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),** **0.10 (-6.04σ)**	**0.93 (-3.07σ),** **0.60 (-4.25σ)**	1.27 (-1.86σ), **-0.06 (-6.61σ)**
Asn 29	Hβ2, Hβ3	2.79±0.34	2.05 (-2.18σ), **-0.64 (-10.1σ)**	2.08 (-2.09σ), **0.11 (-7.88σ)**	1.99 (-2.35σ), **-0.38 (-9.32σ)**
	Hδ21, Hδ22	7.32±0.51, 7.17±0.51	6.70 (-0.92σ), **4.23 (-5.76σ)**	7.21 (+0.08σ), 6.24 (-1.82σ)	6.66 (-1.00σ), **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),** **0.03 (-5.46σ)**	**0.88 (-3.16σ),** **0.88 (-3.16σ)**	**0.94 (-3.00σ),** **0.74 (-3.54σ)**
	Hγ2, Hγ3	1.92±0.35	**0.87 (-3.00σ),** **0.79 (-3.23σ)**	**0.55 (-3.91σ)** **0.08 (-5.26σ)**	**0.46 (-4.17σ)** **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 ΔE_j^{stat} (in prior work, referred to as ΔG_j^{stat}). The units of statistical energy are arbitrary and defined here symbolically as γ^* (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 ΔH_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 ($4^oC \rightarrow 90^oC$, black line) and renaturation ($90^oC \rightarrow 4^oC$, 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.