The Protein-Folding Problem, 50 Years On

Ken A. Dill[1,2,3]* and Justin L. MacCallum[1]

The protein-folding problem was first posed about one half-century ago. The term refers to three broad questions: (i) What is the physical code by which an amino acid sequence dictates a protein's native structure? (ii) How can proteins fold so fast? (iii) Can we devise a computer algorithm to predict protein structures from their sequences? We review progress on these problems. In a few cases, computer simulations of the physical forces in chemically detailed models have now achieved the accurate folding of small proteins. We have learned that proteins fold rapidly because random thermal motions cause conformational changes leading energetically downhill toward the native structure, a principle that is captured in funnel-shaped energy landscapes. And thanks in part to the large Protein Data Bank of known structures, predicting protein structures is now far more successful than was thought possible in the early days. What began as three questions of basic science one half-century ago has now grown into the full-fledged research field of protein physical science.

Protein molecules embody a remarkable relationship between structure and function at the molecular level. Proteins perform many different functions in biochemistry. A protein's biological mechanism is determined by its three-dimensional (3D) native structure, which in turn is encoded in its 1D string of amino acid monomers.

This year marks the 50th anniversary of the 1962 Nobel Prize in Chemistry awarded to Max Perutz and John Kendrew for their pioneering work in determining the structure of globular proteins (*1*–*3*). That work laid the foundation for structural biology, which interprets molecular-level biological mechanisms in terms of the structures of proteins and other biomolecules. Their work also raised the question of how protein structures are explained by physical principles. Upon seeing the structure of myoglobin (Fig. 1) at 6 Å resolution (*1*), Kendrew *et al*. said,

> "Perhaps the most remarkable features of the molecule are its complexity and its lack of symmetry. The arrangement seems to be almost totally lacking in the kind of regularities which one instinctively anticipates, and it is more complicated than has been predicated by any theory of protein structure. Though the detailed principles of construction do not yet emerge, we may hope that they will do so at a later stage of the analysis."

The protein-folding problem came to be three main questions: (i) The physical folding code: How is the 3D native structure of a protein determined by the physicochemical properties that are encoded in its 1D amino-acid sequence? (ii) The folding mechanism: A polypeptide chain has an almost unfathomable number of possible conformations. How can proteins fold so fast? (iii) Predicting protein structures using computers: Can we devise a computer algorithm to predict a protein's native structure from its amino acid sequence? Such an algorithm might circumvent the time-consuming process of experimental protein-structure determination and accelerate the discovery of protein structures and new drugs.

Here, we give our perspective on these questions at the broad-brush level. More detailed reviews can be found elsewhere (*4*–*8*).

The Physical Code of Protein Folding

What forces drive a protein to its 3D folded structure? Much insight comes from the Protein Data Bank (PDB), a collection of now more than 80,000 protein structures at atomic detail (*9*). The following factors appear to contribute (*10*): (i) Hydrogen bonds. Protein structures are composed of α-helices and β-sheets, as was predicted by Linus Pauling on the basis of expected hydrogen bonding patterns (*11*). (ii) van der Waals interactions. The atoms within a folded protein are tightly packed, implying the importance of the same types of close-ranged interactions that govern the structures of liquids and solids. (iii) Backbone angle preferences. Like other types of polymers, protein molecules have preferred angles of neighboring backbone bond orientations. (iv) Electrostatic interactions. Some amino acids attract or repel because of negative and positive charges. (v) Hydrophobic interactions. Proteins ball up into well-packed folded states in which the hydrophobic (H) amino acids are predominantly located in the protein's core and the polar (P) amino acids are more commonly on the folded protein's surface. Theory and experiments indicate that folding is governed by a predominantly binary code based on interactions with surrounding water molecules: There are few ways a given protein sequence of H and P residues can configure to bury its hydrophobic amino acids optimally (*12*, *13*). (vi) Chain entropy. Opposing the folding process is a large loss in chain entropy as the protein collapses into its compact native state from its many open denatured configurations (*12*).

These physical forces are described approximately by "forcefields" (*14*). Forcefields are models of potential energies that are used in computer simulations. They are widely applied to studies of protein equilibria and dynamics. In computer modeling, a protein molecule is put into an initial configuration, often random. Conformations change over the course of the simulation by repeatedly solving Newton's dynamical laws of motion for the atoms of the protein molecule and the solvent by using the forcefield energies. According to the laws of thermodynamics, systems tend toward their states of lowest free energy. Computational protein folding explores the process by which the protein proceeds through conformational states to states of lower free energies. As shown in Fig. 2, the thermodynamically stable states of 12 small protein structures can be reached fairly successfully by means of extensive molecular dynamics (MD) simulations in a bath of

Fig. 1. In 1958, Kendrew and co-workers published the first structure of a globular protein: myoglobin at 6 Å resolution (*1*). Its puzzlingly complex structure lacked the expected symmetry and regularity and launched the protein-folding problem. [With permission from the Medical Research Council Laboratory of Molecular Biology]

[1]Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, NY 11794–5252, USA. [2]Department of Physics, Stony Brook University, Stony Brook, NY 11794–3800, USA. [3]Department of Chemistry, Stony Brook University, Stony Brook, NY 11794–3400, USA.

*To whom correspondence should be addressed. E-mail: dill@laufercenter.org

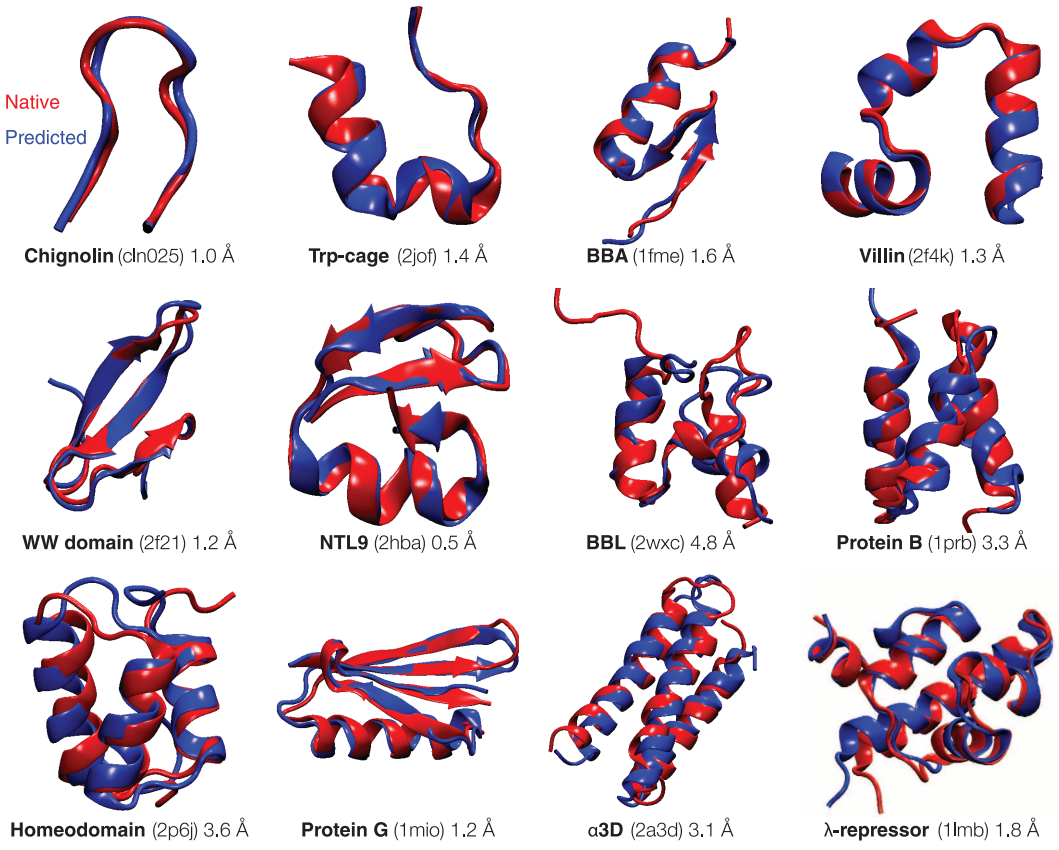

Fig. 2. Modern physical models can compute the folded structures of some small proteins. Using a high-performance custom computer called Anton (*48*), Shaw and co-workers observed reversible folding and unfolding in more than 400 events across 12 small proteins to structures within 4.5 Å of the experimental structure (*15*). The experimental structures are shown in red, and the computed structures are blue. Shown are the name, PDB identifier, and RMSD (root-mean-square deviation between alpha carbon atoms) between the predicted and experimental structures. [Adapted with permission (*15*)]

explicit water molecules (*15*). However, such successes, important as they are, are limited. So far, such modeling succeeds only on a limited set of small simple protein folds (*16*). And, it does not yet accurately predict protein stabilities or thermodynamic properties. Opportunities for the future include better forcefields, better models of the protein-water interactions, and faster ways to sample conformations, which are far too limited, even with today's most powerful computers.

The early days saw hopes of finding simple sequence patterns—say of hydrophobic, polar, charged, and aromatic amino acids—that would predict protein structures and stabilities. That has not materialized. Nevertheless, the results of the detailed atomic simulations described above give optimism that atomically detailed modeling is systematically improving and is contributing to our understanding of protein sequence-structure relationships.

The Rate Mechanism of Protein Folding

At a meeting in Italy in 1968, Cyrus Levinthal raised the question (*17*) of how, despite the huge number of conformations accessible to it, a protein molecule can fold to its one precisely defined native structure so quickly (microseconds, for some proteins). How does the protein "know" what conformations not to search?

This question led to a major experimental quest to characterize the kinetics of protein folding and to find folding intermediates, which are partially structured states along the "folding pathway" (*18*, *19*). The hope was that snapshots of the chain caught in the act of folding would give insights into folding "mechanisms," the rules by which nature performs conformational searching. The experimental challenge was not just to measure atom-by-atom contacts within the heterogeneous interior of a protein molecule, but to do it on the fly, over microsecond-to-second time scales. This drove development of a powerful arsenal of new experimental methods, including mutational studies, hydrogen exchange, fluorescence labeling, laser temperature jumps, and single-molecule methods [reviewed elsewhere (*7*)].

The general-principles solution of the needle-in-a-haystack conundrum emerged from polymer statistical thermodynamics. Studies of the chain entropies in models of foldable polymers showed that more compact, low-energy conformational ensembles have fewer conformations (*12*, *20*–*23*), indicating that protein-folding energy landscapes are funnel-shaped (Fig. 3). Protein folding landscapes are narrower at the bottom; there are few low-energy, native-like conformations and many more open unfolded structures. A protein folds by taking random steps that are mostly incrementally downhill in energy. Steps need only be favorable by one to two times the thermal energy to reach the native structure rapidly (*24*). Insights from funnels, however, have not yet been sufficient to improve computer search methods. A landscape that appears smooth and funnel-shaped on a global scale can be rough on the local scales that are sampled in computer simulations.

But we are still missing a "folding mechanism." By mechanism, we mean a narrative that explains how the time evolution of a protein's folding to its native state derives from its amino acid sequence and solution conditions. A mechanism is more than just the sequences of events followed by any one given protein in experiments or in computed trajectories. We do not yet have in hand a general principle that is applicable to a broad range of proteins, that would explain differences and similarities of the folding routes and rates of different proteins in advance of the data, and that properly average, in some meaningful way, over "irrelevant" thermal motions. One difficulty has been reconciling our "macroscopic" understanding of kinetics (mass-action models) that result from ensemble-averaged experiments with our "microscopic" understanding of the angstrom-by-angstrom changes of each protein conformation in computer simulations (energy landscapes). However, there are a few general conclusions (*25*). Proteins appear to fold in units of secondary structures. A protein's stability increases with its growing partial structure as it folds. And, a protein appears to first develop local structures in the chain (such as helices and turns) followed by growth into more global structures. Even though the folding process is blind, nevertheless it can be fast because native states can be reached by this divide-and-conquer, local-to-global process (*26*, *27*). Funneled landscapes predict that the different individual molecules of the same protein sequence may each follow microscopically different routes to the same native structure. Some paths will be more populated than others.

Computing Protein Structures from Amino Acid Sequences

A grand challenge has been to develop a computer algorithm that can predict a protein's 3D native structure from its amino acid sequence. On the one hand, knowledge of native structures is a starting point for understanding biological mechanisms and for discovering drugs that can

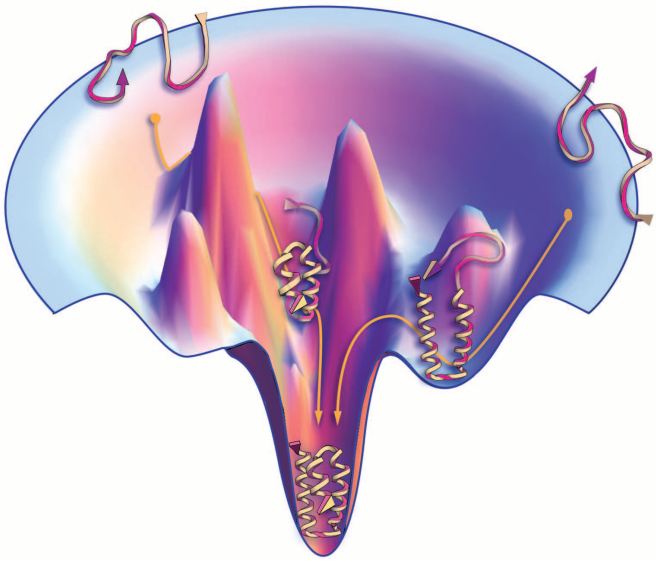

Fig. 3. Proteins have a funnel-shaped energy landscape with many high-energy, unfolded structures and only a few low-energy, folded structures. Folding occurs via alternative microscopic trajectories.

inhibit or activate those proteins. On the other hand, we know 1000-fold more sequences than structures, and this gap is growing because of developments in high-throughput sequencing. So, there is considerable value in methods that could accurately predict structures from sequences.

Computer-based protein-structure prediction has been advanced by Moult and colleagues, in an event initiated in 1994 called CASP: Critical Assessment of protein Structure Prediction (*28*, *29*). Held every second summer, CASP is a community-wide blind competition in which typically more than 100 different "target sequences" (of proteins whose structures are known but not yet publicly available) are made available to a community that numbers more than 150 research groups around the world. Each participating group applies some algorithmic scheme that aims to predict the 3D structures of these target proteins. After each CASP event, the true experimental structures are then revealed, group performances are evaluated, and community evaluations are published.

Currently, all successful structure-prediction algorithms are based on assuming that similar sequences lead to similar structures. These methods draw heavily on the PDB, which now contains more than 80,000 structures. However, many of these structures are similar, and the PDB contains only ~4000 structural families and 1200 folds (*30*).

CASP-wide progress over the past 18 years is summarized in Fig. 4A. Prediction accuracies improved from CASP1 (1994) to CASP5 (2002) on the basis of several advances: (i) PDB expanded from ~1600 structures to 19,000 during that time. (ii) Better sequence search and alignment tools, such as Position-Specific Iterated Basic Local Alignment Search Tool (PSI-BLAST) (*31*), enabled the detection of more remote evolutionary relationships and more accurate sequence alignments. (iii) A strategy, called the "fragment assembly approach" (*32*–*35*), was developed that can often improve predictions when a similar sequence cannot be found in the PDB.

If the target protein's sequence is related to a sequence that is already in the PDB, predicting its structure is usually easy (Fig. 4). In such cases, target protein structures are predicted by using "template-based modeling" (also called homology modeling or comparative modeling). But when there is no protein in the PDB with a sequence resembling the target's, accurately predicting the structure of the target is much more difficult. These latter predictions are called "free modeling" (also called ab initio or de novo prediction). One of the most successful free-modeling techniques is fragment assembly, described below. Many prediction methods are hybrids, combining template-based modeling, fragment assembly, and other strategies.

In fragment assembly (*32*–*35*), a target protein sequence is deconstructed into small, overlapping fragments. A search of the PDB is performed to identify known structures of similar fragment sequences, which are then assembled into a full-length prediction. The qualities of fragments and their assemblies are assessed by using some form of scoring function that aims to select more native-like protein structures from among the many possible combinations. Problems of folding physics described above share more commonality with free modeling than with template-based modeling.

Since CASP6, although overall progress has slowed (Fig. 4A), there has been systematic, incremental progress (*36*). The best groups can now on average produce models that are better than the single best template from the PDB. Progress has been made toward successfully combining multiple templates into a single prediction. Substantial improvements have been observed for free-modeling targets shorter than 100 amino acids, although no single group yet consistently produces accurate models. Larger free-modeling targets remain challenging. Several recent algorithmic developments—to predict residue-residue contacts from sequence alone (*37*–*39*) and to more sensitively and accurately identify remote homologs (*40*)—promise to further improve prediction accuracy.

The performance of two of the best fully automated server predictors during CASP9 (*41*) are shown in Fig. 4B: HHPred, a pure template-based modeling tool (*42*), and ROSETTA, a hybrid tool that combines fragment assembly and template-based modeling with all-atom refinement (*43*). For some fraction of CASP targets [~10%, based on a cutoff of 85 Global Distance Test–Total Score (GDT-TS) (*44*), which is defined in Fig. 4, legend], the best predictions are now accurate enough to interpret biological mechanisms, to guide biochemical studies, or to initiate a drug-discovery program (which requires structural errors of less than 2 to 3 Å). However, it remains a challenge to predict the other 90% of protein structures this accurately. In addition, it is critical to also improve physics-based technologies and to reduce our dependence on knowledge of existing structures, so that we can ultimately study protein motions, intrinsically disordered proteins, induced-fit binding of drugs, and membrane proteins and foldable polymers, for which databases are too limited.

Protein Folding: The Legacy of a Basic Science

Protein folding is a quintessential basic science. There has been no specific commercial target, yet the collateral payoffs have been broad and deep. Specific technical advances are reviewed elsewhere (*7*); below, we describe a few general outgrowths.

Growth of protein-structure databases. Today, more than 80,000 protein structures are known at atomic detail and publically available through the PDB. New structures are being added at a rapid pace, supported by the National Institutes of Health (NIH)–funded Protein Structure Initiative, which was developed in part to inform protein structure prediction.

Advances in computing technology. Understanding protein folding was a key motivation for IBM's development of the Blue Gene supercomputer (*45*), now also used to study the brain, materials, weather patterns, and quantum and nuclear physics. Protein folding has also driven key advances in distributed-grid computing, such as in Folding@home, developed by Pande at Stanford, in which computer users all over the world donate their idle computer time to perform physical simulations of protein systems (*46*). Folding@home, which now has more than one million registered users and an average of 200,000 user-donated CPUs available at any one time, provided some of the earliest simulations showing that MD simulations can accurately predict folding rates (*47*). The Anton computer from DE Shaw Research, custom designed to simulate biomolecules, gives several orders of magnitude better performance than conventional computers (*48*). Advances in computer technology have led to major advances in forcefields and to more reliable atomic-level insights into biological mechanisms.

Improvements in biomolecular forcefields. Computer processing power has advanced at the Moore's law rate, doubling every ~2 years. But equally important, forcefields have kept pace. Increased computer power leads to longer computed time scales, which puts more stringent demands on the accuracies of biomolecular forcefields. In a pioneering paper in 1977, McCammon *et al*.

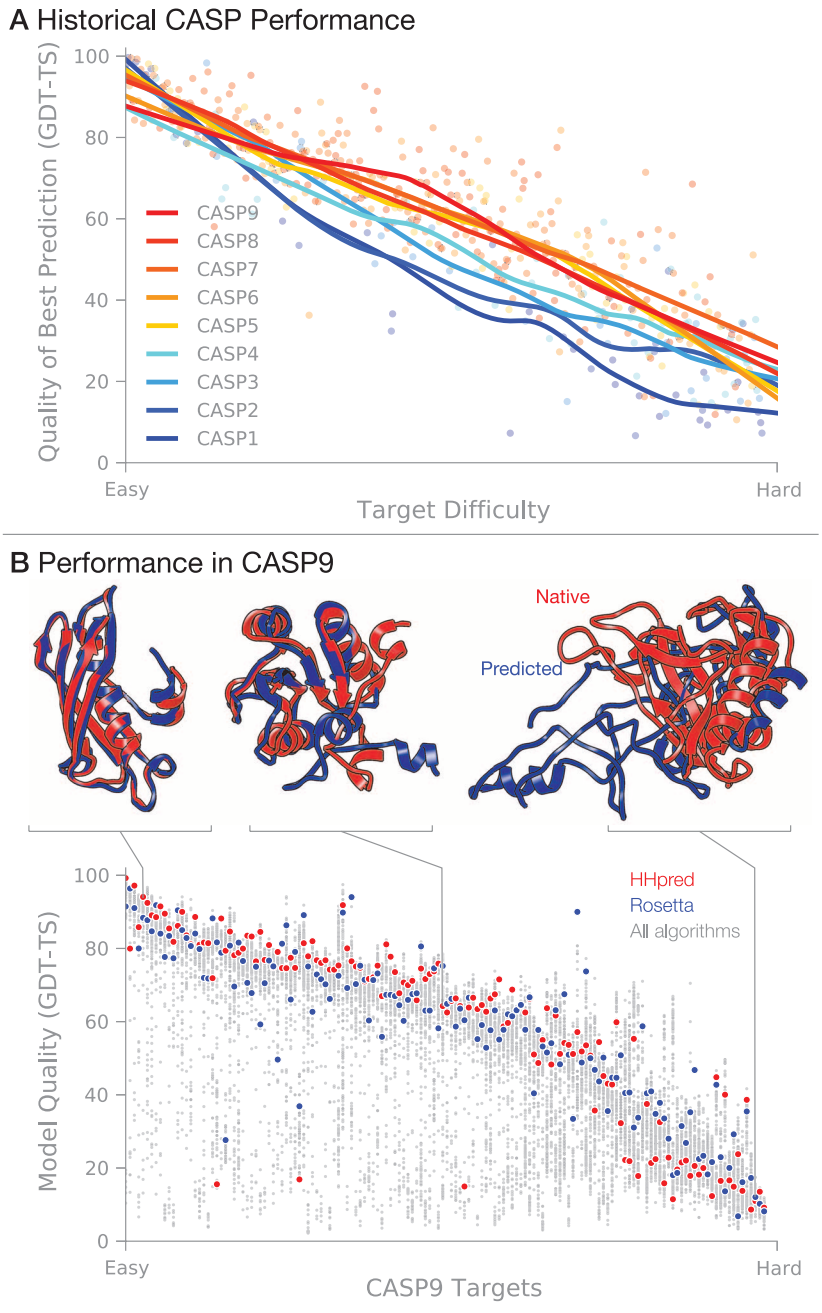

Fig. 4. Historical and present performance in CASP. Model quality is judged by using GDT-TS (*44*), which is approximately the percentage of residues that are located in the correct position. (**A**) Evolution of accuracy over the history of CASP, spanning 18 years. Each target is classified according to an approximate measure of difficulty that incorporates both the structural and sequence similarity to proteins of known structure (*36*). Each dot represents the best prediction (across all participants) for a given target. (**B**) Summary of prediction accuracy in CASP9 (*41*). We highlight the performance of two of the best automated server algorithms. Selected predictions are superimposed on the corresponding native structures to give a visual sense of the accuracy level that can be expected.

showed that the BPTI protein was stable in computer simulations during a computed time of 10 ps (*49*). Today, small proteins are typically stable in explicit-water simulations for 5 to 8 orders of magnitude longer—microseconds to milliseconds of computed time (*50*). Achieving such advances has required continuous improvements in force-field accuracy.

New sociological structures in the scientific enterprise. Protein folding has driven innovations in how science is done. CASP was among the first community-wide scientific competitions/collaborations, a paradigm for how grand-challenge science can be advanced through an organized communal effort. Other such competitions have followed, including Critical Assessment of Prediction of Interactions (CAPRI) (predicting protein-protein docking) (*51*), SAMPL (predicting small-molecule solvation free energies, and ligand binding modes and affinities) (*52*), and GPCR-Dock (predicting structures for G-protein coupled receptors, a pharmaceutically important category of membrane proteins) (*53*), among many others. Protein folding has also pioneered "citizen science," such as in Folding (*46*) and Robetta@home and in a computer game called Foldit (*54*), in which the public engages in protein folding on their home computers.

New materials: Sequence-specific foldable polymers. The principles and algorithms developed for protein folding have led to nonbiological, human-made proteins and to new types of polymeric materials. In particular, proteins have been designed that bind to and inhibit other proteins (fig. 5A) (*55*), have new folds (*56*), have new enzymatic activities (*57*), and act as potential new vaccines (*58*). Also, a class of nonbiological polymers has emerged, called "foldamers," that are intended to mimic protein structures and functions (*59*–*62*). Foldamers already have broad-ranging applications (*63*–*67*) as inhibitors of protein-protein interactions, broad-spectrum antibiotics, lung surfactant mimics, optical storage materials, a zinc-finger–like binder, an RNA-protein binding disrupter for application in muscular dystrophy, gene transfection agents, and "molecular paper" (Fig. 5B). Although such materials have potential applications in biomedicine and materials science, they also provide a way for us to test and deepen our understanding of protein folding.

Protein-folding diseases. Protein-folding research began before it was known that there are diseases of protein folding. Before 1972, it was assumed that all infectious diseases were transmitted through the DNA and RNA carried by viruses and bacteria. But Prusiner's studies of a patient with Creutzfeldt-Jakob disease (CJD) led to a previously unrecognized disease mechanism—namely, protein misfolding (*68*). Protein misfolding is now known to be important in many diseases, including CJD and type II diabetes, as well as neurodegenerative diseases such as Alzheimer's, Parkinson's, Huntington's, and amyotrophic lateral sclerosis. The protein-folding enterprise has provided important underpinnings for our understanding of folding diseases.

Unsolved Problems of Protein Physical Science

Is the protein-folding problem "solved" yet (*69*)? We believe it is no longer useful to frame the question this way. Protein folding is now a whole field of research—a large, growing, and diverse enterprise. A field of science—such as physics, chemistry, or biology—is bigger than any individual research question. A field is self-perpetuating; a few old puzzles generate more new puzzles. For the field of protein physical science, the future is at least as compelling as the past. Here are some of the unsolved problems:

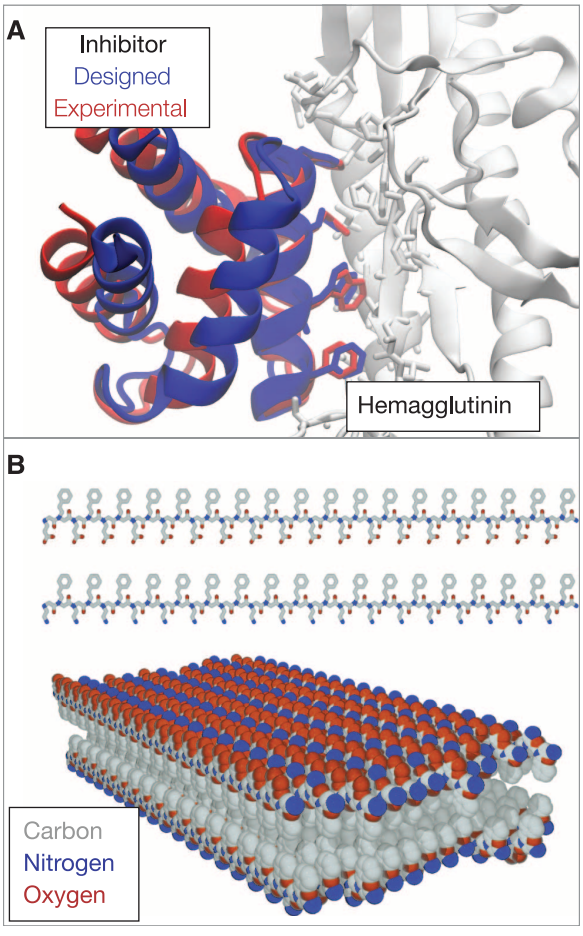

Fig. 5. Designed proteins and foldamers. (**A**) A protein inhibitor that was designed by computer to bind to hemagglutinin, an influenza protein. After design, the inhibitor was crystallized in a complex with hemagglutinin. The designed structure is in remarkably good agreement with experiment, particularly for the side chains involved in binding. [Adapted with permission (*55*)] (**B**) Peptoids are synthetic, foldable, protein-inspired polymers that have various applications. Shown here are peptoids that were designed as chains of alternating hydrophobic (gray) and either positively (blue) or negatively (red) charged side chains that spontaneously form a thin 2D structure called molecular paper. [Reprinted by permission (*67*)]

We have little experimental knowledge of protein-folding energy landscapes.

We cannot consistently predict the structures of proteins to high accuracy.

We do not have a quantitative microscopic understanding of the folding routes or transition states for arbitrary amino acid sequences.

We cannot predict a protein's propensity to aggregate, which is important for aging and folding diseases.

We do not have algorithms that accurately give the binding affinities of drugs and small molecules to proteins.

We do not understand why a cellular proteome does not precipitate, because of the high density inside a cell.

We know little about how folding diseases happen, or how to intervene.

Despite their importance, we still know relatively little about the structure, function, and folding of membrane proteins (*70*, *71*).

We know little about the ensembles and functions of intrinsically disordered proteins (*72*), even though nearly half of all eukaryotic proteins contain large disordered regions. This is sometimes called the "protein nonfolding problem" or "unstructural biology."

Summary

Fifty years ago, the protein-folding problem was born as a grand challenge of basic science. Since then, our understanding has advanced considerably. And, outgrowths of protein folding include the commercial development of new computers, such as IBM's Blue Gene; new modes of citizen science, including Folding@Home and Foldit; the development of communal scientific competitions, such as CASP; a database of now more than 80,000 protein structures; the Moore's-Law advancement in biomolecular simulation forcefields; new areas of materials science based on foldable polymers; and a foundation for understanding whole new classes of diseases—such as Alzheimer's, Parkinson's, and type II diabetes, called folding diseases—that were not even known when the protein-folding problem was first identified.

In times when there are pressures on science budgets for immediate payoffs, it is worth repeating the well-worn point that untargeted basic science often pays off in unexpected ways.

References and Notes

1. J. C. Kendrew *et al.*, *Nature* **181**, 662 (1958).
2. J. C. Kendrew *et al.*, *Nature* **185**, 422 (1960).
3. M. F. Perutz *et al.*, *Nature* **185**, 416 (1960).
4. C. M. Dobson, *Nature* **426**, 884 (2003).
5. A. R. Fersht, *Nat. Rev. Mol. Cell Biol.* **9**, 650 (2008).
6. J. E. Shea, C. L. Brooks III, *Annu. Rev. Phys. Chem.* **52**, 499 (2001).
7. K. A. Dill *et al.*, *Annu. Rev. Biophys.* **37**, 289 (2008).
8. G. R. Bowman *et al.*, *Curr. Opin. Struct. Biol.* **21**, 4 (2010).
9. H. M. Berman *et al.*, *Nucleic Acids Res.* **28**, 235 (2000).
10. K. A. Dill, *Biochemistry* **29**, 7133 (1990).
11. L. Pauling *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **37**, 205 (1951).
12. K. A. Dill, *Biochemistry* **24**, 1501 (1985).
13. S. Kamtekar *et al.*, *Science* **262**, 1680 (1993).
14. J. W. Ponder, D. A. Case, *Adv. Protein Chem.* **66**, 27 (2003).
15. K. Lindorff-Larsen *et al.*, *Science* **334**, 517 (2011).
16. A. Raval *et al.*, *Proteins* **80**, 2071 (2012).
17. C. Levinthal, *J. Chim. Phys.* **65**, 44 (1968).
18. P. S. Kim, R. L. Baldwin, *Annu. Rev. Biochem.* **59**, 631 (1990).
19. C. R. Matthews, *Annu. Rev. Biochem.* **62**, 653 (1993).
20. J. D. Bryngelson, P. G. Wolynes, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 7524 (1987).
21. P. E. Leopold *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 8721 (1992).
22. J. D. Bryngelson *et al.*, *Proteins* **21**, 167 (1995).
23. K. Dill, H. S. Chan, *Nat. Struct. Biol.* **4**, 10 (1997).
24. R. Zwanzig *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 20 (1992).
25. S. W. Englander *et al.*, *Q. Rev. Biophys.* **40**, 287 (2008).
26. J. Hockenmaier *et al.*, *Proteins* **66**, 1 (2006).
27. V. A. Voelz, K. A. Dill, *Proteins* **66**, 877 (2007).
28. J. Moult, *Curr. Opin. Struct. Biol.* **15**, 285 (2005).
29. J. Moult *et al.*, *Proteins* **79**, (suppl. 10), 1 (2011).
30. A. G. Murzin *et al.*, *J. Mol. Biol.* **247**, 536 (1995).
31. S. F. Altschul *et al.*, *Nucleic Acids Res.* **25**, 3389 (1997).
32. D. T. Jones, *Proteins* **45**, 127 (2001).
33. D. T. Jones, L. J. McGuffin, *Proteins* **53**, (Suppl 6), 480 (2003).
34. K. T. Simons *et al.*, *Proteins* **37**, 171 (1999).
35. R. Bonneau *et al.*, *Proteins* **45**, 119 (2001).
36. A. Kryshtafovych *et al.*, *Proteins* **79**, (suppl. 10), 196 (2011).
37. F. Morcos *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **108**, E1293 (2011).
38. D. S. Marks *et al.*, *PLoS ONE* **6**, e28766 (2011).
39. D. T. Jones *et al.*, *Bioinformatics* **28**, 184 (2012).
40. M. Remmert *et al.*, *Nat. Methods* **9**, 173 (2011).
41. V. Mariani *et al.*, *Proteins* **79**, (suppl. 10), 37 (2011).
42. J. Söding *et al.*, *Nucleic Acids Res.* **33**, (Web Server), W244 (2005).
43. P. Bradley *et al.*, *Science* **309**, 1868 (2005).
44. A. Zemla, *Nucleic Acids Res.* **31**, 3370 (2003).
45. F. Allen *et al.*, *IBM Syst. J.* **40**, 310 (2001).
46. M. Shirts, V. S. Pande, *Science* **290**, 1903 (2000).
47. C. D. Snow *et al.*, *Nat. Struct. Mol. Biol.* **420**, 102 (2002).
48. D. E. Shaw *et al.*, *Commun. ACM* **51**, 91 (2008).
49. J. A. McCammon *et al.*, *Nature* **267**, 585 (1977).
50. D. E. Shaw *et al.*, *Science* **330**, 341 (2010).
51. J. Janin *et al.*, *Proteins* **52**, 2 (2003).
52. J. P. Guthrie, *J. Phys. Chem. B* **113**, 4501 (2009).
53. M. Michino *et al.*, *Nat. Rev. Drug Discov.* **8**, 455 (2009).
54. S. Cooper *et al.*, *Nature* **466**, 756 (2010).
55. S. J. Fleishman *et al.*, *Science* **332**, 816 (2011).
56. B. Kuhlman *et al.*, *Science* **302**, 1364 (2003).
57. J. B. Siegel *et al.*, *Science* **329**, 309 (2010).
58. M. L. Azoitei *et al.*, *Science* **334**, 373 (2011).
59. S. H. Gellman, *Acc. Chem. Res.* **31**, 173 (1998).
60. W. S. Horne, S. H. Gellman, *Acc. Chem. Res.* **41**, 1399 (2008).
61. K. Kirshenbaum *et al.*, *Curr. Opin. Struct. Biol.* **9**, 530 (1999).
62. B.-C. Lee *et al.*, *J. Am. Chem. Soc.* **127**, 10999 (2005).
63. S. A. Fowler, H. E. Blackwell, *Org. Biomol. Chem.* **7**, 1508 (2009).
64. M. T. Dohm *et al.*, *Curr. Pharm. Des.* **17**, 2732 (2011).
65. B.-C. Lee *et al.*, *J. Am. Chem. Soc.* **130**, 8847 (2008).
66. N. H. Shah, K. Kirshenbaum, *Org. Biomol. Chem.* **6**, 2516 (2008).
67. K. T. Nam *et al.*, *Nat. Mater.* **9**, 454 (2010).
68. S. B. Prusiner, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 13363 (1998).
69. R. F. Service, *Science* **321**, 784 (2008).
70. Y. Arinaminpathy *et al.*, *Drug Discov. Today* **14**, 1130 (2009).
71. J. L. MacCallum, D. P. Tieleman, *Trends Biochem. Sci.* **36**, 653 (2011).
72. V. N. Uversky, A. K. Dunker, *Biochim. Biophys. Acta* **1804**, 1231 (2010).

Acknowledgments: We thank the NIH for support from NIH GM34993. We thank D. Baker, H. S. Chan, J. Chodera, W. Englander, D. Farrell, C. Fennell, S. Gellman, L. Gierasch, C. R. Matthews, D. Mobley, J. Moult, G. Rocklin, G. Rollins, and R. Zuckermann for helpful comments.

10.1126/science.1219021