Mathematical Modeling and Optimization Methods for De Novo Protein Design

C.A. Floudas[1] and H. K. Fung[1]

[1] Department of Chemical Engineering, Princeton University, Princeton, NJ 08544-5263

Abstract

A major challenge in computational peptide and protein design is the systematic generation of novel peptides and proteins which are either compatible with existing target template structures or with arbitrarily postulated new three dimensional structural folds. This chapter presents an account of the recent advances in mathematical modeling and optimization methods for de novo protein design. It will be followed by a novel integrated framework based on global optimization and mixed-integer optimization for the computational design of peptides and proteins. Compstatin, a 13-residue cyclic peptide that binds to complement component C3 and inhibits complement activation, will be employed as the peptide target for testing the novel de novo protein design framework. Experimental functional analysis provides validation to the in silico predicted novel peptide sequences (e.g. $Ac - I[CVYQDWGAHRC]T - NH_2$) which are shown to exhibit 16-fold improved activity over the synthetic therapeutic peptide Compstatin. Further validation of the in silico sequence prediction framework is obtained by considering tryptophan in position 4 of Compstatin. It is shown that the mutation of valine to tryptophan is preferred (e.g. $Ac - I[CVWQDWGAHRC]T - NH_2$) compared to the mutation from valine to tyrosine, in agreement with recent experimental results (Mallik *et al.*, 2005) which demonstrated 45-fold higher inhibitory activity.

Keywords

Peptide and protein design and discovery; Drug design; Complement inhibitor; Structure prediction; Optimization

Introduction

The de novo peptide and protein design, first suggested almost two decades ago, begins with a postulated or known flexible protein three-dimensional structure and aims at identifying amino acid sequence(s) compatible with this structure. Initially, the problem was denoted as the “inverse folding problem” (Drexler, 1981; Pabo, 1983) since protein design has intimate links to the well-known protein folding problem (C. Hardin and Luthey-Schulten, 2002). In contrast to the characteristic of protein folding to associate a given protein sequence with its own unique shape, the inverse folding problem exhibits high levels of degeneracy; that is, a large number of sequences will be compatible with a given protein structure, although the sequences will vary with respect to properties such as activity and stability.

In silico protein design allows for the screening of overwhelmingly large sectors of sequence space, with this sequence diversity subsequently leading to the possibility of a much broader range of properties and degrees of functionality among the selected sequences. Allowing for all 20 possible amino acids at each position of a small 50 residue protein results in 20^{50} combinations, or more than 10^{65} possible sequences. From this large number of sequences, the computational sequence selection process aims at selecting those sequences that will be compatible with a given structure using efficient optimization of energy functions that model the molecular interactions.

In an effort to make the difficult nature of the energy modeling and combinatorial optimization manageable, the first attempts at computational protein design focused only on a subset of core residues and explored steric van der Waals based energy functions through exhaustive searches for compatible sequences (Ponder and Richards, 1987; Hellinga and Richards, 1991). Over time, the models have evolved to incorporate improved rotamer libraries in combination with detailed energy models and interaction potentials. Although the consideration of packing effects on structural specificity is sometimes sufficient, as shown through the design of compatible structures using backbone-dependent rotamer libraries with only van der Waals energy evaluations for a subset of hydrophobic residues (Desjarlais and Handel, 1995; Dahiyat and Mayo, 1996), there has been extensive research to develop models including hydrogen bonding, solvent and electrostatic effects (Dahiyat *et al.*, 1997; Raha *et al.*, 2000; Street and Mayo, 1998; Nohaile *et al.*, 2001). These functional additions to the design models are especially important for full sequence design since packing interactions no longer dominate for non-core residues (e.g., surface and intermediate residues). The incorporation of these additional non-core residues increases the potential for diversity, and therefore enhances the probability for improving functionality when compared to the parent system. An additional complication is the need to account for changes in amino acid compositions and inherent propensities through the appropriate definition of a reference state (Koehl and Levitt, 1999; Wernisch *et al.*, 2000; Raha *et al.*, 2000).

Template Flexibility

Many computational protein design efforts were based on the premise that the three-dimensional coordinates of the template or backbone were fixed. This assumption was first proposed by Ponder and Richards (1987), and was appealing because it greatly reduced the search space and thus the time required to converge to a solution for the minimum energy sequence, regardless of the kind of search method employed. However, the assumption was also highly questionable. Protein backbones had been observed to allow residues that would not have been permissable had the backbone been fixed (Lim *et al.*, 1994). In the Protein Data Bank, there exist numerous examples of proteins which exhibit multiple NMR structures. Though commonly assumed as rigid bodies as a first approximation, the secondary structures of α-helices and β-sheets actually display some twisting and bending in the protein fold, and Emberly *et al.* (2003)(2004) had applied principal component analysis of database protein structures to quantify the degree and modes of their flexibility.

Su and Mayo (1997) (2001) claimed that their ORBIT (Optimization of Rotamers By Iterative Techniques) computational protein design process was robust against 15 per cent change in the backbone. Nevertheless, they found out on a later case study on T4 lysosome that core repacking to stabilize the fold was difficult to achieve without considering a flexible template (Mooers *et al.*, 2003). Therefore, to ensure that good sequence solutions are not rejected, it is more desirable to assume backbone flexibility in de novo protein design.

Researchers have formulated several methods to incorporate template variability. First, backbone flexibility can simply be modeled by using a smaller atomic radii in the van der Waals potential. One common practice has been to scale down the radii by five to ten per cent (Handel and Desjarlais, 1995; Kuhlman and Baker, 2000) and thus permitting slight overlaps between atoms due to backbone movements. Key disadvantages of this simple approach include overestimation of the attractive forces and also the possibility of hydrophobic core overpacking.

Another way to allow for backbone flexibility is through considering a discrete set of templates by using genetic algorithms and Monte Carlo sampling. This is the approach adopted by both Handel and Desjarlais (1999) and Kraemer-Pecore *et al.* (2003). Under this approach an ensemble of related backbone conformations close to the template are generated at random. Then a sequence will be designed for each of them under the rigid backbone assumption, and finally the backbone-sequence combination with the lowest energy will be selected. For symmetric proteins backbone structure can actually be modeled by parametric fitting and this will enhance computational efficiency. However, the vast majority of protein structures are non-symmetric which make this parametric approach infeasible. Su and Mayo (1997) overcame this difficulty by treating α-helices and β-sheets as rigid bodies and designing sequences for several template variations of the protein Gβ1. Farinas and Regan (1998) considered a discrete set of templates when they designed the metal binding sites in Gβ1, and they identified varied residue positions that would have been missed if average three-dimensional coordinates had been used for calculations. Harbury *et al.* (1998) incorporated template flexibility through an algebraic parameterization of the backbone when they designed a family of α-helical bundle proteins with right-handed superhelical twist. They were able to achieve a root mean square coordinate deviation between the predicted structure and the actual structure of the de novo designed protein of around $0.2\mathring{A}$.

One natural approach to incorporate backbone flexibility is to allow for variability in each position in the template. The deterministic *in silico* sequence selection method recently proposed by Klepeis *et al.* (2003) (2004) using integer linear optimization technique takes into account template flexibility via the introduction of a distance dependent force field in the sequence selection stage. Pairwise amino acid interaction potential, which depends on both the types of the two amino acids and the distance between them, were used to calculate the total energy of a sequence. Instead of being a continuous function, the dependence of the interaction potential on distance is discretized into bins. With typical bin sizes of 0.5 to $1\mathring{A}$, the overall protein design model Klepeis *et al.* (2004) developed implicitly incorporated backbone movements of roughly the same order of magnitude.

Mathematical Modeling and Optimization Methods

Once an energy function has been defined, sequence selection is accomplished through an optimization based search designed to minimize the energy objective. Both stochastic and deterministic methods have been applied to the computational protein design problem. The Self-Consistent Mean Field (SCMF) (Lee, 1994) and dead-end-elimination (DEE) (Desmet *et al.*, 1992) are both good examples of deterministic methods. The key limitations imposed on the SCMF and DEE are (i) the backbone/template is fixed, and (ii) sequence search is restricted to discrete set of rotamers. In their application of the SCMF method, Koehl and Delarue (1994) (1995)(1996) refined iteratively a conformational matrix whose element $CM(i, j)$ gives the probability that side chain i of a protein takes on rotamer j. Hence $CM(i, j)$ sums to unity over all possible rotamers for a given side chain i. With an initial guess for the conformational matrix, which is usually based on the assumption that all rotamers had the same probability, that is, for rotamer k of residue i:

$$CM(i,k) \;=\; \frac{1}{K_i} \;\; k = 1, 2, ..., K_i \tag{1}$$

the mean-field potential $E(i, k)$ is calculated using (Koehl and Delarue, 1994):

$$E(i,k) \;=\; U(x_{ikC}) \;+\; U(x_{ikC}, x_{0C}) \;+\; \sum_{j=1,j\neq i}^{N} \sum_{l=1}^{K_j} CM(j,l)U(x_{ikC}, x_{jlC}) \tag{2}$$

where x_{0C} corresponds to the coordinates of the atoms in the template, and x_{ikC} corresponds to the coordinates of the atoms of residue i whose conformation is described by rotamer k. Lennard-Jones (12-6) potential can be used for the potential energy U (Koehl and Delarue, 1994). Energies of the K_i possible rotamers of residue i can subsequently be converted into probabilities using Boltzmann law:

$$CM_1(i,k) \;=\; \frac{e^{\frac{-E(i,k)}{RT}}}{\sum_{l=1}^{K_i} e^{\frac{-E(i,l)}{RT}}} \tag{3}$$

$CM_1(i, k)$ provides an update on $CM(i, k)$ which can be used to repeat the calculation of energies and another update on the conformational matrix until convergence is attained. The convergence criterion is usually set as 10^{-4} to define self-consistency (Koehl and Delarue, 1994). In addition, oscillations during convergence could be removed by updating $CM_1(i, k)$ with a "memory" of the previous step (Koehl and Delarue, 1994):

$$CM \;=\; \lambda CM_1 \;+\; (1 - \lambda)CM \tag{4}$$

where optimal step size λ was found to be 0.9 (Koehl and Delarue, 1994). The main disadvantage of SCMF is that though deterministic in nature, it does not guarantee to yield a global minimum in energy (Lee, 1994).

In contrast, DEE assures the convergence to a globally optimal solution consistently. DEE operates on the systematic elimination of rotamers that are not allowable to 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):

$$E \;=\; \sum_{i=1}^{N} E(i_r) \;+\; \sum_{i=1}^{N-1}\sum_{j>i}^{N} E(i_r, j_s) \tag{5}$$

where $E(i_r)$ is the rotamer-template energy for rotamer i_r of amino acid i, $E(i_r, j_s)$ is the rotamer-rotamer energy of rotamer i_r and rotatmer j_s of amino acids i and j respectively, and N is the total number of residues in the protein (Pierce *et al.*, 2000). The pruning criterion in DEE is based on the concept that if the pairwise energy between rotamer i_r and rotamer j_s is higher than that between rotamer i_t and j_s for all j_s in a certain rotamer set $\{S\}$, then i_r cannot be the global energy minimum conformation (GMEC) and thus can be eliminated. Mathematically the idea can be expressed as the following inequality (Voigt *et al.*, 2000):

$$E(i_r) \;+\; \sum_{j\neq i}^{N} E(i_r, j_s) \;>\; E(i_t) \;+\; \sum_{j\neq i}^{N} E(i_t, j_s) \;\; \forall\{S\} \tag{6}$$

So rotamer i_r can be pruned if the above holds true. Bounds implied by (6) can be utilized to generate the following computationally more tractable inequality (Voigt *et al.*, 2000):

$$E(i_r) \;+\; \sum_{j\neq i}^{N} \min_{s} E(i_r, j_s) \;>\; E(i_t) \;+\; \sum_{j\neq i}^{N} \max_{s} E(i_t, j_s) \tag{7}$$

The above inequality can be extended to eliminate pairs of rotamers. This is done by determining a rotamer pair i_r and j_s which always contributes higher energies than rotamer pair i_u and j_v for all possible rotamer combinations. The analogous computationally tractable inequality is (Voigt *et al.*, 2000):

$$\varepsilon(i_r, j_s) \;+\; \sum_{k\neq i,j}^{N} \min_{t} \varepsilon(i_r, j_s, k_t) \;>\; \varepsilon(i_u, j_v) \;+\; \sum_{k\neq i,j}^{N} \max_{t} \varepsilon(i_u, j_v, k_t) \tag{8}$$

where ε is the total energies of rotamer pairs:

$$\varepsilon(i_r, j_s) \;=\; E(i_r) \;+\; E(j_s) \;+\; E(i_r, j_s) \tag{9}$$

$$\varepsilon(i_r, j_s, k_t) \;=\; E(i_r, k_t) \;+\; E(j_s, k_t) \tag{10}$$

The Mayo group has pioneered the development of DEE and has applied the method to design a variety of proteins (Malakauskas and Mayo, 1998) (Strop and Mayo, 1999) (Shimaoka *et al.*, 2000) (Bolon and Mayo, 2001) (Mooers *et al.*, 2003). Goldstein (1994) improved the original DEE criterion by stating that rotamer i_r can be pruned if the energy contribution is always reduced by an alternative rotamer i_t:

$$E(i_r) \;-\; E(i_t) \;+\; \sum_{j\neq i}^{N} \min_{s}[E(i_r, j_s) - E(i_t, j_s)] \;>\; 0 \tag{11}$$

For rotamer pair elimination, the corresponding inequality is (Voigt *et al.*, 2000):

$$\varepsilon(i_r, j_s) \;-\; \varepsilon(i_u, j_v) \;+\; \sum_{k\neq i,j}^{N} \min_{t}[\varepsilon(i_r, j_s, k_t) - \varepsilon(i_u, j_v, k_t)] \;>\; 0 \tag{12}$$

In general, rotamer pair elimination is computationally more expensive than single rotamer elimination, and methods have been developed by Gordon and Mayo (1998) to predict which doubles elimination inequalities are the strongest.

Pierce *et al.* (2000) introduced Split DEE which split the conformational space into partitions and thus eliminated the dead-ending rotamers more efficiently:

$$E(i_r) \; - \; E(i_t) \; + \sum_{j,j\neq k\neq i}^{N} \{\min_{u}[E(i_r, j_u) - E(i_t, j_u)]\} \; + \; [E(i_r, k_v) - E(i_t, k_v)] \; > \; 0 \qquad (13)$$

Further revisions and improvements on DEE had been performed by Wernisch *et al.* (2000) and Gordon *et al.* (2003).

The protein design problem has been proved to be *NP*-hard (Pierce and Winfree, 2002), which means the time required to solve the problem varies exponentially according to n^m, where n is the average number of amino acids to be considered per position and m is the number of residues. Hence as the protein becomes big enough, deterministic methods may reach a plateau, and this is when stochastic methods come into play. Monte Carlo methods and genetic algorithms are the most commonly used stochastic methods for de novo protein design. In Monte Carlo methods, a mutation is performed at a certain position in the sequence and the Boltzmann probability calculated from the energies before and after the mutation, as well as temperature is compared to a random number. The mutation is allowed if the Boltzmann probability is higher than the random number, and rejected otherwise. Dantas *et al.* (2003)'s protein design computer program, RosettaDesign, applied Monte Carlo optimization algorithms. In completely redesigning nine globular proteins, RosettaDesign yielded sequences of $70-80\%$ identity as the final results of energy optimization when multiple runs were started with different random sequences (Dantas *et al.*, 2003). Originated in genetics and evolution, genetic algorithms generate a multitude of random amino acid sequences and exchange for a fixed template. Sequences with low energies form hybrids with other sequences while those with high energies are eliminated in an iterative process which only terminates when a converged solution is attained (Tuffery *et al.*, 1991). Handel and Desjarlais (1999) have applied a two-stage combination of Monte Carlo and genetic algorithms to design the hydrophobic core of protein 434cro. Both Monte Carlo methods and genetic algorithms can search larger combinatorial space compared to deterministic methods, but they share the common disadvantage of lacking consistency in finding the global minimum in energy.

Recent methods attempt to avoid the problem of optimizing residue interactions by manipulation of the shapes of free energy landscapes (Jin *et al.*, 2003). Another class of methods focus on a statistical theory for combinatorial protein libraries which provides probabilities for the selection of aminoacids in each sequence position (Zhou and Saven, 2000; Kono and Saven, 2001; Saven, 2003). The set of site-specific amino acid probabilities obtained at the end actually represents the sequence with the maximum entropy subject to all of the constraints imposed (Zhou and Saven, 2000; Kono and Saven, 2001; Park *et al.*, 2004). This statistical computationally assisted design strategy (*scads*) has been employed to characterize the structure and functions of membrane protein KcsA and to enhance the catalytic activity of a protein with dinuclear metal center (Park *et al.*, 2004). It has also been used to calculate the identity probabilities of the varied positions in the immunoglobulin light chain-binding domain of protein L (Kono and Saven, 2001). *Scads* serves as a useful framework for interpreting and designing protein combinatorial libraries, as it provides clues about the regions of the sequence space that are most likely to produce well-folded structures (Hecht *et al.*, 2004).

Several sequence selection approaches have been tested and validated by experiment, thereby firmly establishing the feasibility of computational protein design. The first computational design of a full sequence to be experimentally characterized was the achievement of a stable zinc-finger fold ($\beta\beta\alpha$) using a combination of a backbone-dependent rotamer library with atomistic level modeling and a dead-end elimination based algorithm (Dahiyat and Mayo, 1997). Recently, Kuhlman *et al.* (2003) introduced a computational framework that iterates between sequence design and structure prediction, designed a new fold for a 93-residue α/β protein, and validated its fold and stability experimentally. Despite these accomplishments, the development of a computational protein design technique to rigorously address the problems of fold stability and functional design remains a challenge. One important reason for this is, as mentioned earlier, either the almost universal specification of a fixed backbone or the use of a discrete set of backbones, which does not allow for the true flexibility that would afford more optimal sequences and more robust predictions of stability. Moreover, several models which attempt to incorporate backbone flexibility highlight a second difficulty, namely, inadequacies inherent to energy modeling (Desjarlais and Handel, 1999). The need for empirically derived weighting factors, and the dependence on specific heuristics limit the generic nature of these computational protein design methods. Such modeling based assumptions also raise issues regarding the appropriateness of the optimization method and underscore the question of whether it is sufficient to merely identify the globally optimal sequence or, more likely, a subset of low lying energy sequences. An even more difficult problem relevant to both flexibility and energy modeling is to correctly model the interactions which control the functionality and activity of the designed sequences.

De Novo Protein Design Framework

In Klepeis *et al.* (2003) (2004), a novel two-stage computational peptide and protein design method is presented to not only select and rank sequences for a particular fold but also to validate the stability and specificity of the fold for these selected sequences. The sequence selection phase relies on a novel integer linear programming (ILP) model with several important constraint modifications that improve the tractability of the problem and enhance its deterministic convergence to the global minimum. In addition, a rank-ordered list of low lying energy sequences are identified along with the global minimum energy sequence. Once such a subset of sequences have been identified, the fold validation stage is employed to verify the stabilities and specificities of the designed sequences through a deterministic global optimization approach that allows for backbone flexibility. The selection of the best designed sequences is based on rigorous quantification of energy based probabilities. In the sequel, we will discuss the two stages in detail.

In silico Sequence Selection

To correctly select a sequence compatible with a given backbone template, an appropriate energy function must first be identified. Desirable properties of energy models for protein design include both accuracy and rapid evaluation. Moreover, the functions should not be overly sensitive to fixed backbone approximations. In certain cases, additional requirements, such as the pairwise decomposition of the potential for application of the dead-end elimination algorithm (Desmet *et al.*, 1992), may be necessary.

Instead of employing a detailed atomistic level model, which requires the empirical reweighting of energetic terms, the proposed sequence selection procedure is based on optimizing a pairwise *distance-dependent* interaction potential. Such a statistically based empirical energy function assigns energy values for interactions between amino acids in the protein based on the alpha-carbon separation distance for each pair of amino acids. Such structure based pairwise potentials are fast to evaluate, and have been used in fold recognition and fold prediction (Park and Levitt, 1996). One advantage of this approach is that there is no need to derive empirical weights to account for individual residue propensities. Moreover, the possibility that such interaction potentials lack sensitivity to local atomic structure are addressed within the context of the overall two-stage approach. In fact, the coarser nature of the energy function in the *in silico* sequence selection phase may prove beneficial in that it allows for an inherent flexibility to the backbone.

A number of different parameterizations for pairwise residue interaction potentials exist. The simplest approach is the development of a binary version of the model such that each contact between two amino acids is assigned according to the residues types and the requirement that a contact is defined as the separation between the side chains of two amino acids being less than 6.5 $\mathring{A}$ (Meller and Elber, 2001). An improvement of this model is based on the incorporation of a distance dependence for the energy of each amino acid interaction. Specifically, the alpha-carbon distances are discretized into a set of 13 bins to create a finite number of interactions, the parameters of which were derived from a linear optimization formulated to favor native folds over optimized decoy structures (Tobi and Elber, 2000; Tobi *et al.*, 2000). The use of a distance dependent potential allows for the implicit inclusion of side chains and the specificity of amino acids. The resulting potential, which involves 2730 parameters, was shown to provide higher Z scores than other potentials and place native folds lower in energy (Tobi and Elber, 2000; Tobi *et al.*, 2000).

The linearity of the resulting formulation based on this distance-dependent interaction potential (Loose *et al.*, 2003) is also an attractive characteristic of the *in silico* sequence selection procedure. The development of the formulation can be understood by first describing the variable set over which the energy function is optimized. First, consider the set $i = 1, \ldots, n$ which defines the number of residue positions along the backbone. At each position i there can be a set of mutations represented by $j\{i\} = 1, \ldots, m_i$, where, for the general case $m_i = 20 \forall i$. The equivalents sets $k \equiv i$ and $l \equiv j$ are defined, and $k > i$ is required to represent all unique pairwise interactions. With this in mind, the binary variables y_i^j and y_k^l can be introduced to indicate the possible mutations at a given position. That is, the y_i^j variable will indicate which type of amino acid is active at a position in the sequence by taking the value of 1 for that specification. Then, the formulation, for which the goal is to minimize the energy according to the parameters that multiply the binary variables, can be expressed as :

$$\begin{array}{rc}
\min\limits_{y_i^j, y_k^l} & \sum_{i=1}^{n} \sum_{j=1}^{m_i} \sum_{k=i+1}^{n} \sum_{l=1}^{m_k} E_{ik}^{jl}(x_i, x_k) y_i^j y_k^l \\
\text{subject to} & \sum_{j=1}^{m_i} y_i^j \ = \ 1 \ \forall \ i \\
 & y_i^j \ , \ y_k^l \ = \ 0-1 \ \forall \ i,j,k,l
\end{array}$$

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 the type of amino acids at those positions. The composition constraints require that there is exactly one type of amino acid at each position. For the general case, 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 (Floudas, 1995):

$$\begin{array}{rc}
\min\limits_{y_i^j, y_k^l} & \sum_{i=1}^{n} \sum_{j=1}^{m_i} \sum_{k=i+1}^{n} \sum_{l=1}^{m_k} E_{ik}^{jl}(x_i, x_k) w_{ik}^{jl} \\
\text{subject to} & \sum_{j=1}^{m_i} y_i^j \ = \ 1 \ \forall \ i \\
 & y_i^j \ + \ y_k^l \ - \ 1 \ \leq \ w_{ik}^{jl} \ \leq \ y_i^j \ \forall \ i,j,k,l \\
 & 0 \ \leq \ w_{ik}^{jl} \ \leq \ y_k^l \ \forall \ i,j,k,l \\
 & y_i^j \ , \ y_k^l \ = \ 0-1 \ \forall \ i,j,k,l
\end{array}$$

This reformulation relies on the transformation of the bilinear combinations to a new set of linear variables, w_{ik}^{jl}, while the addition of the four sets of constraints serves to reproduce the characteristics of the original formulation. For example, for a given i, j, k, l combination, the four constraints require w_{ik}^{jl} to be zero when either y_i^j or y_k^l is equal (or when both are equal to zero). If both y_i^j and y_k^l are equal to one then w_{ik}^{jl} is also enforced to be one.

The solution of the integer linear programming problem (ILP) can be accomplished rigorously using branch and bound techniques (CPLEX, 1997) (Floudas, 1995) making convergence to the global minimum energy sequence consistent and reliable. Furthermore, the performance of the branch and bound algorithm is significantly enhanced through the introduction of reformulation linearization techniques (RLT). Here, the basic strategy is to multiply appropriate constraints by bounded non-negative factors (such as the reformulated variables) and introduce the products of the original variables by new variables in order to derive higher-dimensional lower bounding linear programming (LP) relaxations for the original problem (Sherali and Adams, 1999). These LP relaxations are solved during the course of the overall branch and bound algorithm, and thus speed convergence to the global minimum. The following set of constraints illustrates the application of the RLT approach to the original composition constraint. First, the equations are reformulated by forming the product of the equation with some binary variables or their complement. For example, by multiplying by the set of variables y_k^l, the following additional set of constraints $\forall \;\; j, k, l$ is produced:

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

This equation can now be linearized using the same variable substitution as introduced for the objective. The set of RLT constraints then become:

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

Finally, for such an ILP problem it is straightforward to identify a rank ordered list of the low lying energy sequences through the introduction of integer cuts (Floudas, 1995), and repetitive solution of the ILP problem. By using the enhancements outlined above, in combination with the commercial (LP) solver CPLEX (CPLEX, 1997), a globally optimal (ILP) solution is generated.

Fold Specificity

Once a set of low lying energy sequences have been identified via the sequence selection procedure, the fold stability and specificity validation stage is used to identify the most optimal sequences according to a rigorous quantification of conformational probabilities. The foundation of the approach is grounded on the development of conformational ensembles for the selected sequences 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 are likely to be incorporated (in the case of compstatin and its analogs only the disulfide bridge constraint is enforced) and with the underlying template structure not being enforced. In terms of practical considerations, the distance constraints introduced for the template constrained simulation can be based on the structural boundaries defined by the NMR ensemble (in the case of compstatin and its analogs a deviation of 1.5 angstroms is allowed for each non-consecutive Cα-Cα distance from the known NMR structures), or simply by allowing some deviation from a subset of distances provided by the structural template, and hence they allow for a flexible template on the backbone.

The formulations for the folding calculations are reminiscent of structure prediction problems in protein folding (Klepeis *et al.*, 2002). In particular, a novel constrained global optimization problem first introduced for structure prediction using NMR data (Klepeis *et al.*, 1999), and later employed in a generic framework for the structure prediction of proteins (Klepeis and Floudas, 2003) is employed. The global minimization of a detailed atomistic energy forcefield E_{ff} is performed over the set of independent dihedral angles, ϕ, which can be used to describe any possible configuration of the system. The bounds on these variables are enforced by simple box constraints. Finally, a set of distance constraints, $E_l^{dis} \;\; l = 1, \ldots, N$, which are nonconvex in the internal coordinate system, can be used to constrain the system. The formulation is represented by the following set of equations:

$$\begin{array}{rrcll} \min\limits_{\phi} & & E_{ff} & & \\ \text{subject to} & E_j^{dis}(\phi) & \leq & E_j^{ref} & j = 1, \ldots, N \\ & \phi_i^L \;\leq & \phi_i & \leq \; \phi_i^U & i = 1, \ldots, N_\phi \end{array}$$

Here, $i = 1, \ldots, N_\phi$ corresponds to the set of dihedral angles, ϕ_i, with ϕ_i^L and ϕ_i^U representing lower and upper bounds on these dihedral angles. In general, the lower and upper bounds for these variables are set to -π and π. E_j^{ref} are reference parameters for the distance constraints, which assume the form of typical square well potential for both upper and lower distance violations. The set of constraints are completely general, and can represent the full combination of distance constraints or smaller subsets of the defined restraints. The forcefield energy function, E_{ff} can take on a number of forms, although the current work employs the ECEPP/3 model (Némethy *et al.*, 1992).

The folding formulation represents a general nonconvex constrained global optimization problem, a class of problems for which several methods have been developed. In this work, the formulations are solved via the αBB deterministic global optimization approach, a branch and bound method applicable to the identification of the global minimum of nonlinear optimization problems with twice–differentiable functions (Adjiman *et al.*, 1998a,b, 2000; Klepeis *et al.*, 1999; Klepeis and Floudas, 1999; Floudas, 2000; Klepeis *et al.*, 2002). A converging sequence of upper and lower bounds is generated, with the upper bounds on the global minimum obtained by local minimizations of the original nonconvex problem, while the lower bounds belong to the set of solutions of the convex lower bounding problems that are constructed by augmenting the objective and constraint functions through the addition of separable quadratic terms.

In addition to identifying the global minimum energy conformation, the global optimization algorithm provides the means for identifying a consistent ensemble of low energy conformations (Klepeis and Floudas, 1999; Klepeis *et al.*, 2003a,b). Such ensembles are useful in deriving quantitative comparisons between the free folding and template-constrained simulations. In this way, the complications inherent to the specification of an appropriate reference state are avoided because a relative probability is calculated for each sequence studied during this stage of the approach. The relative probability for template stability, p_{temp}, can be found by summing the statistical weights for those conformers from the free folding simulation that resemble the template structure (denote 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*):

$$p_{temp} \quad = \quad \frac{\sum_{i \in temp} \exp[-\beta E_i]}{\sum_{i \in total} \exp[-\beta E_i]}$$

where $\exp[-\beta E_i]$ is the statistical weight for conformer i.

Computational and Experimental Findings

Compstatin

The target chosen to test the novel protein design framework proposed by Klepeis *et al.* (2003) is Compstatin. Compstatin is a 13-residue cyclic peptide that has the ability to inhibit the cleavage of C3 to C3a and C3b. The effect of targeting the C3 cleavage is triple and results to hindrance in: (i) the generation of the pro-inflammatory peptide C3a, (ii) the generation of opsonin C3b (or its fragment C3d), and (iii) further complement activation of the common pathway (beyond C3) with end result the generation of the membrane attack complex (MAC). A C3-binding complement inhibitor was identified as a 27-residue peptide using a phage-displayed random peptide library (Sahu *et al.*, 1996). This peptide was truncated to an equally active 13-residue peptide named compstatin with sequence I[CVVQDWGHHRC]T-NH2 , where the brackets denote cyclization through a disulfide bridge formed by Cys2-Cys12 (Sahu *et al.*, 1996) (Morikis *et al.*, 1998). Acetylation of the N-terminus of compstatin (Ac-compstatin) resulted to a 3-fold increase in activity (Sahu *et al.*, 2000) (Morikis *et al.*, 2002) (Soulika *et al.*, 2003).

Compstatin blocked the cleavage of C3 to the pro-inflammatory peptide C3a and the opsonin C3b in hemolytic assays and in human normal serum (Sahu *et al.*, 1996) (Sahu *et al.*, 2000), prevented heparine/protamine-induced complement activation in baboons in a situation resembling heart surgery (Soulika *et al.*, 2000), inhibited complement activation during the contact of blood with biomaterial in a model of extra-corporeal circulation (Nillson *et al.*, 1998), increased the lifetime of survival of porcine kidneys perfused with human blood in a hyper-acute rejection xenotransplantation model (Fiane *et al.*, 1999), blocked the E coli -induced oxidative burst of granulocytes and monocytes (Mollnes *et al.*, 2002), and inhibited complement activation by cell lines SH-SY5Y, U-937, THP-1 and ECV304 (Klegeris *et al.*, 2002). Compstatin was stable in biotransformation studies in vitro in human blood, normal human plasma and serum, with increased stability upon N-terminal acetylation (Sahu *et al.*, 2000). Compstatin showed little or low toxicity and no adverse effects when these were measured (Fiane *et al.*, 1999) (Nillson *et al.*, 1998) (Soulika *et al.*, 2000). Finally, compstatin showed species-specificity and is active only with human and primate C3 (Sahu *et al.*, 2003).

In silico Sequence Selection

The first stage of the design approach involves the selection of sequences compatible with the backbone template through the solution of the ILP problem. The formulation relies only on the alpha-carbon coordinates of the backbone residues, which were taken from the NMR-average solution structure of compstatin (Morikis *et al.*, 1998).

A full computational design study from compstatin would result in a combinatorial search of $20^{13} \approx 8 \times 10^{16}$ sequences. However, in light of the results of the experimental studies of the rationally designed peptides, a directed, rather than full, set of computational design studies were performed. First, since the disulfide bridge was found to be essential for aiding in the formation of the hydrophobic cluster and prohibiting the termini from drifting apart, both residues Cys^2 and Cys^{12} were maintained. In addition, because the structure of the type-I β turn was not found to be a sufficient condition for activity, the turn residues were fixed to be those of the parent compstatin sequence; namely Gln^5-Asp^6-Trp^7-Gly^8. In fact, when stronger type I β sequences were constructed, which was supported by NMR data indicating that these sequences provided higher β turn populations than compstatin, these sequences resulted in lower or no activity (Morikis *et al.*, 2002). Therefore, the further stabilization of the turn residues, which would likely be a consequence of the computational peptide design procedure, may not enhance compstatin activity. This is especially true for Trp^7, which was found to be a likely candidate for direct interaction with C3. For similar reasons, Val^3 was maintained throughout the computational experiments.

After designing the compstatin system to be consistent with those features found to be essential for compstatin activity, six residue positions were selected to be optimized. Of these six residues, positions 1, 4, and 13 have been shown to be structurally involved in the formation of a hydrophobic cluster involving residues at positions 1, 2, 3, 4, 12, and 13, a necessary but not sufficient component for compstatin binding and activity. The remaining residues, namely those at positions 9, 10 and 11, span the three positions between the turn residues and the C-terminal cystine. For the wild type sequence these positions are populated by positively charged residues, with a total charge of +2 coming from two histidine residues and one arginine residue.

Based on the structural and functional characteristics of those residues involved in the hydrophobic cluster, a base case was studied with positions 1, 4 and 13 selected only from those residues defined as belonging to the hydrophobic set (A,F,I,L,M,V,Y). In addition, this set included threonine for position 13 to allow for the selection of the wild type residue at this position. For positions 9, 10 and 11 in the base case, all residues were allowed, excluding cystine and tryptophan. In view of the experimental studies on Compstatin by Mallik *et al.* (2005) who proposed tryptophan (W) or fused-ring non-natural amino acids at position 4 would contribute to high inhibitory activity of the peptide, an additional run was performed with the inclusion of tryptophan (W) in the selection set for position 4. Table 2 summarizes the preferred selection at each position according to the composition of the lowest lying energy sequences. Tryptophan (W) was indeed strongly favored at position 4 if it was included in the selection set. This observation agrees with the experimental finding by Mallik *et al.* (2005). It should be noted that if tryptophan (W) is allowed to be in the aforementioned hydrophobic set for all six positions, then sequences with tryptophan (W) in position 4 and alanine (A), or phenyl (F), or tryptophan (W) in position 9 are predicted among the most promising ones by the proposed novel in silico sequence selection framework (position 1 is I, position 10 is R and position 13 is T, as in set D of Figure 1).

The sequence selection results exhibit several important and consistent features. First, position 10 is dominated by the selection of a histidine residue, a result that directly reinforces the composition of the wild type compstatin sequence. In contrast, position 11 is found to have the largest variation in composition, with both polar, hydrophobic and charged residue being part of the set of optimal low lying energy sequences. At position 9, a subset of those residues chosen for position 11, are selected. When considering those positions involved in the hydrophobic cluster of compstatin, it is evident that valine provides strong forces at each position. However, the results for position 4 contrast with those at position 1 and 13 in that tyrosine, rather than valine, is the preferred choice for the lowest as well as a large majority of the low lying energy sequences.

It should be noted that because the compstatin structure was determined via NMR methods, there exists an ensemble of 21 structures for which alternative templates could be derived. These alternative templates were studied as a means of incorporating backbone flexibility into the sequence selection process, and the results proved to be consistent and in qualitative agreement with those for the average template structure.

Fold specificity calculations for selected sequences

Based on the sequence selection results a handful of optimal sequences were constructed for use in the second stage of the computational design procedure. Figure 1 presents that peptides studied which are further classified into sets A, B, C and D.

For all sequences further characterized via the fold stability calculations, residue 10 was set to histidine, a prediction consistent with the composition of the parent peptide sequence. Moreover, since the variation in the residue composition for position 11 is predicted to be rather broad, position 11 was restricted to be arginine in subsequent sequences (except Set C). The first set of sequences was constructed to better analyze the effect of the tyrosine substitution at position 4, with the justification to focus on this substitution being an attempt to assess the unusually dominant selection of tyrosine at position 4. The consistent element of the sequences belonging to set A is the assignment of tyrosine to position 4. To further isolate any substitution with respect to the parent peptide sequence, sequences A1, A2 and A3 assume the parent compstatin composition of histidine at position 9. Moreover, sequence A1 resembles the parent peptide sequence at positions 1 and 13 as well, while sequences A2 and A3 are constructed so as to add the valine substitutions incrementally; first at position 13 for sequence A2 and then at both positions 1 and 13 for sequence A3. Sequences A1 and A3 exhibit substantial increases in fold stability over the parent peptide sequence (Table 1). These results highlight the significance of the tyrosine substitution at position 4, and may help to further clarify certain features of the proposed binding model for the compstatin-C3 complex (Morikis *et al.*, 2002).

To further explore the combination of position 9 substitutions with the presence of tyrosine at position 4, several additional sequences were constructed. The B1 and B2 constructions represent a reduction in the number of simultaneous mutations from the parent peptide sequence. In effect these two sequences correspond to the individual combinations of sequence A2 with both sequence A4 and sequence A5 such that position 1 is taken from sequence A2, while position 9 matches the substitutions incorporated into sequences A4 and A5. An additional sequence, B3, is formulated as a combination of sequence A3 and the position 9 substitution of histidine to tryptophan as taken from control sequence X2. Each of the three designed sequences demonstrate significant increases in fold stability relative to the original compstatin sequence (Table 1).

Another set of two additional sequences were identified with the only difference between them being the specification of the residue at position 4. For sequence C1, tyrosine was assigned to position 4, while sequence C2 was selected to have valine at this position. For both sequences, threonine was specified at positions 9 and 11, while positions 1 and 13 were set to isoleucine and valine, respectively. The choice of isoleucine for position 1 helps to reduce the number of simultaneous changes from the parent peptide sequence.

For both sequence C1 and sequence C2 the stability calculations indicate a substantial decrease in stability when compared to the parent peptide sequence. Nevertheless, between sequence C1 and C2 there is strong evidence for the preference of tyrosine at position 4. This prompted closer examination of the residue selections at position 9 and position 11, the two remaining positions not involved in the hydrophobic clustering of compstatin. In particular, the specification of threonine at both positions 9 and 11 results in a negative net charge balance due to the aspartate at position 6, especially because of the replacement of arginine by threonine at position 11. This validates further the placement of arginine at position 11 for the previous set of sequences (Table 1).

The final set of sequences was designed in accordance with additional reductions in the number of simultaneous mutations relative to the parent peptide sequence. Specifically, sequence D1 and sequence D2, resemble sequence B1 and sequence B2 with threonine instead of valine as the C-terminal residue, a specification matching the composition of the original parent peptide sequence. Both sequences provide significant increases in fold stability. For sequences D1 and D2 the differences with respect to the parent peptide sequence are isolated to the residue before and after the β turn. Both the position 4 tyrosine and position 9 phenylalanine substitutions provide enhancements to the fold stability of the compstatin structure, and represent unforeseen and unpredictable enhancements over the parent peptide sequence (Table 1).

Experimental Validation

A number of the designed sequences presented above were constructed and tested experimentally for their activity, without performing NMR-based structural analyses. Since the ultimate goal is to enhance the functional activity of compstatin, such achievements must be complemented and verified through experimental studies. Rather than performing massive chemical synthesis of peptide analogs, a few selected analogs were tested against the theoretical prediction. Table 1 shows the experimentally measured percent complement inhibition and peptide D1 is currently the most active compstatin analog available. The C2A/C12A analog is inactive (Morikis *et al.*, 2002) and has been used as a negative control for the inhibition measurements. Table 1 summarizes the results from the inhibitory activity experiments in comparison to the theoretical fold stability results.

Qualitatively, the predicted increases in fold stability and specificity are in excellent agreement with the results from the experimental studies. This is especially significant given that the predictions correspond more directly to fold stability enhancements while the experiments directly test inhibitory function.

The comparison between experimental and computational results indicate that the most active compstatin analogs are sequences D1 and B1, as suggested by the optimization study. The common characteristic of these two sequences is the substitutions at positions 4 and 9, the two positions flanking the β turn residues, Gln^5-Asp^6-Trp^7-Gly^8. In particular, the combination of tyrosine at position 4 and alanine at position 9 are key residues for increased activity and lead to an 16-fold improvement over the parent peptide compstatin (see Table 1).

Conclusions and Future Work

A novel computational structure-activity based methodology for the de novo design of peptides and proteins was presented. The method is completely general in nature, with the main steps of the approach being the availability of NMR-derived structural templates, combinatorial selection of sequences based on optimization of parameterized pairwise residue interaction potentials and validation of fold stability and specificity using deterministic global optimization. The optimization study led to the identification of many active analogs including a 16-fold more active analog, as validated through immunological activity measurements. Allowing tryptophan in position 4, the in silico sequence prediction framework demonstrates that tryptophan is preferred over tyrosine and tyrosine is preferred over valine. This is in agreement with recent experimental results (Mallik *et al.*, 2005) which showed a 45-fold improvement in the inhibitory activity of the peptide ($Ac - I[CVWQDWGAHRC]T - NH_2$). These results are extremely impressive and represent significant enhancements in inhibitory activity over analogs identified by either purely rational or experimental combinatorial design techniques. The work provides direct evidence that an integrated experimental and theoretical approach can make the engineering of compounds with enhanced immunological properties possible. Future work will be focused on algorithmic improvement on the novel de novo protein design framework to enhance computational efficiency, trial and incorporation of non-energy-based formulations to increase accuracy of predictions, and application of the framework on more protein targets.

Acknowledgments

CAF gratefully acknowledges financial support from the National Science Foundation and the National Institutes of Health (R01 GM52032).

References

C. Adjiman, I. Androulakis, and C. A. Floudas. Global Optimization of Mixed-Integer Nonlinear Problems. *AiChE Journal*, 46:1769–1797, 2000.

C. Adjiman, I. Androulakis, and C.A. Floudas. A Global Optimization Method, αBB, for General Twice-differentiable Constrained NPLs - I. Theoretical Advances. *Computers Chem. Eng.*, 22:1137, 1998a.

C. Adjiman, I. Androulakis, and C.A. Floudas. A Global Optimization Method, αBB, for General Twice-differentiable Constrained NPLs - II. Implementation and Computational Results. *Computers Chem. Eng.*, 22:1159, 1998b.

D.N. Bolon and S.L. Mayo. Enzyme-like Proteins by Computational Design. *Proc. Natl. Acad. Sci. USA*, 98:14274–14279, 2001.

T.V. Pogorelov C. Hardin and Z. Luthey-Schulten. Ab initio protein structure prediction. *Curr. Opin. Struc. Biol.*, 12:176–181, 2002.

CPLEX. *Using the CPLEX Callable Library.* ILOG, Inc., 1997.

B.I. Dahiyat, D.B. Gordon, and S.L. Mayo. Automated design of the surface positions of protein helices. *Protein Sci.*, 6:1333–1337, 1997.

B.I. Dahiyat and S.L. Mayo. Protein design automation. *Protein Sci.*, 5:895–903, 1996.

B.I. Dahiyat and S.L. Mayo. De novo protein design: fully automated sequence selection. *Science*, 278:82–87, 1997.

G. Dantas, B. Kuhlman, D. Callender, M. Wong, and D. Baker. A Large Scale Test of Computational Protein Design: Folding and Stability of Nine Completely Redesigned Globular Proteins. *J. Mol. Biol.*, 332:449–460, 2003.

J.R. Desjarlais and T.M. Handel. De novo design of the hydrophobic cores of proteins. *Protein Sci.*, 4:2006–2018, 1995.

J.R. Desjarlais and T.M. Handel. Side chain and backbone flexibility in protein core design. *J. Mol. Biol.*, 290:305–318, 1999.

J. Desmet, M. Maeyer, B. Hazes, and I. Lasters. The Dead-end Elimination Theorem and Its Use in Protein Side-chain Positioning. *Nature*, 356:539–542, 1992.

K.E. Drexler. Molecular engineering: an approach to the development of general capabilities for molecular manipulation. *Proc. Natl. Acad. Sci. USA*, 78:5275–5278, 1981.

E.G. Emberly, R. Mukhopadhyay, C. Tang, and N.S. Wingreen. Flexibility of α-helices: Results of a Statistical Analysis of Database Protein Structures. *J. Mol. Biol.*, 327:229–237, 2003.

E.G. Emberly, R. Mukhopadhyay, C. Tang, and N.S. Wingreen. Flexibility of β-Sheets: Principal Component Analysis of Database Protein Structures. *Proteins: Structure, Function, and Genetics*, 55:91–98, 2004.

E. Farinas and L. Regan. The De Novo Design of a Rubredoxin-like Fe Site. *Protein Science*, 7:1939–1946, 1998.

A.E. Fiane, T.E. Mollnes, V. Videm, T. Hovig, K. Hogasen, O.J. Mellbye, L. Spruce, W.T. Moore, A. Sahu, and J.D. Lambris. Compstatin, a peptide inhibitor of C3, prolongs survival of ex-vivo perfused pig xenografts. *Xenotransplantation*, 6:52–65, 1999.

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

C.A. Floudas. *Determistic Global Optimization: Theory, Methods and Applications.* Kluwer Academic Publishers, 2000.

R.F. Goldstein. Efficient Rotamer Elimination Applied to Protein Side-chains and Related Spin Glasses. *Biophysics Journal*, 66:1335–1340, 1994.

B.B. Gordon, G.K. Hom, S.L. Mayo, and N.A. Pierce. Exact Rotamer Optimization for Protein Design. *J. Computational Chemistry*, 24:232–243, 2003.

D.B. Gordon and S.L. Mayo. Radical Performance Enhancements for Combinatorial Optimization Algorithms Based on the Dead-end Elimination Theorem. *J. Comput. Chem.*, 19:1505–1514, 1998.

T.M. Handel and J.R. Desjarlais. De Novo Design of the Hydrophobic Cores of Proteins. *Protein Science*, 4:2006–2018, 1995.

T.M. Handel and J.R. Desjarlais. Side-chain and Backbone Flexibiity in Protein Core Design. *J. Mol. Biol.*, 290:305–318, 1999.

P.B. Harbury, J.J. Plecs, B. Tidor, T. Alber, and P.S. Kim. High-resolution Protein Design With Backbone Freedom. *Science*, 282:1462–1467, 1998.

M.H. Hecht, A. Das, A. Go, L.H. Bradley, and Y. Wei. De Novo Proteins From Designed Combinatorial Libraries. *Protein Science.*, 13:1711–1723, 2004.

H.W. Hellinga and F.M. Richards. Construction of new ligand binding sites in proteins of known structure I. Computer aided modeling of sites with predefined geometry. *J. Mol. Biol.*, 222:763–785, 1991.

W. Jin, O. Kambara, H. Sasakawa, A. Tamura, and S. Takada. De Novo Design of Foldable Proteins with Smooth Folding Funnel: Automated Negative Design and Experimental Verification. *Structure*, 11:581–590, 2003.

A. Klegeris, E.A. Singh, and P.L. McGeer. Effects of C-reactive protein and pentosan polysulphate on human complement activation. *Immunology*, 106:381–388, 2002.

J. L. Klepeis, C. A. Floudas, D. Morikis, and J. Lambris. Predicting Peptide Structures Using NMR Data and Deterministic Global Optimization. *J. Comput. Chem.*, 20:1354–1370, 1999.

J. L. Klepeis, C. A. Floudas, D. Morikis, C. G. Tsokos, E. Argyropoulos, L. Spruce, and J. D. Lambris. Integrated Structural, Computational and Experimental Approach for Lead Optimization: Deisgn of Compstatin Variants with Improved Activity. *J. Am. Chem. Soc.*, 125:8422–8423, 2003.

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 R. A. Friesner, editor, *Advances in Chemical Physics*, volume 120, pages 254–457. Wiley, 2002.

J.L. Klepeis and C.A. Floudas. Free Energy Calculations for Peptides Via Deterministic Global Optimization. *J. Chem. Phys.*, 110:7491, 1999.

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

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.*, 125:8422–8423, 2003.

J.L. Klepeis, C.A. Floudas, D. Morikis, C.G. Tsokos, and J.D. Lambris. Design of Peptide Analogs with Improved Activity Using a Novel De Novo Protein Design Approach. *Industrial & Engineering Chemistry Research*, 43:3817, 2004.

J.L. Klepeis, M.T. Pieja, and C.A. Floudas. A New Class of Hybrid Global Optimization Algorithms for Peptide Structure Prediction: Integrated Hybrids. *Comp. Phys. Comm.*, 151:121–140, 2003a.

J.L. Klepeis, M.T. Pieja, and C.A. Floudas. Hybrid Global Optimization Algorithms for Protein Structure Prediction : Alternating Hybrids. *Biophysical J.*, 84:869–882, 2003b.

P. Koehl and M. Delarue. Application of a Self-consistent Mean Field Theory to Predict Protein Side-chains Conformation and Estimate Their Conformational Entropy. *J. Mol. Biol.*, 239:249–275, 1994.

P. Koehl and M. Delarue. A Self Consistent Mean Field Approach to Simultaneouos Gap Closure and Side-chain Positioning in Homology Modeling. *Nature Struct. Biol.*, 2:163–170, 1995.

P. Koehl and M. Delarue. Mean-field Minimization Methods for Biological Macromolecules. *Current Opinion in Structural Biology*, 6:222–226, 1996.

P. Koehl and M. Levitt. De novo protein design I. In search of stability and specificity. *J. Mol. Biol.*, 293:1161–1181, 1999.

H. Kono and J.G. Saven. Statistical Theory for Protein Combinatorial Libraries. Packing Interactions, Backbone Flexibility, and the Sequence Variability of a Main-chain Structure. *J. Mol. Biol.*, 306:607–628, 2001.

C.M. Kraemer-Pecore, J.T. Lecomte, and J.R. Desjarlais. A De Novo Redesign of the WW Domain. *Protein Science.*, 12:2194–2205, 2003.

B. Kuhlman and D. Baker. Native Protein Sequences Are Close to Optimal for Their Structures. *Proc. Natl. Acad. Sci. USA*, 97:10383–10388, 2000.

B. Kuhlman, G. Dantae, G.C. Ireton, G. Verani, B. Stoddard, and D. Baker. Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. *Science*, 302:1364–1368, 2003.

C. Lee. Predicting Protein Mutant Energetics by Self-Consistent Ensemble Optimization. *J. Mol. Biol.*, 236:918–939, 1994.

W.A. Lim, A. Hodel, R.T. Sauer, and F.M. Richards. The Crystal Structure of a Mutant Protein With Altered But Improved Hydrophobic Core Packing. *Proc. Natl. Acad. Sci. USA*, 91:423–427, 1994.

C. Loose, J. Klepeis, and C. Floudas. A new pairwise folding potential based on improved decoy generation and side chain packing. *Proteins*, 2003. in press.

S.M. Malakauskas and S.L. Mayo. Design, Structure, and Stability of a Hyperthermophilic Protein Variant. *Nat. Struct. Biol.*, 5:470–475, 1998.

B. Mallik, M. Katragadda, L.A. Spruce, C. Carafides, C.G. Tsokos, D. Morikis, and J.D. Lambris. Design and NMR Characterization of Active Analogues of Compstatin Containing Non-natural Amino Acids. *Journal of Medicinal Chemistry.*, 2005. in press.

J. Meller and R. Elber. Linear programming optimization and a double statistical filter for protein threading protocols. *Proteins*, 45:241–261, 2001.

T.E. Mollnes, O.L. Brekke, M. Fung, H. Fure, D. Christiansen, G. Bergseth, V. Videm, K.T. Lappegard, J. Kohl, and J.D. Lambris. Essential role of the C5a receptor in E coli-induced oxidative burst and phagocytosis revealed by a novel lepirudin-based human whole blood model of inflammation. *Blood*, 100:1869–1877, 2002.

B.H.M. Mooers, D. Datta, W.A. Baase, E.S. Zollars, S.L. Mayo, and B.W. Matthews. Repacking the Core of T4 Lysozyme by Automated Design. *J. Mol. Biol.*, 332:741–756, 2003.

D. Morikis, N. Assa-Munt, A. Sahu, and J. D. Lambris. Solution Structure of Compstatin, a Potent Complement Inhibitor. *Protein Sci.*, 7:619–627, 1998.

D. Morikis, M. Roy, A. Sahu, A. Torganis, P.A. Jennings, G.C. Tsokos, and J.D. Lambris. The structural basis of compstatin activity examined by structure-function-based design of peptide analogs and NMR. *J. Biol. Chem.*, 277:14942–14953, 2002.

G. Némethy, K. D. Gibson, K. A. Palmer, C. N. Yoon, G. Paterlini, A. Zagari, S. Rumsey, and H. A. Scheraga. Energy Parameters in Polypeptides. 10. *J. Phys. Chem.*, 96:6472–6484, 1992.

B. Nillson, R. Larsson, J. Hong, G. Elgue, K.N. Ekdahl, A. Sahu, and J.D. Lambris. Compstatin inhibits complement and cellular activation in whole blood in two models of extracorporeal circulation. *Blood*, 92:1661–1667, 1998.

M.J. Nohaile, Z.S. Hendsch, B. Tidor, and R.T. Sauer. Altering dimerization specificity by changes in surface electrostatics. *Proc. Natl. Acad. Sci. USA*, 98:3109–3114, 2001.

C. Pabo. Molecular technology. Designing proteins and peptides. *Nature*, 301:200, 1983.

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

S. Park, X. Yang, and J.G. Saven. Advances in Computational Protein Design. *Current Opinion in Structural Biology*, 14:487–494, 2004.

N.A. Pierce, J.A. Spriet, J. Desmet, and S.L. Mayo. Conformational Splitting: A More Powerful Criterion for Dead-end Elimination. *Journal of Computational Chemistry.*, 21:999–1009, 2000.

N.A. Pierce and E. Winfree. Protein Design is NP-hard. *Protein Engineering.*, 15:779–782, 2002.

J.W. Ponder and F.M. Richards. Tertiary templates for proteins. *J. Mol. Biol.*, 193:775–791, 1987.

K. Raha, A.M. Wollacott, M.J. Italia, and J.R. Desjarlais. Prediction of amino acid sequence from structure. *Protein Sci.*, 9:1106–1119, 2000.

S.A. Ross, C.A. Sarisky, A. Su, and S.L. Mayo. Designed Protein G Core Variants Fold to Native-like Structures: Sequence Selection by ORBIT Tolerates Variation in Backbone Specification. *Protein Science*, 10:450–454, 2001.

A. Sahu, B.K. Kay, and J.D. Lambris. Inhibition of human complement by a C3-binding peptide isolated from a phage displayed random peptide library. *J. Immunol.*, 157:884–891, 1996.

A. Sahu, D. Morikis, and J.D. Lambris. Compstatin, a peptide inhibitor of complement, exhibits species-specific binging to complement component C3. *Mol. Immunology*, 39:557–566, 2003.

A. Sahu, A.M. Soulika, D. Morikis, L. Spruce, W.T. Moore, and J.D. Lambris. Binding kinetics, structure activity relationship and biotransformation of the complement inhibitor compstatin. *J. Immunol.*, 165:2491–2499, 2000.

J.G. Saven. Connecting statistical and optimized potentials in protein folding via a generalized foldability criterion. *J. Chemical Physics*, 118:6133–6136, 2003.

H.D. Sherali and W.P. Adams. *A reformulation linearization technique for solving discrete and continuous nonconvex problems*. Kluwer Academic Publishing, Boston, MA, 1999.

M. Shimaoka, J.M. Shifman, H. Jing, L. Takagi, S.L. Mayo, and T.A. Springer. Computational Design of an Intergrin I Domain Stabilized in the Open High Affinity Conformation. *Nat. Struct. Biol.*, 7:674–678, 2000.

A.M. Soulika, M.M. Khan, T. Hattori, F.W. Bowen, B.A. Richardson, C.E. Hack, A. Sahu, L.H. Edmunds, and J.D. Lambris. Inhibition of heparin/protamine complex-induced complement activation by Comsptatin in baboons. *Clin. Immunology*, 96:212–221, 2000.

A.M. Soulika, D. Morikis, M.R. Sarias, M. Roy, L. Spruce, A. Sahu, and J.D. Lambris. Studies of Structure-Activity Relations of Complement Inhibitor Compstatin. *J. Immunology*, 170:1881–1890, 2003.

A.G. Street and S.L. Mayo. Pairwise calculation of protein solvent-accessible surface areas. *Fold. Des.*, 3:253–258, 1998.

P. Strop and S.L. Mayo. Rubredoxin Variant Folds Without Irons. *J. Am. Chem. Soc.*, 121:2341–2345, 1999.

A. Su and S.L. Mayo. Coupling Backbone Flexibility and Amino Acid Sequence Selection in Protein Design. *Protein Science*, 6:1701–1707, 1997.

D. Tobi and R. Elber. Distance-dependent pair potential for protein folding: results from linear optimization. *Proteins*, 41:40–46, 2000.

D. Tobi, G. Shafran, N. Linial, and R. Elber. On the design and analysis of protein folding potentials. *Proteins*, 40:71–85, 2000.

P. Tuffery, C. Etchebest, S. Hazout, and R. Lavery. A New Approach to the Rapid Determination of Protein Side Chain Conformations. *J. Biomol. Struct. Dyn.*, 8:1267–1289, 1991.

C.A. Voigt, D.B. Gordon, and S.L. Mayo. Trading accuracy for speed: a quantitative comparison of search algorithms in protein sequence design. *J. Mol. Biol.*, 299:789–803, 2000.

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

J. Zhou and J.G. Saven. Statistical Theory of Combinatorial Libraries of Folding Proteins: Energetic Discrimination of a Target Structure. *J. Molecular Biology*, 296:281–294, 2000.

Table 1: Sequence and experimental relative activity of compstatin analogs with improved activity that were identified by rational design, experimental combinatorial design, and the novel in silico de novo protein design approach. Boldface is used to indicate that amino acids were fixed. Brackets indicate the disulfide bridge. Relative complement inhibitory activity is derived from IC_{50} measurements.

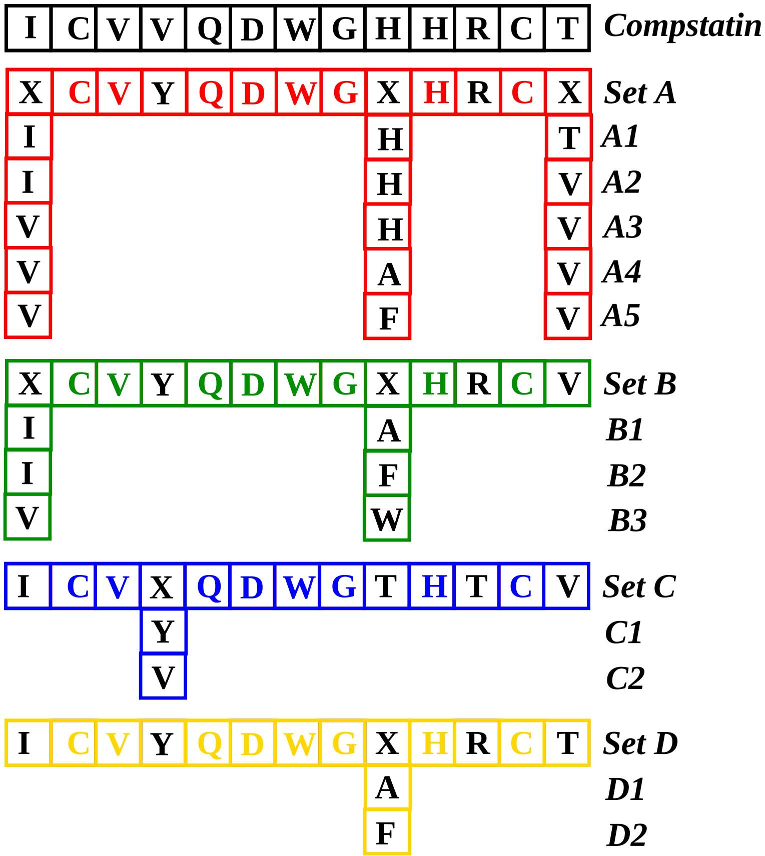
Peptide	Sequence	Relative activity	Reference
Compstatin	$I[\mathbf{CV}V\mathbf{QDWG}HHR\mathbf{C}]T-NH_2$	1	(Sahu *et al.*, 1996)
Ac-Compstatin	$Ac-I[\mathbf{CV}V\mathbf{QDWG}HHR\mathbf{C}]T-NH_2$	3	(Sahu *et al.*, 2000)
Ac-H9A	$Ac-I[\mathbf{CV}V\mathbf{QDWG}AHR\mathbf{C}]T-NH_2$	4	(Morikis *et al.*, 2002)
Ac-I1L/H9W/T13G	$Ac-L[\mathbf{CV}V\mathbf{QDWG}WHR\mathbf{C}]G-NH_2$	4	(Soulika *et al.*, 2003)
Ac-I1V/V4Y/H9F/T13V	$Ac-V[\mathbf{CV}Y\mathbf{QDWG}FHR\mathbf{C}]V-NH_2$	6	(Klepeis *et al.*, 2003)
Ac-I1V/V4Y/H9A/T13V	$Ac-V[\mathbf{CV}Y\mathbf{QDWG}AHR\mathbf{C}]V-NH_2$	9	(Klepeis *et al.*, 2003)
Ac-V4Y/H9F/T13V	$Ac-I[\mathbf{CV}Y\mathbf{QDWG}FHR\mathbf{C}]V-NH_2$	11	(Klepeis *et al.*, 2003)
Ac-V4Y/H9A/T13V	$Ac-I[\mathbf{CV}Y\mathbf{QDWG}AHR\mathbf{C}]V-NH_2$	14	(Klepeis *et al.*, 2003)
Ac-V4Y/H9A	$Ac-I[\mathbf{CV}Y\mathbf{QDWG}AHR\mathbf{C}]T-NH_2$	16	(Klepeis *et al.*, 2003)
Ac-V4W/H9A	$Ac-I[\mathbf{CV}W\mathbf{QDWG}AHR\mathbf{C}]T-NH_2$	45	(Mallik *et al.*, 2005)

Table 2: Preferred residue selection for positions 1, 4, 9, 10, 11 and 13 of compstatin, as compared to the wild type sequence. Only residues with greater than 10 % representation among the lowest lying energy sequences are considered optimal. Provided in decreasing order.

Position	Wild type	Optimal[1]	Optimal[2]
1	I	A,V	V,A
4	V	Y,V	W,Y,V
9	H	T,F,A	F,T
10	H	H	H,K,S
11	R	T,V,A,F,H	H,F,T
13	T	V,A,F	V,A,F

[1] Base case: positions 1 and 4 selected from {A,F,I,L,M,V,Y}; position 13 selected from {A,F,I,L,M,V,Y,T}; positions 9,10 and 11 selected from all residues except C and W.

[2] Base case with position 4 among {A,F,I,L,M,V,Y,W}.



I	C	V	V	Q	D	W	G	H	H	R	C	T	*Compstatin*
X	C	V	Y	Q	D	W	G	X	H	R	C	X	*Set A*
I								H				T	*A1*
I								H				V	*A2*
V								H				V	*A3*
V								A				V	*A4*
V								F				V	*A5*
X	C	V	Y	Q	D	W	G	X	H	R	C	V	*Set B*
I								A					*B1*
I								F					*B2*
V								W					*B3*
I	C	V	X	Q	D	W	G	T	H	T	C	V	*Set C*
			Y										*C1*
			V										*C2*
I	C	V	Y	Q	D	W	G	X	H	R	C	T	*Set D*
								A					*D1*
								F					*D2*

Figure 1: Set of sequences tested for fold specificity.