Commentary

Lysozyme among the Lilliputians
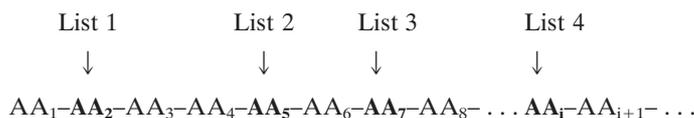
George D. Rose*

Department of Biophysics and Biophysical Chemistry, Johns Hopkins University School of Medicine, 725 North Wolfe Street, Baltimore, MD 21205

Site-directed mutagenesis (1), slightly more than two decades old, changed the way protein biochemists do their business. Now comes the sequel—patterned library analysis—in another landmark paper, reported in a recent issue of PNAS by Lahr *et al.* (2), and closed the millennium on a high note. But what is patterned library analysis?

A protein's conformation is fixed by a large number of weak interactions. Like Gulliver immobilized by the Lilliputians, the fold is substantially overdetermined, and, typically, the breaking of a few individual bonds here and there fails to produce global change. There are 306 structures of phage lysozyme (3) in the protein databank (4), all within a few residues of the wild-type protein. Local details aside, all look just like wild-type lysozyme. For both lysozyme and Gulliver, conformation is inherently a multifactorial issue.

At the opposite extreme, numerous examples demonstrate that proteins with scant sequence identity can also have similar structure and function. Thus, the stereochemical code that determines protein conformation must be highly degenerate. How can such a multifactorial and degenerate code be probed effectively? Approaches using single-site mutagenesis, which seemed so promising initially, were quickly overwhelmed by the combinatorics of the problem. Later approaches using saturation mutagenesis of various sorts (e.g., ref. 5) were a large step in the right direction, but a gap remains between current methods and the full repertoire of site-diversity and combinatorial complexity that is ultimately desired. That gap has now been filled by Lahr *et al.* (2).

Patterned library analysis was devised to test hypotheses about protein folding and function. The method uses resin-splitting technology to create libraries of arbitrary complexity by generating mutants at any number of residue positions, with each position varied according to its own menu of residue choices. For example, in the sequence below, four residue positions (2, 5, 7, and i) are selected for mutation. Each position is varied within a different list of residue options (lists 1, 2, 3, and 4, with n_1, n_2, n_3, and n_4 members, respectively), resulting in a library with $n_1 \times n_2 \times n_3 \times n_4 = N$ possible transformants.

List 1	List 2	List 3	List 4
$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$

$$\text{AA}_1-\textbf{AA}_2-\text{AA}_3-\text{AA}_4-\textbf{AA}_5-\text{AA}_6-\textbf{AA}_7-\text{AA}_8- \ldots \textbf{AA}_i-\text{AA}_{i+1}- \ldots$$

Given multiple libraries, a population of transformants can be selected at random from each library and then compared by using regression analysis. In particular, each transformant is evaluated for some activity of interest, using any convenient assay. For an activity that can be quantified, the transformants are, in effect, vectors, which are given by a linear combination of residue substitutions over available positions. This approach is a natural way to turn mutation data into a statistical model, and it is well suited to multifactorial phenomena like protein structure and function.

To validate their approach, Lahr *et al.* (2) assessed helix propensities in eglin c, a 70-residue proteinase inhibitor with a solitary helix. A close eglin homologue, CI2, has been utilized extensively in folding studies (6). Using three separate libraries, four surface positions in the helix were varied, with seven possible substitutions at each position. Six of the seven—K, Q, E, D, N, and H—were identical, but the seventh was chosen to be either P, A, or G, respectively. Binding activity was used as

Fig. 1. X-ray structure of eglin c (1CSE) (24), a 70-residue proteinase inhibitor. The solitary helix is at the opposite end of the molecule from the large binding loop.

an assay. In eglin, the helix is situated at the opposite end of the molecule from the inhibitor binding site (Fig. 1), supporting the assumption that the only route from increased helix stability to tighter binding is via global molecular stabilization. Accordingly, representative populations were chosen from each library and were screened for binding activity.

Helix propensity is a well studied phenomenon (7–10). Do results from patterned library analysis correlate with those obtained in biophysical methods? In Fig. 3 of their paper, Lahr *et al.* (2) show that their rank-ordered helix propensities agree with those of Chakrabartty *et al.* (11) to the same degree that the biophysically determined scales agree with each other.

Correlation with reliable biophysical methods is reassuring. But the real power of the approach lies in the companion regression analysis, with its ability to provide a numerical estimate of the degree to which the experimental data account for the model under consideration. Here, the helix propensities account for only 31% of the variance. The analysis informs us both that the result is significant and that the hypothesis is incomplete. We are tipped-off, as it were, that answers to further, unasked questions can be extracted from this system.

Experimental design is constrained not only by available methodology but also by our preconceptions. What we think affects what we think to measure. Mutations in the eglin c helix, phage lysozyme, and numerous other proteins alter the equilibrium position of the folding reaction, but rarely, if ever, do they change the overall fold in any appreciable way. Indeed, most mutagenesis studies seek to
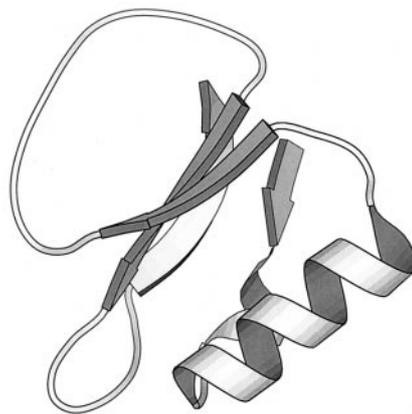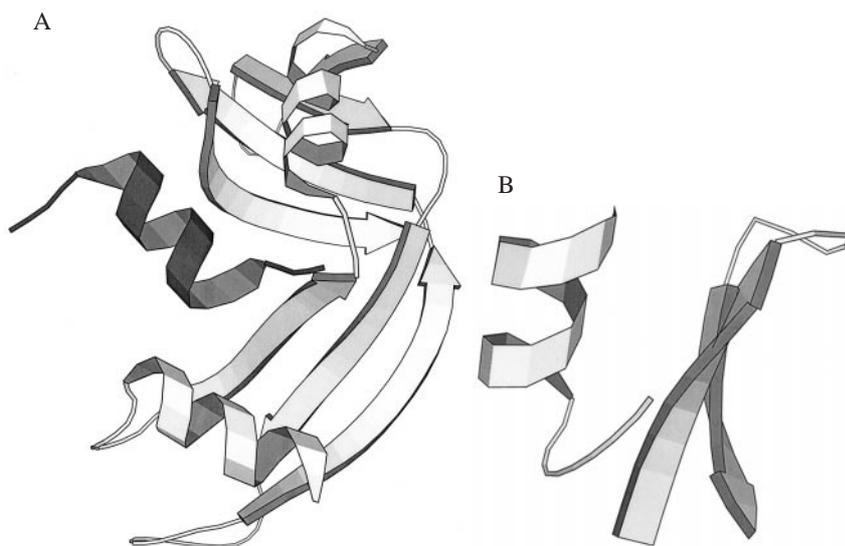
Fig. 2. (*A*) X-ray structure of ribonuclease S (2RNS) (25). S-peptide (residues 1–20, dark shading), a helix, has been cleaved from S-protein (residues 21–124, light shading), the remainder of the molecule. (Residues 16–23, which are disordered in the crystal structure, are not shown.) (*B*) X-ray structure of two fragments from intact BPTI (1BPT) (26). The β-hairpin fragment, Pβ, comprises residues 20–33, and the helical fragment, Pα, comprises residues 43–58. Apart from these two segments, the rest of the molecule is not displayed.

measure changes in the equilibrium position of the folding reaction

$$\text{unfolded protein (U)} \rightleftarrows \text{native protein (N)}$$

as a function of residue type and/or position. These studies can provide useful information about protein stability, the free energy difference between the folded and unfolded populations. But the measurements are indifferent to protein specificity, the extent to which the unfolded population is preorganized. Yet, both thought experiments (12) and actual experiments (13) concur that the reason lysozyme mutants all look like lysozyme originates primarily in their preorganization, and not in the subsequent folding reaction. Like the assembly of a prefabricated house, the folding reaction organizes subassemblies progressively, leading ultimately to a unique final structure—but much of the organization was already there at the onset (14, 15).

To put it blandly, not everyone agrees that unfolded proteins are already preorganized. Why then do experimentalists not simply design telling experiments that can distinguish among the possibilities?

Some do, of course (13, 16–18). But the unfolded population is too disperse for crisp characterization using available tools (like optical spectroscopy and NMR). Until now.

Now the challenge is to design informative experiments that can exploit patterned library analysis. The process will provide us with an invaluable opportunity to test our premises in fundamental ways. Here's an example, just to start the ball rolling.

In the progenitor of all protein fragment complementation studies, Richards (19) showed that limited proteolysis of ribonuclease A cleaved the molecule into two fragments: S-peptide (residues 1–20) and S-protein (21–124) (Fig. 2*A*). Neither fragment exhibited structure in isolation, but, when mixed in stoichiometric proportions, the two fragments reassociated with exquisite fidelity, with recovery of native-like structure and biological activity. Later studies (20, 21) did, in fact, find evidence of helical structure in isolated S-peptide, which contains residues 1–12, a helix in the intact molecule.

In a similar vein, Oas and Kim (22) selected two fragments from bovine pancreatic trypsin inhibitor (Fig. 2*B*)—residues 43–58, dubbed Pα, and residues 20–33, dubbed Pβ. The isolated fragments appeared to lack structure, but the reassociated dimer was native-like.

Experiments such as these have been widely interpreted to mean that the long range interactions needed to stabilize the intact structure are contributed by sites from the separated fragments. Absent fragment association, these crucial interactions are necessarily thwarted.

However, there is another possible interpretation: Recovery of structure in the association complex is largely a nonspecific solvation effect. In this view, the separate fragments are preorganized (23), but their fluctuations are large in water, and their structures evade detection by optical spectroscopy or NMR. Upon complementation, nonspecific surfaces of compatible size and polarity associate, providing the fragments with a mutually stabilizing opportunity to be sequestered from solvent water. This energetically favored association pushes the equilibrium toward the folded form—but not as a consequence of specific long-range interactions.

In principle, these different interpretations could be tested quite easily by simple mix and match experiments. Does S-peptide complement Pβ, and, conversely, does Pα complement S-protein? Of course, there are technical reasons why such an experiment might fail even if the basic idea is correct, but the use of patterned libraries could overcome such obstacles. Specifically, libraries could be designed to test whether substitutions that favor association can be found without corresponding regression toward each fragment's native counterpart. If these complexes can associate successfully but nonspecifically, then our ideas about protein folding may require further evaluation.

Lahr *et al.* (2) introduce the beginning of an open-ended methodology, one that lends itself to chip technology and Baysean analysis. As Dr. Gulliver was soon to learn, many an adventure awaits beyond the shores of Lilliput.

1. Hutchison, C. A., III, Phillips, S., Edgell, M. H., Gillam, S., Janke, P. & Smith, M. (1978) *J. Biol. Chem.* **253,** 6551–6560.

2. Lahr, S. J., Broadwater, A., Carter, C. W., Jr., Collier, M. L., Hensley, L., Waldner, J. C., Pielak, G. J. & Edgell, M. H. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 14860–14865.

3. Matthews, B. W. (1995) in *Studies on Protein Stability with T4 Lysozyme*, eds. Eisenberg, D. S. & Richards, F. M. (Academic, San Diego), Vol. 46, pp. 249–278.

4. Bernstein, F. C., Koetzle, T. G., Williams, G. J. B., Meyer, E. F., Jr., Brice, M. D., Rogers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977) *J. Mol. Biol.* **112,** 535–542.

5. Lim, W. A. & Sauer, R. T. (1991) *J. Mol. Biol.* **219,** 359–376.

6. Itahaki, L. S., Otzen, D. E. & Fersht, A. R. (1995) *J. Mol. Biol.* **254,** 260–288.

7. O'Neil, K. T. & DeGrado, W. F. (1990) *Science* **250,** 646–651.

8. Lyu, P. C., Liff, M. I., Marky, L. A. & Kallenbach, N. R. (1990) *Science* **250,** 669–673.

9. Padmanabhan, S., Marqusee, S., Ridgeway, T., Laue, T. M. & Baldwin, R. L. (1990) *Nature (London)* **344,** 268–270.

10. Merutka, G., Lipton, W., Shalongo, W., Park, S. H. & Stellwagen, E. (1990) *Biochemistry* **29,** 7511–7515.

11. Chakrabartty, A., Kortemme, T. & Baldwin, R. L. (1994) *Protein Sci.* **3,** 843–852.

COMMENTARY

12. Lattman, E. E. & Rose, G. D. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 439–441.
13. Sinclair, J. F. & Shortle, D. (1999) *Protein Sci.* **8,** 991–1000.
14. Baldwin, R. L. & Rose, G. D. (1999) *Trends Biochem. Sci.* **24,** 26–33.
15. Baldwin, R. L. & Rose, G. D. (1999) *Trends Biochem. Sci.* **24,** 77–83.
16. Cocco, M. J. & Lecomte, J. T. (1990) *Biochemistry* **29,** 11067–11072.
17. Feng, Y., Sligar, S. G. & Wand, A. J. (1994) *Nat. Struct. Biol.* **1,** 30–35.
18. Eliezer, D., Yao, J., Dyson, H. J. & Wright, P. E. (1998) *Nat. Struct. Biol.* **5,** 148–155.
19. Richards, F. M. (1958) *Proc. Natl. Acad. Sci. USA* **44,** 162–166.
20. Brown, J. E. & Klee, W. A. (1971) *Biochemistry* **10,** 470–476.
21. Bierzynski, A., Kim, P. S. & Baldwin, R. L. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 2470–2474.
22. Oas, T. G. & Kim, P. S. (1988) *Nature (London)* **336,** 42–48.
23. Srinivasan, R. & Rose, G. D. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 14258–14263.
24. Bode, W., Papamokos, E. & Musil, D. (1987) *Eur. J. Biochem.* **166,** 673–692.
25. Kim, E. E., Varadarajan, R., Wyckoff, H. W. & Richards, F. M. (1992) *Biochemistry* **31,** 12304–12314.
26. Danishefsky, A. T., Housset, D., Kim, K.-S., Tao, F., Fuchs, J., Woodward, C. & Wlodawer, A. (1993) *Protein Sci.* **4,** 577–587.