Journal of Visualized Experiments

Protein WISDOM: A Workbench for In Silico De novo design Of bioMolecules

--Manuscript Draft--

Manuscript Number:	JoVE50476R2
Article Type:	Invited Methods Article - JoVE Produced Video
Corresponding Author:	James Smadbeck Princeton University Princeton, NJ UNITED STATES
First Author:	James Smadbeck
Order of Authors:	James Smadbeck
	Meghan B. Peterson
	George A. Khoury
	Marty S. Taylor
	Christodoulos A. Floudas

Dear Editor,

Thank you for the consideration of our manuscript entitled "Protein WISDOM: A Workbench for In Silico De novo design Of bioMolecules" for the Journal of Visualized Experiments. 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 multimeric proteins for increased binding affinity. For the dissemination of these methods for broader use we present Protein WISDOM (http://www.proteinwisdom.org), a tool that provides automated design methods for a variety of protein design problems. Structural templates are submitted to initialize the design process. The first stage of the methods is an optimization sequence selection stage that aims at stability through minimization of free 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 design structures, provides the user with comprehensive quantitative assessment of the design.

Please consider the following scientists qualified to act as reviewers:

- **Yang Zhang** (zhng@umich.edu), Department of Biological Chemistry, Palmer Commons, University of Michigan, Ann Arbor, MI 48104
- **Jeffrey Skolnick** (skolnick@gatech.edu), Center for the Study of Systems Biology, School of Biology, Georgia Institute of Technology, Atlanta, GA 30332
- **Costas Maranas** (costas@psu.edu), Department of Chemical Engineering, Pennsylvania State University, Fenske Laboratory, University Park, PA 16802
- **David Baker** (dabaker@uw.edu), Department of Biochemistry, University of Washington, Molecular Engineering & Sciences, 3946 W Stevens Wy NE, Seattle, WA 98195
- **Ryan H. Lilien** (ryan.lilien@utoronto.ca), Department of Computer Science, University of Toronto, 10 Kings College Rd., Toronto, Ontario M5S-3G4, Canada
- **Dimitrios Morikis** (dimitrios.morikis@ucr.edu), Department of Bioengineering, University of California, Riverside, Materials Science & Engineering, Riverside, CA 92521

We appreciate your time and consideration of this manuscript for publication, and are looking forward to receiving your comments.

Sincerely,

Christodoulos A. Floudas
Stephen C. Macaleer '63 Professor in Engineering and Applied Science
Professor of Chemical and Biological Engineering
Department of Chemical and Biological Engineering
Princeton University
A325 Engineering Quad
Princeton, NJ 08544
P 609-258-4595 F 609-258-0211
floudas@titan.princeton.edu

Protein WISDOM: A Workbench for In Silico De novo design Of bioMolecules

James Smadbeck[1], Meghan B. Peterson[2], George A. Khoury[3], Marty S. Taylor[4], and Christodoulos A. Floudas[5]

[1] Department of Chemical and Biological Engineering, Princeton University, Princeton, NJ 08544, U.S.A., E-mail: jsmadbec@princeton.edu

[2] Department of Chemical and Biological Engineering, Princeton University, Princeton, NJ 08544, U.S.A., E-mail: meghan.peterson@sandia.gov

[3] Department of Chemical and Biological Engineering, Princeton University, Princeton, NJ 08544, U.S.A., E-mail: gkhoury@princeton.edu

[4] Department of Chemical and Biological Engineering, Princeton University, Princeton, NJ 08544, U.S.A., E-mail: mstaylor@jhmi.edu

[5] **Corresponding Author.** Department of Chemical and Biological Engineering, Princeton University, Princeton, NJ 08544, U.S.A., E-mail: floudas@titan.princeton.edu; Tel.: (609) 258-4595, Fax: (609) 258-0211

KEYWORDS: De novo protein and peptide design; Drug design; In silico sequence selection; Optimization; Fold specificity; Binding affinity

SHORT ABSTRACT

We developed computational de novo protein design methods capable of tackling several important areas of protein design. To disseminate these methods we present Protein WISDOM, an online tool for protein design (http://www.proteinwisdom.org). Starting from a structural template, design of monomeric proteins for increased stability and complexes for increased binding affinity can be performed.

LONG 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 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.

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.

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.

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	-	-

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.

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

Name	IC_{50}	EC_{50}	Protein Target	Applicable Diseases
SQ027	0.94 μM		human C3c	stroke, heart attack, Alzheimer's
SQ086	1.98 μM		human C3c	disease, asthma, rheumatoid
SQ059	4.73 μM	15.2 nM	human C3c	arthritis, systemic lupus
SQ110-4			C3aR	erythematosus, multiple
SQ060-4		36.4 nM	C3aR	sclerosis, psoriasis, diabetes
SQ007-5	15.4 nM		C3aR	type I, Crohn's disease,
SQ002-5	26.1 nM		C3aR	pancreatitis, and cystic fibrosis
SQ435	29 - 253 μM		HIV-1 gp41	AIDS
SQ037	13.57 μM		EZH2	prostate, breast, lymphoma,
SQ011-1	0.521 μM		LSD1	myeloma, bladder, colon,
SQ016-1	0.249 μM		LSD1	skin, liver, endometrial,
SQ026-1	2.51 μM		LSD2	lung, and gastric cancers

SQ015-1	1.332 µM	LSD2

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.

$$
\begin{aligned}
&\min_{y_i^j,\, y_k^l} && \sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{k=i+1}^{n}\sum_{l=1}^{m_k} E_{ik}^{jl}\left(x_i, x_j\right) w_{ik}^{jl} \\
&\text{Subject to} && \sum_{j=1}^{m_i} y_i^j = 1 \;\forall\; i \\
& && \sum_{j=1}^{m_i} w_{ik}^{jl} = y_k^l \;\;\forall\; i, k > i, l \\
& && \sum_{l=1}^{m_k} w_{ik}^{jl} = y_i^j \;\;\forall\; i, k > i, j \\
& && y_i^j, y_k^l, w_{ik}^{jl} \in \{0,1\} \;\forall\; i, j, k > i, l
\end{aligned}
\qquad (1)
$$

Set $i = 1,\ldots,n$ defines the residue positions in the design template. At each position i, mutations are represented by $j\{i\} = 1,\ldots,m_i$, 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 $E_{ik}^{jl}\left(x_i, x_k\right)$ which is the energy interaction between position i occupied by amino acid j and position k occupied by amino acid l, depends on the distance between the α-carbons or side chain centroids at the two positions $\left(x_i, x_k\right)$ 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 $3.4x10^{45}$ and the other at $6.4x10^{37}$ 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) seconds and 4,578 CPU seconds 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 seconds and 14 CPU seconds 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, $\sum_{d=1}^{b_m} E_{ik}^{jl}\left(x_i, x_k\right) wt\left(x_i, x_k, d\right)$, in place of the energy parameter $E_{ik}^{jl}\left(x_i, x_k\right)$ in the Single Structure Model (Eq. 1). The weights $wt\left(x_i, x_k, d\right)$ 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.

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

$$y_i^j, y_k^l, w_{ik}^{jl} \in \{0,1\} \ \forall \ i, j, k > i, l \qquad (2)$$

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, $E_{ik}^{jl}(x_i, x_k)$ 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.

$$\begin{aligned}
\min_{y_i^j, y_k^l} \quad & \sum_{i=1}^{n} \sum_{j=1}^{m_i} \sum_{k=i+1}^{n} \sum_{l=1}^{m_k} \sum_{d:disbin(x_i,x_k,d)=1}^{b_m} E_{ik}^{jl}(x_i, x_k) b_{ikd} w_{ik}^{jl} \\
\text{Subject to} \quad & \sum_{j=1}^{m_i} y_i^j = 1 \ \forall \ i \\
& \sum_{j=1}^{m_i} w_{ik}^{jl} = y_k^l \ \forall \ i, k > i, l \\
& \sum_{l=1}^{m_k} w_{ik}^{jl} = y_i^j \ \forall \ i, k > i, j \\
& \sum_{d:disbin(x_i,x_k,d)=1}^{b_m} b_{ikd} = 1 \ \forall \ i, k > i \\
& y_i^j, y_k^l, w_{ik}^{jl}, b_{ikd} \in \{0,1\} \ \forall \ i, j, k > i, l, d \qquad (3)
\end{aligned}$$

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).

$$f_{spec} = \frac{\sum_{i \in novel} e^{-\beta E_i}}{\sum_{i \in native} e^{-\beta E_i}} \qquad (4)$$

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.

$$K^* = \frac{q_{PL}}{q_P q_L} \qquad (5)$$

Here q_{PL} is the partition function of the protein-ligand complex, q_P 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.

$$q_{PL} = \sum_{b \in B} e^{-\frac{E_b}{RT}},\ q_{PL} = \sum_{b \in B} e^{-\frac{E_b}{RT}},\ q_{PL} = \sum_{b \in B} e^{-\frac{E_b}{RT}} \qquad (6)$$

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, 2000 rotamer conformations are generated for the single starting structure, so the final ensemble size is 2000 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.

1) User Registration

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

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

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

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

2) Stage One: Sequence Selection

2.1) Submission of Protein Sequence and Template Structure(s)

2.1.1) Click on the User Login button to begin the protein design experiment. The user is presented with their "User Homepage" (Fig. 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.

2.1.2) Start a new design job by clicking "Create New Job." The user is taken to the "Job Submission" page (Fig. 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."

2.1.3) Upload the protein structure(s) of the design template (Fig. 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."

2.1.4) 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."

2.1.5) Once the template has been successfully uploaded and confirmed, the user is taken to the "Main Control Page" (Fig. 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.

2.2) Selection of Mutation Sets

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

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

2.2.3) 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."

2.3) Selection of Biological Constraints

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

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

2.3.3) 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.

2.4) Submission of Stage One: Sequence Selection

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

2.4.2) Select the chain to design (Fig. 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.

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

2.4.4) 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.

3) Stage Two: Fold Specificity Calculations

3.1) Fold Specificity Submission

3.1.1) 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.

3.1.2) Click the "Submit" button.

3.1.3) 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.

3.1.4) 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.

3.1.5) 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" (Fig. 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.

4) Stage Three: Approximate Binding Affinity Calculations for Protein-Peptide Complexes

4.1) 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.

4.2) 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."

4.3) 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 (Fig. 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 (Fig. 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 (Fig. 8). Here further design constraints can be set and the Stage One Sequence Selection can be started. Entering the Mutation Sets section (Fig. 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 (Fig. 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" (Fig. 14), the "Sequence Results" (Fig. 15), or the "Energy Results" (Fig. 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 (Fig. 17) or individually in the "Fold Specificity Results" section (Fig. 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 (Fig. 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 (Fig. 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] (Fig. 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 (Fig. 22). This section provides the values for the peptide, protein, and complex partition functions along with the final Approximate Binding Affinity Value.

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.

ACKNOWLEDGMENTS

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.

DISCLOSURES

The authors declare that they have no competing financial interests.

FIGURE LEGENDS

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.
25. Bellows M. L., Fung H. K., Floudas C. A. in Molecular Systems Engineering, Process Systems Engineering, eds Adjiman C. S., Galindo A. (Wiley-VCH Verlag GmbH & Co. KGaA) Vol. 6, pp 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. Mönnigmann 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 Å C_α 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/*

Figure 1
Click here to download high resolution image

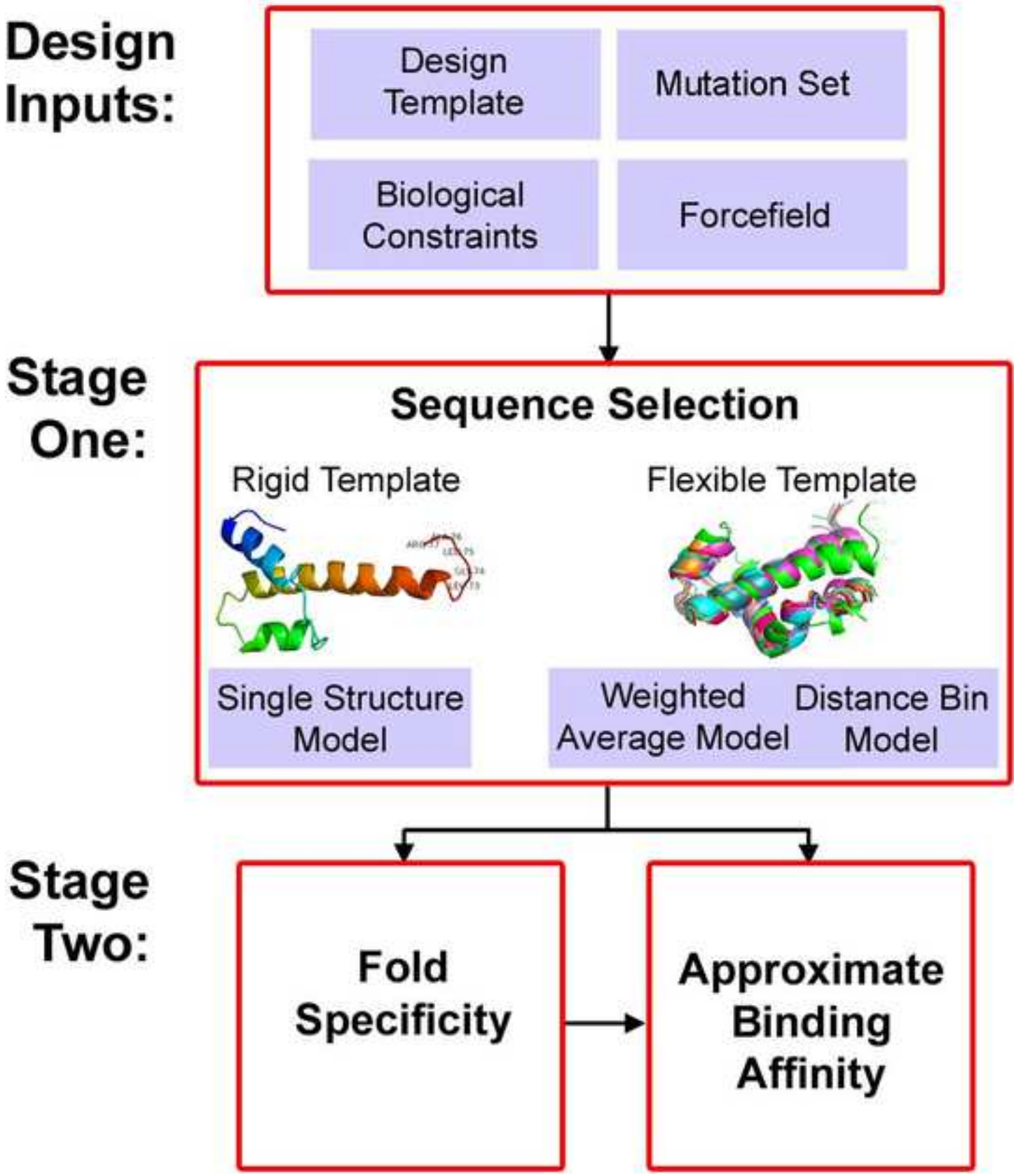


Figure 2
Click here to download high resolution image

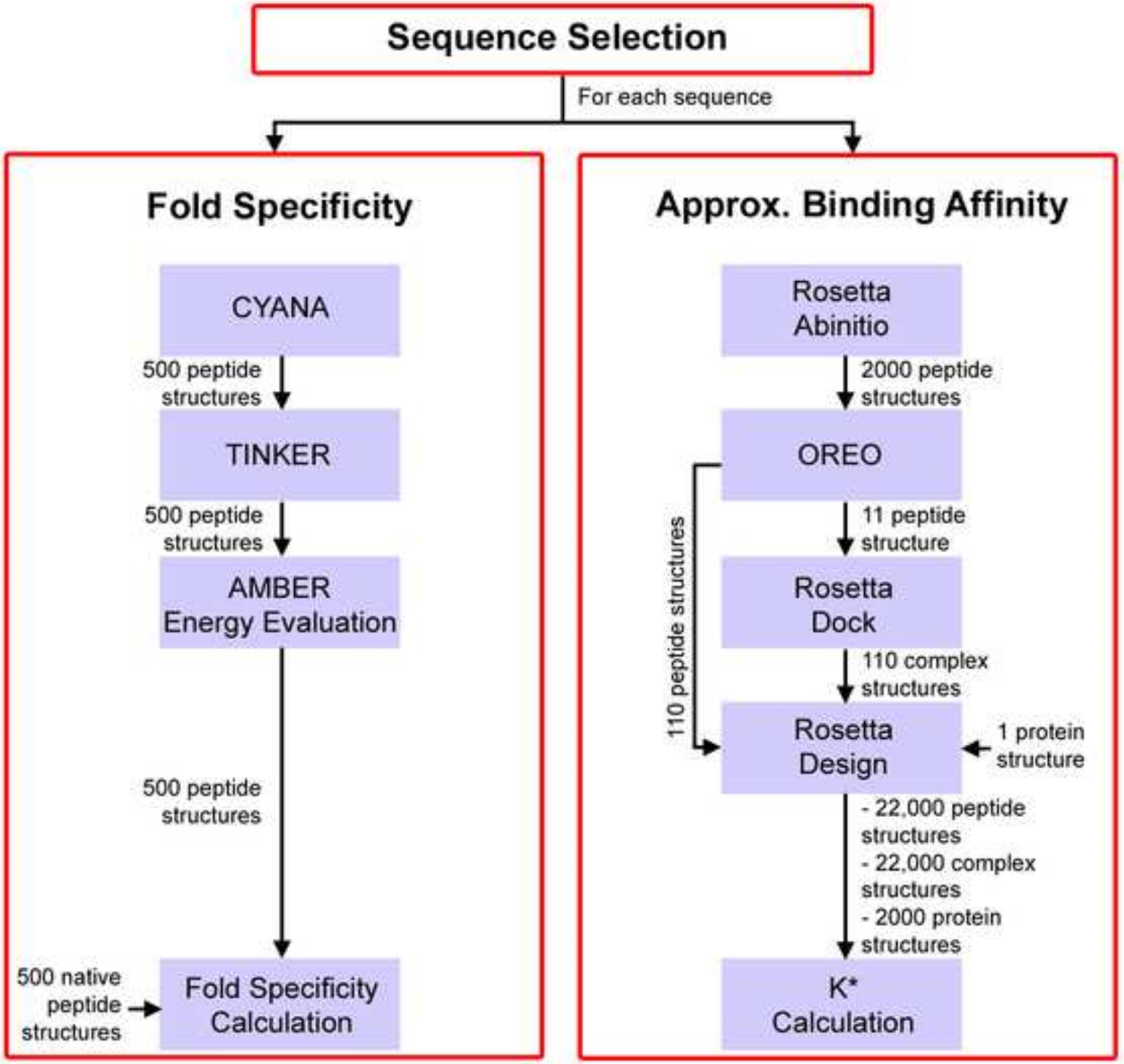
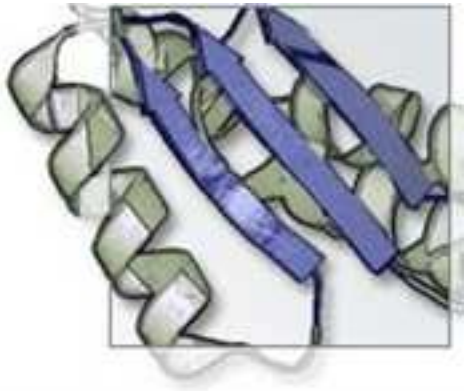


Figure 3
Click here to download high resolution image



protein WISDOM

Workbench for In Silico De novo design Of bioMolecules

User Section | Your Jobs | Create New Job | Queue Status | Logout

C.

User Homepage for James Smadbeck (jamsmad)

Your Jobs: You have submitted 28 jobs. Click here to view them. A.

Your Structures: You have submitted 13 structures.

B.

PDB ID	Description
1DLH	Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptid
3OVJ	Structure of an amyloid forming peptide KLVFFA from amyloid beta
N/A	EZH2 SQ037 cd peptide
1GZL	Crystal structure of C14LINKMID/IQN17: a cross-linked inhibitor of HIV-1 entry bound to gp41
3OOI	Crystal Structure of Human Histone-Lysine N-methyltransferase NSD1 SET domain in Complex with S-aden

Figure 4
Click here to download high resolution image

Figure 5
Click here to download high resolution image

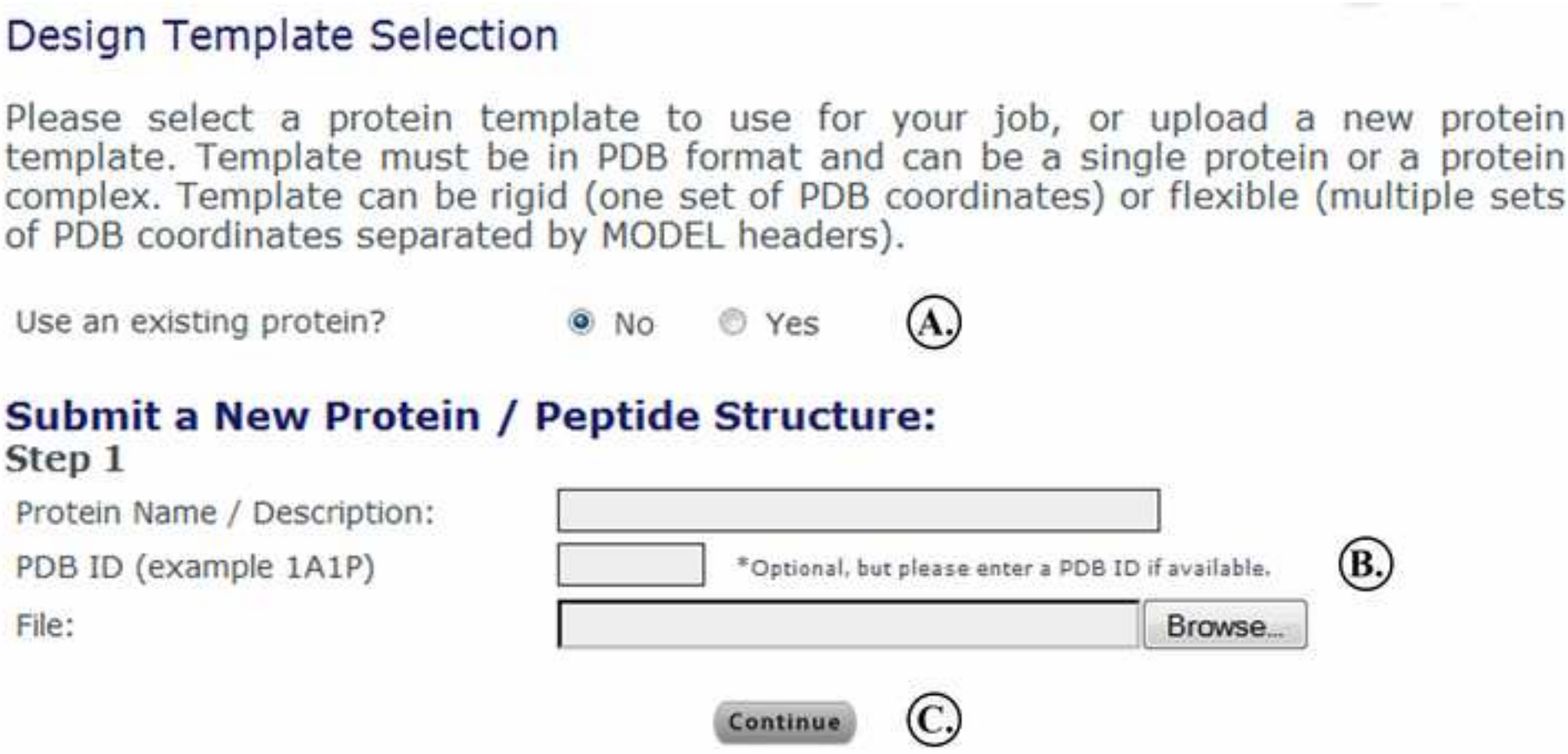


Figure 6
Click here to download high resolution image

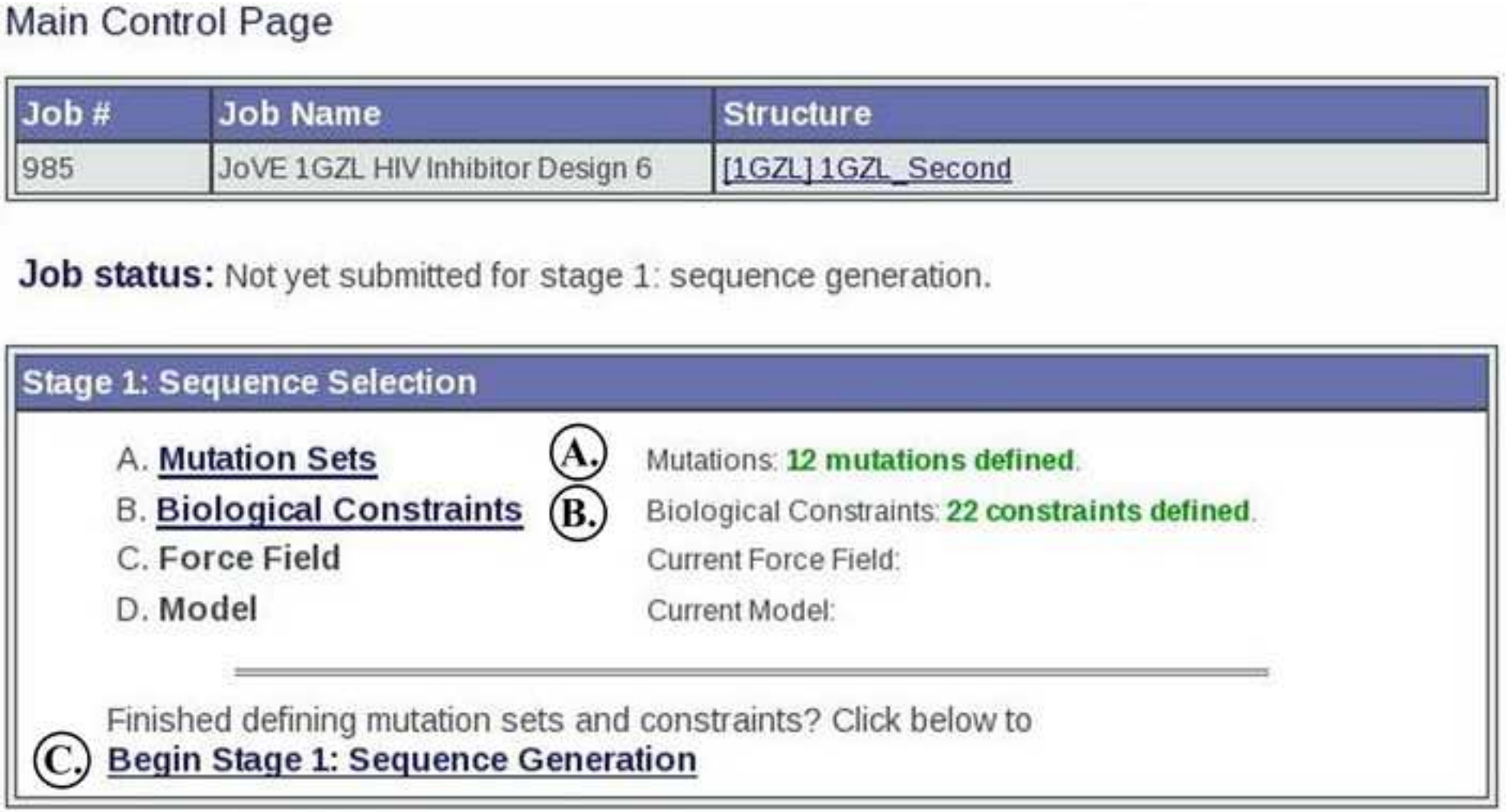


Figure 7
Click here to download high resolution image

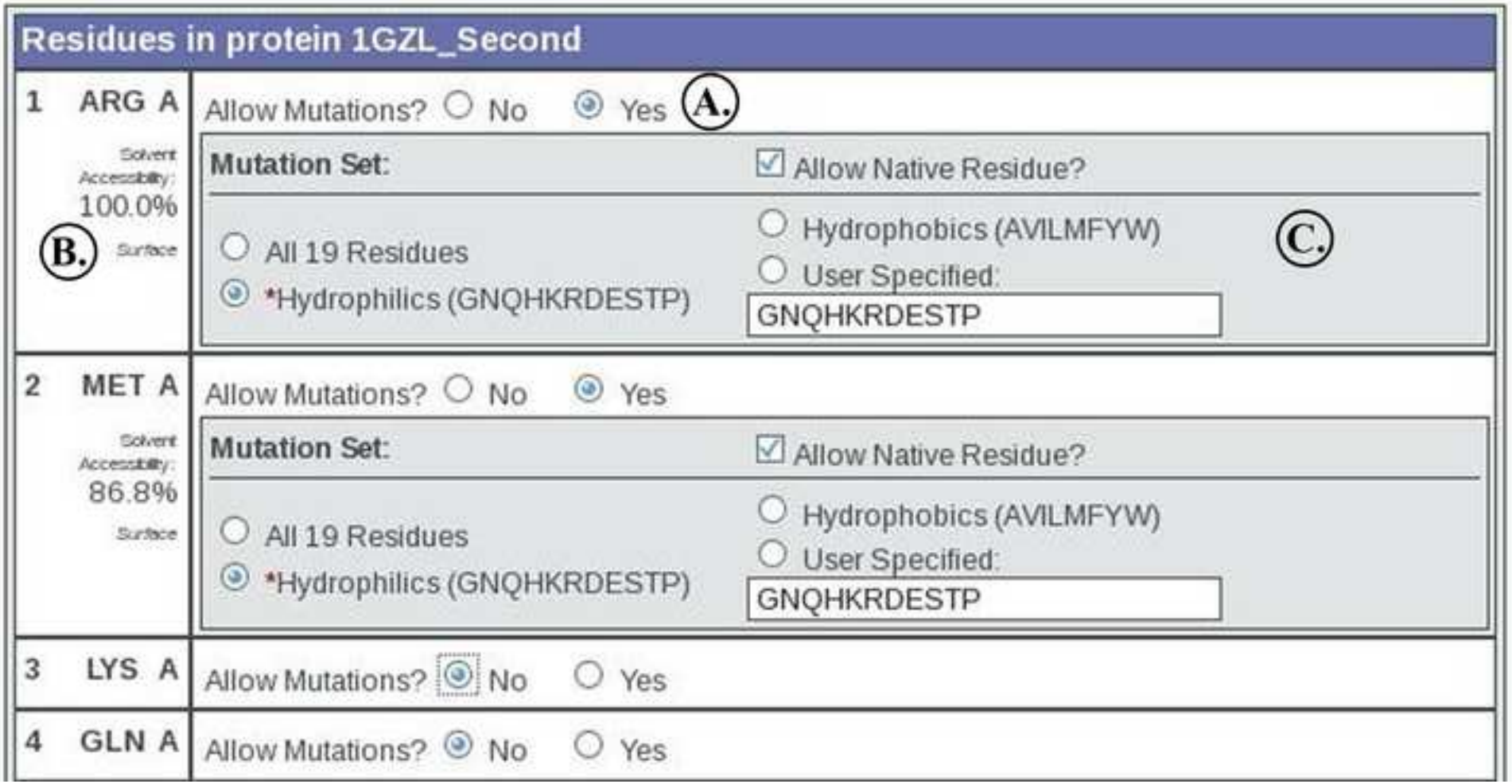


Figure 8
Click here to download high resolution image

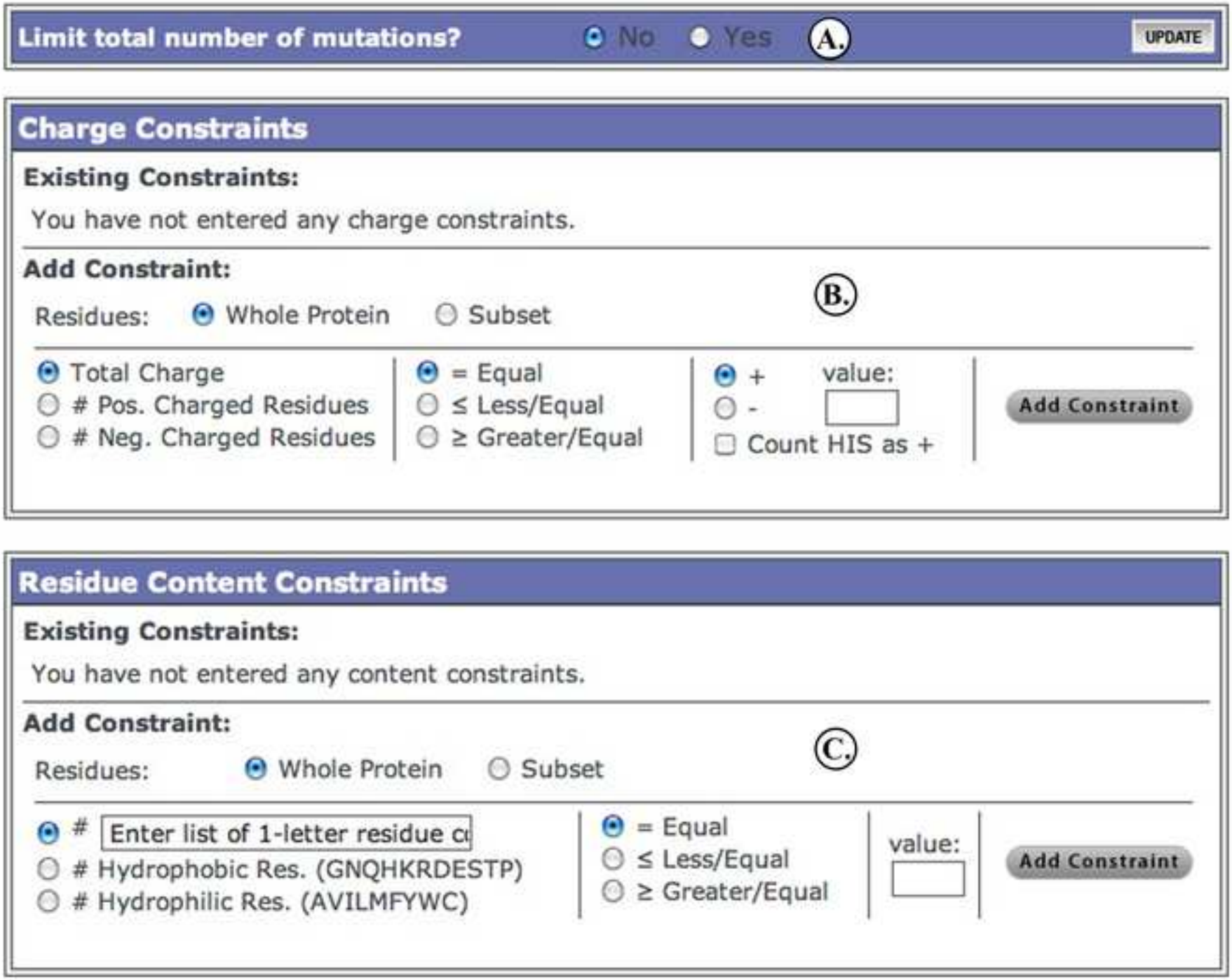


Figure 9
Click here to download high resolution image

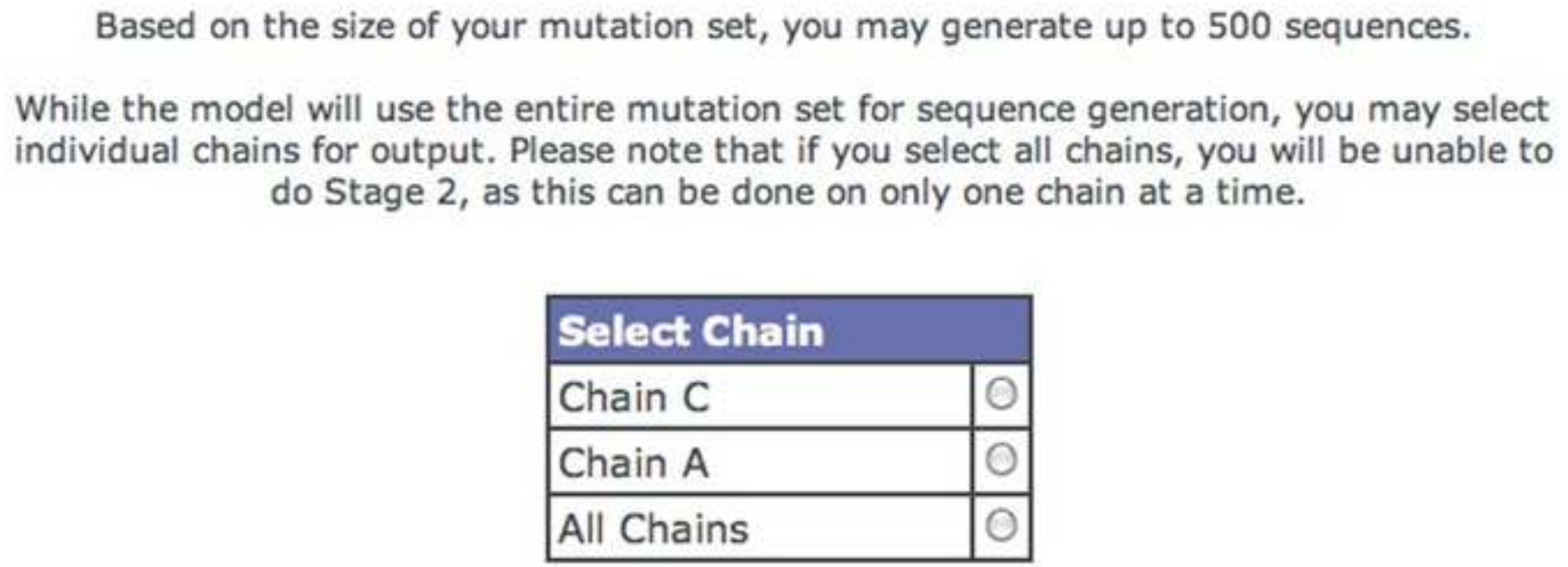


Figure 10
Click here to download high resolution image

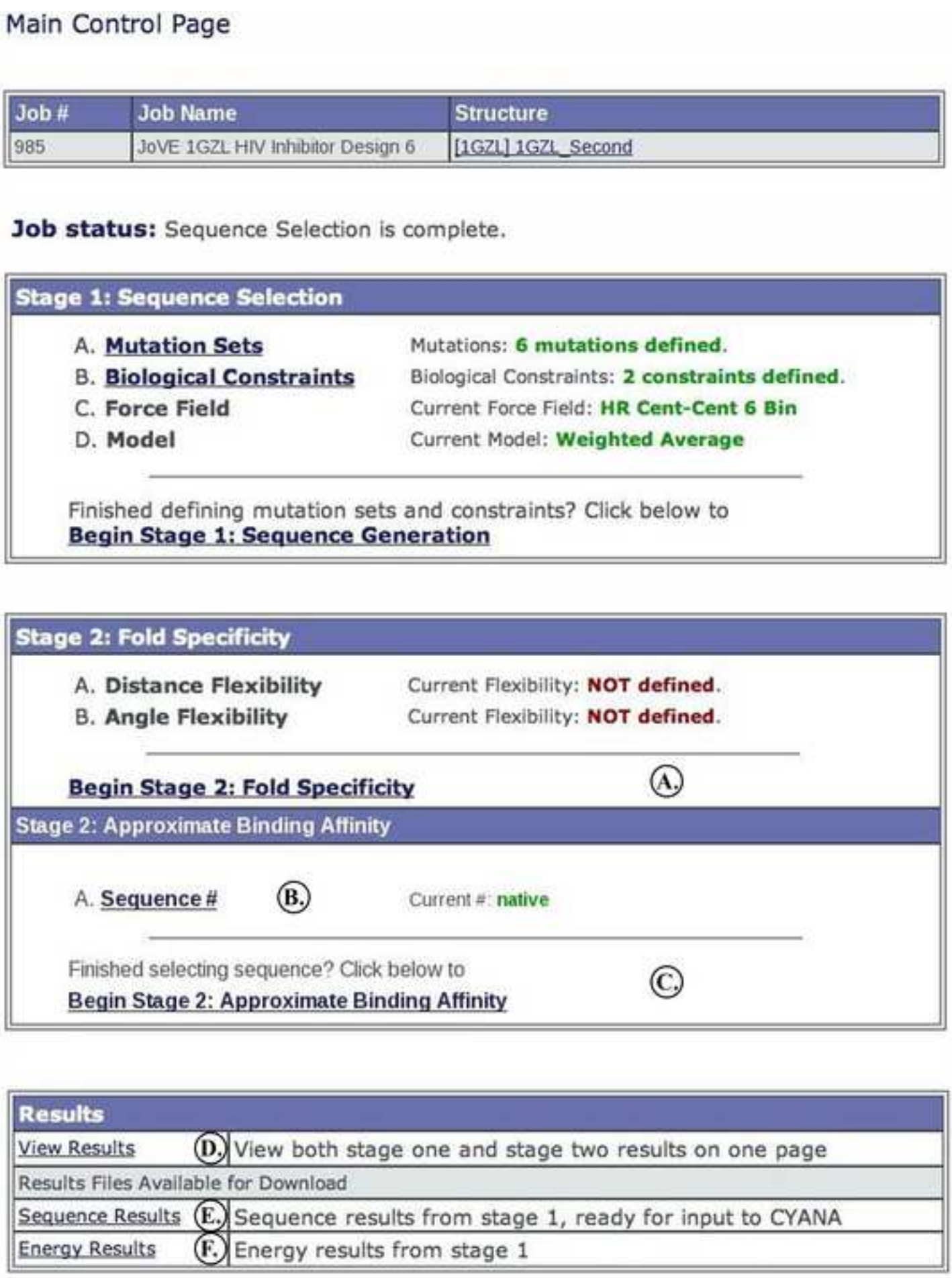


Figure 11
Click here to download high resolution image

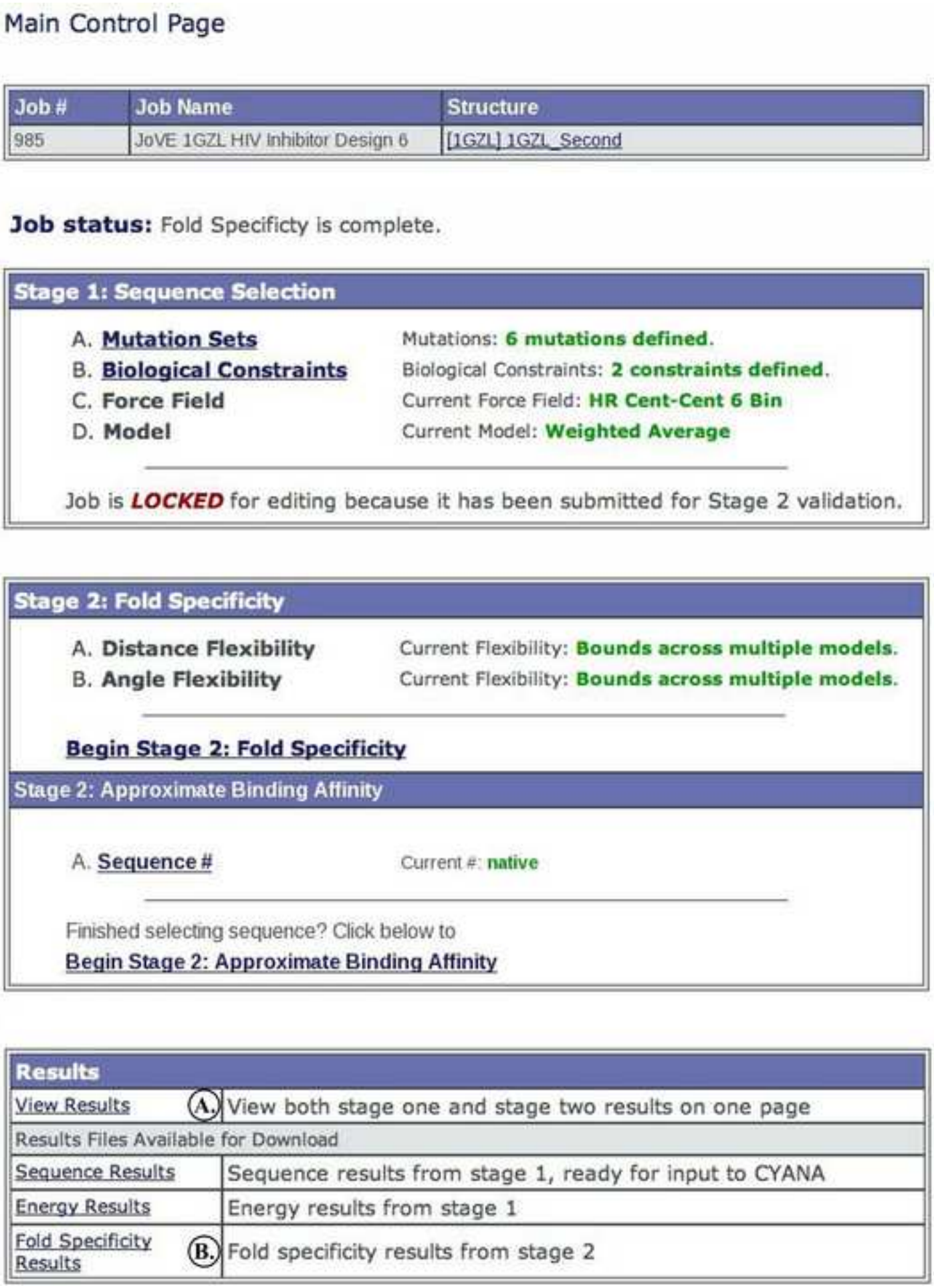


Figure 12
Click here to download high resolution image

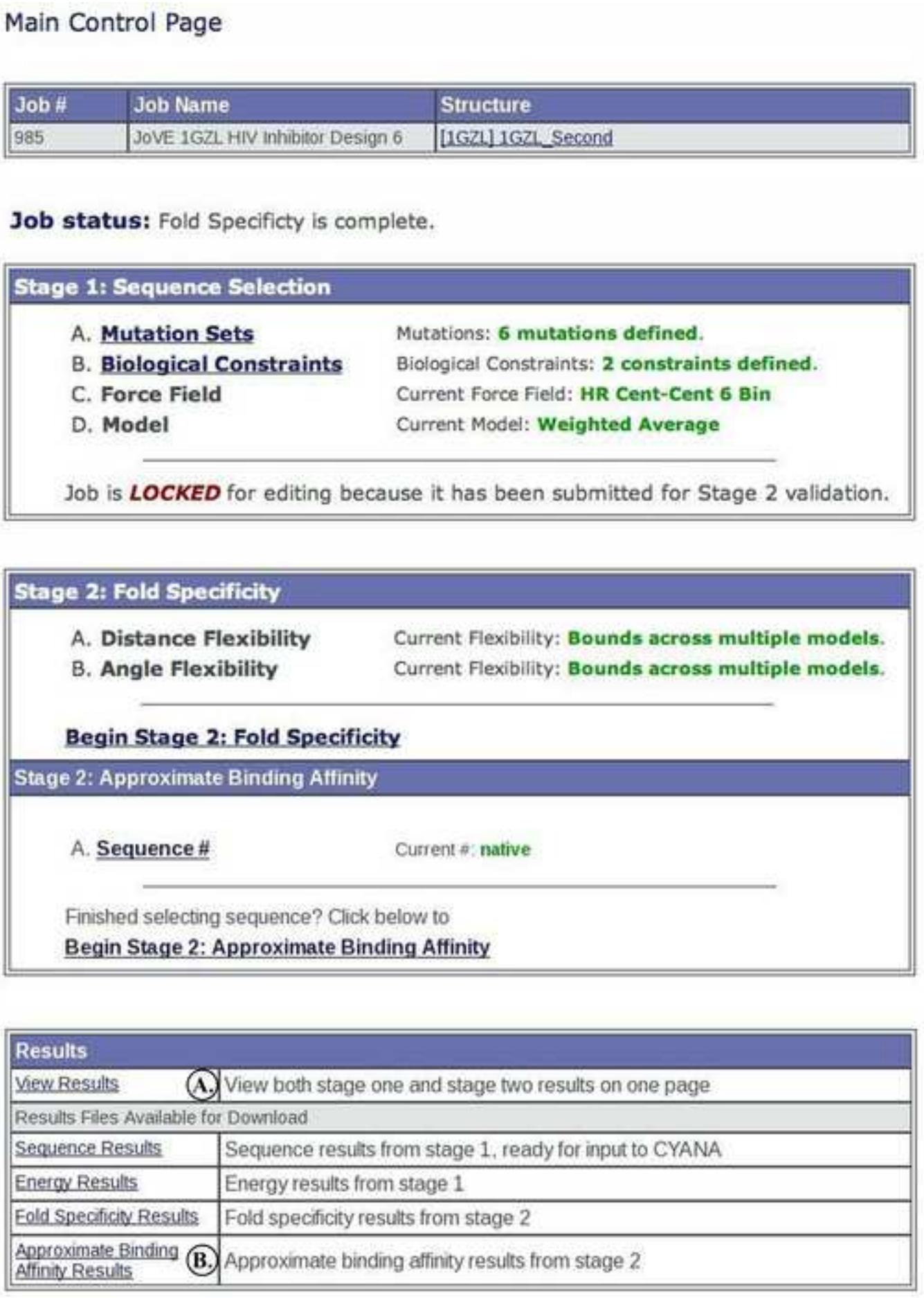


Figure 13
Click here to download high resolution image

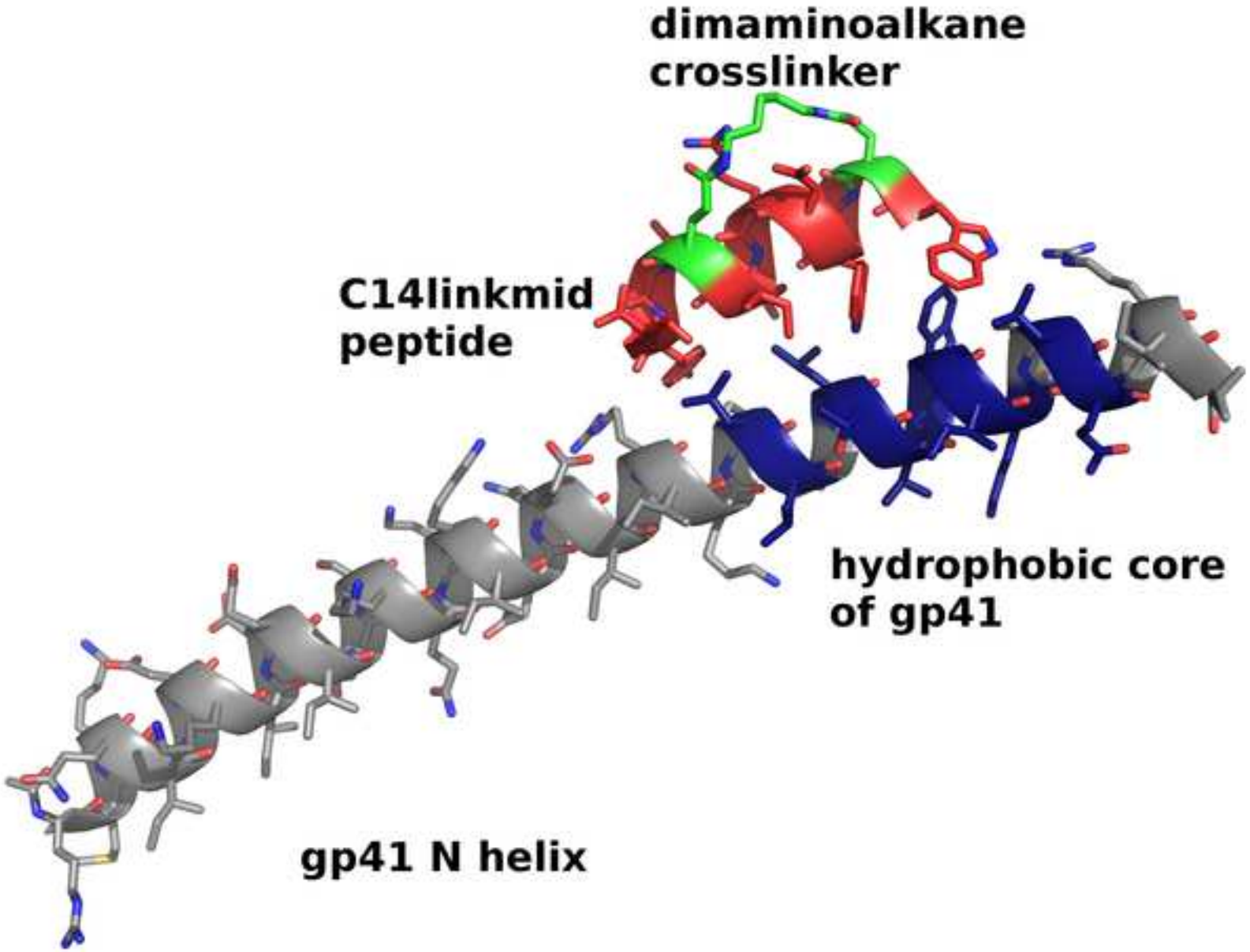


Figure 14
Click here to download high resolution image

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.

A. B. C.

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 15
Click here to download high resolution image

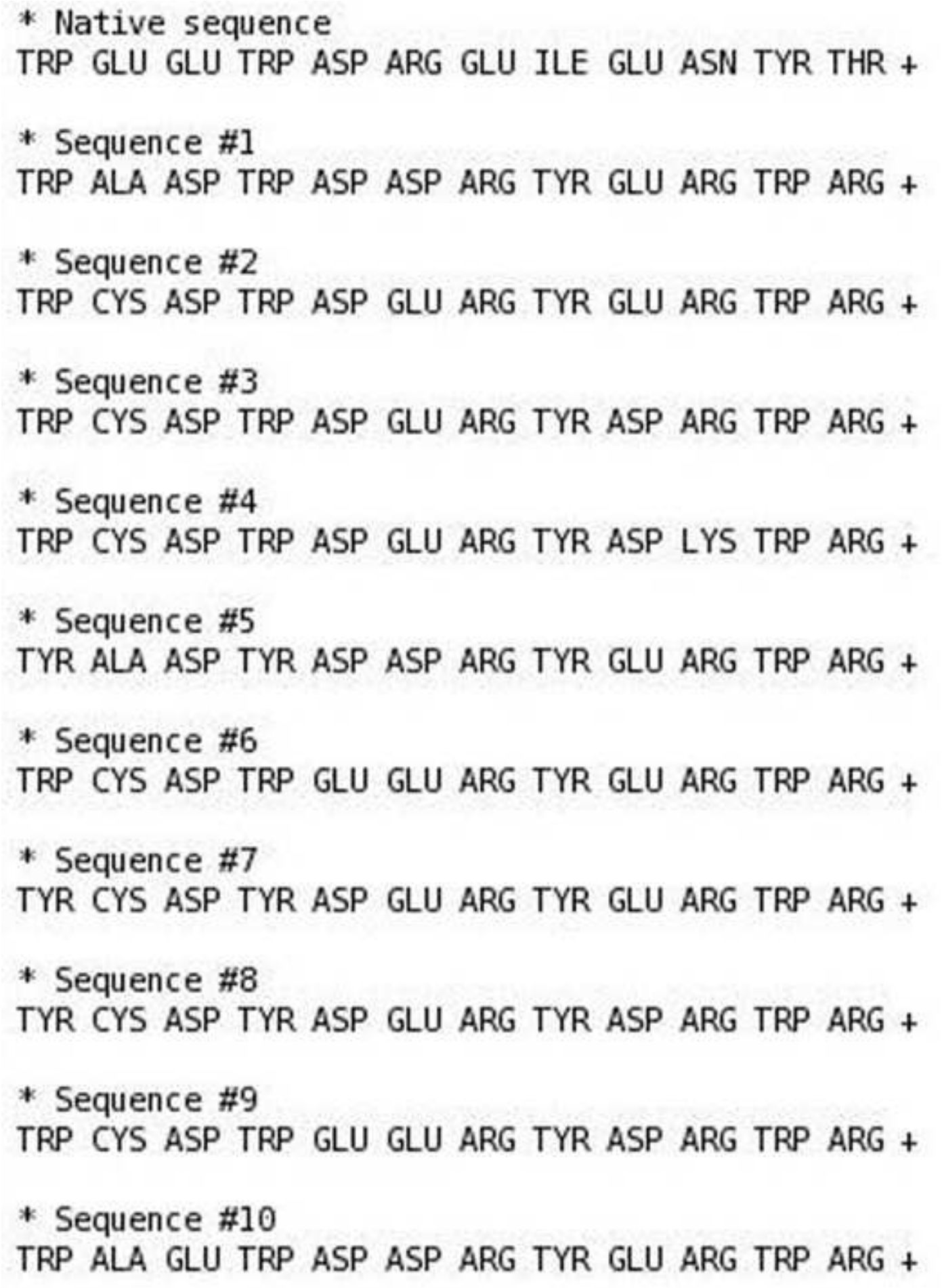


Figure 16
Click here to download high resolution image

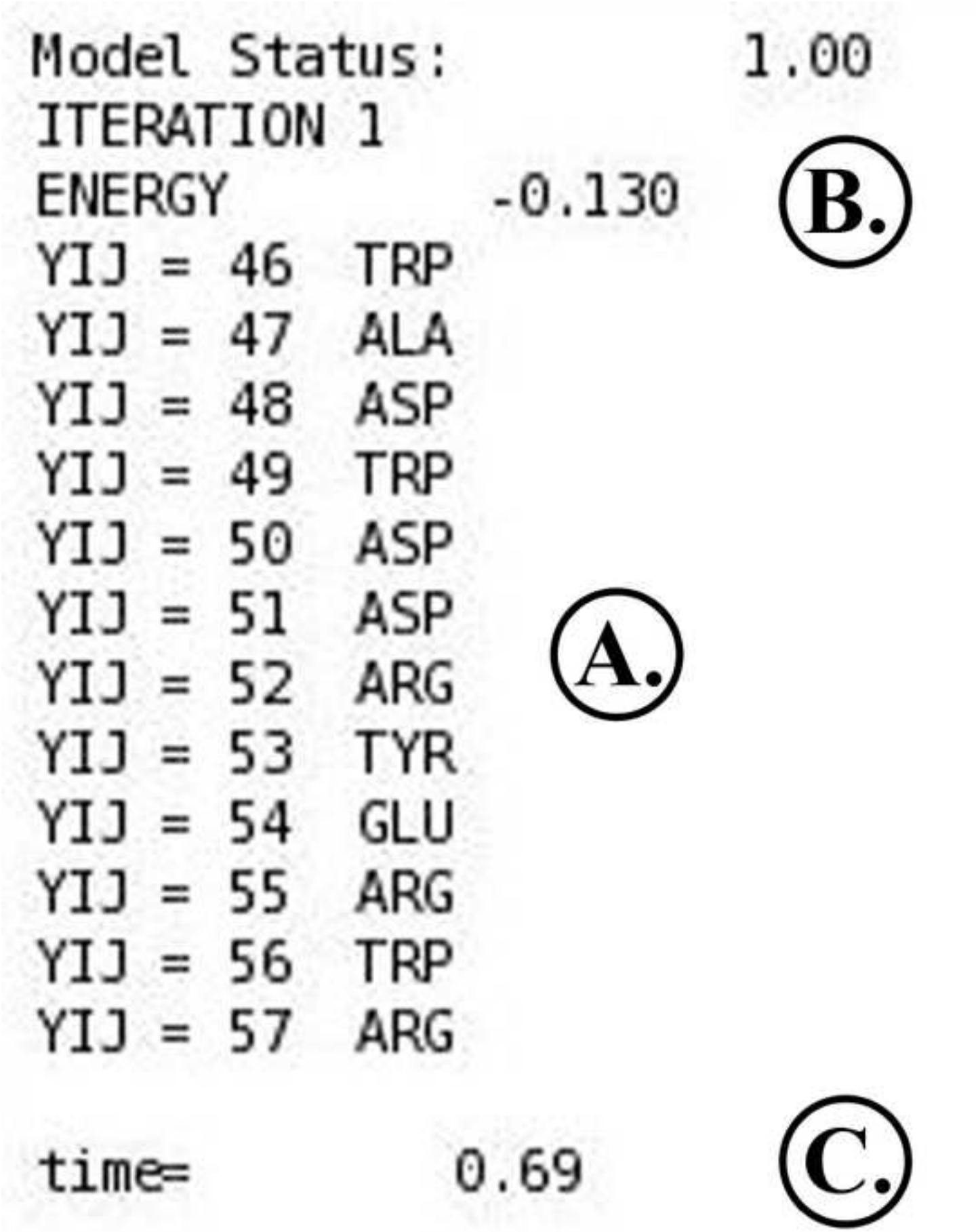


Figure 17
Click here to download high resolution image

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.

A. B.

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 19
Click here to download high resolution image

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.

A. B. C.

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 18
Click here to download high resolution image

Sequence #	Fold Specificity
1	598.91429566
2	508.40415652
3	596.75267706
4	36.70253639
5	1298.95864153
6	487.21592754
7	1190.84602260
8	1268.48498724
9	532.77845376
10	493.77987025
11	600.76236923
12	83.41034332
13	35.42400190
14	458.14358557
15	38.11489321
16	555.86426817
17	1040.62843257
18	549.64143335
19	38.93603645
20	847.06403799

Figure 20
Click here to download high resolution image

Design Information

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

A.

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

B.

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

C.

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

D.

Figure 21
Click here to download high resolution image

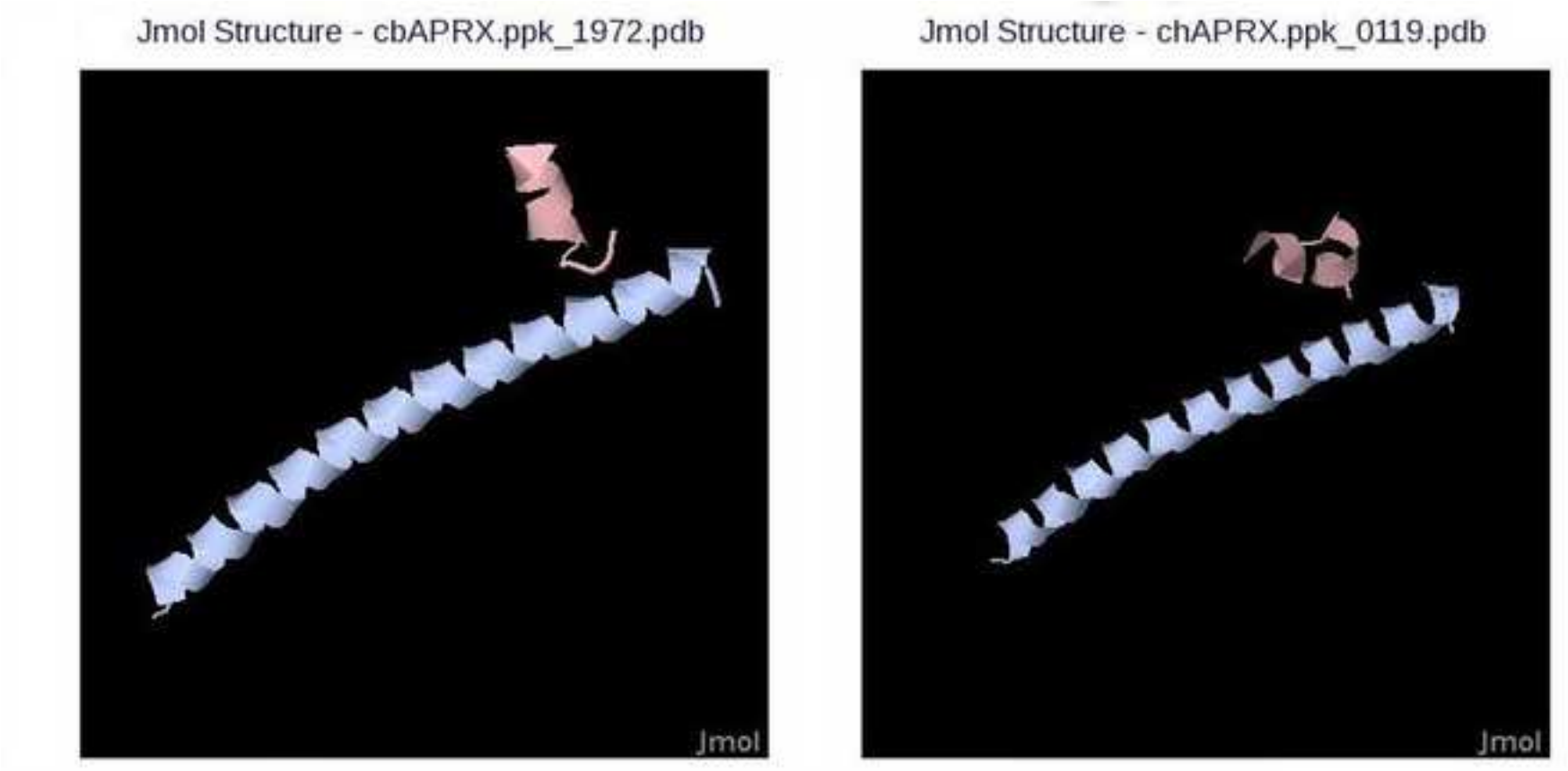


Figure 22
Click here to download high resolution image

Sequence #	qPL	qL	qP	K*	offset		
native	3.452417977032 57e+67		6.85156922691794e+18		1.10830233013821e+49	4.55e-01	0
439	1.289373450441e+67		3.57859328300014e+17		1.12658990768818e+49	3.20e+00	0

Table of Reagents/ Materials Used

Click here to download Table of Reagents/ Materials Used: JoVE_Materials.xlsx

Name of Reagent/Material	Company	Catalog Number	Comments

17 Sellers Street
Cambridge, MA 02139
tel. +1.617.945.9051
www.JoVE.com

ARTICLE AND VIDEO LICENSE AGREEMENT

Title of Article: _____

Author(s): _____

Item 1 (check one box): The Author elects to have the Materials be made available (as described at http://www.jove.com/publish) via: ☐ Standard Access ☐ Open Access

Item 2 (check one box):

☐ The Author is NOT a United States government employee.

☐ The Author is a United States government employee and the Materials were prepared in the course of his or her duties as a United States government employee.

☐ The Author is a United States government employee but the Materials were NOT prepared in the course of his or her duties as a United States government employee.

ARTICLE AND VIDEO LICENSE AGREEMENT

1. Defined Terms. As used in this Article and Video License Agreement, the following terms shall have the following meanings: "**Agreement**" means this Article and Video License Agreement; "**Article**" means the article specified on the last page of this Agreement, including any associated materials such as texts, figures, tables, artwork, abstracts, or summaries contained therein; "**Author**" means the author who is a signatory to this Agreement; "**Collective Work**" means a work, such as a periodical issue, anthology or encyclopedia, in which the Materials in their entirety in unmodified form, along with a number of other contributions, constituting separate and independent works in themselves, are assembled into a collective whole; "**CRC License**" means the Creative Commons Attribution-Non Commercial-No Derivs 3.0 Unported Agreement, the terms and conditions of which can be found at: http://creativecommons.org/licenses/by-nc-nd/3.0/legalcode; "**Derivative Work**" means a work based upon the Materials or upon the Materials and other pre-existing works, such as a translation, musical arrangement, dramatization, fictionalization, motion picture version, sound recording, art reproduction, abridgment, condensation, or any other form in which the Materials may be recast, transformed, or adapted; "**Institution**" means the institution, listed on the last page of this Agreement, by which the Author was employed at the time of the creation of the Materials; "**JoVE**" means MyJove Corporation, a Massachusetts corporation and the publisher of *The Journal of Visualized Experiments;* "**Materials**" means the Article and / or the Video; "**Parties**" means the Author and JoVE; "**Video**" means any video(s) made by the Author, alone or in conjunction with any other parties, or by JoVE or its affiliates or agents, individually or in collaboration with the Author or any other parties, incorporating all or any portion of the Article, and in which the Author may or may not appear.

2. Background. The Author, who is the author of the Article, in order to ensure the dissemination and protection of the Article, desires to have the JoVE publish the Article and create and transmit videos based on the Article. In furtherance of such goals, the Parties desire to memorialize in this Agreement the respective rights of each Party in and to the Article and the Video.

3. Grant of Rights in Article. In consideration of JoVE agreeing to publish the Article, the Author hereby grants to JoVE, subject to **Sections 4** and **7** below, the exclusive, royalty-free, perpetual (for the full term of copyright in the Article, including any extensions thereto) license (a) to publish, reproduce, distribute, display and store the Article in all forms, formats and media whether now known or hereafter developed (including without limitation in print, digital and electronic form) throughout the world, (b) to translate the Article into other languages, create adaptations, summaries or extracts of the Article or other Derivative Works (including, without limitation, the Video) or Collective Works based on all or any portion of the Article and exercise all of the rights set forth in (a) above in such translations, adaptations, summaries, extracts, Derivative Works or Collective Works and (c) to license others to do any or all of the above. The foregoing rights may be exercised in all media and formats, whether now known or hereafter devised, and include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. If the "Open Access" box has been checked in **Item 1** above, JoVE and the Author hereby grant to the public all such rights in the Article as provided in, but subject to all limitations and requirements set forth in, the CRC License.

4. Retention of Rights in Article. Notwithstanding the exclusive license granted to JoVE in **Section 3** above, the

Author shall, with respect to the Article, retain the non-exclusive right to use all or part of the Article for the non-commercial purpose of giving lectures, presentations or teaching classes, and to post a copy of the Article on the Institution's website or the Author's personal website, in each case provided that a link to the Article on the JoVE website is provided and notice of JoVE's copyright in the Article is included. All non-copyright intellectual property rights in and to the Article, such as patent rights, shall remain with the Author.

5. Grant of Rights in Video – Standard Access. This **Section 5** applies if the "Standard Access" box has been checked in **Item 1** above or if no box has been checked in **Item 1** above. In consideration of JoVE agreeing to produce, display or otherwise assist with the Video, the Author hereby acknowledges and agrees that, Subject to **Section 7** below, JoVE is and shall be the sole and exclusive owner of all rights of any nature, including, without limitation, all copyrights, in and to the Video. To the extent that, by law, the Author is deemed, now or at any time in the future, to have any rights of any nature in or to the Video, the Author hereby disclaims all such rights and transfers all such rights to JoVE.

6. Grant of Rights in Video – Open Access. This **Section 6** applies only if the "Open Access" box has been checked in **Item 1** above. In consideration of JoVE agreeing to produce, display or otherwise assist with the Video, the Author hereby grants to JoVE, subject to **Section 7** below, the exclusive, royalty-free, perpetual (for the full term of copyright in the Article, including any extensions thereto) license (a) to publish, reproduce, distribute, display and store the Video in all forms, formats and media whether now known or hereafter developed (including without limitation in print, digital and electronic form) throughout the world, (b) to translate the Video into other languages, create adaptations, summaries or extracts of the Video or other Derivative Works or Collective Works based on all or any portion of the Video and exercise all of the rights set forth in (a) above in such translations, adaptations, summaries, extracts, Derivative Works or Collective Works and (c) to license others to do any or all of the above. The foregoing rights may be exercised in all media and formats, whether now known or hereafter devised, and include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. For any Video to which this Section 6 is applicable, JoVE and the Author hereby grant to the public all such rights in the Video as provided in, but subject to all limitations and requirements set forth in, the CRC License.

7. Government Employees. If the Author is a United States government employee and the Article was prepared in the course of his or her duties as a United States government employee, as indicated in **Item 2** above, and any of the licenses or grants granted by the Author hereunder exceed the scope of the 17 U.S.C. 403, then the rights granted hereunder shall be limited to the maximum rights permitted under such statute. In such case, all provisions contained herein that are not in conflict with such statute shall remain in full force and effect, and all provisions contained herein that do so conflict shall be deemed to be amended so as to provide to JoVE the maximum rights permissible within such statute.

8. Likeness, Privacy, Personality. The Author hereby grants JoVE the right to use the Author's name, voice, likeness, picture, photograph, image, biography and performance in any way, commercial or otherwise, in connection with the Materials and the sale, promotion and distribution thereof. The Author hereby waives any and all rights he or she may have, relating to his or her appearance in the Video or otherwise relating to the Materials, under all applicable privacy, likeness, personality or similar laws.

9. Author Warranties. The Author represents and warrants that the Article is original, that it has not been published, that the copyright interest is owned by the Author (or, if more than one author is listed at the beginning of this Agreement, by such authors collectively) and has not been assigned, licensed, or otherwise transferred to any other party. The Author represents and warrants that the author(s) listed at the top of this Agreement are the only authors of the Materials. If more than one author is listed at the top of this Agreement and if any such author has not entered into a separate Article and Video License Agreement with JoVE relating to the Materials, the Author represents and warrants that the Author has been authorized by each of the other such authors to execute this Agreement on his or her behalf and to bind him or her with respect to the terms of this Agreement as if each of them had been a party hereto as an Author. The Author warrants that the use, reproduction, distribution, public or private performance or display, and/or modification of all or any portion of the Materials does not and will not violate, infringe and/or misappropriate the patent, trademark, intellectual property or other rights of any third party. The Author represents and warrants that it has and will continue to comply with all government, institutional and other regulations, including, without limitation all institutional, laboratory, hospital, ethical, human and animal treatment, privacy, and all other rules, regulations, laws, procedures or guidelines, applicable to the Materials, and that all research involving human and animal subjects has been approved by the Author's relevant institutional review board.

10. JoVE Discretion. If the Author requests the assistance of JoVE in producing the Video in the Author's facility, the Author shall ensure that the presence of JoVE employees, agents or independent contractors is in accordance with the relevant regulations of the Author's institution. If more than one author is listed at the beginning of this Agreement, JoVE may, in its sole discretion, elect not take any action with respect to the Article until such time as it has received complete, executed Article and Video License Agreements from each such author. JoVE reserves the right, in its absolute and sole discretion and without giving any reason therefore, to accept or decline any work submitted to JoVE. JoVE and its employees, agents and independent contractors shall have full, unfettered access to the facilities of the Author or of the Author's institution as necessary to make the Video, whether actually published or not. JoVE has sole discretion as to the method of making and publishing the Materials, including,

without limitation, to all decisions regarding editing, lighting, filming, timing of publication, if any, length, quality, content and the like.

11. Indemnification. The Author agrees to indemnify JoVE and/or its successors and assigns from and against any and all claims, costs, and expenses, including attorney's fees, arising out of any breach of any warranty or other representations contained herein. The Author further agrees to indemnify and hold harmless JoVE from and against any and all claims, costs, and expenses, including attorney's fees, resulting from the breach by the Author of any representation or warranty contained herein or from allegations or instances of violation of intellectual property rights, damage to the Author's or the Author's institution's facilities, fraud, libel, defamation, research, equipment, experiments, property damage, personal injury, violations of institutional, laboratory, hospital, ethical, human and animal treatment, privacy or other rules, regulations, laws, procedures or guidelines, liabilities and other losses or damages related in any way to the submission of work to JoVE, making of videos by JoVE, or publication in *JoVE* or elsewhere by JoVE. The Author shall be responsible for, and shall hold JoVE harmless from, damages caused by lack of sterilization, lack of cleanliness or by contamination due to the making of a video by JoVE its employees, agents or independent contractors. All sterilization, cleanliness or decontamination procedures shall be solely the responsibility of the Author and shall be undertaken at the Author's expense. All indemnifications provided herein shall include JoVE's attorney's fees and costs related to said losses or damages. Such indemnification and holding harmless shall include such losses or damages incurred by, or in connection with, acts or omissions of JoVE, its employees, agents or independent contractors.

12. Fees. To cover the cost incurred for publication, JoVE must receive payment before production and publication the Materials. Payment is due in 21 days of invoice. Should the Materials not be published due to an editorial or production decision, these funds will be returned to the Author. Withdrawal by the Author of any submitted Materials after final peer review approval will result in a US$1,200 fee to cover pre-production expenses incurred by JoVE. If payment is not received by the completion of filming, production and publication of the Materials will be suspended until payment is received.

13. Transfer, Governing Law. This Agreement may be assigned by JoVE and shall inure to the benefits of any of JoVE's successors and assignees. This Agreement shall be governed and construed by the internal laws of the Commonwealth of Massachusetts without giving effect to any conflict of law provision thereunder. This Agreement may be executed in counterparts, each of which shall be deemed an original, but all of which together shall be deemed to me one and the same agreement. A signed copy of this Agreement delivered by facsimile, e-mail or other means of electronic transmission shall be deemed to have the same legal effect as delivery of an original signed copy of this Agreement.

A signed copy of this document must be sent with all new submissions. Only one Agreement required per submission.

AUTHOR:

Name:

Department:

Institution:

Article Title:

Signature: Date:

Please submit a signed and dated copy of this license by one of the following three methods:

1) Upload a scanned copy as a PDF to the *JoVE* submission site upon manuscript submission (preferred);
2) Fax the document to +1.866.381.2236; or
3) Mail the document to *JoVE* / Attn: *JoVE* Editorial / 17 Sellers St / Cambridge, MA 02139

For questions, please email editorial@jove.com or call +1.617.945.9051.

MS # (internal use):

Reviewer's Comments

Dear Editors,

Thank you very much for the editorial and reviewer comments made on our manuscript. We hope the changes we have made in response to the comments are sufficient and that our manuscript will continue to be considered for publication in the Journal of Visualized Experiments. All editorial and reviewer modifications to the manuscript were considered and changes made to improve the manuscript. We would like to comment on each editorial or reviewer comment separately.

Editorial comments:
1) Protocol section is beyond JoVEs 3 page guideline, please highlight less than 3 pages of text to identify which portions of the protocol are most important to film; i.e. which steps should be visualized to best supplement the written section of the protocol. Please see JoVEs instructions for authors for more clarification. Remember that the non-highlighted protocol steps will remain in the manuscript and therefore will still be available to the reader.

We have carefully gone over the steps in the protocol and only steps we feel are necessary for the effective use of the webtool were retained. This left approximately 2.5 pages of highlighted text. Hopefully this has been reduced to a level appropriate for the journal.

Reviewer #1:
This manuscript describes a protocol for protein design. It does a nice job in describing the force field for sequence selection, and the structure prediction for validation of design. However, the description lacks a key step on how the sequences are designed, i.e. the move sets for the sequence mutations and searching. Are they generated by randomly mutating the native sequence or from some sort of Monte Carlo/genetic algorithm searching? In either approach, the procedure for sequence space searching needs to be described clearly so that the readers/users can have a complete idea how the sequences were designed.

The paper is generally well-written and highlighted with several successfully designed protein examples.

This comment concerns the clarity of the document in conveying how the sequences are designed in the first stage of the method. The purpose of an optimization method of sequence selection is to make sure that one is not required to search sequence space through random mutation or genetic algorithm based method. Even for proteins of reasonable size this type of search can be computationally prohibitive and there is no guarantee that the sequence one finds is the global minimum in energy. We introduce a deterministic global optimization method which 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. We have added the following paragraphs describing what we state above:

"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." (page 7)

"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." (page 9)

We recognize that this method may not have been explained in enough detail in the manuscript and hope that the changes made are enough to improve clarity.

Reviewer #2:

Summary:

The authors describe their software workbench for computational protein design. The method has been validated on several protein:ligand complexes and detailed background has already been published. The present article is a nice overview that illustrates the possibilities of the method and will allow people to use the tools themselves.

Major Concerns:

No major concerns.

Minor Concerns:

Several figures are non-essential.

Additional Comments to Authors:

none.

No changes were deemed necessary from Reviewer #2's comments.

Thank you very much for the comments and we hope the changes we have made are sufficient.

James Smadbeck
Floudas Lab
Princeton University