Evolutionary Information for Specifying a Protein Fold

Michael Socolich, Steve W. Lockless, William P. Russ, Heather Lee, Kevin H. Gardner & Rama Ranganathan

Louis Yuk

Fundamental Tenet of Biochemistry

Amino Acid sequence of a protein specifies its atomic structure and biochemical function

Goals

- determine sequence rules for specifying a protein fold by computationally creating artificial sequences
- use only statistical information encoded in a Multiple Sequence Alignment (MSA) and NO tertiary structure information

Statistical Coupling Analysis (SCA)

- Based on the assumption that regardless of spatial location or underlying mechanism, the conserved functional coupling of sites in a protein should drive their mutual coevolution

Reveals Two Conclusion

1. **Global pattern of coevolutionary interactions is sparse, so that a small set of positions mutually coevolves among a majority that are largely decoupled**

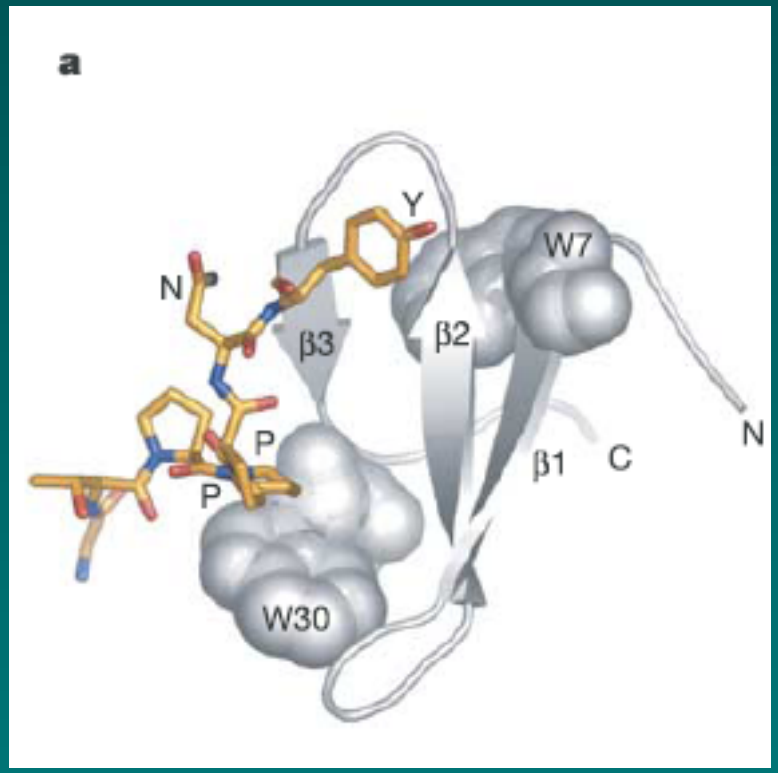
2. **Strongly coevolving residues are spatially organized into physically connected networks linking distant functional sites in the structure through packing interactions**

Testing Hypothesis

If (and only if) the information contained in the statistical coupling analysis (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.

SCA-based protein design

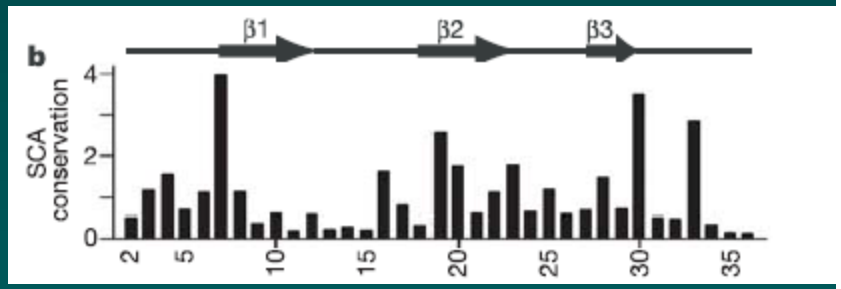
- Carried out the SCA alignment of 120 members of the WW domain family



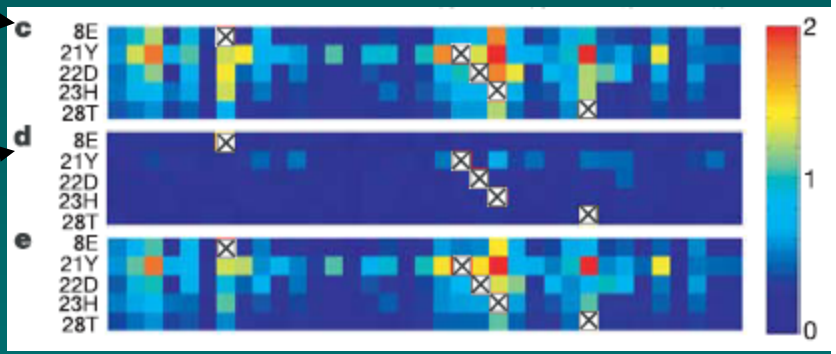
Small, independently folding protein interaction module
• adopt a curved three stranded β-sheet configuration
•bind to proline-containing target sequences

SCA-based protein design

- SCA Conservation scores for each position in the WW domain family alignment (arbitrary units) measures deviation of distribution of A.A from mean values found in all proteins



- A matrix representation of statistical coupling values from perturbation analysis of 5 positions in WW domain MSA.



- **IC (site independent conservation)** sequences
  - built by randomly selecting a.a's at each site from the observed frequency distributions in the natural alignment.
- **CC (coupled conservation)** sequences
  - derived from a design algorithm where both the conservation pattern and the pattern of statistical coupling in the natural alignment are preserved.

Perturbation experiments expose the same redundant pattern of coevolution between moderately conserved positions

Statistical properties of natural and artificial WW domains

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*	35.3 ± 6.1	80.4 ± 14.5	42
(folded)	(36.2 ± 5.8)	(78.2 ± 15.9)	(28)
IC†	36.1 ± 3.0	55.7 ± 5.6	43
CC‡	35.0 ± 4.8	58.6 ± 7.2	43
(folded)	(37.4 ± 5.4)	(63.1 ± 6.0)	(12)
Random§	6.3 ± 2.5	14.9 ± 4.3	19

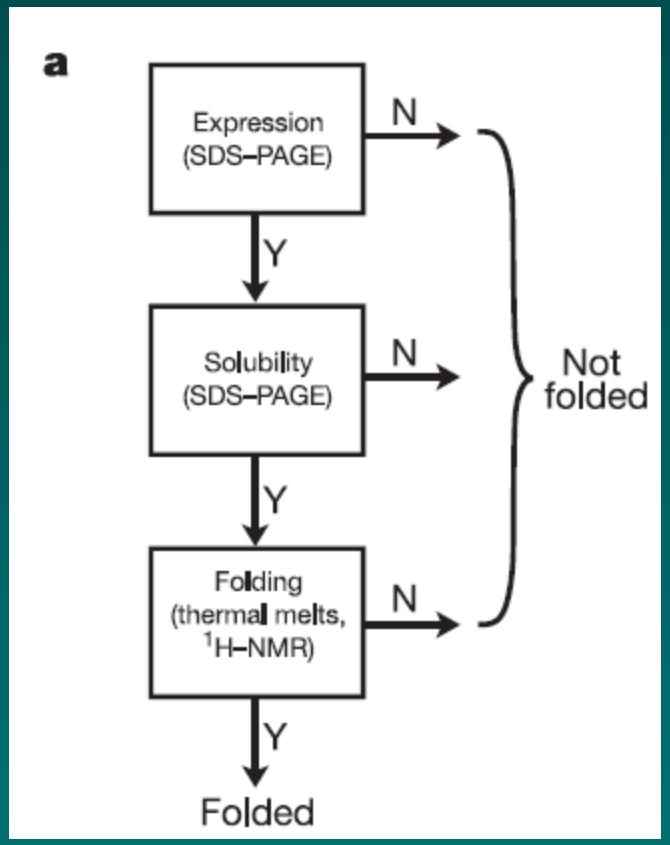
*Drawn randomly from the natural WW MSA.
†Created by randomly selecting amino acids at each site from the observed frequency distribution at that site in the natural WW MSA.
‡Created by Monte-Carlo simulation.
§Created by randomly selecting amino acids at each site from their overall mean frequencies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

Random sequences show much weaker identities to natural WW domains (~6%)

IC/CC show similar top hit identities

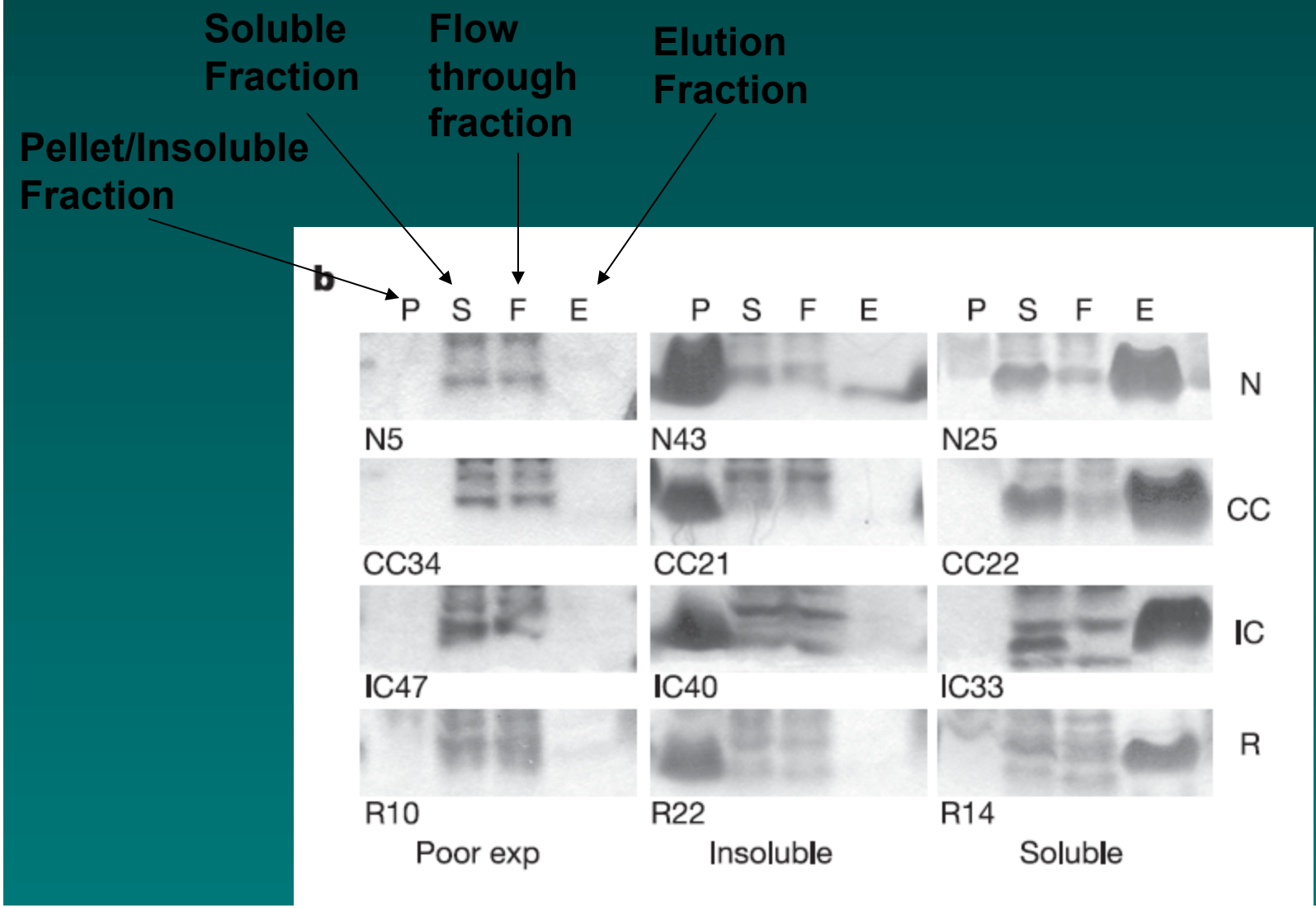
- **Despite additional constraints in the design, CC sequences are about as diverged as IC sequences**

Flow chart of experiments for evaluating WW sequences



- **Proteins expressed as His tagged fusions purified using Ni+-NTA affinity Chromatography and subjected to SDS page to evaluate**
- **Expression**
- **Solubility**

Experimental evaluation of artificial proteins



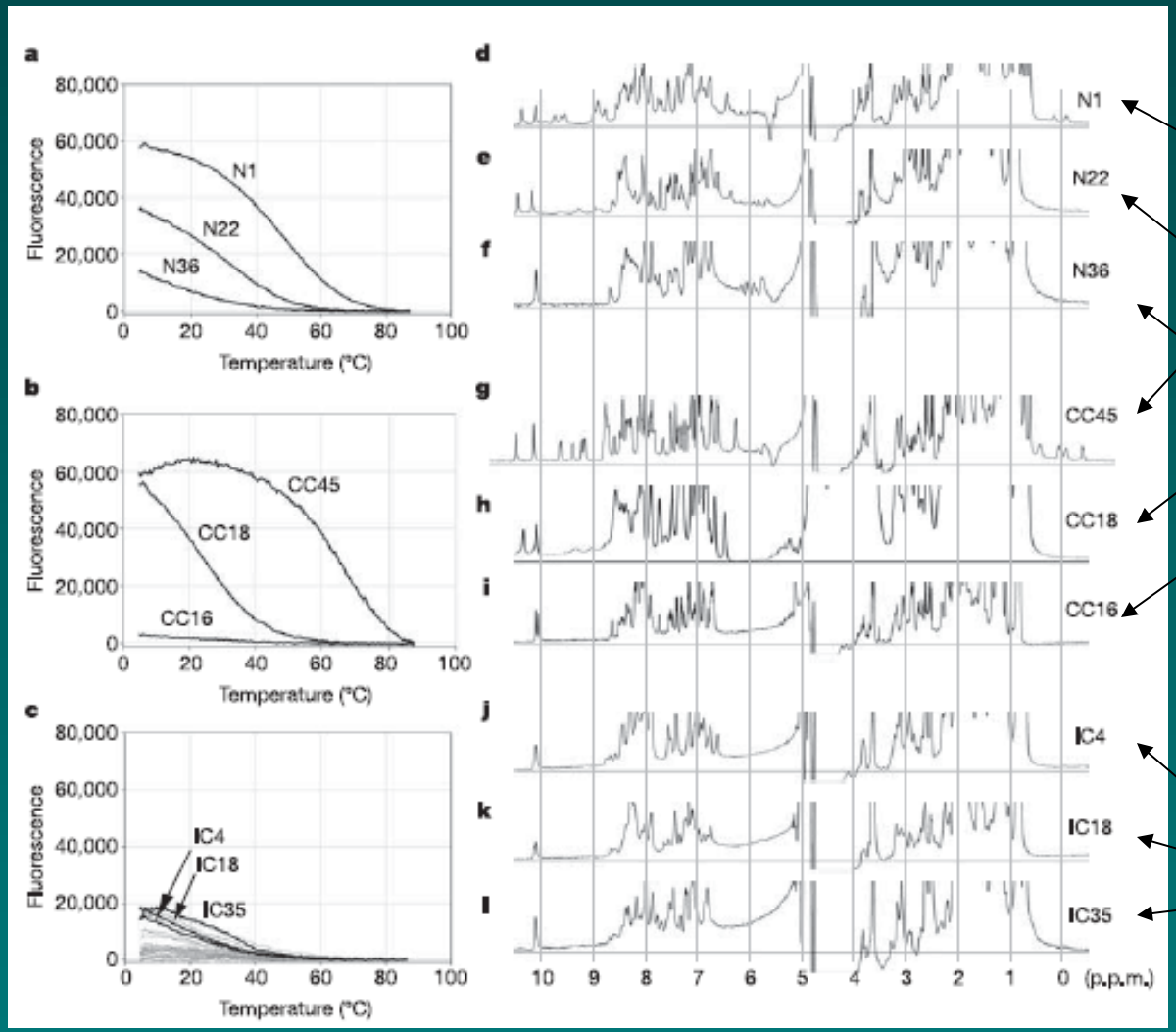
Thermal denaturation studies + NMR Spectra

A hallmark of natively folded small proteins is:

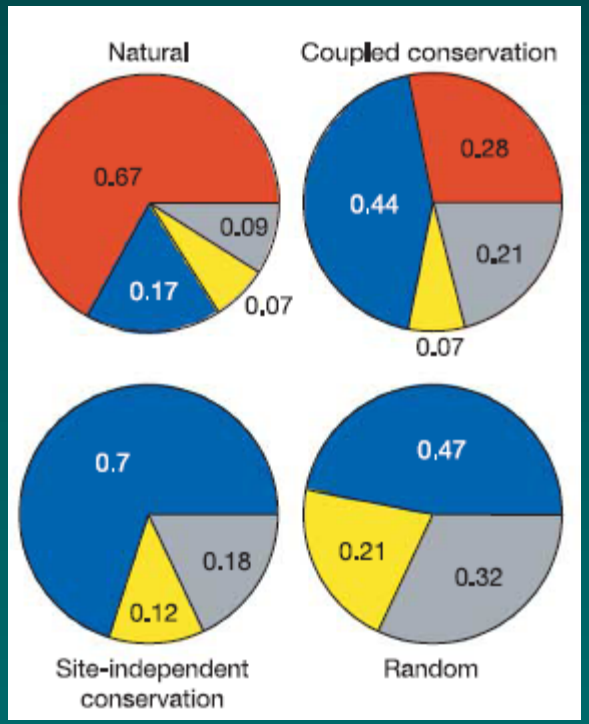
- **Cooperative and reversible transition between folded and unfolded states**
- **Followed the folding reaction by monitoring the fluoresence of a buried tryptophan**
  - **Tryptophan becomes quenched due to solvent exposure upon thermal denaturation and reports the fraction of protein folded as a function of temperature**
- **Tested all 105 well expressed + soluble proteins from their libraries for cooperative and reversible transitions**

Folded WW Proteins have good chemical shift dispersion of peaks corresponding to backbone amide protons, often accompanied by distinct chemical shifts of two indole N – protons downfield of 10 p.p.m

Thermal denaturation studies + NMR Spectra



Summary of experimental analysis of all 147 WW sequences tested



Key

Red-Natively Folded

Blue-soluble but unfolded

Yellow-insoluble

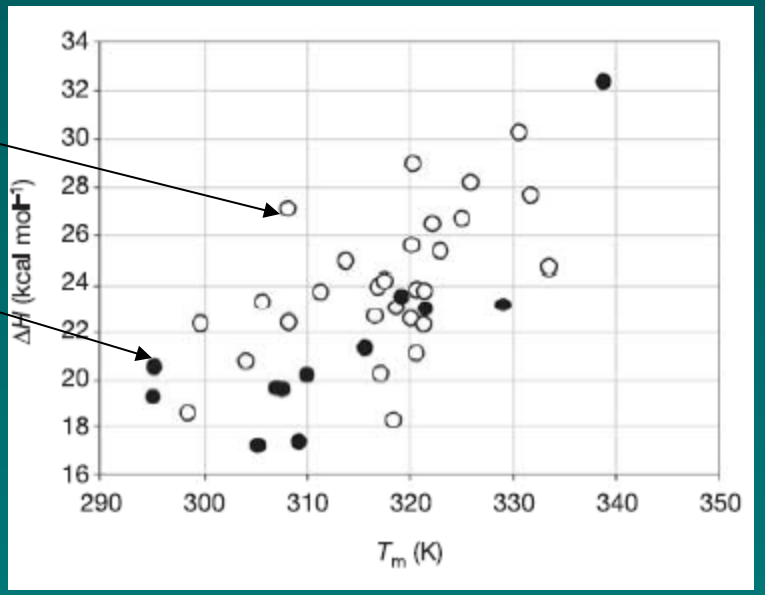
Grey-Poor expressing

CC Sequences fall into the same range of thermodynamic parameters as their natural counterparts!

Unfolding Enthalpy Vs. Melting Temperatures

Natural Sequences

CC Sequences



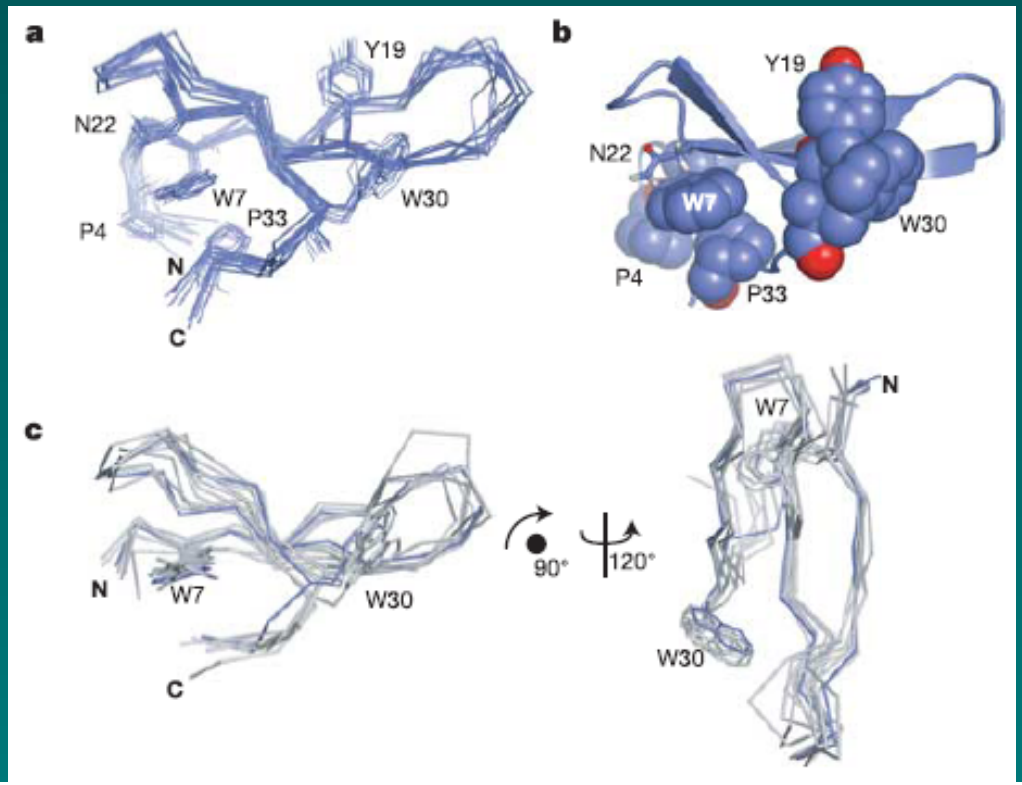
NMR structure of CC45

Do artificial proteins adopt the canonical WW fold?

- **39% mean identity**
- **61% top hit identity to natural WW sequences**
- **Well folded**

- **Confirmed the curved 3 stranded anti-parallel β sheets**

- **Tertiary structure motifs also found that were the same**
  - **Centrally located trp (W7) upon platform of 2 proline side chains**
  - **Contains several sites that display unusual proton chemical shifts that two natural WW domains (Pin1 and Nedd4.3) also contained**



Summary

- Amino acid interactions specifying the atomic structure are conserved throughout members of a protein family rather than site-idiosyncratic
- Conservation is a distributed rather than site-independent property because it fundamentally arises from cooperativity of energetic interactions
- The SCA method estimates the analysis of parsing of conservation pretty well

It is the specific distribution of conservation rather than the quantity of conservation that dictates native folding