Toward Full-Sequence De Novo Protein Design with Flexible Templates for Human Beta-Defensin-2

Ho Ki Fung,* Christodoulos A. Floudas,* Martin S. Taylor,† Li Zhang,‡ and Dimitrios Morikis§

*Department of Chemical Engineering, Princeton University, Princeton, New Jersey; †School of Medicine, The Johns Hopkins University, Baltimore, Maryland; and ‡Department of Chemistry and §Department of Bioengineering, University of California, Riverside, California

ABSTRACT In this article, we introduce and apply our de novo protein design framework, which observes true backbone flexibility, to the redesign of human β-defensin-2, a 41-residue cationic antimicrobial peptide of the innate immune system. The flexible design templates are generated using molecular dynamics simulations with both Generalized Born implicit solvation and explicit water molecules. These backbone templates were employed in addition to the x-ray crystal structure for designing human β-defensin-2. The computational efficiency of our framework was demonstrated with the full-sequence design of the peptide with flexible backbone templates, corresponding to the mutation of all positions except the native cysteines.

INTRODUCTION

Recently there has been growing attention to the importance of antimicrobial peptides (AmPs), which are small proteins of fewer than 100 amino acids that are found in the innate immune system as defense against bacterial infection. This is evidenced by the publication of several important reviews and articles about AmPs (1–6). One of the main reasons is believed to be the better capability of AmPs to combat bacterial resistance compared to conventional antibiotics (3,7).

The various families of AmPs that have been identified in humans so far include histatins, granulysin, lactoferricin, defensins, and cathelicidins, with α- and β-defensins being the most common AmPs (8). The α- and β-defensin classes differ by the positions and connectivity of their six native cysteine residues (9). Human α-defensins, HNP-1 to -4, are found in the storage granules of neutrophils for the killing of ingested microorganisms (10). On the other hand, human β-defensins are expressed in the salivary glands (11,12), the skin (13), and the epithelial tissues (14,15). Six human β-defensins (hβD-1 to -6) have been identified thus far (8,16), and in this research article we focus on human β-defensin-2 (hβD-2) and its de novo computational design using our novel framework (17–20).

The cationic 41-residue peptide hβD-2 was first discovered in 1997 in the human skin (13). It has one α-helix, a β-sheet made of three β-strands, and three disulfide bonds between Cys^{8} and Cys^{37}; Cys^{15} and Cys^{30}; and Cys^{20}, and Cys^{38}, respectively. Since its discovery, it has been shown to be a potent AmP effective against a large variety of microbes, including both Gram-negative bacteria and fungi (21,22). Its antimicrobial property is partly attributed to its high positive charge (+6) which provides a strong electrostatic force between the peptide and the negatively charged outermost leaflet of the microbial membrane bilayer. Based on the Shai-Matsuzaki-Huang mechanism (23–25) by which most other AmPs function, the electrostatic force drives the interaction of the molecule with the membrane, alters the membrane structure, and sometimes even leads to the entry of the peptide into the interior of the microorganism. In addition, hβD-2 serves as a chemotactic agent for T-cells, immature dentritic cells, mast cells, and tumor necrosis factor-α-treated neutrophils (8,26,27). Most importantly, it suppresses the oral transmission of HIV-1, the mechanism of which is still poorly understood, at doses that are compatible with those in the oral cavity (28,29). These characteristics make hβD-2 an ideal candidate as an antimicrobial gene therapy study model (30) and a new generation antibiotic.

Computational de novo design approaches use either rigid templates or flexible templates (31,32). In the former case, the sequence search method is driven either by deterministic methods like the dead-end elimination (33–38) and the self-consistent mean field method (39–41), or by stochastic methods like Monte Carlo methods (42–44) and genetic algorithms (45) based on a single fixed backbone. In the case of flexible templates, de novo design was performed using the same search methods by considering discrete rotamers on discrete templates with fixed backbone assumption for each template (41,46–54), or considering discrete rotamers on a continuum template via backbone parameterization (55–57). Our recently proposed de novo design strategy also employed flexible templates, but via a continuum template and NMR structure refinement instead of discrete rotamers, so that all continuous C$^{\alpha}$-C$^{\alpha}$ distance and dihedral angle values between preset upper and lower bounds are considered (19,20,58). With the study of human β-defensin-2, we aim at illustrating how our framework can be applied to the full-sequence de novo design of proteins.

Submitted April 12, 2007, and accepted for publication August 21, 2007.

Address reprint requests to C. A. Floudas, Tel.: 609-258-4595; E-mail: floudas@titan.princeton.edu.

Editor: Costas D. Maranas.

0006-3495/08/01/584/16 $2.00

doi: 10.1529/biophysj.107.110627

The objective of the de novo design of hβD-2 is to enhance the peptide's antimicrobial property. Krishnakumari et al. (59) investigated the antibacterial activity of a synthetic 19-residue peptide corresponding with position 19 to position 39 of hβD-2 without Cys^{30} and Cys^{37}. Loose et al. (6) designed new AmPs against *Escherichia coli*, *Bacillus anthracis*, and *Staphylococcus aureus* using a purely linguistic approach and obtained favorable experimental results. In our de novo design, instead of fixing the carboxy-terminal region as Krishnakumari et al. (59) did, we considered two separate cases:

1. Up to 10 mutations along hβD-2.
2. Full-sequence design of hβD-2 by mutating all positions except Cys^{8}, Cys^{15}, Cys^{20}, Cys^{30}, Cys^{37}, and Cys^{38}, so as to keep the original S-S bridge architecture and thus the overall structure of the peptide.

Unlike Loose et al. (6)'s approach, we employed the structures of hβD-2 as design templates and identified sequences of new peptides that have the lowest potential energies and thus highest specificities to the templates. Ideally, we should have used the structures of the microbial membrane-peptide complex as design templates. However, they are not readily available in the open literature and are hard to predict with high accuracy. Therefore, we resorted to minimizing the potential energy of the peptide only. Such a strategy has proven to be highly successful in the design of compstatin, a synthetic 13-residue cyclic peptide that binds to complement protein 3 (C3) and inhibits the activation of the complement system, in which the design template was confined to compstatin only (17,18). In this de novo design of hβD-2, in addition to the crystal structure of hβD-2 elucidated by Hoover et al. (60), we generated flexible templates using molecular dynamics simulations with implicit solvation and explicit water molecules and used them as design templates so as to allow for true backbone flexibility.

In this article, our new de novo protein design methodology will be presented first. It will be followed by the study of its application to the design of hβD-2. Finally, we will present the predictions corresponding to the different set of backbone templates employed.

A new de novo protein design framework

In this article, a novel two-stage framework is introduced and applied to the de novo design of human β-defensin-2. The first stage selects amino-acid sequences into either a single template or multiple templates defined by either the C$^{\alpha}$ positions or the side-chain centroids in the template(s). As proven by Pierce and Winfree (61) and by Fung et al. (19), this is an *NP*-hard problem. The second stage calculates and ranks the fold specificities of the sequences selected in the first stage based on the full-atomistic force field AMBER (62), and torsional angle dynamics with restraints through CYANA (63,64).

Stage one: in silico sequence selection

Sequence selection based on a single template structure

The basic sequence selection model for single template structure, recently proposed by Fung et al. (20), has the mathematical formulation of

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

and it is an integer linear programming model. Set $i = 1, \ldots, n$ defines the number of residue positions along the template. At each position i there can be a set of mutations represented by $j\{i\} = 1, \ldots, m_i$, where, for the general case, $m_i = 20 \forall i$. The equivalent sets $k \equiv i$ and $l \equiv j$ are defined, and $k > i$ is required to represent all unique pairwise interactions. Binary variables y_i^j and y_k^l are introduced to indicate the possible mutations at a given position. That is, the y_i^j variable will indicate which type of amino acid is active at a position in the sequence by taking the value of 1 for that specification. The composition constraints in the formulation require that there is exactly one type of amino acid at each position. Noting that binary variable w_{ik}^{jl} is simply the product of y_i^j and y_k^l, the RLT constraints, namely $\sum_{j=1}^{m_i} w_{ik}^{jl} = y_k^l \; \forall\, i,k>i,l$ and $\sum_{l=1}^{m_k} w_{ik}^{jl} = y_i^j \; \forall\, i,k>i,j$, can be derived by multiplying the composition constraints $\sum_{j=1}^{m_i} y_i^j = 1 \; \forall\, i$ by y_k^l and $\sum_{l=1}^{m_k} y_k^l = 1 \; \forall\, k>i$ by y_i^j, respectively.

The objective function to be minimized represents the sum of pairwise amino-acid energy interactions in the template. Parameter $E_{ik}^{jl}(x_i,x_k)$, 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 (x_i, x_k) as well as the type of amino acids j and l. These energy parameters were derived based on solving a linear programming parameter estimation problem subject to constraints which were in turn constructed by requiring the energies of a large number of low-energy decoys to be larger than the corresponding native protein conformation for each member of a set of proteins (65). The resulting potential, which contains 1680 energy parameters for different amino-acid pairs and distance bins, was shown to rank the native fold as the lowest in energy in a large set of proteins tested and also yield very good Z-scores (65–67).

Equation 1 was proved to be significantly more computationally efficient than 12 other equivalent quadratic assignments like models for sequence selection (19,20). In particular, it outperformed the original model proposed by Klepeis et al. (17) on two sequence selection problems for human β-defensin-2: one at a complexity level of 3.4×10^{45}, and the other at 6.4×10^{37} with 49 additional linear biological constraints. The original model proposed by Klepeis et al. (17) was found to take 53,263 CPU seconds and 4578 CPU seconds, respectively, to solve the two problems to global optimality using CPLEX 9.0 (68) on a Pentium IV 3.2 GHz processor. Equation 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.

Sequence selection based on multiple template structures

In an effort to handle the typical case of de novo protein design in which the design template possesses multiple crystal or NMR solution structures, Fung et al. (20) proposed two new sequence selection formulations. One uses a weighted average force field in place of the energy parameters in the single structure model (Eq. 1), with the weights given by the occurrence frequencies of each C^{α}-C^{α} or centroid-centroid distance belonging to a certain distance bin as observed from the template structures. With the aid of binary variables, the other formulation allows the inclusion of all distance bins that each C^{α}-C^{α} or centroid-centroid distance covers according to the template structures. It also imposes constraints that disallow the selection of distance bin combinations which suggest physically meaningless results.

Weighted average force field formulation

In the case when there is only one structure, the energy parameter $E_{ik}^{jl}(x_i, x_k)$ in the objective function can be immediately determined by the coordinates of the two C^{α} or side-chain centroid positions, that is, x_i and x_k, as well as the amino acid at each of those two positions. There is no ambiguity as to which distance bin d it belongs to. In the case of multiple structures, the term $E_{ik}^{jl}(x_i, x_k)$ can be replaced by a weighted average energy term, $\sum_{d=1}^{b_m} E_{ik}^{jl}(x_i, x_k) wt(x_i, x_k, d)$, where the weights $wt(x_i, x_k, d)$ are given by

$$wt(x_i, x_k, d) = \frac{\text{number of structures in which distance between } x_i \text{ and } x_k \text{ is in bin } d}{\text{total number of structures of the template}} \forall\, i, k, d. \qquad (2)$$

The idea can also be examined this way: the distance between x_i and x_k is now replaced by a weighted average distance over all the structures, with the weights given by the above formula. The energy parameters $E_{ik}^{jl}(x_i, x_k)$ can be found using this weighted average distance and simple table lookup in the corresponding force field. With all components other than the energy term kept, the new formulation becomes

$$\begin{aligned}
&\min_{y_i^j, y_k^l} && \sum_{i=1}^{n-1}\sum_{j=1}^{m_i}\sum_{k=i+1}^{n}\sum_{l=1}^{m_k}\sum_{d=1}^{b_m} E_{ik}^{jl}(x_i, x_k) wt(x_i, x_k, d) w_{ik}^{jl} \\
&\text{subject to} && \sum_{j=1}^{m_i} y_i^j = 1\, \forall\, i \\
& && \sum_{j=1}^{m_i} w_{ik}^{jl} = y_k^l\, \forall\, i, k > i, l \\
& && \sum_{l=1}^{m_k} w_{ik}^{jl} = y_i^j\, \forall\, i, k > i, j \\
& && y_i^j, y_k^l, w_{ik}^{jl} = 0 - 1\, \forall\, i, j, k, l. \qquad (3)
\end{aligned}$$

Equation 3 is an integer linear programming model.

Binary distance bin variable formulation

This new formulation was derived by first replacing the energy parameter $E_{ik}^{jl}(x_i, x_k)$ in the objective function of Eq. 1 by $\sum_{d:disbin(x_i, x_k, d)=1} E_{ik}^{jl}(x_i, x_k) b_{ikd}$, where b_{ikd} is a binary variable which assumes the value of one if the distance between x_i and x_k falls into distance bin d and the value of zero otherwise, and $disbin(x_i, x_k, d)$ is a parameter defined as

$disbin(x_i, x_k, d)$
$= 1$ if the distance between x_i and x_k in any of the template structures falls into bin d;
$= 0$ otherwise $\forall\ i,\ k > i,\ d.$

The constraints of $\sum_{d:disbin(x_i, x_k, d)=1} b_{ikd} = 1\, \forall i, k>i$ were imposed to let the energy minimization model pick only one of the distance bins that all the structures cover. After replacing the energy term, the model takes the form

$$\begin{aligned}
&\min_{y_i^j, y_k^l} && \sum_{i=1}^{n-1}\sum_{j=1}^{m_i}\sum_{k=i+1}^{n}\sum_{l=1}^{m_k}\sum_{d:disbin(x_i, x_k, d)=1} E_{ik}^{jl}(x_i, x_k) b_{ikd} w_{ik}^{jl} \\
&\text{subject to} && \sum_{j=1}^{m_i} y_i^j = 1\, \forall\, i \\
& && \sum_{j=1}^{m_i} w_{ik}^{jl} = y_k^l\, \forall\, i, k > i, l \\
& && \sum_{l=1}^{m_k} w_{ik}^{jl} = y_i^j\, \forall\, i, k > i, j \\
& && \sum_{d:disbin(x_i, x_k, d)=1} b_{ikd} = 1\, \forall\, i, k > i \\
& && y_i^j, y_k^l, w_{ik}^{jl}, b_{ikd} = 0 - 1\, \forall\, i, j, k > i, l, d. \qquad (4)
\end{aligned}$$

Equation 4 is nonconvex because of the bilinear term $b_{ikd} w_{ik}^{jl}$ in the objective function. Fung et al. (20) linearized the formulation by declaring $z_{ikd}^{jl} = b_{ikd} w_{ik}^{jl}$ as binary variables and using the RLT equations:

$$w_{ik}^{jl} \sum_{d:disbin(x_i,x_k,d)=1} b_{ikd} = 1 \, \forall \, i,j,k>i,l$$

$$\text{or:} \sum_{d:disbin(x_i,x_k,d)=1} z_{ikd}^{jl} = w_{ik}^{jl} \, \forall \, i,j,k>i,l$$

$$w_{ik}^{jl}, b_{ikd}, z_{ikd}^{jl} = 0-1 \, \forall \, i,j,k>i,l,d. \quad (5)$$

Moreover, Fung et al. (20) also derived novel constraints on the binary distance bin variables to eliminate results in which there is no overlap between regions where the same C^{α} or side-chain centroid position can possibly be located:

$$\begin{aligned}
& b_{ikd} + b_{kpd'} \leq 1 \\
& \quad \text{if} \\
& (l_{mid}(d') < dis(i,p) - l_{mid}(d) \\
& \quad \text{or} \\
& \quad l_{mid}(d') > dis(i,p) + l_{mid}(d)) \\
& \quad \text{and} \\
& \quad \sum_{d''=d+1}^{b_m} disbin(x_i,x_k,d'') \geq 1 \quad \text{and} \quad disbin(x_i,x_k,d) = 1 \\
& \quad \text{and} \\
& \quad disbin(x_k,x_p,d') = 1 \, \forall \, i,k>i,p,d,d',i \neq k \neq p. \qquad (6)
\end{aligned}$$

Without these constraints, the design template defined only by C^{α} or side-chain centroid positions would be given too much flexibility.

With the linearization components and these new constraints on distance bin variables, the whole model for sequence selection into multiple templates takes the form of

$$\begin{aligned}
\min_{y_i^j, y_k^l} \quad & \sum_{i=1}^{n-1} \sum_{j=1}^{m_i} \sum_{k=i+1}^{n} \sum_{l=1}^{m_k} \sum_{d:disbin(x_i,x_k,d)=1} E_{ik}^{jl}(x_i,x_k) z_{ikd}^{jl} \\
\text{subject to} \quad & \textstyle\sum_{j=1}^{m_i} y_i^j = 1 \, \forall \, i \\
& \sum_{j=1}^{m_i} w_{ik}^{jl} = y_k^l \, \forall \, i,k>i,l \\
& \sum_{l=1}^{m_k} w_{ik}^{jl} = y_i^j \, \forall \, i,k>i,j \\
& \sum_{d:disbin(x_i,x_k,d)=1} b_{ikd} = 1 \, \forall \, i,k>i \\
& b_{ikd} + w_{ik}^{jl} - 1 \leq z_{ikd}^{jl} \leq b_{ikd} \quad \forall \, i,j,k>i,l,d \\
& \sum_{d:disbin(x_i,x_k,d)=1} z_{ikd}^{jl} = w_{ik}^{jl} \, \forall \, i,j,k>i,l \\
& b_{ikd} + b_{kpd'} \leq 1 \\
& \text{if} \\
& (l_{mid}(d') < dis(i,p) - l_{mid}(d) \quad \text{or} \\
& \quad l_{mid}(d') > dis(i,p) + l_{mid}(d)) \\
& \text{and} \\
& \sum_{d''=d+1}^{b_m} disbin(x_i,x_k,d'') \geq 1 \quad \text{and} \quad disbin(x_i,x_k,d) = 1 \\
& \text{and} \\
& disbin(x_k,x_p,d') = 1 \, \forall \, i,k>i,p,d,d',i \neq k \neq p \\
& y_i^j, y_k^l, w_{ik}^{jl}, b_{ikd}, b_{kpd'}, z_{ikd}^{jl} = 0-1 \, \forall \, i,j,k>i,l, \\
& p \neq k \neq i,d,d'. \qquad (7)
\end{aligned}$$

The high resolution force fields for sequence selection

In this work, the high resolution C^{α}-C^{α} force field and the high resolution centroid-centroid force field were employed for addressing the sequence selection models (69,70). These force fields are derived from a large training set of 1250 proteins based on high resolution decoys. The force fields are obtained by solving a linear programming parameter estimation problem, requiring that the native conformation of each protein in the training set to be ranked energetically more favorable than their decoys. By using a novel decoy generation method, Rajgaria et al. (69) constructed high resolution decoys which possessed close structural resemblance to the native conformations, and these decoys were subsequently used for training the force fields.

Backbone flexibility at stage one

True backbone flexibility, defined by bounded continuous values of dihedral angles and C^{α}-C^{α} distances (71), is explicitly incorporated into Eqs. 1, 3, and 7 for sequence selection. In all models it is achieved by discretizing the C^{α}-C^{α} or centroid-centroid distance-dependent energy potential $E_{ik}^{jl}(x_i, x_k)$ into a number of bins based on the distance between the two positions (x_i, x_k). For example, in the high resolution C^{α}-C^{α} force field developed by Rajgaria et al. (69), if the pair of amino acids selected at positions i and k are Arg and Glu, respectively, and the corresponding α-carbons are 3.5 Å apart in a single template structure or a weighted average template, their energy contribution to the objective function will be −7.77 kcal/mol. This energy value is constant for all Arg and Glu residues with a C^{α}-C^{α} distance between 3 and 4 Å (bin 1), thus making the objective function insensitive to limited continuous distance variation due to protein backbone motion. A higher degree of backbone flexibility is included in the binary distance bin variable formulation than in other models, because the whole distance range for each position pair (x_i, x_k) spanned by all structures is considered. For instance, if the distance between the same selected amino acid pair Arg-Glu at positions i and k covers bin 1 (3–4 Å with energy −7.77 kcal/mol), bin 2 (4–5 Å with energy −3.77 kcal/mol), and bin 3 (5–5.5 Å with energy −5.61 kcal/mol) according to the flexible template structures, any of the three distance bins can be chosen by the model, and the two C^{α} positions (x_i, x_k) are thus allowed to move within a range of 2.5 Å.

Stage two: fold specificity

The second stage of the new framework provides a more rigorous assessment of the specificity of the low energy sequences within the context of the flexible template. One approach for fold specificity uses the ASTRO-FOLD method and performs first-principles-based protein folding calculations to generate two sets of conformational ensembles: one in which the protein is constrained to a region around the backbone and a second in which the protein is allowed to fold freely (72–78). This requires the use of the deterministic global optimization approach, αBB (79–86). The relative probability of specificity for the protein to assume the target fold is then calculated from the RMSD and free energy of these two ensembles based on the Boltzmann distribution. For rigorous ensemble generation, this method requires that a large number of freefolding calculations be performed, which can be computationally demanding.

A new second-stage method has been developed to handle larger proteins. This method for fold validation is outlined in Fig. 1. First, a flexible template is defined based on the upper and lower bounds on both the distances between α-carbons and the ϕ- and ψ-angles between residues. An ensemble of hundreds of random structures is then generated (conformers) within the confines of the flexible template using the CYANA 2.1 software package for NMR structure refinement (63,64). CYANA 2.1 is then used to perform annealing calculations that simulate a rapid heating of the protein followed by a slow cooling in which high temperature torsion dynamics and annealing torsion dynamics are performed. Violations of van der Waals radii and of the flexible template are minimized, minimizing the energy of the target structures. Hundreds of these structures are generated within the confines of the flexible template.

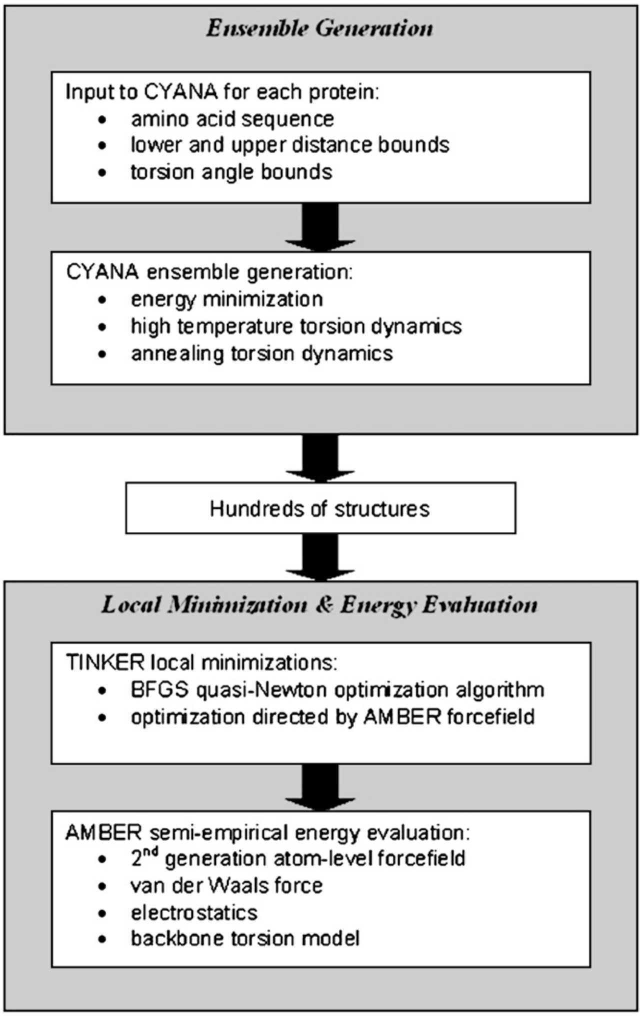
Ensemble Generation
Input to CYANA for each protein:
• amino acid sequence
• lower and upper distance bounds
• torsion angle bounds
CYANA ensemble generation:
• energy minimization
• high temperature torsion dynamics
• annealing torsion dynamics
Hundreds of structures
Local Minimization & Energy Evaluation
TINKER local minimizations:
• BFGS quasi-Newton optimization algorithm
• optimization directed by AMBER forcefield
AMBER semi-empirical energy evaluation:
• 2nd generation atom-level forcefield
• van der Waals force
• electrostatics
• backbone torsion model

FIGURE 1 Workflow for the new method for fold specificity.

For each structure in the ensemble, local minimizations are then performed by the TINKER (87) package as directed by gradients in the fully atomistic force field AMBER (62). AMBER is then used to evaluate the potential energy of the structure. These ensembles are generated for the native sequence of the fold and for each candidate mutant sequence. The specificity of each mutant sequence to the target fold is then calculated relative to the native sequence using the Boltzmann distribution from statistical mechanics. Both the predicted energy of each conformer and its RMSD from the template structure are used in this calculation.

ANALYSIS AND SELECTION OF RESULTS

A method similar to that used by Klepeis et al. (17) for ensemble comparison in fold validation was employed to give a relative ranking for specificity. First, the mean and standard deviation of both RMSD and AMBER energies were found for the native sequence. Upper bounds on both RMSD and energy were then established; for RMSD, the upper bound was selected as one and a half standard deviations above the mean, and in the energy, the upper bound was selected as two standard deviations from the mean. A structure is considered to make a contribution to the ensemble only if its energy and RMSD both fall under these upper bounds. This is illustrated in Fig. 2.

To calculate the relative factor for specificity, we define the set native as the set of all data points from the native sequence that are below both upper bounds, and select the set of all data points from the novel sequence that meet the same criterion. The factor for specificity, $f_{\mathrm{specificity}}$, is then calculated using Boltzmann probabilities as shown in the equation

$$f_{\mathrm{specificity}} = \frac{\sum_{\mathrm{i}\in\mathrm{novel}} \exp[-\beta E_{\mathrm{i}}]}{\sum_{\mathrm{i}\in\mathrm{native}} \exp[-\beta E_{\mathrm{i}}]}, \tag{8}$$

where $\beta = 1/\mathrm{k_B T}$.

Backbone flexibility at stage two

True protein backbone flexibility is incorporated in stage two by CYANA's ability to select any continuous values for the dihedral angles and C^{α}-C^{α} distances between preset bounds

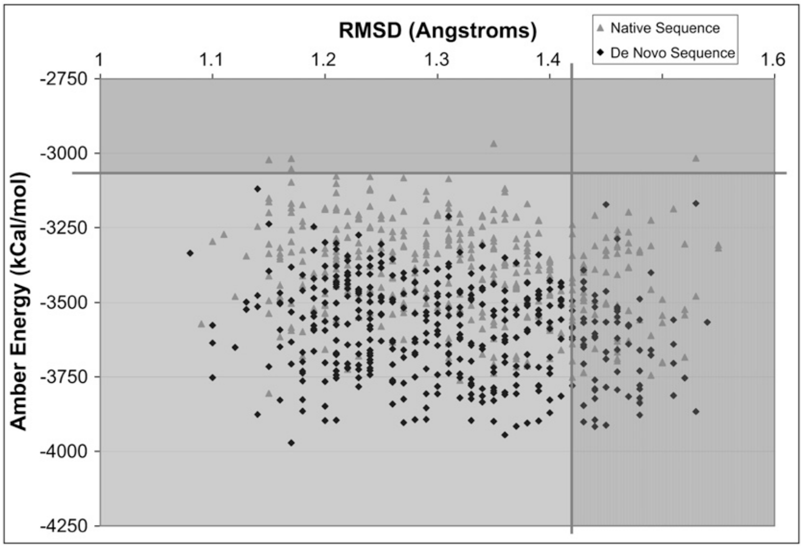

FIGURE 2 Illustration of upper bounds on RMSD and AMBER energy. Lines indicate upper bounds. Data points in the shaded regions are not considered in further calculations.

when it does the simulated annealing calculations. The bounds are input by the user to the program, and they can be based on the observation about the flexible design template(s). The outgoing protein conformations can thus have any possible combination of continuous angle and distance values between the bounds.

DE NOVO DESIGN OF HUMAN β-DEFENSIN-2

This section outlines the details of the de novo design of human β-defensin-2 using the two-stage de novo protein design framework aforementioned.

Design templates

Three different sets of design templates were used for the de novo design of human β-defensin-2.

Single template structure from x-ray crystallography

This design template corresponds to chain A of the x-ray crystal structure elucidated by Hoover et al. (60) (PDB code: 1FD3) at a resolution of 1.35 Å (Fig. 3). Human β-defensin-2 possesses an octameric quaternary structure constituted by eight identical chains: chain A, B, C, D, E, F, G, and H, each of which has the natural sequence of GIGDPVTCLKSGAICHPVFCPRRYKQIGTCGLPGTKCCKKP (88). The identical monomer units of hβD-2 are grouped into units of four that are oriented in such a way that their N-termini are in the core of the octamer. The core is sealed off from solvent by hydrogen bonds between Gly^1, Gly^3, Asp^4, and Thr^7. The overall tertiary structure is maintained by a mix of hydrophobic and hydrogen-bonding interactions between the residues Gly^1, Asp^4, Thr^7, Lys^{10}, Gly^{31}, Leu^{32}, Pro^{33}, and Lys^{39}. Among the eight identical chains, only chain A was used for the de novo design.

The surface of hβD-2 is mostly amphiphilic. Like other human β-defensins, hβD-2 has an *N*-terminus α-helix located at Pro^5-Lys^{10} that is held against the β-sheet by an S-S bond between Cys^8 and Cys^{37}. Two other S-S bonds that stabilize the β-sheet are located at Cys^{15}-Cys^{30} and Cys^{20}-Cys^{38}. The β-sheet is made up of three anti-parallel β-strands held together by hydrophobic interactions. The structural properties of hβD-2 are summarized in Table 1.

Flexible templates from molecular dynamics simulations with Generalized Born implicit solvation model

In an effort to generate a flexible design template for human β-defensin-2, molecular dynamics (MD) simulations were employed to capture different structures of the peptide along the MD trajectory. MD simulations were performed using

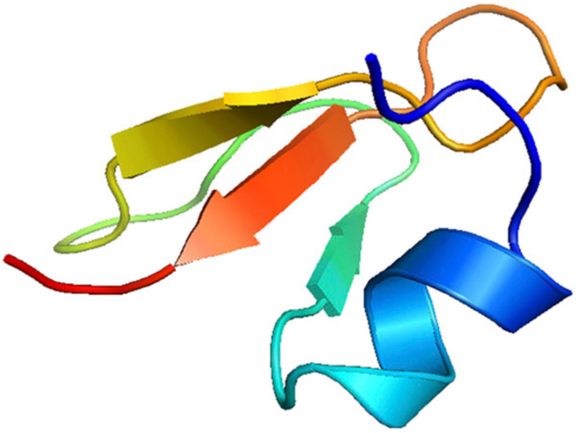

FIGURE 3 Structure of human β-defensin-2 (chain A) as elucidated by Hoover et al. (60). Its secondary structure consists of a β-sheet made up of three β-strands and an α-helix.

TABLE 1 Structural features of human β-defensin-2

Structural features	Positions
β-strands	14–16
	25–28
	36–39
α-helix	5–10
	8–37
S-S bonds	15–30
	20–38
	16–19
β-turns	21–24
	32–35
Hairpins	25–29
Bulges	27, 28, 37

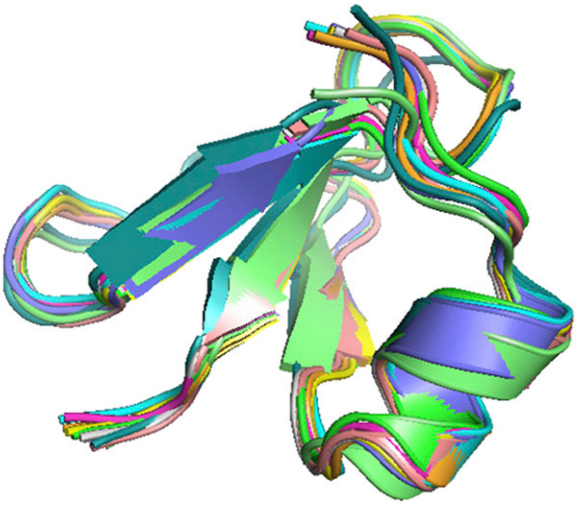

FIGURE 4 Overlay of the 10 structures of human β-defensin-2 used for the flexible design template from MD simulations with the GB implicit solvation model. The structures are snapshots with 1-ns increment.

the program CHARMM (89) version 31b1, with implementation of the Generalized Born (GB) implicit solvent model (90). The CHARMM19 force field was employed with the GBORn model. The dielectric constant was set to 1.0 for the interior of the protein and 80.0 for the solvent. All nonbonded interactions were computed without cutoffs. The SHAKE algorithm (91) was used to fix the length of covalent bonds of hydrogen atoms. The time step was set to 2 fs. Monomer A of the crystal structure 1FD3 was used. The structure was first energy-minimized for 300 steps, using the adopted-basis Newton-Raphson (ABNR) method. Then, the system was subjected to 5 ps of constant volume molecular dynamics, during which the temperature was raised from 0 K to 300 K with velocity rescaled every 0.1 ps. At 300 K, 30-ps equilibrium phase was performed, with velocity rescaled every 0.1 ps, during the first 10 picoseconds. In the middle 10 picoseconds, velocity was rescaled only if the temperature of the system deviated more than 5 K from 300 K. During the last 10 picoseconds, energy and temperature were stable, and no velocity rescaling was necessary. After the equilibration, a 10-ns trajectory of NVT MD at 300 K was generated. Coordinate sets were sampled every 10 ps to generate 1000 snapshots of structures. A total of 10 structures with 1-ns increment were extracted from the MD trajectory to constitute the set of flexible templates, which is shown in Fig. 4.

Flexible templates from molecular dynamics simulations with explicit water molecules

Here MD simulations were done in a more computationally demanding manner by the explicit treatment of water molecules. The program CHARMM (89) version 31b1, with the CHARMM27 force field, was used for the MD simulation with explicit solvation. Monomer A of the crystal structure 1FD3 was used. The protein was immersed in a $50 \times 50 \times 50$ Å^3 cubic box of TIP3P water models. Water molecules were deleted when their oxygen atoms were within 2.8 Å of any heavy atom of the protein. One sodium ion and seven chloride ions were added to represent ~100 mM salt concentration and to neutralize the overall system. The dielectric constant was set to 1.0. Nonbonded interaction cutoff of 12.0 Å was used, with a force-switching function for electrostatic interactions and shifting function for van der Waals interactions between 9 and 12 Å. First, the protein was minimized using 500 steps of the ABNR method with water molecules being held fixed. Subsequently, the entire system was relaxed without any constraints with 500 steps of ABNR minimization. The final system comprises 11,328 atoms including water molecules, protein (610 atoms), and ions. During the equilibration and production molecular dynamics, the protein was restrained to the center of the water box. The system was first heated from 0 K to 300 K with 30 ps of molecular dynamics, during which velocities were scaled every 1 ps and was allowed to fluctuate within 5 K of the target temperature. After heating, the system was equilibrated at 300 K for an additional 100 ps. After the equilibration, a 2-ns MD trajectory was generated at 300 K. Coordinate sets were sampled every 10 ps to generate 200 snapshots of structures. A total of 10 structures with 0.2 ns increment were extracted from the MD trajectory to constitute the set of flexible templates, which is shown in Fig. 5. Alignment of the three different sets of design templates (Fig. 6) indicates flexibility and high structural similarity among them.

The de novo design

Stage one: in silico sequence selection

Models. Different models were employed for sequence selection, depending on the nature of the design template(s). The basic model for single structure (Eq. 1) suffices for the single crystal structure by Hoover et al. (60). For the other

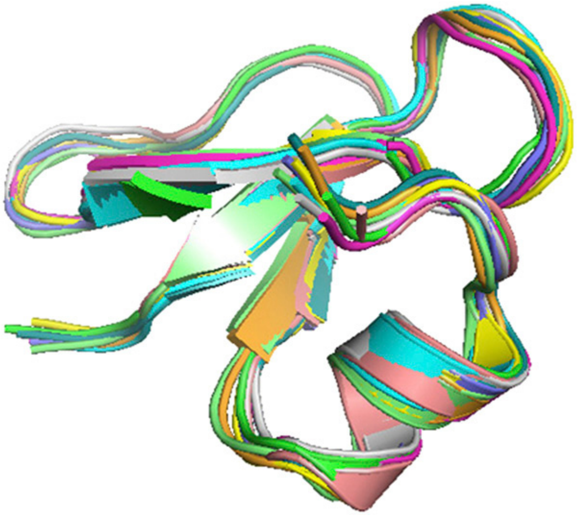

FIGURE 5 Overlay of the 10 structures of human β-defensin-2 for the flexible design template from MD simulations with explicit water molecules. The structures are snapshots with 0.2-ns increment.

two flexible design templates generated from MD simulations which have multiple structures, both the weighted average force field formulation (Eq. 3) and the binary distance bin variable formulation (Eq. 7) were utilized for sequence selection.

Force fields. The high resolution C$^{\alpha}$-C$^{\alpha}$ force field (69) was employed for sequence selection based on the design template from x-ray crystallography, while the high resolution centroid-centroid force field (70) was used for the two sets of templates from MD simulations.

Number of sequence solutions. One-hundred sequences were generated for the crystal structure template in the first stage, and they were ranked by their fold specificities based on the full-atomistic force field AMBER in the second stage. For the flexible templates from MD simulations, ~1000 sequences and ~500 sequences were solved using the weighted average model (Eq. 3) and distance bin model (Eq. 7), correspondingly.

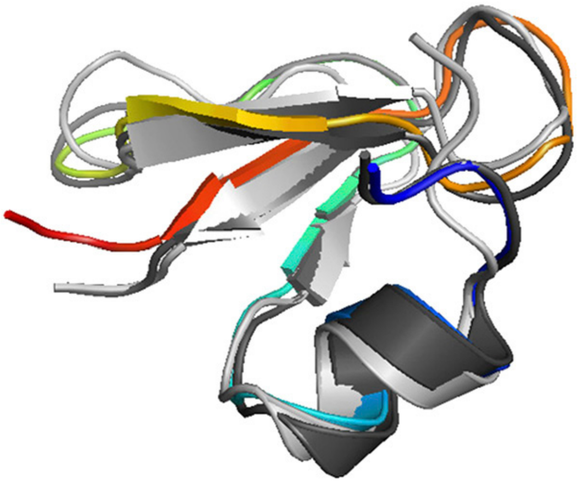

FIGURE 6 Structural alignment of the crystal structure of human β-defensin-2 (chain A) by Hoover et al. (60) (*rainbow color*), the 5-ns MD-GB structure (*light gray*), and the 1-ns explicit MD structure (*dark gray*).

Mutation set. SASA patterning was applied to restrict the sequence search space for the de novo design of hβD-2. The 41 positions in hβD-2 are classified into the core, surface, and intermediate categories which determine the mutation set for each position. The native residue for each position is also included in its mutation set. Proline is excluded from the list for surface and intermediate positions to avoid unnecessary rigidity imposed on the backbone, except when it is the native residue for the position. The mutation set for human β-defensin-2 is tabulated in Table 2. This SASA patterning strategy aims at conserving the natural amphiphi-

TABLE 2 Mutation set of human β-defensin-2 given by SASA patterning

Position	Native residue	Side-chain accessibility	Position type	Varied position?	Allowed mutations
1	G	139.6%	surface	√	R,N,D,Q,E,G,H,K,S,T
2	I	20.7%	intermediate	√	all except C and P
3	G	69.0%	surface	√	R,N,D,Q,E,G,H,K,S,T
4	D	52.5%	surface	√	R,N,D,Q,E,G,H,K,S,T
5	P	52.1%	surface	√	R,N,D,Q,E,G,H,K,P,S,T
6	V	99.9%	surface	√	R,N,D,Q,E,G,H,K,S,T,V
7	T	54.9%	surface	√	R,N,D,Q,E,G,H,K,S,T
8	C	0.0%	buried	×	none
9	L	64.5%	surface	√	R,N,D,Q,E,G,H,K,S,T,L
10	K	94.2%	surface	√	R,N,D,Q,E,G,H,K,S,T
11	S	52.2%	surface	√	R,N,D,Q,E,G,H,K,S,T
12	G	97.3%	surface	√	R,N,D,Q,E,G,H,K,S,T
13	A	1.8%	buried	√	A,I,L,M,F,Y,W,V
14	I	49.6%	intermediate	√	all except C and P
15	C	18.9%	buried	×	none
16	H	24.9%	intermediate	√	all except C and P
17	P	66.4%	surface	√	R,N,D,Q,E,G,H,K,S,T,P
18	V	79.0%	surface	√	R,N,D,Q,E,G,H,K,S,T,V
19	F	69.1%	surface	√	R,N,D,Q,E,G,H,K,S,T,F
20	C	10.7%	buried	×	none
21	P	32.0%	intermediate	√	all except C
22	R	92.2%	surface	√	R,N,D,Q,E,G,H,K,S,T
23	R	84.6%	surface	√	R,N,D,Q,E,G,H,K,S,T
24	Y	24.7%	intermediate	√	all except C and P
25	K	82.3%	surface	√	R,N,D,Q,E,G,H,K,S,T
26	Q	46.7%	intermediate	√	all except C and P
27	I	42.1%	intermediate	√	all except C and P
28	G	45.8%	intermediate	√	all except C and P
29	T	54.1%	surface	√	R,N,D,Q,E,G,H,K,S,T
30	C	2.6%	buried	×	none
31	G	60.3%	surface	√	R,N,D,Q,E,G,H,K,S,T
32	L	87.1%	surface	√	R,N,D,Q,E,G,H,K,S,T,L
33	P	86.1%	surface	√	R,N,D,Q,E,G,H,K,S,T,P
34	G	96.5%	surface	√	R,N,D,Q,E,G,H,K,S,T
35	T	13.9%	buried	√	A,I,L,M,F,Y,W,V,T
36	K	33.2%	intermediate	√	all except C and P
37	C	0.0%	buried	×	none
38	C	0.0%	buried	×	none
39	K	45.8%	intermediate	√	all except C and P
40	K	61.2%	surface	√	R,N,D,Q,E,G,H,K,S,T
41	P	58.5%	surface	√	R,N,D,Q,E,G,H,K,S,T,P

licity of hβD-2, which is considered important for the antimicrobial activity of the peptide (9). Complexity of the problem amounts to 6.40×10^{37}.

Biological constraints. Homology search was executed to determine what fundamental properties of hβD-2 have been highly conserved in evolution. These conserved fundamental properties are maintained in the de novo designed protein as they can contribute significantly to the protein structure and function (92,93).

Such a homology search on hβD-2 was performed using a position-specific iterative basic local alignment search tool (PSI-BLAST 2.0) (l). A total of 96 human β-defensin homologs were identified. The conserved properties about their charges and amino acid content are tabulated in Tables 3 and 4, respectively, and they correspond to the upper and lower bounds on the charges and amino-acid composition.

The upper bound and lower bound on the amino-acid content were set equal to the maximum and minimum occurrences found in the hβD-2 homologs, respectively, except for cysteine, glycine, and tryptophan. The number of cysteines was fixed to six in view of the three disulfide bonds in hβD-2. The number of glycines was limited to be ≤ 6, which is its occurrence in the native sequence, while tryptophan content was allowed to have an upper bound of two instead of one as suggested by homology search, so that the overall hydrophobicity can be enhanced for higher molecular stability.

The conserved charge characteristics of hβD-2 homologs were converted to the constraints below and added to the sequence selection models:

$$
\begin{aligned}
&0 \leq \sum_i y_i^{\mathrm{Arg}} + \sum_i y_i^{\mathrm{Lys}} - \sum_i y_i^{\mathrm{Asp}} - \sum_i y_i^{\mathrm{Glu}} \leq 3\, \forall 5 \leq i \leq 10 \\
&5 \leq \sum_i y_i^{\mathrm{Arg}} + \sum_i y_i^{\mathrm{Lys}} \leq 10\, \forall i \\
&0 \leq \sum_i y_i^{\mathrm{Asp}} + \sum_i y_i^{\mathrm{Glu}} \leq 2\, \forall i \\
&4 \leq \sum_i y_i^{\mathrm{Arg}} + \sum_i y_i^{\mathrm{Lys}} - \sum_i y_i^{\mathrm{Asp}} - \sum_i y_i^{\mathrm{Glu}} \leq 9\, \forall i. \qquad (9)
\end{aligned}
$$

These linear constraints would restrict the charges on any sequence solution generated from stage one to be within the bounds stated in Table 3.

Residue frequencies on the whole sequence were also constrained to be between the maximum and minimum occurrences found in the hβD-2 homologs, by means of the following equations:

TABLE 3 Charge frequencies of homologs of human β-defensin-2

	Lower bound	Upper bound
Net charge on α-helix	0	+3
Total positive charges	5	10
Total negative charges	0	−2
Total net charges	+4	+9

TABLE 4 Occurrence of each amino acid in human β-defensin-2 homologs; residues with asterisks do not follow the maximum and minimum occurrences found in the sequences from the homology search (refer to the text for the actual constraints imposed)

Amino acid	Lower bound	Upper bound
Ala	0	3
Arg	1	9
Asn	0	6
Asp	0	2
Cys*	4	7
Gln	0	3
Glu	0	3
Gly*	3	7
His	0	4
Ile	0	6
Leu	0	4
Lys	0	7
Met	0	3
Phe	0	4
Pro	0	5
Ser	0	6
Thr	0	4
Trp†	0	1
Tyr	0	4
Val	0	6

$$
\begin{aligned}
&0 \leq \sum_i y_i^{\mathrm{Ala}} \leq 3\, \forall i 0 \leq \sum_i y_i^{\mathrm{Gln}} \leq 3\, \forall i \\
&0 \leq \sum_i y_i^{\mathrm{Leu}} \leq 4\, \forall i 0 \leq \sum_i y_i^{\mathrm{Ser}} \leq 6\, \forall i \\
&1 \leq \sum_i y_i^{\mathrm{Arg}} \leq 9\, \forall i 0 \leq \sum_i y_i^{\mathrm{Glu}} \leq 3\, \forall i \\
&0 \leq \sum_i y_i^{\mathrm{Lys}} \leq 7\, \forall i 0 \leq \sum_i y_i^{\mathrm{Thr}} \leq 4\, \forall i \\
&0 \leq \sum_i y_i^{\mathrm{Asn}} \leq 6\, \forall i \sum_i y_i^{\mathrm{Gly}} \leq 6\, \forall i \\
&0 \leq \sum_i y_i^{\mathrm{Met}} \leq 3\, \forall i 0 \leq \sum_i y_i^{\mathrm{Trp}} \leq 2\, \forall i \\
&0 \leq \sum_i y_i^{\mathrm{Asp}} \leq 2\, \forall i 0 \leq \sum_i y_i^{\mathrm{His}} \leq 4\, \forall i \\
&0 \leq \sum_i y_i^{\mathrm{Phe}} \leq 4\, \forall i 0 \leq \sum_i y_i^{\mathrm{Tyr}} \leq 4\, \forall i \\
&\sum_i y_i^{\mathrm{Cys}} = 6\, \forall i 0 \leq \sum_i y_i^{\mathrm{Ile}} \leq 6\, \forall i \\
&\sum_i y_i^{\mathrm{Pro}} \leq 5\, \forall i 0 \leq \sum_i y_i^{\mathrm{Val}} \leq 6\, \forall i. \qquad (10)
\end{aligned}
$$

Lastly, β-strands were restricted to have at least two hydrophobic residues to ensure enough hydrophobic interaction between β-strands for stability purpose. The requisite constraints are

$$
\begin{aligned}
&\sum_i y_i^{\mathrm{Cys}} + \sum_i y_i^{\mathrm{Ile}} + \sum_i y_i^{\mathrm{Leu}} + \sum_i y_i^{\mathrm{Met}} + \sum_i y_i^{\mathrm{Phe}} \\
&\quad + \sum_i y_i^{\mathrm{Trp}} + \sum_i y_i^{\mathrm{Tyr}} + \sum_i y_i^{\mathrm{Val}} + \sum_i y_i^{\mathrm{Ala}} \geq 2\, \forall 14 \leq i \leq 16 \\
&\sum_i y_i^{\mathrm{Cys}} + \sum_i y_i^{\mathrm{Ile}} + \sum_i y_i^{\mathrm{Leu}} + \sum_i y_i^{\mathrm{Met}} + \sum_i y_i^{\mathrm{Phe}} \\
&\quad + \sum_i y_i^{\mathrm{Trp}} + \sum_i y_i^{\mathrm{Tyr}} + \sum_i y_i^{\mathrm{Val}} + \sum_i y_i^{\mathrm{Ala}} \geq 2\, \forall 25 \leq i \leq 28. \qquad (11)
\end{aligned}
$$

Note that no constraint was imposed on the third β-strand that already has two nonvaried cysteines, which are hydrophobic.

Number of mutations. In the sequence selection for the design template from x-ray crystallography, the number of mutations was set to be not more than 10 by the constraint below:

$$\sum_{i=1}^{n} \sum_{j=1, j \neq \text{native residues}}^{m_i} y_i^j \leq 10. \tag{12}$$

For the other two sets of flexible templates, the maximum number of mutations was either set to 10, or unlimited. The latter case corresponds to a full-sequence design of human β-defensin-2, since all positions, except for the six native cysteines, were varied.

Stage two: fold specificity

In calculating the fold specificities using the AMBER force field, the angle and distance bounds input to the CYANA 2.1 package were ±10° around the template and ±10% of those in the template, respectively, for the sequences from the single crystal structure template. For the sequences from the flexible templates from MD simulations, the bounds were set to be the maximum and minimum as observed from all template structures. Five-hundred low energy conformations were generated for each sequence by the simulated annealing algorithm in CYANA and their energies were minimized further by the TINKER program. Finally, the fold specificity for each sequence was computed using formula (8), and the sequences were then ranked according to their specificities.

RESULTS AND DISCUSSION

The top sequences ranked by their fold specificities for the design template from x-ray crystallography, MD simulations with generalized Born implicit solvent, and MD simulations with explicit water molecules are listed in the Supplementary Material in Table S1, Tables S2–S5, and Tables S6–S9, respectively.

Results based on the single template structure from x-ray crystallography

As shown in Table S1 (see Supplementary Material), the high resolution C^{α}-C^{α} force field suggests the mutations of G3T, D4R, I14V, H16V, P17H, R22N, Q26F, G28F, K36(V/F), K39(A/F/Y), and K40N, when the number of mutations is limited to be not more than 10.

Results based on the flexible templates from molecular dynamics simulations with Generalized Born implicit solvation model

In this run, when an upper bound of 10 is imposed on the number of mutations, the weighted average sequence selection model driven by the centroid-centroid force field suggests the mutations of P5R, H16(F/I), P17(Q/N/R), P21I, Q26(I/L/Y), G28L, G31(K/Q), G34R, T35W, K36W, and K39(L/Y) (see Table S2 in Supplementary Material); the binary distance bin sequence selection model with the same force field prefers P5R, G12(H/D), A13F, H16(I/F/W), P17(R/N), P21I, Q26(L/I), G28(L/Y), G31(K/Q), G34R, T35W, K36(W/Y), and K39Y (see Table S3 in Supplementary Material).

For full-sequence design, the following predictions given by the weighted average model yield the highest fold specificity: G1D, I2(F/M), G3H, D4(E/G), P5(K/Q), T7(R/H/K), K10(G/H), S11(K/H), G12(N/H), A13(I/F), H16I, P17R, P21I, R22(T/H/Q/K), R23(G/Q), Y24I, K25(G/R/N), Q26(I/L), I27(L/Y), G28(Y/L), T29K, G31(K/Q), P33E, G34R, T35W, K36W, K39(L/Y), K40(N/Q/R), and P41(Q/T) (see Table S4 in Supplementary Material); the corresponding predictions made by the distance bin model are: I2M, D4(N/K), P5N, T7(N/G), K10G, S11(R/E/D), G12(D/H/K), A13F, I14F, H16W, P17(R/K), P21F, R22H, R23H, Y24I, K25H, Q26I, I27Y, G28Y, T29(N/Q/E/K), G31Q,

TABLE 5 For either up to 10 mutations or full-sequence design, common mutations suggested by both the weighted average model (Eq. 3) and the binary distance bin model (Eq. 7) that are found in both sets of templates are underlined; those underlined mutations that are found in both cases of ≤10 mutations and full-sequence design are added an asterisk each

	Templates from MD simulations with GB implicit solvent			Templates from MD simulations with explicit water molecules		
Up to 10 mutations	P5R	H16(F/I)	P17(N/R)	H16(F/I)	P21(I/Y)	Q26(I/F)
	P21I	Q26(I/L)	G28L	G28L	G34R*	T35W
	G31(K/Q)	G34R*	T35W	K36W	K39Y*	
	K36W	K39Y*				
Full sequence design	I2M	K10G	G12H	G1D	I2F	D4K
	A13F	P17R	R22H	T7N	K10G	A13I
	Y24I	Q26I	I27Y	I14L	P17R	P21I
	G28Y	T29K	G31Q	R22Q	R23G	Y24L
	G34R*	T35W	K39Y*	K25R	Q26F	I27Y
				G28Y	G34R*	K36W
				K39Y*	K40R	P41(G/Q)

TABLE 6 Different mutations suggested by the weighted average model (Eq. 3) and the binary distance bin model (Eq. 7) in each of the cases of ≤10 mutations and full-sequence design and in each of the two sets of flexible templates based on MD simulations

	Templates from MD simulations with GB implicit solvent			Templates from MD simulations with explicit water molecules		
Up to 10 mutations	Weighted average model			Weighted average model		
	P17Q	K39L		G3D	P5(R/N)	H16Y
				P21V	G31(Q/N)	K39L
	Distance bin model			Distance bin model		
	G12(H/D)	A13F	H16W	G12(D/E)	A13F	H16L
	Q26Y	G28Y	K36Y	Q26L	G28Y	
Full sequence design	Weighted average model			Weighted average model:		
	G1D	I2F	G3H	G1E	G3D	P5N
	D4(E/G)	P5(K/Q)	T7(R/H/K)	S11H	G12H	A13Y
	K10H	S11(K/H)	G12N	H16(Y/I/F)	R22G	K25G
	A13I	H16I	P21I	T29K	G31N	P33N
	R22(T/Q/K)	R23(G/Q)	K25(G/R/N)	T35W	K39I	P41T
	Q26L	I27L	G28L			
	G31K	P33E	K36W			
	K39L	P41(Q/T)				
	Distance bin model:			Distance bin model:		
	D4(N/K)	P5N	T7(N/G)	G3H	P5H	T7K
	S11(R/E/D)	G12(D/K)	I14F	S11(N/G)	G12(K/E)	A13F
	H16W	P17K	P21F	H16W	R22(N/T)	Y24I
	R23H	K25H	T29(N/Q/E)	T29H	G31H	P33R
	P33Q	K36Y	P41(K/N)	T35Y	K40(N/Q)	

P33Q, G34R, T35W, K36Y, K39Y, and P41(K/N) (see Table S5 in Supplementary Material).

Results based on the flexible templates from molecular dynamics simulations with explicit water molecules

For this set of templates, when the number of mutations is restricted to be ≤10, the weighted average model suggests the mutations of G3D, P5(R/N), H16(I/F/Y), P21(I/Y/V), Q26(I/F), G28L, G31(Q/N), G34R, T35W, K36W, and K39(L/Y) (see Table S6 in Supplementary Material); the distance bin model selects G12(D/E), A13F, H16(L/F/I), P21(Y/I), Q26(I/F/L), G28(L/Y), G34R, T35W, K36W, and K39Y (see Table S7 in Supplementary Material).

In the case of full-sequence design, the weighted average model suggests the following mutations: G1(D/E), I2F, G3D, D4K, P5N, T7N, K10G, S11H, G12H, A13(Y/I), I14L, H16(Y/I/F), P17R, P21I, R22(G/Q), R23G, Y24L, K25(R/G), Q26F, I27Y, G28Y, T29K, G31N, P33N, G34R, T35W, K36W, K39(Y/I), K40R, and P41(G/T/Q) (see Table S8 in Supplementary Material); the corresponding mutations selected by the distance bin model are: G1D, I2F, G3H, D4K, P5H, T7(K/N), K10G, S11(N/G), G12(K/E), A13(F/I), I14L, H16W, P17R, P21I, R22(N/Q/T), R23G, Y24(I/L), K25R, Q26F, I27Y, G28Y, T29H, G31H, P33R, G34R, T35Y, K36W, K39Y, K40(N/R/Q), and P41(G/Q) (see Table S9 in Supplementary Material).

Similarities and differences between results based on the two sets of flexible templates from molecular dynamics simulations

While the results from the single x-ray crystal structure are based on the high resolution C^{α}-C^{α} force field and should stand alone by themselves, results based on the two sets of flexible templates from MD simulations are all produced using the centroid-centroid force field. In the latter case, high

TABLE 7 For either the weighted average model (Eq. 3) or the binary distance bin model (Eq. 7) for sequence selection, common mutations in the cases of ≤10 mutations and full-sequence design that are found in both sets of templates are underlined; those underlined mutations that are found in both the weighted average model and the binary distance bin model are added an asterisk each

	Templates from MD simulations with GB implicit solvent			Templates from MD simulations with explicit water molecules		
Weighted average model	H16I	P17R	P21I	G3D	P5N	H16(I/F/Y)
	Q26(I/L)	G28L	G31(K/Q)	P21I	Q26F	G31N
	G34R*	T35W	K36W	G34R*	T35W	K36W
	K39(Y*/L)			K39Y*		
Binary distance bin model	G12(H/D)	A13F	H16W	G12E	A13F	P21I
	P17R	Q26I	G28Y	Q26F	G28Y	G34R*
	G31Q	G34R*	K36Y	K36W	K39Y*	
	K39Y*					

level of similarity is observed between the weighted average model predictions and the distance bin model predictions, as well as those for up to 10 mutations and those corresponding to the full-sequence design.

Weighted average model versus binary distance bin model

In each of the cases of up to 10 mutations and full-sequence design and for each of the two flexible templates from MD simulations, the common mutations predicted by both the weighted average model and the binary distance bin model are tabulated in Table 5. Those common mutations that are found in both sets of flexible templates are underlined. It should be noted that G34R and K39Y are seen in all cases.

The high level of similarity clearly suggests that the weighted average model can be used as a good approximation for the distance bin model, which is more computationally demanding, in de novo designs where the problem complexities are high.

The different mutations from the two models in each of the cases of up to 10 mutations and full-sequence design for each of the two flexible templates from MD simulations are listed in Table 6.

Up to 10 mutations versus full-sequence design

Here we identify the mutations in the case of up to 10 mutations that are also found in the full-sequence design case, and we perform this for the weighted average formulation

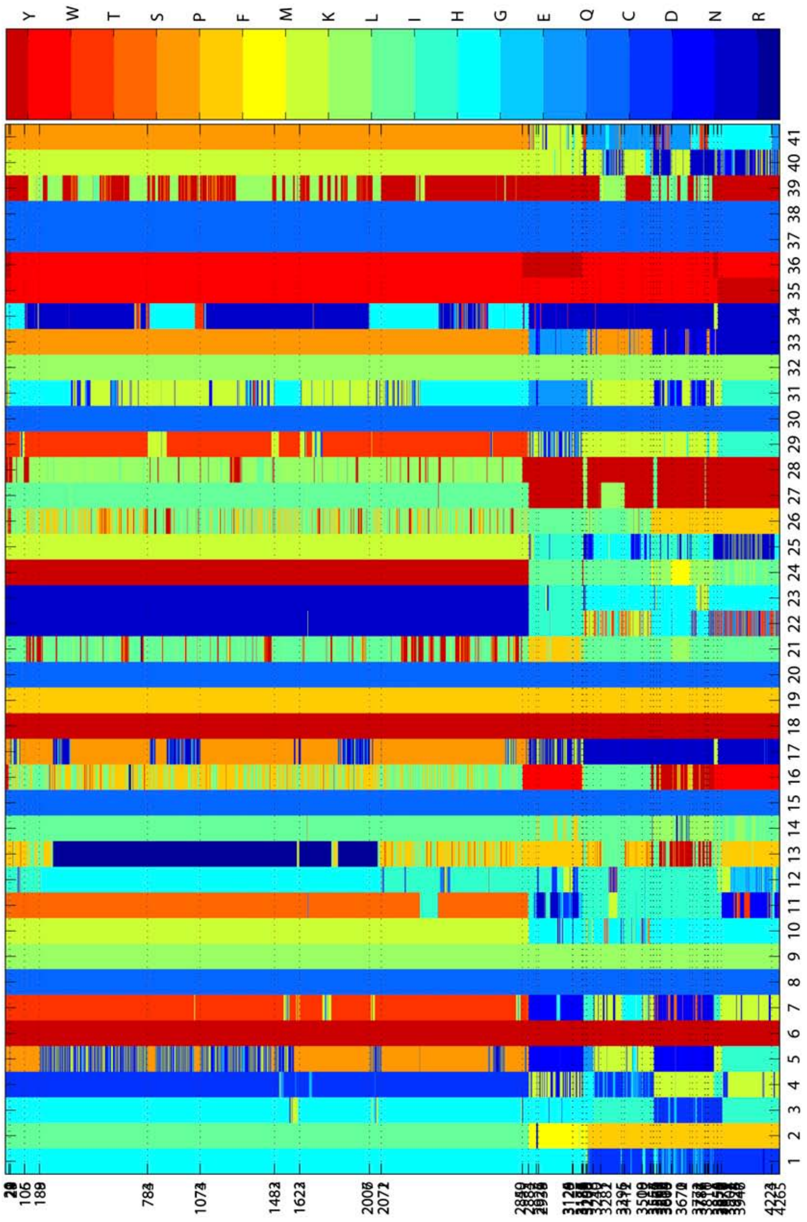

FIGURE 7 Clustering and optimal reordering of the 4266 sequences predicted from all sequence selection models with the flexible templates. Dotted lines indicate cluster boundaries. Different amino acids at the 41 positions are illustrated with different colors.

and the distance bin formulation for each of the two template sets. The results can be found in Table 7. The common mutations that are the same for both sets of flexible templates are underlined. In each of the four cases shown, the number of common mutations is close to 10, indicating that predictions for up to 10 mutations are usually also found in the full-sequence design.

Clustering of predicted sequences

We performed sequence clustering to assess the similarity among the sequences predicted from each of the models using the flexible templates. The substitution matrix PAM70 was used to quantify the distance between sequences because it is recommended for query lengths between 35 and 50 amino acids (94,95). The diagonal of this matrix was modified so that exact matches between amino-acid residues had equivalent scores. Combining the sequences predicted from all the models results in 4266 protein sequences with 41 amino acids in each. We determined the best rearrangement of the protein sequences by minimizing the sum of the overall residue-pair distances for each position using an optimal reordering method (unpublished data). The results for the reordered proteins are presented in Fig. 7. Cluster boundaries are subsequently identified from the reordered proteins using the following method. In the final ordering, each sequence is assigned to its own cluster. We examined the average distance between each cluster to its neighboring clusters in the final ordering and then merged the two clusters that are of minimum distance apart. This is done iteratively until the maximum number of clusters (to be specified by the user) is reached.

As shown in the figure, the largest clusters correspond to the following sequences: G^{1}-I^{2}-G^{3}-D^{4}-(P/R/K/N)^{5}-V^{6}-T^{7}-C^{8}-L^{9}-K^{10}-S^{11}-G^{12}-A^{13}-I^{14}-C^{15}-(F/I/L)^{16}-(P/R)^{17}-V^{18}-F^{19}-C^{20}-(I/Y)^{21}-R^{22}-R^{23}-Y^{24}-K^{25}-(I/L/V/F)^{26}-I^{27}-(I/L/Y)^{28}-T^{29}-C^{30}-(K/G)^{31}-L^{32}-P^{33}-(G/R)^{34}-W^{35}-W^{36}-C^{37}-C^{38}-(L/Y)^{39}-K^{40}-P^{41}. This is obviously contributed from the runs with not more than 10 mutations. Within these clusters, results from the four different sets using the weighted average model and the distance bin model with either the flexible templates from MD simulations with GB implicit solvent or the templates from MD simulations with explicit water molecules were observed to be interspersed, suggesting a high level of conservation among these sequences.

In addition, we compared our predictions to 90 human β-defensin homologs obtained by running the sequence alignment tool of PSI-BLAST, which was created by the National Center for Biotechnology Information of the National Institute of Health, with the default threshold of 0.005 for the position conservation score. Residues at each of the 41 positions found among these homologs are listed in Table 8. To note, except for positions 28, 35, 36, and 39, residues in the major clusters shown above are found in these homologs. This reveals that while our predictions are natural-looking to a large extent, some positions are diverse enough for favorable potential energy contributions.

TABLE 8 Residues at each of the 41 positions among the human β-defensin homologs obtained by using PSI-BLAST; those with an asterisk agree with the major clusters of all our predicted sequences

Position	Residues
1	G*
2	I*,A,V
3	G*,S,M,R,N,I,K,E,T
4	D*,N,G,S,T,E
5	P*,S,H,R*,F,Y,T
6	V*,I,L,R,Q,F,A,K
7	T*,S,K,Q,A
8	C*,Y
9	L*,I,A,S,V,R,Y,H,G,M,W,F,C
10	K*,R,T,L,G,I,Q,W,M,S,E,A,V
11	S*,N,K,H,Y,I,R,A
12	G*,R,K,S,M,N,I
13	A*,G,N,D,R
14	I*,V,F,R,T,Y,A,S
15	C*
16	H,Y,I*,V,M,W,L*,A,F*,Q,R
17	P*,R*,S,G,L,N,A,Y,T,F
18	V*,R,I,P,G,S,A,D,T,F,N,L,Y,K,M
19	F*,S,G,K,R,W,C,L,Y,Q,E,N,D,T
20	C*,I
21	P,L,A,T,I*,G,S,K,R,N
22	R*,G,P,V,T,H,L,Y,W
23	R*,S,G,N,K,P,H,A,T,F,L
24	Y*,M,S,F,L,T,R,H,Q,I,E
25	K*,R,E,I,D,L,Y,N,T
26	Q,E,R,S,V*,L*
27	I*,V,N,L,G
28	G
29	T*,V,N,R,S,I,H
30	C*
31	G*,S,V,L,F,H,R,I,Y
32	L*,V,G,M,T,H,A,F,S,R,E,P,D,K
33	P*,S,R,G,A,F,T,Y,K,L
34	G*,V,A,Q,P,R*,S,K,I,F,L
35	T,I,S,V,Q,L,F,G,A,R
36	K,R,P,N
37	C*
38	C*
39	K,R,Q,H
40	K*,R
41	P*

CONCLUSIONS

A new de novo protein design methodology, which incorporates true backbone flexibility (71) as defined by bounded continuous values of dihedral angles and Cα-Cα distances, is presented. Its application on full-sequence design of small proteins is demonstrated with the study of redesigning hβD-2, which is a 41-residue cationic peptide central to the defense of innate immune system against microbial attack. This study about hβD-2 also shows that the framework can serve as a useful predictive tool for screening peptide/protein drugs and speeding up their development process.

SUPPLEMENTARY MATERIAL

To view all of the supplemental files associated with this article, visit www.biophysj.org.

This work has not been reviewed by and does not represent the opinions of the United States Environmental Protection Agency.

C.A.F. gratefully acknowledges support from the National Science Foundation, the National Institutes of Health (grant No. R01 GM52032 and No. R24 GM069736), and the Environmental Protection Agency (grant No. GAD R 832721-010). D.M. gratefully acknowledges support from the National Science Foundation and the National Institutes of Health (grant No. R24 GM069736).

REFERENCES

1. Fernandez-Lopez, S., H. Kim, E. C. Choi, M. Delgado, J. R. Granja, A. Khasanov, K. Kraehenbuehl, G. Long, D. A. Weinberger, K. M. Wilcoxen, and M. R. Ghadiri. 2001. Antibacterial agents based on the cyclic D,L-α-peptide architecture. *Nature.* 412:452–455.
2. Nizet, V., T. Ohtake, X. Lauth, J. Trowbridge, J. Rudisill, R. A. Dorschner, V. Pestonjamasp, J. Piranino, K. Huttner, and R. L. Gallo. 2001. Innate antimicrobial peptide protects the skin from invasive bacterial infection. *Nature.* 414:454–457.
3. Zasloff, M. 2002. Antimicrobial peptides of multicellular organisms. *Nature.* 415:389–395.
4. Ganz, T. 2003. Defensins: antimicrobial peptides of innate immunity. *Nature.* 3:710–720.
5. Mygind, P. H., R. L. Fischer, K. M. Schnorr, M. T. Hansen, C. P. Sönksen, S. Ludvigsen, D. Raventós, S. Buskov, B. Christensen, L. D. Maria, O. Taboureau, D. Yaver, S. G. Elvig-Jørgensen, M. V. Sørensen, B. E. Christensen, S. Kjærulff, N. Frimodt-Moller, R. I. Lehrer, M. Zasloff, and H.-H. Kristensen. 2005. Plectasin is a peptide antibiotic with therapeutic potential from a saprophytic fungus. *Nature.* 437:975–980.
6. Loose, C., K. Jensen, I. Rigoutsos, and G. Stephanopoulos. 2006. A linguistic model for the rational design of antimicrobial peptides. *Nature.* 443:867–869.
7. Hancock, R. E., and A. Patrzykat. 2002. A Clinical development of cationic antimicrobial peptides: from natural to novel antibiotics. *Curr. Drug Targets Infect. Disord.* 2:79–83.
8. Niyonsaba, F., H. Ogawa, and I. Nagaoka. 2004. Human β-defensin-2 functions as a chemotactic agent for tumor necrosis factor-α-treated human neutrophils. *Immunology.* 111:273–281.
9. Schröder, J., and J. Harder. 1999. Molecules in focus: human β-defensin-2. *Int. J. Biochem. Cell Biol.* 31:645–651.
10. Ganz, T., and R. I. Lehrer. 1994. Defensins. *Curr. Opin. Immunol.* 6:584–589.
11. Dale, B., and S. Krisanaprakornkit. 2001. Defensin antimicrobial peptides in the oral cavity. *J. Oral Pathol. Med.* 30:321–327.
12. Dunsche, A., Y. Acil, R. Siebert, J. Harder, J. M. Schroder, and S. Jepsen. 2001. Expression profile of human defensins and antimicrobial proteins in oral tissues. *J. Oral Pathol. Med.* 30:154–158.
13. Harder, J., J. Bartels, E. Christophers, and J.-M. Schröder. 1997. A peptide antibiotic from human skin. *Nature.* 387:861.
14. Lehrer, R. I., T. Ganz, and M. E. Selsted. 1991. Defensins: endogenous antibiotic peptides of animal cells. *Cell.* 64:229–230.
15. Ouellette, A. J., and M. E. Selsted. 1996. Paneth cell defensins: endogenous peptide components of intestinal host defense. *FASEB J.* 10:1280–1289.
16. Yadava, P., C. Zhang, J. Sun, and J. A. Hughes. 2006. Antimicrobial activities of human β-defensins against *Bacillus* species. *Int. J. Antimicrob. Agents.* 28:132–137.
17. Klepeis, J. L., C. A. Floudas, D. Morikis, C. G. Tsokos, E. Argyropoulos, L. Spruce, and J. D. Lambris. 2003. Integrated structural, computational and experimental approach for lead optimization: design of compstatin variants with improved activity. *J. Am. Chem. Soc.* 125:8422–8423.
18. Klepeis, J. L., C. A. Floudas, D. Morikis, C. G. Tsokos, and J. D. Lambris. 2004. Design of peptide analogs with improved activity using a novel de novo protein design approach. *Ind. Eng. Chem. Res.* 43: 3817–3826.
19. Fung, H. K., S. Rao, C. A. Floudas, O. Prokopyev, P. M. Pardalos, and F. Rendl. 2005. 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.
20. Fung, H. K., M. S. Taylor, and C. A. Floudas. 2007. Novel formulations for the sequence selection problem in de novo protein design with flexible templates. *Optim. Methods Software.* 22:51–71.
21. Mineshiba, F., S. Takashiba, J. Mineshiba, K. Matsuura, and S. Kokeguchi. 2003. Antibacterial activity of synthetic human β-defensin-2 against periodontal bacteria. *J. Int. Acad. Periodontol.* 5:35–40.
22. Huang, G. T., H. B. Zhang, C. Yin, and S. H. Park. 2004. Human beta-defensin-2 gene transduction of dental pulp cells: a model for pulp antimicrobial gene therapy. *Int. J. Oral Biol.* 29:7–12.
23. Matsuzaki, K. 1999. Why and how are peptide-lipid interactions utilized for self-defense? Magainins and tachyplesins as archetypes. *Biochim. Biophys. Acta.* 1462:1–10.
24. Shai, Y. 1999. Mechanism of the binding, insertion and destabilization of phospholipid bilayer membranes by α-helical antimicrobial and cell non-selective membrane-lytic peptides. *Biochim. Biophys. Acta.* 1462:55–70.
25. Yang, L., T. M. Weiss, R. I. Lehrer, and H. W. Huang. 2000. Crystallization of antimicrobial pores in membranes: magainin and protegrin. *Biophys. J.* 79:2002–2009.
26. Yang, D., A. Biragyn, D. M. Hoover, J. Lubkowski, and J. J. Oppenheim. 2004. Multiple Roles of antimicrobial defensins, cathelicidins, and eosinophil-derived neurotoxin in host defense. *Annu. Rev. Immunol.* 22:181–215.
27. Klotman, M. E., and T. L. Chang. 2006. Defensins in innate antiviral immunity. *Nature Rev. Immunol.* 6:447–456.
28. Quiñones Mateu, M. E., M. M. Lederman, Z. Feng, B. Chakraborty, J. Weber, H. R. Rangel, M. L. Marotta, M. Mirza, B. Jiang, P. Kiser, K. Medvik, S. F. Sieg, and A. Weinberg. 2003. Human epithelial β-defensins 2 and 3 inhibit HIV-1 replication. *AIDS.* 17:F39–F48.
29. Sun, L., C. M. Finnegan, T. Kish-Catalone, R. Blumenthal, P. Garzino-Demo, G. M. L. T. Maggiore, S. Berrone, C. Kleinman, Z. Wu, S. Abdelwahab, W. Lu, and A. Garzino-Demo. 2005. Human β-defensins suppress human immunodeficiency virus infection: potential role in mucosal protection. *J. Virol.* 79:14318–14329.
30. Yin, C., H. N. Dang, F. Gazor, and G. T.-J. Huang. 2006. Mouse salivary glands and human β-defensin-2 as a study model for antimicrobial gene therapy: technical considerations. *Int. J. Antimicrob. Agents.* 28:352–360.
31. Fung, H. K., and C. A. Floudas. 2007. Computational de novo peptide and protein design: rigid templates versus flexible templates. *Ind. Eng. Chem. Res.* Accepted for publication.
32. Floudas, C. A., H. K. Fung, S. R. McAllister, M. Mönnigmann, and R. Rajgaria. 2006. Advances in protein structure prediction and de novo protein design: a review. *Chem. Eng. Sci.* 61:966–988.
33. Desmet, J., M. D. Maeyer, B. Hazes, and I. Lasters. 1992. The dead-end elimination theorem and its use in side-chain positioning. *Nature.* 356:539–542.
34. Goldstein, R. 1994. Efficient rotamer elimination applied to protein side-chains and related spin glasses. *Biophys. J.* 66:1335–1340.
35. Pierce, N., J. Spriet, J. Desmet, and S. Mayo. 2000. Conformational splitting: a more powerful criterion for dead-end elimination. *J. Comput. Chem.* 21:999–1009.
36. Wernisch, L., S. Hery, and S. Wodak. 2000. Automatic protein design with all atom force-fields by exact and heuristic optimization. *J. Mol. Biol.* 301:713–736.
37. Looger, L., and H. Hellinga. 2001. Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction

tractable: implications for protein design and structural genomics. *J. Mol. Biol.* 307:429–445.

38. Gordon, B., G. Hom, S. Mayo, and N. Pierce. 2003. Exact rotamer optimization for protein design. *J. Comput. Chem.* 24:232–243.

39. Koehl, P., and M. Delarue. 1994. Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *J. Mol. Biol.* 239:249–275.

40. Zou, J. M., and J. Saven. 2000. Statistical theory of combinatorial libraries of folding proteins: energetic discrimination of a target structure. *J. Mol. Biol.* 296:281–294.

41. Kono, H., and J. Saven. 2001. Statistical theory of protein combinatorial libraries: packing interactions, backbone flexibility, and the sequence variability of a main-chain structure. *J. Mol. Biol.* 306:607–628.

42. Koehl, P., and M. Levitt. 1999. De novo protein design I. In search of stability and specificity. *J. Mol. Biol.* 293:1161–1181.

43. Dantas, G., B. Kuhlman, D. Callender, M. Wong, and D. Baker. 2003. A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins. *J. Mol. Biol.* 332: 449–460.

44. Zou, J., and J. Saven. 2003. Using self-consistent fields to bias Monte Carlo methods with applications to designing and sampling protein sequences. *J. Chem. Phys.* 118:3843–3854.

45. Tuffery, P., C. Etchebest, S. Hazout, and R. Lavery. 1991. A new approach to the rapid determination of protein side chain conformations. *J. Biomol. Struct. Dyn.* 8:1267–1289.

46. Su, A., and S. Mayo. 1997. Coupling backbone flexibility and amino acid sequence selection in protein design. *Protein Sci.* 6:1701–1707.

47. Desjarlais, J., and T. Handel. 1999. Side chain and backbone flexibility in protein core design. *J. Mol. Biol.* 290:305–318.

48. Raha, K., A. Wollacott, M. Italia, and J. Desjarlais. 2000. Prediction of amino acid sequence from structure. *Protein Sci.* 9:1106–1119.

49. Ross, S., C. Sarisky, A. Su, and S. Mayo. 2001. Designed protein G core variants fold to native-like structures: sequence selection by ORBIT tolerates variation in backbone specification. *Protein Sci.* 10:450–454.

50. Larson, S. M., J. L. England, J. R. Desjarlais, and V. S. Pande. 2002. Thoroughly sampling sequence space: large-scale protein design of structural ensembles. *Protein Sci.* 11:2804–2813.

51. Larson, S. M., A. Garg, J. R. Desjarlais, and V. S. Pande. 2003. Increased detection of structural templates using alignments of designed sequences. *Proteins.* 51:390–396.

52. Kraemer-Pecore, C., J. Lecomte, and J. Desjarlais. 2003. A de novo redesign of the WW domain. *Protein Sci.* 12:2194–2205.

53. Kuhlman, B., G. Dantae, G. Ireton, G. Verani, B. Stoddard, and D. Baker. 2003. Design of a novel globular protein fold with atomic-level accuracy. *Science.* 302:1364–1368.

54. Saunders, C. T., and D. Baker. 2005. Recapitulation of protein family divergence using flexible backbone protein design. *J. Mol. Biol.* 346: 631–644.

55. Harbury, P., B. Tidor, T. Alber, and P. Kim. 1995. Repacking protein cores with backbone freedom: structure prediction for coiled coils. *Proc. Natl. Acad. Sci. USA.* 92:8408–8412.

56. Harbury, P., J. Plecs, B. Tidor, T. Alber, and P. Kim. 1998. High-resolution protein design with backbone freedom. *Science.* 282:1462–1467.

57. Plecs, J., P. B. Harbury, P. Kim, and T. Alber. 2004. Structural test of the parameterized-backbone method for protein design. *J. Mol. Biol.* 342:289–297.

58. Floudas, C. A., H. K. Fung, D. Morikis, M. S. Taylor, and L. Zhang. 2007. Overcoming the key challenges in de novo protein design: enhancing computational efficiency and incorporating true backbone flexibility. *In* Modeling of Biosystems: An Interdisciplinary Approach. R. Mondaini, editor. Springer Verlag., Heidelberg.

59. Krishnakumari, V., S. Singh, and R. Nagaraj. 2006. Antibacterial activities of synthetic peptides corresponding to the carboxy-terminal region of human β-defensins 1–3. *Peptides.* 27:2607–2613.

60. Hoover, D., K. Rajashankar, R. Blumenthal, A. Puri, J. Oppenheim, O. Chertov, and J. Lubkowski. 2000. The structure of human β-defensin-2 shows evidence of higher order oligomerization. *J. Biol. Chem.* 275: 32911–32918.

61. Pierce, N., and E. Winfree. 2002. Protein design is NP-hard. *Protein Eng.* 15:779–782.

62. Cornell, W. D., P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. 1995. A 2nd generation force-field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* 117:5179–5197.

63. Guntert, P., C. Mumenthaler, and K. Wuthrich. 1997. Torsion angle dynamics for NMR structure calculation with the new program DYANA. *J. Mol. Biol.* 273:283–298.

64. Guntert, P. 2004. Automated NMR structure calculation with CYANA. *J. Mol. Biol.* 278:353–378.

65. Loose, C., J. Klepeis, and C. Floudas. 2004. A new pairwise folding potential based on improved decoy generation and side chain packing. *Proteins Struct. Funct. Bioinformatics.* 54:303–314.

66. Tobi, D., and R. Elber. 2000. Distance-dependent pair potential for protein folding: results from linear optimization. *Proteins Struct. Funct. Bioinformatics.* 41:40–46.

67. Tobi, D., G. Shafran, N. Linial, and R. Elber. 2000. On the design and analysis of protein folding potentials. *Proteins Struct. Funct. Bioinformatics.* 40:71–85.

68. CPLEX. 1997. Using the CPLEX Callable Library. ILOG, Sunnyvale, CA.

69. Rajgaria, R., S. R. McAllister, and C. A. Floudas. 2006. A novel high resolution C^{α}-C^{α} distance dependent force field based on a high quality decoy set. *Proteins Struct. Funct. Bioinformatics.* 65:726–741.

70. Rajgaria, R., S. R. McAllister, and C. A. Floudas. 2007. Improving the performance of a high resolution distance dependent force field by including protein side chains. *Proteins Struct. Funct. Bioinformatics.* In press.

71. Floudas, C. A. 2005. Research challenges, opportunities and synergism in systems engineering and computational biology. *AIChE J.* 51:1872–1884.

72. Klepeis, J. L., C. A. Floudas, D. Morikis, and J. Lambris. 1999. Predicting peptide structures using NMR data and deterministic global optimization. *J. Comput. Chem.* 20:1354–1370.

73. Klepeis, J. L., and C. A. Floudas. 1999. Free energy calculations for peptides via deterministic global optimization. *J. Chem. Phys.* 110: 7491–7512.

74. Klepeis, J. L., and C. A. Floudas. 2003. Ab initio tertiary structure prediction of proteins. *J. Glob. Optim.* 25:113–140.

75. Klepeis, J. L., and C. A. Floudas. 2003. Prediction of β-sheet topology and disulfide bridges in polypeptides. *J. Comput. Chem.* 24:191–208.

76. Klepeis, J., and C. Floudas. 2003. Ab initio tertiary structure prediction of proteins. *J. Glob. Optim.* 25:113–140.

77. Klepeis, J., and C. Floudas. 2003. 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.

78. Androulakis, I., C. Maranas, and C. A. Floudas. 1997. Prediction of oligopeptide conformations via deterministic global optimization. *J. Glob. Optim.* 11:1–34.

79. Androulakis, I., C. Maranas, and C. A. Floudas. 1995. Alpha BB: a global optimization method for general constrained nonconvex problems. *J. Glob. Optim.* 7:337–363.

80. McDonald, C., and C. Floudas. 1995. Global optimization for the phase and chemical-equilibrium problem—application to the NRTL equation. *Comput. Chem. Eng.* 19:1111–1139.

81. Floudas, C., and P. Pardalos. 1995. State-of-the-art in global optimization—computational methods and applications—preface. *J. Glob. Optim.* 7:113–113.

82. Adjiman, C., I. Androulakis, C. Maranas, and C. A. Floudas. 1996. A global optimization method, αBB, for process design. *Computers Chem. Eng.* 20:S419–S424 (Suppl. A.).

83. Adjiman, C., I. Androulakis, and C. A. Floudas. 1997. Global optimization of MINLP problems in process synthesis and design. *Computers Chem. Eng.* 21:S445–S450 (Suppl. S.).

84. Maranas, C., and C. A. Floudas. 1997. Global optimization in generalized geometric programming. *Computers Chem. Eng.* 21:351–369.

85. Adjiman, C., I. Androulakis, and C. A. Floudas. 1998. A global optimization method, αBB, for general twice-differentiable constrained NPLs. I. Theoretical advances. *Computers Chem. Eng.* 22:1137–1158.

86. Adjiman, C., I. Androulakis, and C. A. Floudas. 2000. Global optimization of mixed-integer nonlinear problems. *AIChE J.* 46:1769–1797.

87. Ponder, J. 1998. TINKER, Software Tools for Molecular Design. Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, St. Louis, MO.

88. García, J., J. Florian, S. Schulz, A. Krause, F. Rodríguez-Jiménez, U. Forssmann, K. Adermann, E. Klüver, C. Vogelmeier, D. Becker, R. Hedrich, W. Forssmann, and R. Bals. 2001. Identification of a novel, multifunctional β-defensin (human β-defensin 3) with specific antimicrobial activity. *Cell Tissue Res.* 306:257–264.

89. Brooks, B. R., R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. 1983. CHARMM—a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* 4:187–217.

90. Dominy, B., and C. L. Brooks III. 1999. Development of a Generalized Born model parameterization for proteins and nucleic acids. *J. Phys. Chem.* 103:3765–3773.

91. Ryckaert, J. P., G. Ciccotti, and H. J. C. Berendsen. 1977. Numerical integration of Cartesian equations of motion of a system with constraints—molecular dynamics of *n*-alkanes. *J. Comput. Phys.* 23: 327–341.

92. Russ, W. P., and R. Ranganathan. 2002. Knowledge-based potential functions in protein design. *Curr. Opin. Struct. Biol.* 12:447–452.

93. Hoover, D., O. Chertov, and J. Lubkowski. 2001. The structure of human β-defensin-1. *J. Biol. Chem.* 276:39021–39026.

94. Dayhoff, M. O., R. M. Schwartz, and B. C. Orcutt. 1978. A model of evolutionary change in proteins. *In* Atlas of Protein Sequence and Structure, Vol. 5. M. O. Dayhoff, editor. National Biomedical Research Foundation, Washington, DC.

95. Schwartz, R. M., and M. O. Dayhoff. 1978. Matrices for detecting distant relationships. *In* Atlas of Protein Sequence and Structure, Vol. 5. M. O. Dayhoff, editor. National Biomedical Research Foundation, Washington, DC.