Protein design holds promise for applications such as the control of cells, therapeutics, new enzymes and protein-based materials. Recently, there has been progress in rational design of protein molecules, and a lot of attempts have been made to create proteins with functions of our interests. The key to the progress is the development of methods for controlling desired protein tertiary structures with atomic-level accuracy. A theory for protein folding, the consistency principle, proposed by Nobuhiro Go in 1983, was a compass for the development. Anfinsen hypothesized that proteins fold into the free energy minimum structures, but Go further considered that local and non-local interactions in the free energy minimum structures are consistent with each other. Guided by the principle, we proposed a set of rules for designing ideal protein structures stabilized by consistent local and non-local interactions. The rules made possible designs of amino acid sequences with funnel-shaped energy landscapes toward our desired target structures. So far, various protein structures have been created using the rules, which demonstrates significance of our rules as intended. In this review, we briefly describe how the consistency principle impacts on our efforts for developing the design technology.

Key words: computational protein design, ideal proteins, funnel-shaped energy landscapes, consistent local and non-local interactions, Go model

Understanding of protein folding is important to develop the methodology for creating our desired proteins. Anfinsen hypothesized that proteins fold into the free energy minimum structures [1]. However, the folding problem—How do amino acid sequences determine the folded structures?—has been a long-standing problem for more than a half century. Researches for protein folding or structure prediction from amino acid sequences have attempted to address the problem by studying complicated proteins created by nature spending billions of years, which have energetically unfavorable non-ideal features such as kinked α-helices, bulged β-strands and buried polar residues. Protein design studies provide an alternative approach to tackle the problem by creating simple protein structures not having such unfavorable features from scratch with hypotheses about protein folding and experimentally testing how the designs fold.

Corresponding authors: Rie Koga, Protein Design Group, Exploratory Research Center on Life and Living Systems (ExCELLS), National Institutes of Natural Sciences, Okazaki, Aichi 444-8585, Japan. e-mail: riekoga@ims.ac.jp; Nobuyasu Koga, Protein Design Group, Exploratory Research Center on Life and Living Systems, National Institutes of Natural Sciences, Okazaki, Aichi 444-8585, Japan. e-mail: nkoga@ims.ac.jp
Protein design in early days

Protein design work was started in the late 1980s by the design of helical bundle structures. DeGrado, W. H., et al. [2] attempted to design dimeric helical bundle structures based on a hydrophobic (H) and hydrophilic (P) amino acid sequence pattern using the helical wheel model. Hecht, M. H., et al. [3] tried to design a four helix bundle, taking into account various structural features for helical proteins, derived from statistics of known protein structures, such as favorable amino acid types in α-helix or at the N- and C-terminal α-helix capping positions. However, these designed proteins were experimentally found to be in a molten globule state, in which proteins are compact with native-like secondary structure contents but without tight core packing [4]. The designs of TIM-barrel fold were also challenged by considering sequence preference for each residue position based on natural TIM-barrel proteins, but also found to be in a molten globule state [5]. All these efforts in early days tell us that protein folding is not determined just by simple amino acid patterning such as HP pattern: detailed atomistic modeling of tertiary structures is essential for designing amino acid sequences that have folding ability toward a unique tertiary structure. For example, the core in natural protein structures is apparently made of densely packed hydrophobic atoms. Actually, Hecht, M. H., et al. [3] tried to achieve the packed hydrophobic core in the design using a physical model, but it would have been difficult to capture the atomistic detail by hand.

Computational protein design from sidechain-redesign to full-scratch

The pioneer work of computational protein design with atomic resolution modeling was done by Dahiyat, B. I., et al. in the late 1990s, focusing on the redesign of sidechains of a naturally occurring protein structure using the backbone as a scaffold [6]. The group redesigned sidechains of zinc finger domain by stripping off the native sidechains of the protein and rebuilding new sidechains (amino acids) with a set of discretely represented sidechain conformations (rotamer library): new sidechain conformations that have the lowest energy for the zinc finger backbone were explored with the rotamer library. The design was found to form a compact well-ordered structure of zinc finger domain in the solution NMR structures, in which the packing of the hydrophobic core was similar to the design model. Since then, successful sidechain-redesigns of natural proteins have been reported for lambda-Cro [7], tenasin [8], homeodomain [9], etc. In these days, various designs such as novel enzymes [10], an influenza binder [11], and cage-like symmetric oligomers [12–15] were created using naturally occurring proteins as scaffolds, which can be considered as applications of the rotamer-based sidechain designs.

The first de novo design of a globular protein structure was achieved by Kuhlman, B., et al. in 2003 [16]. The authors created a novel protein fold, Top7, from scratch (Fig. 4). In this study, the authors developed a protocol in the software, Rosetta, to design protein structures from the backbone, in which the backbone structure of Top7 was built by assembling short fragments of known protein structures [17] and then sidechains that stabilize the built backbone were explored by an iterative approach between the rotamer-based sidechain design for a fixed backbone and gradient-based optimization of the entire structure for a fixed sequence [16]. The developed protocol enabled to identify sidechain-backbone pairs that have very low energies in computation, and one of the sequences of the pairs was found to have folding ability to the designed Top7 structure with atomic-level accuracy. Since then, however, no one made a success of the de novo design of protein structures until our work, indicating that the protocol was not enough to design protein structures. For designing proteins folded into the desired structures, there should be other essential factors than exploring low energy structures with tight hydrophobic core packing. Indeed, the paper did not describe how the lengths of the secondary structures and loops in the Top7 structure were determined. If we can design proteins with well-packed low energy structures whatever lengths of the secondary structures and loops are used, are the designs foldable? We started our work [18] in Baker group by investigating the folding abilities of proteins depending on the lengths of secondary structures and loops with folding simulations and statistical analysis of naturally occurring protein structures. Before describing the work, some hypotheses suggested by protein folding studies need to be introduced, which were the compass for developing our design methods.

Funnel-shaped energy landscape

The theoretical studies for protein folding from the late 1980s to 1990s suggested a hypothesis that natural proteins have evolved to have funnel-shaped energy landscapes toward the native state from the denatured state, in which proteins decrease their energies along with the formation of the native state from the denatured state, in which proteins decrease their energies along with the formation of the native state. This funnel hypothesis is supported by the fact that protein folding studies using Go-like models can explain the folding mechanisms (cooperative folding-unfolding transition, pathways, rates, etc.) found by experiments for small proteins [23–28]. The original Go model (not Go-like), a lattice model, proposed by Go [29] to embody the consistency principle [30] was applied to study the cooperative protein folding-unfolding transition, not in the context of the above energy landscape discussion. In either case of the Go or Go-like models, the essence of the
Go's consistency principle and local backbone preference for shaping funnel

Go proposed a hypothesis for protein folding, the consistency principle [30]. He considered that the local and non-local interactions are consistent with each other, in which local (non-local) interactions are those between amino acids close (distant) along a sequence. For example, if a tertiary structure is stabilized by non-local interactions such as hydrophobic and vdw interactions, but has local steric clashes or amino acids with low propensity for their secondary structures, the interactions of the tertiary structure are regarded as inconsistent. For the consistent case, all the local and non-local interactions consistently stabilize the tertiary structure with each other. The Go’s consistency principle is a paradigm shift for protein folding next to the Anfinsen’s thermodynamic principle: a folded structure has consistent interactions as well as corresponding to a free energy minimum. Indeed, as of 2019, more than 150,000 structures are deposited in the protein structure database (PDB), and when observing the structures, various interactions are surprisingly consistent: for example, buried polar groups without making hydrogen bonds are very rare [32]. About 20 years after the Go’s consistency principle, Chikenji, G., et al. discussed how funnel-shaped energy landscapes arise, using exact enumeration with a HP lattice model [31]. As described above, the Go model considers only the specific interactions formed in the native structure as an ideal limit to satisfy the consistency principle and has smooth-funneled energy landscape. On the other hand, the HP model used in their study has nonspecific nonlocal hydrophobic interactions and has a rugged energy landscape with multiple stable non-native structures. The authors demonstrated that by introducing local interactions into the HP model through a prohibition of one conformation for each local sequence (Fig. 2b), the rugged energy landscape get sculpted to be funneled into the native structure, in which any disallowed local conforma-

Go’s consistency principle and local backbone preference for shaping funnel

Go proposed a hypothesis for protein folding, the consistency principle [30]. He considered that the local and non-local interactions are consistent with each other, in which local (non-local) interactions are those between amino acids close (distant) along a sequence. For example, if a tertiary structure is stabilized by non-local interactions such as hydrophobic and vdw interactions, but has local steric clashes or amino acids with low propensity for their secondary structures, the interactions of the tertiary structure are regarded as inconsistent. For the consistent case, all the local and non-local interactions consistently stabilize the tertiary structure with each other. The Go’s consistency principle is a paradigm shift for protein folding next to the Anfinsen’s thermodynamic principle: a folded structure has consistent interactions as well as corresponding to a free energy minimum. Indeed, as of 2019, more than 150,000 structures are deposited in the protein structure database (PDB), and when observing the structures, various interactions are surprisingly consistent: for example, buried polar groups without making hydrogen bonds are very rare [32]. About 20 years after the Go’s consistency principle, Chikenji, G., et al. discussed how funnel-shaped energy landscapes arise, using exact enumeration with a HP lattice model [31]. As described above, the Go model considers only the specific interactions formed in the native structure as an ideal limit to satisfy the consistency principle and has smooth-funneled energy landscape. On the other hand, the HP model used in their study has nonspecific nonlocal hydrophobic interactions and has a rugged energy landscape with multiple stable non-native structures. The authors demonstrated that by introducing local interactions into the HP model through a prohibition of one conformation for each local sequence (Fig. 2b), the rugged energy landscape get sculpted to be funneled into the native structure, in which any disallowed local conforma-

models is an assumption to consider only native interactions formed in the native conformation as the energy gain ignoring non-native interactions, which makes the energy landscape smoothly funneled.

The funnel theory directly leads to the idea that for designing foldable proteins, it is essential to obtain amino acid sequences with funnel-shaped energy landscapes toward target structures. Such sequences could be acquired by exploring sequences that simultaneously stabilize the target structure and destabilize all of non-native structures. However, it is practically impossible to identify such sequences considering myriad non-native structures. How do we then find out such sequences? Clues for solving the question were the consistency principle proposed by Nobuhiro Go in 1983 [30] and the discussion by Chikenji, G., et al. on how funnel-shaped energy landscapes arise [31].

Go’s consistency principle and local backbone preference for shaping funnel

Go proposed a hypothesis for protein folding, the consistency principle [30]. He considered that the local and non-local interactions are consistent with each other, in which local (non-local) interactions are those between amino acids close (distant) along a sequence. For example, if a tertiary structure is stabilized by non-local interactions such as hydrophobic and vdw interactions, but has local steric clashes or amino acids with low propensity for their secondary structures, the interactions of the tertiary structure are regarded as inconsistent. For the consistent case, all the local and non-local interactions consistently stabilize the tertiary structure with each other. The Go’s consistency principle is a paradigm shift for protein folding next to the Anfinsen’s thermodynamic principle: a folded structure has consistent interactions as well as corresponding to a free energy minimum. Indeed, as of 2019, more than 150,000 structures are deposited in the protein structure database (PDB), and when observing the structures, various interactions are surprisingly consistent: for example, buried polar groups without making hydrogen bonds are very rare [32]. About 20 years after the Go’s consistency principle, Chikenji, G., et al. discussed how funnel-shaped energy landscapes arise, using exact enumeration with a HP lattice model [31]. As described above, the Go model considers only the specific interactions formed in the native structure as an ideal limit to satisfy the consistency principle and has smooth-funneled energy landscape. On the other hand, the HP model used in their study has nonspecific nonlocal hydrophobic interactions and has a rugged energy landscape with multiple stable non-native structures. The authors demonstrated that by introducing local interactions into the HP model through a prohibition of one conformation for each local sequence (Fig. 2b), the rugged energy landscape get sculpted to be funneled into the native structure, in which any disallowed local conforma-

models is an assumption to consider only native interactions formed in the native conformation as the energy gain ignoring non-native interactions, which makes the energy landscape smoothly funneled.

The funnel theory directly leads to the idea that for designing foldable proteins, it is essential to obtain amino acid sequences with funnel-shaped energy landscapes toward target structures. Such sequences could be acquired by exploring sequences that simultaneously stabilize the target structure and destabilize all of non-native structures. However, it is practically impossible to identify such sequences considering myriad non-native structures. How do we then find out such sequences? Clues for solving the question were the consistency principle proposed by Nobuhiro Go in 1983 [30] and the discussion by Chikenji, G., et al. on how funnel-shaped energy landscapes arise [31].

Go’s consistency principle and local backbone preference for shaping funnel

Go proposed a hypothesis for protein folding, the consistency principle [30]. He considered that the local and non-local interactions are consistent with each other, in which local (non-local) interactions are those between amino acids close (distant) along a sequence. For example, if a tertiary structure is stabilized by non-local interactions such as hydrophobic and vdw interactions, but has local steric clashes or amino acids with low propensity for their secondary structures, the interactions of the tertiary structure are regarded as inconsistent. For the consistent case, all the local and non-local interactions consistently stabilize the tertiary structure with each other. The Go’s consistency principle is a paradigm shift for protein folding next to the Anfinsen’s thermodynamic principle: a folded structure has consistent interactions as well as corresponding to a free energy minimum. Indeed, as of 2019, more than 150,000 structures are deposited in the protein structure database (PDB), and when observing the structures, various interactions are surprisingly consistent: for example, buried polar groups without making hydrogen bonds are very rare [32]. About 20 years after the Go’s consistency principle, Chikenji, G., et al. discussed how funnel-shaped energy landscapes arise, using exact enumeration with a HP lattice model [31]. As described above, the Go model considers only the specific interactions formed in the native structure as an ideal limit to satisfy the consistency principle and has smooth-funneled energy landscape. On the other hand, the HP model used in their study has nonspecific nonlocal hydrophobic interactions and has a rugged energy landscape with multiple stable non-native structures. The authors demonstrated that by introducing local interactions into the HP model through a prohibition of one conformation for each local sequence (Fig. 2b), the rugged energy landscape get sculpted to be funneled into the native structure, in which any disallowed local conforma-
strongly dependent on the lengths of the secondary structures and the connecting loop, not the detail of amino acid sequences, and that these dependencies are described in the simple rules (Fig. 3a). Our succeeding paper [34] further identified that the rules can be extended with the discretized backbone torsion bins, ABEGO, for the loops (torsion bins A and B are the α-helix and β-sheet regions; G and E are the positive phi regions equivalent to A and B; and O is the cis peptide bond). The major origin of the backbone structure preferences found in the rules is backbone strain arising from the polypeptide’s molecular geometry and the local steric hindrance in phi-psi angles of each residue. The discovered rules allow us to control protein topologies: selecting lengths or ABEGO patterns of the secondary structures and loops that favor the tertiary motifs present in the desired topology, many of the non-native topologies are disfavored by local backbone strain, resulting in funnel-shaped energy landscapes.

We have finally reached a general method to design amino acid sequences with funnel-shaped energy landscapes toward a unique structure [18]. De novo protein designs proceed in two steps: the backbone building [17] and the side-
Acknowledgements

We gratefully thank Nobuhiro Go, a pioneer for protein folding studies. His proposed concept, the consistency principle, has been the center of our research. Our protein design work was done in the Baker Group. We thank David Baker and all members in his group. We also thank all researchers who had been and continue to work on protein folding and design.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

R. K and N. K reviewed the protein design and folding studies and wrote the manuscript.

References

[1] Anfinsen, C. B. Principles that govern the folding of protein chains. *Science* **181**, 223–230 (1973).
[2] DeGrado, W. F., Regan, L., & Ho, S. P. The design of a four-helix bundle protein. *Cold Spring Harb. Symp. Quant. Biol.* **52**, 521–526 (1987).
