Proc. Natl. Acad. Sci. USA
Vol. 94, pp. 10015–10017, September 1997

Commentary

Rational protein design: Combining theory and experiment

H. W. Hellinga[*]

Department of Biochemistry, Duke University Medical Center, Durham, NC 27710

The rational design of protein structure and function is rapidly emerging as a powerful approach to test general theories in protein chemistry (1). *De novo* creation of a protein or an active site requires that all the necessary interactions are provided. The design approach is therefore a way to test the limits of completeness of understanding experimentally. Furthermore, if the experiments are devised in a progressive fashion, such that the simplest possible designs are tried first, followed by iterative additions of more complex interactions until the desired result is achieved, then it may be possible to identify a minimally sufficient set of components. At the center of the design approach is the "design cycle," in which theory and experiment alternate. The starting point is the development of a molecular model, based on rules of protein structure and function, combined with an algorithm for applying these. This is followed by experimental construction and analysis of the properties of the designed protein. If the experimental outcome is failure or partial success, then a next iteration of the design cycle is started in which additional complexity is introduced, rules and parameters are refined, or the algorithms for applying them are modified. The paper by Dahiyat and Mayo (2) in the current issue of these *Proceedings* describes such a design cycle. Sequences predicted to repack the interior of a small protein were generated by a computer design algorithm using different sets of parameters describing the packing interactions, thereby establishing a direct experimental correlation between the design parameters and the properties of the resulting proteins. This work is the latest addition to a series of efforts in which objective computational techniques developed to create protein structure (3–8) or function (9, 10) are being tested directly by experiment. The ultimate goal of such procedures is to develop a fully automated protein design method (6).

Design of a protein requires that both a structure and a sequence are specified. The basic forces that determine the noncovalent interactions within the polypeptide chain, with the surrounding solvent, and with ligands are relatively well understood: van der Waals and electrostatic interactions, hydrogen bonds, the hydrophobic effect, and the favorable packing interactions associated with the condensed state of protein interiors (11). However, the number of conformations a particular polypeptide can potentially adopt as well as the number of different sequences that can be built into even a small protein is vast.[†] Furthermore, many of these sequences and their conformations are distinguished only by relatively small energy differences. The combination of the immense combinatorial complexity and subtle energetic differences turns the seemingly simple basic interactions into a dauntingly complex landscape of virtually infinite possibilities. The ability of an algorithm to explore this vast landscape and seek out preferred solutions that have to be distinguished from closely related inferior possibilities is therefore a crucial component of any rational design approach. All design methods use the same general approach to reduce the immense complexity of the search problem. The structure of a protein backbone is chosen *a priori*, kept fixed, and redecorated with different amino acid sequences that are predicted to be structurally compatible with that fold. This "inverse folding" approach (12) therefore removes the backbone conformational degrees of freedom from the design problem.

The first rational design approaches used qualitative rules of protein structure applied by inspection (13). These experiments established that it is possible to create sequences *de novo* that adopt defined structures (1, 14). Furthermore, they demonstrated that, by following a progressive design strategy [or "hierachic design" (1)] in which increasing levels of complexity are iteratively introduced, new insights into the fundamentals of protein structure and function can be gained. One of the remarkable observations of these experiments was that it is surprisingly easy to obtain globally correct folds. However, the local details were found to be difficult to get correct. The interiors of these designed proteins show a high degree of disorder, which does not resemble the tightly packed, unique arrangement of natural systems. Global correctness in these designs apparently resulted from incorporation of the correct "binary pattern" of hydrophobic and hydrophilic residues, which sets up the geometric specification of the protein interior and exterior for the hydrophobic effect to act on (15, 16). The difficulty in designing well-ordered cores can be viewed as a problem in specificity. The side chains in a disordered core adopt many alternative conformations of approximately equal energy, instead of assuming a single, specific arrangement.

To achieve specificity, the desired state (well-ordered core) has to have the lowest free energy of all possible states (ground state), and there has to be a large free energy difference between the next available state: the free energy of specificity, ΔG_{spec} (Fig. 1). There are two ways to achieve such a free energy gap: *raising* the free energy of competing states, or *lowering* that of the desired state. One approach is to introduce specific features that prevent the formation of alternative conformations, thereby raising their energy ["negative design" (14)]. Constructing protein interiors out of sequences that increase the degree of geometric irregularity, making it less likely for alternative isoenergetic conformations to exist, results in better-ordered cores (1, 17). Another approach is to lower the free energy of the desired state by searching for a core-forming sequence with the lowest possible free energy that can be located in the entire space of sequences and their conformations ("target state optimization"). This is difficult to achieve by inspection, because of the combinatorial vastness of the search space. Here the computational approaches come into their own. Dahiyat and Mayo (2) use their version of the Dead-End Elimination algorithm (18) to identify the sequence with the lowest free energy minimum that repacks the core of the B1 domain of protein G. By varying the sizes of the atomic

[*] To whom reprint requests should be addressed at: Department of Biochemistry, Box 3711, Duke University Medical Center, Durham, NC 27710. e-mail: hellinga@linnaeus.biochem.duke.edu.

[†] A 100-residue protein can potentially sample $(2\pi/\theta)^{100 n_b n_s}$ conformations (where θ is the chosen sampling interval for a given torsional degree of freedom, $n_b = 2$ is the number of main-chain angles, and $n_s \approx 2$ is the average number of side-chain angles), and can assume 20^{100} sequences.

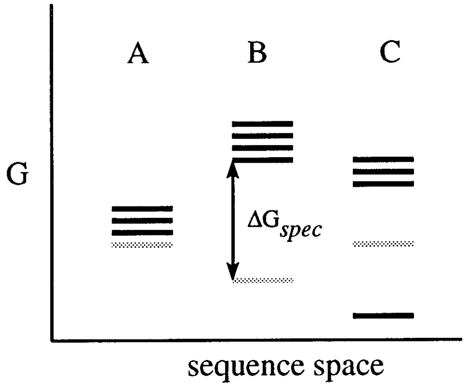

FIG. 1. The requirements for specificity. Three different hypothetical sequences are shown along the *x*-axis. Each sequence can adopt many different states (in a disordered core, for instance). The free energy of each state is given by a horizontal line. The target state is shown in gray. Sequence A is nonspecific, because all the states are approximately isoenergetic. Sequence C has the incorrect specificity, because there is a competing state of lower free energy. Sequence B is specific, because the target state corresponds to the ground state, and there is a large free energy gap, ΔG_{spec}, between it and the next available state. Note that to improve A by moving to B, the free energy of the target state was lowered (Target State Optimization) and the competing states were raised (Negative Design).

radii in their calculations, they are able to artificially tune the packing density, or degree of precision with which the jig-saw puzzle of the core is put together. The experimental behavior of their B1 variants convincingly shows that to get well-ordered cores and folded proteins, it is necessary to predict sequences that fit exquisitely. The strategy of achieving ordered cores by target state optimization therefore works remarkably well. Other algorithms (3, 4), applied to other proteins (19), have also successfully predicted hydrophobic core sequences, demonstrating that precise packing details matter.

The inverse folding concept of redecorating a fixed protein backbone with amino acids can also be used for the design of function in proteins. Algorithms have been developed and tested to rebuild the surface of existing binding sites to change their specificity (20), or to introduce active sites *de novo* (9, 10). Such algorithms, as well as qualitative designs by inspection, have resulted in the construction of a number of metalloproteins (21) where the interplay between the protein fold and the reactivity of the metal center can be studied. Several primitive but functional enzymes have also been constructed (22, 23). Progressive designs and iterative cycles are beginning to elucidate a number of global features that are necessary to create controlled activity.

Most of the computational approaches developed so far have focused on well defined regions of a protein frame, or an area where an active site can be (re)constructed. Furthermore, the backbone is typically left untouched, in strict interpretation of the inverse folding concept. To move toward the ultimate goal of fully automated design, entire protein chains have to be redecorated, and it is also necessary to start considering relaxation of the backbone without altering the overall topology to better explore fitting of allowed sequences. Algorithms for redesigning surface positions have been developed (7). Systematic backbone deformation is much more problematic, but can be done if the geometry of the backbone can be described by parametric equations, as has been proven by experiment in some cases (5, 8).

So far the automated design algorithms work by optimizing the compatibility of the sequence with the structure of the desired state (folded protein, or protein/ligand complex), without explicit consideration of other potential states and maximizing of ΔG_{spec}. This strategy of considering only target state optimization has worked surprisingly well for the successful design of hydrophobic cores. It actually does not work so well for the design of metal centers, if a metal can readily adopt different coordination numbers, geometries, or activities (24). Similar considerations come into play in automated redesign of ligand-binding sites, where it is difficult to discriminate between closely related ligands (20). In both cases it is clear that other states need to be considered explicitly, and that negative design as well as target state optimization plays an important role. All the important states have to be taken into consideration in these more challenging situations. In the terminology of statistical mechanics, a proper partition function has to be integrated over all possible states.

Theoretical studies with lattice models of proteins (25) and experiments have demonstrated that explicit consideration of alternative folds will be necessary in the design process of entire protein sequences. For instance, core mutations can change the oligomerization state of a coiled coil (26). Even more dramatic is a qualitative design experiment in which the B1 domain of protein G (one α-helix, four β-strands) was transformed into Rop (a four-helix bundle), by changing no more than 50% of the sequence (27). Both experiments show that similar sequences can adopt dramatically different folds. To reliably calculate entire sequences *de novo* for such structures it is necessary to consider many more states than just the target.

It is, of course, impossible to construct a partition function over all the possible folds that a sequence of a reasonable length can adopt, using the type of high-resolution model necessary for calculating the final packing details in an automated design program. However, it is probably not necessary to go to such extremes. Binary patterns composed of hydrophobic and hydrophilic residues are likely to play a dominant role in the selection of the overall geometry of many protein folds (16). It may therefore be possible to develop a hierarchic design algorithm in which the first step is to calculate binary patterns that uniquely specify the desired topology (28) by explicitly considering and destabilizing alternative topologies, followed by the detailed calculations necessary for core packing and surface decoration. Such a strategy has worked well in an empirical design experiment in which combinatorial libraries of a four-helix bundle were constructed (29).

The natural interplay of theory and experiment in rational design makes this approach a powerful method for testing general theories of structure and function. As the questions that are being asked become increasingly sophisticated, use of automated design algorithms to solve the tremendous combinatorial challenges inherent in conformational and sequence spaces will become a standard approach. It is clear that one of the main challenges is the development of algorithms that can deal directly with structural and functional specificity. The statistical mechanical concepts developed in the simple exact lattice models (30) will have to be applied to the high-resolution modeling needed for calculating sequences used in experiments.

1. Bryson, J. W., Betz, S. F., Lu, H. S., Suich, D. J., Zhou, H. X., O'Neil, K. T. & DeGrado, W. F. (1995) *Science* **270,** 935–941.
2. Dahiyat, B. I. & Mayo, S. L. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 10172–10177.
3. Hurley, J. H., Baase, W. A. & Matthews, B. W. (1992) *J. Mol. Biol.* **224,** 1143–1159.
4. Desjarlais, J. R. & Handel, T. M. (1995) *Protein Sci.* **4,** 2006–2018.
5. Harbury, P. B., Tidor, B. & Kim, P. S. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 8408–8412.
6. Dahiyat, B. I. & Mayo, S. L. (1996) *Protein Sci.* **5,** 895–903.
7. Dahiyat, B. I., Gordon, B. & Mayo, S. L. (1997) *J. Mol. Biol.* **6,** 1333–1337.
8. Su, A. & Mayo, S. L. (1997) *Protein Sci.* **6,** in press.
9. Hellinga, H. W. & Richards, F. M. (1991) *J. Mol. Biol.* **222,** 763–785.

10. Clarke, N. D. & Yuan, S.-M. (1995) *Proteins Struct. Funct. Genet.* **23,** 256–263.
11. Dill, K. A. (1990) *Biochemistry* **29,** 7133–7155.
12. Pabo, C. A. (1983) *Nature (London)* **301,** 200.
13. Richardson, J. S., Richardson, D. C., Tweedy, N. B., Gernert, K. M., Quinn, T. P., Hecht, M. H., Erickson, B. W., Yan, Y., McClain, R. D., Donlan, M. E. & Surles, M. C. (1992) *Biophys. J.* **63,** 1186–1209.
14. Hecht, M. H., Richardson, J. S., Richardson, D. C. & Ogden, R. C. (1990) *Science* **249,** 884–891.
15. Cordes, M. H. J., Davidson, A. R. & Sauer, R. T. (1996) *Curr. Opin. Struct. Biol.* **6,** 3–10.
16. Beasley, J. R. & Hecht, M. H. (1997) *J. Biol. Chem.* **272,** 2031–2034.
17. Quinn, T. P., Tweedy, N. B., Williams, R. W., Richardson, J. S. & Richardson, D. C. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 8747–8751.
18. Desmet, J., Maeyer, M. D., Hazes, B. & Lasters, I. (1992) *Nature (London)* **356,** 539–542.
19. Lazar, G. A., Desjarlais, J. R. & Handel, T. M. (1997) *Protein Sci.* **6,** 1167–1178.
20. Wilson, C., Mace, J. E. & Agard, D. A. (1991) *J. Mol. Biol.* **220,** 495–506.
21. Hellinga, H. W. (1997) *Curr. Opin. Biotechnol.* **7,** 437–441.
22. Johnsson, K., Allemann, R. K., Widmer, H. & Benner, S. A. (1993) *Nature (London)* **365,** 530–532.
23. Pinto, A. L., Hellinga, H. W. & Caradonna, J. P. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 5562–5567.
24. Hellinga, H. W., Caradonna, J. P. & Richards, F. M. (1991) *J. Mol. Biol.* **222,** 787–803.
25. Yue, K., Fiebig, K. M., Thomas, P. D., Chan, H. S., Shakhnovich, E. I. & Dill, K. A. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 325–329.
26. Harbury, P. B., Zhang, T., Kim, P. S. & Alber, T. (1993) *Science* **262,** 1401–1407.
27. Dalal, S., Balasubramanian, S. & Regan, L. (1997) *Nat. Struct. Biol.* **4,** 548–552.
28. West, M. E. & Hecht, M. H. (1995) *Protein Sci.* **4,** 2032–2039.
29. Kamtekar, S., Schiffer, J. M., Xiong, H., Babik, J. M. & Hecht, M. H. (1993) *Science* **262,** 1680–1685.
30. Dill, K. A., Bromberg, S., Yue, K., Fiebig, K. M., Yee, D. P., Thomas, P. D. & Chan, H. S. (1995) *Protein Sci.* **4,** 561–602.