426f A Novel Systems Engineering Approach for in Silico Sequence Selection in De Novo Protein Design

Ho Ki Fung and Christodoulos A. Floudas

De novo peptide or protein design starts with a flexible 3-dimensional protein backbone and involves the search for all amino acid sequences that will fold into such a template ([1]-[4]) and exhibit higher stability and/or higher activity. Our current de novo protein design framework consists of two stages: (i) in silico sequence selection, and (ii) fold validation. Confirmed with experimental data, the framework proved to be highly successful as it predicted de novo sequences with up to 16-fold higher activity than the mother sequence for Compstatin ([1] and [2]), a 13-residue therapeutic peptide that binds to the complement component C3 and inhibits complement activation. Current work is performed along two major directions: (i) enhancing the optimization algorithms in the sequence selection stage so that the lowest energy sequence can be computed in shorter time, and (ii) extending the application of the de novo protein design framework to larger peptides/proteins.

Computational efficiency for in silico sequence selection is important because of the *NP*-hard nature of the problem ([5]) and the ultimate goal of full-sequence-full-combinatorial optimization of proteins of practical size (i.e., over 100 residues). Algorithmic enhancement is done first by comparing the computational times for the original $O(n^2)$ formulation, which means the number of linear constraints scales with the square of the number of binary variables, n, and those for other $O(n)$ formulations to execute sequence search for the same set of test problems ([5]). The equivalent $O(n)$ formulations were obtained through literature search ([6]-[8]). Promising new components of relaxing the reformulation linearization techniques (RLTs) to inequality constraints, adding triangle inequalities, and doing Dead-End Elimination (DEE) type preprocessing are also examined. Finally, we will illustrate how tighter inequalities on the binary variables can be generated by solving smaller sub-problems with novel energy comparison constraints. With all the algorithmic improvements, we can currently solve problems of complexity 20^{35}, which correspond to the full combinatorial optimization of 35 positions in a protein, within reasonable timeframe on just a single processor.

In addition to Compstatin, de novo sequences were generated for human beta defensin (hβD-2), a 41-residue cationic peptide crucial to innate immunity in the human immune system, as well as C3a, a 77-residue fragment of the human complement component C3 and a potent mediator of inflammation. The novel sequences were submitted to our research collaborators for experimental validation on their fold, stability, and activity. It should be noted that in our work on hβD-2, homology search was performed to elucidate all properties that are highly conserved among the hβD-2 homologs, and the conserved properties were in turn converted to biological constraints in our sequence selection model so that all the de novo sequences produced would observe the conservation. Examples of the biological constraints are constraints on the number of positive charges, negative charges, and total charges on certain portions of as well as the whole peptide, those on the total number of hydrophobic residues in β-strands and in the whole peptide for stability purpose, and others on the occurrence frequency for each type of amino acid. Other peptide/protein candidates under consideration for testing the framework include C5a, a peptide similar in both structure and function to C3a, and G-protein coupled receptors.

[1]- J.L. Klepeis and C.A. Floudas and D. Morikis and C.G. Tsokos and E. Argyropoulos and L. Spruce and J.D. Lambris. "Integrated Computational and Experimental Approach for Lead Optimization and Design of Compstatin Variants with Improved Activity." *J. Am. Chem. Soc.* 125 (2003): 8422-8423. [2]- J.L. Klepeis and C.A. Floudas and D. Morikis and C.G. Tsokos and J.D. Lambris. "Design of Peptide Analogs with Improved Activity Using a Novel de Novo Protein Design Approach." *Ind. Eng. Chem. Res.* 43 (2004): 3817-3826. [3]- J.L. Klepeis and Y. Wei and M.H. Hecht and C.A. Floudas. "Ab Initio Prediction of the Three-Dimensional Structure of a De Novo Designed Protein: A Double-Blind Case

Study." *Proteins.* 58 (2005): 560-570. [4]- C.A. Floudas and H.K. Fung and S.R. McAllister and M. Mönnigmann and R. Rajgaria. "Advances in Protein Structure Prediction and De Novo Protein Design: A Review." *Chem. Eng. Sci.* (2005): in press. [5]- H.K. Fung and S. Rao and C.A. Floudas and O. Prokopyev and P.M. Pardalos and F. Rendl. "Computational Comparison Studies of Quadratic Assignment Like Formulations for the In Silico Sequence Selection Problem in De Novo Protein Design." *J. Comb. Optim.* (2005): in press. [6]- M. Oral and O. Kettani. "A Linearization Procedure for Quadratic and Cubic Mixed-Integer Problems." *Opns. Res.* 40 (1990): S109-S116. [7]- M. Oral and O. Kettani. "Reformulating Nonlinear Combinatorial Optimization Problems for Higher Computational Efficiency." *Eur. J. Oper. Res.* 58 (1992): 236-249. [8]- P.M. Pardalos and H.X. Huang and O. Prokopyev. "Multi-quadratic Binary Programming." University of Florida, Research Report (2004).