From *SIAM News*, Volume 37, Number 1, January/February 2004

In Silico Protein Design: A Combinatorial and Global Optimization Approach

By John L. Klepeis and Christodoulos A. Floudas

The use of computational techniques to create peptide- and protein-based therapeutics is an important challenge in medicine. The ultimate goal, defined about two decades ago, is to use computer algorithms to identify amino acid sequences that not only adopt particular three-dimensional structures but also perform specific functions. To those familiar with the field of structural biology, it is certainly not surprising that this problem has been described as "inverse protein folding" [16]. That is, while the grand challenge of protein folding is to understand how a particular protein, defined by its amino acid sequence, finds its unique three-dimensional structure, protein design involves the discovery of sets of amino acid sequences that form functional proteins and fold into specific target structures.

Experimental, computational, and hybrid approaches have all contributed to advances in protein design. Applying mutagenesis and rational design techniques, for example, experimentalists have created enzymes with altered functionalities and increased stability. The coverage of sequence space is highly restricted for these techniques, however [4]. An approach that samples more diverse sequences, called directed protein evolution, iteratively applies the techniques of genetic recombination and in vitro functional assays [1]. These methods, although they do a better job of sampling sequence space and generating functionally diverse proteins, are still restricted to the screening of $10^3 - 10^6$ sequences [22].

Challenges of Generic Computational Protein Design

The limitations of experimental techniques serve to highlight the importance of computational protein design. Practically speaking, in silico methods can sample astronomically large numbers of sequences; the resulting diversity in the selected sequences leads to a much broader spectrum of functional proteins. Computational methods have already been used successfully to alter existing proteins so that they have better stability and functionality, and to combine or modify proteins for aggregate functionality. The ultimate goal is the de novo computational design of proteins—that is, a systematic way to create proteins that have both new structural templates and better properties.

The success of an approach to computational protein design depends on two main ingredients: (i) the method used to search sequence space, and (ii) the principles on which the modeling is based. To better understand the combinatorial nature of a search through sequence space, consider a relatively small protein, 50 residues long. Allowing for any one of the 20 possible amino acids at each residue position of this protein results in 20^{50}, or more than 10^{65} possible amino acid sequences. Clearly, clever optimization techniques are needed to deal with this level of combinatorial complexity. Although stochastic methods have been used [9, 23], the first successful computational design of a full protein was achieved with a deterministic branch-and-bound technique based on the dead-end elimination theorem [3, 6]. Even in the case of dead-end elimination, however, heuristics must be incorporated to make convergence reasonably fast for large proteins.

Given a method that can effectively search through such large numbers of sequences, the question becomes one of distinguishing between protein sequences. In other words, what is the target function that we would like to optimize? Of course, this will depend on the representation of the protein. Initial efforts focused only on the replacement of core residues, a condition under which the steric van der Waals and hydrophobic forces are expected to dominate [5, 18]. As computational protein design has been extended to full proteins, requiring the addition of hydrogen bonding and solvent and electrostatic effects in various forms and flavors, the models have become more and more complex [2, 15, 19]. Overall, there is no consensus among these models, and it is unclear which methods are more valid and suitable for generic computational protein design. More fundamental concerns still to be addressed include the realization that imposing a rigid template is a severe constraint.

Our recent efforts in the area of computational protein design [12] consist of two separate stages: (i) in silico sequence selection, and (ii) validation of fold stability and specificity. In stage (i) we use a novel mixed-integer formulation that incorporates amino acid side-chain specificity to model sequence space. We devised a method for solving this new mixed-integer optimization problem; the solutions provide a set of candidate sequences for input to stage (ii). In stage (ii) we study the reduced set of sequences in more detail, using the principles of ASTROFOLD [11], a method for ab initio prediction of three-dimensional protein structures within a combinatorial and global optimization framework. The approach allows backbone flexibility, and the final results reflect a quantitative ranking of fold stability and specificity for each amino acid sequence. Figure 1 depicts the full computational design approach.

Modeling Sequence Space

The formulation of stage (i) depends on the representation of the protein system. Initially, rather than describe the amino acids by the coordinates of all atoms, we describe the backbone template only by the coordinates of the alpha-carbon atoms. A pairwise distance-dependent interaction potential is used to calculate the energy of an amino acid sequence on this template. The statistically based energy function assigns energy values according to the alpha-carbon separation distance for each pair of amino acids. Similar

structure-based interaction potentials have been used in fold recognition and fold prediction [17].

The advantages of this representation are the simplicity of the model and the robustness of the system with respect to the rigid backbone approximation. In other words, while the interaction potential implicitly includes amino acid and side-chain specificity, its coarseness allows for an inherent flexibility in the backbone. For the interaction potential used in our study, alpha-carbon distances are discretized into a set of 13 bins, with the 2730 parameters of the model being derived from the solution to a linear optimization formulation that favors native folds over decoy structures [14, 21].

An explanation of the development of the new mixed-integer formulation begins with a description of the variable set over which the energy function is optimized. First, consider the set $i = 1, \ldots, n$, which defines the residue positions along the backbone. At each position i there can be a set of amino acid substitutions represented by $j\{i\} = 1, \ldots, m_i$ where, for the general case $m_i = 20$ $\forall\ i$. The equivalent sets $k \equiv i$ and $l \equiv j$ are defined, and k must be greater than i to ensure that only unique pairwise interactions are represented. Binary variables y_i^j and y_k^l are introduced to indicate the possible substitutions at a given position. That is, y_i^j indicates the amino acid j at a position i in the sequence by taking the value of 1 for one amino acid and 0 for all others. The formulation, in which the goal is to minimize the energy according to the pairwise interaction parameters that multiply the binary variables, can then be expressed as:

$$\min_{y_i^j, y_k^l} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \sum_{k=i+1}^{n} \sum_{l=1}^{m_k} E_{ik}^{jl}\left(x_i, x_k\right) y_i^j y_k^l$$

$$\text{subject to } \sum_{j=1}^{m_i} y_i^j = 1 \ \forall\ i$$

$$y_i^j, y_k^l = 0 - 1 \ \forall\ i, j, k, l. \tag{1}$$

The parameters $E_{ik}^{jl}(x_i, x_k)$ depend on the distance between the alpha-carbons at the two backbone positions (x_i, x_k), as well as on the specific amino acids at those positions. The composition constraints require that at most one amino acid appear at each position. Notice that the binary variables appear as bilinear combinations in the objective function. Fortunately, this objective can be reformulated as a strictly linear (integer linear programming) problem [7]:

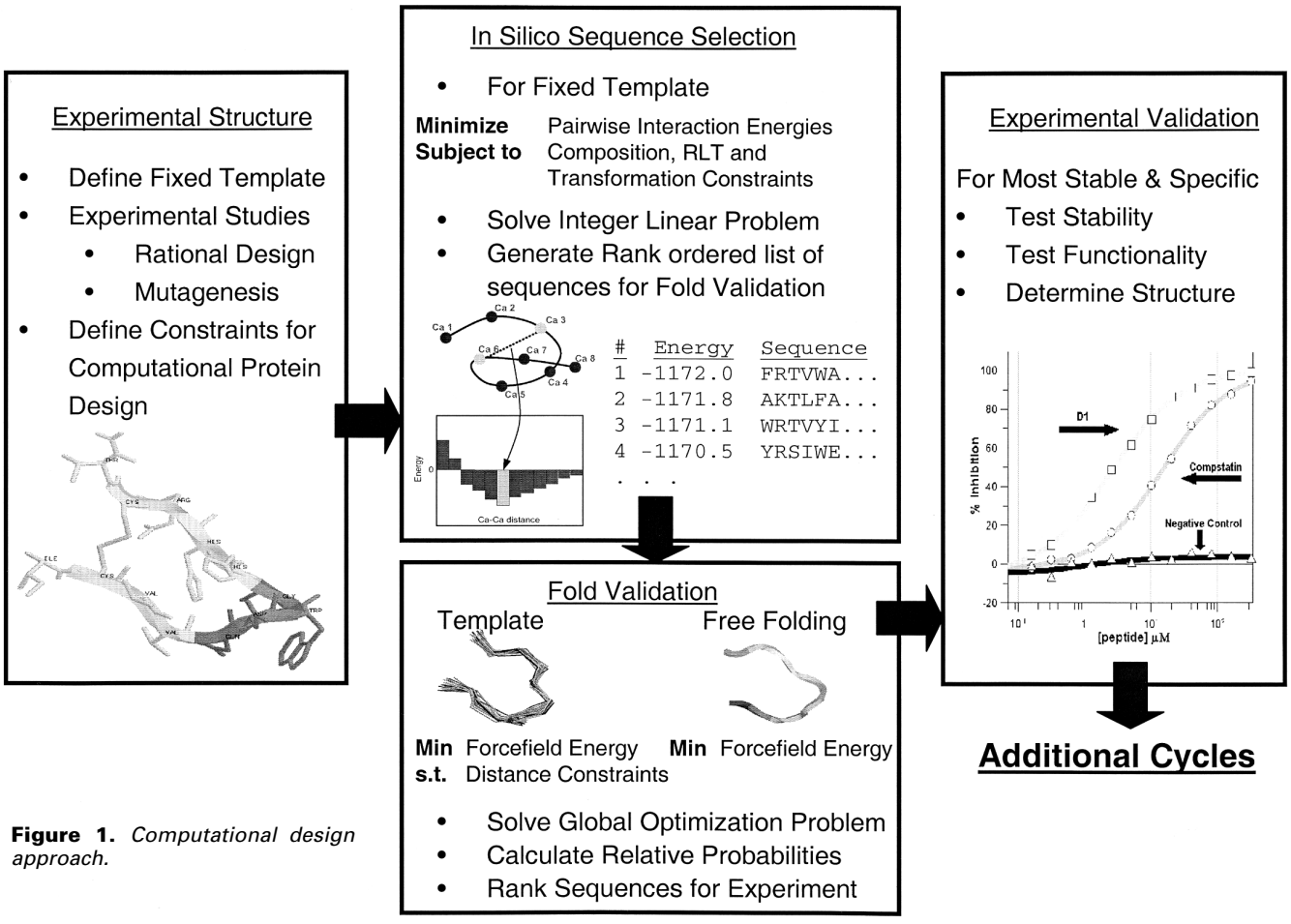

Figure 1. *Computational design approach.*

$$
\begin{aligned}
&\min_{y_i^j, y_k^l} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \sum_{k=i+1}^{n} \sum_{l=1}^{m_k} E_{ik}^{jl}\left(x_i, x_k\right) w_{ik}^{jl} \\
&\text{subject to } \sum_{j=1}^{m_i} y_i^j = 1 \;\forall\; i \\
&y_i^j, y_k^l - 1 \geq w_{ik}^{jl} \leq y_i^j \;\forall\; i,j,k,l \\
&0 \leq w_{ik}^{jl} \leq y_k^l \;\forall\; i,j,k,l \\
&y_i^j, y_k^l = 0-1 \;\forall\; i,j,k,l.
\end{aligned}
\tag{2}
$$

This reformulation relies on the transformation of the bilinear combinations into a new set of linear variables w_{ik}^{jl}, while the addition of the four sets of constraints provides the equivalence to the original formulation.

Although the integer linear programming problem can be solved by standard branch-and-bound techniques [7], convergence is prohibitively slow for large systems. An important finding is that the performance of the branch-and-bound algorithm improves significantly when the principles of reformulation linearization techniques (RLT) are applied. The basic strategy is to multiply appropriate constraints by bounded non-negative factors and then replace the products of the original variables by new variables; in this way, we derive higher-dimensional lower-bounding linear programming relaxations to the original problem [20]. The tighter LP relaxations are included in the course of the overall branch-and-bound search and speed convergence to the global minimum.

Application of the RLT approach to the composition constraint begins with a reformulation of the equations; we form the product of the constraint equations with some binary variables (or their complements). For example, by multiplying the composition constraint by the set of variables y_k^l , we produce the following additional set of constraints $\forall$ j, k, l:

$$
y_k^l \sum_{i=1}^{m_i} y_i^j = y_k^l \;\forall\; j,k,l. \tag{3}
$$

The variable substitution already introduced to linearize the objective function can now be used to linearize equation (3). The set of RLT constraints becomes:

$$
\sum_{i=1}^{m_i} w_{ik}^{jl} = y_k^l \;\forall\; j,k,l. \tag{4}
$$

These additional constraints are added to the formulation given by (2). It is then straightforward to identify a rank-ordered list of the low-energy sequences through the introduction of integer cuts [7] and repetitive solution of the integer linear programming problem.

Predicting 3D Protein Structures

Once a set of sequences has been identified in stage (i), we proceed to stage (ii), using a flexible template to rigorously quantify the stability and specificity of each sequence. The approach is based on the generation of all atomistic structural ensembles for the selected sequences under two sets of conditions. Under the first, the structures are constrained to vary, with some imposed fluctuations, around the template structure; under the second, free folding calculations are performed. The formulations are reminiscent of the structure-prediction problems in protein folding [13]. Specifically, the problems are formulated as constrained global optimizations of a detailed atomistic energy forcefield E_{ff} over a set of internal coordinates ϕ, which describe any conformation of the system. The bounds on these variables are enforced by simple box constraints. Finally, a set of constraints, E_r^{dis} $r = 1, \ldots, N_R$, which are nonconvex in the internal coordinate space, can be used to constrain interatomic distances. The formulation is represented by the following set of equations:

$$
\begin{aligned}
&\min_{\phi} E_{ff} \\
&\text{subject to } E_r^{dis}\left(\phi\right) \leq 0 \quad r = 1,...,N_R \\
&\phi_s^L \leq \phi_s \leq \phi_s^U \quad s = 1,...,N_\phi.
\end{aligned}
\tag{5}
$$

Here, $s = 1, \ldots, N_\phi$ corresponds to the set of internal coordinates ϕ_s, with ϕ_s^L and ϕ_s^U representing lower and upper bounds on these variables. The forms of the distance constraints and the forcefield energy function E_{ff} are completely general. In practice, we use square-well quadratic functions for the distance constraints, and an atomistic-level forcefield that includes van der Waals, hydrogen bonding, and electrostatic and torsional terms for the objective.

These formulations belong to the class of general nonconvex constrained global optimization problems, and are solved via the principles of an αBB deterministic global optimization approach, a branch-and-bound method applicable to the identification of the global minimum of nonconvex optimization problems with twice-differentiable functions [8]. In addition to identifying the global minimum-energy conformation, this global optimization approach has been adapted to locate an ensemble of low-energy conformations [10–11]. These ensembles are used to quantify the fold stability and specificity by summing the statistical weights for the conformers from the free-folding simulation that resemble the template structure, and dividing this sum by the total partition function; that is, statistical weights are summed for all conformers from the free-folding simulation. The analysis is an unambiguous

method for ranking the fold stability and specificity among a set of different amino acid sequences.

The approach described here has been successfully tested on the design of improved analogs for Compstatin, a synthetic therapeutic peptide that prevents the autoimmune-mediated damage of organs during transplantation and in various inflammatory diseases. The new computational design approach yielded a version of Compstatin seven times more efficacious and stable than the original peptide [12] (see Figure 2 for a summary of the results). The result is a significant improvement over analogs identified by either purely rational or experimental combinatorial design techniques.

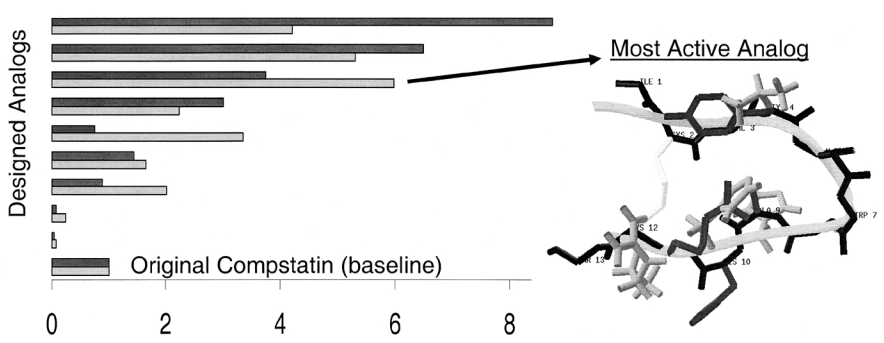

Figure 2. *Stability and activity relative to the synthetic therapeutic peptide Compstatin. Experimental results are in light gray; predicted results are in dark gray.*

References

[1] J. Bowie, J. Reidhaar-Olson, W. Lim, and R. Sauer, *Deciphering the message in protein sequences: Tolerance to amino acid substitutions*, Science, 247 (1990), 1306–1310.

[2] B. Dahiyat, D. Gordon, and S. Mayo, *Automated design of the surface positions of protein helices*, Protein Sci., 6 (1997),1333–1337.

[3] B. Dahiyat and S. Mayo, *De novo protein design: Fully automated sequence selection*, Science, 278 (1997), 82–87.

[4] W. DeGrado, Z. Wasserman, and I. Lear, *Protein design, a minimalist approach*, Science, 243 (1989), 622–628.

[5] J. Desjarlais and T. Handel, *De novo design of the hydrophobic cores of proteins*, Protein Sci., 4 (1995), 2006–2018.

[6] J. Desmet, M.D. Maeyer, B. Hazes, and I. Lasters, *The dead-end elimination theorem and its use in side-chain positioning*, Nature, 356 (1992), 539–542.

[7] C.A. Floudas, *Nonlinear and Mixed-Integer Optimization: Fundamentals and Applications*, Oxford University Press, Oxford, UK, 1995.

[8] C.A. Floudas, *Deterministic global optimization: Theory, methods and applications,* in *Nonconvex Optimization and its Applications*, Kluwer Academic Publishers, New York, 2000.

[9] D. Jones, *De novo protein design using pairwise potentials and a genetic algorithm,* Protein Sci., 3 (1994), 567–574.

[10] J.L. Klepeis and C.A. Floudas, *Ab initio tertiary structure prediction of proteins*, J. Global Optim., 25 (2003), 113–140.

[11] J.L. Klepeis and C.A. Floudas, *Astro-fold : A combinatorial and global optimization framework for ab initio prediction of three-dimensional structures of proteins from the amino-acid sequence*, Biophysical J., 32 (2003), 2119–2146.

[12] J.L. Klepeis, C.A. Floudas, D. Morikis, C.G. Tsokos, E. Argyropoulos, L. Spruce, and J.D. Lambris, *Integrated computational and experimental approach for lead optimization and design of compstatin variants with improved activity*, J. Am. Chem. Soc., 125 (2003), 8422–8423.

[13] J.L. Klepeis, H.D. Schafroth K.M. Westerberg, and C.A. Floudas, *Deterministic global optimization and ab initio approaches for the structure prediction of polypeptides, dynamics of protein folding and protein-protein interaction,* in *Advances in Chemical Physics*, R.A. Friesner, ed., 120, John Wiley & Sons, New York, 2002, 254–457.

[14] C. Loose, J. Klepeis , and C. Floudas, *A new pairwise folding potential based on improved decoy generation and side chain packing*, Proteins, to appear, 2003.

[15] M. Nohaile, Z. Hendsch, B. Tidor, and R. Sauer, *Altering dimerization specificity by changes in surface electrostatics*, Proc. Natl. Acad. Sci. USA, 98 (2001), 3109–3114.

[16] C. Pabo, *Molecular technology: Designing proteins and peptides*, Nature, 301 (1983), 200.

[17] B. Park and M. Levitt, *Energy functions that discriminate x-ray and near native folds from well-constructed decoys*, J. Mol. Biol., 258 (1996), 367–392.

[18] J. Ponder and F. Richards, *Tertiary templates for proteins*, J. Mol. Biol., 193 (1987), 775–791.

[19] K. Raha, A. Wollacott, M. Italia, and J. Desjarlais, *Prediction of amino acid sequence from structure*, Protein Sci., 9 (2000), 1106–1119.

[20] H. Sherali and W. Adams, *A Reformulation Linearization Technique for Solving Discrete and Continuous Nonconvex Problems*, Kluwer Academic Publishing, Boston, 1999.

[21] D. Tobi and R. Elber, *Distance-dependent pair potential for protein folding: Results from linear optimization*, Proteins, 41 (2000), 40–46.

[22] C. Voigt, S. Mayo, and Z.G. Wang, C*omputational method to reduce the search space for directed protein evolution*, Proc. Natl. Acad. Sci. USA, 98 (2001), 3778–3783.

[23] L. Wernisch, S. Hery, and S. Wodak, *Automatic protein design with all atom force-fields by exact and heuristic optimization*, J. Mol. Biol., 301 (2000), 713–736.

John L. Klepeis is a researcher at D.E. Shaw & Co. in New York. Christodoulos A. Floudas is a professor of chemical engineering at Princeton University.