19. The following surface areas buried between domains in the Inv497, Fn-III 7–10, Fn-III 12–14, and VCAM-1 structures (*14*) were calculated with XPLOR (*12*) with a 1.4 Å probe radius: Inv497 D1–D2, 411 Å^2; Inv497 D2–D3, 454 Å^2; Inv497 D3–D4, 564 Å^2; Inv497 D4–D5, 1925 Å^2; Fn-III 7–8, 608 Å^2; Fn-III 8–9, 481 Å^2; Fn-III 9–10, 342 Å^2; Fn-III 12–13, 450 Å^2; Fn-III 13–14, 696 Å^2; and VCAM-1 D1–D2 (molecule B), 696 Å^2.
20. J. M. Leong, P. E. Morrissey, R. R. Isberg, *J. Biol. Chem.* **268**, 20524 (1993); L. H. Saltman, Y. Lu, E. M. Zaharias, R. R. Isberg, *ibid.* **271**, 23438 (1996).
21. R. D. Bowditch *et al.*, *ibid.* **269**, 10856 (1994); S.-I. Aota, M. Nomizu, K. M. Yamada, *ibid.*, p. 24756; T. P. Ugarova *et al.*, *Biochemistry* **34**, 4457 (1995).
22. T. A. Jones and M. Kjeldgaard, *Methods Enzymol.* **277**, 173 (1997).
23. J. M. Casasnovas, T. Stehle, J. Liu, J. Wang, T. A. Springer, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 4134 (1998).
24. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; D, Asp; G, Gly; L, Leu; N, Asn; P, Pro; R, Arg; T, Thr; and W, Trp.
25. R. R. Isberg, Y. Yang, D. L. Voorhis, *J. Biol. Chem.* **268**, 15840 (1993).
26. We thank S. M. Soltis and the staff at the Stanford Synchrotron Radiation Laboratory (SSRL) for help with xenon derivatization and data collection; M. J. Bennett, A. J. Chirino, L. M. Sánchez, D. E. Vaughn, and A. P. Yeh for discussions and help with crystallographic software; S. Matthews for intimin coordinates; P. D. Sun for CD94 coordinates; W. I. Weis for helpful discussions about C-type lectin structures; and W. I Weis, J. M. Leong, and members of the Bjorkman lab for critical reading of the manuscript. Inv497 coordinates have been deposited in the PDB (PDB code 1CWV).

16 June 1999; accepted 1 September 1999

Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families

Steve W. Lockless and Rama Ranganathan*

For mapping energetic interactions in proteins, a technique was developed that uses evolutionary data for a protein family to measure statistical interactions between amino acid positions. For the PDZ domain family, this analysis predicted a set of energetically coupled positions for a binding site residue that includes unexpected long-range interactions. Mutational studies confirm these predictions, demonstrating that the statistical energy function is a good indicator of thermodynamic coupling in proteins. Sets of interacting residues form connected pathways through the protein fold that may be the basis for efficient energy conduction within proteins.

Many cellular processes depend on the sequential establishment of protein-protein interactions that underlies the propagation of information through a signaling system. The interaction of one protein with another can be thought of as an energetic perturbation to each binding surface that distributes through the three-dimensional structure to cause specific changes in protein function (*1*). The structural basis for this process is largely unknown, but large-scale mutagenesis has begun to define some basic principles of energy parsing in proteins. Studies of the interaction of human growth hormone with its receptor show that binding energy is not smoothly distributed over the interaction surface; instead, a few residues comprising only a small fraction of the interaction surface account for most of the free-energy change (*2*). Similarly, high-affinity interaction of K^+ channel pores with peptide scorpion toxins buries ~15 residues on the toxin molecule, but most of the binding energy depends on only two amino acid positions (*3, 4*). Thus, protein interaction surfaces contain functional epitopes or hot spots of binding energy that are generally not predictable from the atomic structure.

In addition, a large body of evidence suggests that the change in free energy at a protein interaction surface propagates through the tertiary structure in a seemingly arbitrary manner. Studies addressing mechanisms of substrate specificity in serine proteases show that many sites distantly positioned from the active site contribute to a determination of the energetics of catalytic residues (*5*). The conversion of trypsin to chymotrypsin specificity required a large set of simultaneous mutations, many at unexpected positions. Similarly, mutations introduced during maturation of antibody specificity have been shown to occur at sites that are distant in tertiary structure from the antigen-binding site, despite substantial increases in binding energy (*6*).

An important step in understanding the problem of energy distribution in proteins is the full-scale mapping of energetic coupling between amino acid positions. Thermodynamic mutant cycle analysis (*3, 7*), a technique that measures the energetic interaction of two mutations, provides a direct method to systematically probe such relations of protein sites. However, practical considerations limit this technique to small-scale studies, precluding a full mapping of all energetic interactions on a complete protein. We report a study that uses evolutionary data for a protein family to measure energetic coupling between positions on a multiple sequence alignment (MSA).

Evolution of a protein fold is the result of large-scale random mutagenesis, with selection constraints imposed by function. The theory described below is based on two hypotheses that derive from the empirical observation of sequence evolution. The lack of evolutionary constraint at one position should cause the distribution of observed amino acids at that position in the MSA to approach their mean abundance in all proteins, and deviances from the mean values should quantitatively represent conservation. In addition, the functional coupling of two positions, even if distantly positioned in the structure, should mutually constrain evolution at the two positions, and these should be represented in the statistical coupling of the underlying amino acid distributions (*8*).

Two definitions guide the development of statistical parameters used in our analysis: (i) Conservation at a given site in a MSA is defined as the overall deviance of amino acid frequencies at that site from their mean values, and (ii) statistical coupling of two sites, i and j, is defined as the degree to which amino acid frequencies at site i change in response to a perturbation of frequencies at another site, j. This definition of coupling does not require that the overall conservation of site i change upon perturbation at j, but only that the amino acid population be rearranged. Therefore, we describe a site by a vector of 20 binomial probabilities of individual amino acid frequencies instead of the scalar multinomial probability of the overall amino acid distribution (*9*). This approach uniquely represents all changes in an amino acid distribution regardless of conservation at a given site.

For an evolutionarily well sampled MSA, where additional sequences do not significantly change the distribution at sites, the probability of any amino acid x at site i relative to that at another site, j, is related to the statistical free energy separating sites i and j for amino acid x ($\Delta G^x_{i\rightarrow j}$) by the Boltzmann distribution (*10*)

$$\frac{P^x_i}{P^x_j} = e^{\frac{\Delta G^x_{i\rightarrow j}}{kT^*}} \quad (1)$$

where kT^* is an arbitrary energy unit (*11*). The probability of any amino acid x at site i (P^x_i) is given by the binomial probability of the observed number of x amino acids, given its mean frequency in all proteins (Fig. 1A) (*12*). The full distribution of amino acids at a site i can then be characterized by a 20-element vector of P^x_i for all x ($\overrightarrow{P^x_i}$). If we take

Howard Hughes Medical Institute and Department of Pharmacology, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, TX 75235–9050, USA.

*To whom correspondence should be addressed. E-mail: rama@chop.swmed.edu

as site j a hypothetical site where all amino acids are found at their mean frequencies in the MSA as a reference state for all sites, we can use Eq. 1 to transform $\vec{P}_i^x$ into a vector of statistical energies that represents the evolutionary constraint at site i. We then define an overall empirical evolutionary conservation parameter (ΔG^{stat}) for site i

$$\Delta G_i^{stat} = kT^* \sqrt{\sum_x \left(\ln \frac{P_i^x}{P_{MSA}^x} \right)^2} \qquad (2)$$

which amounts to taking the magnitude of the vector of amino acid statistical energies for site i (*13*).

To test the theory, we chose the PDZ domain family as a model system for the analyses described below. PDZ domains are a

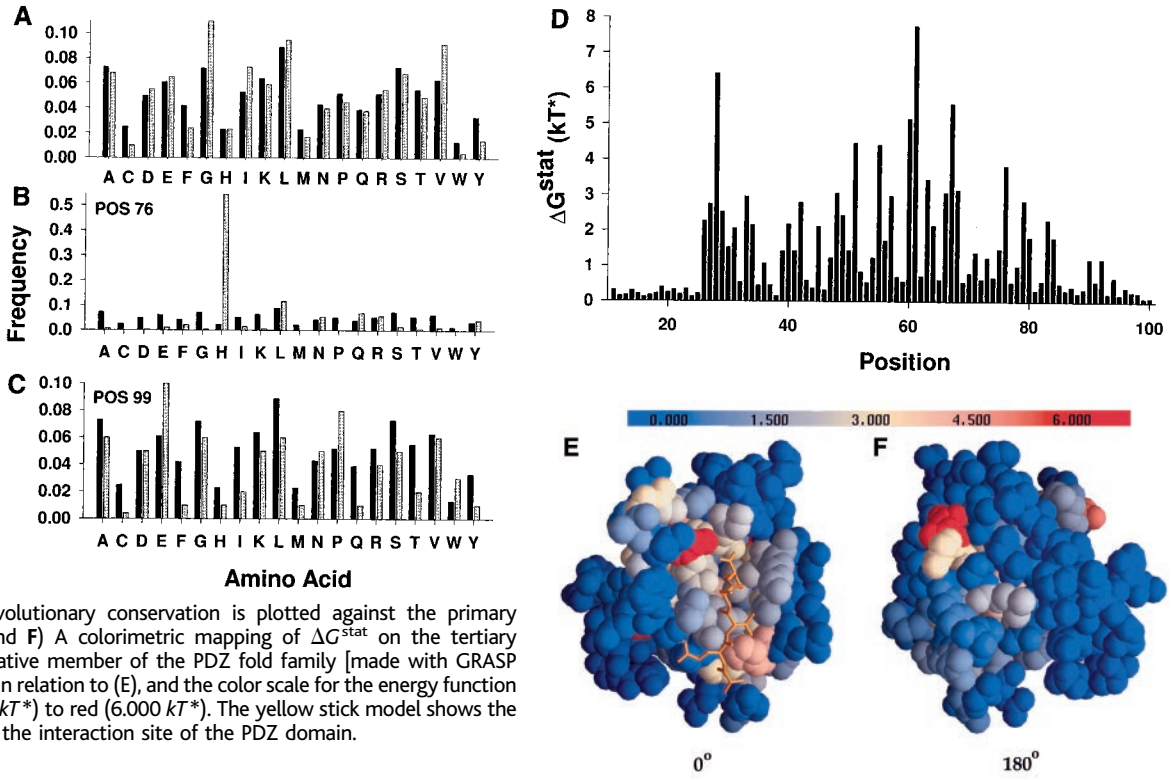

Fig. 1. A statistical energy function describing evolutionary conservation. (**A**) A comparison of amino acid distributions (*20*) in 36,498 unique eukaryotic proteins from the Swiss-Prot database (black bars) and 274 members of the PDZ family (gray bars). Examples of amino acid distributions in (**B**) moderately conserved and (**C**) weakly conserved positions (POS) in the PDZ family (gray bars) in comparison to the Swiss-Prot mean distribution (black bars). There is a difference in scale between (B) and (C). (**D**) The statistical energy (ΔG^{stat}) representing evolutionary conservation is plotted against the primary structure position. (**E** and **F**) A colorimetric mapping of ΔG^{stat} on the tertiary structure of a representative member of the PDZ fold family [made with GRASP (*29*)]. (F) is rotated 180° in relation to (E), and the color scale for the energy function ranges from blue (0.000 kT^*) to red (6.000 kT^*). The yellow stick model shows the peptide ligand bound at the interaction site of the PDZ domain.

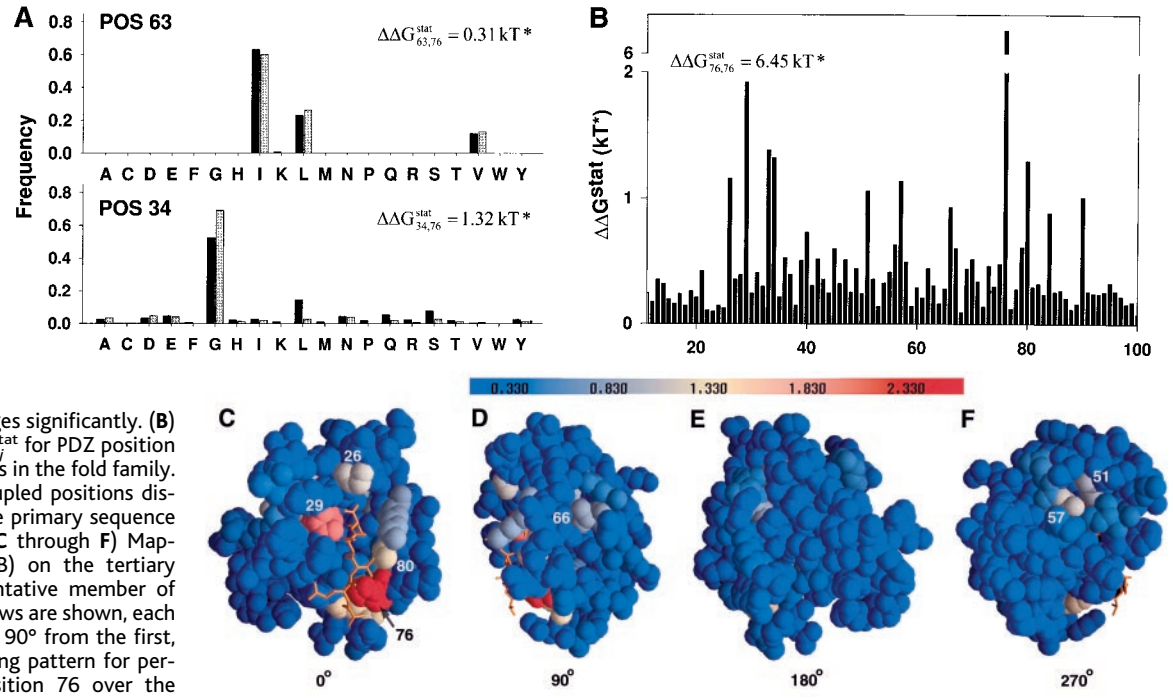

Fig. 2. Statistical coupling for a single site in the PDZ domain family. (**A**) Examples of amino acid distributions for two PDZ domain sites before (black bars) and after (gray bars) a 6.45 kT^* perturbation at position 76. The distribution at position 63 changes very little upon perturbation at position 76, despite high overall conservation, and the distribution at position 34 changes significantly. (**B**) A full mapping of $\Delta\Delta G_{i,j}^{stat}$ for PDZ position 76 for all other positions in the fold family. Only a small set of coupled positions distributed throughout the primary sequence emerge above noise. (**C** through **F**) Mapping of the data in (B) on the tertiary structure of a representative member of the fold family. Four views are shown, each successively rotated by 90° from the first, of the statistical coupling pattern for perturbations at PDZ position 76 over the entire PDZ domain. Coupled positions describe energetic interactions at sites spatially close to and distant from the point of perturbation. The color scale ranges from blue (0.330 kT^*) to red (2.330 kT^*).

family of small, evolutionarily well represented protein binding motifs for which four high-resolution structures of distantly related members exist (*14–16*). The structures are quite similar (root mean square deviation in C_α atoms of 1.4 Å), although the average sequence identity between pairs of domains is only 24% and, in many cases, is indistinguishable from random sequence identity. We used structure-based alignment techniques to generate a MSA of 274 eukaryotic PDZ domains (*17*). Overall amino acid distributions for all proteins and for PDZ domains alone differed only slightly, a fact that derives from the large sequence divergence of this fold family (Fig. 1A). Distributions at sites that represent moderately conserved [position 76, $\Delta G^{stat} = 3.83\ kT^*$, $\sigma = 0.4\ kT^*$ (Fig. 1B)] and weakly conserved [position 99, $\Delta G^{stat} = 0.1\ kT^*$, $\sigma = 0.4\ kT^*$ (Fig. 1C)] positions show that even moderate conservation skewed the mean amino acid distribution significantly, and lack of conservation was correlated with distributions closer to the mean.

Using Eq. 2, we calculated ΔG^{stat} for all sites on the PDZ domain alignments. These data plotted on the primary structure show a dispersed pattern that describes the overall conservation profile of the fold family (Fig. 1D). The same data plotted on a representative three-dimensional structure of a member of the family show that this pattern simplifies into a rough description of the protein interaction surface of the fold (Fig. 1, E and F). For example, the groove on the surface of the PDZ domain that contains the co-crystallized peptide ligand (*14, 16*) (Fig. 1E) emerges as the most conserved portion of the protein family. This finding is consistent with the intuitive expectation that a proper measure of conservation should be able to map functionally important sites on a protein (*18*).

To measure the functional coupling of sites, we performed an experiment in which the statistical energy vector at a given site i was measured for two conditions: (i) the full MSA (ΔG_i^{stat}) or (ii) a selected subset of the MSA representing a perturbation of the amino acid frequencies at another site, j ($\Delta G_{i|\delta j}^{stat}$). The magnitude of the difference in these two energy vectors gives a statistical coupling energy ($\Delta\Delta G_{i,j}^{stat}$) between sites i and j

$$\Delta\Delta G_{i,j}^{stat} = kT^* \sqrt{\sum_x \left(\ln \frac{P_{i|\delta j}^x}{P_{MSA|\delta j}^x} - \ln \frac{P_i^x}{P_{MSA}^x} \right)^2} \quad (3)$$

which quantitatively represents the degree to which the probability of individual amino acids at i is dependent on the perturbation at j (*13*). A systematic calculation of $\Delta\Delta G_{i,j}^{stat}$ at all sites, i, for a given site, j, gives the full-scale mapping of statistical coupling for position j over all protein sites.

We chose one functionally important site in the PDZ domain family as a test case for the perturbation analysis. The PDZ domain family is divided into distinct classes on the basis of target sequence specificity; class I domains bind to peptide ligands of the form -S/T-X-V/I-COO$^-$, and class II domains bind to sequences of the form -F/Y-X-V/A-COO$^-$ (*19, 20*). An important determinant of ligand specificity is domain position 76 (*14, 21*), which appears to select the identity of the antepenultimate peptide position. In class I domains, a histidine at this position hydrogen bonds to the serine or threonine hydroxyl of the characteristic recognition motif (*14*).

To examine the full pattern of energetic connectivity for PDZ position 76, we made a perturbation to the amino acid distribution at this site by extracting the subset of the MSA that contains only histidine at this position. The statistical energetic consequence of this perturbation is a 6.45-kT^* change at position 76 from the full MSA. We illustrate statistical coupling to position 76 through two examples. Position 63 is highly conserved in all PDZ domains, showing a distribution that is virtually exclusive for leucine, isoleucine, or valine (Fig. 2A, top) but is largely unaffected by the perturbation at position 76. Consequently, this position displays a low coupling energy ($\Delta\Delta G_{63,76}^{stat} = 0.31\ kT^*$, $\sigma = 0.3\ kT^*$) with respect to position 76. In contrast, the distribution at position 34 changes for several amino acids upon perturbation at position 76 (Fig. 2A, bottom), resulting in significant statistical coupling ($\Delta\Delta G_{34,76}^{stat} = 1.32\ kT^*$, $\sigma = 0.3\ kT^*$).

A full primary sequence mapping of statistical coupling for PDZ position 76 shows that most positions in the fold family were not coupled to the perturbed site; instead, only a small set of statistical couplings emerged from noise (Fig. 2B). Mapping the data on the PDZ domain tertiary structure shows that the coupled sites fall into three classes (Fig. 2, C through F). A small set of residues [80, 84, 33, 34] are in the immediate environment of position 76, a finding consistent with expected local propagation of energy from a site of perturbation. In addition,

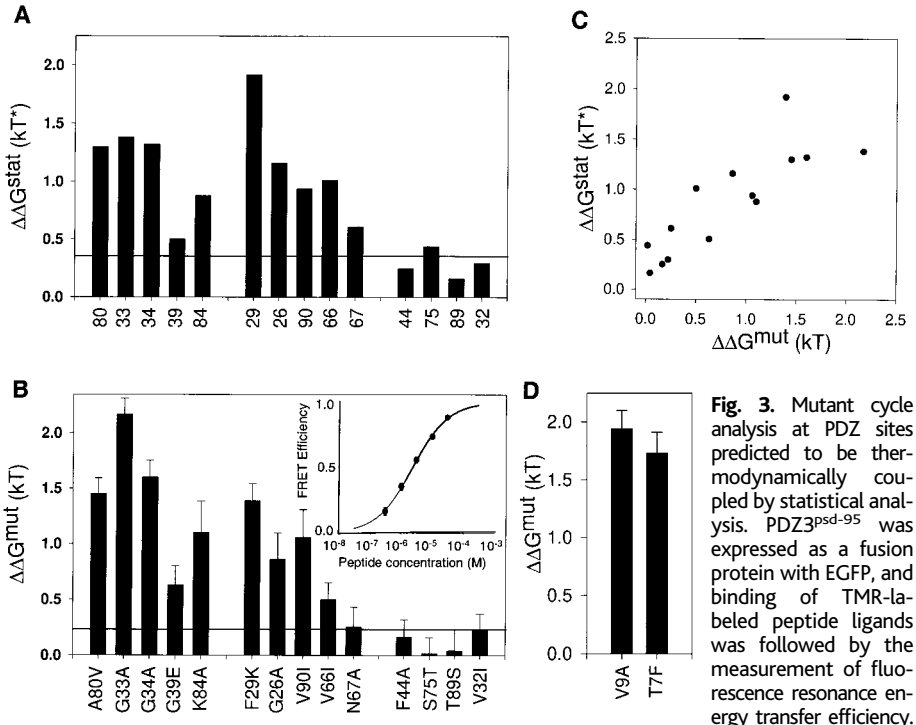

Fig. 3. Mutant cycle analysis at PDZ sites predicted to be thermodynamically coupled by statistical analysis. PDZ3$^{psd\text{-}95}$ was expressed as a fusion protein with EGFP, and binding of TMR-labeled peptide ligands was followed by the measurement of fluorescence resonance energy transfer efficiency. The inset in (B) shows an example of a binding isotherm for wild-type PDZ3^{psd95} protein and a class I binding peptide (*22*). An average and standard deviation of five measurements are shown for each ligand concentration tested (solid circles), with the smooth curve showing a fit to the Hill equation. Using this binding energy assay, we carried out double mutant cycle analysis for mutations at PDZ position 76 (H76Y) and at a set of statistically coupled and uncoupled positions. (**A**) and (**B**) show a comparison of statistical coupling ($\Delta\Delta G^{stat}$) and mutational coupling ($\Delta\Delta G^{mut}$), with sites categorized in three groups: those sites that are statistically coupled and near to position 76 [33, 34, 39, 80, 84], those sites that are statistically coupled but distant from position 76 [26, 29, 66, 67, 90], and those that are statistically uncoupled [32, 44, 75, 89]. The energy units (kT and kT^*) are different for the two analyses and cannot be quantitatively compared (*11*). In each graph, the horizontal line depicts the 1σ error in kT or kT^*. (**C**) A scatterplot of mutational coupling energies and statistical coupling energies. (**D**) Thermodynamic mutant cycle analysis between mutations at PDZ position 76 (H76Y) and mutations at ligand positions at the directly interacting position (T7F) and at the COOH-terminal position (V9A).

other interaction surface residues implicated in target sequence recognition [29, 26] emerged as coupled. This result suggests energy propagation through bound substrate and would be an expected consequence of cooperative interaction of binding site residues. Finally, we observed unexpected coupling at long range from sites in the core and on the opposite side of the PDZ domain [51, 57, 66, 90].

To determine how statistical coupling patterns are related to physical energetic coupling of sites, we used the technique of thermodynamic mutant cycle analysis (*3, 7*) to measure mutational coupling energies for position 76 for one PDZ domain [the third PDZ domain from PSD-95 ($PDZ3^{psd\text{-}95}$)] and compared these data to the statistical predictions. In the mutant cycle method, the energetic effect of one mutation, *m*1, is measured for two conditions: (i) the wild-type background (ΔG_{m1}) or (ii) the background of a second mutation, *m*2 ($\Delta G_{m1|m2}$). The difference in these two energies gives the coupling energy ($\Delta\Delta G_{m1,m2}$) between the two mutations. If *m*1 does not have the same effect in conditions 1 and 2 ($\Delta G_{m1|m2} \neq \Delta G_{m1}$), then $\Delta\Delta G_{m1,m2}$ is nonzero and indicates thermodynamic coupling of the two mutations.

To follow energetic coupling, we developed an equilibrium binding energy assay based on fluorescence resonance energy transfer between green fluorescent protein (GFP)-PDZ domain fusion proteins and tetramethylrhodamine (TMR)-labeled interacting peptides (*22, 23*). Figure 3B (inset) shows a binding isotherm for interaction of a wild-type GFP-$PDZ3^{psd\text{-}95}$ protein and a TMR-labeled class I peptide, demonstrating that this assay is capable of high-resolution mapping of binding energies.

Using this assay, we measured coupling energies for a mutation at position 76 [$His^{76} \rightarrow Tyr^{76}$ (H76Y)] against mutations at a set of 14 PDZ domain positions and two peptide positions (*23*). The mutations chosen were designed to test a range of statistical couplings on the PDZ domain, including a set of sites that are not significantly statistically coupled. Statistical energies at coupled sites, whether spatially near to or distant from position 76 are in fact well correlated to the thermodynamic coupling through mutagenesis (Fig. 3); statistically uncoupled sites display mutational coupling energies near to noise. Thus, patterns of statistical energetic coupling for a protein site are likely to describe the thermodynamic energetic connectivity for that position.

The statistical analysis for perturbation at position 76 indicated that other binding site positions [29 and 26] are energetically coupled and suggested the possibility of propagated coupling through the substrate peptide (Fig. 2, B and C). Indeed, mutations at the peptide position directly interacting with PDZ position 76 [$Thr^{7} \rightarrow Phe^{7}$ (T7F)] and at the position carrying the terminal carboxylate [$Val^{9} \rightarrow Ala^{9}$ (V9A)] were also thermodynamically coupled to the H76Y mutation (Fig. 3D).

With regard to the process of energy propagation in proteins, the overall mapping of coupling shows that residues interacting at long range are connected by a set of juxtaposed coupled residues. These residues form a pathway of energetic connectivity linking these positions (Fig. 4). For example, PDZ position 76 couples through the substrate peptide to the floor of the binding site (position 29), to position 51 within the core of the protein, and to position 57 on the opposite face from the ligand-binding pocket. Although residues composing the pathway occur in some cases along secondary structure elements (residues 76, 80, and 84 fall along one face of a helix), the pathway as a whole is an inherent property of the tertiary structure.

As an independent study, we examined patterns of statistical coupling for a MSA of 178 POZ domains, a fold family that mediates homo- and hetero-oligomerization of ion channel subunits and transcription factors (*24*). A perturbation at a position at the interaction surface [77] shows a pattern of statistical coupling that forms a sterically connected path [77, 118, 149, 148] that connects the oligomerization interface with the opposing protein surface through four aromatic-aromatic interactions (*25*), two of which are fully buried in the core of the protein (Fig. 4, C and D). Position 148, which forms one end of the pathway, has been implicated in the binding of K^+ channel β subunits to the POZ domain of Shaker-class K^+ channels; these subunits confer properties of rapid stimulus-dependent channel inactivation (*26*). The apparent energetic connectivity between two POZ protein interaction surfaces may represent a mechanism of β subunit–dependent

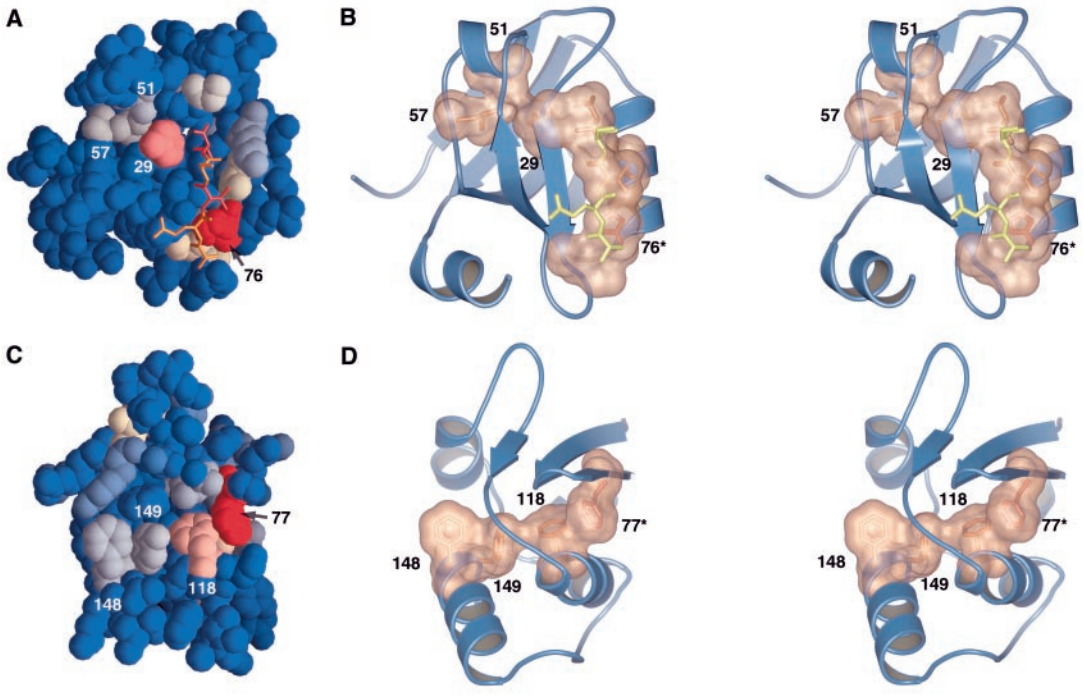

Fig. 4. Pathways of physical connectivity through the core underlie long-distance propagation of energetic coupling in two fold families. Sections through the protein core of PDZ and POZ domains show that, like at the protein surface, energetically coupled positions in the interior for perturbations at (**A**) PDZ position 76 and at (**C**) POZ position 77 are mostly surrounded by uncoupled positions. The color scale is the same as in Fig. 2. (**B**) and (**D**) show stereo images of the pattern of energetic coupling for these two positions superimposed on ribbon models of representative members of PDZ or POZ fold families, respectively. A continuous pathway of van der Waals interaction connects distantly coupled sites through the interior of each domain. The pathway is composed of residues in several secondary structure elements and is therefore an inherent property of the tertiary structure. The figure was prepared with Molscript (*33*), GLRender (*34*), Povray (*35*), and Raster3D (*36*).

regulation of K^+ channel activity.

The ability to efficiently propagate energy through tertiary structure is a fundamental property of many proteins and is the physical basis for key biological properties such as allostery and signal transmission. The coupled pathways may represent conduits along which energy distributes through a protein structure to generate these functional features. Protein interaction modules such as the PDZ and POZ domains are known to play key roles as organizing centers for multiprotein signaling complexes in which proteins are assembled into functional macromolecular units (*27*). In addition to this established role in cellular scaffolding, the finding of energetically coupled pathways within these domains raises the possibility that the interaction modules may also act as conductors of signaling. In the PDZ domain, evidence for such a role comes from the finding that interaction of the guanylate kinase domain of the multi-PDZ protein PSD-95 with MAP1A depends on the binding of target peptides to the PSD-95 PDZ domains (*28*).

As with any thermodynamic mapping, the approach described here can identify couplings, but it does not itself reveal the physical mechanism of the energetic coupling. Nevertheless, the arrangement of coupled residues into ordered pathways through the core of the PDZ and POZ protein folds suggests that one mechanism may be simple mechanical deformation of the structure along coupled pathways. Given the evolutionary basis of the statistical analysis, we infer that these pathways of energetic connectivity have emerged early in the evolution of the protein folds and, much like the atomic structure, are fundamentally conserved features of the domain families. With growing sequence data for evolutionarily distant genomes, the mapping of energetic connectivity for many fold families should be a realistic goal.

References and Notes

1. J. M. Holt and G. K. Ackers, *FASEB J.* **9**, 210 (1995); J. Monod, J. Wyman, J.-P. Changeux, *J. Mol. Biol.* **12**, 88 (1965); K. M. Perry, J. J. Onuffer, M. S. Gittelman, L. Barmat, C. R. Matthews, *Biochemistry* **28**, 7961 (1989); D. W. Pettigrew *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **79**, 1849 (1982); V. J. LiCata and G. K. Ackers, *Biochemistry* **34**, 3133 (1995); G. J. Turner *et al.*, *Proteins* **14**, 333 (1992); E. Freire, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 10118 (1999).
2. S. Atwell, M. Ultsch, A. M. De Vos, J. A. Wells, *Science* **278**, 1125 (1997); T. Clackson and J. A. Wells, *ibid.* **267**, 383 (1995); J. A. Wells, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 1 (1996); *Biotechnology* **13**, 647 (1995).
3. P. Hidalgo and R. MacKinnon, *Science* **268**, 307 (1995).
4. S. A. Goldstein, D. J. Pheasant, C. Miller, *Neuron* **12**, 1377 (1994); R. Ranganathan, J. H. Lewis, R. MacKinnon, *ibid.* **16**, 131 (1996); P. Stampe, L. Kolmakova-Partensky, C. Miller, *Biochemistry* **33**, 443 (1994).
5. L. Hedstrom, *Biol. Chem.* **377**, 465 (1996); L. Hedstrom, L. Szilagyi, W. J. Rutter, *Science* **255**, 1249 (1992); J. J. Perona, L. Hedstrom, W. J. Rutter, R. J. Fletterick, *Biochemistry* **34**, 1489 (1995).
6. P. A. Patten *et al.*, *Science* **271**, 1086 (1996).
7. P. J. Carter, G. Winter, A. J. Wilkinson, A. R. Fersht, *Cell* **38**, 835 (1984); G. Schreiber and A. R. Fersht, *J. Mol. Biol.* **248**, 478 (1995).
8. E. Neher, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 98 (1994).
9. The multinomial probability for all 20 amino acids gives the probability of randomly observing a given amino acid distribution at a site but is degenerate given the redistribution of amino acids with a similar mean frequency. For example, consider a site that displays a distribution of 0.4 Ala, 0.4 Asp, and 0.2 Ile in the overall alignment and changes to 0.4 Ala, 0.2 Asp, and 0.4 Ile upon perturbation at another site. Because the mean frequency of Asp and Ile is nearly identical (Fig. 1A), the multinomial probability of these two distributions is the same, although the significant reorganization of chemical character suggests that these positions are indeed coupled. Description of the site as vectors of individual amino acid probabilities accounts for all such cases because each amino acid distribution maps to a unique vector.
10. R. C. Tolman, *The Principles of Statistical Mechanics* (Dover, New York, 1938).
11. For conventional statistical mechanical systems at equilibrium, the temperature (T) of an ensemble is proportional to the mean velocity of state transitions and defines the fundamental energy unit kT, where k is Boltzmann's constant (*10*). Sites on a MSA can be seen as individual statistical mechanical systems that represent discrete states in an overall state space of amino acid frequencies. The "temperature" (T^*) of an ensemble of such systems is again related to the mean transition rates between states, but the energy unit in such a system (kT^*) is not necessarily related to that for conventional mechanical systems.
12. Each position in a MSA can be described as a 20-element vector of individual amino acid frequencies. Each element is transformed into a probability for that amino acid with the binomial density function

$$P(x) = \frac{N!}{n_x!(N - n_x)!} p_x^{n_x} (1 - p_x)^{N - n_x}$$

N is the total number of sequences, n_x is the number of sequences with amino acid x, and p_x is the mean frequency of amino acid x in all proteins. To determine p_x, we created histograms of amino acids for all 36,498 entries (as of October 1998) in the Swiss-Prot database of eukaryotic nonredundant proteins and calculated the mean values (Fig. 1A). Because all structural and functional information has been scrambled in this analysis, the frequencies of amino acids should represent that which is expected without any functional evolutionary constraint. Stirling's approximation was used for the evaluation of large factorials ($>$170).
13. For visualization and analysis, statistical energies were arbitrarily scaled by 0.01 for compatibility with GRASP and output in Microsoft Excel format or were written to a Protein Data Bank file of a representative member of the fold family. Mapping of statistical energies onto tertiary structures was done with GRASP (*29*). In evaluating statistical coupling, distributions at sites before and after perturbation were normalized for comparison.
14. D. A. Doyle *et al.*, *Cell* **85**, 1067 (1996).
15. J. H. Cabral *et al.*, *Nature* **382**, 649 (1996).
16. D. L. Daniels, A. R. Cohen, J. M. Anderson, A. T. Brunger, *Nature Struct. Biol.* **5**, 317 (1998).
17. Eukaryotic PDZ domains were collected from the nonredundant database of protein sequences with PSI-BLAST (*30*) (*e* score $\leq$ 0.001); four PDZ domains with known structures [(*14–16*); M. Socolich and R. Ranganathan, unpublished data] were used in initial searches. Alignments were created with PILEUP (Genetics Computer Group, Madison, WI), followed by structure-based manual alignment (*31*).
18. O. Lichtarge, H. R. Bourne, F. E. Cohen, *J. Mol. Biol.* **257**, 342 (1996).
19. Z. Songyang *et al.*, *Science* **275**, 73 (1997); C. P. Ponting, C. Phillips, K. E. Davies, D. J. Blake, *Bioessays* **19**, 469 (1997).
20. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; X, any amino acid; and Y, Tyr.
21. The numbering scheme for both PDZ and POZ domains used is consistent with that reported for the structures used for mapping statistical energies (*14, 24*).
22. A final concentration of 100 nM enhanced green fluorescent protein (EGFP)–PDZ fusion protein in storage buffer (*23*) was used for peptide titrations. Ligand peptides were synthesized with an NH_2-terminal TMR adduct and were freshly diluted from a single batch of 6 μM frozen aliquots for binding measurements. For all measurements, we used the binding peptide (or mutants thereof, as indicated) co-crystallized in the original structure determination (*14*). Energy transfer was followed by quenching of fluorescence at 508 nm (corrected for peptide fluorescence). Transfer efficiencies measured for four or five peptide concentrations covering a 2 log-order range around the dissociation constant were used for each binding-energy calculation; each individual measurement was made three to five times. Data were fit to the Hill equation (Origin, MicroCal Software, Northampton, MA).
23. Site-directed mutagenesis on the rat PSD-95 third PDZ domain (residues 294 through 402) was carried out with standard polymerase chain reaction–based techniques. Domains were expressed as COOH-terminal fusions with EGFP (*32*) using the pRSET-B vector (Invitrogen) in *Escherichia coli* [strain BL21(DE3), Stratagene]. Cultures (500 ml) in Terrific broth were grown to an optical density (600 nm) of 1.2 at 37°C, induced for 4 hours with 100 μM isopropyl-β-D-thiogalactopyranoside and harvested. Cells were lysed with B-PER (Pierce, Rockford, IL); cleared supernatants were batch-bound to a 0.5-ml bed volume of Ni–nitrilotriacetic acid agarose beads (Qiagen, Valencia, CA), prewashed in binding buffer (25 mM tris at pH 8.0, 500 mM NaCl, and 10 mM imidazole) and 0.1% Tween-20, washed with 50 column volumes of binding buffer, and eluted with elution buffer (50 mM tris at pH 8.0, 1 M NaCl, and 200 mM imidazole). The protein was dialyzed overnight into storage buffer (50 mM tris at pH 8.0, 100 mM NaCl, and 1 mM dithiothreitol) at 4°C and used immediately for binding assays or flash frozen and stored at −80°C for later use.
24. A. Kreusch, P. J. Pfaffinger, C. F. Stevens, S. Choe, *Nature* **392**, 945 (1998); L. Aravind and E. V. Koonin, *J. Mol. Biol.* **285**, 1353 (1999); V. J. Bardwell and R. Treisman, *Genes Dev.* **8**, 1664 (1994); N. V. Shen, X. Chen, M. M. Boyer, P. J. Pfaffinger, *Neuron* **11**, 67 (1993).
25. S. K. Burley and G. A. Petsko, *Science* **229**, 23 (1985).
26. S. Sewing, J. Roeper, O. Pongs, *Neuron* **16**, 455 (1996).
27. A. S. Fanning and J. M. Anderson, *J. Clin. Invest.* **103** 767 (1999); R. V. Schillace and J. D. Scott, *ibid.*, p. 761; R. Ranganathan and E. M. Ross, *Curr. Biol.* **7**, R770 (1997); S. Tsunoda *et al.*, *Nature* **388**, 243 (1997).
28. J. E. Brenman *et al.*, *J. Neurosci.* **18**, 8805 (1998).
29. A. Nicholls, K. Sharp, B. Honig, *Proteins* **11**, 281 (1991).
30. S. F. Altschul *et al.*, *Nucleic Acids Res.* **25**, 3389 (1997).
31. R. Doolittle, *Methods Enzymol.* **266**, 1 (1996).
32. R. Heim and R. Y. Tsien, *Curr. Biol.* **6**, 178 (1996).
33. P. J. Kraulis, *J. Appl. Crystallogr.* **24**, 946 (1991).
34. L. Esser, unpublished material.
35. D. Bacon and W. F. Anderson, *J. Mol. Graphics* **6**, 219 (1998).
36. E. A. Merrit and M. E. P. Murphy, *Acta Crystallogr.* **D50**, 869 (1994).

37. We thank M. Wall for help with figures, N. Grishin for advice regarding manual sequence alignments, A. Pertsemlidis for help with parsing the Swiss-Prot database, and L. Aravind for communication of data before publication. We are indebted to C. F. Stevens for teaching and important discussions. R.R. is a recipient of the Burroughs-Wellcome Fund New Investigator Award in the Basic Pharmacological Sciences and is an Assistant Investigator of the Howard Hughes Medical Institute.

7 April 1999; accepted 3 September 1999