nature Vol 437|22 September 2005|doi:10.1038/nature03991

ARTICLES

# Evolutionary information for specifying a protein fold

Michael Socolich[1,2*], Steve W. Lockless[1,2*†], William P. Russ[1,2], Heather Lee[1,2], Kevin H. Gardner[2,3] & Rama Ranganathan[1,2]

**Classical studies show that for many proteins, the information required for specifying the tertiary structure is contained in the amino acid sequence. Here, we attempt to define the sequence rules for specifying a protein fold by computationally creating artificial protein sequences using only statistical information encoded in a multiple sequence alignment and no tertiary structure information. Experimental testing of libraries of artificial WW domain sequences shows that a simple statistical energy function capturing coevolution between amino acid residues is necessary and sufficient to specify sequences that fold into native structures. The artificial proteins show thermodynamic stabilities similar to natural WW domains, and structure determination of one artificial protein shows excellent agreement with the WW fold at atomic resolution. The relative simplicity of the information used for creating sequences suggests a marked reduction to the potential complexity of the protein-folding problem.**

A fundamental tenet of biochemistry is that the amino acid sequence of a protein specifies its atomic structure and biochemical function[1]. But exactly what information in the sequence of a protein is necessary and sufficient for producing the fold and its biological activity? Despite considerable progress in understanding the mechanisms of protein folding[2,3], the answer to this fundamental question remains unknown. The main problem is the vast potential complexity of cooperative interactions between amino acids—processes by which the free energy contribution of one residue depends on those of other residues[4,5]. These amino acid couplings could be pairwise and local in the three-dimensional structure, but could also involve more complex cooperativities in which collections of residues interact through three-way or higher-order couplings[6–8]. Given that protein structures are typically compact and well packed[9,10], proteins could be dense and complex networks of inter-atomic interactions, requiring specification of a great number of mutual constraints between amino acid positions to define the fold.

An approach to defining the architecture of amino acid interactions in proteins is suggested by an evolution-based method known as the statistical coupling analysis (SCA). This method postulates that regardless of spatial location or underlying mechanism, the conserved functional coupling of sites in a protein should drive their mutual coevolution[11,12]. Given a sufficiently large and diverse multiple sequence alignment (MSA) of a protein family, the mutual dependencies should be evident in the conserved statistical correlations between amino acid distributions at sites. Application of the SCA in several different protein families reveals two general conclusions: (1) the global pattern of coevolutionary interactions is sparse, so that a small set of positions mutually coevolves among a majority that are largely decoupled, and (2) the strongly coevolving residues are spatially organized into physically connected networks linking distant functional sites in the structure through packing interactions[12–15]. Studies involving directed mutagenesis[12–14], structure determination[16], NMR dynamics[17], computational modelling[18,19] and literature study[15] implicate these networks of coevolving residues in contributing to core aspects of protein function.

More surprising, however, is the finding of sparseness. The SCA implies an unexpected degree of simplicity in amino acid interactions, with far fewer important constraints between residue pairs than would be expected from inspection of the atomic structure. Indeed, as has been pointed out[20,21], the evolution-based mapping of amino acid interactions does not look like the contact graph of the protein structure; many direct packing interactions show coevolution scores close to zero, and some distant sites linked through networks of coevolving residues are predicted to be coupled. Thus, the SCA mapping provides a picture of proteins as sparsely coupled architectures with redundant strong constraints linking a few sites, and a great deal of near-independent variation at most sites.

To test the overall hypothesis, we reasoned that if (and only if) the information contained in the SCA is a good estimate of the total sequence information for specifying a protein, it should be possible to computationally build artificial members of the protein family using no information except the SCA-based parameters of sequence conservation and coupling. In principle, these artificial sequences should fold into a structure representative of the family, and should function in a manner indistinguishable from their natural counterparts. In this and the accompanying paper[22], we test this hypothesis in a computationally and experimentally facile model system, the WW domain.

## SCA-based protein design

We began by carrying out the SCA for an alignment of 120 members of the WW domain family. WW domains are small, independently folding protein interaction modules that adopt a curved three-stranded β-sheet configuration and bind to proline-containing target sequences[23,24] (Fig. 1a). In the SCA, conservation at each position in the MSA is given in an energy-like statistical parameter that measures the deviation of the observed distribution of amino acids from their mean values found in all proteins (Fig. 1b). The evolutionary

[1]Howard Hughes Medical Institute and Departments of [2]Pharmacology and [3]Biochemistry, University of Texas Southwestern Medical Center, Dallas, Texas 75390-9050, USA.
†Present address: The Rockefeller University, 1230 York Avenue, New York, New York 10021, USA.
*These authors contributed equally to this work.

correlations between sequence positions are extracted by carrying out a perturbation analysis on the MSA; the conservation of amino acids at one site is perturbed (usually by restricting the site to one amino acid), and the impact of this perturbation on the conservation of amino acids at each of the other sites is measured in a distinct energy-like statistical parameter. Supplementary Fig. 1 shows an example of this calculation for one site in the WW MSA. Figure 1c shows a matrix representation of statistical coupling values for five different site-specific perturbations in the WW MSA that demonstrates the core results of the SCA. Perturbation experiments at all these sites simply expose the same redundant pattern of coevolution between a small set of moderately conserved positions. Not all moderately conserved positions are coupled—several other sites showing similar conservation scores (for example, positions 16, 17 or 25; Fig. 1b) show little coevolution with the sites chosen for perturbation.

The simplicity of these findings suggests the possibility that just the frequency distribution of amino acids at sites (conservation) plus the few rules of coupling in the SCA matrix may amount to the total constraints on WW sequences. To test this idea, we developed two computational algorithms for designing novel protein sequences using only SCA information. The first algorithm tests the necessity of the information in the SCA matrix by building artificial sequences that preserve the amino acid composition at sites but eliminate all statistical couplings between sites. Thus, this algorithm presumes that conservation can be adequately described as an intrinsic property of each site. To implement this algorithm, we simply select amino acids at each site independently based on the observed frequency distribution in the natural alignment. Accordingly, the SCA matrix for an alignment of 120 such sequences (termed the site-independent conservation (IC) model) shows no statistical coupling between positions (Fig. 1d). The second algorithm tests the sufficiency of the SCA matrix by building artificial sequences that preserve both the conservation pattern and the pattern of statistical couplings. These sequences are built using a Monte-Carlo-based simulated annealing protocol in which amino acids are completely shuffled within every column of the MSA while minimizing the difference between all coupling values in the SCA matrix for the artificial alignment and for the natural alignment. At convergence, this algorithm produces alignments of novel sequences ($\sim 10^5$ substitutions per sequence) that display the same conservation pattern and also closely reproduce the pattern of statistical couplings observed in the natural alignment (termed coupled conservation (CC) model; compare Fig. 1c and e).

### Construction of designed sequences

To evaluate the designed sequences for folding, we constructed libraries of synthetic genes for expression of the artificial proteins in *Escherichia coli*. Four libraries were built using a DNA-oligonucleotide-based gene synthesis protocol (Supplementary Fig. 2): (1) 42 natural WW sequences (N), drawn randomly from the MSA; (2) 43 IC sequences, built on the premise that conservation is strictly an intrinsic property of each site; (3) 43 CC sequences, built on the premise that conservation is a distributed property, parsed among sites in the manner described by the SCA matrix; and (4) 19 random sequences (R), in which amino acids at all sites were randomly drawn from their mean frequencies in the WW MSA. The natural sequences were built as positive controls because we do not know how many of these will fold when expressed as recombinant proteins in bacteria. The random sequences are negative controls; they contain no site-specific information and are not expected to fold.

Table 1 shows the comparative statistical properties of the designed sequences. Natural, IC and CC WW sequences show a mean amino acid identity to all natural WW domains of ~36%, an expected result because all of these sequences contain the same pattern of conservation at sites. Another expected finding was that random sequences show much weaker identities to natural WW domains (~6%). However, the IC and CC sequences also show similar 'top-hit' identities—the per cent identity to their closest counterpart in the natural world (55.7 ± 5.6% (IC) and 58.6 ± 7.2% (CC), $P = 0.11$, Kolmogorov–Smirnov test; mean ± s.d. is shown). This finding shows that despite additional constraints in their design, CC sequences are about as diverged as IC sequences; by this measure, the number of extra constraints arising from the SCA matrix is small.

### Experimental analysis of designed sequences

Figure 2a shows a flowchart of experiments for each WW protein. Proteins were expressed as $His_8$-tagged fusions, purified using $Ni^+$-NTA affinity chromatography, and subjected to SDS–polyacrylamide gel electrophoresis (PAGE) analysis to evaluate expression and solubility. All libraries contained sequences that expressed poorly despite multiple attempts, or produced insoluble aggregates (Fig. 2b, first and second columns); these were scored as not folded. In

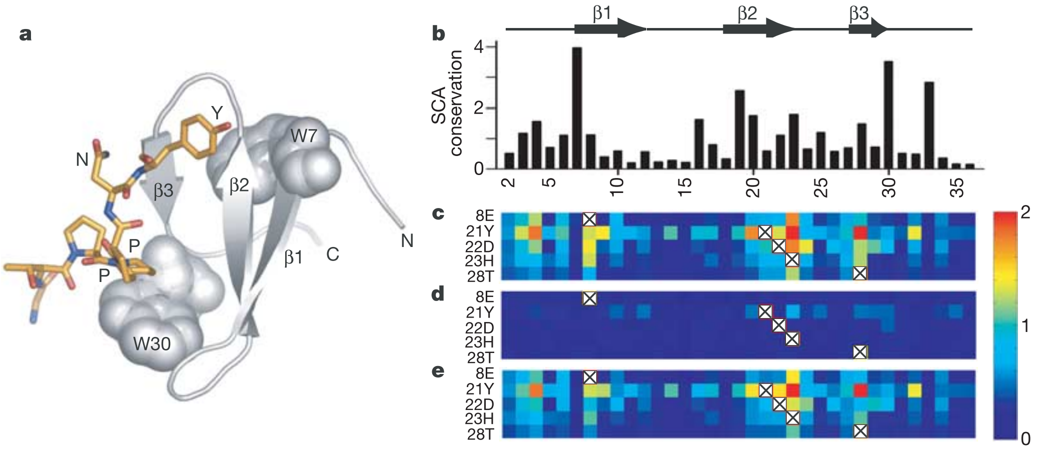

**Figure 1 | SCA-based protein design.** **a**, Structure of a representative WW domain (Nedd4.3, Protein Data Bank 1I5H) in complex with a target peptide (in stick representation). The two canonical tryptophans are shown as space-filling side chains. The figure was prepared using PyMol[51]. **b**, SCA conservation scores for each position in the WW alignment in arbitrary units of statistical energy[12]. Position numbers (*x* axis) and the secondary structure diagram at the top coincide with matrix columns in **c**–**e**. **c**, A matrix representation of statistical coupling values from perturbation analysis of five positions (rows) in the WW domain MSA. **d**, The matrix for an alignment of IC sequences, built by randomly selecting amino acids at each site from the observed frequency distributions in the natural alignment. **e**, The matrix for an alignment of CC sequences, derived from a design algorithm where both the conservation pattern and the pattern of statistical couplings in the natural alignment are preserved. Scale bar shows the SCA coevolution score, ranging from 0 (blue) to 2 (red).

**Table 1 | Statistical properties of natural and artificial WW domains**

| WW library | Identity to natural WW domains in MSA (%; mean ± s.d.) | Identity to closest natural WW domain in MSA (%; mean ± s.d.) | Number of sequences |
|---|---|---|---|
| Natural*<br>(folded) | 35.3 ± 6.1<br>(36.2 ± 5.8) | 80.4 ± 14.5<br>(78.2 ± 15.9) | 42<br>(28) |
| IC† | 36.1 ± 3.0 | 55.7 ± 5.6 | 43 |
| CC‡<br>(folded) | 35.0 ± 4.8<br>(37.4 ± 5.4) | 58.6 ± 7.2<br>(63.1 ± 6.0) | 43<br>(12) |
| Random§ | 6.3 ± 2.5 | 14.9 ± 4.3 | 19 |

*Drawn randomly from the natural WW MSA.
†Created by randomly selecting amino acids at each site from the observed frequency distribution at that site in the natural WW MSA.
‡Created by Monte-Carlo simulation.
§Created by randomly selecting amino acids at each site from their overall mean frequencies in the WW MSA.

addition, all libraries contained some fraction of well-expressed and soluble proteins (Fig. 2b, third column). These were further evaluated using thermal denaturation experiments and $^1$H-NMR for evidence of a native state, as described below. Expression data for all 147 synthetic proteins are provided in Supplementary Fig. 3.

A hallmark of natively folded small proteins is cooperative and reversible transition between folded and unfolded states. In the WW domain, this folding equilibrium and the consistency of the two-state approximation have been well described[25,26]. Here, we followed the folding reaction by monitoring the fluorescence of a buried tryptophan (Trp 7), which becomes quenched due to solvent exposure upon thermal denaturation and therefore reports the fraction of protein folded as a function of temperature (for example, Fig. 3a–c). We tested all 105 well-expressed and soluble proteins from the four sequence libraries for cooperative and reversible thermal transitions (Supplementary Fig. 4); a representative sampling of these data is shown in Fig. 3. Natural WW domains show a range of thermal denaturation profiles (Fig. 3a). Some, such as N1, are clearly well folded, showing a cooperative denaturation with thermodynamic parameters typical for WW domains (van't Hoff enthalpy ($\Delta H_{\mathrm{u}}^{\mathrm{VH}}$) = 22.5 kcal mol$^{-1}$, $T_{\mathrm{m}} = 46.8\,^{\circ}\mathrm{C}$, where $T_{\mathrm{m}}$ is the melting temperature). Others, such as N22, cooperatively denature but are less stable ($\Delta H_{\mathrm{u}}^{\mathrm{VH}} = 18.5\,\mathrm{kcal\,mol^{-1}}$, $T_{\mathrm{m}} = 25.2\,^{\circ}\mathrm{C}$). Still others, such as N36, despite good solubility, show no convincing evidence of a native state given the experimental conditions of the assay. To provide independent support for categorization of these denaturation profiles as folded or not folded, we collected $^1$H-NMR spectra for a representative sampling of natural WW proteins. Characteristic features of folded WW proteins include good chemical shift dispersion of peaks corresponding to backbone amide protons ($\delta > 9$ parts per million (p.p.m.)), often accompanied by distinct chemical shifts of the two indole nitrogen protons down-field of 10 p.p.m. (the two canonical Trp residues are in distinct chemical environments—one in the core and one solvent exposed, see Fig. 1a), and up-field chemical shifts ($\delta < 0.5$ p.p.m.) corresponding to side-chain methyl protons[26]. Spectra for N1 (Fig. 3d) and N22 (Fig. 3e) confirm that these are indeed natively folded. As suggested by thermal melts, the NMR spectrum for N36 (Fig. 3f) shows none of these hallmarks, consistent with an unfolded protein. On the basis of the analysis of natural WW sequences, we determined rigorous criteria for folding: artificial sequences were declared as folded only if they showed cooperative and reversible thermal denaturation and $^1$H-NMR spectra consistent with a native state (Supplementary Figs 4 and 5).

Figure 3b, g–i shows similar data for a representative set of CC sequences. Some, such as CC16 (Fig. 3i), are, like N36, soluble but not folded by either criterion. Others, such as CC45 (Fig. 3g) or CC18 (Fig. 3h), show thermal denaturation profiles that fall into the same range as natural WW domains (CC45: $\Delta H_{\mathrm{u}}^{\mathrm{VH}} = 32.4\,\mathrm{kcal\,mol^{-1}}$, $T_{\mathrm{m}} = 65.6\,^{\circ}\mathrm{C}$; CC18: $\Delta H_{\mathrm{u}}^{\mathrm{VH}} = 19.63\,\mathrm{kcal\,mol^{-1}}$, $T_{\mathrm{m}} = 34.3\,^{\circ}\mathrm{C}$), and show strong evidence in $^1$H-NMR spectra of being natively folded. In contrast, no IC sequences showed evidence of native folding. Figure 3c shows thermal denaturation profiles for every soluble IC sequence ($n = 30$, grey) and Fig. 3j–l shows corresponding $^1$H-NMR spectra for the three IC sequences that constitute the best case scenarios by thermal melts. The data demonstrate that some CC sequences are natively folded with thermodynamic properties similar to that of natural WW domains, and that no IC sequences are natively folded.

Figure 4a summarizes the experimental analysis of all 147 WW

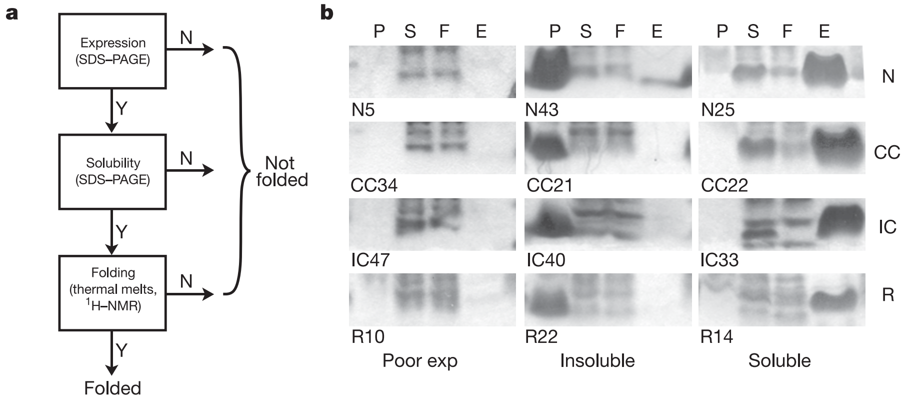

**Figure 2 | Experimental evaluation of artificial proteins (part 1). a**, A flowchart of experiments for evaluating WW sequences. Each sequence was expressed as a His$_8$-tagged fusion in *E. coli*, tested for solubility, and if soluble, subjected to thermal denaturation and, in most cases, $^1$H-NMR. **b**, SDS–PAGE analysis of WW sequences drawn from four different sequence libraries: N, natural; CC, coupled conservation model; IC, site-independent conservation model; R, random. Lanes represent pellet or insoluble fraction (P), soluble fraction (S), flow-through after Ni$^+$-NTA affinity chromatography (F) and eluted proteins (E).

sequences tested. Sixty-seven per cent of natural WW sequences were folded by the criteria described above, a number that simply reflects the efficiency of producing randomly chosen WW domains in our expression system. As expected, no random sequences were folded, although nearly half of these were well expressed and soluble. No IC sequences were natively folded, although a substantial fraction (70%) was soluble. The finding that these sequences showed a greater probability of being soluble in comparison to random sequences suggests that conservation taken at sites alone may provide enough information about the folding process to permit hydrophobic collapse to a molten globule-like state. However, the site-independent conservation model is insufficient to produce the native state. In contrast, over one-quarter of the CC sequences were natively folded. Figure 4b compares the unfolding enthalpy and melting temperatures derived from thermal denaturation experiments for all folded natural WW domains and CC sequences, showing that the CC sequences fall into the same range of thermodynamic parameters as their natural counterparts ($P = 0.1$, multivariate analysis of variance (ANOVA)). Just like natural WW domains, the CC domains are marginally stable folds in which a fine balance of opposing forces probably accounts for the distinction between folded and non-folded states.

Folding to a native state involves efficient packing of hydrophobic atoms within the interior of proteins, a factor that leads to higher average sequence conservation in the core of proteins[27]. Although CC sequences show similar divergence from natural WW domains as IC sequences overall (Table 1), we wondered whether CC sequences might natively fold because they are more similar to WW domains within the core. To examine this, we calculated sequence identities within core residues (3, 7, 20, 22, 33) between CC or IC sequences and natural WW domains. CC sequences show 66.7 ± 7.0% mean and 97 ± 7.1% top-hit identities to natural sequences for these positions (mean ± s.d.), values that reflect the near invariance of some core positions in the MSA. However, IC sequences are no different from CC sequences, showing 68.3 ± 9% mean and 95.4 ± 9.4% top-hit identities to natural sequences for core positions (mean ± s.d.). Also, the folded subset of CC sequences shows the same core sequence identity (67.8 ± 7% mean and 98.3 ± 6% top-hit; mean ± s.d.). Thus, native folding in CC sequences is not explained by a more natural-like composition of core residues.

Taken together, the data argue that in addition to the amino acid distributions at sites, the statistical coupling information is necessary and sufficient to specify native folding. Because no information about the WW domain except the coupling values in the SCA matrix was used to constrain the CC sequences beyond the IC sequences, we conclude that the specific topology of mutual constraints between sites predicted by the SCA is one solution for achieving the folded state of this protein family.

## Atomic structure of an artificial WW domain

Do the artificial proteins adopt the canonical WW fold? To examine this, we pursued in-depth structural analysis of one of the CC proteins, CC45. CC45 shows 39% mean identity and 61% top-hit identity to natural WW sequences, values that are near the average

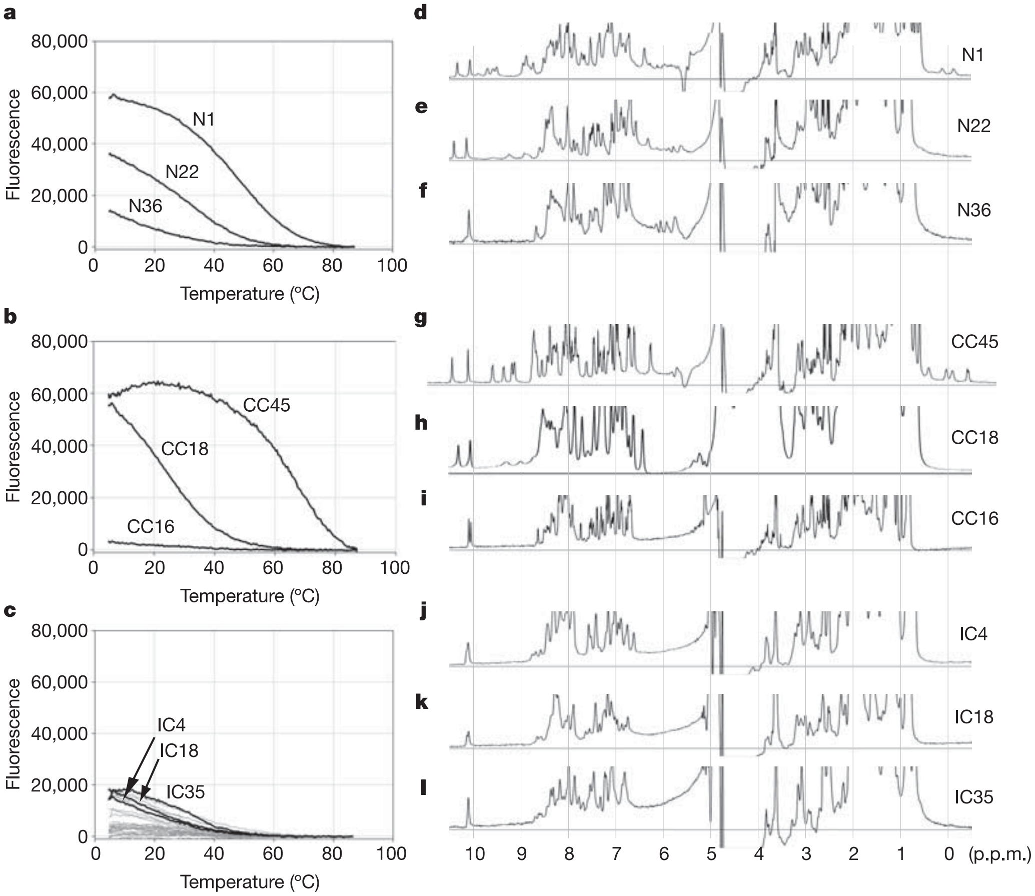

**Figure 3 | Experimental evaluation of artificial proteins (part 2).** Thermal denaturation studies and corresponding $^1$H-NMR spectra for a sampling of WW sequences drawn from the natural (**a**, **d–f**), CC (**b**, **g–i**), or IC model (**c**, **j–l**) alignments. For natural and CC sequences, three sequences are shown that were highly stable (**d**, **g**), moderately stable (**e**, **h**), and unfolded (**f**, **i**). For IC sequences, melts of every soluble sequence ($N = 30$) are overlaid in grey; $^1$H-NMR spectra of three sequences (in black) with the best potential for some native structure are shown in **j–l**. Whereas natural and CC sequences include some that are folded, no IC sequences are folded. Fluorescence is reported in counts s$^{-1}$.

for both CC and IC sequences (Table 1). As described, this protein is well folded by several independent biophysical assays (Fig. 3b, g). We solved the three-dimensional structure of CC45 by solution NMR methods, using 800 distance and dihedral angle restraints (Fig. 5a, b; see also Supplementary Table 2). CC45 was refined to reasonably high precision, clearly confirming the curved three-stranded antiparallel β-sheet structure characteristic of all WW domains. However, the structural similarity of CC45 to other WW domains seems to go well beyond just the fold level. Tertiary structure motifs common to WW domains are found in CC45, including a centrally located tryptophan (W7) that sits upon a platform of two proline side chains (P4, P33, Fig. 5b). In addition, several sites in CC45 display unusual proton chemical shifts based on comparison with the BioMagResBank database of protein NMR data (for example, $\delta = -0.38$ p.p.m. for N22 Hβ2, Fig. 3g; see Supplementary Table 1 for further examples). Such unusual shifts arise from unique tertiary packing that places protons in close proximity to aromatic side chains. Comparison of the chemical shifts of CC45 to those of two natural WW domains, Pin1 (ref. 28) and Nedd4.3 (ref. 29), showed that all three proteins display these same unusual shifts at analogous positions in each sequence. Thus, CC45 adopts a stable WW-like three dimensional structure.

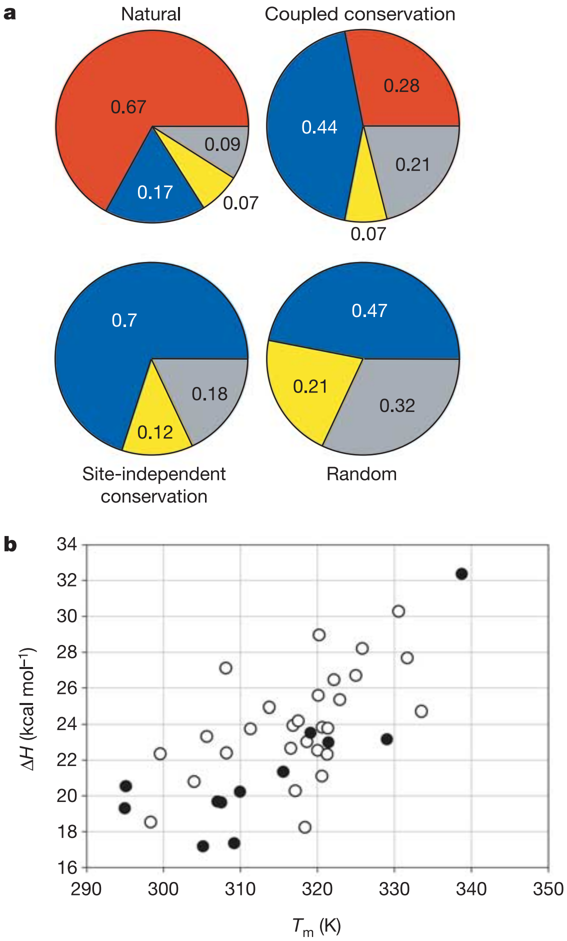

**Figure 4 | Summary of experiments on all natural and artificial WW sequences.** **a**, A pie chart showing the outcomes of folding studies for natural ($n = 42$), CC ($n = 43$), IC ($n = 43$), or random ($n = 19$) WW sequences. Red, natively folded; blue, soluble but unfolded; yellow, insoluble; grey, poor expressing. **b**, Melting temperatures ($T_m$) and van't Hoff enthalpies of unfolding for all folded WW sequences. Open circles indicate natural sequences and filled circles indicate the 12 folded CC sequences. The artificial sequences show thermodynamic parameters that fall into the same range as that of natural WW domains.

To examine how well the SCA-based design recapitulates the native structure of the WW domain, we overlaid the structure of CC45 with those of several different natural WW domains by minimizing the root mean squared deviations (r.m.s.d.) of backbone Cα atoms for all structures (Fig. 5c). All seven structures clearly adopt very similar backbone folds, and no obvious feature seems to distinguish CC45 from the other proteins. To quantify this result, we compared the average pairwise r.m.s.d. values for backbone atoms of all the natural WW domains (1.52 ± 0.4 Å) with the r.m.s.d. values of CC45 from all the natural domains (1.19 ± 0.65 Å). The difference between these distributions is not significant ($P = 0.34$), demonstrating that CC45 is as similar at atomic resolution to natural WW domains as natural domains are to each other.

## Conclusions

Classical studies indicate that protein structure and function results from globally minimizing the free energy of the polypeptide chain under physiological conditions[1]. In this work, we have tested a model for the specific pattern of amino acid interactions that makes up the global free energy minimum using the WW domain as a model system (Fig. 1b, c). The model is based on three core hypotheses: (1) that amino acid interactions specifying the atomic structure are conserved throughout members of a protein family rather than being idiosyncratic; (2) that conservation is a distributed rather than site-independent property because it fundamentally arises from the cooperativity of energetic interactions; and (3) that the parsing of conservation is well estimated by the statistical energy function contained in the SCA method. The finding that a significant fraction of CC sequences are natively folded whereas IC sequences are not provides strong support for these hypotheses. This result is particularly informative because CC sequences are statistically indistinguishable from IC sequences with regard to sequence divergence from natural WW domains. We conclude that it is the specific distribution of conservation rather than the quantity of conservation that dictates native folding.

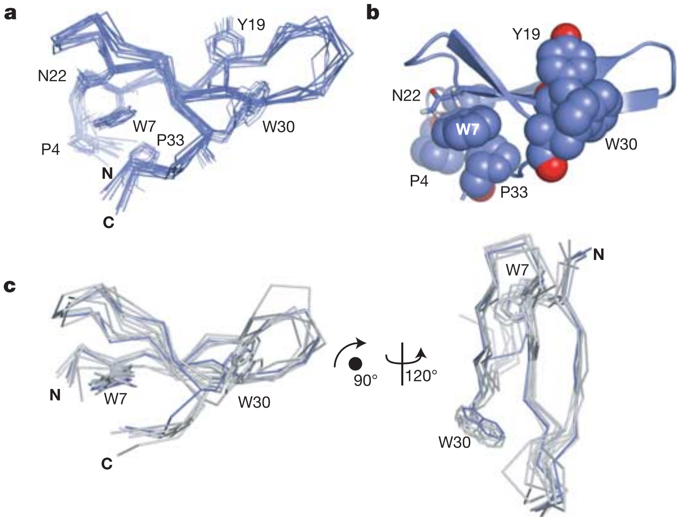

**Figure 5 | NMR structure determination of CC45, an artificial WW domain.** **a**, Ensemble of ten lowest energy structures determined for CC45, showing backbone traces of each structure and side chains of selected residues discussed. **b**, Ribbon diagram of the representative structure of the CC45 ensemble. Residues highlighted in CPK or stick bonds illustrate packing interactions in the core (Trp 7, Pro 33, Asn 22), and in the canonical proline-binding pocket (Tyr 19, Trp 30). **c**, Structure-based alignment of CC45 (blue) with six natural WW domains in white (FBP28WW (1E0L), YJQ8WW (1E0N), dystrophin (1EG3), Nedd4.3 (1I5H), YAP65 (1K9R) and Pin1 (1PIN)). The data show that the CC45 adopts a WW-like tertiary structure.

How is the sparse architecture of residue interactions suggested by the SCA consistent with the fact that proteins are well packed, and with the intuitive sense that residues closer to each other in the structure should interact more strongly? Indeed, mutations made in spatially local sites ($<5$ Å apart) are more likely to be thermodynamically coupled[30], a finding that has been used to argue that the energetic architecture of proteins could be dense and local rather than sparse and distributed[21]. However, mutation-based free energy measurements do not measure the energetic value of wild-type residues; they measure the energetic effect due to mutation. When close in space, it becomes difficult to separate the average spatial correlation of mutation-induced effects in proteins from the intrinsic interactions between the wild-type residues in proteins. For example, alanine mutations are often thought of as loss-of-function mutations in proteins, but these mutations can cause local cavities to form in structures, and the energetic value of such mutations includes a cavity-dependent contribution[31]. In such cases, it is not surprising that pairs of mutations in local regions show energetic coupling. In contrast, SCA-based protein design is a massive perturbation experiment in which the mutagenesis at every site is constrained by its own frequency distribution in the MSA and by coupling to those few other sites that evolution shows can provide compensatory mutations. Consequently, a large number of packing interactions are decoupled in the design of CC sequences and vary independently, consistent with findings that proteins are generally tolerant to mutation[32]. We suggest that this experiment provides a more informative and extensive test of the architecture of amino acid interactions in proteins.

Nevertheless, it is also clear that the density of atoms within the core of proteins is high, similar to that in crystals of free amino acids[9,10]. At first glance, this finding seems to highlight the importance of local interactions in proteins to create this high level of order. However, recent studies show that although the mean density of atoms in the core of proteins suggests a solid-like arrangement, the distribution of packing densities is much more like that of a liquid[33,34]. That is, packing is heterogeneous in protein cores with some highly ordered regions mixed with many regions that are not as highly ordered. A number of studies are now beginning to add experimental support for the heterogeneous and distributed nature of mechanical interactions in proteins[35–37]. It will be important to test further the physical model that emerges from these studies and the SCA: proteins are heterogeneous energetic architectures with a network of coupled residues existing in the environment of many weaker interactions.

## METHODS

**Statistical coupling analysis.** Statistical coupling analysis was performed as previously described[12]. WW domain sequences were collected using PSI-BLAST[38] (*e*-scores $\leq 0.001$) from the non-redundant database of protein sequences (release date October 2000), and aligned using ClustalW[39]. The five statistical perturbations used in this study (8E, 21Y, 22D, 23H and 28T) were chosen by knowledge of structural or functional importance in the WW domain[40,41].

**Protein design algorithms.** IC model sequences were created by randomly drawing amino acids for each site from the corresponding amino acid frequency distributions in the natural WW alignment. Random sequences were created by randomly selecting residues at every site from the overall frequency distribution in the MSA. CC model sequences were generated using a Monte-Carlo simulated annealing algorithm, described in the Supplementary Methods section.

**Gene construction and protein expression.** Genes encoding artificial sequences were designed by back-translating designed protein sequences using *E. coli* codon optimization (Vector NTI Suite, Informax Inc), built on overlapping DNA oligonucleotides, assembled using the polymerase chain reaction, and cloned into the pHIS8-3 expression vector (provided by J. Noel). Constructs were verified by DNA sequencing. Proteins were expressed in JM109(DE3) cells grown at 37 °C in Terrific broth to an absorbance at 600 nm of ~1.2, and induced with 0.5 mM IPTG at 18 °C overnight. For fluorescence experiments, 50 ml culture was lysed in 3 ml binding buffer (25 mM TrisHCl, pH 8.0, 0.5 M NaCl, 5 mM imidazole) by sonication followed by centrifugation and incubation of the cleared lysate with 75 μl bed volume $Ni^+$-NTA (Qiagen) for 30 min at 4 °C. After washing (three times with 15 ml binding buffer), WW proteins were eluted in elution buffer (100 mM TrisHCl, pH 8.0, 1 M NaCl, 400 mM imidazole). Cultures were scaled up for NMR experiments, using 1–2 l of Terrific broth and 0.5 ml of affinity resin.

**Thermal denaturation assay.** Purified proteins were diluted into 2 ml elution buffer for a final WW domain concentration of ~1–10 μM. Protein fluorescence was monitored at 340 nm (excitation at 295 nm) using a spectrofluorometer (Photon Technologies Inc.) outfitted with a Peltier temperature controller. Data were acquired while changing the temperature from 4 °C to 90 °C and back to 4 °C, at a rate of $2\,°\mathrm{C}\,\mathrm{min}^{-1}$ with 5 min equilibration periods at 4 °C and 90 °C. Melting curves for 10 μM free tryptophan were subtracted to account for the intrinsic temperature dependence of Trp fluorescence. Melting temperature ($T_\mathrm{m}$) and enthalpy of unfolding at the $T_\mathrm{m}$ were calculated by fitting the first derivative of the denaturation curves to the differential form of the van't Hoff equation[42].

**NMR spectroscopy.** All NMR spectra were recorded on 500 and 600 MHz Varian Inova spectrometers. One-dimensional $^1$H-NMR spectra of various WW domains were obtained on samples containing 100 μM to 1 mM protein in 100 mM NaCl, 100 mM sodium phosphate buffer (pH 7.0) and 90%:10% $H_2O$:$D_2O$. All spectra were recorded at 25 °C using a 3-9-19 Watergate sequence[43] for water suppression. For the CC45 chemical shift assignments and solution structure determination, a combination of two-dimensional double-quantum filtered correlation spectroscopy (DQF-COSY), total correlated spectroscopy (TOCSY; $\tau_\mathrm{mix} = 60$ ms) and nuclear Overhauser effect spectroscopy (NOESY; $\tau_\mathrm{mix} = 150$ ms) were recorded on 600 μM unlabelled protein samples in the above buffer at 25 °C and 38 °C, along with two-dimensional TOCSY ($\tau_\mathrm{mix} = 60$ ms) and NOESY ($\tau_\mathrm{mix} = 60$, 150 ms) spectra recorded at 25 °C on a 1 mM protein sample made in the same buffer with 99% $D_2O$ solvent. Additional spectra were recorded on an 850 μM sample of uniformly $^{15}$N-labelled CC45 in 90%:10% $H_2O$:$D_2O$, including three-dimensional $^{15}$N-edited TOCSY ($\tau_\mathrm{mix} = 60$ ms), three-dimensional $^{15}$N-edited NOESY ($\tau_\mathrm{mix} = 150$ ms), three-dimensional HNHA and two-dimensional $^{15}$N/$^1$H heteronuclear single quantum correlation (HSQC) experiments. All spectra were processed using the NMRPipe package[44] and analysed using nmrView[45]. The CC45 sequence is $NH_3^+$-MPLPPGWERRTDVEGKVYYFNVRTLTTTWERPTIILE-$COO^-$.

**Structure determination.** Distance restraints were obtained by analysing the three-dimensional $^{15}$N-edited NOESY and two-dimensional homonuclear NOESY spectra recorded in 99% $D_2O$ (both $\tau_\mathrm{mix} = 150$ ms) using the CNS package[46] with the ARIA 1.2 extension[47]. Additional restraints on the backbone dihedral angles $\phi$ and $\psi$ were derived from a TALOS-based analysis of backbone chemical shifts[48] and supplemented by 3J(HN-Hα) coupling constants measured in a three-dimensional HNHA spectrum. Hydrogen bond restraints for a total of six backbone amides were generated for sites showing β-sheet structure by dihedral angle analysis and typical inter-strand nuclear Overhauser effects (NOEs). One hundred structures were generated in the last ARIA iteration, out of which the ten with the lowest energies were selected for a final refinement stage in water. The resulting ensemble of ten structures was used for all analyses as implemented in the programs PROCHECK-NMR[49] and MOLMOL[50]. Ramachandran statistics for CC45 are: 80.7 ± 6.3% in most favoured regions, 18.0 ± 7.3% in additionally allowed regions, 1.0 ± 1.6% in generously allowed regions, and 0.3 ± 1.0% in disfavoured regions.

**Received 25 January; accepted 30 June 2005.**

1. Anfinsen, C. B. Principles that govern the folding of protein chains. *Science* **181**, 223–230 (1973).
2. Daggett, V. & Fersht, A. The present view of the mechanism of protein folding. *Nature Rev. Mol. Cell Biol.* **4**, 497–502 (2003).
3. Dinner, A. R., Sali, A., Smith, L. J., Dobson, C. M. & Karplus, M. Understanding protein folding via free-energy surfaces from theory and experiment. *Trends Biochem. Sci.* **25**, 331–339 (2000).
4. Hidalgo, P. & MacKinnon, R. Revealing the architecture of a $K^+$ channel pore through mutant cycles with a peptide inhibitor. *Science* **268**, 307–310 (1995).
5. Horovitz, A. & Fersht, A. R. Co-operative interactions during protein folding. *J. Mol. Biol.* **224**, 733–740 (1992).
6. Chen, J. & Stites, W. E. Higher-order packing interactions in triple and quadruple mutants of staphylococcal nuclease. *Biochemistry* **40**, 14012–14019 (2001).
7. LiCata, V. J. & Ackers, G. K. Long-range, small magnitude nonadditivity of mutational effects in proteins. *Biochemistry* **34**, 3133–3139 (1995).
8. Luque, I., Leavitt, S. A. & Freire, E. The linkage between protein folding and functional cooperativity: two sides of the same coin? *Annu. Rev. Biophys. Biomol. Struct.* **31**, 235–256 (2002).

9. Gerstein, M. & Chothia, C. Packing at the protein-water interface. *Proc. Natl Acad. Sci. USA* **93,** 10167–10172 (1996).
10. Richards, F. M. & Lim, W. A. An analysis of packing in the protein folding problem. *Q. Rev. Biophys.* **26,** 423–498 (1993).
11. Lichtarge, O., Bourne, H. R. & Cohen, F. E. An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257,** 342–358 (1996).
12. Lockless, S. W. & Ranganathan, R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* **286,** 295–299 (1999).
13. Hatley, M. E., Lockless, S. W., Gibson, S. K., Gilman, A. G. & Ranganathan, R. Allosteric determinants in guanine nucleotide-binding proteins. *Proc. Natl Acad. Sci. USA* **100,** 14445–14450 (2003).
14. Shulman, A. I., Larson, C., Mangelsdorf, D. J. & Ranganathan, R. Structural determinants of allosteric ligand activation in RXR heterodimers. *Cell* **116,** 417–429 (2004).
15. Suel, G. M., Lockless, S. W., Wall, M. A. & Ranganathan, R. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature Struct. Biol.* **10,** 59–69 (2003).
16. Peterson, F. C., Penkert, R. R., Volkman, B. F. & Prehoda, K. E. Cdc42 regulates the Par-6 PDZ domain through an allosteric CRIB-PDZ transition. *Mol. Cell* **13,** 665–676 (2004).
17. Fuentes, E. J., Der, C. J. & Lee, A. L. Ligand-dependent dynamics and intramolecular signalling in a PDZ domain. *J. Mol. Biol.* **335,** 1105–1115 (2004).
18. Estabrook, R. A. *et al.* Statistical coevolution analysis and molecular dynamics: identification of amino acid pairs essential for catalysis. *Proc. Natl Acad. Sci. USA* **102,** 994–999 (2005).
19. Ota, N. & Agard, D. A. Intramolecular signalling pathways revealed by modeling anisotropic thermal diffusion. *J. Mol. Biol.* **351,** 345–354 (2005).
20. Fodor, A. A. & Aldrich, R. W. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins* **56,** 211–221 (2004).
21. Fodor, A. A. & Aldrich, R. W. On evolutionary conservation of thermodynamic coupling in proteins. *J. Biol. Chem.* **279,** 19046–19050 (2004).
22. Russ, W. P., Lowery, D. M., Mishra, P., Yaffe, M. B. & Ranganathan, R. Natural-like function in artificial WW domains. *Nature* doi:10.1038/nature03990 (this issue).
23. Macias, M. J. *et al.* Structure of the WW domain of a kinase-associated protein complexed with a proline-rich peptide. *Nature* **382,** 646–649 (1996).
24. Bork, P. & Sudol, M. The WW domain: a signalling site in dystrophin? *Trends Biochem. Sci.* **19,** 531–533 (1994).
25. Jager, M., Nguyen, H., Crane, J. C., Kelly, J. W. & Gruebele, M. The folding mechanism of a β-sheet: the WW domain. *J. Mol. Biol.* **311,** 373–393 (2001).
26. Koepf, E. K., Petrassi, H. M., Sudol, M. & Kelly, J. W. WW: An isolated three-stranded antiparallel β-sheet domain that unfolds and refolds reversibly; evidence for a structured hydrophobic cluster in urea and GdnHCl and a disordered thermal unfolded state. *Protein Sci.* **8,** 841–853 (1999).
27. Bashford, D., Chothia, C. & Lesk, A. M. Determinants of a protein fold. Unique features of the globin amino acid sequences. *J. Mol. Biol.* **196,** 199–216 (1987).
28. Wintjens, R. *et al.* 1H NMR study on the binding of Pin1 Trp-Trp domain with phosphothreonine peptides. *J. Biol. Chem.* **276,** 25150–25156 (2001).
29. Kanelis, V., Rotin, D. & Forman-Kay, J. D. Solution structure of a Nedd4 WW domain-ENaC peptide complex. *Nature Struct. Biol.* **8,** 407–412 (2001).
30. Schreiber, G. & Fersht, A. R. Energetics of protein-protein interactions: analysis of the barnase-barstar interface by single mutations and double mutant cycles. *J. Mol. Biol.* **248,** 478–486 (1995).
31. Eriksson, A. E. *et al.* Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. *Science* **255,** 178–183 (1992).
32. Bowie, J. U., Reidhaar-Olson, J. F., Lim, W. A. & Sauer, R. T. Deciphering the message in protein sequences: tolerance to amino acid substitutions. *Science* **247,** 1306–1310 (1990).
33. Liang, J. & Dill, K. A. Are proteins well-packed? *Biophys. J.* **81,** 751–766 (2001).
34. Lindorff-Larsen, K., Best, R. B., Depristo, M. A., Dobson, C. M. & Vendruscolo, M. Simultaneous determination of protein structure and dynamics. *Nature* **433,** 128–132 (2005).
35. Benkovic, S. J. & Hammes-Schiffer, S. A perspective on enzyme catalysis. *Science* **301,** 1196–1202 (2003).
36. Eisenmesser, E. Z., Bosco, D. A., Akke, M. & Kern, D. Enzyme dynamics during catalysis. *Science* **295,** 1520–1523 (2002).
37. Frauenfelder, H., McMahon, B. H. & Fenimore, P. W. Myoglobin: the hydrogen atom of biology and a paradigm of complexity. *Proc. Natl Acad. Sci. USA* **100,** 8615–8617 (2003).
38. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).
39. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22,** 4673–4680 (1994).
40. Huang, X. *et al.* Structure of a WW domain containing fragment of dystrophin in complex with β-dystroglycan. *Nature Struct. Biol.* **7,** 634–638 (2000).
41. Kasanov, J., Pirozzi, G., Uveges, A. J. & Kay, B. K. Characterizing Class I WW domains defines key specificity determinants and generates mutant domains with novel specificities. *Chem. Biol.* **8,** 231–241 (2001).
42. John, D. M. & Weeks, K. M. van't Hoff enthalpies without baselines. *Protein Sci.* **9,** 1416–1419 (2000).
43. Sklenar, V., Piotto, M., Leppik, R. & Saudek, V. Gradient-tailored water suppression for $^{1}$H-$^{15}$N HSQC experiments optimized to retain full sensitivity. *J. Magn. Reson. A* **102,** 241–245 (1993).
44. Delaglio, F. *et al.* NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6,** 277–293 (1995).
45. Johnson, B. A. & Blevins, R. A. NMRView: a computer program for the visualization and analysis of NMR data. *J. Biomol. NMR* **4,** 603–614 (1994).
46. Brünger, A. T. *et al.* Crystallography and NMR system (CNS): a new software system for macromolecular structure determination. *Acta Crystallogr. D* **54,** 905–921 (1998).
47. Linge, J. P., O'Donoghue, S. I. & Nilges, M. *Methods in Enzymology* 71–90 (Academic Press, 2001).
48. Cornilescu, G., Delaglio, F. & Bax, A. Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J. Biomol. NMR* **13,** 289–302 (1999).
49. Laskowski, R. A., Rullman, J. A. C., MacArthur, M. W., Kaptein, R. & Thornton, J. M. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR* **8,** 477–486 (1996).
50. Koradi, R., Billeter, M. & Wüthrich, K. MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.* **14,** 51–55 (1996).
51. Delano, W. L. *The PyMOL Molecular Graphics System* ⟨http://www.pymol.org⟩ (2002).

**Supplementary Information** is linked to the online version of the paper at www.nature.com/nature.

**Acknowledgements** We thank A. G. Gilman, M. Rosen, and members of the Ranganathan laboratory for advice and critical review of the manuscript, and E. Olson for contributing to this project. R.R. and S.W.L. thank the students of the 2001 Woods Hole physiology course for participating in an initial phase of this work. This study was supported by the Robert A. Welch foundation (R.R. and K.H.G.), the Mallinckrodt Foundation Scholar Award (R.R.), and the NIH (K.H.G.). W.P.R. is an associate and R.R. is an investigator of the Howard Hughes Medical Institute.

**Author Information** Atomic coordinates for CC45 have been deposited in the Protein Data Bank under accession code 1YMZ. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.R. (rama.ranganathan@utsouthwestern.edu).

**MS# 2005-01-00888A (Socolich et al.), Supplementary Online Materials.**

## I. Supplementary Table 1: Distinctive proton chemical shifts in WW domains

| CC45 residue | atoms | BMRB mean±s.d. | Pin1 | Nedd4 | CC45 |
|---|---|---|---|---|---|
| Arg 18 | Hβ2, Hβ3 | 1.79±0.28 | **0.10 (-6.04σ),**<br>**0.10 (-6.04σ)** | **0.93 (-3.07σ),**<br>**0.60 (-4.25σ)** | 1.27 (-1.86σ),<br>**-0.06 (-6.61σ)** |
| | | | | | |
| Asn 29 | Hβ2, Hβ3 | 2.79±0.34 | 2.05 (-2.18σ),<br>**-0.64 (-10.1σ)** | 2.08 (-2.09σ),<br>**0.11 (-7.88σ)** | 1.99 (-2.35σ),<br>**-0.38 (-9.32σ)** |
| | Hδ21, Hδ22 | 7.32±0.51,<br>7.17±0.51 | 6.70 (-0.92σ),<br>**4.23 (-5.76σ)** | 7.21 (+0.08σ),<br>6.24 (-1.82σ) | 6.66 (-1.00σ),<br>**4.10 (-6.02σ)** |
| | | | | | |
| Arg 39 | Hα | 4.29±0.45 | **2.68 (-3.58σ)** | **2.89 (-3.11σ)** | **2.74 (-3.44σ)** |
| | | | | | |
| Pro 40 | Hβ2, Hβ3 | 2.05±0.37 | **0.62 (-3.86σ),**<br>**0.03 (-5.46σ)** | **0.88 (-3.16σ),**<br>**0.88 (-3.16σ)** | **0.94 (-3.00σ),**<br>**0.74 (-3.54σ)** |
| | Hγ2, Hγ3 | 1.92±0.35 | **0.87 (-3.00σ),**<br>**0.79 (-3.23σ)** | **0.55 (-3.91σ)**<br>**0.08 (-5.26σ)** | **0.46 (-4.17σ)**<br>**0.09 (-5.23σ)** |

Proton chemical shifts (in ppm) for each identified position were compared among the equivalent residues in the Pin1 (1), Nedd4 (2) and CC45 WW domains. Chemical shift assignments for Pin1 and Nedd4 were obtained from the BioMagResBank database (Pin1: BMRB entry 4882; Nedd4: entry 4963 for Nedd4/ENaC bP2 complex). Means and standard deviations of the proton chemical shift values were also obtained from BMRB, using the Dec. 16, 2004 version of the restricted set database containing ~1.2 million ppm entries. For methylene protons, no attempts are made to take stereospecific assignments into account and shifts are simply listed in numerical order. Shifts listed in bold are a minimum of three standard deviations away from mean values.

References:

1. Wintjens, R., Wieruszeski, J.-M., Drobecq, H., Rousselot-Pailley, P., Buee, L., Lippens, G. and Landrieu, I. (2001) $^1$H NMR study on the binding of Pin1 Trp-Trp domain with phosphothreonine peptides. *J. Biol. Chem.* **276**: 25150-25156.

2. Kanelis, V., Rotin, D., Forman-Kay, J.D. (2001) Solution structure of a Nedd4 WW domain–ENaC peptide complex. *Nat. Struct. Biol.* **8**: 407-412.

**Supplementary Table 2** NMR and refinement statistics for protein structures

| | **Protein** |
|---|---|
| **NMR distance and dihedral constraints** | |
| Distance constraints | |
| Total NOE | 756 |
| Intra-residue | 320 |
| Inter-residue | 436 |
| Sequential (\|i-j\| = 1) | 155 |
| Medium-range (\|i-j\| < 4) | 67 |
| Long-range (\|i-j ≥ 4) | 214 |
| Intermolecular | n/a |
| Hydrogen bonds | 12 |
| Total dihedral angle restraints (phi, psi and chi1) | 48 |
| | |
| **Structure statistics** | |
| Violations (mean and s.d.) | |
| Distance constraints (Å) | 0.012±0.003Å |
| Dihedral angle constraints (°) | 0.51±0.11 |
| Max. dihedral angle violation (°) | 3.7° |
| Max. distance constraint violation (Å) | 0.317Å |
| Deviations from idealized geometry | |
| Bond lengths (Å) | 0.0037±0.0002Å |
| Bond angles (°) | 0.48±0.03° |
| Impropers (°) | 1.28±0.12 |
| Average pairwise r.m.s.d.** (Å) | |
| Heavy | 1.75±0.31Å |
| Backbone | 0.86±0.20Å |

**** Pairwise r.m.s.d. was calculated among 10 refined structures.**

## II. Supplementary Methods

### Statistical coupling analysis

Given an alignment with $N$ positions, the total dataset produced by the SCA method is a $20 \times N$ matrix of conservation values and a $20 \times N \times M$ matrix of pairwise coupling values, where the impact of every perturbation $j \in \{1..M\}$ is measured at every MSA position $i \in \{1..N\}$ by a vector of amino acid specific statistical coupling energies. The datasets shown in Fig. 1b and Fig. 1c-e are $N$-dimensional vectors and $N \times M$ matrices, respectively, where for simplicity of presentation, we have taken the magnitudes along the amino acid dimension[14]. All calculations were carried out using MATLAB 7.0.1 (Mathworks Inc.). Alignments for natural and artificial WW sequences are available for download from our laboratory web site (http://www.hhmi.swmed.edu/Labs/rr), and code is available upon request.

### Statistical significance of coupling values

To calculate the statistical significance of statistical coupling values for a perturbation (Fig. S1f), we calculated the coupling scores for 100 independent trials of randomly selecting subalignments of the same size and mean change in conservation, and assessed whether observed coupling values are greater than from the random expectation (Z-test, significance threshold was $p < 0.05$).

### SCA-based protein design

The simulation initially introduces substantial variation in the MSA that scrambles the pattern of statistical coupling, and then searches for a new alignment solution that minimizes the difference SCA matrix between the artificial and natural alignments. The basic iterative step in this sequence space simulated annealing (SA) method is to choose two sequences (rows) in the MSA and one position (column) at random and swap the corresponding residues. By only swapping residues within a column in the MSA, the conservation pattern at sites is never changed, but couplings between sites may change. The swap is either accepted or not accepted based on the Metropolis criterion using a statistical energy function that describes the amino acid couplings during the progress of the simulation. For computational speed, the energy function describing couplings is slightly modified from the standard SCA procedure. Each element of the statistical energy matrix for the MSA at iteration $n$ (**E(n)**) is:

$$E_{i,j}^{x}(n) = \ln \frac{f_{i,j}^{x}(n)}{f_{i,j,nat}^{x}},$$

where $f_{i,j}^{x}(n)$ is the frequency of the $x^{th}$ amino acid at site $i$ given perturbation $j$ for the MSA at iteration $n$, and $f_{i,j,nat}^{x}$ is the same quantity for the natural MSA. Thus, **E(n)** is a quantitative measure of how different the alignment at the $n^{th}$ iteration is from the target (natural) MSA. We collapse the matrix **E(n)** by taking the magnitudes along each of its dimensions to finally produce $e(n)$, the scalar energy difference for the alignment at the $n^{th}$ iteration. In evaluating acceptance of the swap, we compare the energy of the alignment at the $n^{th}$ iteration with that of the $n-1^{th}$ iteration:

$$\Delta e = e(n) - e(n-1).$$

If $\Delta e \leq 0$, the swap is accepted, since the alignment draws closer to the natural MSA. If $\Delta e > 0$, the swap is accepted with a Boltzmann-weighted probability ($P_{accept} = e^{-\frac{\Delta e}{\beta}}$), where β is a statistical equivalent to temperature in a physical system and controls the likelihood of accepting swaps that diverge from the optimization target. In designing the CC sequences, β is initially set high to randomize the alignment, and then is exponentially cooled to convergence. The simulation exits if a preset number of swaps are not accepted upon three sequential attempts at 100-fold coverage of the alignment.

The alignment at exit comprises the CC sequences.  Simulations were carried out on a four-processor 500 MHz Dec Alpha cluster, and converged after 1.5 hours after roughly one billion swaps.

## III. Supplementary Figure Legends:

**Supplementary Figure 1:** Statistical coupling analysis for one site in the WW domain family. The SCA is based on two simple postulates about sequence conservation. First, the lack of evolutionary constraint at one site should cause the observed frequencies of amino acids in a large and diverse multiple sequence alignment (MSA) to approach their mean values found in all proteins. As a corollary, the evolutionary constraint (conservation) of any site $j$ is the degree to which the observed distribution deviates from these mean frequencies; in the SCA this is measured in an energy-like statistical parameter called $\Delta E_j^{stat}$ (in prior work, referred to as $\Delta G_j^{stat}$). The units of statistical energy are arbitrary and defined here symbolically as $\gamma *$ (in prior work, referred to as $kT*$). Second, the functional coupling of two sites $i$ and $j$, regardless of underlying mechanism or structural location, should drive their correlated evolution. In the SCA, this co-evolution of two sites is measured by introducing a change to the frequency of an amino acid at one site $i$ in the MSA and measuring the impact of this perturbation on amino acid $x$ at another site $j$. This impact is quantitatively measured by another statistical parameter, $\Delta\Delta E_{j,i}^{stat,x}$ (in prior work, referred to as $\Delta\Delta G_{j,i}^{stat,x}$). Calculated for all sites $j$, this so-called statistical perturbation experiment maps how the perturbation at $i$ is felt by all other sites in the protein, and is a global prediction of amino acid interactions for site $i$ derived from the evolutionary record of the protein family. **a**, A schematic representation of a multiple sequence alignment (MSA) of 120 WW domains, showing the frequencies of the dominant amino acids at four sites: 8 (62% Glu), 11 (an unconserved site with residues close to their mean frequencies in the non-redundant database), 16 (79% Gly), and 23 (64% His). Residue numbering is per alignment positions. Sequences with Glu at position 8 are indicated with red boxes. **b**, The subalignment resulting from a perturbation at site 8 (restricting it to Glu). Since the parent alignment showed 62% Glu, the 8E subalignment contains 0.62*120, or 74 sequences. **c-e**, The impact of the 8E perturbation on the three other sites (11, 16, and 23). The frequency distribution of amino acids in the full alignment is in black bars and that in the 8E subalignment is in white bars. For comparison, the mean frequencies of amino acids in the non-redundant database are shown (gray bars). **c**, Position 11 is unconserved since it shows amino acid frequencies close to their mean values in all proteins (compare black and gray bars, $\Delta E_{11}^{stat} = 0.2\gamma *$). Unconserved sites by definition show little evolutionary constraint, and thus are expected to remain unconserved upon perturbation as long as the subalignment produced upon perturbation remains sufficiently large and diverse. Indeed, position 11 shows nearly the same frequency distribution upon making the 8E perturbation (compare white and gray bars), and shows a weak coupling score in the SCA ($\Delta\Delta E_{11,8E}^{stat} = 0.1\gamma *$). This coupling score is insignificant because it does not exceed scores for position 11 derived from randomly selected subalignments of the MSA with the same average conservation changes as the 8E subalignment (p=0.99). **d**, Positions 16 is a moderately conserved site since its amino acid distribution differs strongly from the mean in all proteins ($\Delta E_{16}^{stat} = 1.63\gamma *$), but interestingly, the distribution remains the same in the 8E subalignment as in the parent alignment. Thus position 16 is uncoupled to the 8E perturbation ($\Delta\Delta E_{16,8E}^{stat} = 0.082\gamma *$, p=0.89). This result indicates that conservation *per se* is not tantamount to coupling; sites can be evolutionarily constrained for independent reasons. **e**, However, some conserved positions are coupled to the 8E perturbation. Position 23, a residue in direct packing contact with position 8, shows an amino acid distribution that is not only conserved ($\Delta E_{23}^{stat} = 1.8\gamma *$), but that is strongly influenced by the 8E perturbation ($\Delta\Delta E_{23,8E}^{stat} = 1.5\gamma *$, p<<.001). Thus, this site is evolutionarily coupled to position 8. **f**, The complete set of statistical coupling values for the 8E perturbation for all other sites $i$ ($\Delta\Delta E_{i,8E}^{stat}$) in the MSA. Statistically significant couplings (p<0.05) are marked with an asterisk.

**Supplementary Figure 2:** Summary of all sequences constructed and experimentally tested. Shown are alignments of natural WW domains (**A**), CC (**B**), IC (**C**), and random sequences (**D**). For each sequence, the table at right gives the mean sequence identity to the MSA of natural WW domains (mean_id), the maximum ("top-hit") identity (max_id), assessments of protein expression (exp), solubility (sol), and thermal denaturation (melt) experiments, and if applicable, thermodynamic parameters (the melting temperature $T_m$, and the vant Hoff enthalpy of unfolding $\Delta H_u^{VH}$ ). Expression and solubility were assessed by SDS-PAGE and thermodynamic parameters were determined by fitting the denaturation data to a two-state model for folding.

**Supplementary Figure 3:** Assessment of expression and solubility of all sequences constructed. SDS-PAGE analysis of natural WW domains (**A**), CC sequences (**B**), IC sequences (**C**), and random sequences (**D**). Lane order is P – pellet or insoluble fraction, S – supernatant after lysis, or soluble fraction, F – flow through after incubation with Ni-NTA affinity matrix, and E – eluted product after washing the Ni-NTA beads.

**Supplementary Figure 4:** Thermal denaturation and renaturation studies. Graphs plot the fluorescence at 340nm against temperature in degrees Celsius for all soluble natural WW domains (**A**), CC (**B**), IC (**C**), and random (**D**) sequences upon denaturation ($4^o C \rightarrow 90^o C$, black line) and renaturation ($90^o C \rightarrow 4^o C$, gray line). For cases in which proteins were only partially reversible, the inset shows normalized fluorescence curves to aid in assessing the similarity of the unfolding and refolding processes.

**Supplementary Figure 5:** One-dimensional proton NMR spectra for natural WW domains (**A**), CC sequences (**B**), and IC sequences (**C**). For this analysis, natural sequences were selected to represent a range of stabilities as measured by thermal denaturation experiments in order to provide a standard for assessing artificial sequences. All CC sequences that showed some potential for native folding were studied by $^1$H-NMR with the exception of CC14, which has a thrombin cleavage site within loop 1 of the fold and was therefore difficult to prepare these studies. Though no IC sequences showed convincing evidence in thermal denaturation experiments of producing the native state, a number of these sequences with non-zero signals were studied to definitively rule out folding to a native state.