De Novo Protein Design: An Interplay of Global Optimization, Mixed-Integer Optimization and Experiments

C.A. Floudas[1], J.L. Klepeis[1], J.D. Lambris[2], and D. Morikis[3]

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

[2] Department of Pathology & Laboratory Medicine, University of Pennsylvania, Philadelphia, PA 19104

[3] Department of Chemical and Environmental Engineering, University of California at Riverside, Riverside, CA, 92521

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. In this paper, an account of the recent advances in de novo protein design is presented, followed by background and previous work on the design of Compstatin, a 13-residue cyclic peptide that binds to complement component C3 and inhibits complement activation. A novel integrated framework based on global optimization, mixed-integer optimization, in vitro and in silico characterization via NMR experiments, as well as experiments for the synthesis and functional characterization of peptides, is introduced for the computational design of peptides and proteins. The experimental functional analysis provides validation to the in silico predicted novel peptide sequences which are shown to exhibit 16-fold improved activity over the synthetic therapeutic peptide Compstatin. This overview paper is based on advances reported in Klepeis et al. (2003a), Morikis et al. (2004), and Klepeis et al. (2004).

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.

Computational Methods: 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 astronomical 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). Overall, there is no consensus between model parameterizations, and it is unclear which methods are more valid and suitable for generic protein design.

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. Stochastic approaches are appealing because their heuristic nature can be used to control termination, and both genetic algorithm (Jones, 1994) and Monte Carlo methods (Wernisch et al., 2000; Desjarlais and Handel, 1999) have been applied to the protein design problem. However, these methods involve some element of chance and thus may lack consistency and reliability in locating the global minimum (Voigt et al., 2000). Deterministic methods, such as the dead-end elimination algorithm (Desmet et al., 1992), offer the advantage of convergence to a consistent solution. These methods may not be globally deterministic in that heuristic modifications must be applied to make convergence tractable for complex systems (Gordon and Mayo, 1999; Wernisch et al., 2000). Recent advances in the dead-end elimination algorithms include a hybrid exact rotamer optimization method which improves the computational performance (Gordon et al., 2003) and the introduction of conformational splitting which expedites the roramer elimination process and allows for complete protein design (Pierce et al., 2000). Restrictions on methods based on the dead-end elimination are the requirement for a pairwise representation of the energy function, and more importantly the postulation of a fixed template. 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, 2001, 2003).

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 the almost universal specification of a fixed backbone, 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.

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 biotranformation 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). A recent mini-review provides a detailed account of the advances using rational design methods experimental combinatorial design approaches, molecular dynamics, and novel optimization methods (Morikis et al., 2004). In the following section, we outline these advances.

Rational design of compstatin analogs: The three-dimensional structure of compstatin in solution revealed the presence of a major conformer consisting of a Type I β-turn located at a position opposite to the disulfide bridge (Morikis et al., 1998). The molecular surface of compstatin consists of a polar part that includes the β-turn and a hydrophobic part that includes the disulfide bridge.

The rational design of analogs with higher inhibitory activity has been discussed and compared to similar efforts for other low-molecular mass complement inhibitors in a recent mini-review (Morikis and Lambris, 2002). The rational or SAR design was based on the available three-dimensional structure of compstatin, structural NMR studies of the designed new analogs, kinetic binding studies to C3 and its fragments, and complement inhibitory activity measurements. The three-dimensional structure revealed the overall fold of compstatin and intra-molecular interactions involving hydrogen bonding, hydrophobicity, electrostatics, van der Waals forces, disulfide bridge, and polar interactions with solvent molecules. These data provided insights into the structural stability of compstatin and, in combination with additional NMR studies, into the structural stabilities of the designed analogs. Radical site-specific replacements were used to determine the effect of gross aminoacid differences in structure, binding, and activity, and conservative replacements were used for fine-tuning of the design, together with additions/deletions, alanine scan, incorporation of non-natural aminoacids with directed properties, methylation, and alternative cyclization (Sahu et al., 1996), (Morikis et al., 1998), (Sahu et al., 2000), (Morikis et al., 2002), (Soulika et al., 2003).

The analog with highest inhibitory activity identified using this method, named Ac-H9A, had 4-fold higher activity than the parent peptide compstatin (see Table 1). These efforts include a prior benchmark of acetylation of the N-terminus that resulted to a 3-fold increase of inhibitory activity.

Experimental combinatorial design of compstatin analogs: The technique of phage-displayed

random peptide libraries to randomly identify peptides that are capable of binding to specific targets and altering their functionality is widely used. Compstatin was identified using a phage-displayed peptide library and screened for binding against C3b (Sahu et al., 1996). This technique was used again for peptide binding against C3 (Soulika et al., 2003), but this time incorporating findings from our rational design. Specifically, 7 aminoacids were kept fixed while 6 were allowed to randomly vary.

Four binding clones to native C3 were identified using this method. Complement inhibitory activity measurements of synthetic acetylated peptides with the sequences of the binding clones identified one analog with 4-fold higher activity than compstatin (Table 1). This analog was named Ac-I1L/H9W/T13G and its sequence is given in Table 1. NMR experiments of this analog demonstrated similar structural characteristics as compstatin, Ac-compstatin, and the equally active rationally designed analog Ac-H9A (Table 1). The hydrophobic cluster and the Type I β-turn were preserved in Ac-I1L/H9W/T13G and a novel feature was observed by the introduction of a second Trp at position 9.

The experimental combinatorial design identified an equally active analog as the rational design, but in combination two important features for activity were revealed: (i) position 9 was amenable to further optimization, and (ii) side chain ring stacking involving one residue inside the β-turn and one outside could be important to optimize activity. The case of the latter was re-enforced by the computational combinatorial design, which will be described below.

Molecular dynamics studies of compstatin: Small peptides in solution form ensembles of interconverting conformers. Compstatin showed better defined structure when the disulfide bridge between Cys2 and Cys12 was intact and less defined structure when the disulfide bridge was broken. The flexibility of compstatin in solution was shown by analysis of NMR parameters such as spin-spin coupling constants, chemical shifts, temperature dependence of chemical shifts, and NOEs (Morikis et al., 1998). The structure of a major conformer of compstatin was determined using NMR data and computational modeling and global optimization (Morikis et al., 1998), (Klepeis et al., 1999).

Molecular dynamics (MD) simulations of the entire NMR ensemble of 21 structures, the average minimized structure, and the global optimization structure revealed the presence of five families of interconverting conformers at 1 ns of simulation time (Mallik et al., 2003). The major population of these conformers was a coil conformation with a Type I β-turn with probability of 44%. This is in agreement with the estimated population of a major conformer of compstatin from the original NMR data using spin-spin coupling constant analysis that was 42-63% (Morikis et al., 1998). The remaining MD conformers (and their populations) were β-hairpin with Type II β-turn (22%), β-hairpin with Type I β-turn (17%), β-hairpin with Type VIII β-turn (9%), and partial a-helix-partial coil (9%). It should be noted that 91% of the MD conformers contained some type of a β-turn and 61% contained a Type I β-turn (Mallik et al., 2003). This demonstrates the significance of the presence of a turn for structural stability of compstatin. These data introduce the concept of a dynamic peptide in the drug design process as opposed to the widely-used, yet overly simplified, static view.

De Novo Protein Design Framework

In Klepeis et al. (2003a), 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 $\AA$ (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). Recent work has resulted in improvements through the use of physical constraints and extension of the parameterization to include β-carbon interactions to better represent side-chain placement (Loose et al., 2003).

The linearity of the resulting formulation based on this distance-dependent interaction potential 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 :

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

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):

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

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 in less than 5 CPU minutes on an HP J-2240.

Fold Stability and 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 nonconsecutive 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:

$$\begin{aligned} \min_{\phi} \quad & E_{ff} \\ \text{subject to} \quad & E_j^{dis}(\phi) \ \leq \ E_j^{ref} \ \ j = 1, \ldots, N \\ & \phi_i^L \ \leq \ \phi_i \ \leq \ \phi_i^U \ \ i = 1, \ldots, N_\phi \end{aligned}$$

Here, $i = 1, \ldots, N_\phi$ corresponds to the set of dihe-

dral 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., 2003b,c). 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*).

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

Here $\exp[-\beta E_i]$ is the statistical weight for conformer i. For compstatin, the template constrained optimizations required approximately six CPU hours on a single P-III 600 MHz processor running Linux. The free folding optimizations were run on a cluster of 64 P-III 600 MhZ processors running Linux, and the parallelized branch-and-bound algorithm utilized about 4-5 wallclock hours per sequence.

Peptide Synthesis and Complement Inhibition Assays

Peptide synthesis and purification was performed as described previously (Sahu et al., 1996, 2000). Inhibitory activity of compstatin and its analogs on the complement system was studied by measuring their effect on the classical pathway. Complement activation inhibition was assessed by measuring the inhibition of C3 fixation to OVA-anti-OVA complexes in normal human plasma. Briefly, microtiter plates were coated with ovalbumin, followed with anti-ovalbumin antibodies and normal human plasma (generally diluted 1/160) in the presence or absence of peptides diluted in gelatin Veronal buffer2+ (VBS, 0.1% gelatin, 0.5 mM MgCl2, 2 mM CaCl2). Complement activation was assessed using a goat anti-human C3 HRP conjugated antibody to detect deposition of activated C3b/iC3b. Color was developed by adding peroxidase substrate and optical density measured at 405 nm. The concentration of the peptide causing 50% inhibition of C3b/iC3b deposition was taken as the IC50 and used to compare the activities of various peptides. All peptides were analyzed at least three times.

Computational and Experimental Findings

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, positions 1, 4 and 13 were allowed to select 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. In positions 9, 10 and 11 all residues were allowed, excluding cystine and tryptophan. Table 2 summarizes the preferred selection at each position according to the composition of the lowest lying energy sequences. It should be noted that if tryptophan (W) is allowed to be in the aforementioned hydrophobic set, 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 stability and 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.

Mutations in Set A: 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 compo-

sition 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).

Mutations in Set B: 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).

Mutations in Set C: 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).

Mutations in Set D: 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. 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.

Acknowledgments

CAF gratefully acknowledges financial support from the National Science Foundation and the National Institutes of Health (R01 GM52032). JDL and DM gratefully acknowledge financial support from the National Institutes of Health (AI 30040 and GM 62134).

References

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

Morikis, D.; Soulika, A.; Mallik, B.; Klepeis, J.; Floudas, C.; Lambris, J. Improvement of the anti-C3 activity of compstatin using rational and combinatorial approaches. *Biochem. Soc. Trans.* **2004**, *32*.

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**, .

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

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

C. Hardin, T. P.; Luthey-Schulten, Z. Ab initio protein structure prediction. *Curr. Opin. Struc. Biol.* **2002**, *12*, 176.

Ponder, J.; Richards, F. Tertiary templates for proteins. *J. Mol. Biol.* **1987**, *193*, 775.

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

Desjarlais, J.; Handel, T. De novo design of the hydrophobic cores of proteins. *Protein Sci.* **1995**, *4*, 2006.

Dahiyat, B.; Mayo, S. Protein design automation. *Protein Sci.* **1996**, *5*, 895.

Dahiyat, B.; Gordon, D.; Mayo, S. Automated design of the surface positions of protein helices. *Protein Sci.* **1997**, *6*, 1333.

Raha, K.; Wollacott, A.; Italia, M.; Desjarlais, J. Prediction of amino acid sequence from structure. *Protein Sci.* **2000**, *9*, 1106.

Street, A.; Mayo, S. Pairwise calculation of protein solvent-accessible surface areas. *Fold. Des.* **1998**, *3*, 253.

Nohaile, M.; Hendsch, Z.; Tidor, B.; Sauer, R. Altering dimerization specificity by changes in surface electrostatics. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 3109.

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

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

Jones, D. De novo protein design using pairwise potentials and a genetic algorithm. *Protein Sci.* **1994**, *3*, 567.

Desjarlais, J.; Handel, T. Side chain and backbone flexibility in protein core design. *J. Mol. Biol.* **1999**, *290*, 305.

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

Desmet, J.; Maeyer, M. D.; Hazes, B.; Lasters, I. The dead-end elimination theorem and its use in side-chain positioning. *Nature* **1992**, *356*, 539.

Gordon, D.; Mayo, S. Branch-and-terminate. *Structure Fold. Des.* **1999**, *7*, 1089.

Gordon, B.; Hom, G.; Mayo, S.; Pierce, N. Exact Rotamer Optimization for Protein Design. *J. Computational Chemistry* **2003**, *24*, 232.

Pierce, N.; Spriet, J.; Desmet, J.; Mayo, S. Conformational Splitting: A More Powerful Criterion for Dead-End Elimination. *J. Computational Chemistry* **2000**, *21*, 999.

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

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

Kono, H.; Saven, J. Statistical Theory of Protein Combinatorial Libraries: Packing Interactions, Backbone Flexibility, and the Sequence Variability of a Main-chain Structure. *J. Molecular Biology* **2001**, *306*, 607.

Saven, J. Designing Protein Energy Landscapes. *Chem. Rev.* **2001**, *101*, 3113.

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

Dahiyat, B.; Mayo, S. De novo protein design: fully automated sequence selection. *Science* **1997**, *278*, 82.

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

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

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

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

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

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

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

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

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

Mollnes, T.; Brekke, O.; Fung, M.; Fure, H.; Christiansen, D.; Bergseth, G.; Videm, V.; Lappegard, K.; Kohl, J.; Lambris, J. 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* **2002**, *100*, 1869.

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

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

Morikis, D.; Lambris, J. Structural aspects and design of low molecular mass complement inhibitors. *Biochem. Soc. Trans.* **2002**, *30*, 1026.

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

Mallik, B.; Morikis, D.; Lambris, J. Conformational inter-conversion of compstatin probed with molecular dynamics simulations. *Proteins: Structure, Function and Genetics* **2003**, *53*, 130.

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

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

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

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

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

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

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

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

Klepeis, J. L.; Schafroth, H. D.; Westerberg, K. M.; Floudas, C. A. 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*; Friesner, R. A., Ed.; vol. 120 Wiley, 2002, p 254–457.

Klepeis, J.; Floudas, C. Ab initio tertiary structure prediction of proteins. *J. Global. Optim.* **2003**, *25*, 113.

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

Adjiman, C.; Androulakis, I.; Floudas, C. A. A Global Optimization Method, αBB, for General Twice-Differential Constrained NPLs - I. Theoretical Advances. *Computers Chem. Engng.* **1998**a, *22*, 1137.

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., 2003a)
Ac-I1V/V4Y/H9A/T13V	$Ac-V[\mathbf{CV}Y\mathbf{QDWG}AHR\mathbf{C}]V-NH_2$	9	(Klepeis et al., 2003a)
Ac-V4Y/H9F/T13V	$Ac-I[\mathbf{CV}Y\mathbf{QDWG}FHR\mathbf{C}]V-NH_2$	11	(Klepeis et al., 2003a)
Ac-V4Y/H9A/T13V	$Ac-I[\mathbf{CV}Y\mathbf{QDWG}AHR\mathbf{C}]V-NH_2$	14	(Klepeis et al., 2003a)
Ac-V4Y/H9A	$Ac-I[\mathbf{CV}Y\mathbf{QDWG}AHR\mathbf{C}]T-NH_2$	16	(Klepeis et al., 2003a)

Adjiman, C.; Androulakis, I.; Floudas, C. A. A Global Optimization Method, αBB, for General Twice-Differentiable Constrained NLPs - II. Implementation and Computational Results. *Computers Chem. Engng.* **1998**b, *22*, 1159.

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

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

Floudas, C. A. *Deterministic Global Optimization : Theory, Methods and Applications*; Nonconvex Optimization and its Applications Kluwer Academic Publishers, 2000.

Klepeis, J.; Pieja, M.; Floudas, C. A New Class of Hybrid Global Optimization Algorithms for Peptide Structure Prediction: Integrated Hybrids. *Comp. Phys. Comm.* **2003**b, *151*, 121.

Klepeis, J.; Pieja, M.; Floudas, C. Hybrid Global Optimization Algorithms for Protein Structure Prediction : Alternating Hybrids. *Biophysical J.* **2003**c, *84*, 869.

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	I	A,V
4	V	Y,V
9	H	T,F,A
10	H	H
11	R	T,V,A,F,H
13	T	V,A,F

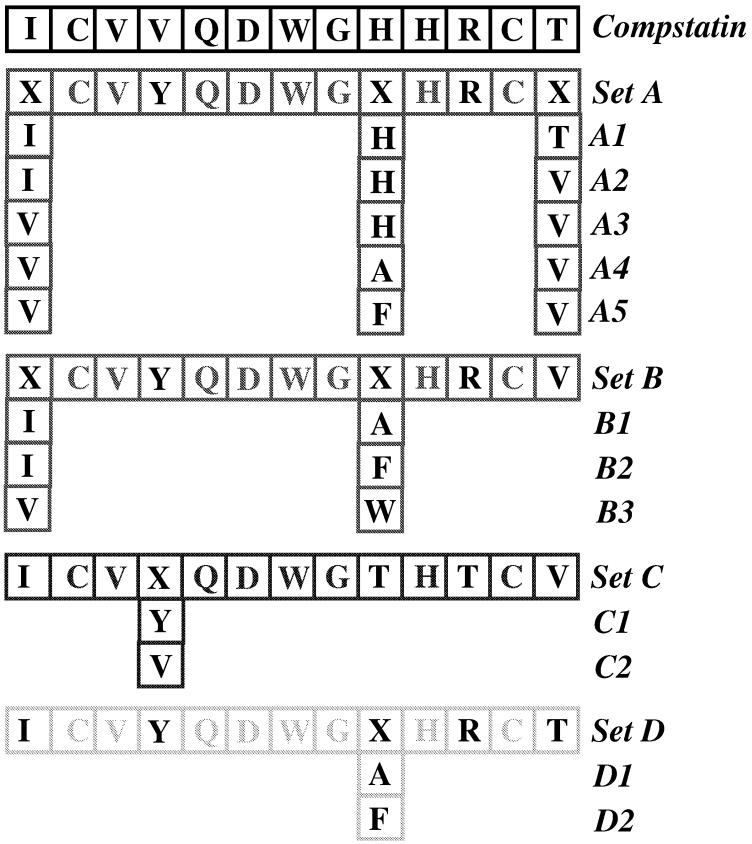

Figure 1: Set of sequences tested for fold stability.