1838. doi:10.1016/j.jmb.2013.02.013.

All Repeats are Not Equal: A Module-Based Approach to Guide Repeat Protein Design

Nicholas Sawyer^{#1,2}, Jieming Chen^{#1,3}, and Lynne Regan^{b,1,2,3,4}

¹ Integrated Graduate Program in Physical and Engineering Biology, Yale University, 266 Whitney Avenue, New Haven, CT 06511, USA

² Department of Molecular Biophysics and Biochemistry, Yale University, 266 Whitney Avenue, New Haven, CT 06511, USA

³ Program in Computational Biology and Bioinformatics, Yale University, 266 Whitney Avenue, New Haven, CT 06511, USA

⁴ Department of Chemistry, Yale University, 266 Whitney Avenue, New Haven, CT 06511, USA

[#] These authors contributed equally to this work.

Abstract

Repeat proteins composed of tandem arrays of a short structural motif often mediate protein-protein interactions. Past efforts to design repeat protein-based molecular recognition tools have focused on the creation of templates from the consensus of individual repeats, regardless of their natural context. Such an approach assumes that all repeats are essentially equivalent. In this study we present the results of a ‘module-based’ approach, in which modules composed of tandem repeats are aligned to identify repeat-specific features. Using this approach to analyze tetratricopeptide repeat modules that contain 3 tandem repeats (3TPRs), we identify two classes of 3TPR modules with distinct structural signatures that are correlated with different sets of functional residues. Our analyses also reveal a high degree of correlation between positions across the entire ligand-binding surface, indicative of a coordinated, coevolving binding surface. Extension of our analyses to different repeat protein modules reveals more examples of repeat-specific features, especially in armadillo repeat (ARM) modules. In summary, the module-based analyses that we present effectively capture key repeat-specific features that will be important to include in future repeat protein design templates.

Introduction

Repeat proteins mediate protein-protein and protein-nucleic acid interactions that are central to a variety of cellular processes.¹⁻⁵ They contain tandem repetitions of a short structural motif. Examples include the tetratricopeptide repeat (TPR), the leucine-rich repeat (LRR), the WD40 repeat, the ankyrin repeat (ANK), and the armadillo repeat (ARM). Each of these motifs has its own “signature” – a small set of highly conserved (typically hydrophobic) amino acids that specify the repeat structure.^{4,6-8} Here, we define a repeat protein “module”

^b Corresponding author: Department of Molecular Biophysics and Biochemistry, Yale University, 266 Whitney Avenue, New Haven, CT 06511 (lynne.regan@yale.edu, (203)-432-9843).

URLs

SMART : <http://smart.embl.de/>

WebLogo3 : <http://weblogo.threeplusone.com/>

Cytoscape : <http://www.cytoscape.org/>

PyMol : <http://pymol.org/>

as an array of tandem repeats and classify modules based on their repeat type and the number of repeats.

Designed repeat protein modules have been used as molecular recognition tools in a wide variety of applications, including affinity chromatography, Western blotting, histochemistry, and targeted inhibition of protein-protein interactions *in vivo*.⁹⁻¹¹ Designed repeat protein modules have many advantages over traditional molecular recognition tools (e.g. antibodies), including thermal and chemical robustness, facile and cheap production in large quantities from *E. coli*, and disulfide-independent folding for intracellular applications.¹²⁻¹⁴ Furthermore, repeat protein modules have a well-defined subset of ligand-binding residues,¹⁵⁻¹⁶ which allows for focused design and/or selection approaches to obtain new modules with desired function.

An important element in designing any protein to have a desired binding activity is the selection of a template onto which functional residues can be grafted.¹⁷ Although natural repeat proteins have been used successfully as design templates, they may lack the necessary stability to tolerate the introduction of substantial changes because the template protein is presumably only as stable as it needs to be to perform its *in vivo* function. Many groups have shown that a protein representing a consensus sequence is often more stable than any individual protein from which the consensus is derived.¹⁸⁻²² Extending this consensus-based approach to repeat proteins, researchers have designed consensus repeats for many different repeat motifs and created modules by assembling tandem arrays of these consensus repeats.^{12,15,23-27}

One fundamental assumption of the consensus repeat-based designs is that all repeats are equal. The validity of this notion is supported by the observation that the structures of repeat proteins composed of tandem consensus repeats reproduce the structures of their natural counterparts.^{12,13,27-29} Several repeat protein designs incorporated repeat-specific elements on a somewhat *ad hoc*, or protein-by-protein, basis. Inclusion of special N- and/or C-terminal capping repeats to improve solubility has been a fairly common strategy.^{12,30-33} However, introduction of functional residues onto consensus repeat proteins, either by randomization and selection or statistical analysis of repeat protein modules with known function, has yielded mixed results. Cortajarena and colleagues showed that Hsp90-binding residues could be grafted onto a consensus TPR module to create a protein that specifically bound the C-terminal peptide of Hsp90.³⁴ However, the binding affinity of the interaction was over an order of magnitude weaker than natural Hsp90-binding TPR modules. Higher affinity was accomplished by incorporation of additional design features.³⁵ A similar result was reported by Zahnd and colleagues,³⁶ whose designed ankyrin repeat protein (DARPin) exhibited the desired specificity for the Her2 extracellular domain but weaker affinity than that of known high-affinity ankyrin-ligand interactions.³⁷ In this example, the binding affinity of the DARPin for Her2 was greatly increased using random mutagenesis and selection. Intriguingly, several of the mutations that increased binding affinity occurred at ankyrin signature positions.³⁸ These results suggest that repeat-specific modifications are very important for designing high-affinity repeat protein modules.

We have developed a novel “module-based approach” in which we treat repeat protein sequences as modules of tandem repeats instead of separating and aligning all individual repeats. This approach involves an important change of mindset in understanding repeat protein sequences, namely that repeat-specific features can be distilled from repeat protein module sequence alignments and incorporated into designs from the beginning. Incorporation of these features early in the design process can be very powerful because it limits the number of rounds of design and characterization of individual proteins required the desired binding properties. We applied this module-based approach to the analysis of

different classes of repeat proteins. As an example of how important repeat-specific information can be extracted from repeat protein module alignments, our most expansive analyses focused on modules composed of 3 tandem TPR motifs (3TPRs). We discovered a number of key repeat-specific sequence features that are overlooked with a consensus repeat-based approach. In particular, we found that there are two distinct classes of naturally-occurring 3TPR modules, one of which has a significant, repeat-specific signature change that is highly correlated with, but not exclusive to, a particular set of functional residues. This observation is of relevance for protein design because it suggests that certain modules are more favorable templates for hosting certain ligand-binding functions. We also performed preliminary module-based analysis on two other types of repeat proteins, ANK and ARM proteins, to demonstrate the generality of our approach. We found that most ANK repeats have very high sequence similarity regardless of module size, whereas ARM repeats are highly divergent and are prime candidates for design improvement using the module-based approach. Together, these findings suggest that a module-based approach is useful in generating templates for repeat protein design because not all repeats are equal.

Results and Discussion

We examined 3TPR modules because they are the most common TPR modules in natural proteins (40% of TPR modules in all TPR-containing proteins have 3 TPR repeats, see Supplementary Figure 1A). Several 3TPR modules have been well-studied in terms of their structure and ligand binding specificity, which allows us to interpret sequence variations in the context of known structure and function. We identified 4171 proteins containing exactly 3 canonical TPR motifs (i.e. a canonical TPR motif is exactly 34 amino acids long). Of these, we obtained a database of 974 non-redundant proteins that had no gaps between adjacent repeats (see Materials and Methods for further description of database construction). For subsequent discussion of *equivalent* positions in different 3TPR repeats, we use the sequence position number in the context of a single repeat (i.e. a number from 1 to 34). To refer to *specific* positions in 3TPR modules, two identifiers will be given. The first identifier is simply the sequence position number in the context of a 3TPR module (i.e. a number between 1 and 102). To clarify comparisons between equivalent positions in different repeats, a second identifier is given in the form of x@TPRy, where x indicates the sequence position within a single repeat (i.e. a number between 1 and 34) and y indicates the repeat number within a 3TPR module (i.e. a number between 1 and 3). For example, 3@TPR2 denotes position 3 in repeat 2 of a 3TPR module.

All Repeats are Not Equal

To compare each of the 3TPR repeats to an alignment of all canonical TPRs (allTPR), we calculated the relative entropy (a measure of sequence entropy) for each position (Equation 1) using the difference between the observed amino acid frequency distribution at each position and the amino acid frequency distribution of the SMART database (the reference distribution). To fully present the nuances of these data, we show both bar graph and sequence logo representations. The bar graphs show the degree of conservation (or variability) at each position, whereas the sequence logos show which amino acids are present at each position.³⁹

We find that the highly conserved TPR signature positions (e.g. 8, 20, 27) stand out with high relative entropy values and strong preferences for one or two amino acids in the allTPR alignment and each 3TPR repeat alignment (Figure 1). There are also positions (e.g. 22) that are highly variable in the allTPR alignment and each 3TPR repeat alignment. However, for most positions, we find that the relative entropy values for one or more of the 3TPR repeats differ significantly from the relative entropy values for allTPR and the other 3TPR repeats.

For simplicity, we grouped the trends that distinguish 3TPR repeats from allTPR and from each other into three types. Type I trends, exemplified by position 9, have higher relative entropy values in the first 3TPR repeat only. Type II trends, exemplified by position 23, have higher relative entropy values in the second 3TPR repeat only. Finally, type III trends, exemplified by position 4, have higher relative entropy values in both the second and third repeats of 3TPR but not the first. Together, these three types of trends distinguish 3TPR repeats from allTPR, providing the first evidence that not all repeats are equal in the context of a repeat protein module. Moreover, these trends typically show high relative entropy values and significant amino acid preferences, suggesting that potential functional or structural features of 3TPRs would be missed in a consensus repeat-based approach because of the inherent averaging of such an approach.

Two Structural Classes of 3TPR Modules

What is the significance of the repeat-specific features of 3TPRs that our analyses have revealed? To investigate this issue in greater depth, we compared the amino acid distributions for TPR signature positions in allTPR and each 3TPR repeat (examples in Supplementary Figure 2). The amino acid distributions for most of the signature positions match well for allTPR and each of the 3TPR repeats (for example, see position 17 in Supplementary Figure 2). However, we also identified 3 signature positions in the 3TPR repeats whose distributions differed significantly from the allTPR distributions – positions 7 (7@TPR1), 11 (11@TPR1), and 58 (24@TPR2) (Supplementary Figure 2). In all three cases, previously underrepresented amino acids are now observed at much higher frequency. The fact that only three TPR signature positions deviate from expectation suggests that, on the whole, the TPR signature positions are conserved and specify the TPR fold. However, the three deviations represent repeat-specific signatures and strongly suggest critical roles for these positions in the stability and/or function of 3TPR modules. To investigate how these repeat-specific signature changes might impact the rest of a 3TPR module, we used mutual information (MI, see Equation 2) to calculate the mutual dependence of two positions on each other.^{40,41} Here we use MI scores to rank residue pairs for possible covariation: a high MI score relative to the mean suggests high possibility of covariation for a residue pair while a low MI score relative to the mean indicates little, if any, covariation. Of the 3 positions with changes to the TPR signature (7 or 7@TPR1, 11 or 11@TPR1, and 58 or 24@TPR2), only position 58 (24@TPR2) shows high covariation with other 3TPR positions (with a maximum MI score almost 4.5 standard deviations above the mean MI score). The maximum MI score for position 58 (24@TPR2) is obtained with position 74 (6@TPR3), a position located on the concave ligand-binding face of natural 3TPR modules.

We investigated the amino acid distributions at position 74 for each of the 5 most frequent residues at position 58 (A – 25%, C – 24%, F – 11%, L – 14%, and Y – 17%). We found that A or C at position 58 (24@TPR2) is strongly correlated with R at position 74 (6@TPR3) whereas F, L, or Y at position 58 (24@TPR2) is strongly correlated with N at position 74 (6@TPR3) (Figure 2A). By considering structural models, we can readily understand the covariation between residue pairs at positions 58 (24@TPR2) and 74 (6@TPR3) in terms of side chain packing (Figure 2B and 2C). For the A58-R74 pair, the aliphatic part of the R74 side chain packs closely against A58 while the guanidinium group remains exposed for ligand binding (Figure 2B). A large hydrophobic residue at position 58 (24@TPR2) would prevent this conformation of the R74 side chain and thus could impair interaction with a ligand. On the other hand, the smaller N74 side chain would not clash with a large hydrophobic residue at position 58 (24@TPR2) but could still interact with ligands, as others have hypothesized (Figure 2C).⁴² The strong correlation between positions 58 (24@TPR2) and 74 (6@TPR3) suggests a novel interplay between structure and function

in 3TPRs in which fold-specifying signature residues (e.g. A or C at position 58 (24@TPR2)) can co-evolve with functional residues (e.g. R at position 74 (6@TPR3)).

Dichotomy of 3TPR Module Functions

We further examined the influence of position 58 (24@TPR2) on 3TPR module composition by following the natural divide of our 3TPR dataset into two large subsets: the AC58 subset of 473 sequences with A or C at position 58 (24@TPR2) and the FLY58 subset of 406 sequences with F, L, or Y at position 58 (24@TPR2).

In the AC58 subset (Figure 3A), we discovered that there are actually two distinct TPR signatures – a “classic” signature with F, L, or Y at position 24 and an “alternative” signature with A or C at position 24 and a R-D salt bridge between positions 7 and 23 of the same repeat. In this subset, there are a number of positions with even higher relative entropy than many of the TPR signature positions, the most notable of which are positions 41 (7@TPR2) and 57 (23@TPR2) (asterisks in Figure 3A). Greater than 94% of AC58 sequences have R at position 41 (7@TPR2) and D at position 57 (23@TPR2). Furthermore, there are only 3 sequences (out of 448) that have one of these residues without the other. In many natural 3TPR modules of known structure (e.g. PDB: 3SZ7, 1ELW, 2VYI), these residues form a salt bridge in repeat 2 (Figure 3B). This unique combination of a R41-D57 salt bridge in the same repeat as A or C at position 58 (24@TPR2) defines an alternative TPR signature for the second repeat of 3TPR modules in this subset. Because this alternative signature is only observed in the second repeat of 3TPR modules, it appears that there are two distinct classes of natural 3TPR modules – one with three classic repeats and a second with an alternative repeat 2.

Within the AC58 subset, we also discovered that the majority (59%) of the sequences have a set of amino acids previously named the “dicarboxylate clamp”, which is known to interact with the C-terminal EEVD peptides of the chaperones Hsp70 and Hsp90 (arrows in Figure 3A, Figure 3C).^{43,44} Specifically, the “dicarboxylate clamp” consists of K at position 5 (5@TPR1), N at position 9 (9@TPR1), N at position 40 (6@TPR2), and K at position 70 (2@TPR3). This set of functional residues is one of very few sets of residues in repeat proteins for the function is well-characterized. While each of these amino acids is present at approximately 70% frequency at its respective position, the 59% of sequences observed to have all four residues is significantly greater than the expected co-occurrence of 24% if these residues were independent. Given that almost 60% of the AC58 sequences have the “dicarboxylate clamp” residues, it is apparent that the AC58 is particularly well-suited, but by no means limited, to binding C-terminal EEVD peptides.

In stark contrast, there are very few positions in the FLY58 subset (Figure 3D) with high relative entropy other than the TPR signature positions. As described previously,¹⁶ the high degree of variability at ligand-binding positions (e.g. positions 2 and 9) in repeat proteins has been previously interpreted to mean that these protein possess a wide variety of ligand binding specificities. Our results support this interpretation and also suggest that certain modules are well-suited for certain binding functions.

Covariation Across the 3TPR Concave Face: Coordinated Ligand Binding

The significant covariation revealed by the initial MI analysis led us to hypothesize that there might be additional covariation in 3TPR modules that does not involve signature positions but which may still have a significant impact on 3TPR function. We therefore expanded the MI analysis to include all pairs of positions in 3TPR modules and discovered strong covariation between many non-signature positions on the concave ligand-binding face. Figure 4A shows a network where each 3TPR position is represented by a node and

each edge connects pairs of positions with MI scores greater than 3 standard deviations above the mean MI score (see Methods for more information). In this network, we identified three clusters of highly connected nodes, or “hubs”, that indicate covariation between residues on the concave ligand-binding surface (Figure 4B). The “dicarboxylate clamp” residues (5 or 5@TPR1, 9 or 9@TPR1, 40 or 6@TPR2, and 70 or 2@TPR3) and position 74 (6@TPR3) all appear prominently among these correlated positions. Position 74 (6@TPR3), in fact, is the most well-connected node, suggesting that this position has the most pervasive impact on the amino acid preferences at other 3TPR positions. Overall, the strong correlation between ligand-binding positions that we observe strongly suggests that ligand binding is coordinated across the entire binding surface in 3TPR modules.

Extension of Module-Based Analyses to Other Repeat Proteins

We applied our methodology to TPR modules containing different numbers of tandem repeats (4TPR and 5TPR). For 4TPR and 5TPR modules, we found that the relative entropy values of each repeat are quite similar to allTPR, much like the repeats in the FLY58 subset of 3TPRs (Supplementary Figure 3). Thus, unlike 3TPR modules, 4TPR and 5TPR modules do not appear to have a significant number of modules with changes to the classic TPR signature. This information is valuable in that it confirms that, for certain module types, the assumption that all repeat are roughly equal is a valid one.

We also examined ankyrin (ANK) modules with emphasis on the two most common ANK modules in SMART (2ANK and 3ANK, see Supplementary Figure 1B and ref. 22). For both of these module types, we observed that each repeat is quite similar to an alignment of ANK repeats (allANK) (3ANK results shown in Figure 5). As with TPRs, signature positions (i.e. 4-7, 9, 13, 21-22, and 25) are readily distinguishable in the sequence logos. However, unlike with 3TPR modules, we observed only small fluctuations in relative entropy for equivalent positions in different repeats. This result suggests that all repeats are roughly equivalent in 2ANK and 3ANK, similar to the case for 4TPRs and 5TPRs. It is important to note, though, that this analysis likely does not include the N- and C-terminal repeats of natural ANK modules because they tend to be so highly divergent that they are not detected by the SMART algorithm.

Armadillo (ARM) repeats present a contrasting picture to either TPRs or ANKs. In 8ARM modules, which are one of the two most common ARM modules (Supplementary Figure 1C), the individual 8ARM repeats are vastly different from each other and from an alignment of individual ARM repeats (allARM) (Figure 6). Unlike TPR and ANK repeats, the ARM signature is relatively sparse.⁴⁵ This is obvious in the allARM sequence logo, where very few positions have significant amino acid preferences. Nonetheless, positions 4, 8, 9, 12, 17, 20, 21, and 38 all show some degree of conservation for either a single amino acid or a class of amino acids (e.g. hydrophobic) in allARM and can thus serve as starting points for comparison against the 8ARM repeats. However, even at some of these partially conserved positions in allARM, there are often striking differences between individual 8ARM repeats. For instance, position 21 has a very strong bias for leucine in 7 of the 8 repeats but has a strong preference for histidine, tyrosine, and phenylalanine in repeat 4. We also discovered that our 8ARM dataset had an abundance of importin sequences (approximately 18%), which share a common function. Despite their common function, alignments of each importin repeat are clearly different from an alignment of all importin repeats (Supplementary Figure 4). Because of the large number of differences between the 8ARM repeats, it is difficult to infer the structural or functional consequences of these sequence differences. Thus, 8ARM modules are complex and not just simple arrays of idealized ARM repeats.

Methods

Motif Sequences and Sequence Dataset Construction

All motif sequences (TPR, ANK, and ARM) were extracted from the genomic mode of SMART database on November 28, 2011.^{46,47} For each dataset described below, we filtered the SMART results to obtain datasets where all sequences in a given dataset are the same length. Thus, when we refer to an “alignment”, we simply lined up all sequences by position for each dataset.

There are a total of 12,766 proteins with TPR repeats in the SMART database, distributed across 956 species from all 3 domains of life. Of these sequences, 4171 proteins contain exactly 3 canonical TPR repeats, where canonical is defined as exactly 34 amino acids in length. The TPRs in some of these sequences are separated from adjacent repeats by gaps. To obtain a conservative set of natural 3TPR modules for this analysis, we filtered the 4171 sequences to obtain only those sequences with no gaps between adjacent TPRs. Subsequently, we removed redundant sequences, leaving a final dataset of 974 non-identical 3TPR modules. We constructed 4TPR and 5TPR module sets using the same set of filters and obtained 287 and 124 sequences, respectively.

There are a total of 17,387 proteins with ANK repeats in the SMART database, distributed across 629 species from all 3 domains of life. The most commonly cited canonical size of an ANK repeat is 33 amino acids.¹³ However, the most common ANK repeat length in the SMART database is 30 amino acids, coupled with the most common gap size of 3 between adjacent repeats. This is probably a result of the hidden Markov model algorithm in the SMART database. Based on previous studies on ANK repeats, the major secondary structure elements are within the first 30 amino acids of each repeat, giving us confidence that studying the 30 amino acid repeats will capture the major features of ANK modules. Hence, our ANK dataset is strictly obtained by the criteria that all ANK repeats are exactly 30 amino acids long and adjacent repeats are separated by exactly 3 amino acids. After applying these filters to sequences containing 2 or 3 ANK repeats (2ANK and 3ANK modules, respectively) and removing redundant sequences, we obtained 850 and 575 unique sequences, respectively.

Compared to the large number of TPR- and ANK-containing proteins, there are only 1,912 proteins with ARM repeats in the SMART database, all of which are eukaryotic. The length of an ARM repeat has been estimated to be approximately 42 amino acids.⁴⁸ However, the most common length in the SMART database is 41. As with ANK repeats, we noted that the three core helices of an ARM repeat are found in the first 40 amino acids. Hence, we included only sequences in which all repeats were greater than 40 amino acids long. Unlike the TPR and ANK proteins, there were two distinct, commonly-occurring ARM repeat modules in the SMART database, composed of 3 ARM repeats (3ARM, 345 sequences) and 8 ARM repeats (8ARM, 359 sequences). Because the most frequent size of ARM repeats was not the expected 42 amino acids, we defined “tandem” repeats as repeats with 2 or fewer amino acids between adjacent repeats. However, approximately 83% of 3ARM proteins did not fulfill this criterion. Hence, we decided to focus on 242 8ARM proteins that had repeats of at least 40 amino acids (mean=41) and a gap allowance of 2 amino acids between adjacent repeats.

Sequence Logos

Sequence logos were generated using the *WebLogo 3* software⁴⁹ and colored by amino acid chemistry, where polar residues (G, S, T, Y, C) are colored green, neutral residues (Q, N)

purple, basic residues (K, R, H) blue, acidic residues (D, E) red, and hydrophobic residues (A, V, L, I, P, W, F, M) black.

Relative Entropy and Mutual Information

Relative entropy at position i , D_i , is given by:

$$D_i(p|f) = \sum_x p(x) \ln \frac{p(x)}{f(x)} \quad (1)$$

where $p(x)$ represents the frequency of amino acid x at position i and $f(x)$ denotes the reference amino acid frequency computed from all the sequences in the non-redundant genomic mode of the SMART database (in descending order): Leu (0.099), Ala (0.078), Gly (0.071), Val (0.070), Glu (0.065), Ser (0.064), Arg (0.060), Ile (0.058), Asp (0.058), Lys (0.054), Thr (0.054), Asn (0.041), Pro (0.041), Gln (0.040), Phe (0.038), Tyr (0.030), His (0.027), Cys (0.023), Met (0.021), Trp (0.011).

Mutual information between positions i and j , $M_{i,j}$, is given by:

$$M_{i,j}(X;Y) = \sum_x \sum_y p(x_i, y_j) \ln \frac{p(x_i, y_j)}{p(x_i) p(y_j)} \quad (2)$$

where $p(x_i, y_j)$ denotes the frequency of the simultaneous occurrence of amino acid x at position i and amino acid y at position j , $p(x_i)$ denotes the frequency of amino acid x at position i , and $p(y_j)$ denotes the frequency of amino acid y at position j . These calculations were carried out using a custom Matlab script, which is available upon request. In Equations 1 and 2, the natural logarithm is used to calculate information in nats, while the original information theoretic metric uses a logarithm of base 2 to calculate information in bits.⁵⁰ All relative entropy and mutual information calculations were applied to the entire 3TPR dataset of 974 sequences.

In the mutual information network, each node represents one of the 102 positions in a 3TPR module. An edge is drawn between any pair of positions for which the MI score of that pair is more than three standard deviations greater than the mean MI score for all position pairs. The network was constructed using the software *Cytoscape* 2.8.3.

Protein structure visualization

Protein models were constructed using *Pymol* 1.3.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank the Regan lab and Roger Alexander for critical reading of and suggestions on this manuscript. Thanks to Thomas Magliery and Ajit Divakaruni for background work in the Regan lab related to this project. This work was supported, in part, by the Raymond and Beverly Sackler Institute for Biological, Physical and Engineering Sciences.

References

1. Grove TZ, Cortajarena AL, Regan L. Ligand binding by repeat proteins: natural and designed. *Curr. Opin. Struct. Biol.* 2008; 18:507–515. [PubMed: 18602006]

2. Kopan R, Ilagan M. The canonical Notch signaling pathway: unfolding the activation mechanism. *Cell*. 2009; 137:216–233. [PubMed: 19379690]
3. Schmid AB, Lagleder S, Gräwert MA, Röhl A, Hagn F, Wandinger SK, Cox MB, Demmer O, Richter K, Groll M, Kessler H, Buchner J. The architecture of functional modules in the Hsp90 co-chaperone Sti1/Hop. *EMBO J*. 2012; 31:1506–1517. [PubMed: 22227520]
4. Peifer M, Berg S, Reynolds AB. A repeating amino acid motif shared by proteins with diverse cellular roles. *Cell*. 1994; 76:789–791. [PubMed: 7907279]
5. Akira S, Takeda K. Toll-like receptor signalling. *Nat. Rev. Immunol*. 2004; 4:499–511. [PubMed: 15229469]
6. Sikorski RS, Boguski MS, Goebel M, Hieter P. A repeating amino acid motif in CDC23 defines a family of proteins and a new relationship among genes required for mitosis and RNA synthesis. *Cell*. 1990; 60:307–317. [PubMed: 2404612]
7. Takahashi N, Takahashi Y, Putnam FW. Periodicity of leucine and tandem repetition of a 24-amino acid segment in the primary structure of leucine-rich alpha 2-glycoprotein of human serum. *Proc. Natl. Acad. Sci. USA*. 1985; 82:1906–1910. [PubMed: 3856868]
8. Lux SE, John KM, Bennett V. Analysis of cDNA for human erythrocyte ankyrin indicates a repeated structure with homology to tissue-differentiation and cell-cycle control proteins. *Nature*. 1990; 344:36–42. [PubMed: 2137557]
9. Jackrel ME, Valverde R, Regan L. Redesign of a protein-peptide interaction: Characterization and applications. *Prot. Sci*. 2009; 18:762–774.
10. Veessler D, Dreier B, Blangy S, Lichère J, Tremblay D, Moineau S, Spinelli S, Tegoni M, Plückthun A, Campanacci V, Cambillau C. Crystal structure and function of a DARPin neutralizing inhibitor of lactococcal phage TP901-1: comparison of DARPins and camelid VHH binding mode. *J. Biol. Chem*. 2009; 284:30718–30726. [PubMed: 19740746]
11. Theurillat J-P, Dreier B, Nagy-Davidescu G, Seifert B, Behnke S, Zürrer-Härdi U, Ingold F, Plückthun A, Moch H. Designed ankyrin repeat proteins: a novel tool for testing epidermal growth factor receptor 2 expression in breast cancer. *Mod. Pathol*. 2010; 23:1289–1297. [PubMed: 20495541]
12. Main ERG, Xiong Y, Cocco MJ, D'Andrea L, Regan L. Design of Stable α -Helical Arrays from an Idealized TPR Motif. *Structure*. 2003; 11:497–508. [PubMed: 12737816]
13. Kohl A, Binz HK, Forrer P, Stumpp MT, Plückthun A, Grütter MG. Designed to be stable: Crystal structure of a consensus ankyrin repeat protein. *Proc. Natl. Acad. Sci. USA*. 2003; 100:1700–1705. [PubMed: 12566564]
14. Parmeggiani F, Pellarin R, Larsen AP, Varadamsetty G, Stumpp MT, Zerbe O, Caflisch A, Plückthun A. Designed armadillo repeat proteins as general peptide-binding scaffolds: consensus design and computational optimization of the hydrophobic core. *J. Mol. Biol*. 2008; 376:1282–1304. [PubMed: 18222472]
15. Binz HK, Stumpp MT, Forrer P, Amstutz P, Plückthun A. Designing Repeat Proteins: Well-expressed, Soluble and Stable Proteins from Combinatorial Libraries of Consensus Ankyrin Repeat Proteins. *J. Mol. Biol*. 2003; 332:489–503. [PubMed: 12948497]
16. Magliery TJ, Regan L. Beyond consensus: statistical free energies reveal hidden interactions in the design of a TPR motif. *J. Mol. Biol*. 2004; 343:731–745. [PubMed: 15465058]
17. Fleischman SJ, Whitehead TA, Ekiert DC, Dreyfus C, Corn JE, Strauch E-M, Wilson IA, Baker D. Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science*. 2011; 332:816–821. [PubMed: 21566186]
18. Steipe B, Schiller B, Plückthun A, Steinbacher S. Sequence Statistics Reliably Predict Stabilizing Mutations in a Protein Domain. *J. Mol. Biol*. 1994; 240:188–192. [PubMed: 8028003]
19. Lehmann M, Kostrewa D, Wyss M, Brugger R, D'Arcy A, Pasamontes L, van Loon APGM. From DNA sequence to improved functionality: using protein sequence comparisons to rapidly design a thermostable consensus phytase. *Protein Eng*. 2000; 13:49–57. [PubMed: 10679530]
20. Loening AM, Fenn TD, Wu AM, Gambhir SS. Consensus guided mutagenesis of Renilla luciferase yields enhanced stability and light output. *Protein Eng. Des. Sel*. 2006; 19:391–400. [PubMed: 16857694]

21. Jackel C, Bloom JD, Kast P, Arnold FH, Hilvert D. Consensus protein design without phylogenetic bias. *J. Mol. Biol.* 2010; 399:541–546. [PubMed: 20433850]
22. Sullivan BJ, Nguyen T, Durani V, Mathur D, Rojas S, Thomas M, Syu T, Magliery TJ. Stabilizing Proteins from Sequence Statistics: The Interplay of Conservation and Correlation in Triosephosphate Isomerase Stability. *J. Mol. Biol.* 2012; 420:384–399. [PubMed: 22555051]
23. Stumpp MT, Forrer P, Binz HK, Plückthun A. Designing Repeat Proteins: Modular Leucine-rich Repeat Protein Libraries Based on the Mammalian Ribonuclease Inhibitor Family. *J. Mol. Biol.* 2003; 332:471–487. [PubMed: 12948496]
24. Urvoas A, Guellouz A, Valerio-Lepiniec M, Graille M, Durand D, Desravines DC, van Tilbeurgh H, Desmardil M, Minard P. Design, Production and Molecular Structure of a New Family of Artificial Alpha-helical Repeat Proteins (α Rep) Based on Thermostable HEAT-like Repeats. *J. Mol. Biol.* 2010; 404:307–327. [PubMed: 20887736]
25. Yadid I, Tawfik DS. Functional β -propeller lectins by tandem duplications of repetitive units. *Protein Eng. Des. Sel.* 2011; 24:185–195. [PubMed: 20713410]
26. Broom A, Doxey AC, Lobsanov YD, Berthin LG, Rose DR, Howell PL, McConkey BJ, Meiering EM. Modular Evolution and the Origins of Symmetry: Reconstruction of a Three-Fold Symmetric Globular Protein. *Structure.* 2012; 20:161–171. [PubMed: 22178248]
27. Mosavi LK, Cammett TJ, Desrosiers DC, Peng Z-Y. The ankyrin repeat as molecular architecture for protein recognition. *Prot. Sci.* 2004; 13:1435–1448.
28. Das AK, Cohen PTW, Barford D. The structure of the tetratricopeptide repeats of protein phosphatase 5: implications for TPR-mediated protein-protein interactions. *EMBO J.* 1998; 17:1192–1199. [PubMed: 9482716]
29. Kajander T, Cortajarena AL, Mochrie S, Regan L. Structure and stability of designed TPR protein superhelices: unusual crystal packing and implications for natural TPR proteins. *Acta Crystallogr D.* 2007; 63:800–811. [PubMed: 17582171]
30. Madhurantakam C, Varadamsetty G, Grütter MG, Plückthun A, Mittl PRE. Structure-based optimization of designed Armadillo-repeat proteins. *Prot. Sci.* 2012; 21:1015–1028.
31. Kramer MA, Wetzel SK, Plückthun A, Mittl PR, Grütter MG. Structural determinants for improved stability of designed ankyrin repeat proteins with a redesigned C-capping module. *J. Mol. Biol.* 2010; 404:381–391. [PubMed: 20851127]
32. Kloss E, Barrick D. C-terminal deletion of leucine-rich repeats from YopM reveals a heterogeneous distribution of stability in a cooperatively folded protein. *Prot. Sci.* 2009; 18:1948–1960.
33. Zweifel ME, Barrick D. Studies of the Ankyrin Repeats of the *Drosophila melanogaster* Notch Receptor. 2. Solution Stability and Cooperativity of Unfolding. *Biochemistry.* 2001; 40:14357–14367. [PubMed: 11724547]
34. Cortajarena AL, Kajander T, Pan W, Cocco MJ, Regan L. Protein design to understand peptide ligand recognition by tetratricopeptide repeat proteins. *Protein Eng. Des. Sel.* 2004; 17:399–409. [PubMed: 15166314]
35. Cortajarena AL, Yi F, Regan L. Designed TPR modules as novel anticancer agents. *ACS Chem. Biol.* 2008; 3:161–166. [PubMed: 18355005]
36. Zahnd C, Pecorari F, Straumann N, Wyler E, Plückthun A. Selection and characterization of Her2 binding-designed ankyrin repeat proteins. *J. Biol. Chem.* 2006; 281:35167–35175. [PubMed: 16963452]
37. Suzuki F, Goto M, Sawa C, Ito S, Watanabe H, Sawada J, Handa H. Functional interactions of transcription factor human GA-binding protein subunits. *J. Biol. Chem.* 1998; 273:29302–29308. [PubMed: 9792629]
38. Zahnd C, Wyler E, Schwenk JM, Steiner D, Lawrence MC, McKern NM, Pecorari F, Ward CW, Joos TO, Plückthun A. A Designed Ankyrin Repeat Protein Evolved to Picomolar Affinity to Her2. *J. Mol. Biol.* 2007; 369:1015–1028. [PubMed: 17466328]
39. Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* 1990; 18:6097–6100. [PubMed: 2172928]

40. Atchley WR, Wollenberg KR, Fitch WM, Terhalle W, Dress AW. Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. *Mol. Biol. Evol.* 2000; 17:164–178. [PubMed: 10666716]
41. White RA, Szurmant H, Hoch JA, Hwa T. Features of Protein-Protein Interactions in Two-Component Signaling Deduced from Genomic Libraries. *Methods Enzymol.* 2007; 422:75–101. [PubMed: 17628135]
42. Jinek M, Rehwinkel J, Lazarus BD, Izaurralde E, Hanover JA, Conti E. The superhelical TPR-repeat domain of O-linked GlcNAc transferase exhibits structural similarity to importin alpha. *Nat. Struct. Mol. Biol.* 2004; 1001–1007. [PubMed: 15361863]
43. Scheufler C, Brinker A, Bourenkov G, Pegoraro S, Moroder L, Bartunik H, Hartl FU, Moarefi I. Structure of TPR Domain-Peptide Complexes: Critical Elements in the Assembly of the Hsp70-Hsp90 Multichaperone Machine. *Cell.* 2000; 101:199–210. [PubMed: 10786835]
44. Brinker A, Scheufler C, von der Mülbe F, Fleckenstein B, Herrmann C, Jung G, Moarefi I, Hartl FU. Ligand discrimination by TPR domains: Relevance and selectivity of EEVD-recognition in Hsp70-Hop-Hsp90 complexes. *J. Biol. Chem.* 2002; 277:19265–19275. [PubMed: 11877417]
45. Andrade MA, Petosa C, O'Donoghue SI, Müller CW, Bork P. Comparison of ARM and HEAT Protein Repeats. *J. Mol. Biol.* 2001; 309:1–18. [PubMed: 11491282]
46. Schultz J, Milpetz F, Bork P, Ponting CP. SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl. Acad. Sci.* 1998; 95:5857–5864. [PubMed: 9600884]
47. Letunic I, Doerks T, Bork P. SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Res.* 2012; 40:D302–D305. [PubMed: 22053084]
48. Riggleman B, Wieschaus E, Schedl P. 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.* 1989; 3:96–113. [PubMed: 2707602]
49. Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: a sequence logo generator. *Genome Res.* 2004; 14:1188–1190. [PubMed: 15173120]
50. Cover, TM.; Thomas, JA. *Elements of Information Theory*. Second Edition. John Wiley & Sons, Inc.; Hoboken, NJ, USA: 2006. Entropy, Relative Entropy, and Mutual Information..

Figure 1. A comparison of sequence variability for allTPR and 3TPR

(A) allTPR sequence logos show the amino acid preferences for each sequence position in an alignment of all canonical TPR sequences. Positions 8 and 20 are examples of highly conserved positions while positions 2 and 9 are examples of highly variable positions. (B) 3TPR sequence logos show the amino acid preferences for each sequence position in an alignment of our dataset of 3TPR sequences. Position 20 is an example of similarly high conservation in allTPR and each 3TPR repeat while position 22 is an example of similarly low conservation in allTPR and each 3TPR repeat. In contrast, positions 4, 9, and 23 are examples where there are significantly higher amino acid preferences in only one or two of the 3TPR repeats. (C) The bar graph shows the relative entropy values of each sequence position in allTPR (dark blue) and 3TPR (repeat 1 – aqua, repeat 2 – yellow, repeat 3 – maroon). High relative entropy indicates a highly skewed amino acid distribution while low relative entropy indicates an amino acid distribution similar to the reference distribution (Equation 1). Positions 20 and 22 are examples of high and low relative entropy in allTPR

and each 3TPR repeat, respectively. Examples of novel 3TPR trends are indicated at representative positions: position 9 –higher relative entropy for repeat 1 of 3TPR only (Type I), position 23 – higher relative entropy values for repeat 2 of 3TPR only (Type II), and position 4 – higher relative entropy values for repeats 2 and 3 of 3TPR only (Type III).

Figure 2. Covariation of position 58 (24@TPR2) and position 74 (6@TPR3) in 3TPR modules
 (A) The 3D bar plot shows the frequency of all 20 amino acids at position 74 (6@TPR3) when position 58 (24@TPR2) is A (red), C (yellow), F (blue), L (green), and Y (gray). When position 58 (24@TPR2) is A or C, there is a clear preference for R at position 74 (6@TPR3). When position 58 (24@TPR2) is F, L, or Y, there is a clear preference for N at position 74 (6@TPR3). (B) A model of the interaction between A58 and R74 side chains in a representative 3TPR module (from the crystal structure of SGT2 from *Aspergillus fumigatus*, PDB: 3SZ7). For the models in Figure 4B and 4C, amino acid side chains are shown as semi-transparent, space-filling spheres with sticks showing side chain connectivity. (C) A model of the interaction between Y58 and H74 in a 3TPR module (from the crystal structure of SycD from *Yersenia enterocolitica*, PDB: 4AM9). No structures have been solved for 3TPRs with F, L, or Y at position 58 (24@TPR2) and N at position 74 (6@TPR3); however, asparagine can replace histidine at position 74 (6@TPR3) in the structure without steric clashes with position 58 (24@TPR2) or any other part of the 3TPR module.

Figure 3. Comparison of sequence variability in the 3TPR subsets AC58 and FLY58

(A) Sequence logos show the amino acid preferences for each sequence position in a multiple alignment of 3TPR modules in the AC58 subset. Notable sequence differences compared to the full dataset include the “dicarboxylate clamp” residues (K at position 5 or 5@TPR1, N at position 9 or 9@TPR1, N at position 40 or 6@TPR2, and K at position 70 or 2@TPR3) (arrows) and the salt bridge between R at position 41 (7@TPR2) and D at position 57 (23@TPR2) (asterisks). (B) A model of the salt bridge between R at position 41 (7@TPR2) and D at position 57 (23@TPR2) in a 3TPR module (from the crystal structure of SGT2 from *Aspergillus fumigatus*, PDB: 3SZ7). Salt bridge side chains are shown as sticks with yellow dashed lines indicating interaction between these residues. Alanine residues at positions 42 (8@TPR2) and 58 (24@TPR2) are shown as sticks with semi-transparent, space filling spheres to illustrate the close packing between helices within repeat 2 of this 3TPR module. (C) A model of the “dicarboxylate clamp” residues in a 3TPR module interacting with the C-terminal aspartate residue of the peptide ligand GPTIEEVD (from the crystal structure of the TPR1-GPTIEEVD complex, PDB: 1ELW). Carbon atoms are colored green for the 3TPR module and pink for the peptide. A pink ribbon indicates the direction of the peptide backbone. Yellow dashed lines indicate ionic/hydrogen bonding interactions between the TPR residues and the C-terminal aspartate. (D) In the sequence logos for the FLY58 subset, all of the repeats are very similar to each other and to an alignment of all individual TPR repeats (see Figure 1A).

Figure 4. Network of covarying residues in 3TPRs

(A) In the network, each 3TPR position is represented as a node and colored by repeat (repeat 1 – red, repeat 2 – blue, repeat 3 – green). An edge connects two nodes if the mutual information (MI) between those sequence positions is greater than three standard deviations above the mean MI for all pairs of sequence positions. Only the positions with edges are numbered. Node size increases with an increasing number of edges. The α -helices and loop regions of a 3TPR module are shown as wide and thin arcs, respectively, around the outside of the network. The A helices are labeled and colored in a light shade of their respective repeat colors while B helices are labeled and colored in a darker shade of their respective repeat colors. The N-terminus of the 3TPR module is indicated with an arrow. The C-terminus of the 3TPR module is indicated with a black circle. “Hubs,” defined here as nodes with 4 or more connections, are localized to the A helices with approximately α -helical periodicity ($i, i+3/i+4$) (e.g. 5 and 9; 36, 39, and 43; 70 and 74). (B) A model of the network “hubs” on a representative 3TPR module (PDB: 1ELW) shows the positions of these residues on the concave face. Repeats are colored as in the network. Hubs are shown as yellow spheres labeled with corresponding position numbers.

Figure 5. A comparison of sequence variability for allANK and 3ANK

(A) allANK sequence logos show the amino acid preferences for each sequence position in a multiple alignment of all 30-amino acid ANK sequences. Signature positions (4-7, 9, 13, 21-22, and 25) show high degrees of conservation while positions 11 and 15 are examples of highly variable positions. (B) 3ANK sequence logos show the amino acid preferences for each sequence position in a multiple alignment of our dataset of 3ANK modules. Similar amino acid preferences and levels of variability are observed for almost all equivalent positions in allANK and each 3ANK repeat.

Figure 6. A comparison of sequence variability for allARM and 8ARM

(A) allARM sequence logos show the amino acid preferences for the first 40 sequence positions in a multiple alignment of all ARM sequences that are 40 or more amino acids long. Positions 4, 8, 9, 12, 17, 20, 21, and 38 show moderate conservation for either a single amino acid or a class of amino acids (e.g. hydrophobic). However, the amino acid preferences for these “signature” positions are not as strong as the amino acid preferences at TPR and ANK signature positions. (B) 8ARM sequence logos show the amino acid preferences for the first 40 sequence positions for each repeat in a multiple alignment of our dataset of 8ARM modules. While many positions in each repeat show very high preferences

for certain amino acids, these positions are different for each repeat. Furthermore, each repeat shares some preferences with the allARM sequence logos but not to the same degree.