Optimization Methods and Software
Vol. 00, No. 00, Month 200x, 1–27

Novel Formulations for the Sequence Selection Problem in De Novo Protein Design with Flexible Templates

H. K. Fung, M. S. Taylor, and C. A. Floudas†
Department of Chemical Engineering
Princeton University
Princeton, NJ 08544-5263
(accepted February 2006)

This paper presents two novel formulations for solving the sequence selection problem in de novo protein design with highly flexible templates, each of which exhibits a crystal or NMR structure. The first formulation applies weighted average energy parameters to incorporate information about every structure, with the weights, which are parameters dependent on a pair of C^{α} positions and a particular distance bin, given by the probability that the distance between the two positions is found to belong to that distance bin in any of the structures. The second formulation allows the distance between the two positions considered to fall into any distance bin that all the structures span over, but imposes novel linear constraints to ensure a physically consistent structure. Both formulations were tested on redesigning Compstatin, the template of which has 21 NMR structures from the Protein Data Bank.

Keywords
Peptide and protein design and discovery; Drug design; In silico sequence selection; Structure prediction; De novo protein design; Protein backbone flexibility; Optimization

1 Introduction

1.1 *De Novo Protein Design*

De novo protein design, or engineering proteins from "scratch", requires the determination of the amino acid sequences that will fold into a certain three-dimensional template structure with higher stability or functionality than the native sequence. The problem lies in the core of the fruitful field of protein structural bioinformatics, which contributes significantly to our understanding

† Author to whom all correspondence should be addressed;

Tel: (609) 258-4595; Fax: (609) 258-0211; E-mail: floudas@titan.princeton.edu.

of protein structure prediction [1,2], protein folding kinetics [3], protein-ligand interactions [4,5], and protein-protein docking [6], and thus enhances our process of drug discovery. By applying de novo protein design techniques, successes have been achieved in modulating protein-protein interactions [7], promoting stability and specificity of the target protein [3,8–10], conferring novel binding sites or properties onto the template [11, 12], as well as locking proteins into certain useful conformations [13, 14].

The major challenges in de novo protein design stem from: (1) the NP-hard nature of the problem in terms of computational efficiency [15,16], and (2) the incorporation of protein backbone flexibility into the design model. The former challenge requires any formulation or algorithm used for *in silico* protein design be computationally efficient; otherwise the maximum problem complexity level that the design model can handle will not be high enough for practical purpose. While guaranteeing the convergence to the global optimal solution, the novel formulations presented in this article for performing sequence selection in de novo protein design can be shown by case studies, which are also outlined in this paper, to be fast enough to run for real applications. However, it is mainly addressing the second issue of incorporating high degree of protein backbone flexibility that the formulations were designed for.

1.2 *Approaches for Incorporating Protein Backbone Flexibility*

The starting point for any de novo protein design method is the definition of the three-dimensional template structure, and the ultimate goal of the design is to obtain sequences that adopt such template structure as their native fold. Very often the template is the native fold of a protein which already exists in Nature. Being a macromolecule, protein is known to exhibit a multitude of stable conformations as its backbone moves. Backbone flexibility proves to complicate the de novo protein design problem. [17] demonstrated that backbone flexibility can allow residues that would not have been permissable had a rigid template been considered. [18] provided convincing evidence for the superiority of backbone flexibility in their successful design of metal binding sites in the Gβ1 protein. They noted that elements of their design would have been overlooked using a single, averaged template because the required conformations would have been deemed infeasible. Backbone flexibility was also found to be of fundamental importance to obtaining stable folds [19].

Different approaches were adopted to address the issue of backbone flexibility in de novo protein design. In one approach atomic radii in calculating the van der Waals potential were scaled down, typically by five to ten per cent, to allow for small overlap between atoms during backbone movements [20,21]. However, this method has the intrinsic disadvantages of overestimation of attractive forces and possible hydrophobic core overpacking. In another approach

either a fixed set of rotamers was considered or some super-secondary-structure parameters were changed to adjust the relative orientation and distances between secondary structures [22]. [23] also constructed ensemble of random structures from the template and solved each structure in the ensemble for the low energy sequence using the fixed backbone assumption. These two similar methods only take into account either a chosen subset or a random subset of all possible conformations of the protein template. Lately, backbone flexibility was treated by the Baker's group by iterating between sequence space and structure space until the algorithm converged to a final solution [24, 25]. Again their approach only addressed backbone flexibility indirectly by movements along the structure space during iteration.

2 General Definition of Backbone Template Flexibility

[1] defined the meaning of true backbone template flexibility that should be incorporated into de novo protein design. The template backbone structure can be (1) a single rigid backbone (e.g., the average NMR structure for a protein), or (2) a set of rigid backbone structures (e.g., all NMR structures for a protein or a discrete number of randomly selected rigid structures, based on some algorithmic procedure or a discrete set of rigid structures, based on parametrization of the backbone), or (3) a flexible backbone structure defined by lower and upper bounds on the distances between the alpha carbons and the backbone dihedral angles. Apparently, true backbone template flexibility is reflected in (3) since it allows for all possible combinations of distances and backbone dihedral angles within their specified ranges, while as previously mentioned, (2) considers only a small subset of flexible structures, and (1) is restricted to a single structure only.

3 In Silico Sequence Selection: Basic Model for Single Template Structure

Recently [9,10] proposed a two-stage framework for performing de novo protein design. In the first stage of *in silico* sequence selection promising low energy amino acid sequences were chosen using an integer linear programming (ILP) model. In the second stage, fold stability for each sequence from the first stage was quantified by performing two separate runs of protein structure prediction: one for configurations constrained around the template and the other for free folding. The probability for a sequence to fold into the template would be given by the Boltzmann type distribution for those low energy conformers from the free folding calculation that fell into the overlap between the template-

constrained conformers and the free folding conformers on an energy-versus-rmsd plot. Structure prediction was done with the aid of the αBB deterministic global optimization solver [26–32] with an objective function of a full-atomistic force field over the set of independent dihedral angles which can be used to describe any possible configuration of the system.

This article only focuses on the optimization model for the first stage of sequence selection and outlines improved versions of it. However, before proceeding to the new formulations, it should be worthwhile to present what the old model was and how it was derived.

The original form of the sequence selection model used by [9, 10] takes the form:

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

Note that this formulation corresponds to a quadratic assignment like model. It differs, however, in the set of constraints. Set $i = 1, \ldots, n$ 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 equivalent sets $k \equiv i$ and $l \equiv j$ are defined, and $k > i$ is required to represent all unique pairwise interactions. Binary variables y_i^j and y_k^l are introduced to indicate the possible mutations at a given position. Specifically, variable y_i^j will be one if position i is occupied by amino acid j, and zero otherwise. Similarly, variable y_k^l will assume the value of one if position k is taken by amino acid l, and the value of zero otherwise. The composition constraints in the formulation require that there is exactly one type of amino acid at each position. Energy parameter E_{ik}^{jl} indicates the pairwise interaction between the amino acid j at position i and the amino acid l at position k. It only contributes to the objective function of total energy of the system if both y_i^j and y_k^l are one, i.e., position i and k are really taken by amino acid j and l respectively.

Bilinear terms in the objective function of the original formulation were linearized using an equivalent representation [33]:

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

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

$$y_i^j + y_k^l - 1 \leq w_{ik}^{jl} \leq y_i^j \; \forall \; i, j, k, l \tag{2}$$

$$0 \leq w_{ik}^{jl} \leq y_k^l \; \forall \; i, j, k, l$$

$$y_i^j \; , \; y_k^l = 0 - 1 \; \forall \; i, j, k, l$$

As indicated in the formula above, bilinear terms were transformed into a new set of linear variables, w_{ik}^{jl}, with the addition of the four sets of constraints 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 [33, 34] making convergence to the global minimum energy sequence consistent and reliable.

Performance of the branch and bound algorithm can be significantly enhanced through the use of the reformulation linearization techniques (RLTs). 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 and tighter bounding linear programming (LP) relaxations for the original problem [35]. In this case RLTs are introduced by multiplying the composition constraints by the binary variables y_k^l to produce:

$$y_k^l \sum_{j=1}^{m_i} y_i^j = y_k^l \quad \forall i, k, l \quad \text{or} \quad \sum_{j=1}^{m_i} w_{ik}^{jl} = y_k^l \forall i, k, l \tag{3}$$

In summary, the whole basic model for performing sequence selection in [9, 10]'s de novo protein design framework is:

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

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

$$
\begin{aligned}
y_i^j \; + \; y_k^l \; - \; 1 \; \leq \; w_{ik}^{jl} \; \leq \; y_i^j \; \forall \; i,j,k,l \qquad (4)\\
0 \; \leq \; w_{ik}^{jl} \; \leq \; y_k^l \; \forall \;\; i,j,k,l \\
\textstyle\sum_{j=1}^{m_i} w_{ik}^{jl} \; = \; y_k^l \;\; \forall \;\; i,k,l \\
y_i^j \; , \; y_k^l \; = \; 0-1 \;\; \forall \;\; i,j,k,l
\end{aligned}
$$

As explained in the subsection that immediately follows, this model allows true backbone template flexibility by the use of energy parameters E_{ik}^{jl} which are discretized according to the distance bin they belong rather than continuously dependent on C^α-C^α distance. The details of how this works are outlined in the subsection below.

Incorporation of true backbone template flexibility

It should be highlighted that [9, 10] incorporated protein backbone flexibility explicitly into both stage one and stage two of their framework. Rather than being a continuous function, the dependence on C^α-C^α distance of the energy parameter $E_{ik}^{jl}(x_i, x_k)$ in the objective function of (4) is discretized into bins. In both the force field of [36] and that of [37], the distance bins are classified in the way as shown in Table 1. Bin 1 is for any distance that is between $l_{beg}(1) = 3.0\mathring{A}$ and $l_{end}(1) = 4.0\mathring{A}$, bin 2 for any distance between $l_{beg}(2) = 4.0\mathring{A}$ and $l_{end}(2) = 5.0\mathring{A}$, and so forth. Given a certain pair of amino acids, any distance between the alpha carbon of the two amino acids falling within the range bounded by the upper bound of $l_{end}(n)$ and lower bound of $l_{beg}(n)$ will belong to the distance bin n, thus giving the same energy value. In this way the energy function can tolerate backbone motion to a certain degree. With the bin sizes varying between 0.5 and $1\mathring{A}$, this discretization of the force field allows backbone movements of similar magnitude. In the second stage of fold stability quantification, backbone flexibility is explicitly included by treating C^α-C^α distances and dihedral angles as bounded continuous variables in the template-constrained structure prediction run. In addition, [9, 10] chose to set the lower and upper bounds to be $\pm 10\%$ of those in the template for the C^α-C^α distances and $\pm 10^o$ around the template for the phi and psi angles.

Possible improvement for the model

Allowing backbone flexibility by the use of distance bins, formulation (4) for sequence selection can also be applied to highly flexible templates where all the distances between any position pair i and k fall into the same distance bin. If the distances between a position pair i and k span over different different

Distance Bin d	$l_{beg}(d)[\AA]$	$l_{mid}(d)[\AA]$	$l_{end}(d)[\AA]$
Bin 1	3.0	3.5	4.0
Bin 2	4.0	4.5	5.0
Bin 3	5.0	5.25	5.5
Bin 4	5.5	5.75	6.0
Bin 5	6.0	6.25	6.5
Bin 6	6.5	6.75	7.0
Bin 7	7.0	7.5	8.0
Bin 8	8.0	8.5	9.0

Table 1. Distance bin classification in the high resolution force field developed by [36] for the sequence selection of de novo protein design.

bins, significant modification has to be done before the formulation becomes applicable. To the best of our knowledge, there is currently no de novo protein design model from open literature that explicitly deals with such kind of highly flexible templates. Therefore novel formulations to fill this void are needed and the effort of development would be worthwhile.

4 Algorithmic Improvement of the Basic Sequence Selection Formulation for Designing Proteins into a Template with a Single Structure

This section presents the algorithmic improvement efforts to hasten the convergence of formulation (4) to the global optimal solution. The rationale behind the efforts is to identify all superfluous equations in the model, and by doing so computational performance can hopefully be enhanced with the solution space unchanged. First, it is noticed that the RLT constraints $\sum_{j=1}^{m_i} w_{ik}^{jl} = y_k^l \;\; \forall \;\; i,k,l$ could have been written as two separate equations, namely $\sum_{j=1}^{m_i} w_{ik}^{jl} = y_k^l \;\; \forall \;\; i,k > i,l$ and $\sum_{l=1}^{m_k} w_{ik}^{jl} = y_i^j \;\; \forall \;\; i,k > i,j$. By doing so it becomes more apparent that the inequalities $w_{ik}^{jl} \leq y_i^j \;\; \forall \;\; i,j,k,l$ and $w_{ik}^{jl} \leq y_k^l \;\; \forall \;\; i,j,k,l$ are already implied by the RLTs, due to the fact that w_{ik}^{jl} is positive. The inequalities $y_i^j + y_k^l - 1 \leq w_{ik}^{jl} \;\; \forall \;\; i,j,k,l$ is active when both y_i^j and y_k^l are one, which will turn variable w_{ik}^{jl} into one. Had w_{ik}^{jl} been declared as binary variables instead of a continuous variables, these inequalities would have been made superfluous also. This can be shown as below:

PROPOSITION 4.1 *With y_i^j, y_k^l, and w_{ik}^{jl} declared as binary variables, the following set of equations:*

$$\sum_{j=1}^{m_i} y_i^j \;\; = \;\; 1 \; \forall \; i$$

$$\begin{aligned} \textstyle\sum_{j=1}^{m_i} w_{ik}^{jl} &= y_k^l \quad \forall \quad i,k > i,l \\ \textstyle\sum_{l=1}^{m_k} w_{ik}^{jl} &= y_i^j \quad \forall \quad i,k > i,j \end{aligned}$$

already implies that if y_i^j and y_k^l are one, then w_{ik}^{jl} has to be one.

Proof

- first notice that $\sum_{l=1}^{m_k}\sum_{j=1}^{m_i} w_{ik}^{jl} = \sum_{l=1}^{m_k} y_k^l = \sum_{j=1}^{m_i}\sum_{l=1}^{m_k} w_{ik}^{jl} = \sum_{j=1}^{m_i} y_i^j = 1 \ \forall \ i,k > i$
- expansion on $\sum_{j=1}^{m_i}\sum_{l=1}^{m_k} w_{ik}^{jl}$ gives $\sum_{j=1}^{m_i}\sum_{l=1}^{m_k} w_{ik}^{jl} = \sum\limits_{j:y_i^j=0}\sum\limits_{l:y_k^l=0} w_{ik}^{jl} + \sum\limits_{j:y_i^j=0} w_{ik}^{jl}|_{l:y_k^l=1} + \sum\limits_{l:y_k^l=0} w_{ik}^{jl}|_{j:y_i^j=1} + w_{ik}^{jl}|_{j:y_i^j=1,l:y_k^l=1} = 1 \ \forall \ i,k > i$
- $\sum_{j:y_i^j=0}\sum_{l=1}^{m_k} w_{ik}^{jl} = \sum_{j:y_i^j=0} y_i^j = 0$, and $\sum_{l:y_k^l=0}\sum_{j=1}^{m_i} w_{ik}^{jl} = \sum_{l:y_k^l=0} y_k^l = 0$. Obviously $\sum\limits_{j:y_i^j=0}\sum\limits_{l:y_k^l=0} w_{ik}^{jl}$ is also zero.
- $\sum_{j=1}^{m_i}\sum_{l=1}^{m_k} w_{ik}^{jl} = 2\sum_{j:y_i^j=0}\sum_{l:y_k^l=0} w_{ik}^{jl} + \sum_{j:y_i^j=0} w_{ik}^{jl}|_{l:y_k^l=1} + \sum_{l:y_k^l=0} w_{ik}^{jl}|_{j:y_i^j=1} + w_{ik}^{jl}|_{j:y_i^j=1,l:y_k^l=1} = 1 \ \forall \ i,k > i$
- it follows that $w_{ik}^{jl}|_{j:y_i^j=1,l:y_k^l=1} = 1 \ \forall \ i,k > i,j,l$

□

By taking out all the superfluous constraints aforementioned, and by declaring w_{ik}^{jl} as binary variables, a novel formulation, which is totally equivalent to formulation (4), is obtained:

$$\begin{aligned} \min_{y_i^j, y_k^l} \ & \textstyle\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{k=i+1}^{n}\sum_{l=1}^{m_k} E_{ik}^{jl}(x_i,x_k) w_{ik}^{jl} \\ \text{subject to} \quad & \textstyle\sum_{j=1}^{m_i} y_i^j = 1 \ \forall \ i \\ & \textstyle\sum_{j=1}^{m_i} w_{ik}^{jl} = y_k^l \quad \forall \quad i,k > i,l \\ & \textstyle\sum_{l=1}^{m_k} w_{ik}^{jl} = y_i^j \quad \forall \quad i,k > i,j \\ & y_i^j \ , \ y_k^l \ w_{ik}^{jl} = 0-1 \quad \forall \quad i,j,k > i,l \end{aligned} \tag{5}$$

The computational performance of this novel formulation was compared to that of formulation (4) by recording their CPU times to solve two benchmark sequence selection problems. Both problems tried to redesign the template of chain A of human beta defensin 2 (PDB code: 1FD3). The first problem used

only the basic model to search for the global optimal solution from a mutation set which was generated by fixing the native CYS at position 8, 15, 20, 30, 37, and 38, and allowing all other positions to choose from any of the 20 amino acids. This corresponds to a sequence search space of $20^{35} = 3.4 \times 10^{45}$. The second problem used 49 linear biological constraints in addition to the basic model for the sequence search. These biological constraints aim at improving the quality of the solutions by ensuring that all sequences observe certain properties which are important to molecular function. They include charge constraints:

$$\begin{aligned} 0 \le \textstyle\sum_i y_i^{Arg} + \sum_i y_i^{Lys} - \sum_i y_i^{Asp} - \sum_i y_i^{Glu} \le 3 \ \forall\, 5 \le i \le 10 \\ 5 \le \textstyle\sum_i y_i^{Arg} + \sum_i y_i^{Lys} \le 10 \ \forall i \\ 0 \le \textstyle\sum_i y_i^{Asp} + \sum_i y_i^{Glu} \le 2 \ \forall i \\ 4 \le \textstyle\sum_i y_i^{Arg} + \sum_i y_i^{Lys} - \sum_i y_i^{Asp} - \sum_i y_i^{Glu} \le 9 \ \forall\, i \end{aligned} \tag{6}$$

and constraints which place bounds on the occurrence of each amino acid in each sequence:

$$\begin{array}{ll} 0 \le \sum_i y_i^{Ala} \le 3 \ \forall i & 0 \le \sum_i y_i^{Gln} \le 3 \ \forall i \\ 0 \le \sum_i y_i^{Leu} \le 4 \ \forall i & 0 \le \sum_i y_i^{Ser} \le 6 \ \forall i \\ 1 \le \sum_i y_i^{Arg} \le 9 \ \forall i & 0 \le \sum_i y_i^{Glu} \le 3 \ \forall i \\ 0 \le \sum_i y_i^{Lys} \le 7 \ \forall i & 0 \le \sum_i y_i^{Thr} \le 4 \ \forall i \\ 0 \le \sum_i y_i^{Asn} \le 6 \ \forall i & \sum_i y_i^{Gly} \le 6 \ \forall i \\ 0 \le \sum_i y_i^{Met} \le 3 \ \forall i & 0 \le \sum_i y_i^{Trp} \le 2 \ \forall i \\ 0 \le \sum_i y_i^{Asp} \le 2 \ \forall i & 0 \le \sum_i y_i^{His} \le 4 \ \forall i \\ 0 \le \sum_i y_i^{Phe} \le 4 \ \forall i & 0 \le \sum_i y_i^{Tyr} \le 4 \ \forall i \\ \sum_i y_i^{Cys} = 6 \ \forall i & 0 \le \sum_i y_i^{Ile} \le 6 \ \forall i \\ \sum_i y_i^{Pro} \le 5 \ \forall i & 0 \le \sum_i y_i^{Val} \le 6 \ \forall i \end{array} \tag{7}$$

and constraints that restrict β strands to have at least two hydrophobic residues to ensure enough hydrophobic interaction for stability purpose:

Table 2. Comparison of computational performance of two different formulations for designing proteins into a template with a single structure.

First Problem			
	Number of	CPU times† [sec]	
Problem complexity	biological constraints	Formulation (4)	Formulation (5)
3.4×10^{45}	none	$53,263$	649
Second Problem			
	Number of	CPU times [sec]	
Problem complexity	biological constraints	Formulation (4)	Formulation (5)
6.4×10^{37}	49	$4,578$	14

†Generated using CPLEX 9.0 on one single Pentium IV 3.2 GHz processor.

$$\begin{aligned} &\textstyle\sum_i y_i^{Cys} + \sum_i y_i^{Ile} + \sum_i y_i^{Leu} + \sum_i y_i^{Met} + \sum_i y_i^{Phe} + \\ &\textstyle\sum_i y_i^{Trp} + \sum_i y_i^{Tyr} + \sum_i y_i^{Val} + \sum_i y_i^{Ala} \geq 2 \;\; \forall\, 14 \leq i \leq 16 \\ &\textstyle\sum_i y_i^{Cys} + \sum_i y_i^{Ile} + \sum_i y_i^{Leu} + \sum_i y_i^{Met} + \sum_i y_i^{Phe} + \\ &\textstyle\sum_i y_i^{Trp} + \sum_i y_i^{Tyr} + \sum_i y_i^{Val} + \sum_i y_i^{Ala} \geq 2 \;\; \forall\, 25 \leq i \leq 28 \end{aligned} \tag{8}$$

Lastly, the number of mutations on each solution sequence is permitted to be ten at maximum by the following equation:

$$\textstyle\sum_{i=1}^{n} \sum_{j=1}^{m_i} y_i^j \leq 10 \;\; \forall\, i \in domain\,, j \neq native_i \tag{9}$$

The conserved properties can be generated by running a sequence alignment tool like PSI-BLAST, which was created by the National Center for Biotechnology Information (NCBI) of the National Institute of Health, on the human beta defensin homologs. The details for the mutation set of the second problem were also outlined in the case studies section. The mutation set was derived from Solvent Accessible Surface Area (SASA) patterning, and it corresponds to a sequence search space of 6.4×10^{37}. In both problems the forcefield developed by [37] was employed for the energy parameters in the model.

The CPU times it took for the two formulations to converge to the global optimal solution for the two problems are tabulated in Table 2. As shown by the comparison, it is clear that the new formulation with all the superfluous equations taken out outperforms formulation (4) by a large margin. Its CPU times to solve the first and second problem were 82-fold and 327-fold shorter respectively. With its amazing computational performance, the new formulation (formulation (5)) was used as a starting point for the major task

of developing novel formulations for sequence selection into a template with multiple structures.

5 Novel Formulations for Designing Proteins into a Template with Multiple Structures

Derived based on the basic model for single template structure (formulation (5)), the novel formulations outlined in this section represent improvements that select amino acid sequences compatible with a template with multiple crystal or NMR structures. In their derivation two different approaches have been applied: one uses a weighted average forcefield with the weights given by the occurrence frequencies of the C^α-C^α distance between a position pair i and k falling into different distance bins, and the other allows the possibility of spanning all the distance bins that the C^α-C^α distance between i and k covers by the use of binary distance bin variables. The basic ideas behind the development are explained in more detail below.

5.1 *Formulation using a Weighted Average Forcefield*

This approach is relatively simple and straightforward to follow from the model for single template structure. In the case when there is only one structure, the energy parameter $E_{ik}^{jl}(x_i, x_k)$ in the objective function can be immediately determined by the coordinates of the two C^α positions, i.e., x_i and x_k, as well as the amino acid at each of those two positions. There is no ambiguity as to which distance bin d it belongs to. In the case of multiple structures, the term $E_{ik}^{jl}(x_i, x_k)$ can be replaced by a weighted average energy term, $\sum_{d=1}^{b_m} E_{ik}^{jl}(x_i, x_k) wt(x_i, x_k, d)$, where the weights $wt(x_i, x_k, d)$ are given by:

$$wt(x_i, x_k, d) \;=\; \frac{\#\text{ of structures where dist.}(x_i, x_k)\text{ falls into bin } d}{\text{total }\#\text{ of template structures}} \;\forall\; i, k, d$$

The idea can also be examined this way: the distance between x_i and x_k is now replaced by a weighted average distance over all the structures, with the weights given by the above formula. The energy parameters $E_{ik}^{jl}(x_i, x_k)$ can be found using this weighted average distance and simple table lookup in the corresponding forcefield. For instance, in Compstatin (PDB code: 1A1P), a synthetic 13-residue peptide and a pharmaceutical candidate that interferes with complement activation with its details quoted in the case studies section that immediately follows, the distribution for the distance between the alpha carbon of the first residue and the third residue is as follows: bin 4 (5.5 to 6.0 $\mathring{A}$): 1 structure; bin 5 (6.0 to 6.5 $\mathring{A}$): 9 structures; bin 6 (6.5 to 7.0 $\mathring{A}$):

10 structures; and bin 7 (7.0 to 8.0 $\mathring{A}$): 1 structure. Data for the 21 structures were deposited in the Protein Data Bank for Compstatin. Therefore, $wt(x_1, x_3, 4) = \frac{1}{21} = 0.0476$, $wt(x_1, x_3, 5) = \frac{9}{21} = 0.429$, $wt(x_1, x_3, 6) = \frac{10}{21} = 0.476$, $wt(x_1, x_3, 7) = \frac{1}{21} = 0.0476$, and $wt(x_1, x_3, d) = 0 \ \forall d \neq 4, 5, 6, 7$. It should be noticed that in the case of the force field used for generating results presented in this paper [36], the sum of the weights $wt(x_i, x_k, d)$ over the distance bin set $d = 1, \ldots, b_m = 8$ does not equal to one. This is simply because the distance bins only cover the range of 3 to 9 $\mathring{A}$ in the force field. Had the bins covered the full positive distance range, the weights would have added up to one.

All the other components in formulation (5) for single template structure can be kept for this new weighted average forcefield formulation. Therefore in summary, the novel weighted average forcefield formulation for designing proteins into multiple highly flexible templates takes the form:

$$\begin{aligned} \min_{y_i^j, y_k^l} \ & \textstyle\sum_{i=1}^{n} \sum_{j=1}^{m_i} \sum_{k=i+1}^{n} \sum_{l=1}^{m_k} \sum_{d=1}^{b_m} E_{ik}^{jl}(x_i, x_k) wt(x_i, x_k, d) w_{ik}^{jl} \\ \text{subject to} \quad & \textstyle\sum_{j=1}^{m_i} y_i^j = 1 \ \forall \ i \\ & \textstyle\sum_{j=1}^{m_i} w_{ik}^{jl} = y_k^l \ \ \forall \ \ i, k > i, l \\ & \textstyle\sum_{l=1}^{m_k} w_{ik}^{jl} = y_i^j \ \ \forall \ \ i, k > i, j \\ & y_i^j \ , \ y_k^l \ , \ w_{ik}^{jl} = 0 - 1 \ \ \forall \ \ i, j, k, l \end{aligned} \tag{10}$$

Like formulation (4) or formulation (5), it is an integer linear programming (ILP) model.

5.2 *Formulation using Binary Distance Bin Variables*

Another more elegant and advanced approach to incorporate distance information from multiple structures is by using a binary distance bin variable, b_{ikd}, which assumes the value of one if the distance between x_i and x_k falls into distance bin d and the value of zero otherwise. A parameter, $disbin(x_i, x_k, d)$, which will be used in the derivation of the constraints, needs to be defined:

$disbin(x_i, x_k, d)$
$= 1$ if the distance between x_i and x_k in ANY of the template structures falls into bin d
$= 0$ otherwise $\forall \ \ i, k > i, d$

Hence for the first and third residue of Compstatin, parameter $disbin(x_1, x_3, d)$ equals one for $d = 4, 5, 6, 7$ and zero for other distance bins. With this new parameter, the constraint $\sum_{d:disbin(x_i,x_k,d)=1} b_{ikd} = 1 \; \forall i, k > i$ can be imposed. This constraint essentially lets the energy minimization model free to pick only one of the distance bins that all the structures cover. Thus in the same example of Compstatin, the constraint $\sum_{d=4,5,6,7} b_{13d} = 1$ is to be added. Since only the distance bin d with b_{ikd} assigned to be 1 will contribute to the total energy of the protein, when deriving this new formulation, the term $E_{ik}^{jl}(x_i, x_k)$ in the objective function of formulation (5) can be replaced by $\sum_{d:disbin(x_i,x_k,d)=1} E_{ik}^{jl}(x_i, x_k) b_{ikd}$, leading to a new model that looks like:

$$\min_{y_i^j, y_k^l} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \sum_{k=i+1}^{n} \sum_{l=1}^{m_k} \sum_{d:disbin(x_i,x_k,d)=1} E_{ik}^{jl}(x_i, x_k) b_{ikd} w_{ik}^{jl}$$

$$\begin{aligned} \text{subject to} \quad \sum_{j=1}^{m_i} y_i^j &= 1 \; \forall \; i \\ \sum_{j=1}^{m_i} w_{ik}^{jl} &= y_k^l \;\; \forall \;\; i, k > i, l \\ \sum_{l=1}^{m_k} w_{ik}^{jl} &= y_i^j \;\; \forall \;\; i, k > i, j \\ \sum_{d:disbin(x_i,x_k,d)=1} b_{ikd} &= 1 \;\; \forall \;\; i, k > i \\ y_i^j \;,\; y_k^l \;,\; w_{ik}^{jl} \;,\; b_{ikd} &= 0 - 1 \;\; \forall \;\; i, j, k > i, l, d \end{aligned} \tag{11}$$

Formulation (11) is non-convex because of the bilinear term $b_{ikd} w_{ik}^{jl}$ in the objective function. The formulation could have been linearized in the same way as for formulation (1), i.e., by using a positive continuous variable $z_{ikd}^{jl} = b_{ikd} w_{ik}^{jl}$ with the addition of four sets of inequalities to reproduce the original characteristics:

$$\begin{aligned} b_{ikd} + w_{ik}^{jl} - 1 \leq z_{ikd}^{jl} \leq b_{ikd} \; \forall \; i, j, k > i, l, d \\ 0 \leq z_{ikd}^{jl} \leq w_{ik}^{jl} \; \forall \;\; i, j, k > i, l, d \end{aligned} \tag{12}$$

However, based on the observation about the superior computational performance of formulation (5), linearization was done by declaring $z_{ikd}^{jl} = b_{ikd} w_{ik}^{jl}$ as a binary variable and using the RLT equations:

$$w_{ik}^{jl} \sum_{d:disbin(x_i,x_k,d)=1} b_{ikd} = 1 \;\; \forall \;\; i, j, k > i, l \;\; \text{or :}$$

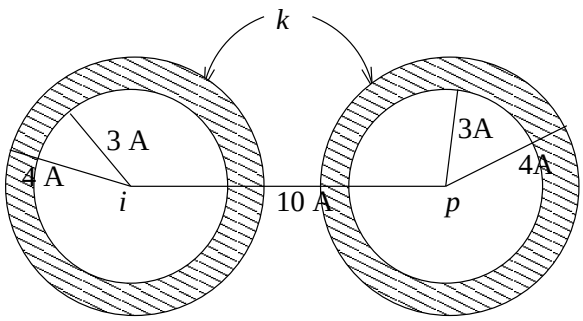

Figure 1. No overlap between the shaded regions where position x_k can possibly be.

$$\sum_{d:disbin(x_i,x_k,d)=1} z_{ikd}^{jl} \;=\; w_{ik}^{jl} \;\; \forall \;\; i,j,k > i,l \tag{13}$$

$$w_{ik}^{jl} \;,\; b_{ikd} \;,\; z_{ikd}^{jl} \;=\; 0-1 \;\; \forall \;\; i,j,k > i,l,d$$

This RLT equation already implies $z_{ikd}^{jl} \leq w_{ik}^{jl} \;\; \forall \;\; i,j,k > i,l,d$, and declaring z_{ikd}^{jl} as a binary variable means $z_{ikd}^{jl} \geq 0 \;\; \forall \;\; i,j,k > i,l,d$. Both of these equations can thus be dropped from the formulation.

Constraints on distance bin variables

Since two alpha carbons are now free to pick any distance bin that the template structures cover, additional constraints have to be imposed on their distance bin variables to avoid physically meaningless results. This requirement can be proved necessary by considering the case shown in Figure 1, in which x_i, x_k, and x_p are three distinct positions and both the distance between x_i and x_k and that between x_k and x_p select to be in bin 1 in the energy minimization model. Assume the average distance between x_i and x_p over all the template structures is $10\mathring{A}$. The selections will result in no overlap between the two shaded regions, each of which corresponds to the area where position x_k can possibly be. The constraints on the distance bin variables serve to ensure there is some kind of consistency about the possible location of any alpha carbon throughout the distance bin selection process.

There are two cases in which the constraints should come into play. The first case is illustrated by Figure 2, where the areas corresponding to the binary variables b_{ikd} and $b_{kpd'}$ do not overlap. The condition for no overlap is: $l_{mid}(d) < dis(i,p) - l_{mid}(d')$, where $dis(i,p)$ is the average distance between x_i and x_p over all template structures. Since if both variables are one there will be no consistency about the location of position x_k, the necessary constraint

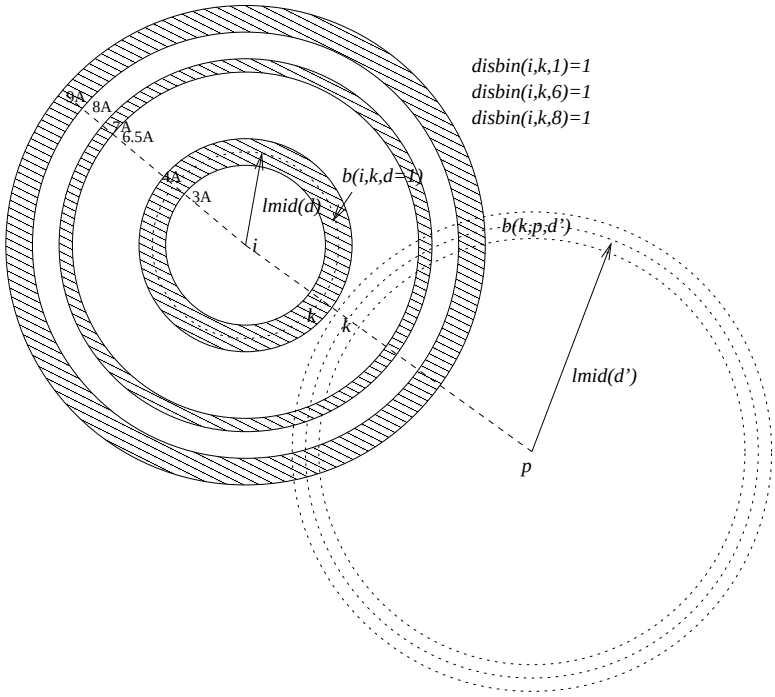

Figure 2. First case in which constraints on distance bin variables are applicable: no overlap between the areas corresponding to the binary variables $b(x_i, x_k, d)$ and $b(x_k, x_p, d')$ because $l_{mid}(d) < dis(i,p) - l_{mid}(d')$. Condition $\sum_{d"=d+1}^{b_m} disbin(x_i, x_k, d") \geq 1$ has to hold to avoid infeasibility.

is: $b_{ikd} + b_{kpd'} \leq 1$. However, it should be highlighted that infeasibility problem may occur to the model if only the no overlap condition is used for checking constraint applicability. This is because an average value has been used for $dis(i,p)$, with which the areas corresponding to the non-zero $disbin(x_i, x_k, d)$'s and that corresponding to the variable $b_{kpd'}$ may not overlap at all. Hence an additional condition has to be imposed besides the no overlap criterion: $\sum_{d"=d+1}^{b_m} disbin(x_i, x_k, d") \geq 1$. It means that there is at least one non-zero $disbin(x_i, x_k, d")$ for $d" > d$, whose area may overlap with that corresponding to $b_{kpd'}$. The model can thus select any of these bins for the distance between x_i and x_k and avoid the problem of infeasibility.

The second case in which the constraints are applicable is shown in Figure 3. This case differs from the first one in its no overlap condition, which can be expressed as the equation $l_{mid}(d') > dis(i,p) + l_{mid}(d)$. Again, to get around the problem of infeasibility, the constraints are only to be applied when $\sum_{d"=d+1}^{b_m} disbin(x_i, x_k, d") \geq 1$, in addition to the no overlap criterion.

In summary, the constraints on binary distance bin variables take the form of:

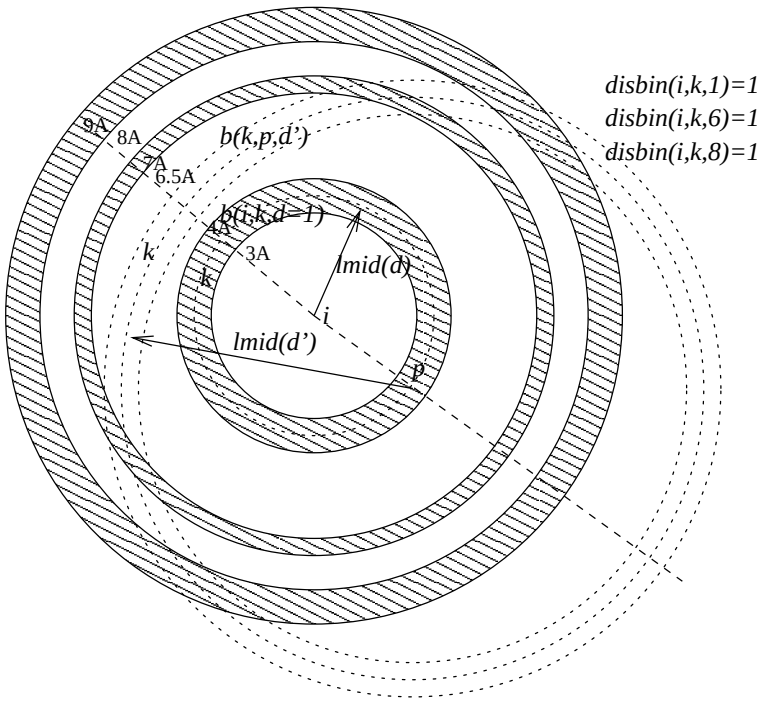

Figure 3. Second case in which constraints on distance bin variables are applicable: no overlap between the areas corresponding to the binary variables $b(x_i, x_k, d)$ and $b(x_k, x_p, d')$ because $l_{mid}(d') > dis(i,p) + l_{mid}(d)$. Condition $\sum_{d"=d+1}^{b_m} disbin(x_i, x_k, d") \geq 1$ has to hold to avoid infeasibility.

$$
\begin{aligned}
& b_{ikd} \;+\; b_{kpd'} \;\leq\; 1 \\
& \text{if } (l_{mid}(d') < dis(i,p) - l_{mid}(d) \;\text{ or }\; l_{mid}(d') > dis(i,p) + l_{mid}(d)) \\
& \text{and } \textstyle\sum_{d"=d+1}^{b_m} disbin(x_i, x_k, d") \;\geq\; 1 \text{ and } disbin(x_i, x_k, d) \;=\; 1 \\
& \text{and } disbin(x_k, x_p, d') \;=\; 1 \;\; \forall \;\; i, k > i, p, d, d', i \neq k \neq p
\end{aligned}
\tag{14}
$$

and the novel formulation for designing proteins into a template with multiple structures, by using binary distance bin variables and their constraints, is:

$$
\min_{y_i^j, y_k^l} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \sum_{k=i+1}^{n} \sum_{l=1}^{m_k} \sum_{d:disbin(x_i,x_k,d)=1} E_{ik}^{jl}(x_i, x_k) z_{ikd}^{jl}
$$

subject to

$$
\begin{aligned}
\textstyle\sum_{j=1}^{m_i} y_i^j \;&=\; 1 \;\forall\; i \\
\textstyle\sum_{j=1}^{m_i} w_{ik}^{jl} \;&=\; y_k^l \;\;\forall \;\; i, k > i, l
\end{aligned}
$$

$$\begin{aligned}
\textstyle\sum_{l=1}^{m_k} w_{ik}^{jl} &= y_i^j \;\; \forall \;\; i,k > i,j \\
\textstyle\sum_{d:disbin(x_i,x_k,d)=1} b_{ikd} &= 1 \;\; \forall \;\; i,k > i \\
b_{ikd} \; + \; w_{ik}^{jl} \; - \; 1 \; &\leq \; z_{ikd}^{jl} \; \leq \; b_{ikd} \; \forall \; i,j,k > i,l,d \\
\textstyle\sum_{d:disbin(x_i,x_k,d)=1} z_{ikd}^{jl} &= w_{ik}^{jl} \;\; \forall \;\; i,j,k > i,l \qquad (15) \\
b_{ikd} \; + \; b_{kpd'} \; &\leq \; 1 \\
\text{if } (l_{mid}(d') < dis(i,p) - l_{mid}(d) \;\; &\text{or} \;\; l_{mid}(d') > dis(i,p) + l_{mid}(d)) \\
\text{and } \textstyle\sum_{d"=d+1}^{b_m} disbin(x_i,x_k,d") \; &\geq \; 1 \text{ and } disbin(x_i,x_k,d) \; = \; 1 \\
\text{and } disbin(x_k,x_p,d') \; &= \; 1 \; \forall \;\; i,k > i,p,d,d', i \neq k \neq p \\
y_i^j \;,\; y_k^l \;,\; w_{ik}^{jl} \;,\; b_{ikd} \;,\; b_{kpd'} \;,\; z_{ikd}^{jl} \; &= \; 0-1 \\
\forall \;\; i,j,k > i,l,p \neq k \neq i,d,d'
\end{aligned}$$

Implementation of the models

For higher computational efficiency purpose, the objective function of formulation (5) is written in the GAMS [38] program as the addition of four terms: one for the case when both position i and position k are varied, one for the case when only position i is varied, one for the case when only position k is varied, and the remaining one for the case when neither i nor k is varied. If a non-mutated position i is fixed at amino acid j, a parameter, $yfx(i,j)$, can be used in place of the variable y_i^j. Similarly parameter $yfx(k,l)$ can be written instead of the variable y_k^l if position k is known to be occupied by amino acid l. The objective function $\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{k=i+1}^{n}\sum_{l=1}^{m_k} E_{ik}^{jl}(x_i,x_k)w_{ik}^{jl}$ is thus actually implemented as:

$$\begin{aligned}
&\sum_{i:i\ varied}\sum_{j=1}^{m_i}\sum_{k:k>i,k\ varied}\sum_{l=1}^{m_k} E_{ik}^{jl}(x_i,x_k)w_{ik}^{jl} + \\
&\sum_{i:i\ varied}\sum_{j=1}^{m_i}\sum_{k:k>i,k\ fixed}\sum_{l=1}^{m_k} E_{ik}^{jl}(x_i,x_k)y_i^j yfx(k,l) + \\
&\sum_{i:i\ fixed}\sum_{j=1}^{m_i}\sum_{k:k>i,k\ varied}\sum_{l=1}^{m_k} E_{ik}^{jl}(x_i,x_k)yfx(i,j)y_k^l +
\end{aligned}$$

$$\sum_{i:i\ fixed}\sum_{j=1}^{m_i}\sum_{k:k>i,k\ fixed}\sum_{l=1}^{m_k} E_{ik}^{jl}(x_i,x_k)yfx(i,j)yfx(k,l)$$

This way of implementation proves to be better than treating all positions as mutated positions in saving computation time.

A similar type of expansion on the objective function can be done when using the weighted average structure formulation for designing proteins into a template with multiple structures. For the formulation using distance bin variables, a similar expansion will result in an objective function that looks like:

$$\sum_{i:i\ varied}\sum_{j=1}^{m_i}\sum_{k:k>i,k\ varied}\sum_{l=1}^{m_k}\sum_{d:disbin(x_i,x_k,d)=1} E_{ik}^{jl}(x_i,x_k)b_{ikd}w_{ik}^{jl} +$$

$$\sum_{i:i\ varied}\sum_{j=1}^{m_i}\sum_{k:k>i,k\ fixed}\sum_{l=1}^{m_k}\sum_{d:disbin(x_i,x_k,d)=1} E_{ik}^{jl}(x_i,x_k)y_i^j yfx(k,l)b_{ikd} +$$

$$\sum_{i:i\ fixed}\sum_{j=1}^{m_i}\sum_{k:k>i,k\ varied}\sum_{l=1}^{m_k}\sum_{d:disbin(x_i,x_k,d)=1} E_{ik}^{jl}(x_i,x_k)yfx(i,j)y_k^l b_{ikd} +$$

$$\sum_{i:i\ fixed}\sum_{j=1}^{m_i}\sum_{k:k>i,k\ fixed}\sum_{l=1}^{m_k}\sum_{d:disbin(x_i,x_k,d)=1} E_{ik}^{jl}(x_i,x_k)yfx(i,j)yfx(k,l)b_{ikd}$$

As shown in the equation, all four terms except the last one are non-linear. The bilinear product $b_{ikd}w_{ik}^{jl}$ is, as mentioned, linearized by the binary variable $z_{ikd}^{jl} = b_{ikd}w_{ik}^{jl}$ with the RLT equations: $\sum_{d:disbin(x_i,x_k,d)=1} z_{ikd}^{jl} = w_{ik}^{jl} \quad \forall \quad i,j,k > i,l$. In an analogous fashion, the bilinear terms $y_i^j b_{ikd}$ can be linearized by using the binary variable $u_{ikd}^j = y_i^j b_{ikd}$ with the corresponding RLT equations: $\sum_{d:disbin(x_i,x_k,d)=1} u_{ikd}^j = y_i^j \quad \forall \quad i,j,k > i$ and $\sum_{j=1}^{m_i} u_{ikd}^j = b_{ikd} \quad \forall \quad i,k > i,d$, which are obtained by multiplying variable y_i^j with the constraint $\sum_{d:disbin(x_i,x_k,d)=1} b_{ikd} = 1 \quad \forall \quad i,k > i$, and variable b_{ikd} with the constraint $\sum_{j=1}^{m_i} y_i^j = 1 \quad \forall \quad i$, respectively. The bilinear terms $y_k^l b_{ikd}$ are linearized by using the binary variable $v_{ikd}^l = y_k^l b_{ikd}$ with the corresponding RLT equations: $\sum_{d:disbin(x_i,x_k,d)=1} v_{ikd}^l = y_k^l \quad \forall \quad i,l,k > i$ and $\sum_{l=1}^{m_k} v_{ikd}^l = b_{ikd} \quad \forall \quad i,k > i,d$, which are obtained by multiplying variable y_k^l with the constraint $\sum_{d:disbin(x_i,x_k,d)=1} b_{ikd} = 1 \quad \forall \quad i,k > i$, and variable b_{ikd} with the constraint $\sum_{l=1}^{m_k} y_k^l = 1 \quad \forall \quad k > i$, respectively. With the new

variables u^j_{ikd} and v^l_{ikd} for actual implementation, two more sets of second level RLTs can be added to speed up the branch-and-bound algorithm even further. They are obtained by multiplying variable b_{ikd} to the first level RLTs, $\sum_{j=1}^{m_i} w^{jl}_{ik} = y^l_k \;\; \forall \;\; i,k > i,l$ and $\sum_{l=1}^{m_k} w^{jl}_{ik} = y^j_i \;\; \forall \;\; i,k > i,j$, respectively, yielding:

$$
\begin{aligned}
b_{ikd} \sum_{j=1}^{m_i} w^{jl}_{ik} &= y^l_k \;\; \forall \;\; i,k > i,l,d \quad \text{or} \\
\sum_{j=1}^{m_i} z^{jl}_{ikd} &= v^l_{ikd} \;\; \forall \;\; i,k > i,l,d
\end{aligned} \tag{16}
$$

and:

$$
\begin{aligned}
b_{ikd} \sum_{l=1}^{m_k} w^{jl}_{ik} &= y^j_i \;\; \forall \;\; i,k > i,j,d \quad \text{or} \\
\sum_{l=1}^{m_k} z^{jl}_{ikd} &= u^j_{ikd} \;\; \forall \;\; i,k > i,j,d
\end{aligned} \tag{17}
$$

To summarize, the version of the formulation using distance bin variables for actual implementation should be:

$$
\begin{aligned}
\min_{y^j_i, y^l_k} & \sum_{i:i\ varied} \sum_{j=1}^{m_i} \sum_{k:k>i,k\ varied} \sum_{l=1}^{m_k} \sum_{d:disbin(x_i,x_k,d)=1} E^{jl}_{ik}(x_i,x_k) b_{ikd} w^{jl}_{ik} + \\
& \sum_{i:i\ varied} \sum_{j=1}^{m_i} \sum_{k:k>i,k\ fixed} \sum_{l=1}^{m_k} \sum_{d:disbin(x_i,x_k,d)=1} E^{jl}_{ik}(x_i,x_k) u^j_{ikd} yfx(k,l) + \\
& \sum_{i:i\ fixed} \sum_{j=1}^{m_i} \sum_{k:k>i,kvaried} \sum_{l=1}^{m_k} \sum_{d:disbin(x_i,x_k,d)=1} E^{jl}_{ik}(x_i,x_k) v^l_{ikd} yfx(i,j) + \\
& \sum_{i:i\ fixed} \sum_{j=1}^{m_i} \sum_{k:k>i,kfixed} \sum_{l=1}^{m_k} \sum_{d:disbin(x_i,x_k,d)=1} E^{jl}_{ik}(x_i,x_k) yfx(i,j) yfx(k,l) b_{ikd}
\end{aligned}
$$

subject to

$$
\begin{aligned}
\textstyle\sum_{j=1}^{m_i} y_i^j &= 1 \;\forall\; i \\
\textstyle\sum_{j=1}^{m_i} w_{ik}^{jl} &= y_k^l \;\;\forall\;\; i,k > i,l \\
\textstyle\sum_{l=1}^{m_k} w_{ik}^{jl} &= y_i^j \;\;\forall\;\; i,k > i,j \\
\textstyle\sum_{d:disbin(x_i,x_k,d)=1} b_{ikd} &= 1 \;\;\forall\;\; i,k > i \\
b_{ikd} + w_{ik}^{jl} - 1 &\leq z_{ikd}^{jl} \leq b_{ikd} \;\forall\; i,j,k > i,l,d \\
\textstyle\sum_{d:disbin(x_i,x_k,d)=1} z_{ikd}^{jl} &= w_{ik}^{jl} \;\;\forall\;\; i,j,k > i,l \qquad (18) \\
b_{ikd} + b_{kpd'} &\leq 1 \\
\text{if } (l_{mid}(d') < dis(i,p) - l_{mid}(d) &\;\text{ or }\; l_{mid}(d') > dis(i,p) + l_{mid}(d)) \\
\text{and } \textstyle\sum_{d"=d+1}^{b_m} disbin(x_i,x_k,d") &\geq 1 \text{ and } disbin(x_i,x_k,d) = 1 \\
\text{and } disbin(x_k,x_p,d') &= 1 \;\;\forall\;\; i,k > i,p,d,d',i \neq k \neq p \\
\textstyle\sum_{d:disbin(x_i,x_k,d)=1} u_{ikd}^j &= y_i^j \;\;\forall\;\; i,j,k > i \\
\textstyle\sum_{j=1}^{m_i} u_{ikd}^j &= b_{ikd} \;\;\forall\;\; i,k > i,d \\
\textstyle\sum_{d:disbin(x_i,x_k,d)=1} v_{ikd}^l &= y_k^l \;\;\forall\;\; i,l,k > i \\
\textstyle\sum_{l=1}^{m_k} v_{ikd}^l &= b_{ikd} \;\;\forall\;\; i,k > i,d \\
\textstyle\sum_{l=1}^{m_k} z_{ikd}^{jl} &= u_{ikd}^j \;\;\forall\;\; i,k > i,j,d \\
\textstyle\sum_{j=1}^{m_i} z_{ikd}^{jl} &= v_{ikd}^l \;\;\forall\;\; i,k > i,l,d \\
y_i^j\;,\; y_k^l\;,\; w_{ik}^{jl}\;,\; b_{ikd}\;,\; b_{kpd'}\;,\; u_{ikd}^j\;,\; v_{ikd}^l\;,\; z_{ikd}^{jl} &= 0-1 \\
\forall\;\; i,j,k > i,l,p \neq k \neq i,d,d'
\end{aligned}
$$

6 Case Studies

Both the formulation using a weighted average structure and the formulation using binary distance bin variables were employed for the sequence selection in redesigning Compstatin, a 13-residue peptide. Results generated from the two different formulations will be compared.

Compstatin

Compstatin (PDB code: 1A1P) is a synthetic 13-residue cyclic peptide that inhibits the cleavage of C3 to C3a and C3b in the human complement sys-

tem and thus hinders complement activation. It is a novel drug candidate identified through the screening of a phage-displayed random peptide library with C3b, a proteolytically activated form of complement C3, and was later truncated to its present 13-residue form without loss of activity [39]. Although complement activation is part of normal inflammatory response, inappropriate complement activation can cause host-cell damage, which is the case in more than 25 pathological conditions, including autoimmune diseases, stroke, heart attack, Alzheimer's disease, and burn injuries [40].

Compstatin has shown highly promising results in numerous clinically relevant trials conducted recently. Compstatin blocked the cleavage of C3 to the pro-inflammatory peptide C3a and the opsonin C3b in hemolytic assays and in human normal serum [39] [41], prevented heparine/protamine-induced complement activation in baboons in a situation resembling heart surgery [42], inhibited complement activation during the contact of blood with biomaterial in a model of extra-corporeal circulation [43], increased the lifetime of survival of porcine kidneys perfused with human blood in a hyper-acute rejection xenotransplantation model [44], blocked the E coli -induced oxidative burst of granulocytes and monocytes [45], and inhibited complement activation by cell lines SH-SY5Y, U-937, THP-1 and ECV304 [46]. Compstatin was stable in biotranformation studies in vitro in human blood, normal human plasma and serum, with increased stability upon N-terminal acetylation [41]. Compstatin showed little or low toxicity and no adverse effects when these were measured [42–44]. Finally, compstatin showed species-specificity and is active only with human and primate C3 [47].

De novo design on Compstatin is aimed at acquiring the sequences corresponding to the best inhibitors to C3 and thus the most potent drugs for diseases related to inappropriate complement activation. Recently [48] did experimental studies that revealed some sequence-function relationships about Compstatin. Their experimental findings can be used to match against the computational results from the de novo design.

The flexible templates for Compstatin

The flexible template for Compstatin as defined by its 21 NMR structures available from the PDB is shown in Figure 4. The structures in this case do not deviate from each other by too much. Each of them corresponds to the native sequence of Ile^1-Cys^2-Val^3-Val^4-Gln^5-Asp^6-Trp^7-Gly^8-His^9-His^{10}-Arg^{11}-Cys^{12}-Thr^{13}-NH_2, with a disulfide bond connecting the two cysteines at positions 2 and 12. Similar to several other immunogenic peptides, Compstatin adopts a β-turn structure, which is located across Gln^5-Asp^6-Trp^7-Gly^8. The β-turn is considered to be important for functional recognition as it is where side-chain interactions exist between turn residues and C3 [41].

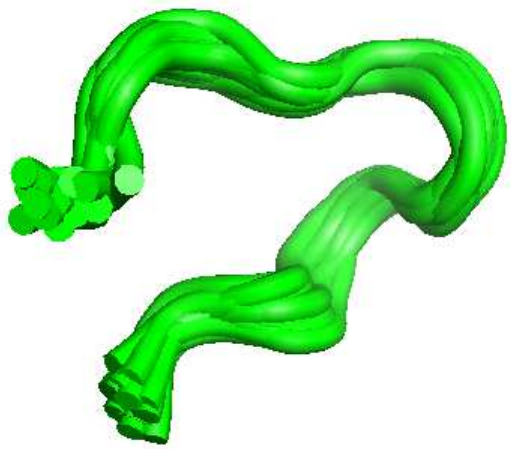

Figure 4. Flexible template of Compstatin for de novo protein design as illustrated by overlapping its 21 NMR structures available from the Protein Data Bank.

Mutation set

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 [49]. 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 se-

lected 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 the hydrophobic amino acid set (A,F,I,L,M,W,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, all residues were allowed, excluding cystine and tryptophan. This mutation set leads to a problem with complexity 3.0×10^6. With both the forcefield developed by [37] and the one by [36], 500 sequences were generated for each of the formulation using a weighted average forcefield and the formulation using binary distance bin variables.

Results

The percentage occurrence of the selected amino acids at each of the 6 varied positions in the 500 sequence solutions is tabulated in Table 3.

The results show that for both forcefields, the formulation using weighted average energy and the formulation using binary distance bin variables gave very similar results in the case of Compstatin. This may be due to the slight deviation from each other among the 21 structures of the Compstatin template. The two forcefields suggested slightly different predictions for each varied position. However, both forcefields suggested tryptophan (W) for position 4, which is in agreement with the experimental finding by [48], who proposed tryptophan or fused-ring non-natural amino acids at position 4 would contribute to high inhibitory activity of the peptide.

7 Conclusions

Two different formulations were derived for handling the common case of de novo protein design into highly flexible templates. One formulation uses weighted average energy parameters, with the weights, which depend on two C^α positions and a particular distance bin, given by the probability that the distance between the two C^α positions is found in that distance bin in any of the structures. By using binary distance bin variables, the other formulation

Formulation using a weighted average forcefield			
Varied position	Native residue	Selections by the model	
		High Resolution forcefield [36]	LKF forcefield [37]
1	I	W, A, F, V	A, Y, V
4	V	F, W, Y	I, L, V, Y, W
9	H	T, H, K, R, F, I	A, N, P, S
10	H	E, F, H, V, N, T, A	Y, F, H
11	R	A, F	A, E, D, N
13	T	A, W	A, Y, V
Formulation using binary distance bin variables			
Varied position	Native residue	Selections by the model	
		High Resolution forcefield [36]	LKF forcefield [37]
1	I	F, W, V, A	Y, A, L, F
4	V	F, W, Y	I, Y, V, L, W
9	H	I, T, F, H, R, M, L	P, A, S, V, N
10	H	F, T, V, E	Y, F, H, V
11	R	A, F, V	A, N, E, D
13	T	W, A, F	A, Y, M, V

Table 3. Sequence selection results for de novo protein design with the flexible template of Compstatin (21 NMR structures) using two different formulations. Selected amino acids with less than 5% occurrence are not listed.

allows the distance between the two C^{α} positions to fall into any distance bin that all the structures span over. In the meantime, by imposing linear constraints on the binary distance bin variables, it also maintains consistency about the location of any C^{α} position and avoids results that might otherwise suggest a physically impossible structure. Both formulations were developed based on a highly computational efficient formulation for solving the sequence selection problem in designing proteins into a template with a single structure. Finally, they were tested on the case study of Compstatin, which has a flexible template of 21 NMR structures.

Acknowledgments

CAF gratefully acknowledges financial support from the National Science Foundation and the National Institutes of Health (R01 GM52032, R24 GM069736).

References

[1] Floudas, C. A., 2005, Research Challenges, Opportunities and Synergism in Systems Engineering and Computational Biology. *AIChE Journal,***51**, 1872–1884.

[2] Floudas, C. A., Fung, H. K., McAllister, S. R., Mönnigmann, M. and Rajgaria, R., 2006, Advances in Protein Structure Prediction and De Novo Protein Design: A Review. *Chemical Engineering Science,***61**, 966–988.

[3] Kuhlman, B. and Baker, D., 2004, Exploring Folding Free Energy Landscapes Using Computational Protein Design. *Current Opinion in Structural Biology,***14**, 89–95.

[4] Lilien, R. H., Stevens, B. W., Anderson, A. C. and Donald, B. R., 2005, A Novel Ensemble-Based Scoring and Search Algorithm for Protein Redesign and Its Application to Modify the Substrate Specificity of the Gramicidin Synthetase A Phenylalanine Adenylation Enzyme. *Journal of Computational Biology,***12**, 740–761.

[5] McFarland, B. J., Kortemme, T., Yu, S. F., Baker, D. and Strong, R. K., 2003, Symmetry Recognizing Asymmetry: Analysis of the Interactions between the C-Type Lectin-like Immunoreceptor NKG2D and MHC Class I-like Ligands. *Structure,***11**,411–422.

[6] Niv, M. Y. and Weinstein, H., 2005, A Flexible Docking Procedure for the Exploration of Peptide Binding Selectivity to Known Structures and Homology Models of PDZ Domains. *Journal of the American Chemical Society,***127**, 14072–14079.

[7] Kortemme, T. and Baker, D., 2004, Computational Design of Protein-Protein Interactions. *Current Opinion in Chemical Biology,***8**, 91–97.

[8] Malakauskas, S. M. and Mayo, S. L., 1998, Design, Structure, and Stability of a Hyperthermophilic Protein Variant. *Nature Structural Biology,***5**, 470–475.

[9] Klepeis, J. L., Floudas, C. A., Morikis, D., Tsokos, C. G., Argyropoulos, E., Spruce, L. and Lambris, J. D., 2003, Integrated Computational and Experimental Approach for Lead Optimization and Design of Compstatin Variants with Improved Activity. *Journal of the American Chemical Society,***125**, 8422–8423.

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

[11] Hellinga, H. W. and Richards, F. M., 1991, Construction of New Ligand Binding Sites in Proteins of Known Structure I. Computer Aided Modeling of Sites with Predefined Geometry. *Journal of Molecular Biology,***222**, 763–785.

[12] Richards, F. M., Caradonna, J. P. and Hellinga, H. W., 1991, Construction of New Ligand Binding Sites in Proteins of Known Structure. II. Grafting of a Buried Transition Metal Binding Site into *Escherichia Coli* Thioredoxin. *Journal of Molecular Biology,***222**, 787–803.

[13] Shimaoka, M., Shifman, J. M., Jing, H., Takagi, L., Mayo, S. L. and Springer, T. A., 2000, Computational Design of an Intergrin I Domain Stabilized in the Open High Affinity Conformation. *Nature Structural Biology,***7**, 674–678.

[14] Kraemer-Pecore, C. M., Wollacott, A. M. and Desjarlais, J. R., 2001, Computational Protein Design. *Current Opinion in Chemical Biology,***5**, 690–695.

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

[16] Fung, H. K., Rao, S., Floudas, C. A., Prokopyev, O., Pardalos, P. M. and Rendl, F., 2005, Computational Comparison Studies of Quadratic Assignment Like Formulations for the In Silico Sequence Selection Problem in De Novo Protein Design. *Journal of Combinatorial Optimization,***10**, 41–60.

[17] Lim, W. A., Hodel, A., Sauer, R. T. and Richards, F. M., 1994, The Crystal Structure of a Mutant Protein With Altered But Improved Hydrophobic Core Packing. *Proceedings of the National Academy of Sciences of the United States of America,***91**, 423–427.

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

[19] Mooers, B. H. M., Datta, D., Baase, W. A., Zollars, E. S., Mayo, S. L. and Matthews, B. W., 2003, Repacking the Core of T4 Lysozyme by Automated Design. *Journal of Molecular Biology,***332**, 741–756.

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

[21] Kuhlman, B. and Baker, D., 2000, Native Protein Sequences Are Close to Optimal for Their Structures. *Proceedings of the National Academy of Sciences of the United States of America,***97**, 10383–10388.

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

[23] Desjarlais, J. R. and Handel, T. M., 1999, Side Chain and Backbone Flexibility in Protein Core Design. *Journal of Molecular Biology,***290**, 305–318.
[24] Kuhlman, B., Dantae, G., Ireton, G. C., Verani, G., Stoddard, B. and Baker, D., 2003, Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. *Science,***302**, 1364–1368.
[25] Saunders, C. T. and Baker, D., 2005, Recapitulation of Protein Family Divergence Using Flexible Backbone Protein Design. *Journal of Molecular Biology,***346**, 631–644.
[26] Klepeis, J. L., and Schafroth, H. D., Westerberg, K. M. and Floudas, C. A., 2002, 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 (Ed), *Advances in Chemical Physics.*New York, NY: Wiley, pp. 254–457.
[27] Klepeis, J. L., Floudas, C. A., Morikis, D. and Lambris, J. D., 1999, Predicting Peptide Structures Using NMR Data and Deterministic Global Optimization. *Journal of Computational Chemistry,***20**, 1354–1370.
[28] Adjiman, C., Androulakis, I. and Floudas, C. A., 1998, A Global Optimization Method, αBB, for General Twice-Differential Constrained NPLs - I. Theoretical Advances. *Computers and Chemical Engineering,***22**, 1137–1158.
[29] Adjiman, C., Androulakis, I. and Floudas, C. A., 1998, A Global Optimization Method, αBB, for General Twice-Differentiable Constrained NLPs - II. Implementation and Computational Results. *Computers and Chemical Engineering,***22**, 1159–1179.
[30] Adjiman, C., Androulakis, I. and Floudas, C. A., 2000, Global Optimization of Mixed-Integer Nonlinear Problems. *AIChE Journal,***46**, 1769–1797.
[31] Klepeis, J. L. and Floudas, C. A., 1999, Free Energy Calculations for Peptides via Deterministic Global Optimization. *Journal of Chemical Physics,***110**, 7491–7512.
[32] Floudas, C. A., 2000, *Deterministic Global Optimization : Theory, Methods and Applications.* Nonconvex Optimization and its Applications. Kluwer Academic.
[33] Floudas, C. A., 1995, *Nonlinear and Mixed-Integer Optimization: Fundamentals and Applications.*New York, NY: Oxford University Press.
[34] CPLEX, 1997, *Using the CPLEX Callable Library.* ILOG, Inc.
[35] Sherali, H. D., and Adams, W. P., 1999,*A Reformulation Linearization Technique for Solving Discrete and Continuous Nonconvex Problems.*Boston, MA: Kluwer Academic.
[36] Rajgaria, R., McAllister, S. R., and Floudas, C. A., 2006, Improved High Resolution Force Fields for de Novo Protein Design. *Proteins: Structure, Function, and Bioinformatics,* accepted for publication.
[37] Loose, C., Klepeis, J. L. and Floudas, C. A., 2004, A New Pairwise Folding Potential based on Improved Decoy Generation and Side Chain Packing. *Proteins: Structure, Function, and Bioinformatics,***54**, 303–314.
[38] GAMS Development Corporation, 2005, *GAMS: A Users Guide.*Washington, D.C.
[39] Sahu, A., Kay, B. K. and Lambris, J. D., 1996, Inhibition of Human Complement by a C3-binding Peptide Isolated from a Phage Displayed Random Peptide Library. *Journal of Immunology,***157**, 884–891.
[40] Sahu, A. and Lambris, J. D., 2001, Structure and Biology of Complement Protein C3, a Connecting Link between Innate and Acquired Immunity. *Immunological Reviews,***180**, 35–48.
[41] Sahu, A. and Soulika, A. M., and Morikis, D. and Spruce, L., Moore, W. T., and Lambris, J. D., 2000, Binding Kinetics, Structure Activity Relationship and Biotransformation of the Complement Inhibitor Compstatin. *Journal of Immunology,***165**, 2491–2499.
[42] Soulika, A. M., and Khan, M. M. and Hattori, T. and Bowen, F. W., and Richardson, B. A., and Hack, C. E., and Sahu, A., and Edmunds, L. H., and Lambris, J. D., 2000, Inhibition of Heparin/Protamine Complex-induced Complement Activation by Comsptatin in Baboons. *Clinical Immunology,***96**, 212–221.
[43] Nillson, B., Larsson, R., Hong, J., Elgue, G., Ekdahl, K. N., Sahu, A. and Lambris, J. D., 1998, Compstatin Inhibits Complement and Cellular Activation in Whole Blood in Two Models of Extracorporeal Circulation. *Blood,***92**, 1661–1667.
[44] Fiane, A. E., Mollnes, T. E., Videm, V., Hovig, T., Hogasen, K., Mellbye, O. J., Spruce, L., Moore, W. T., Sahu, A., and Lambris, J. D., 1999, Compstatin, a Peptide Inhibitor of C3, Prolongs Survival of Ex-vivo Perfused Pig Xenografts. *Xenotransplantation,***6**, 52–65.
[45] Mollnes, T. E., Brekkem, O. L., Fung, M., Fure, H., Christiansen, D., Bergseth, G., Videm, V., Lappegard, K. T., Kohl, J. and Lambris, J. D., 2002, 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.
[46] Klegeris, A., Singh, E. A. and McGeer, P. L., 2002, Effects of C-reactive Protein and Pentosan

Polysulphate on Human Complement Activation. *Immunology,***106**, 381–388.

[47] Sahu, A., Morikis, D. and Lambris, J. D., 2003, Compstatin, a Peptide Inhibitor of Complement, Exhibits Species-specific Binding to Complement Component C3. *Molecular Immunology,***39**, 557–566.

[48] Mallik, B., Katragadda, M., Spruce, L. A., Carafides, C., Tsokos, C. G., Morikis, D., and Lambris, J. D., Design and NMR Characterization of Active Analogues of Compstatin Containing Non-natural Amino Acids. *Journal of Medicinal Chemistry,***48**, 274–286.

[49] Morikis, D., Roy, M., Sahu, A., Torganis, A., Jennings, P. A., Tsokos, G. C. and Lambris, J. D., 2002, The Structural Basis of Compstatin Activity Examined by Structure-function-based Design of Peptide Analogs and NMR. *Journal of Biological Chemistry,***277**, 14942–14953.