

# Beyond Consensus: Statistical Free Energies Reveal Hidden Interactions in the Design of a TPR Motif

Thomas J. Magliery<sup>1\*</sup> and Lynne Regan<sup>1,2</sup>

<sup>1</sup>Department of Molecular  
Biophysics and Biochemistry  
Yale University  
P.O. Box 208114, New Haven  
CT 06520-8114, USA

<sup>2</sup>Department of Chemistry  
Yale University, New Haven  
CT, USA

Consensus design methods have been used successfully to engineer proteins with a particular fold, and moreover to engineer thermostable exemplars of particular folds. Here, we consider how a statistical free energy approach can expand upon current methods of phylogenetic design. As an example, we have analyzed the tetratricopeptide repeat (TPR) motif, using multiple sequence alignment to identify the significance of each position in the TPR. The results provide information above and beyond that revealed by consensus design alone, especially at poorly conserved positions. A particularly striking finding is that certain residues, which TPR-peptide co-crystal structures show are in direct contact with the ligand, display a marked hypervariability. This suggests a novel means of identifying ligand-binding sites, and also implies that TPRs generally function as ligand-binding domains. Using perturbation analysis (or statistical coupling analysis), we examined site–site interactions within the TPR motif. Correlated occurrences of amino acid residues at poorly conserved positions explain how TPRs achieve their near-neutral surface charge distributions, and why a TPR designed from straight consensus has an unusually high net charge. Networks of interacting sites revealed that TPRs fall into two unrecognized families with distinct sets of interactions related to the identity of position 7 (Leu or Lys/Arg). Statistical free energy analysis provides a more complete description of “What makes a TPR a TPR?” than consensus alone, and it suggests general approaches to extend and improve the phylogenetic design of proteins.

© 2004 Elsevier Ltd. All rights reserved.

**Keywords:** protein design; Boltzmann law; TPR; knowledge-based potentials; statistical coupling

\*Corresponding author

## Introduction

Consensus design, the engineering of a protein composed of the most common residues at each position as determined from multiple sequence alignment, has been a successful approach for generating proteins with a specific fold.<sup>1</sup> Some have taken this idea still further, proposing that consensus design is a route to thermostable versions of particular folds.<sup>2,3</sup> In many cases, consensus proteins have in fact been considerably more stable than individual proteins in the multiple sequence alignment. Moreover, it seems intuitively

reasonable that the residues that are conserved among proteins that adopt a certain fold are likely to be the most important for fold stabilities.

For example, thermostable versions of fungal phytases have been engineered by consensus methods, wherein parent sequences with  $T_m$  values around 60 °C resulted in a consensus sequence with a  $T_m$  of 78 °C.<sup>4</sup> When six more sequences were added to the alignment (a total of only 19), ten of the predicted mutations were found to lead to higher thermal stability, eight were neutral and ten were destabilizing. Culling for stabilizing mutations resulted in a protein with a  $T_m$  of 90 °C.<sup>5</sup>

The intracellular expression of murine antibodies was shown to be optimized by increasing the stability of the so-called “intrabodies,” and the stabilizing mutations were derived from consensus analysis.<sup>2,6,7</sup> About 60% of predicted mutations were found to be stabilizing. A set of human antibody scaffolds for use in combinatorial selection

Abbreviations used: Ank, ankyrin; TPRs, tetratricopeptide repeats; LRRs, leucine-rich repeats; MSA, multiple sequence alignment; RMS, root-mean-square.

E-mail address of the corresponding author:  
thomas.magliery@yale.edu

of binding molecules was also designed successfully by consensus methods. The resulting proteins appear to express well in the *Escherichia coli* periplasm, and successful expression of antibodies requires adequate stability.<sup>8</sup> While the fact that about half of consensus-predicted stabilizing mutations are not in fact stabilizing is problematic, one must recall that an arbitrary mutation is overwhelmingly likely to be neutral or destabilizing.

Phylogenetic methods have recently been applied to the design of a number of repeat motifs, including ankyrin (Ank) repeats,<sup>3,9–11</sup> tetratricopeptide repeats (TPRs),<sup>12,13</sup> and leucine-rich repeats (LRRs).<sup>14</sup> (This has recently been reviewed.<sup>15</sup>) For example, Mosavi *et al.* designed a model Ank based on straight consensus for well-conserved residues, although poorly conserved residues were chosen based on a constellation of factors including helical propensity and overall charge distribution. Tandem arrays of three or four of these consensus Ank repeats have  $T_m$  values exceeding those of a handful of well-characterized natural Ank repeat domains, and X-ray crystallography showed that they adopt the intended fold.

However, the consensus Ank proteins were only soluble under highly acidic conditions.<sup>9</sup> Replacement of surface Leu residues with Arg residues ameliorated this problem to some extent.<sup>10</sup> Binz *et al.* took a slightly different approach, using straight consensus for the well-conserved positions, other rational considerations at poorly conserved framework positions, and a combinatorial approach to residues that typically contact ligands. Thus, six of the 33 positions were randomized (in a biased fashion), and the scaffold differed in four positions from the Mosavi *et al.* sequence. Six randomly chosen library members were soluble and stable, with  $T_m$  values again in excess of natural Ank domains (60–85 °C versus 50 °C).<sup>11</sup> The crystal structure of one of these library members was subsequently solved, verifying the correct fold.<sup>3</sup>

The underlying mechanism of consensus-based fold stabilization is that the distribution of amino acid residues observed at a given position in a multiple sequence alignment (MSA) is a reflection of what is tolerated by the fold. If the only selective pressure shared by all the proteins in the MSA were fold stability, then the variation of the distribution of amino acid residues from unbiased expectation should be related to thermodynamic stability.<sup>2</sup> This statement requires some scrutiny, however. First, the members of the MSA are sure to be under other pressures than merely stability. Because homologous proteins often carry out related functions, the distribution will be affected by the selective pressure to carry out that function. Proteins must express and avoid aggregation and proteolysis, which may or may not be related directly to thermodynamic stability, and other factors such as codon usage affect the frequency of amino acid residues.

Secondly, the common selective pressure among proteins with the same fold is not maximal fold

stability, but adequate fold stability. There is no selective pressure on any of the individual proteins in the alignment to be any more stable than is necessary for function. How therefore does additional stability arise from the collection? One suggestion is that functional constraints might limit stability, and so averaging out those constraints (by, for example, aligning proteins that bind to different ligands) may result in higher stability.<sup>16</sup> However, these averaged “functional” residues in fact represent sites with little information content in the MSA, since the conservation is poor at these sites, and experimental evidence suggests that functional residues are not always negatively correlated with stability.<sup>17</sup> A more compelling suggestion comes from the observation that, to a reasonable approximation, point mutations tend to have independent effects on stability. Thus, if different members in the MSA have found alternative, independent routes to adequate stability, the superposition of those modes should be more stable, since the stabilizing mutations will be roughly additive.<sup>2,18,19</sup>

If mutations are roughly independent, and if stabilizing mutations can be predicted from sequence alignment, then it is reasonable to think of the sequences in the alignment as being “in equilibrium.”<sup>20</sup> This implies significant evolutionary time since divergence of the sequences, such that sequence-space is well-sampled. As a result, the probability of the occurrence of a residue at a specific position is correlated with the degree of stabilization conferred. This suggests that there will be a quantitative relationship between the positional deviation of the amino acid distribution from a reference state and the thermodynamic stability of the protein, as given by the Boltzmann law.<sup>2</sup> This hypothesis provides a mathematical model for converting positional amino acid distributions into statistical free energies, which is particularly important, since there will always, of course, be a most-common residue at each position in an MSA, but it is not clear how the degree of conservation is related to stability. If 90% of the members of an MSA have a Phe at a given position, then it may be reasonable to think that the Phe is important to the fold, and probably to the stability of the fold. But how important is it (in terms of  $\Delta G$ )? And what if that number is 30%? or 10%?

A computational approach to statistical free energies has been introduced by Lockless & Ranganathan, in which the binomial probabilities of observing each amino acid at the observed frequency in an MSA are calculated given expected or reference frequencies.<sup>21–23</sup> Effectively, the scalar “length” of a resulting 20-dimensional vector (a dimension for each of the 20 amino acid residues) is a measure of the deviation from the reference state. Since each amino acid is evaluated separately, numerically equivalent distributions of different amino acid residues can be discriminated. The ratio of the probabilities between an observed state  $i$  and a reference state is related to the statistical free

energy separating those states by the Boltzmann law:

$$\frac{P_i}{P_{\text{ref}}} = e^{\frac{\Delta G_{\text{stat}}}{kT^*}}$$

where  $kT^*$  is an arbitrary energy unit. The root-mean-square (RMS) average of the natural log of this ratio of probabilities for each amino acid  $x$  at a given position  $i$  (i.e. the scalar length of the vector) is therefore taken to be proportional to  $\Delta G_{\text{stat}}$ :

$$\Delta G_{\text{stat}} = kT^* \sqrt{\sum_x \left( \ln \frac{P_i^x}{P_{\text{ref}}^x} \right)^2}$$

Of course, the additivity of the effects of mutation is clearly not universal, and it may be an especially poor assumption for the design of small motifs like those found in repeat proteins. Consensus-based alignment treats each position independently, neglecting effects of covariance. Covariance may be very important in replicating key interactions. For example, protein cores leading to similar stability can arise from different arrangements of the same amino acid residues, preserving overall core volume. Thus larger amino acid residues at one site might be compensated by smaller residues at another site. Hydrogen bonds might increase the stability of a protein, but the constituent residues may be radically different in independent sequences in the MSA (e.g. Glu-Lys or Lys-Glu). Clearly, these effects will not be reflected in a first-order analysis of frequency alone, since they result in increased variation and therefore lower conservation. However, covariance is significantly more difficult to calculate than frequency, since the number of possible pairwise correlations is large,  $[400(n^2 - n)]/2$  for  $n$  positions in the protein.

In contrast, statistical free energies can also be used to rapidly identify statistically interacting sites. (We say “statistically interact” to emphasize that a direct physical interaction is not necessarily implied from statistical coupling.) If positions  $i$  and  $j$  do not interact, then one would anticipate that the distribution of amino acid residues at  $j$  will be unaffected by the identity of  $i$ . Thus, the distribution of amino acid residues would be expected to be the same at  $j$  regardless of whether we are considering all proteins in the MSA or just those proteins that have a particular amino acid at  $i$ . The deviation from this expectation (the statistical coupling) can be assessed in the same computational fashion as above, where the reference state is now the distribution at  $j$  for all proteins in the MSA. This is a very powerful test of covariance, since it does not presuppose how the distribution at  $j$  is affected by the identity at  $i$ . One need not calculate individually the correlation between each amino acid at  $i$  and each amino acid at  $j$ , since any such correlation will affect the distribution (including correlations that are not significant for an individual amino acid but are significant for a group of amino acid residues with similar properties, such as hydrophobicity or size).

Our group recently reported the successful design of a TPR motif using straight global propensities at each position in the sequence.<sup>12,13</sup> TPRs are 34 amino acid repeat motifs that are found most commonly in groups of three and are generally thought to mediate protein–protein interactions. However, only a small number of TPRs have known ligands, and only a handful of TPR–ligand complex co-crystal structures have been solved. The TPR is a helical repeat motif (it encodes a helix–turn–helix–turn) that forms a spiral staircase-like structure when arrayed sequentially; the superhelical twist of the sequential array results in a concave surface that is seen to be the ligand-binding groove in the few available co-crystal TPR–ligand structures (e.g. HOP-TPR2A/Hsp90, HOP-TPR1/Hsc70, and PEX5/peroxisomal signaling peptide). Consensus and hidden Markov models have facilitated the identification of thousands of individual repeats, making this an attractive target for phylogenetic design approaches. Moreover, since the TPR-repeat arrays are thought to mediate protein–protein interactions, a consensus TPR is an attractive scaffold for the design of proteins that bind to a peptide target of choice, much as zinc finger arrays have been engineered for DNA binding<sup>24</sup> and, recently, Ank repeats have been engineered as generic binding domain.<sup>25,26</sup>

The maximum-global-propensity TPR sequence was repeated  $n$  times in tandem to give CTPR $n$  proteins. CTPR3 is analogous to well-characterized natural 3-TPR domains (e.g. PP5, and TPR1 and TPR2A from HOP). The crystal structure of CTPR3 shows that it adopts the typical TPR fold. In addition, its thermal and chemical stability exceeds natural 3-TPR domains. However, CTPR3 does not replicate all the features of natural TPRs. A striking example is surface charge. The net charge of each engineered repeat was  $-6$ , in sharp contrast to natural TPRs, which tend to have net charges near zero. Charged residues were found in poorly conserved surface-exposed positions, and it therefore seemed likely that the anomalous net charge might be a result of not including covariance in the design. Using a statistical free energy approach, we have identified positional interactions which could not be identified by consensus alone. The analysis reveals why the consensus sequence has such anomalously high charge. The covariance analysis also shows that known TPRs can be categorized into two distinct, previously unrecognized subfamilies. The statistical free energy and coupling analyses together significantly expand on consensus analysis in answering the question: “What makes a TPR a TPR?”

## Results and Discussion

### Choice of reference state

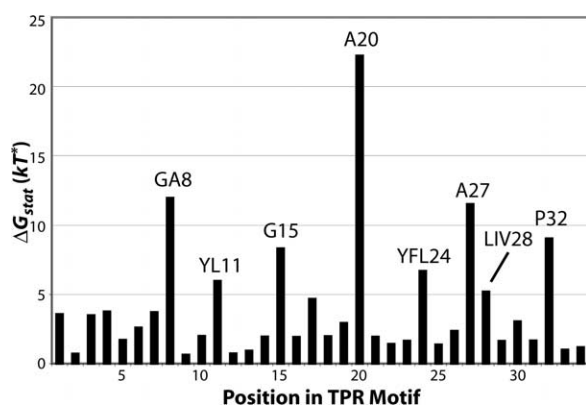
We chose to use yeast codon usage as a reference

state, as opposed to the 36,498 non-redundant eukaryotic proteins from Swiss-Prot used by Lockless & Ranganathan.<sup>21</sup> Yeast is a eukaryote whose complete codon data in all open reading frames is readily available. Within reason, the choice of reference state is not critical to the statistical free energies. This is because the statistical free energy is dominated by the effects of large deviations from the expected frequencies, so any reference state with near-equal usage of all amino acid residues will give approximately the same results. More rigorously, one can calculate the statistical free energy that separates the chosen reference state (yeast codon usage in all proteins) from hypothetical sites with distributions representative of various other states, such as equal usage of all amino acid residues, expected usage from the distribution of the 61 sense codons, or all proteins in Pfam (used by Main *et al.*<sup>12</sup>) Even approximations as crude as equal usage or the genetic code are separated by less than  $kT^*$  ( $0.59 kT^*$  and  $0.36 kT^*$ ), and amino acid usage in Pfam is separated by only  $0.09 kT^*$ . In contrast, each TPR position is separated from the yeast reference state by a mean of  $4.0 kT^*$  (see the next section).

### Statistical free energies for TPR positions

Application of these calculations to the TPR amino acid distributions results in the  $\Delta G_{\text{stat}}$  values in Figure 1. The values vary from  $0.63 kT^*$  to  $22.2 kT^*$ . The significance of the numerical values of  $\Delta G_{\text{stat}}$  are evident from the distributions associated with representative sites possessing a spectrum of  $\Delta G_{\text{stat}}$  values (Figure 2). Below  $kT^*$ , there is virtually no deviation from the reference state, and strong biases are not evident until about  $2.5 kT^*$ . As high as  $5 kT^*$ , nearly every amino acid is observed at some level. Above  $10 kT^*$ , there is a virtual requirement for one or two amino acid residues at the given position.

These values can also be mapped onto the structures of representative TPRs, such as the first



**Figure 1.** Statistical free energies by position in TPRs. The eight most significant residues are labeled with the most common amino acid residues and residue number.

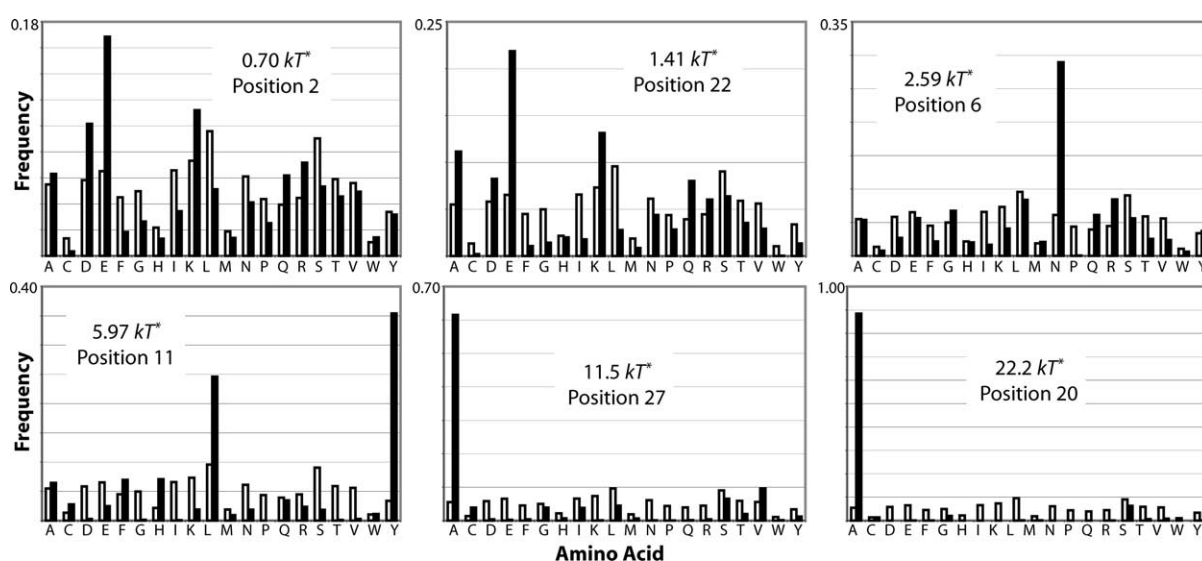
TPR of HOP-TPR1<sup>27</sup> or the central TPR of CTPR3 from Main *et al.*<sup>12</sup> (Figure 3). This allows us to structurally group the positions by deviation from reference distribution. The five most significant positions ( $\Delta G_{\text{stat}} = 8.3\text{--}22 kT^*$ ), in order, are A20, GA8, A27, P32 and G15 (where the letter before the position number is the most common amino acid residue seen at the corresponding position). The first three of the positions lie within the hydrophobic core of the single repeat, and the latter two lie in the second and first turns, respectively. The helices in a TPR motif are separated by a tight (usually Gly) turn, and the helices between motifs are separated by a broader (usually Pro) turn (Figure 3). The next most significant set of residues ( $\Delta G_{\text{stat}} = 2.5\text{--}6.7 kT^*$ ) mostly comprise strips of residues on nearly opposite edges of the TPR, and these are mostly hydrophobic residues that pack against each other between successive TPR repeats (especially YFL24, YL11 and LIV28, Figure 4). The remaining residues with  $\Delta G_{\text{stat}} < 2.5 kT^*$  are mostly along the two solvent-exposed surfaces of the TPR (see below and Figure 5).

### Statistical free energies identify the binding site in TPRs

When the statistical free energies are plotted on the 3-TPR domains of HOP-TPR1, HOP-TPR2A,<sup>27</sup> or either of the PEX5 TPR domains,<sup>28</sup> it is clear that the residues with the least deviation from the reference state ( $\Delta G_{\text{stat}} < 1.25 kT^*$ ) all lie on the concave side of the TPR domains, and make extensive contacts to the ligands and, in the case of PEX5, the opposite TPR domain (Figure 5). The convex surface, which is not known to contain a binding site in any TPR, is dominated by positions with distributions similar to the reference state, but more bias is exhibited than on the binding face ( $\Delta G_{\text{stat}} = 1.25\text{--}2.5 kT^*$ ). That is, the ligand-binding surface is more variable than other solvent-exposed surfaces.

We hypothesize that the reason for this difference is a consequence of the mechanism of evolutionary divergence. TPR domains have likely reached their current ubiquity from duplication events followed by random mutations enforced by functional constraints. Surface-exposed residues are generally the most tolerant of mutation, allowing a higher degree of regression to the reference state over time than with buried positions.<sup>29,30</sup> However, each TPR has been selected for the ability to bind to a different ligand. When all TPR sequences are considered in aggregate, this statistical randomization at ligand-binding positions is more significant than the stochastic process at work on other surface positions, which is limited by evolutionary time. Recently, we have demonstrated that the specificity-determining positions in Ank repeats and Cys<sub>2</sub>His<sub>2</sub> zinc fingers can also be identified in this fashion (our unpublished results).

It is worth noting that in the case of TPRs only a



**Figure 2.** Comparison of statistical free energies with underlying amino acid distributions. Open bars represent the reference state (yeast amino acid usage) and filled bars represent the distribution at various positions. The  $\Delta G_{\text{stat}}$  values are indicated on the graphs. Note that the frequency scales (*y*-axes) vary. Significant biases are evident in the range of about 2.5–5.0  $kT^*$ , and significant preferences for a small subset of amino acid residues are evident at 10–20  $kT^*$ .

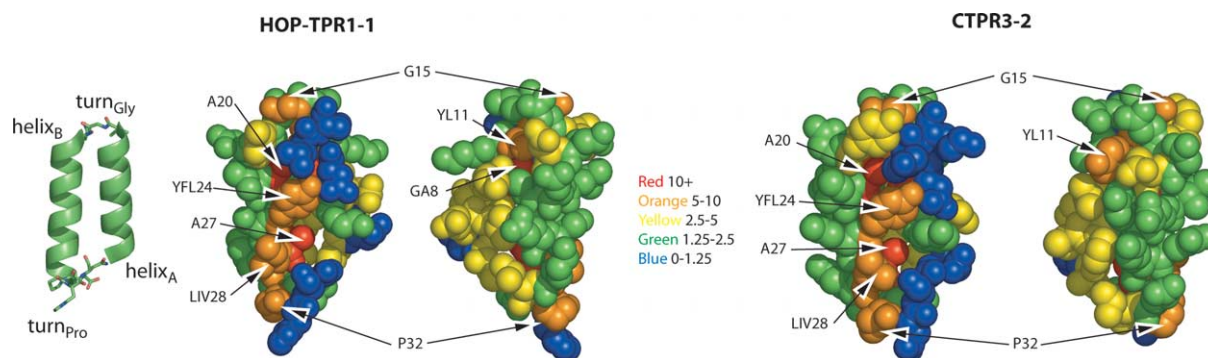
handful of TPR–ligand interactions have been well-characterized, but the additional randomization of the residues on the concave surface of 3-TPR domains implies that these domains are generally involved in binding. While this has been proposed previously based on the handful of TPRs with known ligands, this is, to our knowledge, the first direct evidence of the generality of TPRs as binding domains.

### Statistical free energy, consensus and global propensity

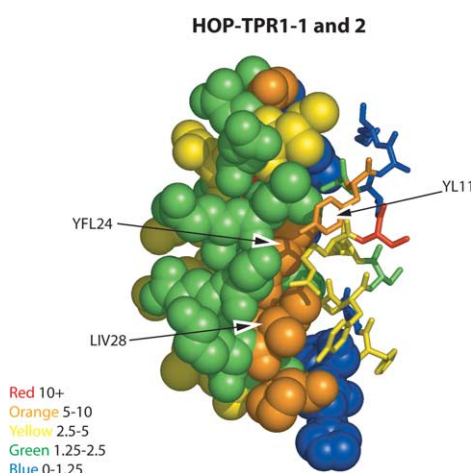
At 12 of the TPR positions, at least 19 of the 20 amino acid residues are commonly observed (Pro is often excluded, since many of the positions are in  $\alpha$ -helices), and fully 18 of the 34 positions show no significant preference for a small subset of the

amino acid residues. The traditional consensus-based definition of a TPR<sup>31</sup> (see Figure 6) does not include G15 or LIV28, which are highly biased positions, and includes positions LYW4, and LR7, which fall into a group of eight residues of only moderate bias. It is worth considering the reason for the differences between sequences derived from maximum frequency (i.e. straight conservation), maximum global propensity, and statistical free energy.

Figure 6 shows the consensus sequence of TPRs, the sites at which consensus differs from global propensity and the degenerate definition of a TPR suggested by the combination of statistical free energies and the underlying amino acid distributions. Maximal global propensity differs from maximal frequency in eight of 34 positions. However, seven of those positions show no marked



**Figure 3.** Statistical free energies mapped onto example TPRs. The molecule on the left is the first TPR from the HOP-TPR1 domain;<sup>27</sup> the molecule on the right is the center TPR from the CTPR3 consensus TPR structure described by Main *et al.*<sup>12</sup> At far left, a ribbon diagram is shown for orientation; the left space-filling images in each pair are in the same orientation. The right images are rotated about 180°. For clarity, the degenerate description of each position is listed rather than the identity of the actual amino acid in each particular structure. The meaning of the coloration ( $\Delta G_{\text{stat}}$  in  $kT^*$ ) is indicated in the Figure. Rendered from PDB entries 1ELW and 1NA0 using PyMOL (<http://www.pymol.org>).



**Figure 4.** Interface between TPR motifs. The orange strip of residues on helix<sub>B</sub> (see Figure 3) makes extensive contacts with the yellow strip on the opposite side of helix<sub>A</sub>. Here, helix<sub>A</sub> of HOP-TPR1-2 is rendered in sticks to visualize the contacts with the previous TPR1-1. This group is critical for packing interactions between repeats, but the added variability relative to the internal hydrophobic residues is likely due to (1) the variety of superhelical packing seen in TPRs and (2) the fact that some “external” TPRs are probably solvent-exposed along these surfaces. Rendered from PDB entry 1ELW using PyMOL.

preferences for any particular amino acid. In position 4, Trp has the highest global propensity but is the third most frequent residue behind Leu and Tyr. This is because Trp is tenfold less common than Leu and threefold less common than Tyr in proteins overall. The CTPR design used global propensity to identify residues at each position to account for amino acid usage in proteins overall. Global propensity reflects how much a particular amino acid prefers being in a particular position, but not how much a particular position prefers having one amino acid *versus* another.<sup>32</sup> If a rare amino acid were actually required in a particular site, evolutionary pressure would have to overcome the rarity of the amino acid. For example, in position 17, Tyr is more common than Leu or Phe, but it is a rarer amino acid in proteins overall.

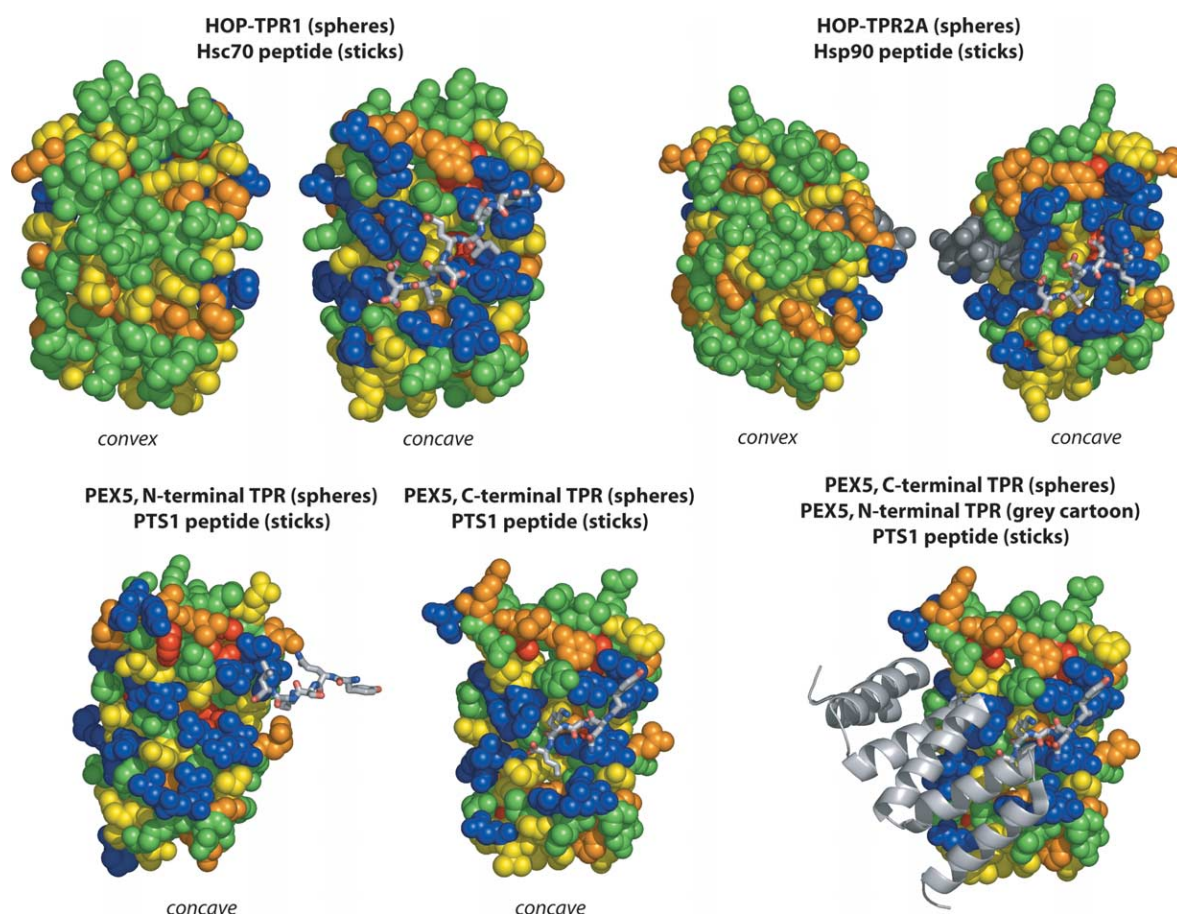
We can quantitatively assess the relationship between maximal frequency and statistical free energy by creating a hypothetical site where the most frequent amino acid is Ala at varying levels from 10% to 99.9%, and all other amino acid residues are in proportion to their frequency in yeast proteins. The relationship between maximum frequency and  $\Delta G_{\text{stat}}$  can be approximated by a quadratic function (Figure 7, left). The maximal frequency and  $\Delta G_{\text{stat}}$  for each site in TPRs is plotted on the same graph. Ala was chosen for the model site because it is at median usage in yeast. As a result, any point above the trendline represents a site where maximum frequency underestimates the

deviation from the reference state, and any point below the trendline represents a site where the maximal frequency overestimates the significance.

For example, the most common residue in position 28 is Leu in 48% of sequences, and one would expect a  $\Delta G_{\text{stat}}$  of about  $7 kT^*$  for such a site based on the Ala model. However, the actual  $\Delta G_{\text{stat}}$  is  $5.2 kT^*$ , and therefore the site appears more significant than it is based on maximal frequency (i.e. consensus) alone. This is because Leu is the most common amino acid in the proteins overall, and having a high frequency of Leu is therefore less significant. Conversely, position 17 is most commonly Tyr (in 31% of sequences), but it has a  $\Delta G_{\text{stat}}$  of  $4.7 kT^*$ , which is higher than the expected value of about  $3 kT^*$ . Consensus makes this site appear less significant than it is, here because Tyr is a relatively rare amino acid in proteins overall, and because 15% of sequences at this position are Phe and 13% are Leu. Thus, 59% of all TPRs have only three amino acid residues at this position, which is a significant bias, but this is obscured by looking at only the single most common amino acid.

Since one of the problems with consensus frequency is that the rarity of the amino acid in proteins overall is not considered, one might hypothesize that maximal global propensity might be more closely related to statistical significance (i.e.  $\Delta G_{\text{stat}}$ ). However, the correlation between maximal global propensity (as calculated by Main *et al.*<sup>12</sup>) and  $\Delta G_{\text{stat}}$  is in fact a great deal worse (correlation coefficients are 0.95 for maximal frequency and 0.65 for maximal global propensity). Using the same hypothetical Ala sites as above, one can approximate the relationship between maximal global propensity and  $\Delta G_{\text{stat}}$  with a quadratic equation (Figure 7, center). Strikingly, nearly all the points from the TPRs are below this line, indicating that global propensity almost always overestimates the significance of a particular position relative to the Ala model, and the spread of the points indicates that the degree of this overestimation is highly variable. The points that conform most closely to the trendline correspond to positions dominated by one or two common amino acid residues; those that are farthest away are generally sites with a mix of rare and common residues.

It should be noted that, although Ala is present at median frequency in yeast proteins, it is quite common in all proteins in Pfam, which explains why the importance of virtually all other sites is overestimated. This is therefore somewhat of an unfair comparison, since the global propensities were calculated using Pfam amino acid usage as a reference state (following Main *et al.*<sup>12</sup>), while the  $\Delta G_{\text{stat}}$  values were based on the yeast reference state. If the global propensities are recalculated using the yeast reference state (Figure 7, right), then the correlation between maximal global propensity and  $\Delta G_{\text{stat}}$  improves, but it is still poorer than consensus (the correlation coefficient is 0.87). Essentially, this is because global propensity grossly overestimates the importance of rare residues, since



**Figure 5.** TPR domains and ligands. At top, domains TPR1 and TPR2A from HOP,<sup>27</sup> with peptide ligands shown in sticks (on the concave surfaces). The ligands are contacted extensively by the least conserved (blue) positions, which are all on the concave surface. In PEX5 (bottom),<sup>28</sup> the ligand peptide binds on the concave surfaces of the two TPR domains, and the concave surfaces also act as a protein–protein interface between the two domains. Coloring of spheres is the same as Figures 3 and 4. Rendered from PDB entries 1ELW, 1ELR and 1FCH using PyMOL.

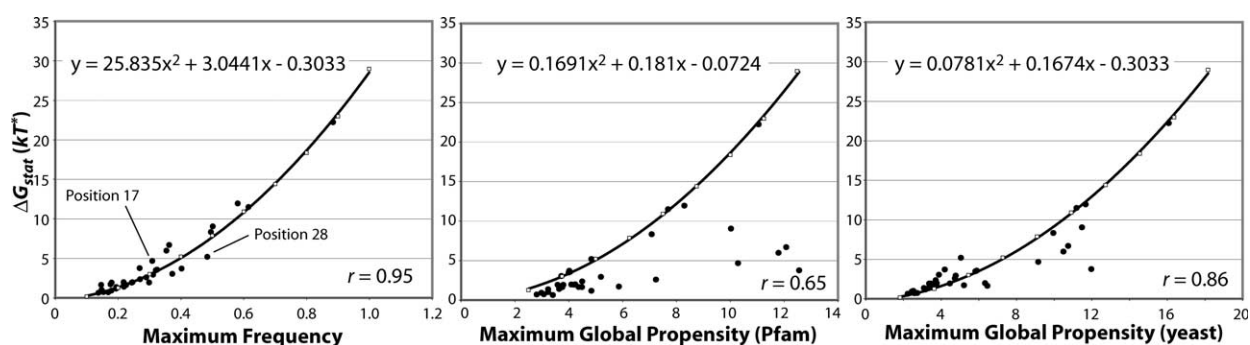
AEALYNLGLAYLK	LG	YEEALEAYEKALEL	DPDN	Consensus
W	NC Y M	I Y	N	Max GP (Pfam)
W	NC Y M R	I C QR		Max GP (Sc)
W	LG Y	A F A	P	Signature
A-SKYNLQ-AY--	LGD YEAL--Y-KALEL	DP--		Degenerate
P LPLNRa LL	Q K EDK I F R IKI Na			
V VVNa V	M R IKQ E I E vQ A			
FG	I n Q v q A E			
	c e a r			
	q			

**Figure 6.** Phylogenetic designs of TPRs. Sequence of TPRs, divided into groups of helix<sub>A</sub>, turn<sub>Gly</sub>, helix<sub>B</sub> and turn<sub>Pro</sub>. Consensus: residues with maximal frequency at each position. MGP (Pfam): residues with maximal global propensity, using amino acid usage in Pfam as a reference state. MGP(Sc): maximal global propensity with yeast amino acid usage as a reference state. Signature: traditional definition of a TPR.<sup>31</sup> Degenerate: combinatorial definition of a TPR implied from statistical free energies and amino acid distributions. Residues in uppercase are those that make up at least 50% of the residues at the indicated positions, in order of frequency from top to bottom. When informative, the next most common residues are indicated in lowercase. A dash (–) indicates that there is no significant preference at that position. Colors are the same as Figures 3 and 4.

it was not designed to compare the preference of different amino acid residues for a common position. We can conclude that: (1) global propensity is very sensitive to reference state, unlike statistical free energies; and (2) global propensity accounts for usage poorly for this purpose and is especially poor at reflecting the importance of sites with a mix of common and uncommon amino acid residues.

### Correlated interactions from perturbation analysis

To investigate the statistical interaction of two positions  $i$  and  $j$ , one can calculate the free energy that separates the distribution of amino acid residues at  $i$  in all TPR sequences and the distribution of amino acid residues at  $i$  in a subset of sequences at  $j$  (for example, wherein all the selected sequences have a particular amino acid at  $j$ ). The principle is that if there is no correlation between residue identity at  $i$  and  $j$ , then the distribution of amino acid residues at  $i$  should be the same regardless of what is at  $j$ , and  $\Delta\Delta G_{\text{stat}}^{i|j}$  will be zero. To identify which correlations are



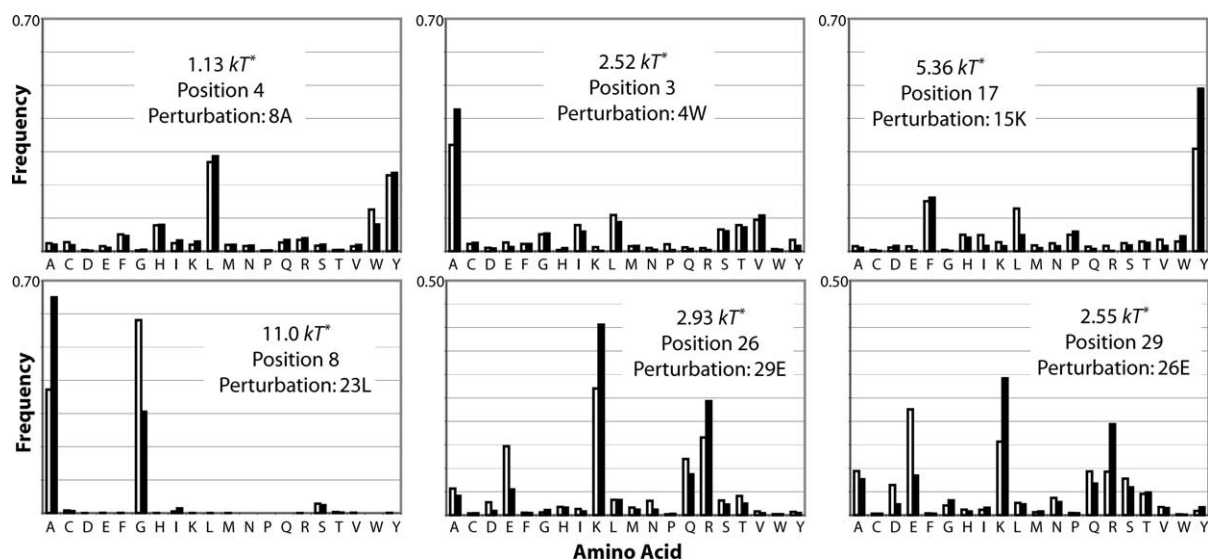
**Figure 7.** Relationship between  $\Delta G_{\text{stat}}$  and maximal frequency or global propensity. Filled circles represent the 34 positions in TPRs. Open squares represent hypothetical sites with varying levels of Ala conservation (see the text). The quadratic trendlines are fitted relationships between maximal frequency or global propensity and  $\Delta G_{\text{stat}}$  for the hypothetical Ala sites; the trendline equations are shown on the graphs. The correlation coefficients  $r$  for the TPR datasets are also indicated.

significant, two things must be true: (1) the subset  $\delta j$  must be statistically significant; and (2)  $\Delta \Delta G_{\text{stat}}^{i|\delta j}$  must be above a threshold value. We arbitrarily decided that any subset  $\delta j$  was significant if at least 10% of TPRs were in the subset (here, 689 sequences). There are 91 such subsets (out of a possible of 34 positions  $\times$  20 amino acid residues = 680). For this analysis, we chose a  $\Delta \Delta G_{\text{stat}}^{i|\delta j}$  threshold of  $2.5 kT^*$ . Figure 8 illustrates four examples of  $\Delta \Delta G_{\text{stat}}^{i|\delta j}$  between 1 and  $10 kT^*$ , from which it is evident that  $2.5 kT^*$  represents a fairly significant level of redistribution. There are 126 such interactions (out of a possible 91 subsets  $\times$  33 possible coupled positions = 3003). See Supplementary Materials for an analysis of the subset size and threshold value used here.

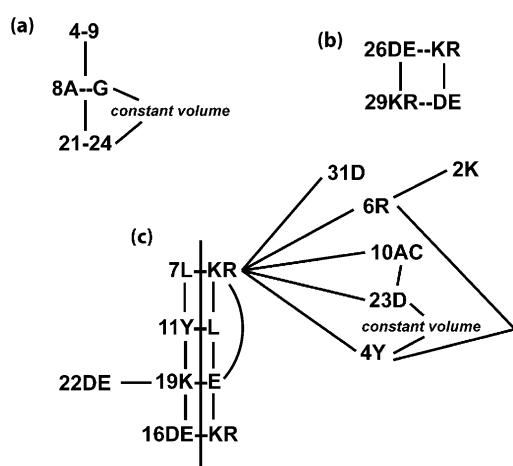
The meaning of these statistical interactions can

be inferred, in part, from the difference between the amino acid distribution in all TPRs and the selected subset (Figure 8). For example, when position 29 is restricted to Glu, the frequency of Lys and Arg increases and the frequency of Asp and Glu decreases at position 26. When position 26 is restricted to Glu, the reciprocal trend is seen at position 29. Therefore, it can be inferred that positions 26 and 29 interact in such a way that having opposite charge is favorable, and a reasonable hypothesis would be that these positions can interact directly through a hydrogen bond or charge–charge interaction. In fact, in the crystal structure of HOP-TPR2A-1, 26K and 29E appear to interact; and in HOP-TPR1-3, 26E and 29K appear to interact (see Figure 11).

Many of the 126 significant interactions can be



**Figure 8.** Examples of  $\Delta \Delta G_{\text{stat}}$  values for various perturbations. Open bars represent the distribution at the indicated position in all TPRs, and filled bars represent the amino acid distribution at the indicated position in the subset (the “perturbation”). Note that the frequency scale ( $y$ -axis) is different for the 26/29E and 29/26E perturbations. In this study, only perturbations resulting in  $\Delta \Delta G_{\text{stat}} > 2.5 kT^*$  were considered.



**Figure 9.** Networks of statistically interacting residues implied from perturbation analysis. Statistically significant interactions can be arranged into “networks” by examining the differences in amino acid distribution for various TPR subsets. Lines between different positions represent direct correlations. (a) The identity of residue 8, almost always Gly or Ala, is affected by residues 4–9 and 21–24. Residue 24 tends to get larger or smaller inversely with residue 8. (b) Positions 26 and 29 tend to have opposite charges. (c) TPRs with Leu7 tend to have Tyr11, DE16, Lys19 and DE22. TPRs with KR7 tend to have Leu11, KR16 and Glu 19, in addition to Lys2, Tyr 4, Arg6, AC10, Asp23 and Asp31.

summarized in networks shown in Figure 9. Position 8 is nearly always Gly or Ala, but the amino acid residues at positions 4–9 and 21–24 affect which amino acid of the two amino acid residues is in position 8. These residues, with the nearly invariant residue A20, essentially compose the “layer” below the Gly turn surrounding residue 8. As noted above, positions 26 and 29, which lie on the convex surface of HOP-TPR1, nearly always have opposite charges. The most extensive network is one in which the canonical Leu7 tends to be seen with Tyr11 and Lys19, which tends to be seen with DE16 and DE22. However, when position 7 is occupied by Lys or Arg, it is commonly seen with Leu11, Glu19 and KR16, in addition to Arg6, Lys2, Asp31, AC10, Asp23 and Tyr4.

### Perturbation analysis reveals two families of TPRs

These networks suggest that there are two previously-unrecognized subfamilies of TPRs, those with a canonical Leu7 and those with a basic residue at position 7. One can extract just those sequences with Leu or Ile, or those with Lys or Arg, at position 7, and compare the resulting consensus sequences to the predictions afforded by the perturbation network analysis (Figure 10). Evidently, these two families do exist, although about 50% of TPRs have LI7 and 30% have KR7 (which is why the canonical definition of a TPR

AEALYNLGLAYLK	LGD	YEEALEAYEKALEL	DPDN	Consensus					
AEALYNLGLLYLK	LGD	YEEALEYEEKALEL	DPDN	LI7 Cons.					
	L	Y	D	K	E	L7 Network			
AKAYYNRGLALLK	LGD	YEEALEDYEKALEL	DPDN	KR7 Cons.					
	K	Y	RR	AL	K	E	D	D	KR7 Network
A-ALYSLG-AI--	LGD	YEEAL--Y-KALEL	DP--	Degenerate					
P LPLRPa LL	Q	K	DR	I	P	R	K	N	
V VwNL	V	M	R	K	E	I	E	Q	A
	FG	I	n	Q	v	q	A	E	
		c	e	a		r			
						q			

**Figure 10.** TPR subfamilies: consensus TPRs with LI7 or KR7. Comparison of the most frequent residues seen in all TPRs, TPRs with Leu or Ile at position 7, TPRs with Lys or Arg at position 7, and the degenerate definition of a TPR described in Figure 6. The networks of residues implied from perturbation analysis for Leu7 and KR7 TPRs are also shown. Note that Lys16 and Lys19 are the second most common residues (to Asp and Glu, respectively) in all groups, and Arg6 is the second most common residue (to Asn) for KR7 TPRs.

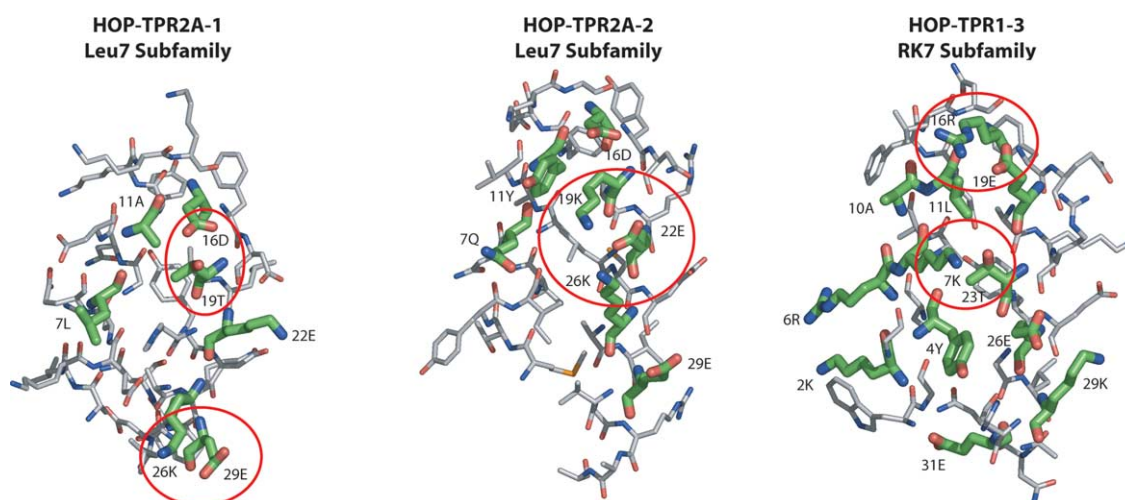
includes Leu7 and why the D’Andrea & Regan CTPR is a Leu7 TPR). Interestingly, crystal structures of members of the two families exist, since both families are present among the six TPRs in two TPR domains of HOP (Figure 11).

HOP-TPR1-3 has nearly the archetypal KR7 sequence. Residue Lys7 hydrogen bonds to Thr23, causing these residues to point toward the center of the TPR, and Arg16, Glu19 and the backbone carbonyl group of Leu11 clearly interact. Residues Lys2 and Arg6 are oriented opposite Lys7 into the solvent-exposed concave binding surface of the TPR domain. HOP-TPR2A-1 and 2 both have Leu7-like sequences, and in both cases the residue at position 7 points away from the center of the TPR. In the first TPR, Asp16 and Thr19 interact, and in the second TPR, Lys19, Glu22 and Lys26 interact (resulting in a loss of the common 26–29 interaction seen in the other TPRs).

Both TPR1 and TPR2A bind to C-terminal EEVD-CO<sub>2</sub><sup>-</sup> peptides,<sup>27,33</sup> so the position 7-subfamily is not clearly related to the function of the TPR domain. It seems likely that these are two evolutionarily divergent solutions to achieving TPR structure, and it would be of interest both to design archetypal members of each family and to analyze their occurrence in nature. It is worth pointing out that these networks could not have been identified from consensus alone, since many of the interactions involve charged residues at poorly conserved positions.

### Charge perturbation analysis and net charge in consensus

If one approximates the net charge of the 6887



**Figure 11.** Leu7 and KR7 networks in HOP TPRs. The first two TPRs from HOP-TPR2A conform closely to the Leu7 predicted network, and the third TPR from HOP-TPR1 possesses the predicted KR7 network. In TPR2A-1, there are 16D-19T and 26K-29E interactions. In contrast, TPR2A-2 shows an interaction network with 19K-22E-26K. In TPR1-3, the 7K-23T interaction results in position 7 pointing toward the center of the motif, as opposed to the two motifs from TPR2A. Residues 16R and 19E interact with each other and the backbone carbonyl group of 11L. Rendered from PDB entries 1ELR and 1ELW with PyMOL.

TPRs in this analysis, the average net charge is  $+0.04$ , approximately normally distributed with a standard deviation of 2.5 (Figure 12). Thus, net charges of  $-6$  or  $-7$  for the maximum global propensity or maximum frequency sequences, respectively, are significant outliers from the data set. It is not initially obvious why the consensus of sequences with near-zero net charge should have such a high charge load. The answer must, however, lie beyond first-order analysis.

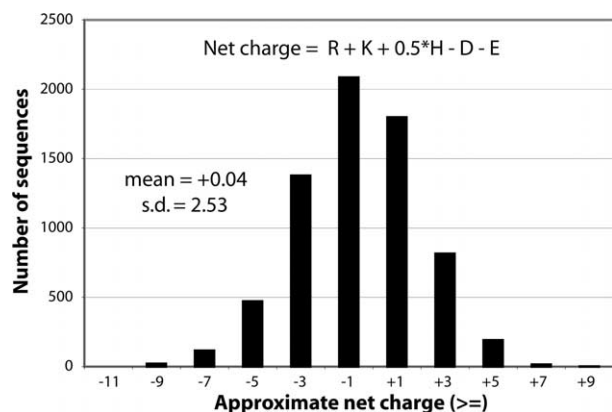
We hypothesized: (1) that most of the charged residues were in surface-exposed positions with poor conservation, where the consensus residue did not reflect very much information; and (2) that natural TPRs would tend to offset the effects of a charged residue in any particular position by having opposite charge distributed over the rest of

the motif. We tested this by selecting those positions which are commonly occupied by charged residues ( $>30\%$  of the time) and extracting the subsets in which these positions are either acidic (Glu or Asp) or basic (Lys or Arg). We then calculated the mean net charge for each of the other 33 positions in the selected TPRs, as well as the total for all 33 positions. The difference between these net charges and the net charges for all TPRs shows how the TPR sequences respond to having a particular charged position.

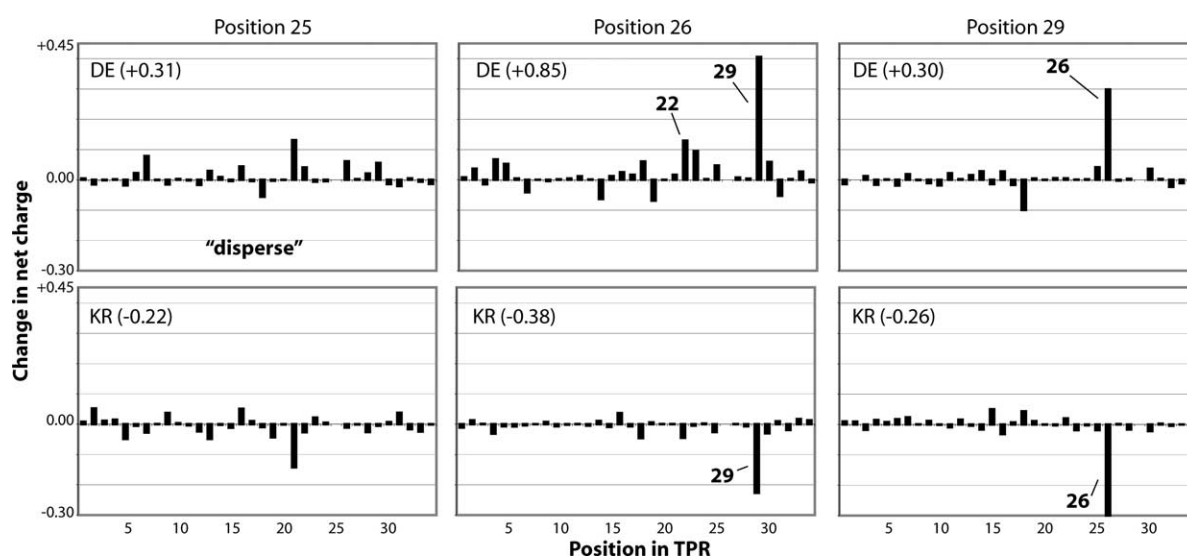
As an example (Figure 13), the DE and KR perturbations of positions 26 and 29 are shown. When position 26 is restricted to DE, position 29 has significantly more mean positive charge than in all TPRs, and position 22 has slightly more positive charge. When 26 is restricted to KR, position 29 is significantly more negative than in all TPRs. On the other hand, DE29 causes 26 to be more positive, and KR29 caused 26 to be more negative. As seen above (Figure 11), some of the HOP TPRs show that 26–29 charge–charge interactions are possible, as well as 22–26 interactions.

Table 1 shows that in the vast majority of the cases where the mean net charge changes by more than 0.1 from all TPRs, it does so to balance the charge perturbation. In some cases, such as with positions 26 and 29, the charge balance is commonly achieved by another specific residue possessing opposite charge. In other cases, such as with position 25, the effect is more disperse, and while the other 33 residues balance the charge, on average, the effect is not dominated by a single (and therefore presumably direct) interaction. It should be noted that nearly all of the interactions where two residues reciprocally balance each other's charge were also identified by the Lockless & Ranganathan method.

The surface charge effect is a good example of



**Figure 12.** Histogram of approximate net charges of TPRs. Charges were calculated with the indicated formula.



**Figure 13.** Examples of charge perturbations. Graphs show the net mean difference in charge between all TPRs and the indicated subset of TPRs (e.g. Position 25, DE, means all TPRs with Asp or Glu at positions 25). The sum of the positional net differences is shown on the graphs. For example, when position 26 is restricted to Asp or Glu (top middle panel), the net charge of the rest of the proteins increases by 0.85 relative to all TPRs (thus balancing the charge). This compensation is dominated by position 29 becoming more positive, but position 22 also becomes more positive. Here, only Asp, Glu, Arg and Lys were used to approximate the net charges.

**Table 1.** Summed positional mean net charge differences between all TPRs and those in the indicated subset

Perturbation	Result	Notes
13DE	-0.23	Disperse effect, two more negative
15DE	-0.26	Disperse effect
16DE	+0.39	Dominated by 19 becoming more positive
19DE	+0.59	Dominated by 16 becoming more positive
22DE	+0.25	Dominated by 19 becoming more positive
25DE	+0.31	Disperse effect, 21 more positive
26DE	+0.85	Dominated by 29 becoming more positive; 22 more positive
29DE	+0.30	Dominated by 26 becoming more positive
31DE	+0.15	Disperse effect, seven more positive
2KR	+0.22	Disperse effect, six more positive
7KR	-0.55	Dominated by 23 becoming more negative; 31 more positive
16KR	-0.37	Dominated by 19 becoming more negative
18KR	+0.23	Disperse; 13 and 29 more positive; 19 more negative
19KR	-0.75	Dominated by 16 and 22 becoming more negative
22KR	-0.52	19, 26 and 31 become more negative
25KR	-0.22	Disperse; 21 more negative
26KR	-0.38	Dominated by 29 becoming more negative
29KR	-0.26	Dominated by 26 becoming more negative
31KR	-0.33	Disperse; two and 33 more negative, 16 more positive

The 33 positions other than the site of perturbation were summed. For example, when position 26 is Asp or Glu, on average, the rest of the TPR becomes more positive to compensate for this, and the effect is dominated by position 29 becoming more positive.

how covariation can be incorporated into a design scheme based on straight consensus. This is especially true at positions with poor conservation, since there is little meaning to the most frequent residue (i.e. straight consensus) at these positions. This is in interesting contrast to the contention of Mosavi *et al.*, who commented that the most frequently observed pairs in Ank repeats were already represented in the consensus.<sup>9</sup> Of course, it is true that the most highly conserved positions will also be frequently observed pairs, but why were no correlated pairs identified in poorly conserved positions?

The answer probably lies mostly in the covariation metric employed in the Mosavi *et al.* study, which is the RMS over all 20 amino acid residues for:

$$f(a_i, b_j) - f(a_i) \times f(b_j)$$

where  $f(a_i)$  and  $f(b_j)$  are the frequencies of amino acid residues  $a$  and  $b$  at positions  $i$  and  $j$ , respectively, in all sequences, and  $f(a_i, b_j)$  is the frequency of observing both amino acid  $a$  in position  $i$  and amino acid  $b$  in position  $j$  in the same sequence. (Although the authors say that they calculated the probability of each amino acid appearing at each position, we believe from context that they mean frequency, and that their design is based on pure consensus rather than some metric adjusted for a reference state.) This metric is highly biased toward conserved positions. If  $f(a_i)$  and  $f(b_j)$  are 0.4 and 0.6, and  $f(a_i, b_j)$  is 0.3, then the covariation is 0.06. If  $f(a_i)$  and  $f(b_j)$  are 0.02 and 0.04, and  $f(a_i, b_j)$  is 0.02, then the covariation is 0.012. However, in the latter case, every time position  $i$  is  $a$ , position  $j$  is  $b$ !

We submit that a more informative metric would have been:

$$\frac{f(a_i, b_j)}{f(a_i) \times f(b_j)}$$

In the example above, the first covariation score would have been 1.25 (meaning that the coincident pair is 25% more common than expected from the underlying frequencies alone), and the second would be 2.5 (meaning that it is 2.5-fold more common than expected by chance).

Nevertheless, this method of examining covariation is much more calculation-intensive than the statistical perturbation approach, it does not address the statistical significance of the covariation, and it only detects one-to-one redistributions. Mosavi *et al.* solved the problem of choosing residues at poorly conserved sites the same way Binz *et al.* did, applying rational design principles, which was obviously successful for both groups. But it would be of interest to interrogate how natural repeat proteins tend to satisfy charge distributions, for example, and to build proteins accordingly.

### What makes a TPR a TPR?

Statistical free energy analysis of TPR sequences reveals that the most critical factors that define a TPR are helical punctuations in the turns and the residues that comprise the buried hydrophobic residues in a single motif. These turn out to be related issues, since the turn between the helices within a single motif (the Gly turn) is tighter than the turn between successive motifs (the Pro turn), and the small GA8 and A20 core positions help permit the tight intramotif turn. The other critical core residues are YL11, which packs against GA8 and A20, and YFL24 and A27, which comprise the core further away from the Gly turn. The next most conserved sets of residues define the intermotif surfaces, and the greater variability of these largely hydrophobic regions reflects the variability in the superhelical characteristics of TPRs.<sup>12</sup> The two most critical residues between repeats are YFL24, which also sits in the core of a single motif, and LIV28, which makes the most significant interdigitation into the hydrophobic core of the next motif. It would be interesting to investigate the effects of mutation of these two residues on the cooperativity of TPR domain folding and unfolding, since repeat protein domains have been found to fold in a two-state cooperative fashion despite their apparent modularity (E. R. G. Main *et al.*, unpublished results).<sup>34–36</sup> The residues on both the concave and convex surfaces of the arrayed TPRs are the most variable, and the extremely high variability of the concave surface suggests that it is commonly involved in binding interactions to diverse ligands.

The statistical coupling analysis allows us to identify higher-order interactions, even in poorly conserved positions. Position 8 is a key core residue

that is nearly always Gly or Ala, but the choice of these two residues is affected by the nature of the more variable residues around it. The most significant correlated effect that is totally obscured by consensus analysis is the nature of charge compensation in natural TPRs. For example, 16–19 and 26–29 interactions often serve to balance overall charge, but other TPRs exhibit a 19–22–26 interaction instead. These residues all lie along a strip of the B helix on the convex, solvent-exposed surface of the TPR domain. The solvent-exposed residues on the A helix, including both those that point out from the convex surface (positions 6 and 10) and into the concave surface (position 2), are related to the nature of the residue at position 7 and, in general, TPRs fall into two subfamilies depending upon the residue at this position (Leu or Lys/Arg). The two position 7 subfamilies also differ in the specific nature of the interactions within the core of the motif (positions 4 and 11) and otherwise between the helices (positions 7 and 23).

The combination of the positional statistical free energies, the underlying amino acid distributions, and the higher-order interactions revealed by perturbation methods allows us to make significant additions to the design strategy for TPRs.

- (1) It is possible to design both Leu7 and KR7 TPRs by including the networks implied from the perturbation analysis. There are likely other subfamilies which might be identified with perturbation analysis methods, such as “external” and “internal” motifs (Tommi Kajander, T.J.M. & L.R., unpublished results). The insolubility of consensus-designed Ank arrays is probably at least partly a result of the alignment of both internal and external motif subfamilies together (resulting in hydrophobic residues in solvent-exposed positions).<sup>9,10</sup> It was noted previously that alignment of antibody sequences from different subfamilies results in the consensus design of the most common subfamily.<sup>15,37</sup>
- (2) The amino acid residues selected at positions with poor conservation can be chosen to balance charge the same way that natural TPRs do. Especially since the highly charged CTPR3 described by Main *et al.* is very stable, the effects of replicating natural surface charge properties will be interesting. Our group has recently shown that neutralizing the convex-face charge of the consensus TPR increases its stability.<sup>38</sup> Also, the extremely high “back face” negative charge of CTPR3 prevents binding to highly charged ligands, like the Hsc70 or Hsp90-derived EEVD-CO<sub>2</sub> peptides, even when the proximal binding residues are designed back into the consensus scaffold.
- (3) The statistical free energies and the amino acid distributions can be combined to define a degenerate TPR sequence, and the validity of that combinatorial design model can be tested rigorously by making libraries in which the

positions vary according to the degeneracy model. This is similar to the approach of the Hecht group in the combinatorial design of four-helix bundles, although the degenerate composition of each position was selected from first principles.<sup>39,40</sup> The Ranganathan group is designing WW domains in this fashion.<sup>41</sup>

## Conclusion

It is clear that phylogenetic design methods are useful for engineering molecules with a specific fold when a reasonable number of examples of the fold are available. However, the relationship between level of conservation and stability is not clear, and the effects of covariance can be significant but are ignored in pure consensus design. In particular, consensus design is compromised by four factors.

- (1) Maximal frequency and maximal global propensity fail both to account for sites where a subset of amino acid residues is common or to reasonably account for the frequency of amino acid residues in proteins overall.
- (2) There is little information encoded at poorly conserved positions, and the most common residue may not always be the best design choice in such cases. In TPRs, as many as half of the positions in the motif are candidates for reconsideration due to poor conservation.
- (3) In part, this is because first-order analysis of sequence does not contain information about correlated occurrences of amino acid residues at poorly conserved positions. These types of interactions lead to fine details of natural proteins, such as surface charge distribution, and these may have an effect on protein stability and other properties (such as solubility or usefulness as a scaffold for further design).
- (4) Even at better conserved positions, consensus methods obscure rarer subfamilies, which may reveal other modes of achieving the same structure. The aggregate consideration of multiple subfamilies may negatively affect the consensus design.

Fortunately, statistical free energy analysis is straightforward, and the calculations are sufficiently simple that they can be carried out in an Excel spreadsheet. The Ranganathan group has compiled a software package that is suitable for this purpose (Rama Ranganathan, University of Texas Southwestern Medical Center, Dallas, TX, personal communication). These methods allow the consensus designer to quantify the significance of the consensus amino acid residues selected at each position and to rapidly uncover correlated changes in amino acid distributions that could not be derived from first-order consensus analysis. This can guide the systematic design of a particular fold,

provide a roadmap for the combinatorial design of folds, and delineate an experimental model for how sequence gives rise to particular structures.

## Materials and Methods

All data manipulations and calculations were carried out using Microsoft Excel XP and Visual Basic scripts, as necessary. Calculations were carried out on a 2.2 GHz Mobile P4 Dell Latitude C640 laptop.

### Data sets

All TPR sequences from Pfam<sup>42†</sup> were downloaded on August 1, 2003 (7386 sequences). There are now (March 2, 2004) 9756 TPRs in the Pfam database. The database was narrowed to only those sequences with the canonical 34 amino acid residues (6887 sequences). No attempt was made to remove sequences that are repeated.

Global propensities for each amino acid at each position were calculated as:

$$GP = \frac{n_i^x / N_{\text{TPRs}}}{N_{\text{ref}}^x / N_{\text{ref}}}$$

where  $n_i^x$  is the number of TPR sequences that contain amino acid  $x$  at position  $i$ ,  $N_{\text{TPRs}}$  is the total number of TPR sequences,  $N_{\text{ref}}^x$  is the total number of each amino acid  $x$  in all positions in the reference set, and  $N_{\text{ref}}$  is the total number of positions in the reference set.<sup>12,32</sup> The reference set for the denominator of this fraction is all proteins in Pfam, and the values were taken directly from Main *et al.*:<sup>12</sup> A (0.08), C (0.02), D (0.05), E (0.06), F (0.04), G (0.07), H (0.02), I (0.06), K (0.06), L (0.1), M (0.02), N (0.04), P (0.05), Q (0.04), R (0.05), S (0.07), T (0.06), V (0.07), W (0.01), Y (0.03).

The reference set used for the statistical free energy calculations was amino acid usage in all proteins in yeast, calculated from codon usage in all ORFs in yeast (*Saccharomyces cerevisiae*, Kyoto Encyclopedia of Genes and Genomes<sup>‡</sup>). The amino acid usage frequencies for each amino acid  $x$  are ( $f_x$ ): A (0.055), C (0.013), D (0.058), E (0.065), F (0.045), G (0.050), H (0.022), I (0.066), K (0.073), L (0.096), M (0.019), N (0.061), P (0.044), Q (0.039), R (0.045), S (0.090), T (0.059), V (0.056), W (0.010), Y (0.034).

### Statistical free energies

The binomial probability  $P$  of each observed amino acid  $x$  frequency at each position  $i$ , given the reference frequency  $f_x$ , was calculated as:

$$P_i^x = \frac{N!}{n_i^x!(N - n_i^x)!} f_x^{n_i^x} (1 - f_x)^{N - n_i^x}$$

where  $N$  is the total number of sequences in the analysis and  $n_i^x$  is the number of sequences with amino acid  $x$  at position  $i$ . Due to the fact that the value of  $P$  is affected by the sample size  $N$ , all usage data were scaled to  $N=1000$ . Due to the difficulty of calculating  $x!$  for large  $x$ , the Stirling approximation was used:

$$\ln x! \approx \text{Stirling}(x) = \left(x + \frac{1}{2}\right) \ln x - x + \frac{1}{2} \ln 2\pi$$

† <http://pfam.wustl.edu>

‡ <http://www.genome.ad.jp/kegg>

such that  $P_i^x$  was actually calculated from:

$$\ln P_i^x \approx \text{Stirling}(N) - \text{Stirling}(n_i^x) - \text{Stirling}(N - n_i^x) \\ + n_i^x \ln f_x + (N - n_i^x) \ln(1 - f_x)$$

To calculate  $\Delta G_{\text{stat}}, P_{\text{ref}}^x$  values were calculated from a hypothetical site where amino acid usage was set to that expected from the yeast usage statistics for  $N=1000$ . For each site, the statistical free energy was then calculated as:

$$\frac{\Delta G_{\text{stat}}}{kT^*} = \sqrt{\sum_x (\ln P_i^x - \ln P_{\text{ref}}^x)^2}$$

The Ranganathan group sets  $N=100$  and arbitrarily divides their statistical free energies by 100. We therefore also arbitrarily divided our free energy values by 100. However, due to the difference in  $N$ , the  $\Delta G_{\text{stat}}$  values here will be higher than those obtained by Lockless & Ranganathan for the same distribution of amino acid residues. In fact, using the Stirling approximation, it can be shown that for large  $N$ ,  $\ln P$  increases proportionally to  $N$ . Thus, the  $\Delta G_{\text{stat}}$  values calculated here with  $N=1000$  will be approximately tenfold greater than those calculated with  $N=100$ . The important point, however, is that  $P$ -values between sites are internally consistent if  $N$  is held constant.

### Perturbation analysis

Perturbation analyses (or statistical coupling analyses) were conducted by calculation of the statistical free energies between the entire MSA of TPRs and a subset with a given amino acid  $x$  at a position  $j$ . Specifically, the amino acid usage frequencies were extracted for a subset  $\delta_j$ , and the statistical free energies of the perturbation at each of the other 33 positions  $i$  were then calculated from:

$$\Delta \Delta G_{\text{stat}}^{i|\delta_j} = kT^* \sqrt{\sum_x \left( \ln \frac{P_{i|\delta_j}^x}{P_{\text{ref}|\delta_j}^x} - \ln \frac{P_i^x}{P_{\text{ref}}^x} \right)^2}$$

To adjust for the arbitrary differences in  $P$  associated with the number of sequences  $N$  in the subsets  $\delta_j$ , the number of sequences were all scaled to  $N=1000$ . However, not all subsets were suitable for this type of analysis due to small sample sizes. For example, Ala or Gly appeared in position 8 in 6562 of the 6887 sequences. What appears in any other position  $i$  where  $j$  is occupied by an amino acid other than Gly or Ala is meaningless, because the samples would not be statistically significant. Therefore, we arbitrarily decided that only subsets with at least 10% of the total number of sequences (689) would be subjected to this analysis. Since  $N$  was scaled to the same value for all  $P$ , the  $P_{\text{ref}}^x$  values are the same and drop out of the equation. The  $P$ -values were calculated as above using the Stirling approximation and the yeast usage frequencies  $f_x$ .

### Charge perturbation analysis

To calculate the approximate net charge of each of the 6887 TPRs, the following formula was applied:

$$\text{Net charge} \approx N_{\text{Lys}} + N_{\text{Arg}} + \frac{1}{2} N_{\text{His}} - N_{\text{Glu}} - N_{\text{Asp}}$$

Charge perturbations were calculated in the following way. For each position of interest (see Results and Discussion), all TPR sequences were extracted with Glu or Asp, or with Lys or Arg, at the specified position.

The average net charge at each position was then approximated using the following formula:

Mean net positional charge  $\approx$

$$(N_{\text{Lys}} + N_{\text{Arg}} - N_{\text{Glu}} - N_{\text{Asp}}) / N_{\text{subset}}$$

The mean net charge of the 33 ‘‘unperturbed’’ positions was then calculated by summing the mean net positional charges of those positions.

## Acknowledgements

The authors thank Rama Ranganathan (University of Texas Southwestern Medical Center, Dallas, TX), Ewan R. G. Main (Cambridge University, UK), and Luca D’Andrea (CNR, Naples, Italy) for helpful discussions. Thanks to Tommi Kajanader, Aitziber López Cortajarena and Janani Vankatraman (Yale University) for critical reading of this manuscript. T.J.M. is an N.I.H. Postdoctoral Fellow (GM065750). This work was supported, in part, by N.I.H. grants GM49146 and GM62413 (L.R.).

## Supplementary Data

Supplementary data associated with this article can be found, in the online version, at [doi:10.1016/j.jmb.2004.08.026](https://doi.org/10.1016/j.jmb.2004.08.026)

## References

- Desjarlais, J. R. & Berg, J. M. (1993). Use of a zinc-finger consensus sequence framework and specificity rules to design specific DNA binding proteins. *Proc. Natl Acad. Sci. USA*, **90**, 2256–2260.
- Steipe, B., Schiller, B., Pluckthun, A. & Steinbacher, S. (1994). Sequence statistics reliably predict stabilizing mutations in a protein domain. *J. Mol. Biol.* **240**, 188–192.
- Kohl, A., Binz, H. K., Forrer, P., Stumpp, M. T., Pluckthun, A. & Grutter, M. G. (2003). Designed to be stable: crystal structure of a consensus ankyrin repeat protein. *Proc. Natl Acad. Sci. USA*, **100**, 1700–1705.
- Lehmann, M., Kostrewa, D., Wyss, M., Brugger, R., D’Arcy, A., Pasamontes, L. & van Loon, A. P. (2000). From DNA sequence to improved functionality: using protein sequence comparisons to rapidly design a thermostable consensus phytase. *Protein Eng.* **13**, 49–57.
- Lehmann, M., Loch, C., Middendorf, A., Studer, D., Lassen, S. F., Pasamontes, L. *et al.* (2002). The consensus concept for thermostability engineering of proteins: further proof of concept. *Protein Eng.* **15**, 403–411.
- Ohage, E. C., Wirtz, P., Barnikow, J. & Steipe, B. (1999). Intrabody construction and expression. II. A synthetic catalytic Fv fragment. *J. Mol. Biol.* **291**, 1129–1134.
- Ohage, E. & Steipe, B. (1999). Intrabody construction and expression. I. The critical role of VL domain stability. *J. Mol. Biol.* **291**, 1119–1128.
- Knappik, A., Ge, L. M., Honegger, A., Pack, P., Fischer, M., Wellnhofer, G. *et al.* (2000). Fully synthetic human combinatorial antibody libraries (HuCAL) based on modular consensus frameworks and CDRs randomized with trinucleotides. *J. Mol. Biol.* **296**, 57–86.

9. Mosavi, L. K., Minor, D. L., Jr & Peng, Z. Y. (2002). Consensus-derived structural determinants of the ankyrin repeat motif. *Proc. Natl Acad. Sci. USA*, **99**, 16029–16034.
10. Mosavi, L. K. & Peng, Z. Y. (2003). Structure-based substitutions for increased solubility of a designed protein. *Protein Eng.* **16**, 739–745.
11. Binz, H. K., Stumpp, M. T., Forrer, P., Amstutz, P. & Pluckthun, A. (2003). Designing repeat proteins: well-expressed, soluble and stable proteins from combinatorial libraries of consensus ankyrin repeat proteins. *J. Mol. Biol.* **332**, 489–503.
12. Main, E. R. G., Xiong, Y., Cocco, M. J., D'Andrea, L. & Regan, L. (2003). Design of stable alpha-helical arrays from an idealized TPR motif. *Structure*, **11**, 497–508.
13. D'Andrea, L. & Regan, L. (2003). TPR proteins: the versatile helix. *Trends Biochem. Sci.* **28**, 655–662.
14. Stumpp, M. T., Forrer, P., Binz, H. K. & Pluckthun, A. (2003). Designing repeat proteins: modular leucine-rich repeat protein libraries based on the mammalian ribonuclease inhibitor family. *J. Mol. Biol.* **332**, 471–487.
15. Forrer, P., Binz, H. K., Stumpp, M. T. & Pluckthun, A. (2004). Consensus design of repeat proteins. *ChemBiochem*, **5**, 183–189.
16. Main, E. R. G., Jackson, S. E. & Regan, L. (2003). The folding and design of repeat proteins: reaching a consensus. *Curr. Opin. Struct. Biol.* **13**, 482–489.
17. Lehmann, M., Pasamontes, L., Lassen, S. F. & Wyss, M. (2000). The consensus concept for thermostability engineering of proteins. *Biochim. Biophys. Acta*, **1543**, 408–415.
18. Steipe, B. (1999). Evolutionary approaches to protein engineering. *Curr. Top. Microbiol. Immunol.* **243**, 55–86.
19. Lehmann, M. & Wyss, M. (2001). Engineering proteins for thermostability: the use of sequence alignments versus rational design and directed evolution. *Curr. Opin. Biotechnol.* **12**, 371–375.
20. Shortle, D. (2003). Propensities, probabilities and the Boltzmann hypothesis. *Protein Sci.* **12**, 1298–1302.
21. Lockless, S. W. & Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, **286**, 295–299.
22. Suel, G. M., Lockless, S. W., Wall, M. A. & Ranganathan, R. (2003). Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature Struct. Biol.* **10**, 59–69.
23. Hatley, M. E., Lockless, S. W., Gibson, S. K., Gilman, A. G. & Ranganathan, R. (2003). Allosteric determinants in guanine nucleotide-binding proteins. *Proc. Natl Acad. Sci. USA*, **100**, 14445–14450.
24. Pabo, C. O., Peisach, E. & Grant, R. A. (2001). Design and selection of novel Cys2His2 zinc finger proteins. *Annu. Rev. Biochem.* **70**, 313–340.
25. Forrer, P., Stumpp, M. T., Binz, H. K. & Pluckthun, A. (2003). A novel strategy to design binding molecules harnessing the modular nature of repeat proteins. *FEBS Letters*, **539**, 2–6.
26. Binz, H. K., Amstutz, P., Kohl, A., Stumpp, M. T., Briand, C. & Forrer, P. (2004). High-affinity binders selected from designed ankyrin repeat protein libraries. *Nature Biotechnol.* **22**, 575–582.
27. Scheufler, C., Brinker, A., Bourenkov, G., Pegoraro, S., Moroder, L., Bartunik, H. *et al.* (2000). Structure of TPR domain–peptide complexes: critical elements in the assembly of the Hsp70-Hsp90 multichaperone machine. *Cell*, **101**, 199–210.
28. Gatto, G. J., Jr, Geisbrecht, B. V., Gould, S. J. & Berg, J. M. (2000). Peroxisomal targeting signal-1 recognition by the TPR domains of human PEX5. *Nature Struct. Biol.* **7**, 1091–1095.
29. Bashford, D., Chothia, C. & Lesk, A. M. (1987). Determinants of a protein fold. Unique features of the globin amino acid sequences. *J. Mol. Biol.* **196**, 199–216.
30. Bowie, J. U., Reidhaar-Olson, J. F., Lim, W. A. & Sauer, R. T. (1990). Deciphering the message in protein sequences: tolerance to amino acid substitutions. *Science*, **247**, 1306–1310.
31. Sikorski, R. S., Boguski, M. S., Goebel, M. & Hieter, P. (1990). A repeating amino acid motif in CDC23 defines a family of proteins and a new relationship among genes required for mitosis and RNA synthesis. *Cell*, **60**, 307–317.
32. Penel, S., Hughes, E. & Doig, A. J. (1999). Side-chain structures in the first turn of the alpha-helix. *J. Mol. Biol.* **287**, 127–143.
33. Brinker, A., Scheufler, C., Von Der Mulbe, F., Fleckenstein, B., Herrmann, C., Jung, G. *et al.* (2002). Ligand discrimination by TPR domains. Relevance and selectivity of EEVD-recognition in Hsp70·Hop·Hsp90 complexes. *J. Biol. Chem.* **277**, 19265–19275.
34. Zweifel, M. E. & Barrick, D. (2001). Studies of the ankyrin repeats of the *Drosophila melanogaster* Notch receptor. 2. Solution stability and cooperativity of unfolding. *Biochemistry*, **40**, 14357–14367.
35. Mosavi, L. K., Williams, S. & Peng, Z. Y. (2002). Equilibrium folding and stability of myotrophin: a model ankyrin repeat protein. *J. Mol. Biol.* **320**, 165–170.
36. Bradley, C. M. & Barrick, D. (2002). Limits of cooperativity in a structurally modular protein: response of the Notch ankyrin domain to analogous alanine substitutions in each repeat. *J. Mol. Biol.* **324**, 373–386.
37. Visintin, M., Settanni, G., Maritan, A., Graziosi, S., Marks, J. D. & Cattaneo, A. (2002). The intracellular antibody capture technology (IACT): towards a consensus sequence for intracellular antibodies. *J. Mol. Biol.* **317**, 73–83.
38. Lopez Cortajarena, A., Kajander, T., Pan, W., Cocco, M. J. & Regan, L. (2004). Protein design to understand peptide ligand-recognition by tetratricopeptide proteins. *Protein Eng. Des. Select*, **17**, 399–409.
39. Wei, Y., Liu, T., Sazinsky, S. L., Moffet, D. A., Pelczar, I. & Hecht, M. H. (2003). Stably folded de novo proteins from a designed combinatorial library. *Protein Sci.* **12**, 92–102.
40. Kamtekar, S., Schiffer, J. M., Xiong, H., Babik, J. M. & Hecht, M. H. (1993). Protein design by binary patterning of polar and nonpolar amino acids. *Science*, **262**, 1680–1685.
41. Ranganathan, R. (2003). *Seventeenth Symposium of the Protein Society*, The Protein Society, Boston, MA.
42. Sonnhammer, E. L., Eddy, S. R. & Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins: Struct. Funct. Genet.* **28**, 405–420.

Edited by J. Thornton

(Received 27 May 2004; received in revised form 20 July 2004; accepted 10 August 2004)