rescue-of-function screens, and more recent procedures for investigating genome-wide genetic interactions (Tong et al., 2004).

The lack of reproducibility is the skeleton in the closet for all high-throughput interaction studies. Indeed it is fair to say that all of the high-throughput methods for measuring protein interactions suffer from significant false-positive and false-negative scoring (von Mering et al., 2002). In fact, there is surprisingly little overlap in the data generated by different detection methods, suggesting that they are nonsaturating, erroneous, or both. For these reasons, large-scale interaction studies are frequently criticized. Another concern with two-hybrid studies is whether the interactions detected are biologically relevant given that they are assessed in yeast nuclei, a nonphysiological milieu for cytoplasmic, membrane, or nonendogenous proteins. This may account for the susceptibility of the two-hybrid assay to false positives. Often, for unknown reasons, the assay exhibits a considerable rate of false-positive detection, perhaps because ectopic expression may lead to fortuitous binding or because of the natural randomness associated with mRNA expression (Raser and O'Shea, 2004). Missed interactions (false negatives) are another concern. Moving past these apprehensions, a goal for the future is to assemble the information from these interactome studies into dynamic models of cellular processes. As George Bernard Shaw wrote, "If you cannot get rid of the family skeleton, you may as well teach it to dance."

Even though interactome studies would benefit from efforts to improve or accelerate data validation, they provide a valuable, previously unseen, view of a major defining feature of cell biology—the protein interaction network—from a global, systems-wide vantage. And bioinformaticians have devised ingenious ways to deal with the limitations of the core assays, principally by combining datasets, but also through analysis of the network properties of the interaction maps and projections of interactions across species. Intriguing, albeit largely theoretical, observations have been made by examining interaction datasets at different levels of abstraction (e.g., Kelley and Ideker, 2005). Yet many unanswered questions remain. How dynamic or hard-wired are protein interaction networks? What is the relationship of networks to cell phenotype or physiology? How plastic are interaction networks across evolution? Conversely, to what extent does the evolution of protein networks drive speciation? How does the modular organization of the protein networks in a cell contribute to its overall interactome? And in light of the current study, what features of the human interactome are unique?

In addition to representing a potentially rich source of newly discovered interactions, the Stelzl et al. dataset provides an intriguing glimpse of the far larger skeleton of human protein interactions that is certain to exist. This study and other imminent reports can help to reveal aspects of human biology that have been hidden from traditional approaches. Based on what we learn from these new perspectives, we may need to revisit the issue of what is the appropriate unit for studying human biology. Ultimately, it may not be the level of protein complexes or pathways, or even phenotypes. Instead, the full assembly of these interactions, both genetic and physical, could produce a breakthrough in understanding what it is to be human.

Ata Ghavidel,[1] Gerard Cagney,[2] and Andrew Emili[1]
[1] Banting and Best Department of Medical Research
University of Toronto
Toronto, Ontario
Canada, M5G 1L6
[2] Conway Institute
University College Dublin
Ireland

Selected Reading

Barrios-Rodiles, M., Brown, K.R., Ozdamar, B., Bose, R., Liu, Z., Donovan, R.S., Shinjo, F., Liu, Y., Dembowy, J., Taylor, I.W., et al. (2005). Science *307*, 1621–1625.

Butland, G., Peregrin-Alvarez, J.M., Li, J., Yang, W., Yang, X., Canadien, V., Starostine, A., Richards, D., Beattie, B., Krogan, N., et al. (2005). Nature *433*, 531–537.

Colland, F., Jacq, X., Trouplin, V., Mougin, C., Groizeleau, C., Hamburger, A., Meil, A., Wojcik, J., Legrain, P., and Gauthier, J.M. (2004). Genome Res. *14*, 1324–1332.

de Lichtenberg, U., Jensen, L.J., Brunak, S., and Bork, P. (2005). Science *307*, 724–727.

Giot, L., Bader, J.S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y.L., Ooi, C.E., Godwin, B., Vitols, E., et al. (2003). Science *302*, 1727–1736.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). Proc. Natl. Acad. Sci. USA *98*, 4569–4574.

Kelley, R., and Ideker, T. (2005). Nat. Biotechnol. *23*, 561–566.

Li, S., Armstrong, C.M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P.O., Han, J.D., Chesneau, A., Hao, T., et al. (2004). Science *303*, 540–543.

Raser, J.M., and O'Shea, E.K. (2004). Science *304*, 1811–1814.

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F.H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., and Koeppen, S. (2005). Cell *122*, this issue, 957–968.

Tong, A.H., Lesage, G., Bader, G.D., Ding, H., Xu, H., Xin, X., Young, J., Berriz, G.F., Brost, R.L., Chang, M., et al. (2004). Science *303*, 808–813.

Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., et al. (2000). Nature *403*, 623–627.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S., and Bork, P. (2002). Nature *417*, 399–403.

DOI 10.1016/j.cell.2005.09.006

Finding the Fittest Fold: Using the Evolutionary Record to Design New Proteins

For many years, the holy grail of protein engineering has been the design of artificial amino acid sequences that fold into stable proteins with desired functions. In the current issue of *Nature*, two papers from the Ranganathan group (Russ et al., 2005; Socolich et al., 2005) report remarkable success in the design of artificial WW domains. Their method, termed statistical coupling analysis (Lockless and Ranganathan, 1999), does not use structural or physicochemi-

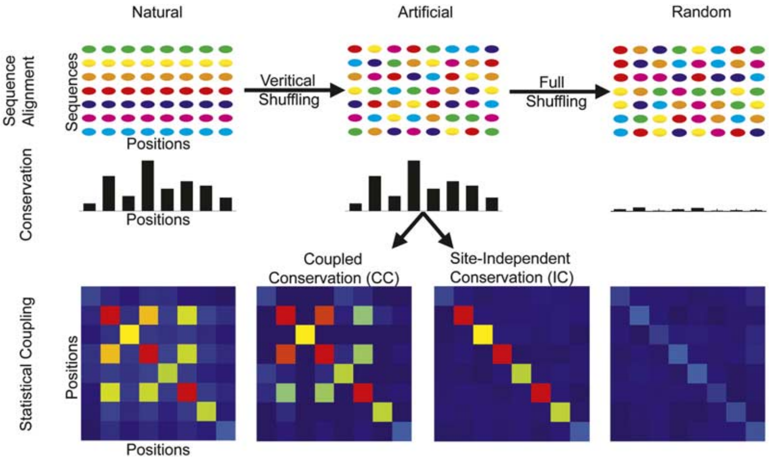

Figure 1. The Amino Acid Composition and Evolutionary Relationships of Natural WW Domain Sequences Guide the Creation of Artificial and Control Sequences

(Top and middle panels) Amino acids are swapped along each position of an alignment of natural WW domain sequences to create an alignment of artificial WW domain sequences. Vertical shuffling preserves conservation at each position, while the statistical coupling between positions can either remain intact (creating artificial sequences termed coupled conservation, or CC) or deviate from the natural alignment (artificial sequences termed site-independent conservation, or IC). In contrast, full shuffling also mixes amino acids between positions, thus creating sequences of relatively random composition (with the same amino acid frequencies as the natural sequence set), and destroys both conservation and coupling. (Bottom panel) Statistical coupling matrices provide color-coding for the coupling between positions (red = high; blue = low). A position's self-coupling reduces to a measure of conservation and is shown along the diagonal of each matrix. The pattern of off-diagonal intensity defines the coupling between the positions of a given sequence alignment.

cal information but instead extracts information about essential patterns of amino acids from the evolutionary record.

Amino acid conservation has long been valued as a primary indicator of the importance of individual residues in the structure and function of a protein. Yet, conservation alone does not describe the cooperative context of an amino acid with other residues in the protein. In previous work, the comparative analysis of correlated mutations in a protein's evolutionary history has successfully identified and predicted functionally important residues (Larson et al., 2000; Lichtarge et al., 1996; Lockless and Ranganathan, 1999; Neher, 1994). In general, these findings have been tested by single or double mutations of a naturally occurring protein sequence.

The two new *Nature* papers from the Ranganathan group (Russ et al., 2005; Socolich et al., 2005) use statistical coupling analysis (SCA), which takes the cooperative nature of amino acid interactions into account (Lockless and Ranganathan, 1999), and expand its scope. They do this by exploiting the predictive potential of SCA to make extensive mutations simultaneously, and they then test the predictions experimentally. They analyzed the WW domain, a small, highly conserved protein fold consisting of a three strand β sheet, which binds to proline-rich peptide motifs. Instead of measuring the effects of small changes in a natural sequence, Socolich et al. (2005) created a large set of extensively mutated WW domains by shuffling amino acids between 120 naturally occurring WW sequences (see Figure 1). Shuffling amino acids "vertically" along their respective alignment positions maintains the conservational distributions of the natural alignment but can disrupt the positional interdependencies between amino acids. Therefore, the vertically shuffled sequences were parsed into two groups based on the preservation of the statistical coupling relationships present in the natural alignment. One set, termed the coupled conservation (CC) variants, preserves the coupling profile seen in the natural alignment, whereas the other set, termed site-independent conservation (IC) variants, deviates from natural coupling. Finally, a set of randomized sequences, created by shuffling amino acids both vertically and horizontally through the natural sequence alignment, destroys both the natural conservation and coupling profiles. The artificial sequences from all three groups showed significant primary sequence differences from their most closely related natural sequence.

Strikingly, Socolich et al. (2005) created artificial WW domains that fold to a native state using only the evolutionary rules deduced from the coupling data of natural WW domains. The statistical coupling information was validated by expression and characterization of 147 proteins randomly extracted from the natural, CC, IC, and random-sequence groups. A multitiered test of protein expression, solubility, ^{1}H NMR spectra, tryptophan burial, and thermal denaturation was used to determine if the pool of artificial WW domains mimics the range of properties seen in natural domains. Both the natural and CC sequence groups displayed a high propensity to fold to the native state in the bacterial expression system used (67% of natural domains and 28% of CC domains), and the NMR structure of a selected CC domain was indistinguishable from its natural counterparts. Perhaps equally impressive, the CC sequences had stabilities in the same range as the natural sequences. This has not been the case for protein design algorithms that use physicochemical relationships to promote desired native contacts between residues and tend to result in proteins with extreme, nonphysiological stability (e.g., Kuhlman et al., 2003).

In contrast to the natural and CC sequence groups, IC and random sequences entirely failed to fold to a native state. Interestingly, the CC and IC sequences with the same positional conservation as natural sequences showed the same solubility as the natural sequences (72% and 70% respectively for CC and IC, compared to 84% for natural and 47% for random sequences), implying that solubility and the capacity to form a native fold are not directly linked. The authors speculate that positional conservation may be sufficient for hydrophobic collapse to a molten globule, which may be soluble. The CC and IC sequences differ only by the existence of natural coupling relationships, which indicates that the statistical coupling analysis appears to serve as a necessary and sufficient criterion for protein design. It is of interest to examine the core hydrophobic residues in the CC and IC groups in light of previous proposals that emphasize the sufficiency of the hydrophobic core in specifying the low-resolution structure of a protein (Cordes et al., 1996). Vertical shuffling of a natural sequence alignment does not create variation at fully conserved positions, and five core residues in CC and IC sequences (including the folded subset) showed a high percentage identity to their most closely related natural sequences. The fact that this occurred in both CC and IC sequences argues that conservation of core residues is not sufficient to determine the native fold and instead that the coupling interactions between residues are essential determinants of structure.

In the companion *Nature* paper, Russ et al. (2005) extended the design study by comparing natural and artificial WW domain function. First, they screened peptide libraries that were based on the different classes of WW domains for functional binding to the artificial WW sequences. They then measured binding affinities of the peptides to the artificial sequences. These approaches, combined with ligand saturation mutagenesis to test the interaction between the artificial domains and the peptides, have shown that the artificial WW domains are not only functional but can be divided into different classes according to their specificity very much like the natural domains. With this validation, the authors explored the amino acid determinants of binding specificity in WW domains, using relationships between the SCA, functional classification of WW domains, and mutational data. This analysis indicated that a distributed, cooperative network of residues is involved in substrate binding, even on the opposite face of the binding pocket. Therefore, artificial sequences that preserve both conservation and coupling showed a high propensity to fold to the native state with physiological stability and functional binding.

One of the most striking results of these two studies is the sparseness of the highly coupled interdependencies in the SCA that are sufficient to specify the WW fold and to confer specific binding functions. Although these two requirements are critical, they are not the only ones that drive sequence evolution. What other pressures are there on a protein sequence? What does the "fitness" of a sequence actually select for? For example, the prevalence of stabilizing native contacts over competing nonnative contacts may be a consequence of natural selection to ensure successful folding (Onuchic and Wolynes, 2004). In these studies, the artificial WW domains have not been subjected to some demands that would have been imposed on their naturally selected counterparts, including synthesis in the native cellular context, correct cellular localization, the ability to interact with partner domains at optimal affinities, and an analysis of whether they turn over at a physiological rate. Other constraints on naturally selected proteins are imposed at the level of the nucleic acid predecessors of the protein sequence. By mining the evolutionary record, SCA results should provide insight into these additional layers of selective history imposed on a set of natural sequences.

From a protein-engineering standpoint, the SCA approach has great promise. It may expand the range of functional sequence space beyond natural sequences and beyond current design approaches by describing permissive and nonpermissive mutations more completely. SCA results could be used in a complementary strategy to balance physicochemical calculations in computational models that already show significant success (Kuhlman and Baker, 2004). Additionally, the SCA results could be combined with other computational and experimental techniques to show regions of proteins that would be tolerant to modification in order to design new functionalities (Voigt et al., 2002) and to further elucidate the relationship between sequence, structure, stability, and function (Magliery and Regan, 2004).

Robert G. Smock and Lila M. Gierasch
Department of Biochemistry & Molecular Biology
University of Massachusetts Amherst
710 North Pleasant Street
Amherst, Massachusetts 01003

Selected Reading

Cordes, M., Davidson, A., and Sauer, R. (1996). Curr. Opin. Struct. Biol. *6*, 3–10.

Kuhlman, B., and Baker, D. (2004). Curr. Opin. Struct. Biol. *14*, 89–95.

Kuhlman, B., Dantas, G., Ireton, G.C., Varani, G., Stoddard, B.L., and Baker, D. (2003). Science *302*, 1364–1368.

Larson, S.M., Di Nardo, A.A., and Davidson, A.R. (2000). J. Mol. Biol. *303*, 433–446.

Lichtarge, O., Bourne, H.R., and Cohen, F.E. (1996). J. Mol. Biol. *257*, 342–358.

Lockless, S.W., and Ranganathan, R. (1999). Science *286*, 295–299.

Magliery, T.J., and Regan, L. (2004). Eur. J. Biochem. *271*, 1595–1608.

Neher, E. (1994). Proc. Natl. Acad. Sci. USA *91*, 98–102.

Onuchic, J.N., and Wolynes, P.G. (2004). Curr. Opin. Struct. Biol. *14*, 70–75.

Russ, W., Lowery, D., Mishra, P., Yaffe, M., and Ranganathan, R. (2005). Nature, in press. Published online September 21, 2005. 10.1038/nature03990.

Socolich, M., Lockless, S., Russ, W., Lee, H., Gardner, K., and Ranganathan, R. (2005). Nature, in press. Published online September 21, 2005. 10.1038/nature03991.

Voigt, C.A., Martinez, C., Wang, Z.-G., Mayo, S.L., and Arnold, F.H. (2002). Nat. Struct. Biol. *9*, 553–558.

DOI 10.1016/j.cell.2005.09.005