100TH ANNIVERSARY
REVIEW

Computational De Novo Peptide and Protein Design: Rigid Templates versus Flexible Templates

Ho Ki Fung,[†] William J. Welsh,[‡] and Christodoulos A. Floudas*,[†]

Department of Chemical Engineering, Princeton University, Princeton, New Jersey 08544-5263, and Department of Pharmacology, University of Medicine & Dentistry of New Jersey (UMDNJ), Robert Wood Johnson Medical School, and the Informatics Institute of UMDNJ, Piscataway, New Jersey 08854

Both rigid and flexible backbone design templates have been used in the numerous computational de novo peptide and protein design efforts reported so far. In this review paper, we use the type of templates (i.e., rigid or flexible) as a criterion to classify and review examples of successes in de novo protein design. For both cases of rigid and flexible templates, we briefly outline the different search methods for exploring the sequence space and quote some notable success examples for each search method. In particular, we divide the case of flexible templates into three subcategories, according to their approaches for incorporating backbone flexibility: (i) discrete rotamers on multiple backbones with fixed backbone assumption for each, (ii) discrete rotamers on a continuum backbone through algebraic parametrization, and (iii) continuum backbone template with continuous ranges of backbone angles.

1. Introduction

De novo protein design is initiated with a postulated or known flexible three-dimensional protein backbone structure and is intended for use to identify amino acid sequences compatible with such a structure. The problem was first denoted as the "inverse folding problem",[1,2] because protein design has intimate links to the well-known protein folding problem.[3] Although the objective of the protein folding problem is to determine the folded structure with the lowest free energy for a given amino acid sequence, the de novo protein design problem exhibits a high level of degeneracy; that is, a large number of sequences are always observed to share a common fold, although the sequences will vary, with respect to properties such as activity and stability.

Traditionally, protein design was performed using experimental techniques such as rational design, mutagenesis, and directed evolution.[4] Although capable of producing good results, they all entail the major drawback of being able to screen only a highly restricted number of mutants. It was estimated that the maximum size of amino acid sequence search space that these experimental approaches can handle is $\sim 10^3-10^6$.[5] On the other hand, the number of sequences through which computational de novo design methods can search is significantly larger. For instance, Gordon et al.[6] reported their redesign of the 74 core residues of the catalytic antibody (Protein Databank (PDB) code: 1HKL), which corresponded to a rotamer search space of 4.7×10^{128}: this is an unimaginable size for experimentalists. This feature has caused in silico de novo protein design approaches to gain popularity.

There have been a large number of reported successes in computational de novo protein design. A few representative examples include achievements in modulating protein−protein interactions,[7] promoting stability and specificity of the target protein,[8−14] and conferring novel binding sites or properties onto the template,[15,16] as well as locking proteins into certain useful conformations.[17,18] To a large extent, in silico methods elucidate protein folding kinetics[9] and protein−ligand interactions,[19,20] assist in protein−protein docking,[21] and most importantly, enhance our peptide and protein drug discovery process.[22]

Nevertheless, despite the large search spaces that computational methods can handle, and the fact that network flow structure may be embedded in computational protein design formulations,[23] full-sequence de novo design of a 100-residue protein is still considered challenging today. Taking an average of ∼100 rotamers for all 20 amino acids to be considered at each position,[24] the complexity of the problem amounts to an overwhelming level of $100^{100} = 10^{200}$. This high level of complexity couples with the NP-hard nature of the problem[25,26] to impose stringent demand on any sequence selection algorithm. Besides improving computational efficiency, as noted by Floudas et al.,[27] incorporating true protein backbone flexibility represents a key challenge in de novo protein design.

This paper reviews the research performed in the field of computational de novo peptide and protein design. It classifies the work according to the design templates that are used, which can be either rigid or flexible. In the latter case, backbone flexibility is incorporated through (i) the consideration of discrete rotamers on multiple discrete backbones with the fixed template assumption imposed on each backbone for de novo design, and

* To whom correspondence should be addressed. Tel.: (609) 258-4595. Fax: (609) 258-0211. E-mail: floudas@titan.princeton.edu.
[†] Department of Chemical Engineering, Princeton University.
[‡] Department of Pharmacology, University of Medicine & Dentistry of New Jersey (UMDNJ), Robert Wood Johnson Medical School, and the Informatics Institute of UMDNJ.

10.1021/ie071286k CCC: $40.75
Published on Web 01/09/2008

the combination of all the results at the end; or (ii) the consideration of discrete rotamers on a continuum template, which is made possible by the algebraic parametrization of the backbone; or (iii) the use of a continuum template where all possible continuous values of $C^{\alpha}-C^{\alpha}$ distances and dihedral angles bounded between upper and lower limits are taken into account for the design. The sequence search methods and successful applications using each method will be presented.

2. De Novo Peptide and Protein Design with Fixed Template

Computational protein design efforts were first initiated with the premise that the three-dimensional coordinates of the design template or backbone were fixed. This simplification was first proposed by Ponder and Richards,[28] and it was appealing because it greatly reduced the combinatorial complexity of the search. Together with consideration of only a limited set of the most frequently observed side-chain conformations (called rotamers),[24,29] the assumption enhanced the efficiency of the initial de novo design efforts, most of which focused on protein cores,[30−34] in exploring search spaces. The reason why protein cores were selected instead of the boundary or surface regions is based on the thesis that protein folding is primarily driven by hydrophobic collapse; thus, a good core has a tendency to provide a well-folded and stable structure for the de novo designed protein.[35] The scope of the de novo design encompassed intermediate and surface residues in subsequent years, and obviously the problem became more challenging. In this section, we outline the different deterministic and stochastic methods that search for sequences specific to the fixed rigid design template. Note that they all discretize the side-chain conformational space into rotamers for tractability of the search problem. After each method is introduced, we also review examples of successes.

2.1. Sequence Search Methods. De novo design algorithms can be classified into two main categories: deterministic and stochastic.[36] The two main methods that fall into the deterministic category are dead-end elimination (DEE) and self-consistent mean field (SCMF), whereas the two major stochastic-type frameworks are Monte Carlo and genetic algorithms (GAs). Some methods search for low-energy sequences, whereas others assign probability to each of the 20 amino acids for each design position in a sequence to maximize the conformational entropy.

2.1.1. Deterministic Methods. 2.1.1.1. The Dead-End Elimination (DEE) Criteria. DEE, which is arguably the most popular rotamer search algorithm now, operates on the systematic elimination of rotamers that cannot be parts of the sequence with the lowest energy. The energy function in DEE is written in the form of a sum of individual term (rotamer−template) and pairwise term (rotamer−rotamer):

$$E = \sum_{i=1}^{N} E(i_a) + \sum_{i=1}^{N-1} \sum_{j>i}^{N} E(i_a,j_b) \tag{1}$$

where $E(i_a)$ is the rotamer−template energy for rotamer i_a of amino acid i; $E(i_a,j_b)$ is the rotamer−rotamer energy of rotamers i_a and j_b of amino acids i and j, respectively; and N is the total number of positions. The original DEE pruning criterion is based on the concept that, if the pairwise energy between rotamers i_a and j_b is higher than that between rotamers i_c and j_b for all j_b in a certain rotamer set {B}, then i_a cannot be in the global energy minimum conformation and, thus, can be eliminated. This was proposed by Desmet et al.[37] and can be expressed in the following mathematical form:

$$E(i_a) + \sum_{j\neq i}^{N} E(i_a,j_b) > E(i_c) + \sum_{j\neq i}^{N} E(i_c,j_b) \qquad \forall\{\mathrm{B}\} \tag{2}$$

Rotamer i_a can be pruned if the previous expression holds true. Bounds implied by expression 2 can be utilized to generate the following computationally more tractable inequality:[37]

$$E(i_a) + \sum_{j\neq i}^{N} \min_b E(i_a,j_b) > E(i_c) + \sum_{j\neq i}^{N} \max_b E(i_c,j_b) \tag{3}$$

The aforementioned equations for eliminating rotamers at a single position (or singles) can be extended to eliminating rotamer pairs at two distinct positions (doubles), rotamer triplets at three distinct positions (triples), or above.[37,38] In the case of doubles, the equation becomes

$$\epsilon(i_a,j_b) + \sum_{k\neq i,j}^{N} \min_c \epsilon(i_a,j_b,k_c) > \epsilon(i_{a'},j_{b'}) + \sum_{k\neq i,j}^{N} \max_c \epsilon(i_{a'},j_{b'},k_c) \tag{4}$$

where ϵ is the total energies of the rotamer pairs:

$$\epsilon(i_a,j_b) = E(i_a) + E(j_b) + E(i_a,j_b) \tag{5}$$

$$\epsilon(i_a,j_b,k_c) = E(i_a,k_c) + E(j_b,k_c) \tag{6}$$

This parameter determines a rotamer pair i_a and j_b that always contributes higher energies than rotamer pair $i_{a'}$ and $j_{b'}$ for all possible rotamer combinations.

Goldstein[39] improved the original DEE criterion by stating that rotamer i_a can be pruned if the energy contribution is always reduced by an alternative rotamer i_c:

$$E(i_a) - E(i_c) + \sum_{j\neq i}^{N} \min_b [E(i_a,j_b) - E(i_c,j_b)] > 0 \tag{7}$$

This can be generalized to the use of a weighted average of C rotamers i_c to eliminate i_a:[39]

$$E(i_a) - \sum_{c=1,\ldots,C} w_c E(i_c) + \sum_{j\neq i}^{N} \min_b [E(i_a,j_b) - \sum_{c=1,\ldots,C} w_c E(i_c,j_b)] > 0 \tag{8}$$

Lasters et al.[40] proposed that the most suitable weights w_c can be determined by solving a linear programming problem.

In addition to these criteria that were proposed by Goldstein,[39] Pierce et al.[38] introduced the Split DEE, which splits the conformational space into partitions and thus eliminates the dead-ending rotamers more efficiently:

$$E(i_a) - E(i_c) + \sum_{j,j\neq k\neq i}^{N} \{\min_{a'}[E(i_a,j_{a'}) - E(i_c,j_{a'})]\} + [E(i_a,k_{b'}) - E(i_c,k_{b'})] > 0 \tag{9}$$

Generally, n splitting positions can be assigned for more efficient but more computationally expensive rotamer elimination:

$$E(i_a) - E(i_c) + \sum_{j,j\neq k_1,\ldots,k_n\neq i}^{N} \{\min_{a'} [E(i_a,j_{a'}) - E(i_c,j_{a'})]\} + \sum_{k=k_1,\ldots,k_n} [E(i_a,k_{b'}) - E(i_c,k_{b'})] > 0 \tag{10}$$

Looger and Hellinga[41] also introduced the generalized DEE by ranking the energy of rotamer clusters, instead of individual rotamers, and they increased the ability of the algorithm to deal with higher levels of combinatorial complexity. Further revisions and improvements on DEE had been performed by Wernisch et al.,[42] Gordon et al.,[6] and Georgiev et al.[130]

Being deterministic in nature, the different forms of DEE reviewed previously all yield the same globally optimal solution upon convergence.

Successes Using Dead-End Elimination: Based on operating the DEE algorithm on a fixed template, the Mayo group devised their optimization of rotamers by iterative techniques (ORBIT) program and applied it to numerous de novo protein designs. Examples are the full-sequence design of the $\beta\beta\alpha$ fold of a zinc finger domain,[43] improvement of calmodulin binding affinity,[44] full core design of the variable domains of the light and heavy chains of catalytic antibody 48G7 FAB, full core/boundary design, full surface design, and full-sequence design of the β1 domain of protein G[6], as well as the redesign of the core of T4 Lysozyme.[33] They also adjusted secondary structure parameters to build the "idealized backbone" and used it as a fixed template to design an α/β-barrel protein.[45] The Hellinga group applied DEE with fixed backbone structure to introduce iron and oxygen binding sites in thioredoxin,[46,47] design receptor and sensor proteins with novel ligand-binding functions,[48] and confer novel enzymatic properties onto ribose-binding protein.[49]

2.1.1.2. The Self-Consistent Mean Field (SCMF) Method. The SCMF optimization method is an iterative procedure that predicts the values of the elements of a conformational matrix $P(i,a)$ for the probability of a design position i adopting the conformation of rotamer a. Note that $P(i,a)$ sums to unity over all rotamers a for each position i. Koehl and Delarue[50] were among the groups who introduced such a method for protein design. They started the iteration with an initial guess for the conformational matrix, which assigns equal probability to all rotamers:

$$P(i,a) = \frac{1}{A} \qquad (\text{for } a = 1, 2, ..., A) \tag{11}$$

Most importantly, they applied the mean field potential, $E(i,a)$, which is dependent on the conformational matrix $P(i,a)$:

$$E(i,a) = U(x_{ia}) + U(x_{ia},x_o) + \sum_{j=1,j\neq i}^{N}\sum_{b=1}^{B} P(j,b)U(x_{i_a},x_{j_b}) \tag{12}$$

where x_o corresponds to the coordinates of atoms in the fixed template, and x_{i_a} and x_{j_b} correspond to the coordinates of the atoms of position i, assuming the conformation of rotamer a and those of position j assuming the conformation of rotamer b, respectively. The classical Lennard-Jones (12−6) potential can be used to describe potential energy (U).[50] The conformational matrix can be subsequently updated using the mean field potential and the Boltzmann law:

$$P_1(i,a) = \frac{\exp[-E(i,a)/(RT)]}{\sum_{a=1}^{A}\exp[-E(i,a)/(RT)]} \tag{13}$$

The updated $P(i,a)$, namely $P_1(i,a)$, can then be used to repeat the calculation of mean field potential and another update can be obtained until convergence is attained. Koehl and Delarue[50] set the convergence criterion to be 10^{-4} to define self-consistency. They also proposed the introduction of memory of the previous step to minimize oscillations during convergence:

$$P(i,a) = \lambda P_1(i,a) + (1-\lambda)P(i,a) \tag{14}$$

with an optimal step size of $\lambda = 0.9$.[50]

The Saven group extended the SCMF theory and formulated de novo design as an optimization problem, maximizing the sequence entropy, subject to composition constraints and mean-field energy constraints.[51−54] In addition to the site probabilities, their method also predicts the number of sequences for a combinatorial library of arbitrary size for the fixed template as a function of energy.

It should be highlighted that, although deterministic in nature, the SCMF method does not guarantee convergence to the global optimal solution.[55]

Successes Using the Self-Consistent Mean Field Method: Koehl and Delarue[56] applied the SCMF approach to design protein loops. In their optimization procedure, they first selected the loop fragment from a database with the highest site probabilities. They then placed side chains on the fixed-loop backbone from a rotamer library. Kono and Doi[57] also used an energy minimization with automata network, which bears some resemblance to the SCMF method, to design the cores of the globular proteins of cytochrome b_{562}, triosephosphate isomerase, and barnase. SCMF is related to the design of combinatorial libraries of new sequences with good folding properties, which was reviewed by several papers.[58−61]

2.1.2. Stochastic Methods. The fact that de novo design is NP-hard[25,26] means that, in the worst case, the time required to solve the problem scales nonpolynomially with the number of design positions. As the problem complexity exceeds a certain level, deterministic methods may reach their limits and, in such instances, we may be forced to resort to stochastic methods, which perform searches for only locally optimal solutions. Monte Carlo methods and genetic algorithms are the two most commonly used types of stochastic methods for de novo protein design.

2.1.2.1. Monte Carlo Methods. Different variants of the Monte Carlo methods have been applied for sequence design. In the classic Monte Carlo method, mutation is performed at a certain position in the sequence and energies of the sequence in the fixed template are calculated before and after the mutation. This usually involves the use of discrete rotamer libraries to simplify the consideration of possible side-chain conformations. The new sequence after mutation is accepted if the energy becomes lower. If the energy is higher, the Metropolis acceptance criterion[62] is used:

$$p_{\text{accept}} = \min(1, \exp(-\beta\Delta E))\ \beta = \frac{1}{kT} \tag{15}$$

The sequence is updated if p_{accept} is larger than a random number uniformly distributed between 0 and 1.

In the configurational bias Monte Carlo method, at each step, a local energy is used that does not include those positions where a mutation has not been attempted.[63] Cootes et al.[64] reported that the method was more efficient in finding good solutions than the conventional Monte Carlo method, especially for complex systems. Zou and Saven[63] also devised the mean-field biased Monte Carlo method, which biases the sequence search with predetermined site probabilities, which, in turn, are calculated using SCMF theory. They claimed their new method converges to low-energy sequences faster than classic Monte Carlo and configurational bias Monte Carlo methods.

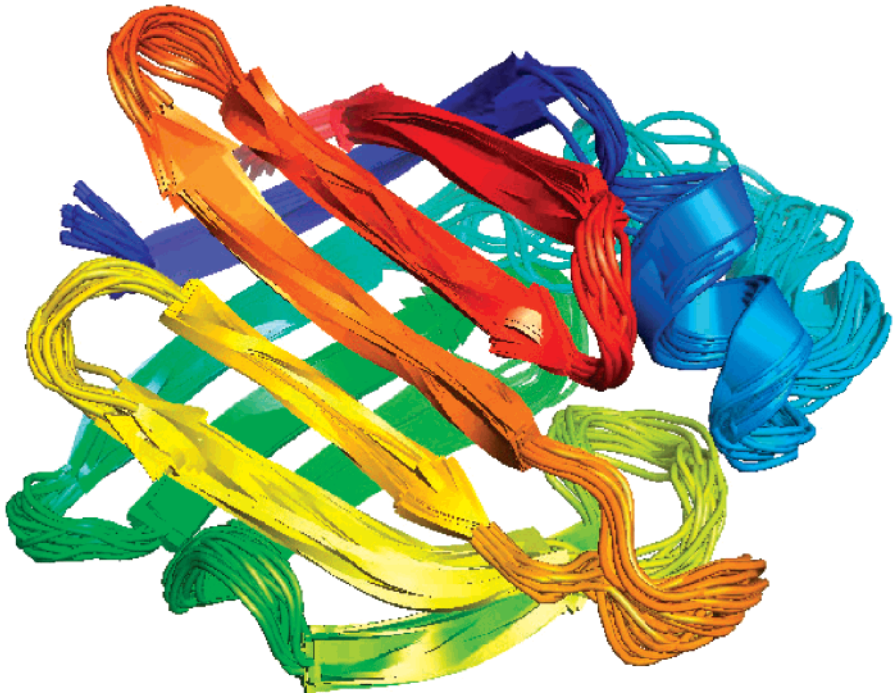

Figure 1. Template flexibility, as illustrated by the superposition of the 20 nuclear magnetic resonance (NMR) structures of apo intestinal fatty acid-binding protein (Protein Databank (PDB) Code: 1AEL).

Successes of Monte Carlo Methods: Imposing sequence specificity by keeping the amino acid composition fixed, which reduced significantly the complexity, Koehl and Levitt[65,66] designed new sequences for the fixed backbones of the β1 domain of protein G, λ repressor, and sperm whale myoglobin, using the conventional Monte Carlo method. The Baker group also utilized the classic Monte Carlo algorithm in their computational protein design program, RosettaDesign. Examples of applications of the program include the redesign of nine globular proteins—the src SH3 domain, λ repressor, U1A, protein L, tenascin, procarboxypeptidase, acylphosphatase, S6, and FKBP12—using fixed templates.[67]

2.1.2.2. Genetic Algorithms. Originated in genetics and evolution, genetic algorithms generate a multitude of random amino acid sequences and exchange them for a fixed template. Sequences with low energies form hybrids with other sequences, whereas those with high energies are eliminated in an iterative process that terminates only when a converged solution is attained.[68]

Successes of Genetic Algorithms: With fixed backbones, Belda et al.[69] applied genetic algorithms to the design of ligands for prolyl oligopeptidase, p53, and DNA gyrase. In addition, with a cubic lattice and empirical contact potentials,[70] Hohm et al.[71] also applied evolutionary methods to design short peptides that resemble the antibody epitopes of thrombin and blood coagulation factor VIII with high stability.

3. De Novo Peptide and Protein Design with Flexible Template

The assumption of fixed template for de novo peptide and protein design is highly questionable,[72] because protein is commonly known to exhibit backbone flexibility, as illustrated by the superposition of NMR structures in Figure 1. De novo design templates were observed to allow residues that would not have been permitted had the backbone been fixed.[73] The Mayo group claimed that their ORBIT protein design program was robust against 15% change in the backbone.[33] Nevertheless, they found, in a later case study on T4 lysozyme, that core repacking to stabilize the fold was difficult to achieve without considering a flexible template.[33] The secondary structures of α-helices and β-sheets actually display twisting and bending in the fold, and Emberly et al.[74,75] had applied principal component analysis of database protein structures to quantify the degree and modes of their flexibilities.

In this section, we classify the various methodologies of incorporating backbone flexibility into the design template into three main types according to their treatment of the backbone and side-chain conformations. The first type involves considering a set of multiple discrete templates and performing de novo design with discrete rotamers on each of the templates under the fixed-backbone assumption. The second type considers a continuum template by means of algebraic parametrization of the backbone and variation of the parameters to allow for backbone movement during sequence selection. However, it still uses rotamer libraries to simplify the side-chain conformations. Through novel sequence selection formulations[76] and pairwise contact potentials, which are discretized over distance bins,[77−79] the third type considers a continuum design template in which the Cα−Cα distances and dihedral angles assume continuous values between upper and lower bounds observed from the template structures,[80] and it confirms sequence specificity to the target fold, based on these bounded continuous distances and angles via nuclear magnetic resonance (NMR) structure refinement methods,[81,82] rather than the discrete rotamer approach. It should be emphasized that the third type is the most general case for treating the flexible template, because it allows for all possible meaningful combinations of distances and dihedral angles within their defined lower and upper bounds (i.e., semi-infinite set). Similar frameworks were introduced for defining flexibility in chemical processes (e.g., see the work of Floudas and co-workers[83−85]). For each category, we will quote some examples of successes of de novo peptide and protein design.

3.1. Flexible Template via Multiple Discrete Templates and Discrete Rotamers. By incorporating protein backbone flexibility via discrete templates and discrete rotamers, de novo protein design frameworks either separate sequence selection and backbone movement explicitly or iterate between sequence space and structure space.[86] Note that, in both cases, the sequence search methods outlined in the previous section are all applicable, because fixed backbones and discrete rotamers are still assumed.

3.1.1. Approaches That Separate Sequence Selection and Backbone Movement. These approaches consider an ensemble of fixed backbones, search for sequences for each of them (assuming a fixed template), and finally identify the best solutions from all the results. Successes using different types of search algorithms include successes using dead-end elimination, the self-consistent mean field (SCMF) method, and Monte Carlo methods/genetic algorithms.

With regard to successes using dead-end elimination, by varying the supersecondary structure parameters, Su and Mayo[87] and Ross et al.[88] generated several sets of perturbed backbones from the native structure and redesigned the core of the β1 domain of the streptococcal protein using the DEE algorithm, under the fixed template assumption for each backbone. As confirmed by NMR experiments, six of the seven sequences tested folded into nativelike structures.

With regard to successes using the SCMF method, Kono and Saven[53] applied their SCMF-based protein combinatorial library design strategy on a set of similar backbone structures to obtain new sequences that are robust to distance changes in the template for the immunoglobulin light-chain binding domain of protein L.

With regard to the successes of Monte Carlo methods/genetic algorithms, the Pande group generated families of 100 fixed templates within a rootmean square deviation (rmsd) of 1 Å from the initial backbone, using Monte Carlo method. With these fixed template ensembles, they performed de novo design, which was based on genetic algorithms, on their Genome@home distributed grid system for 253 naturally occurring proteins. They obtained sequences that exhibited higher diversity than the corresponding natural sequence alignments, as well as good agreement on the sequence entropies of the designed sequences from the same template family.[89,90]

To incorporate protein flexibility, Kraemer-Pecore et al.[91] executed a Monte Carlo simulation to generate 30 fixed backbones that are within 0.3 Å rmsd of the initial template. A genetic-algorithm-based sequence prediction algorithm (SPA),[92] which combines filtering and sampling rotamers and energy minimization, was then used for sequence search on each template, under the fixed backbone assumption. The work led to the identification of a sequence that folded into the WW domain.

In designing protein conformational switches, Ambroggio and Kuhlman[93,94] also used the Monte Carlo-based RosettaDesign to search for sequences for multiple fixed-template structures.

3.1.2. Approaches That Iterate between Sequence Space and Structure Space. There are two good examples that belong to this class. The first example is a genetic algorithm/Monte Carlo-based framework used by Desjarlais and Handel,[95] in which a starting population of backbones is generated by small-angle perturbations to the template, rotamers are randomly selected on each backbone, and a genetic algorithm is subsequently used that exchanges not only rotamers but also backbone torsional information in recombination. The framework is ended with a Monte Carlo stage, which refines the backbone structures. Using this novel approach, Desjarlais and Handel[95] designed three new core variants of the protein 434 cro. They also compared results on 434 cro and T4 lysozyme with those obtained earlier, using fixed-template models, and they found that they were similar, given that the fixed-template models scan over a much larger rotamer space.

The second example is that proposed by Kuhlman et al.[96] and Saunders and Baker.[97] Their method starts with a set of initial backbones, uses the Monte Carlo method to search for the sequence with the lowest energy for each of them, performs atomic-resolution structure prediction for the sequences to allow shifts in the structure space, and continues until the number of iterations hits a predetermined number. They successfully designed a new sequence for Top 7, which is a 93 residue α/β protein with a novel fold.[96] They also claimed that the new method better captures sequence variation than approaches that separate sequence selection and backbone movement explicitly.

3.2. Flexible Template via Continuum Template and Discrete Rotamers. This method of constituting a continuum template via backbone parametrization and performing sequence search from rotamer libraries is proposed by Harbury and co-workers.[98−100] Based on the algebraic parametrization equations developed for coiled coils by Crick,[101] their method allowed backbone movement by treating the parameters as variables during sequence search for energy minimization, which, in turn, is performed by the local optimization methods of steepest descents (SD) minimization and adopted-basis Newton−Raphson (ABNR) minimization.

Successes: Harbury and co-workers[98−100] adopted this approach to design a family of α-helical bundle proteins with a right-handed superhelical twist. The crystal structure of the designed sequences with the optimal specificity was experimentally validated to match the design template.

3.3. Flexible Template via Continuum Template and Continuous Ranges of Backbone Angles. Considering discrete rotamers is certainly not the best approach to adopt in de novo design, because ∼15% of the side-chain conformations are not represented by common rotamer libraries.[102] A recent two-stage de novo design approach proposed by the Floudas group considers a continuum design template without using discrete rotamers for the possible side-chain conformations.[10,11,14,27] Their flexible design template was generated using molecular dynamics simulation with either generalized Born implicit solvation or explicit water molecules based on a standard force field (see Figure 2).[132]

The first stage selects a rank-ordered list of low-energy sequences using novel quadratic assignment-like models[26,76] driven by pairwise residue contact potentials, which were developed by the group by solving a linear programming parameter estimation problem, requiring that the native conformations for a large training set of 1250 proteins be ranked energetically more favorable than their high-resolution decoys.[77−79] The force fields developed were observed to produce very good Z-scores, in regard to recognizing the native folds for a large test set of proteins.[77−79] Rather than being continuous, the dependence of contact potential on distance is discretized into bins. This designed feature serves to make the energy objective function insensitive to a limited degree of backbone movement. For example, in the high-resolution Cα−Cα force field,[78] if the pair of amino acids selected at two positions *i* and *k*, which are 3.5 Å apart in the template, are ARG and GLU, respectively, their energy contribution to the objective function is −7.77 kcal/mol. Despite small distance variations, this energy value is constant for all ARG−GLU interactions, as long as the Cα

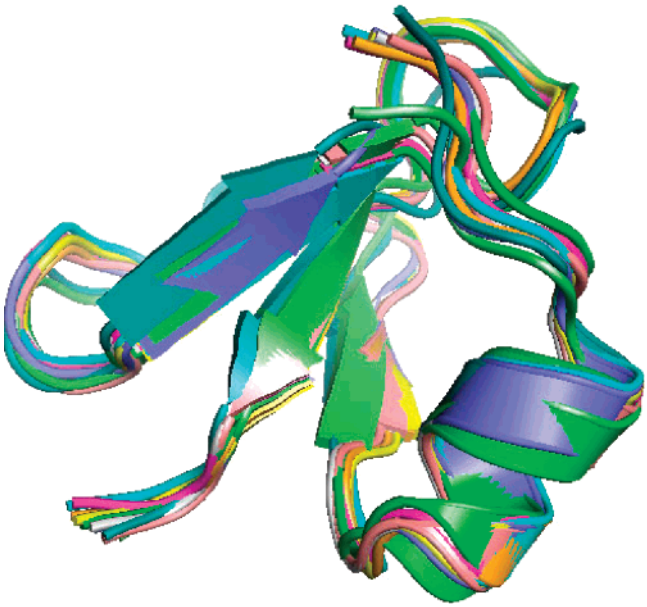

Figure 2. Flexible templates for the de novo design of human beta-defensin-2 generated from molecular dynamics (MD) simulations with the generalized Born implicit solvation model and CHARMM force-field (version 31b1).[129] The 10 structures shown correspond to 10 different snapshots, with increments of 1 ns, along the MD trajectory.

positions of the two residues are 3−4 Å (bin 1) apart. Two classes of force fields have been developed: (i) Cα−Cα distance-based force fields[77,78] and (ii) centroid−centroid distance-based force fields.[79] To perform sequence selection based on a flexible template of multiple structures, Fung et al.[76] also developed two novel formulations: a weighted model, which considers the distance between any two positions as the weighted average of their distances in all structures, and a binary distance bin model that decides which bin the distance falls into during energy optimization. The latter approach is, in a sense, similar to the backbone parametrization approach by Harbury and co-workers,[98−100] where there are distance variables associated with the backbone.

The second stage of the approach confirms fold specificity of the sequences generated in the first stage based on a full-atomistic forcefield. The group used to perform the task via ASTRO-FOLD,[103−111] which is a protein structure prediction methodology via global optimization.[112−127] Conformational ensembles are generated for each sequence under two sets of conditions. In the first circumstance, the structure is constrained to vary, with some imposed fluctuations, around the template structure. In the second condition, a free-folding calculation is performed for which only a limited number of restraints (e.g., disulfide bridges), but not the underlying template structure, are enforced. The relative fold specificity of the sequence (f_{spec}) can be determined by summing the statistical weights for those conformers from the free-folding simulation that resembles the template structure (denoted as set "temp"), and dividing this sum by the summation of statistical weights for all conformers from the free folding simulation (denote as set "total"):

$$f_{spec} = \frac{\sum_{i \in temp} \exp(-\beta E_i)}{\sum_{i \in total} \exp(-\beta E_i)}$$

where $\exp(-\beta E_i)$ is the statistical weight for conformer i.

Note that, in this nonrotamer approach, in both the template-constrained and free-folding calculations, all continuous Cα−Cα and angle values between upper and lower bounds input by the user are considered in sampling the conformers. Thus, true backbone flexibility[80] is conserved.

Lately, the Floudas group developed an approximate fold validation method that is computationally less expensive than ASTRO-FOLD. Through the CYANA 2.1 software for NMR structure refinement,[81,82] an ensemble of several hundred conformers are generated for both a new sequence from the first stage and the native sequence. The energies of the conformers are then minimized using TINKER,[128] and the fold specificity of the new sequence is calculated using the formula

$$f_{spec} = \frac{\sum_{i \in \text{conformers for new sequence}} \exp(-\beta E_i)}{\sum_{i \in \text{conformers for native sequence}} \exp(-\beta E_i)}$$

based on the assumption that the fold specificity to the flexible template is unity for the native sequence.

Similar to the fold validation method via ASTRO-FOLD, all continuous-distance and dihedral-angle values between their upper and lower bounds, which are input into CYANA, based on observations about the template structures, are considered in generating the conformers. This distinguishes the method from the common rotamer approach in which only discrete side-chain conformations are allowed.

Recently, the Donald group derived a novel DEE criterion called flexible-backbone DEE, which guarantees that no rotamers in the global minimum energy conformation compatible with a flexible backbone will be eliminated.[129,130] Discrete rotamers are used for the de novo protein design, and the ideas behind the development of the DEE algorithm are based on continuous-backbone dihedral-angle ranges and real-space restraint volumes on backbone movement.[19,129,130]

Successes: The Floudas group applied their two-stage de novo strategy to (i) the design of new sequences for compstatin, which is a synthetic 13-residue cyclic peptide that binds to complement protein 3 (C3) and inhibits the activation of the complement system (part of innate immunity);[10−14] (ii) the design of a potential peptide-drug candidate derived from the C-terminal sequence of the C3a fragment of C3;[131] and (iii) full-sequence of human β-defensin-2, which is a 41-residue cationic peptide in the immune system.[132] In the case of the compstatin redesign, sequences with 16-fold and 45-fold improvement in specificity over the native sequence were confirmed in experiments.[10,11] For the design of peptide drug from C3a, the best sequence identified corresponds to 15-fold improvement.[131]

Based on experimental results from the redesign of the β1 domain of protein G and the NRPS enzyme GrsA-PheA, the Donald group demonstrated the importance of flexible backbone, because the flexible-backbone DEE is able to design proteins with lower energies than the traditional DEE algorithms.[129]

Acknowledgment

C.A.F. gratefully acknowledges support from the National Science Foundation (NSF) and the National Institutes of Health (NIH) (under Grant Nos. R01 GM52032 and R24 GM069736). W.J.W. and C.A.F. gratefully acknowledge support from the U.S. Environmental Protection Agency (USEPA) (under Grant No. GAD R 832721-010). This work has not been reviewed by and does not represent the opinions of the USEPA.

Literature Cited

(1) Drexler, K. E. Molecular Engineering: An Approach to the Development of General Capabilities for Molecular Manipulation. *Proc. Natl. Acad. Sci., U.S.A.* **1981**, *78*, 5275.

(2) Pabo, C. Molecular Technology: Designing Proteins and Peptides. *Nature* **1983**, *301*, 200.

(3) Pogorelov, T. V.; Hardin, C.; Luthey-Schulten, Z. Ab Initio Protein Structure Prediction. *Curr. Opin. Struct. Biol.* **2002**, *12*, 176.

(4) Bryson, J. W.; Betz, S. F.; Lu, H. S.; Suich, D. J.; Zhou, H. X.; O'Neil, K. T.; DeGrado, W. F. Protein Design, a Heuristic Approach. *Science* **1995**, *270*, 935.

(5) Voigt, C. A.; Arnold, F. H.; Mayo, S. L.; Wang, Z.-G. Computational Method to Reduce the Search Space for Directed Protein Evolution. *Proc. Natl. Acad. Sci., U.S.A.* **2001**, *98*, 3778.

(6) Gordon, B. B.; Hom, G. K.; Mayo, S. L.; Pierce, N. A. Exact Rotamer Optimization for Protein Design. *J. Comput. Chem.* **2003**, *24*, 232.

(7) Kortemme, T.; Baker, D. Computational Design of Protein−Protein Interactions. *Curr. Opin. Chem. Biol.* **2004**, 8, 91−97.

(8) Malakauskas, S. M.; Mayo, S. L. Design, Structure, and Stability of a Hyperthermophilic Protein Variant. *Nat. Struct. Biol.* **1998**, *5*, 470.

(9) Kuhlman, B.; Baker, D. Exploring Folding Free Energy Landscapes Using Computational Protein Design. *Curr. Opin. Struct. Biol.* **2004**, 14, 89−95.

(10) Klepeis, J. L.; Floudas, C. A.; Morikis, D.; Tsokos, C. G.; Argyropoulos, E.; Spruce, L; Lambris, J. D. Integrated Computational and Experimental Approach for Lead Optimization and Design of Compstatin Variants with Improved Activity. *J. Am. Chem. Soc.* **2003**, *125*, 8422.

(11) Klepeis, J. L.; Floudas, C. A.; Morikis, D.; Tsokos, C. G.; Lambris, J. D. Design of Peptide Analogs with Improved Activity Using a Novel De Novo Protein Design Approach. *Ind. Eng. Chem. Res.* **2004**, *43*, 3817.

(12) Morikis, D.; Soulika, A. M.; Mallik, B.; Klepeis, J. L.; Floudas, C. A.; Lambris, J. D. Improvement of the Anti-C3 Activity of Compstatin Using Rational and Combinatorial Approaches. *Biochem. Soc. Trans.* **2004**, *32*, 28.

(13) Morikis, D.; Floudas, C. A.; Lambris, J. D. Structure-Based Integrative Computational and Experimental Approach for the Optimization of Drug Design. *Lect. Notes Comput. Sci.* **2005**, *3515*, 680.

(14) Floudas, C. A.; Fung, H. K. Mathematical Modeling and Optimization Methods for De Novo Protein Design. In *Systems Biology*, Vol. II; Rigoutsos, I., Stephanopoulos, G., Eds.; Oxford University Press: Oxford, U.K., 2006; pp 42−66.

(15) Hellinga, H. W.; Richards, F. M. Construction of New Ligand Binding Sites in Proteins of Known Structure. I. Computer-aided Modeling of Sites with Pre-defined Geometry. *J. Mol. Biol.* **1991**, *222*, 763.

(16) Hellinga, H. W.; Caradonna, J. P.; Richards, F. M. Construction of New Ligand Binding Sites in Proteins of Known Structure. II. Grafting of a Buried Transition Metal Binding Site into *Escherichia coli* Thioredoxin. *J. Mol. Biol.* **1991**, *222*, 787.

(17) Shimaoka, M.; Shifman, J. M.; Jing, H.; Takagi, L.; Mayo, S. L.; Springer, T. A. Computational Design of an Intergrin I Domain Stabilized in the Open High Affinity Conformation. *Nat. Struct. Biol.* **2000**, *7*, 674.

(18) Kraemer-Pecore, C. M.; Wollacott, A. M.; Desjarlais, J. R. Computational Protein Design. *Curr. Opin. Chem. Biol.* **2001**, *5*, 690.

(19) Lilien, R. H.; Stevens, B. W.; Anderson, A. C.; Donald, B. R. A Novel Ensemble-Based Scoring and Search Algorithm for Protein Redesign and Its Application to Modify the Substrate Specificity of the Gramicidin Synthetase a Phenylalanine Adenylation Enzyme. *J. Comput. Biol.* **2005**, *12*, 740.

(20) McFarland, B. J.; Kortemme, T.; Yu, S. F.; Baker, D.; Strong, R. K. Symmetry Recognizing Asymmetry: Analysis of the Interactions Between the C-Type Lectin-like Immunoreceptor NKG2D and MHC Class I-like Ligands. *Structure* **2003**, *11*, 411.

(21) Niv, M. Y.; Weinstein, H. A Flexible Docking Procedure for the Exploration of Peptide Binding Selectivity to Known Structures and Homology Models of PDZ Domains. *J. Am. Chem. Soc.* **2005**, *127*, 14072.

(22) Lazar, G. A.; Marshall, S. A.; Plecs, J. J.; Mayo, S. L.; Desjarlais, J. R. Designing Proteins for Therapeutic Applications. *Curr. Opin. Struct. Biol.* **2003**, *13*, 513.

(23) Zhu, Y. Mixed-Integer Linear Programming Algorithm for a Computational Protein Design Problem. *Ind. Eng. Chem. Res.* **2007**, *46*, 839.

(24) Dunbrack, R. L., Jr.; Cohen, F. E. Bayesian Statistical Analysis of Protein Side-Chain Rotamer Preferences. *Protein Sci.* **1997**, *6*, 1661.

(25) Pierce, N. A.; Winfree, E. Protein Design is NP-Hard. *Protein Eng.* **2002**, *15*, 779.

(26) Fung, H. K.; Rao, S.; Floudas, C. A.; Prokopyev, O.; Pardalos, P. M.; Rendl, F. Computational Comparison Studies of Quadratic Assignment Like Formulations for the In Silico Sequence Selection Problem in De Novo Protein Design. *J. Comb. Optim.* **2005**, *10*, 41.

(27) Floudas, C. A.; Fung, H. K.; Morikis, D.; Taylor, M. S.; Zhang, L. In *Modeling of Biosystems: An Interdisciplinary Approach*; Mondaini, R., Ed.; Springer−Verlag: New York, 2007 (in press).

(28) Ponder, J. W.; Richards, F. M. Tertiary Templates for Proteins. *J. Mol. Biol.* **1987**, *193*, 775.

(29) Lovell, S. C.; Word, J. M.; Richardson, J. S.; Richardson, D. C. The Penultimate Rotamer Library. *Proteins* **2000**, *40*, 389.

(30) Hellinga, H. W.; Richards, F. M. Optimal Sequence Selection in Proteins of Known Structure by Simulated Evolution. *Proc. Natl. Acad. Sci., U.S.A.* **1994**, *91*, 5803.

(31) Handel, T. M.; Desjarlais, J. R. De Novo Design of the Hydrophobic Cores of Proteins. *Protein Sci.* **1995**, *4*, 2006.

(32) Dahiyat, B. I.; Mayo, S. L. Protein Design Automation. *Protein Sci.* **1996**, *5*, 895.

(33) Mooers, B. H. M.; Datta, D.; Baase, W. A.; Zollars, E. S.; Mayo, S. L.; Matthews, B. W. Repacking the Core of T4 Lysozyme by Automated Design. *J. Mol. Biol.* **2003**, *332*, 741.

(34) Rosenberg, M.; Goldblum, A. Computational Protein Design: A Novel Path to Future Protein Drugs. *Curr. Pharm. Des.* **2006**, *12*, 3973.

(35) Dill, K. A. Dominant Forces in Protein Folding. *Biochemistry* **1990**, *29*, 7133.

(36) Desjarlais, J. R.; Clarke, N. D. Computer Search Algorithms in Protein Modification and Design. *Curr. Opin. Struct. Biol.* **1998**, *8*, 471.

(37) Desmet, J.; De Maeyer, M.; Hazes, B.; Lasters, I. The Dead-End Elimination Theorem and Its Use in Side-Chain Positioning. *Nature* **1992**, *356*, 539.

(38) Pierce, N. L.; Spriet, J. A.; Desmet, J.; Mayo, S. L. Conformational Splitting: A More Powerful Criterion for Dead-End Elimination. *J. Comput. Chem.* **2000**, *21*, 999.

(39) Goldstein, R. F. Efficient Rotamer Elimination Applied to Protein Side-Chains and Related Spin Glasses. *Biophys. J.* **1994**, *66*, 1335.

(40) Lasters, I.; De Maeyer, M.; Desmet, J. Enhanced Dead-End Elimination in the Search for the Global Minimum Energy Conformation of a Collection of Protein Side Chains. *Protein Eng.* **1995**, *8*, 815.

(41) Looger, L. L.; Hellinga, H. W. Generalized Dead-End Elimination Algorithms Make Large-Scale Protein Side-Chain Structure Prediction Tractable: Implications for Protein Design and Structural Genomics. *J. Mol. Biol.* **2001**, *307*, 429.

(42) Wernisch, L.; Hery, S.; Wodak, S. J. Automatic Protein Design with All Atom Force-Fields by Exact and Heuristic Optimization. *J. Mol. Biol.* **2000**, *301*, 713.

(43) Dahiyat, B. I.; Mayo, S. L. De Novo Protein Design: Fully Automated Sequence Selection. *Science* **1997**, *278*, 82.

(44) Shifman, J. M.; Mayo, S. L. Modulating Calmodulin Binding Specificity through Computational Protein Design. *J. Mol. Biol.* **2002**, *323*, 417.

(45) Offredi, F.; Dubail, F.; Kischel, P.; Sarinski, K.; Stern, A. S.; Van de Weerdt, C.; Hoch, J. C.; Prosperi, C.; François, J. M.; Mayo, S. L.; Martial, J. A. De Novo Backbone and Sequence Design of an Idealized α/β-Barrel Protein: Evidence of Stable Tertiary Structure. *J. Mol. Biol.* **2003**, *325*, 163.

(46) Benson, D. E.; Wisz, M. S.; Hellinga, H. W. The Development of New Biotechnologies Using Metalloprotein Design. *Curr. Opin. Biotechnol.* **1998**, *9*, 370.

(47) Benson, D. E.; Wisz, M. S.; Hellinga, H. W. Rational Design of Nascent Metalloenzymes. *Proc. Natl. Acad. Sci., U.S.A.* **2000**, *97*, 6292.

(48) Looger, L. L.; Dwyer, M. W.; Smith, J. J.; Hellinga, H. W. Computational Design of Receptor and Sensor Proteins with Novel Functions. *Nature* **2003**, *423*, 185.

(49) Dwyer, M. A.; Looger, L. L.; Hellinga, H. W. Computational Design of a Biologically Active Enzyme. *Science* **2004**, *304*, 1967.

(50) Koehl, P.; Delarue, M. Application of a Self-Consistent Mean Field Theory to Predict Protein Side-Chains Conformation and Estimate Their Conformational Entropy. *J. Mol. Biol.* **1994**, *239*, 249.

(51) Saven, J. G.; Wolynes, P. G. Statistical Mechanics of the Combinatorial Synthesis and Analysis of Folding Macromolecules. *J. Phys. Chem. B.* **1997**, *101*, 8375.

(52) Zou, J.; Saven, J. G. Statistical Theory of Combinatorial Libraries of Folding Proteins: Energetic Discrimination of a Target Structure. *J. Mol. Biol.* **2000**, *296*, 281.

(53) Kono, H.; Saven, J. G. Statistical Theory of Protein Combinatorial Libraries: Packing Interactions, Backbone Flexibility, and the Sequence Variability of a Main-Chain Structure. *J. Mol. Biol.* **2001**, *306*, 607.

(54) Fu, X.; Kono, H.; Saven, J. G. Probabilistic Approach to the Design of Symmetric Protein Quaternary Structures. *Protein Eng.* **2003**, *16*, 971.

(55) Lee, C. Predicting Protein Mutant Energetics by Self-Consistent Ensemble Optimization. *J. Mol. Biol.* **1994**, *236*, 918.

(56) Koehl, P.; Delarue, M. A Self Consistent Mean Field Approach to Simultaneous Gap Closure and Side-Chain Positioning in Homology Modeling. *Nat. Struct. Biol.* **1995**, *2*, 163.

(57) Kono, H.; Doi, J. Energy Minimization Method Using Automata Network for Sequence and Side-Chain Conformation Prediction from Given Backbone Geometry. *Proteins* **1994**, *19*, 244.

(58) Saven, J. G. Combinatorial Protein Design. *Curr. Opin. Struct. Biol.* **2002**, *12*, 453.

(59) Park, S.; Stowell, X. F.; Wang, W.; Yang, X.; Saven, J. G. Computational Protein Design and Discovery. *Annu. Rep. Prog. Chem., Sect. C* **2004**, *100*, 195.

(60) Park, S.; Yang, X.; Saven, J. G. Advances in Computational Protein Design. *Curr. Opin. Struct. Biol.* **2004**, *14*, 487.

(61) Hecht, M. H.; Das, A.; Go, A.; Bradley, L. H.; Wei, Y. De Novo Proteins from Designed Combinatorial Libraries. *Protein Sci.* **2004**, *13*, 1711.

(62) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **1953**, *21*, 389.

(63) Zou, J.; Saven, J. G. Using Self-Consistent Fields to Bias Monte Carlo Methods With Applications to Designing and Sampling Protein Sequences. *J. Chem. Phys.* **2003**, *118*, 3843.

(64) Cootes, A. P.; Curmi, P. M. G.; Torda, A. E. Biased Monte Carlo Optimization of Protein Sequences. *J. Chem. Phys.* **2000**, *113*, 2489.

(65) Koehl, P.; Levitt, M. De Novo Protein Design. I. In Search of Stability and Specificity. *J. Mol. Biol.* **1999**, *293*, 1161.

(66) Koehl, P.; Levitt, M. De Novo Protein Design. II. Plasticity in Sequence Space. *J. Mol. Biol.* **1999**, *293*, 1183.

(67) Dantas, G.; Kuhlman, B.; Callender, D.; Wong, M.; Baker, D. A Large Scale Test of Computational Protein Design: Folding and Stability of Nine Completely Redesigned Globular Proteins. *J. Mol. Biol.* **2003**, *332*, 449.

(68) Tuffery, P.; Etchebest, C.; Hazout, S.; Lavery, R. A New Approach to the Rapid Determination of Protein Side Chain Conformations. *J. Biomol. Struct. Dyn.* **1991**, *8*, 1267.

(69) Belda, I.; Madurga, S.; Llorà, X.; Martinell, M.; Tarragó, T.; Piqueras, M. G.; Nicolás, E.; Giralt, E. ENPDA: An Evolutionary Structure-Based De Novo Peptide Design Algorithm. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 585.

(70) Miyazawa, S.; Jernigan, R. L. Residue-Residue Potentials with a Favorable Contact Pair Term and an Unfavorable High Packing Density Term for Simulation and Threading. *J. Mol. Biol.* **1996**, *256*, 623.

(71) Hohm, T.; Limbourg, P.; Hoffmann, D. A Multiobjective Evolutionary Method for the Design of Peptidic Mimotopes. *J. Comput. Biol.* **2006**, *13*, 113.

(72) Pokala, N.; Handel, T. M. Review: Protein Design-Where We Were, Where We Are, Where We're Going. *J. Struct. Biol.* **2001**, *134*, 269.

(73) Lim, W. A.; Hodel, A.; Sauer, R. T.; Richards, F. M. The Crystal Structure of a Mutant Protein With Altered but Improved Hydrophobic Core Packing. *Proc. Natl. Acad. Sci., U.S.A.* **1994**, *91*, 423.

(74) Emberly, E. G.; Mukhopadhyay, R.; Tang, C.; Wingreen, N. S. Flexibility of α-helices: Results of a Statistical Analysis of Database Protein Structures. *J. Mol. Biol.* **2003**, *327*, 229.

(75) Emberly, E. G.; Mukhopadhyay, R.; Tang, C.; Wingreen, N. S. Flexibility of *β*-sheets: Principal Component Analysis of Database Protein Structures. *Proteins* **2004**, *55*, 91.

(76) Fung, H. K.; Taylor, M. S.; Floudas, C. A. Novel Formulations for the Sequence Selection Problem in De Novo Protein Design with Flexible Templates. *Optim. Methods Software* **2007**, *22*, 51.

(77) Loose, C.; Klepeis, J. L.; Floudas, C. A. A New Pairwise Folding Potential Based on Improved Decoy Generation and Side Chain Packing. *Proteins* **2004**, *54*, 303.

(78) Rajgaria, R.; McAllister, S. R.; Floudas, C. A. A Novel High Resolution Cα-Cα Distance Dependent Force Field Based on a High Quality Decoy Set. *Proteins* **2006**, *65*, 726.

(79) Rajgaria, R.; McAllister, S. R.; Floudas, C. A. Improving the Performance of a High Resolution Distance Dependent Force Field by Including Protein Side Chains. *Proteins* **2007**, in press.

(80) Floudas, C. A. Research Challenges, Opportunities and Synergism in Systems Engineering and Computational Biology. *AIChE J.* **2005**, *51*, 1872.

(81) Guntert, P.; Mumenthaler, C.; Wuthrich, K. Torsion Angle Dynamics for NMR Structure Calculation with the New Program DYANA. *J. Mol. Biol.* **1997**, *273*, 283.

(82) Guntert, P. Automated NMR Structure Calculation with CYANA. *Methods Mol. Biol.* **2004**, *278*, 353.

(83) Floudas, C. A.; Grossmann, I. E. Synthesis of Flexible Heat-Exchanger Networks with Uncertain Flowrates and Temperatures. *Comput. Chem. Eng.* **1987**, *11*, 319.

(84) Grossmann, I. E.; Floudas, C. A. Active Constraint Strategy for Flexibility Analysis in Chemical Processes. *Comput. Chem. Eng.* **1987**, *11*, 675.

(85) Floudas, C. A.; Gumus, Z. H.; Ierapetritou, M. G. Global Optimization in Design Under Uncertainty: Feasibility Test and Flexibility Index Problems. *Ind. Eng. Chem. Res.* **2001**, *40*, 4267.

(86) Butterfoss, G. L.; Kuhlman, B. Computer-Based Design of Novel Protein Structures. *Annu. Rev. Biophys. Biomol. Struct.* **2006**, *35*, 49.

(87) Su, A.; Mayo, S. L. Coupling Backbone Flexibility and Amino Acid Sequence Selection in Protein Design. *Protein Sci.* **1997**, *6*, 1701.

(88) Ross, S. A.; Sarisky, C. A.; Su, A.; Mayo, S. L. Designed Protein G Core Variants Fold to Native-Like Structures: Sequence Selection by Orbit Tolerates Variation in Backbone Specification. *Protein Sci.* **2001**, *10*, 450.

(89) Larson, S. M.; England, J. L.; Desjarlais, J. R.; Pande, V. S. Thoroughly Sampling Sequence Space: Large-Scale Protein Design of Structural Ensembles. *Protein Sci.* **2002**, *11*, 2804.

(90) Larson, S. M.; Garg, A.; Desjarlais, J. R.; Pande, V. S. Increased Detection of Structural Templates Using Alignments of Designed Sequences. *Proteins* **2003**, *51*, 390.

(91) Kraemer-Pecore, C. M.; Lecomte, J. T.; Desjarlais, J. R. A De Novo Redesign of the WW Domain. *Protein Sci.* **2003**, *12*, 2194.

(92) Raha, K.; Wollacott, A. M.; Italia, M. J.; Desjarlais, J. R. Prediction of Amino Acid Sequence from Structure. *Protein Sci.* **2000**, *9*, 1106.

(93) Ambroggio, X. I.; Kuhlman, B. Computational Design of a Single Amino Acid Sequence That Can Switch Between Two Distinct Protein Folds. *J. Am. Chem. Soc.* **2006**, *128*, 1154.

(94) Ambroggio, X. I.; Kuhlman, B. Design of Protein Conformational Switches. *Curr. Opin. Struct. Biol.* **2006**, *16*, 525.

(95) Desjarlais, J. R.; Handel, T. M. Side Chain and Backbone Flexibility in Protein Core Design. *J. Mol. Biol.* **1999**, *290*, 305.

(96) Kuhlman, B.; Dantas, G.; Ireton, G. C.; Verani, G.; Stoddard, B.; Baker, D. Design of a Novel Globular Protein Fold With Atomic-Level Accuracy. *Science* **2003**, *302*, 1364.

(97) Saunders, C. T.; Baker, D. Recapitulation of Protein Family Divergence Using Flexible Backbone Protein Design. *J. Mol. Biol.* **2005**, *346*, 631.

(98) Harbury, P. B.; Tidor, B.; Alber, T.; Kim, P. S. Repacking Protein Cores with Backbone Freedom: Structure Prediction for Coiled Coils. *Proc. Natl. Acad. Sci., U.S.A.* **1995**, *92*, 8408.

(99) Harbury, P. B.; Plecs, J. J.; Tidor, B.; Alber, T.; Kim, P. S. High-Resolution Protein Design with Backbone Freedom. *Science* **1998**, *282*, 1462.

(100) Plecs, J. J.; Harbury, P. B.; Kim, P. S.; Alber, T. Structural Test of the Parameterized-Backbone Method for Protein Design. *J. Mol. Biol.* **2004**, *342*, 289.

(101) Crick, F. H. C. The Fourier Transform of a Coiled-Coil. *Acta Crystallogr.* **1953**, *6*, 685.

(102) De Filippis, V.; Sander, C.; Vriend, G. Predicting Local Structural-Changes That Result From Point Mutations. *Protein Eng.* **1994**, *7*, 1203.

(103) Klepeis, J. L.; Floudas, C. A.; Morikis, D.; Lambris, J. D. Predicting Peptide Structures Using NMR Data and Deterministic Global Optimization. *J. Comput. Chem.* **1999**, *20*, 1354.

(104) Klepeis, J. L.; Floudas, C. A. Free Energy Calculations for Peptides Via Deterministic Global Optimization. *J. Chem. Phys.* **1999**, *110*, 7491.

(105) Klepeis, J. L.; Floudas, C. A. Ab Initio Prediction of Helical Segments in Polypeptides. *J. Comput. Chem.* **2002**, *23*, 245.

(106) Klepeis, J. L.; Floudas, C. A. Ab Initio Tertiary Structure Prediction of Proteins. *J. Global Optim.* **2003**, *25*, 113.

(107) Klepeis, J. L.; Floudas, C. A. Prediction of *β*-Sheet Topology and Disulfide Bridges in Polypeptides. *J. Comput. Chem.* **2003**, *24*, 191.

(108) Klepeis, J. L.; Floudas, C. A. ASTRO-FOLD: A Combinatorial and Global Optimization Framework for Ab Initio Prediction of Three-Dimensional Structures of Proteins From the Amino Acid Sequence. *Biophys. J.* **2003**, *85*, 2119.

(109) Klepeis, J. L.; Wei, Y. N.; Hecht, M. H.; Floudas, C. A. Ab Initio Prediction of the Three Dimensional Structure of a De Novo Designed Protein: A Double-Blind Case Study. *Proteins* **2005**, *58*, 560.

(110) McAllister, S. R.; Mickus, B. E.; Klepeis, J. L.; Floudas, C. A. Novel Approach for Alpha-Helical Topology Prediction in Globular Proteins: Generation of Interhelical Restraints. *Proteins* **2006**, *65*, 930.

(111) Androulakis, I. P.; Maranas, C. D.; Floudas, C. A. Prediction of Oligopeptide Conformations via Deterministic Global Optimization. *J. Global Optim.* **1997**, *11*, 1.

(112) Adjiman, C. S.; Floudas, C. A. Rigorous Convex Underestimators for General Twice-Differentiable Problems. *J. Global Optim.* **1996**, *9*, 23.

(113) McDonald, C.; Floudas, C. A. Global Optimization and Analysis for the Gibbs Free-Energy Function Using the UNIFAC, Wilson, and Asog Equations. *Ind. Eng. Chem. Res.* **1995**, *34*, 1674.

(114) Maranas, C. D.; Floudas, C. A. Finding All Solutions of Nonlinearly Constrained Systems of Equations. *J. Global Optim.* **1995**, *7*, 143.

(115) Adjiman, C. S.; Androulakis, I. P.; Maranas, C. D.; Floudas, C. A. A Global Optimization Method, alpha BB, for Process Design. *Comput. Chem. Eng.* **1996**, *20*, S419.

(116) Visweswaran, V.; Floudas, C. A. A Global Optimization Algorithm (GOP) for Certain Classes of Nonconvex NLPs. 2. Application of Theory and Test Problems. *Comput. Chem. Eng.* **1990**, *14*, 1419.

(117) Floudas, C. A.; Pardalos, P. M. State-of-the-art in Global Optimization—Computational Methods and Applications: Preface. *J. Global Optim*. **1995**, *7*, 113.

(118) McDonald, C. M.; Floudas, C. A. Global Optimization for the Phase and Chemical-Equilibrium Problem: Application to the NRTL Equation. *Comput. Chem. Eng.* **1995**, *19*, 1111.

(119) Maranas, C. D.; Floudas, C. A. Global Optimization in Generalized Geometric Programming. *Comput. Chem. Eng.* **1997**, *21*, 351.

(120) Adjiman, C. S.; Androulakis, I. P.; Floudas, C. A. Global Optimization of MINLP Problems in Process Synthesis and Design. *Comput. Chem. Eng.* **1997**, *21*, S445.

(121) Esposito, W. R.; Floudas, C. A. Global Optimization for the Parameter Estimation of Differential−algebraic Systems. *Ind. Eng. Chem. Res*. **2000**, *39*, 1291.

(122) Esposito, W. R.; Floudas, C. A. Global Optimization in Parameter Estimation of Nonlinear Algebraic Models via the Error-in-variables Approach. *Ind. Eng. Chem. Res*. **1998**, *37*, 1841.

(123) Visweswaran, V.; Floudas, C. A. New Properties and Computational Improvement of the GOP Algorithm for Problems with Quadratic Objective Functions and Constraints. *J. Global Optim*. **1993**, *3*, 439.

(124) Harding, S. T.; Maranas, C. D.; McDonald, C. M.; Floudas, C. A. Locating All Homogeneous Azeotropes in Multicomponent Mixtures. *Ind. Eng. Chem. Res*. **1997**, *36*, 160.

(125) McDonald, C. M.; Floudas, C.A. Decomposition Based and Branch-and-bound Global Optimization Approaches for the Phase-Equilibrium Problem. *J. Global Optim*. **1994**, *5*, 205.

(126) Esposito, W. R.; Floudas, C. A. Deterministic Global Optimization in Nonlinear Optimal Control Problems. *J. Global Optim*. **2000**, *17*, 97.

(127) Westerberg, K. M.; Floudas, C. A. Locating All Transition States and Studying the Reaction Pathways of Potential Energy Surfaces. *J. Chem. Phy*. **1999**, *110*, 9259.

(128) Ponder, J. *TINKER, Software Tools for Molecular Design*; Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine: St. Louis, MO, 1998.

(129) Georgiev, I.; Donald, B. R. Dead-end Elimination with Backbone Flexibility. *Bioinformatics* **2007**, *23*, i185.

(130) Georgiev, I.; Lilien, R. H.; Donald, B. R. Improved Pruning Algorithms and Divide-and-Conquer Strategies for Dead-End Elimination, with Application to Protein Design. *Bioinformatics* **2006**, *22*, e174.

(131) Fung, H. K.; Taylor, M. S.; Floudas, C. A.; Morikis, D.; Lambris, J. D. Redesigning Complement 3a Based on Flexible Templates from Both X-ray Crystallography and Molecular Dynamics Simulation. Unpublished work.

(132) Fung, H. K.; Floudas, C. A.; Taylor, M. S.; Zhang, L.; Morikis, D. Towards Full-Sequence De Novo Protein Design with Flexible Templates for Human Beta-Defensin-2. *Biophys. J*. **2007**, in press (DOI: 10.1529/biophysj.107.110627).

Received for review September 24, 2007
Revised manuscript received October 30, 2007
Accepted November 1, 2007

IE071286K