Supplementary Information for

Integrated Computational and Experimental Approach for Lead Optimization and Design of Compstatin Variants with Improved Activity

J. L. Klepeis, C. A. Floudas,
Department of Chemical Engineering
Princeton University
Princeton, NJ 08544-5263, USA
D. Morikis,
Department of Chemical and Environmental Engineering
University of California at Riverside
Riverside, CA 92521, USA
C. G. Tsokos, E. Argyropoulos, L. Spruce, and J. D. Lambris
Department of Pathology and Laboratory Medicine
University of Pennsylvania
Philadelphia, PA 19104, USA

Methods

Sequence Selection

To correctly select a sequence compatible with a given backbone template, an appropriate energy function must first be identified. The proposed sequence selection procedure is based on solving an optimization problem that employs a pairwise distance-dependent interaction potential. A number of different parameterizations for pairwise residue interaction potentials exist; the one employed here is based on the discretization of alpha carbon distances 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. 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 ($N(N+1)13 = 2730$, where $N = 20$, the number of naturally occurring amino acids), was shown to provide higher Z scores than other potentials and place native folds lower in energy (see reference [8] of paper, and [8] for availability of the parameters).

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 equivalent 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 :

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

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 at most one type of amino acid at each position. For the general case, the binary variables appear as bilinear combinations in the objective function. This objective can be reformulated as a strictly linear (integer linear programming) problem. The solution of the integer linear programming problem (ILP) can be accomplished rigorously and deterministically using branch and bound techniques, making convergence to the global minimum energy sequence consistent and reliable. This is accomplished using CPLEX (available from ILOG, Inc.), a commercially available package for solving such ILP problems. 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, and repetitive solution of the ILP problem.

Fold Stability and Specificity

Once a set of low lying energy sequences have been identified via the sequence selection procedure, the fold validation stage is used to identify the most optimal sequences according to a rigorous quantification of specificity through the calculation 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. On the other hand, the second simulation is treated as a free folding calculation 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 structural restraints defined by the NMR ensemble (in the case of compstatin and its analogs a deviation of 1.5 angstroms is allowed for each non-consecutive Cα-Cα distance from the known NMR structures), or simply by allowing some deviation from a subset of distances provided by the structural template, and hence they allow for a flexible template on the backbone.

The formulations for the folding calculations are reminiscent of structure prediction problems in protein folding, and these formulations can be solved to generate an ensem-

ble of lower energy conformations using a variety of search techniques that allow for the enforcement of distance restraints such as Monte Carlo, Simulated Annealing and Genetic Algorithms. In this work, a novel constrained global optimization problem first introduced for structure prediction of compstatin using NMR data, and later employed in a generic framework for the structure prediction of proteins, is utilized. 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 (see reference [9] of paper).

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. 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*).

$$p_{temp} = \frac{\sum_{i \in temp} \exp[-\beta E_i]}{\sum_{i \in total} \exp[-\beta E_i]} \tag{2}$$

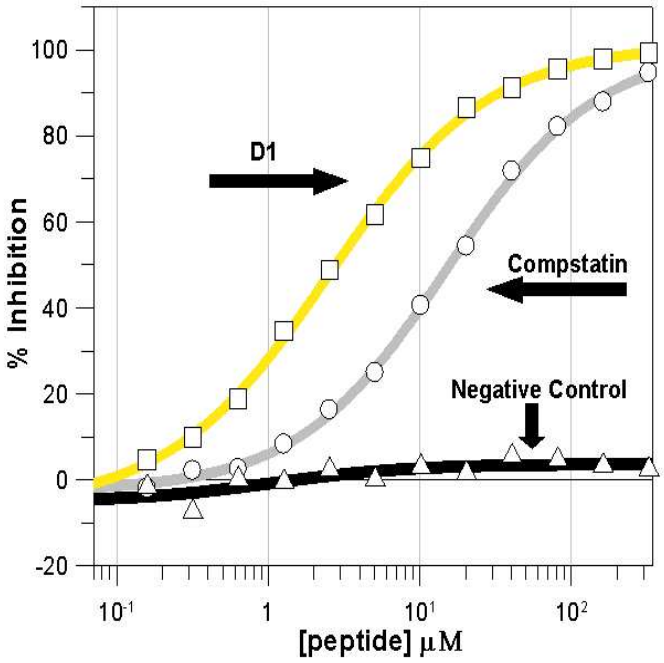
Here $\exp[-\beta E_i]$ is the statistical weight for conformer i, which was calculated at room temperature (i.e., 298 K).

Peptide Synthesis and Complement Inhibition Assays

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.

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

Set of sequences tested for fold stability and specificity.



Percent of complement inhibition as a function of peptide concentration for parent peptide compstatin, the most active analog D1, and an inactive peptide used as negative control.