Video Article

Protein WISDOM: A Workbench for *In silico De novo* Design of BioMolecules

James Smadbeck[1], Meghan B. Peterson[1], George A. Khoury[1], Martin S. Taylor[1], Christodoulos A. Floudas[1]

[1]Department of Chemical and Biological Engineering, Princeton University

Correspondence to: Christodoulos A. Floudas at floudas@titan.princeton.edu

URL: https://www.jove.com/video/50476
DOI: doi:10.3791/50476

Keywords: Genetics, Issue 77, Molecular Biology, Bioengineering, Biochemistry, Biomedical Engineering, Chemical Engineering, Computational Biology, Genomics, Proteomics, Protein, Protein Binding, Computational Biology, Drug Design, optimization (mathematics), Amino Acids, Peptides, and Proteins, *De novo* protein and peptide design, Drug design, *In silico* sequence selection, Optimization, Fold specificity, Binding affinity, sequencing

Date Published: 7/25/2013

Citation: Smadbeck, J., Peterson, M.B., Khoury, G.A., Taylor, M.S., Floudas, C.A. Protein WISDOM: A Workbench for *In silico De novo* Design of BioMolecules. *J. Vis. Exp.* (77), e50476, doi:10.3791/50476 (2013).

Abstract

The aim of *de novo* protein design is to find the amino acid sequences that will fold into a desired 3-dimensional structure with improvements in specific properties, such as binding affinity, agonist or antagonist behavior, or stability, relative to the native sequence. Protein design lies at the center of current advances drug design and discovery. Not only does protein design provide predictions for potentially useful drug targets, but it also enhances our understanding of the protein folding process and protein-protein interactions. Experimental methods such as directed evolution have shown success in protein design. However, such methods are restricted by the limited sequence space that can be searched tractably. In contrast, computational design strategies allow for the screening of a much larger set of sequences covering a wide variety of properties and functionality. We have developed a range of computational *de novo* protein design methods capable of tackling several important areas of protein design. These include the design of monomeric proteins for increased stability and complexes for increased binding affinity.

To disseminate these methods for broader use we present Protein WISDOM (http://www.proteinwisdom.org), a tool that provides automated methods for a variety of protein design problems. Structural templates are submitted to initialize the design process. The first stage of design is an optimization sequence selection stage that aims at improving stability through minimization of potential energy in the sequence space. Selected sequences are then run through a fold specificity stage and a binding affinity stage. A rank-ordered list of the sequences for each step of the process, along with relevant designed structures, provides the user with a comprehensive quantitative assessment of the design. Here we provide the details of each design method, as well as several notable experimental successes attained through the use of the methods.

Video Link

The video component of this article can be found at https://www.jove.com/video/50476/

Introduction

De novo protein design is the identification of protein sequences that will yield a desired tertiary structure with improved properties or function. Since the native fold of a protein is the conformation which lies at the free energy minimum, *de novo* protein design seeks sequences that will have a free energy minimum in the target fold. This problem was first described by Drexler[1] and Pabo[2] and was referred to as the "inverse folding problem." However, unlike the protein folding problem, where a sequence can yield only one folded structure solution, the *de novo* protein design problem exhibits degeneracy. Many different amino acid sequences can yield the same tertiary structure and function.

While protein design has traditionally been performed experimentally through rational design and directed evolution, computational methods have more recently been employed to overcome the limited search space inherent in experimental methods. A variety of computational methods have been used, including deterministic methods, stochastic methods, and probabilistic methods.[3,4] Early computational methods used fixed-backbone templates to make the problem easier to solve.[5-7] With the advent of faster processors, high performance computing, and more efficient algorithms, backbone flexibility has been incorporated by using an ensemble of fixed-backbone templates[8-14] or by incorporating true backbone flexibility by expressing the template in terms of ranges of atom-to-atom distances and dihedral angles.[15,16]

This paper describes in detail Protein WISDOM, an online tool that has been made available to the academic community to utilize our computational *de novo* protein design framework. This framework has been applied to the design of numerous proteins, for therapeutic use targeting diseases such as HIV, cancer, complement diseases, and other autoimmune disorders. Many of the predicted peptides were experimentally validated, demonstrating the power of the method. **Table 1** provides a summary of the different proteins that have been designed including the size of the protein or peptide, the number of predictions, and experimental validation.

Protein Design	Protein Length	# of Computational Predictions	# of Experimental Validations	Reference

Full sequence design of human beta-defensin-2	41	340		(17)
Compstatin inhibitors of human C3	13	28	3/3	(18, 19)
Compstatin analogues that bind to rat C3c	13	5		(20)
Compstatin analogues with di-serine extension	15	8		
Stabilizing structure of compstatin analog W4A9	13	18		
C3a receptor agonists and antagonists	77	20	4/7	(21)
C5a receptor agonists and antagonists	74	61	2/61	
HIV-1 gp14 inhibitors	12	6	4/5	(22)
HIV-1 gp120 inhibitors	9	14		
Bak inhibitors of Bcl-x L and Bcl-2	16-18	10	5/5	(23)
Inhibitors of ERK2	11	25		
Inhibitors of EZH2	21	17	10/10	(24)
Inhibitors of LSD1 and LSD2	16	41	17/20	
Inhibitors of HLA-DR1	13	6		(25)
Inhibitors of PNP	5	13		

Table 1. Summary of designed proteins and peptides using the *de novo* protein design framework. The # of computational predictions is presented as the number of favorable predictions (*i.e.* fold specificities above a certain cutoff or approximate binding affinities greater than the native sequence). The # of experimental validations gives two numbers: the first is the number of predictions that were experimentally validated while the second is the total number of predictions that were tested experimentally.

Design of human-beta-defensin-2 (hβD-2) was performed to enhance the peptide's antimicrobial property.[17] For this design, we considered two cases: 1) up to 10 mutations along hβD-2 and 2) full sequence design of all hβD-2 residue positions except the Cysteines (8, 15, 20, 30, 37, and 38). Three different design templates and three different sequence selection models were utilized in the design. High levels of similarity in mutations were observed between the weighted average and distance bin models for both the 10 mutation design and the full sequence design. Additionally, a large number of sequences were found to have more favorable calculated Fold Specificity values than the native sequence.

Complement system inhibitors (of C3, C3a, and C5a) were designed to combat a number of immune diseases such as stroke, heart attack, Alzheimer's disease, asthma, rheumatoid arthritis, rejection of xenotransplantation, adult respiratory disease, psoriasis, and Crohn's disease. Three compstatin inhibitors of C3c predicted by the protein design framework plus three rationally designed sequences were experimentally validated to be better binders than the native compstatin.[18,19]

Further studies examined the loss of activity of compstatin against non-primate C3c and designed a number of candidate rat and mouse C3c inhibitors. Five sequences were shown to have more favorable association free energies with rat C3c than the W4A9 compstatin mutant known to inhibit C3c. This is due to a new salt bridge formation by Arg1.[20] Eight sequences with an N-terminal extension were predicted to be better binders than W4A9 with a di-Serine extension. Finally, 18 compstatin sequences were predicted to stabilize the bound conformation of W4A9, providing strong candidates for primate and non-primate C3c inhibitors.

In addition to C3c inhibitors, C3a and C5a receptor agonists and antagonists were designed based upon the structures of C3a and C5a. Seven C3a sequences predicted by the model were experimentally tested. Two of the sequences were potent agonists while two others were partial agonists.[21] The two potent agonists showed a 58-fold improvement over a previously discovered "superagonist". The design of C5a receptor agonists and antagonists provided a set of 61 sequences. All the sequences were synthesized and two were found to be novel C5a agonists.

Fusion inhibitors of HIV-1, the virus that causes AIDS, were designed to prevent HIV-1 from infecting cells. The first design targeted gp41, an envelope glycoprotein of HIV-1. The protein design framework predicted six sequences that were better binders than the native sequence. Four of these predicted sequences were experimentally validated to inhibit HIV-1 with the best sequence having an IC_{50} as low as 29 μM. This sequence showed a 3-15 fold improvement over the native sequence and had no loss of activity against an Enfuvirtide-resistant virus strain.[22] The second design targeted gp120, another envelope glycoprotein of HIV-1. Fourteen sequences were predicted to be binders of gp120 and provide additional potential fusion inhibitors of HIV-1.

Numerous proteins linked to cancer provided promising targets for cancer therapeutics. Bcl-2 and Bcl-x_L are anti-apoptotic proteins that prevent cell death. Inhibitors of these two proteins were designed to induce cell death in cancer cells. Ten sequences were predicted to be better binders than the native, and these results captured previous experimental and mutagenesis results.[23] Another target protein, ERK2, is involved in

signal-transduction cascades that make it a promising target for antiproliferative cancer therapies. Twenty-five sequences were predicted to be inhibitors of ERK2.

Histone methyltransferases and demethylases dynamically control histone methylation, which has been linked to many cancer types including prostate, breast, lymphoma, myeloma, bladder, colon, skin, liver, endometrial, lung, and gastric. The *de novo* protein design framework identified 17 inhibitors of EZH2 (a Lysine methyltransferase) and of the ten experimentally tested, all were found to inhibit EZH2.[24] The most potent peptide had an IC_{50} of about 13 μM, was equally effective with elevated enzyme concentrations, and did not compete with the cofactor. These peptides were the first set of inhibitors of EZH2. 53 inhibitors of LSD1 (a demethylase) were predicted by the framework and of the 20 experimentally tested, 17 were inhibitors of LSD1 and 18 were inhibitors of LSD2. The best inhibitors had IC_{50} values below 1 μM, making them the most potent peptidic inhibitors discovered to date.

The final two protein systems provided targets for treating various autoimmune diseases such as Coeliac disease, diabetes mellitus type 1, systemic lupus erythematosus, Sjögren's syndrome, Churg-Strauss Syndrome, Hashimoto's thyroiditis, Graves' disease, idiopathic thrombocytopenic purpura, rheumatoid arthritis, and allergies. None of these potential inhibitors have been experimentally validated, however the framework predicted six sequences that bind to HLA-DR1 and 13 sequences that bind to PNP.

Table 2 summarizes experimentally validated inhibitors and agonists predicted using the *de novo* protein design framework. The approximate binding affinity metric was used to predict nine of the sequences (inhibitors of human C3c, HIV-1 gp41, EZH2, LSD1, and LSD2), while the fold specificity metric was used to identify four of the sequences (agonists/antagonists of C3aR). These peptides highlight the success of the *de novo* protein design framework, particularly the added approximate binding affinity metric. The framework is extremely versatile in its applicability. Six different proteins linked to twenty-five different diseases have been successfully designed and experimentally validated.

Name	IC50	EC50	Protein Target	Applicable Diseases
SQ027	0.94 μM		human C3c	stroke, heart attack, Alzheimer's disease, asthma, rheumatoid arthritis, systemic lupus erythematosus, multiple sclerosis, psoriasis, diabetes type I, Crohn's disease, pancreatitis, and cystic fibrosis
SQ086	1.98 μM		human C3c	
SQ059	4.73 μM		human C3c	
SQ110-4		15.2 nM	C3aR	
SQ060-4		36.4 nM	C3aR	
SQ007-5	15.4 nM		C3aR	
SQ002-5	26.1 nM		C3aR	
SQ435	29 - 253 μM		HIV-1 gp41	AIDS
SQ037	13.57 μM		EZH2	prostate, breast, lymphoma, myeloma, bladder, colon, skin, liver, endometrial, lung, and gastric cancers
SQ011-1	0.521 μM		LSD1	
SQ016-1	0.249 μM		LSD1	
SQ026-1	2.51 μM		LSD2	
SQ015-1	1.332 μM		LSD2	

Table 2. Computationally predicted and experimentally validated peptides targeting various diseases.

Protocol

Method Overview

The *de novo* design framework used in Protein WISDOM consists of two stages. The first stage produces a rank-ordered list of amino acid sequences that will fold into a given template structure. The second stage validated these sequences by calculating either fold specificity or approximate binding affinity, or both. The former is primarily used when the design is of a single protein, while the latter is used when the design is of a complex (a peptide binding to a target protein). **Figure 1** gives an overview of the steps involved in the framework.

Design Inputs: A number of inputs need to be defined for the *de novo* protein design framework. The first is the design template. This is a 3-dimensional (3D) protein structure that contains coordinates for all the atoms in the protein. The structure can be rigid or flexible. Rigid templates are a set of fixed atom coordinates and are obtained from x-ray crystallography structures. Flexible templates can be a set of fixed atom coordinates or upper and lower bounds on the atom coordinates. These templates can be obtained from NMR solution structures, molecular dynamics, or docking simulations.

The design template is used to generate the allowed mutation set of the designed protein. This set defines which positions of the sequence can mutate and to what amino acids. The mutation set is generated by calculating the solvent accessible surface area (SASA) of each residue in the design template. If the residue is more than 50% exposed to solvent, a set of hydrophilic amino acids is allowed (D, E, G, H, K, N, P, Q, R, S, T). If the residue is less than 20% exposed to solvent, a set of hydrophobic amino acids is allowed (A, F, I, L, M, V, W, Y). If the residue's exposure is in between 20% and 50%, all amino acids are allowed. Cysteine is typically excluded from the mutation set unless experimental or literature data deem it appropriate. The small amino acids (A, G, T) are typically included in all mutation sets. When available, experimental or literature insights can be used to manually modify the mutation sets of particular amino acid positions.

A forcefield is chosen to calculate the pairwise interaction energy of the sequences in the design template. While any forcefield can be adapted to be used within the framework, two distance-dependent forcefields have been developed and are used extensively in the *de novo* design framework. The first is a high resolution C^{α}-C^{α} forcefield,[26] where the distances are between the C^{α} carbons of the residues. The second is a high resolution centroid-centroid forcefield[27] where the distances are between the centroids of the residues. The energy parameters in the forcefields were derived by solving a linear programming parameter estimation problem which required the low-energy high-resolution decoys for a large training set of proteins to be energetically less favorable than their native conformations. The high-resolution centroid-centroid forcefield and the C^{α}-C^{α} forcefield were both tested and validated in previous studies on human beta-defensin-2.[17] True backbone flexibility is incorporated into the model by discretizing the forcefields into distance bins. The distance between a pair of amino acids will correspond to a distance bin giving the same energy value to a range of distances. This enables the sequence selection optimization model to account for backbone movement.

Biological constraints, in the form of charge constraints or content constraints, can be included manually by the user as an additional design input. Charge constraints specify a particular charge or range of charges that must be satisfied for the designed sequence or a portion of the designed sequence. The charge is calculated as the sum of the positively charged residues (K and R) minus the sum of the negatively charged residues (D and E). Content constraints specify upper and lower bounds on the occurrence of a particular amino acid in the sequence. Biological Constraints are generally defined through an extensive sequence alignment to the native sequence. This is to capture the known biological limits on charge and amino acid content represented in nature for a family of proteins. Further constraints are manually defined through analysis of known experimental data.

Stage One: Sequence Selection: The original sequence selection method was first developed by Klepeis *et al.*[15,16] It selects and ranks amino acid sequences according to their energies in the design template using an Integer Linear Optimization (ILP) model. The method was later improved by the use of a more computationally efficient sequence selection model for rigid (single) templates and expanded through the development of models for flexible templates. This global optimization method does not rely on random mutations and is theoretically guaranteed to search the complete sequence space and determine a global solution. This is a major advantage of our approach compared to all other existing approaches.

Single Structure Model: The original form of the sequence selection model proposed by Klepeis *et al.*[15,16] was further refined by Fung *et al.*[28] Its final form is given in Eq. 1.

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

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

$$\sum_{j=1}^{m_i} w_{ik}^{jl} = y_k^l \;\;\forall\; i,k>i,l$$

$$\sum_{l=1}^{m_k} w_{ik}^{jl} = y_i^j \;\;\forall\; i,k>i,j$$

$$y_i^j, y_k^l, w_{ik}^{jl} \in \{0,1\} \;\forall\; i,j,k>i,l$$

Set i=1,...,n defines the residue positions in the design template. At each position i, mutations are represented by $j\{i\}$=1,...,m_i, where m_i=20 if position i is allowed to mutate to any of the twenty natural amino acids. The alias sets $k \equiv i$ and $l \equiv j$, with $k>i$, are employed to represent all unique pairwise interactions. Binary variables y_i^j and y_k^l are introduced to model amino acid mutations. The y_i^j variable will assume the value of one if the model assigns amino acid j to position i, and the value of zero otherwise (similarly for y_k^l). The objective function represents the sum of all pairwise energy interactions in the design template. Parameter E_{ik}^{jl} which is the energy interaction between position i occupied by amino acid j and position i occupied by amino acid l , depends on the distance between the α-carbons or side chain centroids at the two positions (x_i,x_j) as well as the type of amino acids j and l . It only contributes to the objective function if both y_i^j and y_k^l are equal to one.

Fung *et al.*[28] found that formulation (1) is significantly more computationally efficient than twelve other equivalent quadratic assignment-like models for sequence selection.[28,29] In particular, it outperformed the original model proposed by Klepeis *et al.*[15,16] on two sequence selection problems for human beta-defensin-2: one at a complexity level of 3.4×10^{45} and the other at 6.4×10^{37} with 49 additional linear biological constraints. The original model proposed by Klepeis *et al.*[15,16] was found to take 53,263 central processing unit (CPU) sec and 4,578 CPU sec respectively to solve the two problems to global optimality using CPLEX 9.0[30] on a Pentium IV 3.2 GHz processor. Formulation (1) only took 649 CPU sec and 14 CPU sec to perform the same tasks, corresponding to an 82-fold and 327-fold improvement in computational efficiency.

Weighted Average Model: Fung *et al.*[28] developed two models to handle the typical case of *de novo* protein design in which the design template is flexible, containing a set of structures. The Weighted Average Model uses a weighted average energy,

$\sum_{d=1}^{b_m} E_{ik}^{jl}(x_i,x_k)\,wt(x_i,x_k,d)$, in place of the energy parameter $E_{ik}^{jl}(x_i,x_k)$ in the Single Structure Model (Eq. 1). The weights $wt(x_i,x_k,d)$ are determined by the frequencies of the distance between x_i and x_k falling into distance bin d in the template structures. The final form of the Weighted Average Model is given in Eq. 2.

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

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

$$\sum_{j=1}^{m_i} w_{ik}^{jl} = y_k^l \;\forall\, i,k>i,l$$

$$\sum_{l=1}^{m_k} w_{ik}^{jl} = y_i^j \;\forall\, i,k>i,j$$

$$\sum_{d:disbin(x_i,x_k,d)=1}^{b_m} b_{ikd} = 1 \;\forall\, i,k>i$$

$$y_i^j, y_k^l, w_{ik}^{jl}, b_{ikd} \in \{0,1\} \;\forall\, i,j,k>i,l,d$$

Distance Bin Model: The second sequence selection model for flexible template structures incorporates the distance information from the multiple structures by introducing a binary variable b_{ikd}. This variable equals one if the distance between x_i and x_k falls into distance bin d , and is zero otherwise. Another parameter introduced, $disbin(x_i, x_k, d)$, equals one if the distance between x_i and x_k in any of the template structures falls into distance bin d and is zero otherwise. Since only one distance bin per amino acid pair will contribute to the total energy, E_{ik}^{jl} in the objective function is replaced with $\sum_{d:disbin(x_i,x_k,d)=1}^{b_m} E_{ik}^{jl}(x_i,x_k)\,b_{ikd}$. This, however, introduces nonlinearity into the objective function.

Further details on linearizing the model and additional constraints that need to be added for feasibility can be found in Fung *et al.*[28] The Distance Bin Model is given in Eq. 3.

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

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

$$\sum_{j=1}^{m_i} w_{ik}^{jl} = y_k^l \;\forall\, i,k>i,l$$

$$\sum_{l=1}^{m_k} w_{ik}^{jl} = y_i^j \;\forall\, i,k>i,j$$

$$\sum_{d:disbin(x_i,x_k,d)=1}^{b_m} b_{ikd} = 1 \;\forall\, i,k>i$$

$$y_i^j, y_k^l, w_{ik}^{jl}, b_{ikd} \in \{0,1\} \;\forall\, i,j,k>i,l,d$$

Any of the above formulated Integer Linear Programming (ILP) problems[15-17] can be solved rigorously using branch-and-bound techniques.[28-30] Such techniques guarantee consistent and reliable convergence to the global minimum energy sequence.

Stage Two: Validation: **Figure 2** provides a detailed overview of the two Stage Two approaches. The figure shows the steps required to calculate the final ranking metric and the number of structures generated in each step.

Fold Specificity: Fold specificity is a metric used for ranking preliminary designs derived in Stage One. The aim of the calculation is to find how well each sequence folds into the template structure relative to the original sequence of the template, based on energy calculations. There are two approaches for how to do this, each with different computational demands.

The first approach was implemented by Klepeis *et al.*[15,16] This approach utilizes the protein structure prediction framework ASTRO-FOLD,[26,27,31-47] which is based on deterministic global optimization. This approach is not currently used in the implementation of Protein WISDOM since it is very computationally demanding. Recognizing computational resource limitations and the need to perform this calculation on potentially hundreds to thousands of sequences in design, Fung *et al.*[17] proposed a more efficient approach using TINKER/CYANA.[48-50] The approach involves defining a flexible template of the structure. The flexible template can be defined using upper and lower bounds on the distances between C^{α} atoms, as well as the ϕ and ψ angles of the residues. For a single structure, the initial distances and dihedral angles are used and bounds are defined either as a fixed distance or a percentage. The default bounds are ±10% for C^{α} distances or ±10° for dihedral angle bounds. For a flexible template, bounds can be obtained from the maximum and minimum values seen across all template structures given as input to design. Once initial bounds are defined for each sequence, ensembles containing hundreds of conformers are generated using CYANA 2.1.[48,49]The conformers are generated using a torsion angle dynamics simulated annealing protocol in CYANA that heats the protein rapidly and slowly cools it, tracking the conformations sampled. After the simulated annealing, a local energy minimization is performed that minimizes the clashes from Van der Waals radii overlapping, as well as violations in the distance and angle constraints. By default, 500 final structures are generated. Each structure in the ensemble for each sequence is subjected to a local minimization in TINKER 3.6,[50] using the AMBER forcefield.[51] The final potential energy of each minimized structure is tabulated. This overall approach is performed for the starting sequence as well as each candidate mutant sequence. Then, the Fold Specificity of each mutant sequence to the target fold can be calculated relative to the native sequence using the following Boltzmann distribution (Eq. 4).

$$f_{spec} = \frac{\sum_{i \in novel} e^{-\beta E_i}}{\sum_{i \in native} e^{-\beta E_i}}$$

Approximate Binding Affinity: The approximate binding affinity calculation method is used to rank the designed sequences that are in complex with a target protein. These calculations can be done on the sequences directly from Stage One or can be performed on the high fold specificity sequences obtained from the fold specificity step.

Lilien *et al.*[52] proposed an approach for the calculation of approximate binding affinities of protein-ligand complexes. It is based on generating rotamerically-based ensembles of the protein, the ligand, and the protein-ligand complex and using those ensembles to calculate partition functions. This approximate binding affinity is denoted as *K** and is defined by Eq. 5.

Here q_{PL} is the partition function of the protein-ligand complex, q_b is the partition function of the free protein, and q_L is the partition function of the free ligand. The partition functions are defined in Eq. 6, where the sets *B, F* and *L* contain the rotamerically-based conformations of the bound protein-ligand complex, the free protein, and the free ligand, respectively. E_n is the energy of conformation *n, R* is the gas constant, and *T* is the temperature.

$$q_{PL} = \sum_{b \in B} e^{-\frac{E_b}{RT}}, \; q_{PL} = \sum_{b \in B} e^{-\frac{E_b}{RT}}, \; q_{PL} = \sum_{b \in B} e^{-\frac{E_b}{RT}}$$

Structure Prediction: In order to begin calculating *K**, a 3D structure of each sequence is needed. This is done using the Rosetta AbRelax function,[53-55] part of the Rosetta 3.4 software package. The strategy behind the AbRelax algorithm is based upon experimental observation that the local structure of the protein is influenced but not uniquely determined by the local sequence of the protein. A Monte Carlo algorithm is used to replace local protein structures with sequence derived structural fragments. This method produces the final compact protein structures that account for non-local interactions such as buried hydrophobic residues, paired β strands, and specific side chain interactions.

Clustering: The structures from AbRelax are then clustered based upon their φ and ψ angles using OREO.[56,57] This clustering method elucidates representative backbone structures of the entire structural ensemble. The average structures from the ten largest clusters and the overall lowest energy structure are chosen for docking to the target protein. This provides 11 unique backbone structures for each peptide sequence, incorporating backbone flexibility into the ensemble generation.

Docking Prediction: Docking prediction is done using RosettaDock.[58-60] For each sequence, each of the 11 peptide backbone structures is docked against the target protein. In this case, since the binding site is known, the peptides are placed near the binding site and allowed to translate 3 Å normal to the binding site, 8 Å parallel to the binding site, and rotate 8°. RosettaDock uses a Monte Carlo algorithm for low and high resolution docking movements. Each docking run generates a large ensemble of complex structures. The ten lowest energy complexes in each of the 11 runs are used as starting structures in the final rotamerically-based conformation ensemble generation (110 starting structures per sequence).

Final Ensemble Generation: RosettaDesign[61] is used to generate the final rotamerically-based conformation ensemble because it can be used to generate a number of structures by only adjusting the rotamers on the side chains through the fixbb function. RosettaDesign is given a number of starting structures, and for each structure, a residue is randomly chosen and the rotamer changed through a Monte Carlo algorithm. This

is repeated until thousands of rotamer substitutions are attempted and gives a final low-energy conformation that will contribute highly to the partition function.

To generate the peptide ensemble, the ten lowest-energy peptide structures from each of the ten largest clusters plus the ten overall lowest-energy peptide structures are used as starting structures for RosettaDesign (110 total starting structures). For each starting structure, 200 rotamer conformers are generated, giving a final ensemble of 22,000 structures (set *L* in Eq. 6). The ensemble incorporates both backbone flexibility and rotamer flexibility.

The complex ensemble is generated similarly by taking the 110 starting structures from the docking prediction step and generating 200 rotamer conformers per starting structure. The final ensemble size is 22,000 structures (set *B* in Eq. 6). Flexibility is taken into account by the various peptide backbone structures used, the various docked conformations, and the rotamer conformers for each starting structure.

The protein ensemble is generated by running RosettaDesign on just the target protein structure. In this case, 2,000 rotamer conformations are generated for the single starting structure, so the final ensemble size is 2,000 structures (set *F* in Eq. 6).

Protein WISDOM

Protein WISDOM, which stands for Protein Workbench for *In Silico De novo* design Of bioMolecules, is an online tool that gives the academic community access to our *de novo* protein design framework in a user-friendly way. It can handle several commonly encountered design objectives, from designing single protein chains to adopt a template fold to designing novel peptides that will bind to a target protein. The next two sections describe the capabilities of Protein WISDOM with regards to the two main types of protein design problems encountered. The first type applies sequence selection to select novel sequences that are favorable in the given design template and then uses fold specificity to validate the novel sequences. The second type uses sequence selection to select novel sequences of a peptide bound in a complex and then uses both fold specificity and approximate binding affinity calculations to validate the novel sequences.

User Registration

Visit the Protein WISDOM web page at http://www.proteinwisdom.org.

Click the User Login button on the top right of the page. Click the "Click here" to register.

Fill out information related to email address and requested username and click continue.

Fill out additional information on name, institution, group, address. Click the checkbox to agree to terms of use. Click the "Submit Registration" button.

Stage One: Sequence Selection

Submission of Protein Sequence and Template Structure(s)

Click on the User Login button to begin the protein design experiment. The user is presented with their "User Homepage" (**Figure 3**) which lists the number of jobs they have submitted, the number of structures (templates) they have uploaded, and a list of the structures they have uploaded so far.

Start a new design job by clicking "Create New Job." The user is taken to the "Job Submission" page (**Figure 4**). Give the job a name, and indicate if it is based on a previous job (*i.e.* the same design template, mutation sets, and biological constrains can be imported into a new job, however the user will have the ability to modify the mutation sets and biological constraints). Click "continue."

Upload the protein structure(s) of the design template (**Figure 5**). This template must be in standard protein data bank (PDB) format. It can be a rigid template (one set of coordinates for every atom) or a flexible template (multiple models, such as obtained from NMR solution structures). For the case of designing a single protein, there can only be one chain in the template. A user can upload a new template or select from existing templates they have previously uploaded. Optionally indicate the pdb ID of the template, if available. If multiple templates are uploaded, be sure each model begins with "MODEL #" and ends with "ENDMDL." Ensure every residue is designated by a natural amino acid. Click "Continue."

Upon successful upload of the template, Protein WISDOM will display the number of residues, chains, and models it found in the template, list the sequence, and ask the user to verify the template. Confirm the template structure if it has been correctly inputted, and click "Continue."

Once the template has been successfully uploaded and confirmed, the user is taken to the "Main Control Page" (**Figure 6**). On this page, the user can view the job status, modify the mutation sets and biological constraints, and submit the job for Stage One: Sequence Selection. At this point, since Stage One has not completed, there are no options for Stage Two. Those appear once results from Stage One are available.

Selection of Mutation Sets

Click on the "Mutation Sets" link on the "Main Control Page" to define mutation sets.

Select which residues will be allowed to mutate, and select which amino acids they are allowed to mutate to (**Figure 7**). By default, the allowable amino acids at any given position are selected based upon Solvent Accessible Surface Area (SASA). Mutation sets are required.

Click "Save Changes" after mutation sets are selected. The user can choose to continue editing the mutation set. When finished editing the mutation set, click to return back to the "Main Control Page."

Selection of Biological Constraints

Click on the "Biological Constraints" link on the "Main Control Page" to define biological constraints.

Specify charge or amino acid content constraints across the whole protein or a portion of the protein (**Figure 8**).

Limit the total number of mutations allowed to occur, if required. Biological constraints are optional. Click to return to the "Main Control Page" when finished.

Submission of Stage One: Sequence Selection

Click on the "Begin Stage 1" link to bring user to "Submit Stage 1" page.

Select the chain to design (**Figure 9**), the number of sequences to generate, the distance-dependent forcefield, and the model. If a complex is being design and a Fold Specificity calculation is desired, one must choose only a single chain to design. If the uploaded template was a single structure, or a "rigid template," only the Single Structure model is allowed. If the uploaded template is flexible, the user has the option to select from all three models: Single Structure, Weighted Average, and Distance Bin. Take note of the computational complexity of the optimization to be solved. There is an upper limit of 20^{25} for computational complexity allowed.

Submit the job. The user is redirected back to the "Main Control Page" (**Figure 10**). The Job Status will be updated to indicate the current progress of the job. The job will become locked for editing after submission.

Upon completion of the job, the user receives an email with the results, which consist of a list of designed sequences. The results are also viewable on the "Main Control Page." A box for Stage 2: Fold Specificity appears on the page to enable the user to perform this validation.

Stage Two: Fold Specificity Calculations

Fold Specificity Submission

Click "Begin Stage 2: Fold Specificity" to enter the "Build Stage 2" page. Define the upper and lower C^{α}-C^{α} distance bounds by specifying the Template flexibility factor either as a percentage of distance, or as a fixed distance. Define upper and lower angle bounds on the φ and ψ dihedral angles by specifying the Template flexibility factor as a percentage. Note that when using a flexible template, the upper and lower distance bounds are taken as the lowest and highest distance values across all the template models. Likewise, upper and lower angle bounds are taken from the highest and lowest angle values across all the models.

Click the "Submit" button.

Specify the number of structures per sequence to generate and click "Continue." Note there is an upper bound of 500 structures per sequence to generate.

Click "Continue" to confirm intent to submit for fold validation. Stage One and Stage Two are locked for editing until the completion of Stage Two.

Upon completion of the job, an email is sent to the user with the results. View the results on Protein WISDOM on the "Main Control Page" (**Figure 11**). Here the text files containing designed sequences, corresponding energy values from Stage One and fold specificity values from Stage Two can be viewed and downloaded. In addition, the user may click the "View Results" link which displays a table in the browser with Stage One ranks and energy values as well as Stage Two ranks and fold specificity values.

Stage Three: Approximate Binding Affinity Calculations for Protein-peptide Complexes

Approximate Binding Affinity calculations calculate the affinity of the designed ligand protein/peptide to the rest of the complex. These calculations can be performed directly after Stage One, or after Fold Specificity calculations have been completed.

Click on "Sequence #" to select the sequence to begin approximate binding affinity calculation. User will be directed to the "Select Sequence" page, which presents a list of the designed sequences along with their sequence selection and fold specificity ranks. Only one sequence can be selected at a time for approximate binding affinity calculation, as the calculations are very computationally demanding. Upon completion of a sequence, the user may select another sequence to have the approximate binding affinity calculated, and this result is added to the previous result, displaying the approximate binding affinity for all completed sequences. Once a sequence is selected and saved, the user is redirected to the "Main Control Page."

Click "Begin Stage 2: Approximate Binding Affinity" to submit the job. Upon completion, results are emailed to the user, which include an attachment containing the sequence number, approximate binding affinity, and values of the partition functions in Eq. 6. For every subsequent approximate binding affinity job, this file contains the results for all the completed sequences. Full results (from sequence selection, fold specificity, and approximate binding affinity) can also be viewed by accessing the "Main Control Page" for the job (**Figure 12**).

Representative Results

***De Novo* Design of Entry Inhibitors for HIV-1**

The *de novo* design framework implemented in Protein WISDOM has been used for the design of inhibitor peptides for several important therapeutic systems (**Tables 1 and 2**). One system of note is the design of peptides to inhibit HIV-1 entry to the host cell receptor CD4, which is here used as a representative system to demonstrate the practical use of the Protein WISDOM interface. The peptides were designed to target the transmembrane subunit gp41, which functions as a key part in the fusion and entry of HIV-1 to the host T helper cells. Note that results will

not necessarily be identical to those presented in the original publication. This is due to the stochastic nature of the Rosetta methods used in this part of the method and the update from Rosetta2.3 to Rosetta3.4 since the original publication.

To initialize the job, the user provides a valid protein design template. This will either be a single protein structure for fold design or a complex for binding design. The design template for entry inhibitors of HIV-1 is the crystal structure of C14linkmid, a 14-residue crosslinked peptide, in complex with the gp41 core, PDB:1GZL (**Figure 13**). This template peptide is a known, potent inhibitor and is submitted with the cross-linker removed.

Once an input structure template is verified, the user is taken to the job home page (**Figure 8**). Here further design constraints can be set and the Stage One Sequence Selection can be started. Entering the Mutation Sets section (**Figure 7**) constraints can be set for each position in this system based on the Solvent Accessible Surface Area of the residue position in the design template. In the case of HIV-1, we allow positions 628, 631, 635, 638 to mutate to hydrophobic amino acids (A,L,I,M,F,W,Y,V), positions 630, 632, 634, 637, and 639 to mutate to hydrophilic amino acids excluding Proline, positions 633 and 636 to mutate to hydrophilic amino acids excluding Proline and allowing for Cysteine, and position 629 to mutate to hydrophobic amino acids plus Cysteine. The choice to disallow Proline in all positions was due to the possibility that the Proline could disrupt the helical structure of the C14linkmid target peptide. The choices to allow or disallow Cysteine mutations were based on sequence alignments.

Entering the Biological Constraints section of Protein WISDOM (**Figure 8**), the user can specify charge and amino acid content constraints for all or portions of the protein design template. In the case of the original HIV-1 design, charge was limited to ± 1 from the native charge of -4 for the section of the designed peptide that was exposed to solvent when bound: positions 629, 630, 632-634, and 636. From sequence alignment analysis, all amino acids types were limited to ≤ 3 present in the full peptide sequence.

With all the necessary design inputs defined, the system can be submitted for Stage One: Sequence Selection. The number of minimum energy protein sequences identified by the method was set to 500 and the force field used was the 6-bin centroid-centroid forcefield.[27] The sorted Sequence Selection results can be accessed either through the "View Results" (**Figure 14**), the "Sequence Results" (**Figure 15**), or the "Energy Results" (**Figure 16**) links on the Main Job Page. The "View Results" section gives an easily readable summary of all the results calculated for a given system, which can be sorted by any of the Stage results for quick analysis. The "Sequence Results" section gives a summary of selected sequences in three-letter amino acid code. The "Energy Results" gives a summary of the optimization model run with the selected sequences, their energies, and the time it took to solve the model for the solution.

Once the Stage One Sequence Selection has completed, the user will be allowed to submit to the Stage Two Fold Specificity and Approximate Binding Affinity Methods. For HIV-1, all 500 sequences were submitted to the Stage Two Fold Specificity Method. For this method, the user has the option to define the flexibility of the distance and angle bounds for template structure production, as well as the number of structures to produce for each mutated sequence. The Fold Specificity results can be accessed and sorted in the summary "View Results" section (**Figure 17**) or individually in the "Fold Specificity Results" section (**Figure 18**). In the "Fold Specificity Results section the sequence number and Fold Specificity values are provided as an array.

If the user is designing a complex, the option to submit for Approximate Binding Affinity Calculation is allowed. Due to the computational resources necessary for the Binding Affinity Calculation, only one sequence can be selected for calculation at any one time. In order to demonstrate the final results of the method, the Native and SQ435 sequences were selected for Binding Affinity Calculation run.

The results of the sample Approximate Binding Affinity Calculation are presented as a sortable list in the "View Results" section (**Figure 19**). All sequences that have been run for the Approximate Binding Affinity Calculation have a highlighted link in the "View" column. The "View" link takes the user to a "Design Information" page (**Figure 20**). This page provides downloadable zip files for all the complex and peptide structures used in the final structure Design step. For both the Complex and Peptide structures, the top 10 lowest energy structures are provided in an rank-ordered list. Each structure has a "View" link which allows the user to view the structure in an interactive Jmol environment[62] (**Figure 21**). A "Download" link is also provided to allow the user to download each structure individually. Further details of the results are presented in the "Approximate Binding Affinity Results" section (**Figure 22**). This section provides the values for the peptide, protein, and complex partition functions along with the final Approximate Binding Affinity Value.

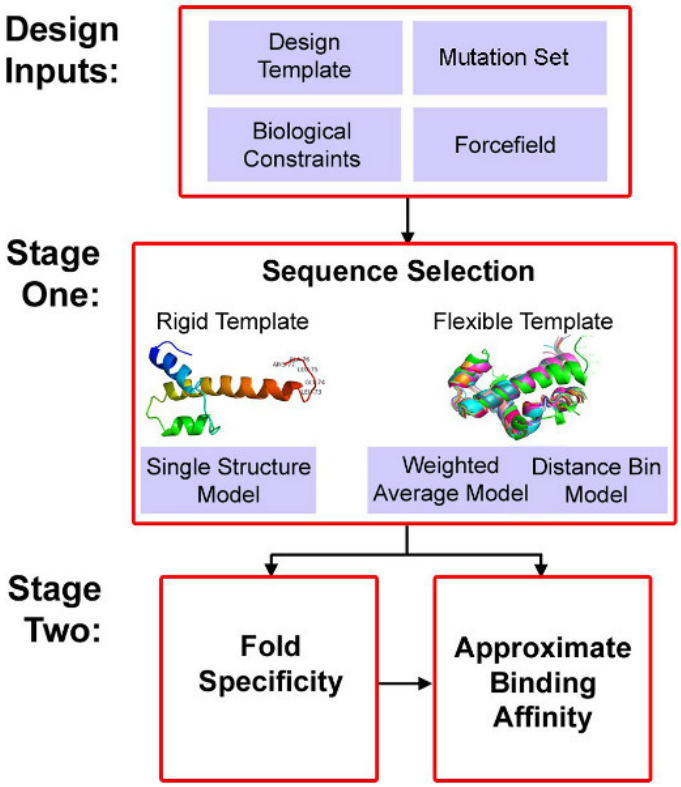


Figure 1. Overview of the *de novo* protein design framework. The *de novo* design method used in Protein WISDOM involves three steps: design inputs, stage one sequence selection, and stage two fold validation steps. Click here to view larger figure.

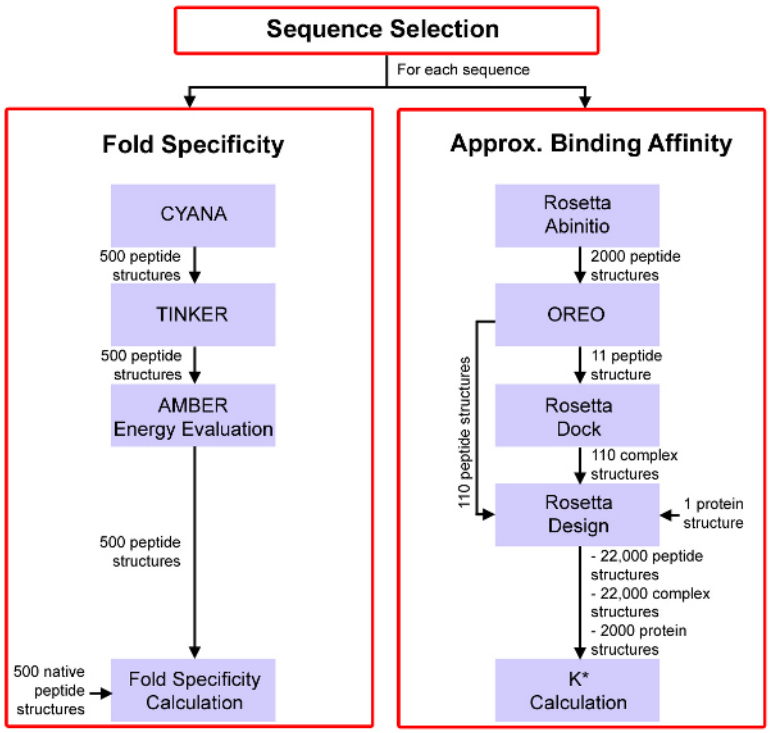


Figure 2. Detailed overview of stage two approaches. Results from the sequence selection stage are used as inputs into the final fold validation steps. Details of both Fold Specificity and Approximate Binding Affinity calculations are shown in this expanded flow diagram.

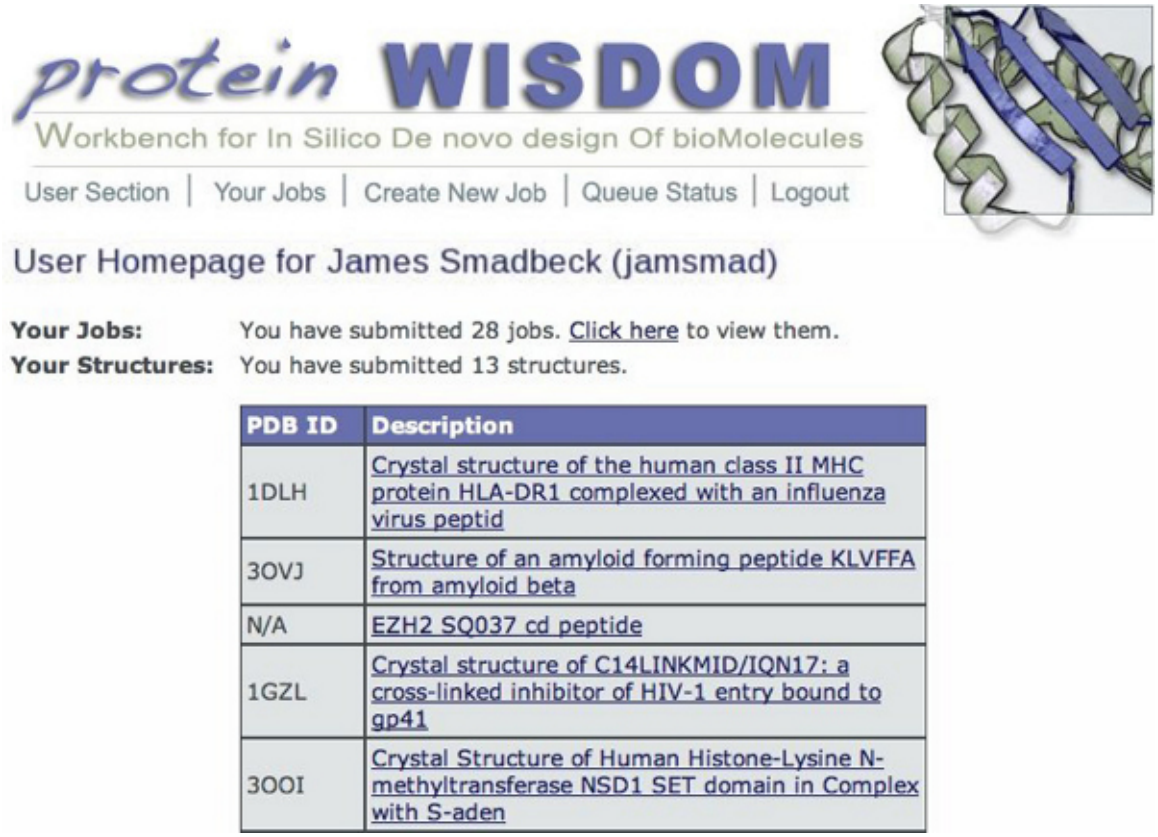


Figure 3. User Homepage. Once the user has registered and logged-in they can view the user homepage. This is the main control page for all of a user's structures and jobs that they have submitted through Protein WISDOM. There are several important links found on this page. **(A.)** The user can view all previous jobs they have submitted by clicking on the "click here" link under "Your jobs". **(B.)** The user can view all protein structures they have submitted to the system directly on this page. **(C.)** Once the user is ready to submit a job they can click "Create New Job" at the top of the page.

Figure 4. Job Creation page. To create a job, **(A.)** the user must first name the job and then **(B.)** indicate whether the job is based on a previously submitted job. If so, **(C.)** a dropdown menu of previously submitted jobs allows the user to choose which one. A job based on a previously submitted job must use the same template structure, but all other design inputs can be modified. **(D.)** Once the user is ready for template submission they can click the "Continue" button.

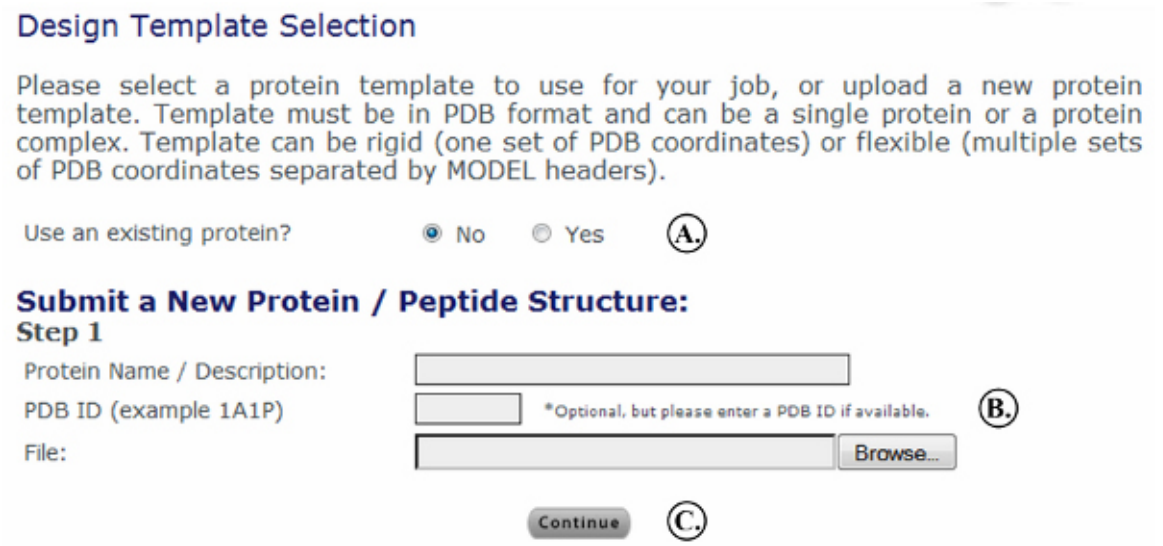


Figure 5. Design Template Submission page. To initialize a new design job, the user must submit a design template. **(A.)** The user first chooses whether the job uses a previously submitted structure. If so, the user must select the template from a table of previously submitted structures. If not, **(B.)** The user must name the structure, indicate the PDB source of the structure, and upload a structure file in PDB format. **(C.)** Once the user is ready for further design specification they can click the "Continue" button.

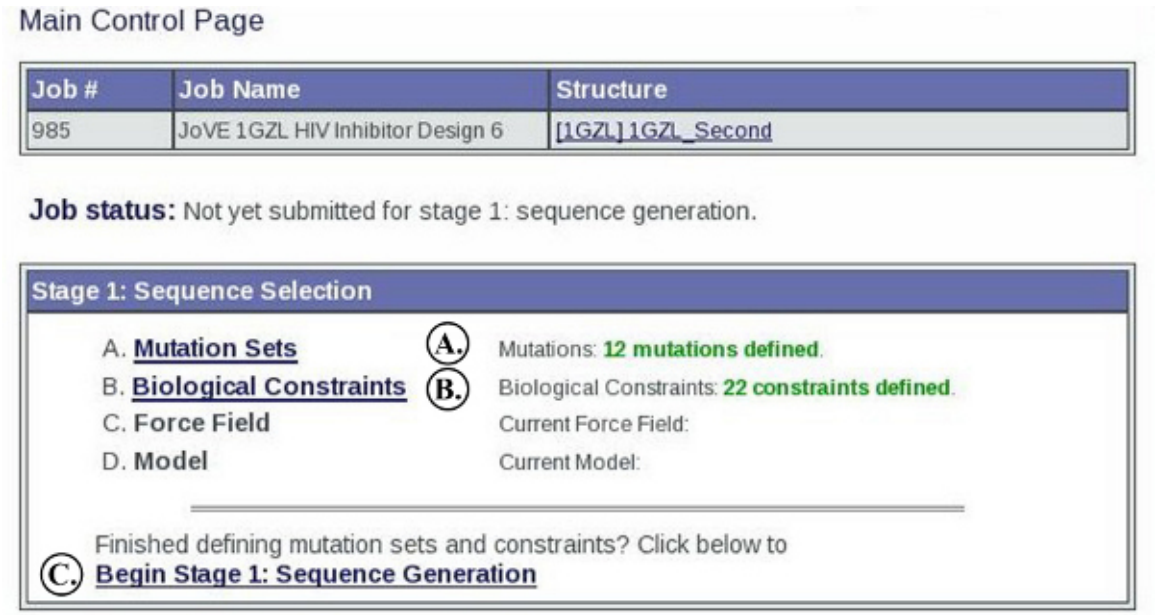


Figure 6. Main Control Page before submission of Stage One. Once a template has been successfully uploaded, the user is taken to the Main Control Page for that job. Before submission of Stage One, only Design Input and Stage One links are active. **(A.)** In order to input the mutation constraints for the job, the user must click on the "Mutation Sets" link. **(B.)** In order to input biological constraints, like charge and amino acid content constraints, the user must click on "Biological Constraints" link. **(C.)** Once the user is ready to start a design job, they can click on "Begin Stage 1: Sequence Generation" to input stage one parameters and submit the job.

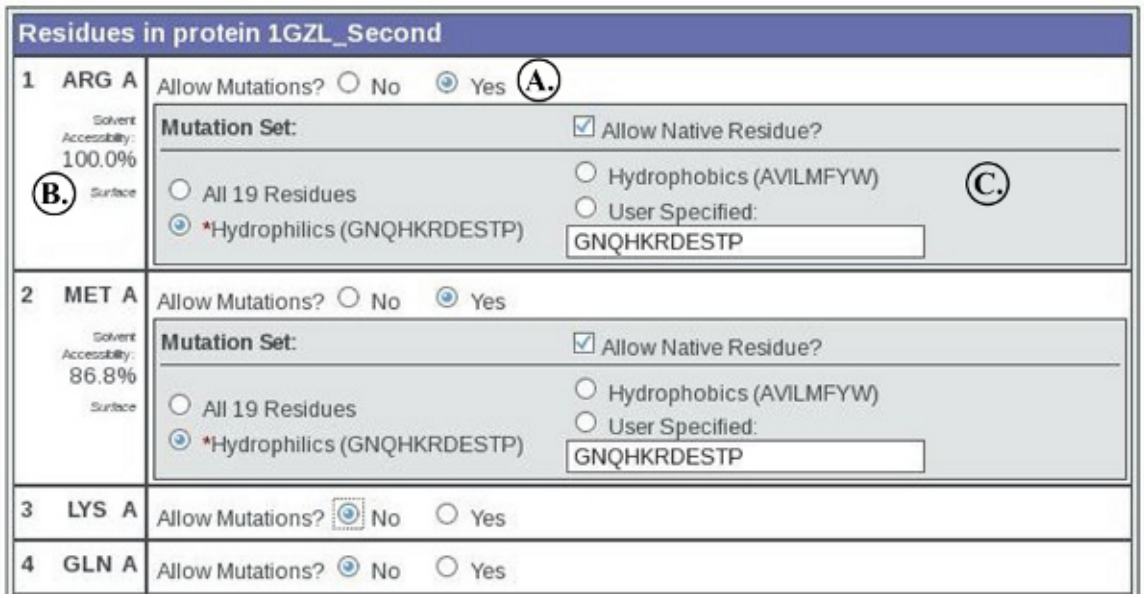


Figure 7. Mutation Sets selection page. To select mutation constraints, the user must go to the "Mutation Sets" page. **(A.)** First, the user must specify which residues are mutatable in the system. **(B.)** Solvent Accessible Surface Area (SASA) calculations are performed for each position in the protein, and a default mutation set is generated automatically. **(C.)** The user can also manually specify allowed mutations for each position. Click here to view larger figure.

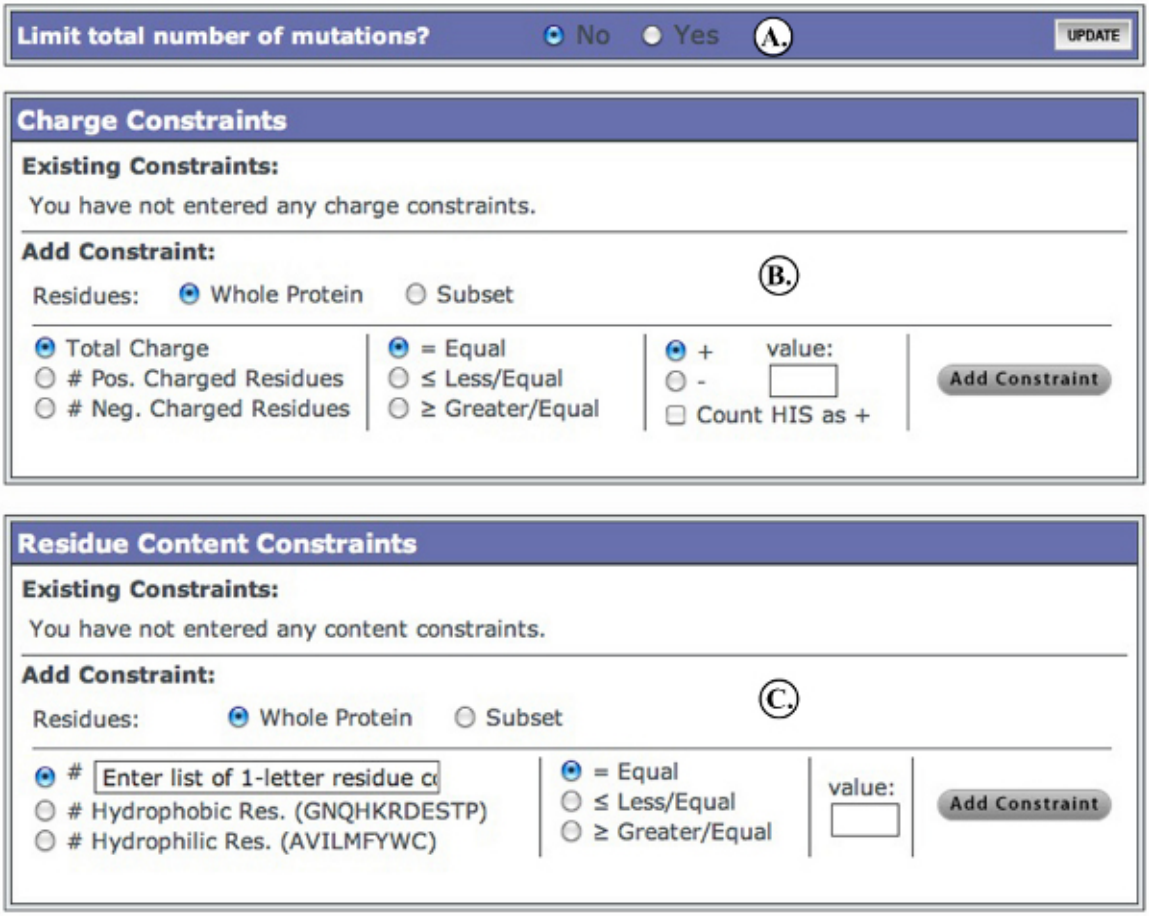


Figure 8. Biological Constraints selection page. To select biological constraints, the user must go to the "Biological Constraints" selection page. There are three types of biological constraints that can be specified by the user. **(A.)** The user may specify a limit on the number of mutations allowed in a given design. **(B.)** The user may specify upper and lower charge constraints for all or part of the design sequence. **(C.)** The user may specify upper and lower amino acid content constraints for individual or sets of amino acids for all or part of the design sequence.

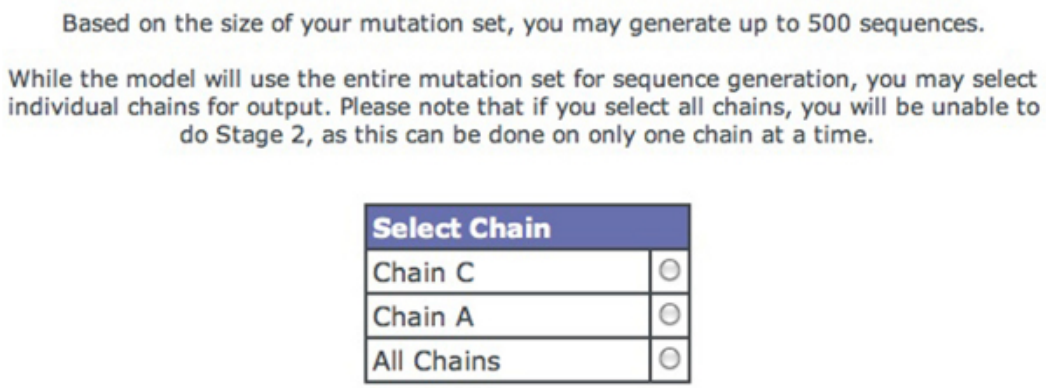


Figure 9. Selecting a specific chain for output for Stage One submission of a complex. Before Stage One submission, the user must specify which chain the design is being performed on. Design can be performed on single or multiple chains. However, if one wishes to design using Fold Specificity, only a single design chain can be selected.

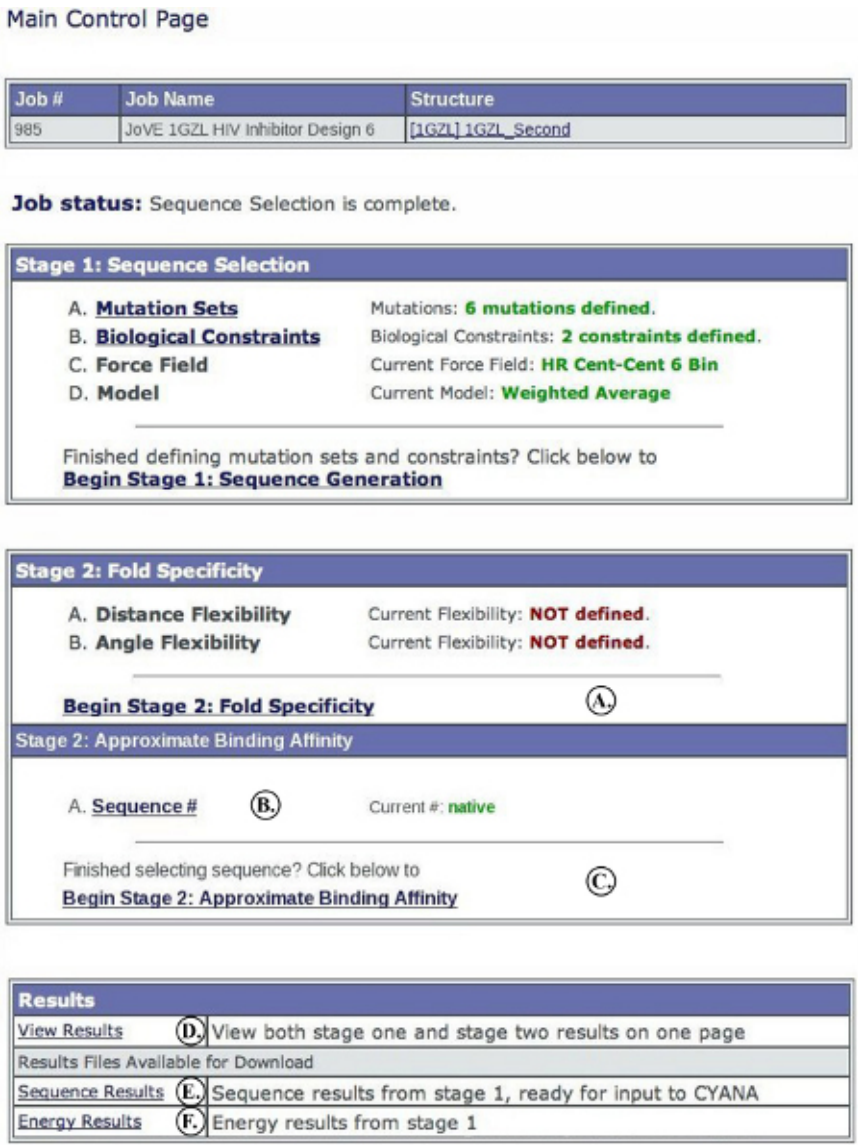


Figure 10. Main Control Page upon completion of Stage One. Upon the completion of Stage One, several new options are unlocked. **(A.)** The user can submit the selected sequences for Fold Specificity calculation by clicking on the "Begin Stage Two: Fold Specificity" link. **(B.)** Before submitting for Approximate Binding Affinity calculation, which can only be run for a single sequence at a time, the user must select a sequence by clicking on the "Sequence #" link. **(C.)** Once a sequence has been selected, the user may submit for Approximate Binding Affinity calculation by clicking on the "Begin Stage 2: Approximate Binding Affinity" link. **(D.)** A rank-ordered list of sequences based on Stage One energy can be viewed by clicking on the "View Results" link at the bottom of the control page. **(E.)** Stage One results in CYANA sequence format can be viewed by clicking on the "Sequence Results" link at the bottom of the control page. **(F.)** Stage one output from the optimization model, with energy and solve time for each sequence selected, can be viewed by clicking on the "Energy Results" link at the bottom of the control page. Click here to view larger figure.

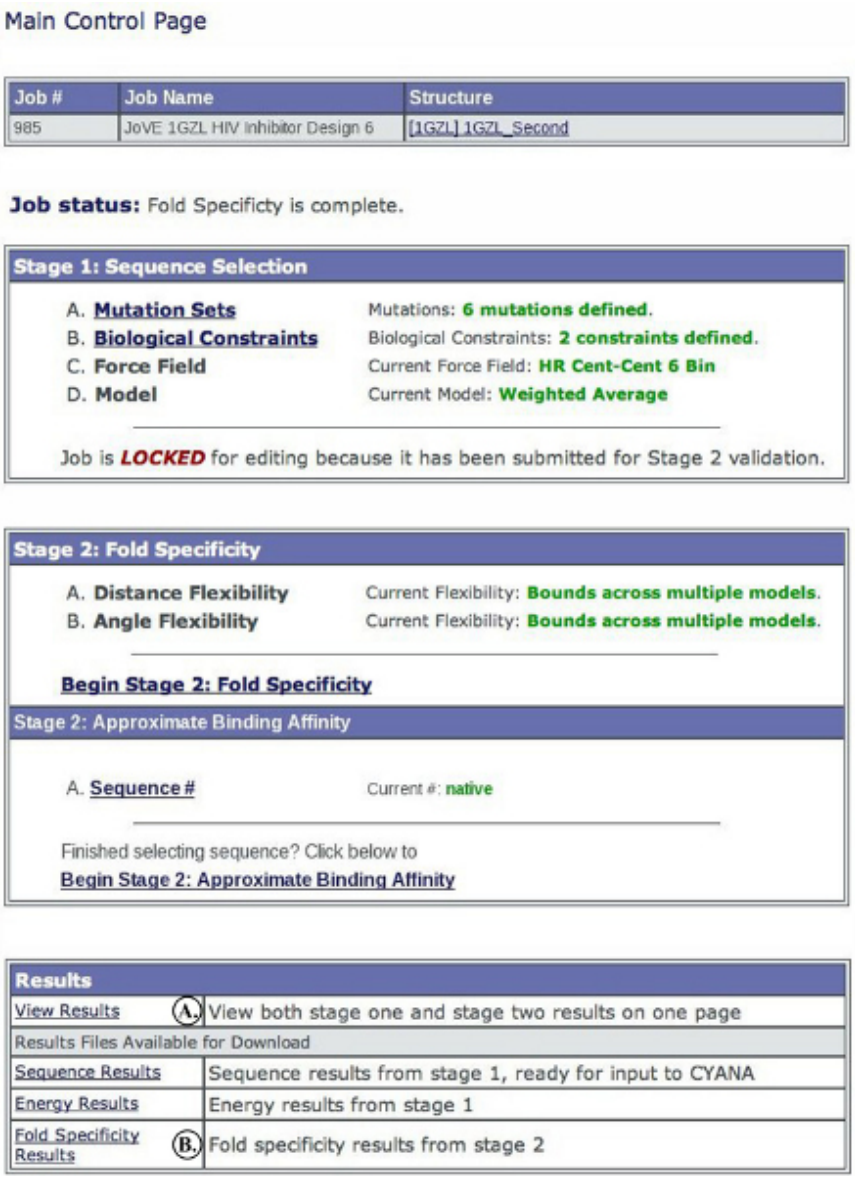


Figure 11. Main Control Page upon completion of Stages One and Two. Upon the completion of the Fold Specificity calculation stage, several new options are unlocked. **(A.)** A rank-ordered list of sequences based on Stage One energy or Fold Specificity can be viewed by clicking on the "View Results" link at the bottom of the control page. **(B.)** Output from the Fold Specificity stage can be viewed by clicking on the "Fold Specificity Results" link at the bottom of the control page. Click here to view larger figure.

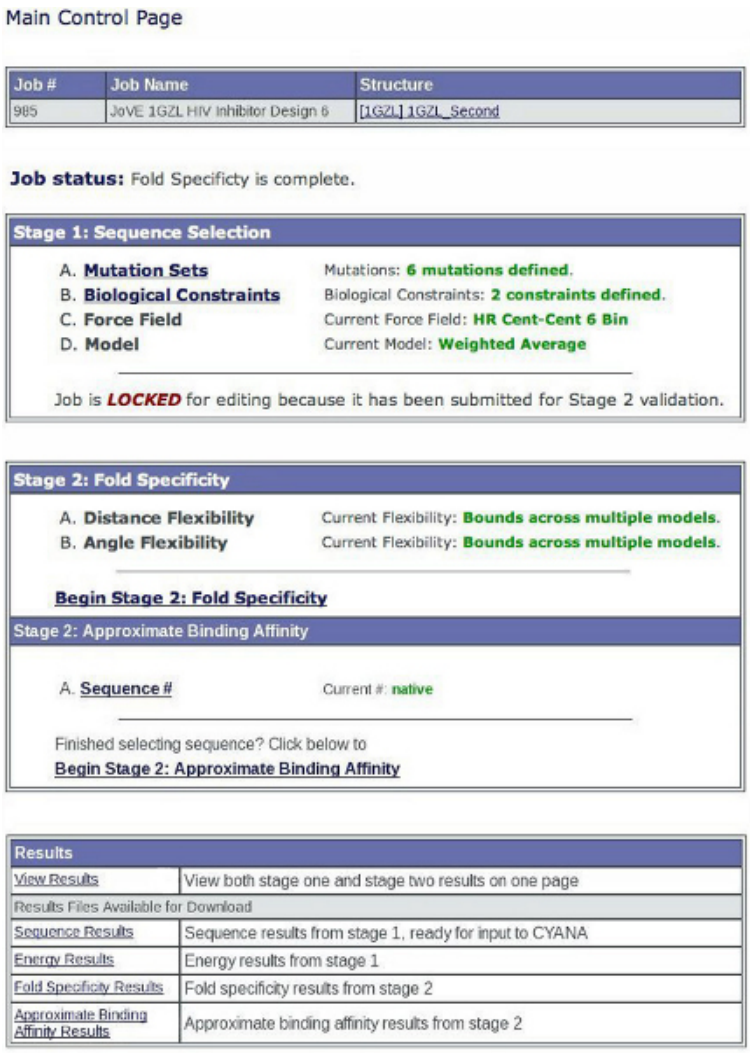


Figure 12. Main Control Page upon completion of Stages One and Two with Binding Affinity Calculation. Upon the completion of Approximate Binding Affinity, several new options are unlocked. **(A.)** A rank-ordered list of sequences based on Stage One energy, Fold Specificity, or Approximate Binding Affinity, as well as designed protein structures, can be viewed by clicking on the "View Results" link at the bottom of the control page. **(B.)** Output from the Approximate Binding Affinity stage can be viewed by clicking on the "Approximate Binding Affinity Results" link at the bottom of the control page. Click here to view larger figure.

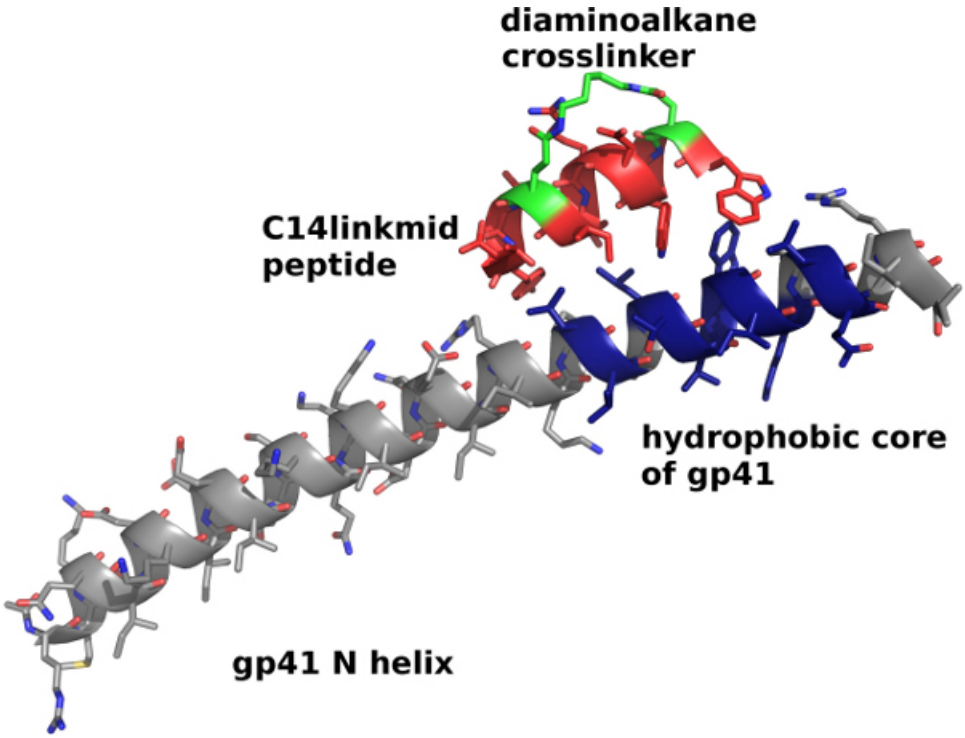


Figure 13. HIV-1 gp41 Complex Template Structure (PDB: 1GZL). HIV-1 template structure derived from PDB structure 1GZL. The linker in the template peptide must be removed before template submission. This template is used for mutation set and distance constraint generation for Stage One and Stage Two calculations.

Job #	Job Name	Structure
985	JoVE 1GZL HIV Inhibitor Design 6	[1GZL] 1GZL_Second

Job status: Sequence Selection is complete.

Ranks are given for sequence selection (low energy (E) = rank 1), fold specificty (high fold specificty (fspec) = rank 1), and approximate binding affinity (high affinity (K*) = rank 1), when applicable.

E Rank	E	fspec Rank	fspec	K* Rank	K*	Sequence	View
native	-	-	-	-	-	WEEWDREIENYT	-
1	-0.130	-	-	-	-	WADWDDRYERWR	-
2	-0.130	-	-	-	-	WCDWDERYERWR	-
3	-0.130	-	-	-	-	WCDWDERYDRWR	-
4	-0.129	-	-	-	-	WCDWDERYDKWR	-
5	-0.129	-	-	-	-	YADYDDRYERWR	-
6	-0.129	-	-	-	-	WCDWEERYERWR	-
7	-0.129	-	-	-	-	YCDYDERYERWR	-
8	-0.129	-	-	-	-	YCDYDERYDRWR	-
9	-0.129	-	-	-	-	WCDWEERYDRWR	-
10	-0.129	-	-	-	-	WAEWDDRYERWR	-

Figure 14. Sortable Design Results upon Completion of Stage One. The "View Results" section of Protein WISDOM allows the user to sort the design results by the output of each design method. **(A.)** By clicking the "E Rank" link, the table will sort by the Stage One Energy output. **(B.)** The "E" column provides the potential energy calculated for the designed sequence in the given structural template. **(C.)** The selected sequences are provided in the "Sequence" column.

```
* Native sequence
TRP GLU GLU TRP ASP ARG GLU ILE GLU ASN TYR THR +

* Sequence #1
TRP ALA ASP TRP ASP ASP ARG TYR GLU ARG TRP ARG +

* Sequence #2
TRP CYS ASP TRP ASP GLU ARG TYR GLU ARG TRP ARG +

* Sequence #3
TRP CYS ASP TRP ASP GLU ARG TYR ASP ARG TRP ARG +

* Sequence #4
TRP CYS ASP TRP ASP GLU ARG TYR ASP LYS TRP ARG +

* Sequence #5
TYR ALA ASP TYR ASP ASP ARG TYR GLU ARG TRP ARG +

* Sequence #6
TRP CYS ASP TRP GLU GLU ARG TYR GLU ARG TRP ARG +

* Sequence #7
TYR CYS ASP TYR ASP GLU ARG TYR GLU ARG TRP ARG +

* Sequence #8
TYR CYS ASP TYR ASP GLU ARG TYR ASP ARG TRP ARG +

* Sequence #9
TRP CYS ASP TRP GLU GLU ARG TYR ASP ARG TRP ARG +

* Sequence #10
TRP ALA GLU TRP ASP ASP ARG TYR GLU ARG TRP ARG +
```

Figure 15. Sequence Results upon Completion of Stage One. The "Sequence Results" page provides downloadable Stage One results in format compatible with input into CYANA, as is used in the Stage Two Fold Specificity method.

```
Model Status:         1.00
ITERATION 1
ENERGY        -0.130
YIJ = 46  TRP
YIJ = 47  ALA
YIJ = 48  ASP
YIJ = 49  TRP
YIJ = 50  ASP
YIJ = 51  ASP
YIJ = 52  ARG
YIJ = 53  TYR
YIJ = 54  GLU
YIJ = 55  ARG
YIJ = 56  TRP
YIJ = 57  ARG

time=         0.69
```

Figure 16. Energy Results upon Completion of Stage One. The "Energy Results" page provides output from the optimization model from Stage One. **(A.)** The designed sequence, restricted to only those positions allowed to be modified, is provided, along with **(B.)** the potential energy of the sequence in the given template structure, and **(C.)** the time it took to solve for the sequence.

Job #	Job Name	Structure
985	JoVE 1GZL HIV Inhibitor Design 6	[1GZL] 1GZL_Second

Job status: Sequence Selection is complete.

Ranks are given for sequence selection (low energy (E) = rank 1), fold specificty (high fold specificty (fspec) = rank 1), and approximate binding affinity (high affinity (K*) = rank 1), when applicable.

E Rank	E	fspec Rank	fspec	K* Rank	K*	Sequence	View
native	-	-	-	-	-	WEEWDREIENYT	-
370	-0.126	1	1782.01	-	-	YVDYDDRYERWR	-
322	-0.126	2	1771.71	-	-	YVEYDDRYDRWR	-
376	-0.126	3	1743.86	-	-	YLDYDDRYERWR	-
323	-0.126	4	1715.58	-	-	YLEYDDRYDRWR	-
292	-0.126	5	1614.40	-	-	YVEYDDRYERWR	-
490	-0.125	6	1612.18	-	-	YVEYDERYDRWR	-
489	-0.125	7	1582.04	-	-	YLEYDERYDRWR	-
291	-0.126	8	1535.50	-	-	YLEYDDRYERWR	-
447	-0.126	9	1376.02	-	-	YLEYDERYERWR	-
372	-0.126	10	1350.77	-	-	YMDYDDRYERWR	-

Figure 17. Sortable Design Results upon Completion of Fold Specificity Stage. Following the completion of the Stage Two Fold Specificity calculation, the "View Results" section of Protein WISDOM allows the user to sort the design results. **(A.)** By clicking the "F Rank" link, the table will sort by the Stage Two Fold Specificity output. **(B.)** The "F" column provides the Fold Specificity calculated for the designed sequence in the given structural template.

Sequence #	Fold Specificity
1	598.91
2	508.40
3	596.75
4	36.70
5	1298.96
6	487.22
7	1190.85
8	1268.48
9	532.78
10	493.78
11	600.76
12	83.41
13	35.42
14	458.14
15	38.11
16	555.86
17	1040.63
18	549.64
19	38.94
20	847.06

Figure 18. Fold Specificity Results upon Completion of Fold Specificity Stage. The "Fold Specificity Results" page provides a downloadable text file with the Fold Specificity results.

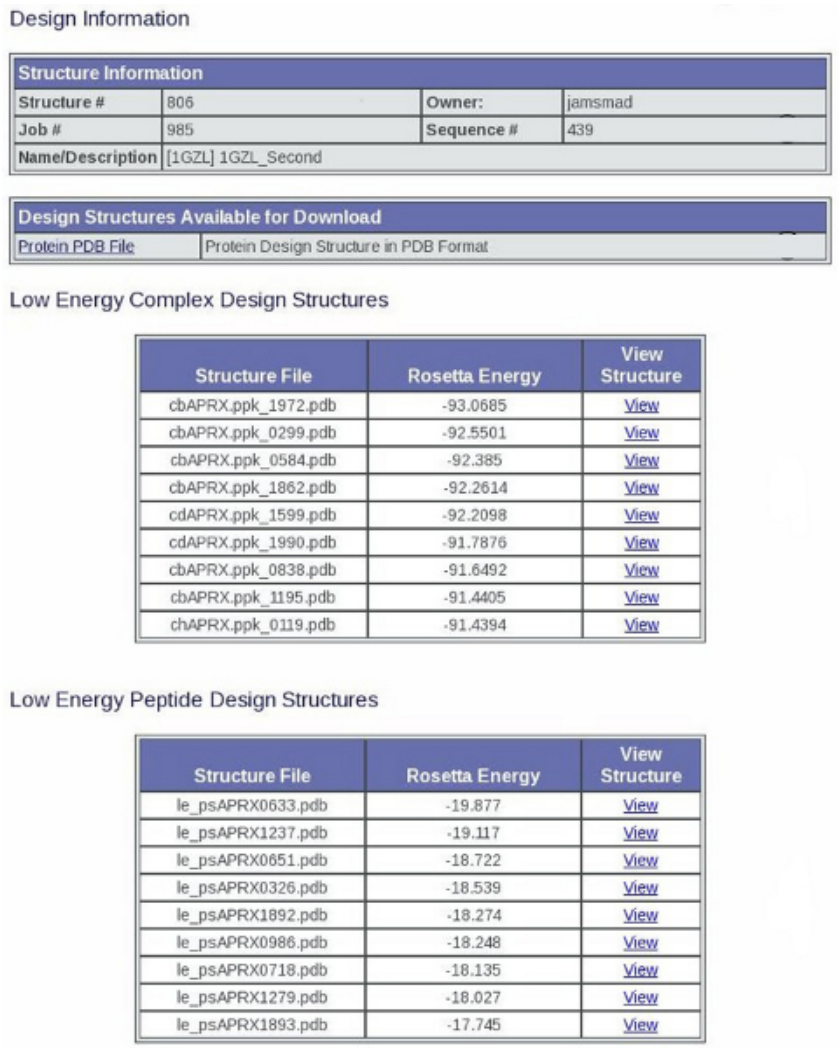
Job #	Job Name	Structure
985	JoVE 1GZL HIV Inhibitor Design 6	[1GZL] 1GZL_Second

Job status: Sequence Selection is complete.

Ranks are given for sequence selection (low energy (E) = rank 1), fold specificty (high fold specificty (fspec) = rank 1), and approximate binding affinity (high affinity (K*) = rank 1), when applicable.

E Rank	E	fspec Rank	fspec	K* Rank	K*	Sequence	View
439	-0.126	150	446.20	1	3.20e+00	WCDWRDEWERYR	439
native	-	-	-	2	4.55e-01	WEEWDREIENYT	native
1	-0.130	109	598.91	-	-	WADWDDRYERWR	-
2	-0.130	139	508.40	-	-	WCDWDERYERWR	-
3	-0.130	110	596.75	-	-	WCDWDERYDRWR	-
4	-0.129	257	36.70	-	-	WCDWDERYDKWR	-
5	-0.129	19	1298.96	-	-	YADYDDRYERWR	-
6	-0.129	145	487.22	-	-	WCDWEERYERWR	-
7	-0.129	26	1190.85	-	-	YCDYDERYERWR	-
8	-0.129	20	1268.48	-	-	YCDYDERYDRWR	-
9	-0.129	127	532.78	-	-	WCDWEERYDRWR	-
10	-0.129	143	493.78	-	-	WAEWDDRYERWR	-

Figure 19. Sortable Design Results upon Completion of Approximate Binding Affinity Calculation. Following the completion of the Stage Two Approximate Binding Affinity calculation, the "View Results" section of Protein WISDOM allows the user to sort the design results. **(A.)** By clicking the "K* Rank" link, the table will sort by the Stage Two Approximate Binding Affinity, K*, values. **(B.)** The "K*" column provides the K* values calculated for the designed sequence docked to the template protein structure. **(C.)** All sequences that finish the Approximate Binding Affinity calculation stage will have a link to a structural data page provided in the "View" column.



Design Information

Structure Information			
Structure #	806	Owner:	jamsmad
Job #	985	Sequence #	439
Name/Description	[1GZL] 1GZL_Second		

Design Structures Available for Download	
Protein PDB File	Protein Design Structure in PDB Format

Low Energy Complex Design Structures

Structure File	Rosetta Energy	View Structure
cbAPRX.ppk_1972.pdb	-93.0685	View
cbAPRX.ppk_0299.pdb	-92.5501	View
cbAPRX.ppk_0584.pdb	-92.385	View
cbAPRX.ppk_1862.pdb	-92.2614	View
cdAPRX.ppk_1599.pdb	-92.2098	View
cdAPRX.ppk_1990.pdb	-91.7876	View
cbAPRX.ppk_0838.pdb	-91.6492	View
cbAPRX.ppk_1195.pdb	-91.4405	View
chAPRX.ppk_0119.pdb	-91.4394	View

Low Energy Peptide Design Structures

Structure File	Rosetta Energy	View Structure
le_psAPRX0633.pdb	-19.877	View
le_psAPRX1237.pdb	-19.117	View
le_psAPRX0651.pdb	-18.722	View
le_psAPRX0326.pdb	-18.539	View
le_psAPRX1892.pdb	-18.274	View
le_psAPRX0986.pdb	-18.248	View
le_psAPRX0718.pdb	-18.135	View
le_psAPRX1279.pdb	-18.027	View
le_psAPRX1893.pdb	-17.745	View

Figure 20. Complete Designed Sequence Structure Summary. By clicking on the link in the "View" column of the "View Results" table the user has access to the "Structure Summary" page for that sequence. **(A.)** The page provides links to the relevant jobs and structures related to that designed sequence. **(B.)** Downloadable .pdb files in .zip format from the "Protein PDB File" link. **(C.)** Low-energy complex structures generated in the Approximate Binding Affinity calculation are provided with "View" links to Jmol interactive viewing. **(D.)** Low-energy peptide structures generated in the Approximate Binding Affinity calculation are provided with "View" links to Jmol interactive viewing.

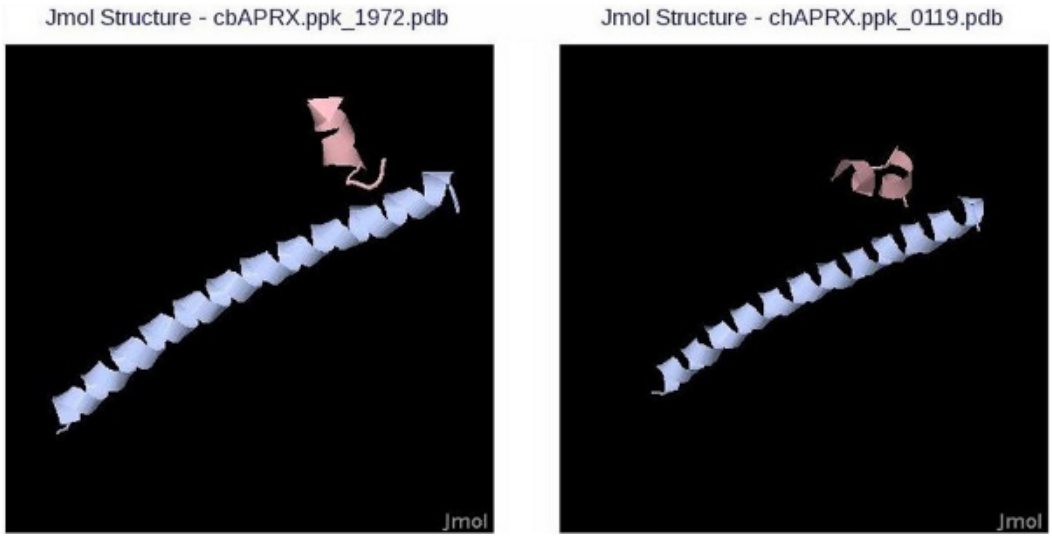


Figure 21. Interactive Jmol Environment for Low-Energy Structure Viewing. An example of low-energy docked structures produced during an Approximate Binding Affinity calculation. In this case, we show two low-energy structures produced during the representative results run using the HIV-1 structure template from PDB:1GZL.

Sequence #	qPL	qL	qP	K*	offset
native	3.45e+67	6.85e+18	1.10e+49	4.55e-01	0
1	1.28e+67	3.57e+17	1.12e+49	3.20e+00	0
2	2.29e+66	4.59e+16	2.16e+49	4.20e+00	0

Figure 22. Binding Affinity Results upon Completion of Approximate Binding Affinity Calculation. The "Approximate Binding Affinity Results" page provides a downloadable text file with the Approximate Binding Affinity results.

Discussion

The *de novo* protein design framework consists of two stages, a sequence selection stage and a validation stage. The framework is robust enough to handle rigid and flexible design templates, and can be applied to single protein design or complex protein design. The framework has been successfully applied to numerous protein systems with applications to dozens of diseases. A number of the designs have been experimentally validated, providing the most potent inhibitors or agonists of some proteins discovered to date. This framework is now available to the academic community via Protein WISDOM.

There are three critical steps in the method. The first is the Sequence Selection stage, which employs global optimization techniques for protein design. The protein design problem is a high complexity problem (20^n possible sequences for *n* mutatable positions). This number is significantly higher than the possible number of sequence that can be considered by experimental design methods. Further inclusion of mutation and biological constraints speeds up the optimization through the reduction of complexity. Overall, this results in a method capable of quickly identifying the biologically relevant sequence with the global minimum potential energy.

The second critical step of the method is Fold Specificity. In this stage, how well the designed sequences from Sequence Selection fold into the desired template structure in comparison to the native sequence is calculated. This stage increases the rigor of the calculations through the determination and minimization of mutated structures in order to rerank the designed sequences.

The final critical step of the method is Approximate Binding Affinity Calculation, which takes a small subset of designed sequences from the first two stages of the method and measures how well they bind to a target protein. This step is completely *ab initio,* as it takes only the designed sequence in and produces large ensembles of peptide, protein, and complex structures. This allows the method to take into account changes in structure and docking poses that could be induced by changes in sequence.

The *de novo* design framework described within addresses the design of single proteins as well as protein-peptide complexes. The generalization of the framework to address multimeric systems, protein-DNA interactions, and the design with post-translational modifications and noncanonical amino acids represent limitations to the web interface described above. Each expansion poses its own unique challenges and are currently under development for inclusion in future versions of Protein WISDOM.

Disclosures

The authors declare that they have no competing financial interests.

Acknowledgements

CAF gratefully acknowledges support from NSF, NIH (R01 GM52032; R24 GM069 736), and the US Environmental Protection Agency, EPA (R 832721-010). A portion of this research was made possible with Government support by DoD, Air Force Office of Scientific Research. JS gratefully acknowledges support from NIH (P50GM071508-06). MLBP gratefully acknowledges support from a National Defense Science and

Engineering Graduate (NDSEG) Fellowship, 32 CFR 168a. GAK gratefully acknowledges support from a National Science Foundation Graduate Research Fellowship under grant number DGE-1148900.

References

1. Drexler, K. Molecular engineering: An approach to the development of general capabilities for molecular manipulation. *Proc. Natl Acad. Sci. U.S.A.* **78**, 5275-5278 (1981).
2. Pabo, C. Molecular technology: Designing proteins and peptides. *Nature*. **301**, 200 (1983).
3. Floudas, C.A. Research challenges, opportunities and synergism in systems engineering and computational biology. *AIChE J.* **51**, 1872-1884 (2005).
4. Fung, H.K., Welsh, W.J., & Floudas, C.A. Computational *de novo* peptide and protein design: Rigid templates versus flexible templates. *Ind. Eng. Chem. Res.* **47** (4), 993-1001 (2008).
5. Ponder, J. & Richards, F. Tertiary templates for proteins. *J. Mol. Biol.* **193**, 775-791 (1987).
6. Dahiyat, B.I. & Mayo, S.L. Protein design automation. *Protein Sci.* **5**, 895-903 (1996).
7. Dahiyat, B.I., Gordon, D.B., & Mayo, S.L. Automated design of the surface positions of protein helices. *Protein Sci.* **6**, 1333-1337 (1997).
8. Su, A. & Mayo, S.L. Coupling backbone flexibility and amino acid sequence selection in protein design. *Protein Sci.* **6**, 1701-1707 (1997).
9. Desjarlais, J. & Handel, T. Side chain and backbone flexibility in protein core design. *J. Mol. Biol.* **290**, 305-318 (1999).
10. Farinas, E., Regan, L. The *de novo* design of a rubredoxin-like Fe site. *Protein Sci.* **7**, 1939-1946 (1998).
11. Harbury, P.B., Plecs, J.J., Tidor, B., Alber, T., & Kim, P.S. High-resolution protein design with backbone freedom. *Science.* **282**, 1462-1467 (1998).
12. Koehl, P. & Levitt, M. *De novo* protein design: I. In search of stability and specificity. *J. Mol. Biol.* **293**, 1161-1181 (1999).
13. Koehl, P. & Levitt, M. *De novo* protein design. II. Plasticity in sequence space. *J. Mol. Biol.* **293**, 1183-1193 (1999).
14. Kuhlman, B., Dantae, G., Ireton, G., Verani, G., Stoddard, B., & Baker, D. Design of a novel globular protein fold with atomic-level accuracy. *Science*. **302**, 1364-1368 (2003).
15. Klepeis, J.L., Floudas, C.A., *et al.* Integrated structural, computational and experimental approach for lead optimization: Design of compstatin variants with improved activity. *J. Am. Chem. Soc.* **125**, 8422-8423 (2003).
16. Klepeis J.L., Floudas C.A., Morikis D., Tsokos C.G., & Lambris J.D. Design of peptide analogs with improved activity using a novel *de novo* protein design approach. *Ind. Eng. Chem. Res.* **43**, 3817-3826 (2004).
17. Fung H.K., Floudas C.A., Taylor M.S., Zhang L., & Morikis D. Toward full-sequence *de novo* protein design with flexible templates for human beta-defensin-2. *Biophys. J.* **94**, 584-599 (2008).
18. Bellows M.L., Fung H.K., Floudas C.A., López de Victoria A., & Morikis D. New compstatin variants through two de novo protein design frameworks. *Biophys. J.* **98** (10), 2337-2346 (2010).
19. López de Victoria, A., Gorham Jr, R.D., *et al.* A new generation of potent complement inhibitors of the compstatin family. *Chem. Biol. Drug Des.* **77**, 431-440 (2011).
20. Tamamis, P., López de Victoria, A., *et al*. Molecular dynamics in drug design: New generations of compstatin analogs. *Chem. Biol. Drug Des.* **79** (5), 703-718 (2012).
21. Bellows-Peterson M.L., Fung H.K., *et al. De novo* peptide design with c3a receptor agonist and antagonist activities: Theoretical predictions and experimental validation. *J. Med. Chem.* **55** (9), 4159-4168 (2012).
22. Bellows, M.L., Taylor, M.S., *et al.* Discovery of entry inhibitors for HIV-1 via a new *de novo* protein design framework. *Biophys. J.* **99**, 3445-3453 (2010).
23. Sun J.-J., Abdeljabbar D.M., Clarke N.L., Bellows M.L., Floudas C.A., & Link A.J. Reconstitution and engineering of apoptotic protein interactions on the bacterial cell surface. *J. Mol. Biol.* **394**, 297-305 (2009).
24. Smadbeck J., Bellows-Peterson M.L., *et al. De novo* protein design and validation of histone methyltranferase inhibitors., In Preparation, (2013).
25. Bellows, M.L., Fung, H.K., Floudas, C.A. In: *Molecular Systems Engineering, Process Systems Engineering.*, Adjiman, C.S. & Galindo, A., eds., Wiley-VCH Verlag GmbH & Co. KGaA, **6**, 207-232 (2010).
26. Rajgaria, R., McAllister, S.R., & Floudas, C.A. A novel high resolution C^{α}-C^{α} distance dependent force field based on a high quality decoy set. *Proteins.* **65**, 726-741 (2006).
27. Rajgaria, R., McAllister, S.R., & Floudas, C.A. Distance dependent centroid to centroid force fields using high resolution decoys. *Proteins.* **70**, 950-970 (2008).
28. Fung H.K., Taylor M.S., & Floudas C.A. Novel formulations for the sequence selection problem in de novo protein design with flexible templates. *Optim. Method. Softw.* **22**, 51-71 (2007).
29. Fung H.K., Rao S., Floudas C.A., Prokopyev O., Pardalos P.M., & Rendl F. 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).
30. CPLEX Using the CPLEX Callable Library, ILOG, Inc., (1997).
31. Klepeis J.L. & Floudas C.A. Free energy calculations for peptides via deterministic global optimization. *J. Chem. Phys.* **110**, 7491-7512 (1999).
32. Klepeis J.L., Floudas C.A., Morikis D., & Lambris J.D. Predicting peptide structures using NMR data and deterministic global optimization. *J. Comput. Chem.* **20**, 1354-1370 (1999).
33. Klepeis J.L., Schafroth H.D., Westerberg K.M., & Floudas C.A. Deterministic global optimization and ab initio approaches for the structure prediction of polypeptides, dynamics of protein folding and protein-protein interactions. *Adv. Chem. Phys.* **120**, 265-457 (2002).
34. Klepeis J.L. & Floudas C.A. Ab initio prediction of helical segments of polypeptides. *J. Comput. Chem.* **23**, 246-266 (2002).
35. Klepeis J.L. & Floudas C.A. Prediction of beta-sheet topology and disulfide bridges in polypeptides. *J. Comput. Chem.* **24**, 191-208 (2003).
36. Klepeis J.L. & Floudas C.A. ASTRO-FOLD: A combinatorial and global optimization framework for ab initio prediction of three-dimensional structures of proteins from the amino acid sequence. *Biophys. J.* **85**, 2119-2146 (2003).
37. Klepeis J.L., Pieja M.T., & Floudas C.A. A new class of hybrid global optimization algorithms for peptide structure prediction: Integrated hybrids. *Comput. Phys. Commun.* **151**, 121-140 (2003).

38. Klepeis J., Pieja M., & Floudas C. Hybrid global optimization algorithms for protein structure prediction : Alternating hybrids. *Biophys. J.* **84**, 869-882 (2003b).
39. Klepeis J.L. & Floudas C.A. Analysis and prediction of loop segments in protein structures. *Comput. Chem. Eng.* **29**, 423-436 (2005).
40. Mo¨nnigmann M. & Floudas C.A. Protein loop structure prediction with flexible stem geometries. *Proteins.* **61**, 748-762 (2005).
41. McAllister S.R., Mickus B.E., Klepeis J.L., & Floudas C.A. A novel approach for alpha-helical topology prediction in globular proteins: Generation of interhelical restraints. *Proteins.* **65**, 930-952 (2006).
42. Floudas, C.A., Fung, H.K., McAllister, S.R., Mönnigmann, M., & Rajgaria, R. Advances in protein structure prediction and *de novo* protein design: A review. *Chem. Eng. Sci.* **61**, 966-988 (2006).
43. Subramani A., Wei Y., & Floudas C.A. ASTRO-FOLD 2.0: An enhanced framework for protein structure prediction. *AIChE J.* **58** (5), 1619-1637 (2012).
44. Wei Y., Thompson J., & Floudas C.A. Concord: a consensus method for protein secondary structure prediction via mixed integer linear optimization. *P. Roy. Soc. A-Math. Phy.* **468**, 831-850 (2011).
45. Subramani A. & Floudas C.A.β-sheet topology prediction with high precision and recall for β and mixed α/β proteins. *PLoS One.* **7** (3), e32461 (2012).
46. Rajgaria R., Wei Y., & Floudas C.A. Contact prediction for beta and alpha-beta proteins using integer linear optimization and its impact on the first principles 3D structure prediction method ASTRO-FOLD. *Proteins*. **78** (8), 1825-1846 (2010).
47. Subramani A. & Floudas C.A. Structure prediction of loops with fixed and flexible stems. *J. Phys. Chem. B.* **116** (23), 6670-6682 (2012).
48. Güntert P., Mumenthaler C., & Wüthrich K. Torsion angle dynamics for NMR structure calculation with the new program DYANA. *J. Mol. Biol.* **273**, 283-298 (1997).
49. Güntert P. Automated NMR structure calculation with CYANA. *Methods Mol. Biol.* **278**, 353-378 (2004).
50. Ponder J. TINKER, software tools for molecular design. 1998, Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine: St. Louis, MO., (1998).
51. Cornell W.D., Cieplak P., *et al.* A 2nd generation forcefield for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **117**, 5179-5197 (1995).
52. Lilien R.H., Stevens B.W., Anderson A.C., & Donald B.R. 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. *J. Comput. Biol.* **12**, 740-761 (2005).
53. Lee M.R., Baker D., & Kollman P.A. 2.1 and 1.8 A°Cα RMSD structure predictions on two small proteins, HP-36 and S15. *J. Am. Chem. Soc.* **123** (6), 1040-1046 (2001).
54. Rohl C.A. & Baker D. *De novo* determination of protein backbone structure from residual dipolar couplings using rosetta. *J. Am. Chem. Soc.* **124** (11), 2723-2729 (2002).
55. Rohl C.A., Strauss C.E.M., Misura K.M.S., & Baker D. Protein structure prediction using rosetta. *Methods Enzymol.* **383**, 66-93 (2004).
56. DiMaggio P.A., McAllister S.R., Floudas C.A., Feng X.J., Rabinowitz J.D., & Rabitz H.A. Biclustering via optimal re-ordering of data matrices in systems biology: Rigorous methods and comparative studies. *BMC Bioinformatics.* **9** (458), (2008).
57. DiMaggio P.A., McAllister S.R., Floudas C.A., Feng X.J., Rabinowitz J.D., & Rabitz H.A. A network flow model for biclustering via optimal re-ordering of data matrices. *J Global Optimization.* **47** (3), 343-354 (2010).
58. Daily M.D., Masica D., Sivasubramanian A., Somarouthu S., & Gray J.J. CAPRI rounds 3-5 reveal promising successes and future challenges for RosettaDock. *Proteins.* **60**, 181-186 (2005).
59. Gray J.J., Moughon S., *et al.* Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J. Mol. Biol.* **331**, 281 -299 (2003).
60. Gray J.J., Moughon S.E., *et al.* Protein-protein docking predictions for the CAPRI experiment. *Proteins.* **52**, 118-122 (2003).
61. Kuhlman B. & Baker D. Native protein sequences are close to optimal for their structures. *Proc. Natl Acad. Sci. U.S.A.* **97**, 10383-10388 (2000).
62. Jmol: an open-source java viewer for chemical structures in 3d. http://www.jmol.org/, (2013).

