



Advances in De Novo Protein Design

Ch. A. Floudas, H. K. Fung, M. S. Taylor

published in

*From Computational Biophysics to Systems Biology (CBSB07),
Proceedings of the NIC Workshop 2007,*
Ulrich H. E. Hansmann, Jan Meinke, Sandipan Mohanty,
Olav Zimmermann (Editors),
John von Neumann Institute for Computing, Jülich,
NIC Series, Vol. 36, ISBN 978-3-9810843-2-0, pp. 9-14, 2007.

© 2007 by John von Neumann Institute for Computing

Permission to make digital or hard copies of portions of this work for personal or classroom use is granted provided that the copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise requires prior specific permission by the publisher mentioned above.

<http://www.fz-juelich.de/nic-series/volume36>

Advances in De Novo Protein Design

Christodoulos A. Floudas¹, Ho Ki Fung¹, and Martin S. Taylor²

¹ Department of Chemical Engineering,
Princeton University, Princeton, NJ 08540
E-mail: floudas@titan.princeton.edu

² School of Medicine,
Johns Hopkins University, Baltimore, MD 21205
E-mail: mstaylor@jhmi.edu

A new de novo protein design framework and its applications to the redesign of compstatin, human beta defensin-2, and the C-terminal analogs of Complement 3a is presented.

1 Introduction

De novo protein design searches for amino acid sequences that are compatible with a three-dimensional protein backbone template. Traditionally the backbone coordinates were treated as fixed in order to reduce the search space and make the design problem tractable. However, this is a highly questionable assumption as proteins are known to exhibit backbone flexibility. In de novo design, backbone flexibility was incorporated through either the consideration of multiple backbones with sequence search performed on each of them under the fixed template assumption, or the parameterization of backbone¹⁶. Recently we have developed a novel framework which performs de novo design on a truly flexible backbone template, which is defined by continuous C^α-C^α distances and dihedral angles between upper and lower bounds, through NMR structure refinement.

2 Our De Novo Protein Design Framework

Our two-stage de novo protein design framework not only selects and ranks amino acid sequences for a particular fold using a novel integer linear programming (ILP) model, but also validates the specificity to the fold for these sequences based on the full-atomistic forcefield AMBER¹. The two stages are outlined as below:

2.1 Stage One: In Silico Sequence Selection

The ILP model we use for sequence selection into a single template structure, which is the most computationally efficient one among 13 equivalent formulations we studied⁹, takes the form:

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

Set $i = 1, \dots, n$ defines the number of residue positions along the template. At each position i there can be a set of mutations represented by $j \in \{i\} = 1, \dots, m_i$, 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 (y_k^l) will be one if position i (k) is occupied by amino acid j (l), and zero otherwise. The composition constraints require that there is exactly one type of amino acid at each position. The pairwise energy interaction parameters E_{ik}^{jl} were empirically derived by solving a linear programming parameter estimation problem, which restricts the low energy high resolution decoys for a large training set of proteins to be ranked energetically less favorable than their native conformations².

Besides the basic model (1), we also developed a weighted average model and a binary distance bin model¹⁰ for de novo design based on a flexible template with multiple crystal or solution structures.

2.2 Stage Two: Approximate Method for Fold Validation

Driven by the full atomistic forcefield AMBER¹, simulated annealing calculations are performed for an ensemble of several hundred random structures generated for each sequence from stage one using CYANA 2.1^{3,4} within the upper and lower bounds on C^α-C^α distances and dihedral angles input by the user. This feature allows our framework to observe true backbone flexibility⁵. The TINKER package⁶ is subsequently used for local energy minimization of these conformers. A fold specificity factor is finally computed for each sequence using the following equation:

$$f_{\text{specificity}} = \frac{\sum_{i \in \text{new sequence conformers}} \exp[-\beta E_i]}{\sum_{i \in \text{native sequence conformers}} \exp[-\beta E_i]} \tag{2}$$

3 Case Studies

3.1 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 system and thus hinders complement activation. It is a novel drug candidate for treating inappropriate complement activation that has shown highly promising results in numerous pre-clinical trials conducted recently. The de novo design on compstatin is aimed at acquiring the sequences for the best

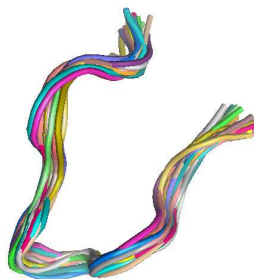


Figure 1. Flexible template of compstatin for de novo protein design as illustrated by the overlapping of its 21 NMR structures available from the Protein Data Bank.

inhibitors to C3. It was performed based on the flexible template of all 21 NMR structures available from the Protein Data Bank (Fig. 1).

As for the mutation set of the design, Cys² and Cys¹² were kept invariant to maintain the disulfide bridge in order to aid the formation of the hydrophobic cluster and prohibit the termini from drifting apart. The type I β -turn, Gln⁵-Asp⁶-Trp⁷-Gly⁸, was also fixed as it was not found to be a sufficient condition for activity. Val³ and Trp⁷ were not mutated either as they were found to interact directly with C3. For the varied positions, 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 cysteine and tryptophan. This mutation set leads to a problem with complexity 3.0×10^6 . Results for the design can be found in¹⁰.

3.2 Human Beta Defensin-2

Human Beta Defensin-2 (h β D-2) is a cysteine-rich 41-residue cationic peptide found in the human immune system. It belongs to the class of small, cationic peptides known as defensins. h β D-2 is crucial to innate immunity¹¹. It possesses antimicrobial property derived from the electrostatic force between the positive charge on the defensin molecule and the negative charge of the anionic head group of the microbe's membrane lipids. This electrostatic force disrupts the microbe's cell membrane and thus kills the cell¹¹.

Three different sets of flexible design templates were employed for the de novo design of h β D-2. The first one corresponds to the X-ray crystal structure elucidated by¹¹(PDB code: 1FD3) at a resolution of 1.35Å. The other two were generated using molecular dynamics simulation with generalized Born implicit solvation (Fig. 2) and molecular dynamics simulation with explicit water molecules (not shown).

In the design of h β D-2, SASA patterning was applied to restrict the sequence search space for the de novo design of h β D-2. The 41 positions of h β D-2 are classified into the core, surface, and intermediate categories which determine the mutation set for each position. This corresponds to the full-sequence design of the antimicrobial peptide with

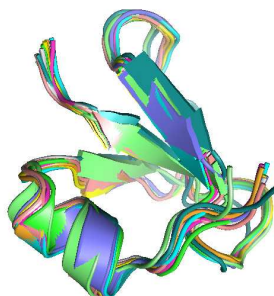


Figure 2. Overlay of the 10 structures of human beta-defensin-2 used for the flexible design template from MD simulations with the GB implicit solvation model. The structures are snapshots with 1 ns increment.

problem complexity of 6.40×10^{37} .

It should be noted that in the sequence selection stage, biological constraints, which were obtained from a homology search using PSI-BLAST, were imposed to ensure certain conserved properties among the sequence solutions. The constraints added cover charge characteristics, hydrophobic content, and amino acid occurrence frequencies of the human beta defensin-2 homologs. Results for the de novo design can be found in¹².

3.3 Complement 3a

Complement 3a (C3a) is a 77-residue small cationic peptide derived from the cleavage of the amino-terminus of the α -chain of complement component C3. It is a potent mediator which controls the pro-inflammatory activities of the complement system¹³. Having small molecular size and high potency, C3a proves to be a strong candidate as a superior therapeutic agent. Our de novo design aims at obtaining a potential peptide-drug candidate based on the C-terminal sequence of the C3a fragment of C3.

Like the design of human beta defensin-2, three different sets of flexible templates were employed. One corresponds to the single crystal structure elucidated by¹⁴, and the other two were generated using molecular dynamics simulations, one with the generalized Born implicit solvation model and the other with explicit water molecules (Fig. 3). The basic sequence selection model (1) was used for the single crystal structure template, whereas both the weighted average formulation and the binary distance bin formulation¹⁰ were employed for the flexible templates from molecular dynamics simulation.

Table 1 shows the mutation set of the de novo design.

Results of the de novo design are tabulated in¹⁵. Several 15-residue peptides were selected to be synthesized based on our predictions from the de novo design framework. The best sequence was experimentally validated to be close in performance to the superpotent peptide synthesized by¹³ in Ca^{2+} mobilization assay.



Figure 3. Overlay of the 10 structures of Complement 3a used for the flexible design template from MD simulations with the GB implicit solvation model.

Table 1. Mutation set of *in silico* sequence selection of C3a.

Positions	Native residue	Allowed mutations
63	L	A,I,L,M,F,Y,W,V
64	R	all except C and P
65	R	R,N,D,Q,E,G,H,K,S,T
66	Q	R,N,D,Q,E,G,H,K,S,T
67	H	R,N,D,Q,E,G,H,K,S,T
68	A	all except C and P
70	A	R,N,D,Q,E,G,H,K,S,T
71	S	R,N,D,Q,E,G,H,K,S,T
72	H	R,N,D,Q,E,G,H,K,S,T

4 Conclusions

In this paper, we presented the advances in our de novo protein design models, as well as our predictions on compstatin, human beta defensin-2, and C3a.

Acknowledgments

CAF gratefully acknowledges support from the National Science Foundation, the National Institutes of Health (R01 GM52032, R24 GM069736), and the US Environmental Protection Agency (GAD R 832721-010). This work has not been reviewed by and does not represent the opinions of USEPA.

References

1. W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, *A 2nd Generation*

- Force-Field For The Simulation Of Proteins, Nucleic-Acids, And Organic-Molecules*, J. Am. Chem. Soc. **117**, 5179–5197, 1995.
2. R. Rajgaria, S. R. McAllister, and C. A. Floudas, *A Novel High Resolution C^α-C^α Distance Dependent Force Field Based on a High Quality Decoy Set*, Proteins **65**, 726–741, 2006.
 3. P. Guntert, C. Mumenthaler, and K. Wuthrich, *Torsion Angle Dynamics for NMR Structure Calculation with the New Program DYANA*, J. Mol. Bio. **273**, 283–298, 1997.
 4. P. Guntert, *Automated NMR structure calculation with CYANA*. *Methods Mol Biol*, J. Mol. Bio. **278**, 353–378, 2004.
 5. C. A. Floudas, *Research Challenges, Opportunities and Synergism in Systems Engineering and Computational Biology*, AIChE Journal **51**, 1872–1884, 2005.
 6. J. Ponder, *TINKER, software tools for molecular design*. 1998. (Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine: St. Louis, MO, 1998).
 7. J. L. Klepeis, C. A. Floudas, D. Morikis, C. G. Tsokos, E. Argyropoulos, L. Spruce, and J. D. Lambris, *Integrated Structural, Computational and Experimental Approach for Lead Optimization: Design of Compstatin Variants with Improved Activity*, J. Am. Chem. Soc. **125**, 8422–8423, 2003.
 8. J. L. Klepeis, C. A. Floudas, D. Morikis, C. G. Tsokos, and J. D. Lambris, *Design of Peptide Analogs with Improved Activity using a Novel de novo Protein Design Approach*, Ind. Eng. Chem. Res. **43**, 3817–3826, 2004.
 9. H. K. Fung, S. Rao, C. A. Floudas, O. Prokopyev, 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. **10**, 41–60, 2005.
 10. H. K. Fung, M. S. Taylor, and C. A. Floudas, *Novel Formulations for the Sequence Selection Problem in De Novo Protein Design with Flexible Templates*, Optim. Methods & Software **22**, 51–71, 2007.
 11. D. Hoover, K. Rajashankar, R. Blumenthal, A. Puri, J. Oppenheim, O. Chertov, and J. Lubkowski, *The Structure of Human β-Defensin-2 Shows Evidence of Higher Order Oligomerization*, J. Biol. Chem. **275**, 32911–32918, 2000.
 12. H. K. Fung, C. A. Floudas, M. S. Taylor, L. Zhang, and D. Morikis, *Towards Full-Sequence De Novo Protein Design with Flexible Templates for Human Beta-Defensin-2*, Biophys. J., submitted for publication (2007).
 13. J. A. Ember, N. L. Johansen, and T. E. Hugli, *Designing Synthetic Superagonists of C3a Anaphylatoxin*, Biochemistry **30**, 3603–3612, 1991.
 14. R. Huber, H. Scholze, E. P. Paques, and J. Deisenhofer, *Crystal Structure Analysis and Molecular Model of Human C3a Anaphylatoxin*, Hoppe-Seylers Z Physiol Chemie **361**, 1389–1399, 1980.
 15. H. K. Fung, C. A. Floudas, M. S. Taylor, L. Zhang, and D. Morikis, *Redesigning Complement 3a based on Flexible Templates from both X-ray Crystallography and Molecular Dynamics Simulation*, in preparation (2007).
 16. H. K. Fung, and C. A. Floudas, *Computational De Novo Peptide and Protein Design: Rigid Template versus Flexible Templates*, Curr. Protein & Peptide Sci., submitted for publication (2007).