PHYSICS-BASED DESIGN OF PROTEIN-LIGAND BINDING

A DISSERTATION

SUBMITTED TO THE DEPARTMENT OF BIOCHEMISTRY

AND THE COMMITTEE ON GRADUATE STUDIES

OF STANFORD UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

F. Edward Boas

May 2008

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Pehr Harbury

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Dan Herschlag

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Tom Wandless

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Axel Brünger

Approved for the Stanford University Committee on Graduate Studies

Abstract

Different potential energy functions have been used in protein dynamics simulations, protein design calculations, and protein structure prediction. Clearly, the same physics applies in all three cases, so the variation in potential energy functions reflects differences in how the calculations are performed. With improvements in computer power and algorithms, the same potential energy function should be applicable to all three problems.

Here we show that a standard molecular-mechanics potential energy function without any modifications can be used to engineer protein-ligand binding. A molecular-mechanics potential is used to reconstruct the coordinates of various binding sites with an average root mean square error of 0.61 Å, and to reproduce known ligand-induced side-chain conformational shifts. Within a series of 34 mutants, the calculation can always distinguish weak ($K_d > 1$ mM) and tight ($K_d < 10$ μM) binding sequences. Starting from partial coordinates of the ribose binding protein lacking the ligand and the ten primary contact residues, the molecular-mechanics potential is used to redesign a ribose binding site. Out of a search space of 2×10^{12} sequences, the calculation selects a point mutant of the native protein as the top solution (experimental $K_d = 17$ μM), and the native protein as the second best solution (experimental $K_d = 210$ nM). The quality of the predictions depends on the accuracy of the generalized Born electrostatics model, treatment of protonation equilibria, high resolution rotamer sampling, a final local energy minimization step, and explicit modeling of the bound, unbound, and unfolded states.

After this initial proof of principle experiment, we next used a standard molecular mechanics potential energy function to redesign ribose binding protein to bind a series of ligands: L-arabinose, D-xylose, indole-3-acetic acid, and estradiol. The resulting proteins have 5 – 10 mutations from the native, are stable, the predicted structures have good hydrogen bonds and shape complementarity, and they use motifs similar to natural binding proteins. All of the designed proteins bind to their target ligands with measurable but weak affinity. The affinity was improved by random mutagenesis and screening.

The application of unmodified molecular-mechanics potentials to protein design links two fields in a mutually beneficial way. Design provides a new avenue to test molecular-mechanics energy functions, and future improvements in these energy functions will presumably lead to more accurate design results.

This is the first time a single model has been used to predict structures, binding constants, and to design new small-molecule binding sites. Using a standard model should improve the generality of protein design, which could enable the creation of custom proteins for a wide variety of applications, including sensors, enzymes, and protein therapeutics.

Acknowledgements

Pehr does the type of science that makes people say "wow, I had no idea you could do that" or "why didn't *I* think of that." He has the remarkable ability to invent crazy new ideas, and turn them into reality. He focuses on the key roadblocks in a field, and comes up with creative new approaches to solving them. That's why I joined Pehr's lab, and I have learned a lot from doing science with him.

The lab is a very social place, where everyone helps each other out. Lance has been my baymate for 5 years, and we've gone through many life changing events together. Since I met Lance, he got married and became a father. We've discussed the mysteries of the universe, and all of the popular culture that I missed while growing up. Jarrett and I are classmates in the MSTP program, and we joined the lab at the same time. He's a remarkably multi-talented person, and I've appreciated his good insights into both scientific and personal topics. Becky always knows the inside story, and she knows how to negotiate the best price for anything. She is a loyal friend and also has been a wonderful collaborator, generating crystals of our designed small molecule binding proteins. Rebecca has a good stories to tell, and is always eager to hear my stories as well. She is smart, curious, and quirky in the best possible way. Becky and Rebecca have both made the lab a much nicer place to work, making sure that everyone's birthday gets celebrated, and generally keeping the lab together socially. Jim and Erica also worked on computational protein design, and they were a good source of ideas. Jim developed the multi-state design framework that the lab

uses. Erica developed a continuous mean field algorithm for improving structural sampling.

Several collaborators have been indispensable for characterizing the designed binding proteins. Rebecca Fenn has spearheaded this effort, from collecting small angle X-ray scattering (SAXS) data to generating crystals. Pavel Strop in Axel Brunger's lab solved a crystal structure of one of the unbound designed proteins. Jan Lipfert from Seb Doniach's lab collected SAXS data on the designed proteins. John King in Suzanne Pfeffer's lab helped with protein purification.

I'd also like to thank my committee — Dan Herschlag, Axel Brunger, and Tom Wandless — for keeping me on track, and getting me focused on realistic goals. Dan Herschlag was particularly instrumental in this.

I've gotten great scientific advice from a few other people. Loren Looger and I talked broadly about protein design and he suggested ways to troubleshoot the design calculations. Michael Levitt and Buzz Baldwin provided encouragement and advice when I needed it.

My Mom and Dad have been a big source of wisdom over the years. My Mom encouraged my interest in science from an early age, and I remember all the times she helped me learn new things and meet interesting people. My mom taught me the value of family, hard work, and perseverence. She has a lot of willpower, doesn't take no for an answer, and I hope I have picked up some of that. My Dad taught me how to tinker and build things, and I enjoyed the times we used to spend together building things. Finally, Katie has been my best friend, source of encouragement, and she has taught me how to relax and enjoy the ride.

Acknowledgements

Thank you to everyone who helped make this possible! I couldn’t have done it without you all.

Table of contents

List of tables

List of figures

Chapter 1: Introduction

Proteins are the ultimate nanotechnology devices. Inside our cells, these molecular machines do all sorts of fantastic things — catalyze chemical reactions, create the electrical signals in neurons, copy DNA, move vesicles around, transmit information, and generally get the job done inside our bodies. Many of these functions have been studied and worked out in great detail, a triumph of modern molecular biology.

The inimitable Richard Feynman, a great source to consult for practical scientific philosophy, wrote that "What I can not create I can not understand." So in that spirit, we believe that if we claim to understand how proteins work, we should be able to make predictions about their behavior, and if we really understand how they work, we should be able to design proteins with new functions. Thus, we employ an engineering approach to protein design, meaning that we start with a physical model of how proteins work, and we use this model to predict how mutations affect a protein's activity, and to design proteins with new functions.

Relationship to the protein folding problem

The protein folding problem asks if you can predict the structure of a protein, given its amino acid sequence. The protein design problem asks if you can find an amino acid sequence that folds into a particular structure. Which problem is harder? Can you solve one without solving the other?

We believe that proteins can be designed computationally without solving the protein folding problem, and that this approach will provide unique insights into how proteins work. From one perspective, design is easier than folding: out of the millions of sequences that fold into your target structure, you only need to find one of them. From another perspective, design is harder than folding: the design algorithm will exploit any flaws in your protein energy calculations if they appear to stabilize the target structure.

Design goal

The overall goal is to develop an algorithm that will take the structure of a scaffold protein, and the structure of a small molecule, and design a set of mutations needed to create a binding site in the scaffold. We only consider mutations at a limited number of "design positions"; the rest of the protein simply serves as a rigid structure for constraining the conformational flexibility of the designed binding site. Thus, we only need to consider a limited range of protein conformations and do not need to solve the full protein folding problem.

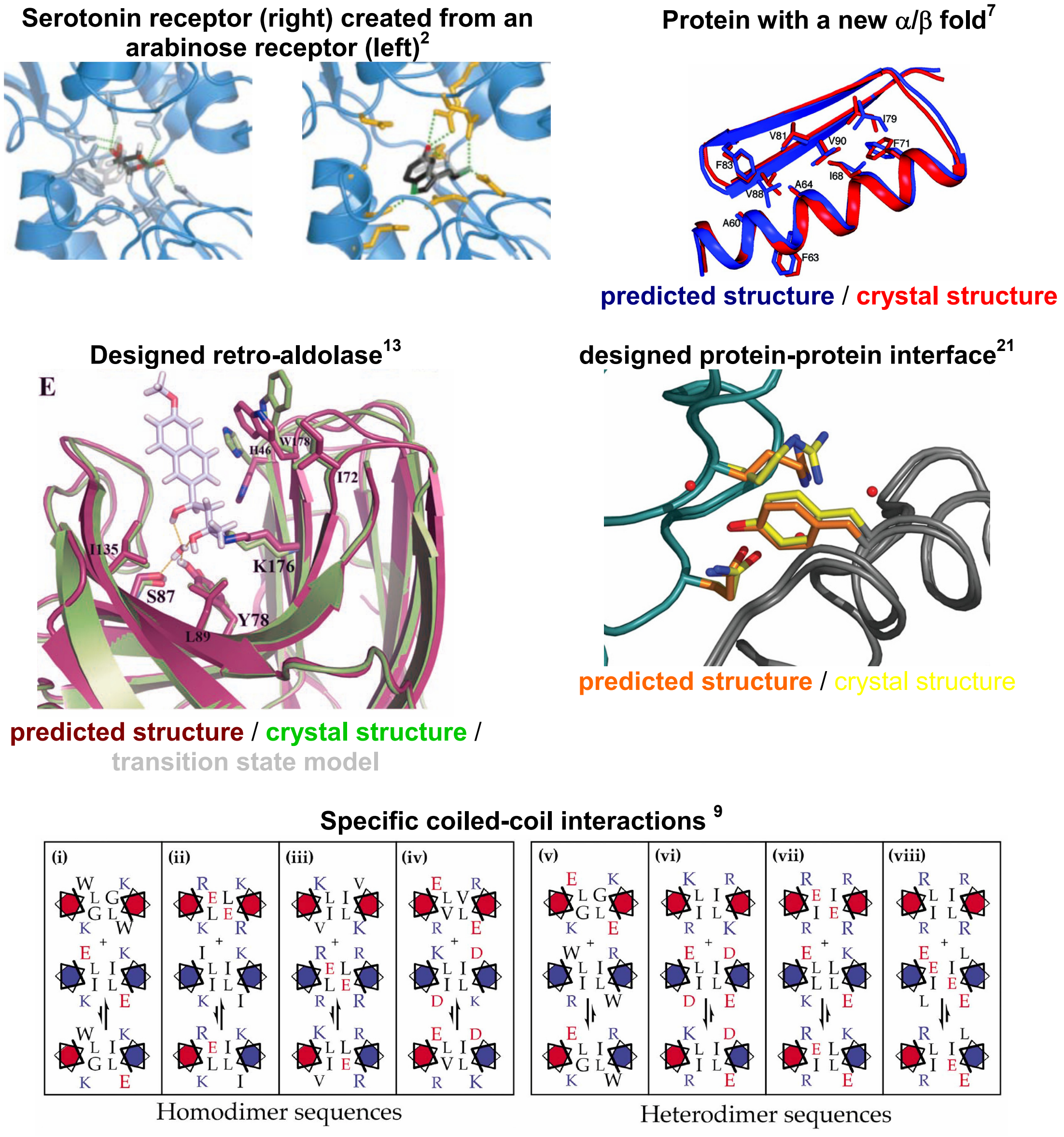
The ideal scaffold protein can host a wide range of different binding sites. It should be stable, so it can accommodate destabilizing mutations. The proposed binding site should also be lined with sidechain contacts, which will be easier to modify by mutation than backbone contacts. Natural scaffold proteins include antibodies, which bind different antigens, and alpha/beta barrel proteins, which host a wide range of enzyme active sites.[1] In this thesis, we use ribose binding protein, based on the pioneering work of Hellinga,[2,3] who used this as a scaffold for computational protein design (although note that ref. [3] was recently retracted[4]).

Innovative aspects of the research

The computational protein design field is small, but has seen some remarkable successes.[2,5-13] In each of these examples, the protein was designed in a computer, then experimentally validated with a crystal structure or activity measurement (Figure 1). However, the generality of these design algorithms is unclear, because each lab typically uses its own custom software, and standard models for protein design have not yet emerged. Furthermore, protein design typically requires multiple iterations of feedback from experimental results before reaching the desired target. We would like to address these problems using better sampling strategies and more accurate energy functions.

Many of the individual components of our design calculation are related to algorithms that have been proposed and validated before in the literature, including the molecular mechanics potential energy function,[14] continuum solvent model,[15,16] side chain rotamer library,[17] ligand docking procedure,[18] probabilistic description of protein conformation and mean field algorithm,[19] multi-state design framework,[9] and genetic algorithm.[20]

The unique aspect of this project is the combination of these existing methods to tackle interesting design problems. This integration has never been achieved before, partly because of the technical difficulty, and partly because others in this field have broken down the design problem into different pieces. Especially notable is our treatment of ligand flexibility, protonation equilibria, conformational entropy, high resolution structural sampling, accurate continuum solvation model, and explicit unbound and unfolded reference states. These factors are often ignored in protein design calculations, despite evidence that they are important in ligand binding.


Serotonin receptor (right) created from an arabinose receptor (left)[2]

Protein with a new α/β fold[7]

predicted structure / crystal structure

Designed retro-aldolase[13]

predicted structure / crystal structure / transition state model

designed protein-protein interface[21]

predicted structure / crystal structure

Specific coiled-coil interactions [9]

Figure 1. Examples of computational protein design.

Models for understanding molecular recognition

Qualitatively, specific molecular recognition occurs between molecules with steric complementarity and chemical complementarity (charge patterning and hydrogen bonding)[22] (see Figure 2). Steric complementarity is important both for producing favorable van der Waals interactions between ligand and receptor, and also for preventing the formation of buried pockets of water. Charge patterning is important for specificity. In water, intermolecular interactions are often strengthened when salt bridges are replaced with uncharged groups. Thus, charge patterning does not always create the most stable complexes. It does, however, confer specificity: a single charge buried in a molecular interface without a salt bridge partner is extremely unfavorable, thus preventing ligands with the wrong charge pattern from binding.

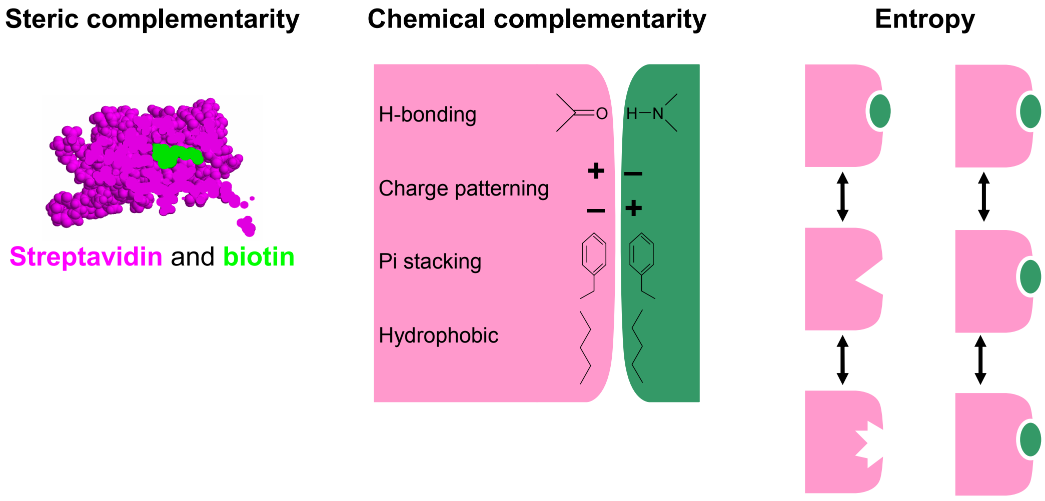


Figure 2. Qualitative model of molecular recognition.

While a qualitative understanding is the most intuitive way to think about molecular recognition, it does have limitations. Each molecular interface contains many favorable and unfavorable interactions, so it is difficult to know whether the interaction is net favorable without adding up the actual strength of each interaction. Furthermore, protein backbone and side chain conformations shift in response to mutations, and these conformational changes are difficult to predict by just eyeballing the structure.

If qualitative understanding is one extreme, the other extreme is a full quantum mechanical treatment of the protein and solvent. From this perspective, the energy of the system is solely based on Coulomb's law interactions between electrons and nuclei, and the evolution of the system is given by the Schrödinger equation. In addition to being totally impractical computationally, this approach also clouds our understanding of the system: all of the energy is electrostatic and is not partitioned into more understandable categories.

Fortunately, over the past thirty years, several groups have developed molecular mechanics potential energy functions (Figure 6), which model proteins as a collection of atoms connected by springs that hold bond lengths and angles near their standard values. A torsion angle energy term penalizes eclipsed conformations. Standard molecular mechanics potentials also include two energy terms for atoms that are not bonded to each other: van der Waals interactions, and Coulomb's law interactions between charges. Some versions also include a hydrogen-bonding term, but often this is handled by the van der Waals and Coulomb's law terms.

These molecular mechanics potentials are typically used in molecular dynamics simulations, which trace protein motions over time. At each time step, the computer uses the molecular mechanics potential to calculate a force on each atom, then uses the force to update its velocity and position. Unfortunately, most of the computer time is spent simulating water molecules, even though the protein is usually the molecule of interest. Water does significantly affect the behavior of the protein, but fortunately many of these effects can be understood in a smeared-out continuum model. First, water solvates charges: it exerts an attractive force pulling both positive and negative charges towards the protein's surface. Second, water screens charges: interactions between charges in a vacuum are weakened by a factor of 80 when they are placed in water. Third, water's hydrogen bond network is disrupted at the protein surface: this is why oil does not dissolve in water. The first two factors can be treated by modeling water as a continuum dielectric, and the third factor can be treated with a surface tension term.

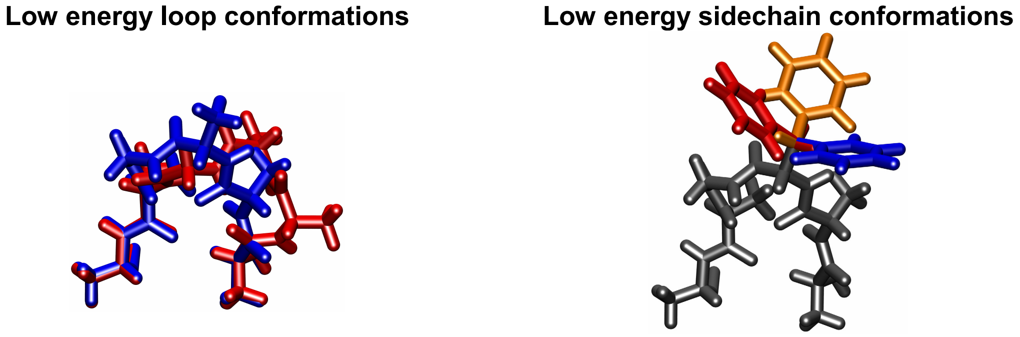
Finding a protein's low energy conformations

Even when water is treated in a continuum model, molecular dynamics is extremely slow, because the simulation typically proceeds in femtosecond time steps. Simulating a 1 second unbinding event would take 1 million years of time on a typical desktop computer.

Most of the time in a molecular dynamics simulation, the protein is simply jiggling around some equilibrium conformation. Every once in a while, the protein crosses an energetic barrier and settles into another conformation. To speed up the calculation, can we just skip these barriers and directly identify the low energy conformations? Our strategy for doing this is to break the protein-ligand system into three parts: protein backbone conformation, protein side chain conformation, and ligand position / conformation. Then, we can find the low energy configurations of each part, and mix and match these low energy conformations to explore the whole system's conformational space.

To find the low energy conformations of the protein backbone at design positions, we use the following rather elaborate scheme (all of the simpler strategies we tried missed some conformations). First, we take snapshots from a high temperature molecular dynamics simulations. Second, we sample the backbone ϕ and ψ angles on a 30° grid. Third, we search the entire Protein Data Bank for loops whose endpoints match the fixed portions of our protein scaffold. Finally, we feed all of these conformations into a genetic algorithm search, which randomly perturbs and splices together structures to generate new structures.

To find the low energy protein side chain conformations, we clustered side chain conformations observed in the Protein Data Bank into a rotamer library, placed rotamers at each design position, and applied an energy cutoff to eliminate unfavorable rotamers (Figure 3).



Low energy loop conformations

Low energy sidechain conformations

Figure 3. Low energy protein conformations.

To find the low energy docked ligand positions / conformations, we move the ligand over a translational and rotational grid, and eliminate ligand orientations that clash with the scaffold, or do not make sufficient contact with side chains at design positions.

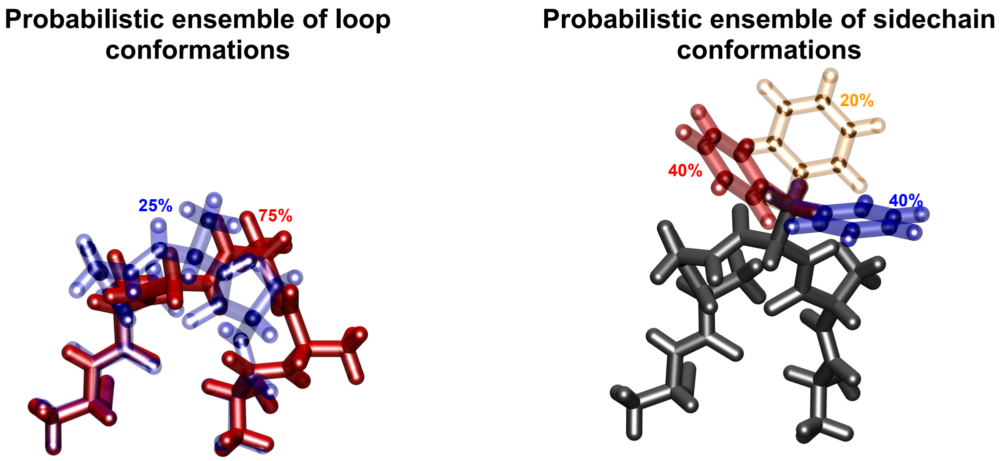
One advantage of identifying all of the low-energy configurations of the protein-ligand system at the beginning of the calculation is that all of the energy terms can be precomputed (both intrinsic energies and interaction energy matrices). These are calculated using a standard molecular mechanics potential with continuum water, as described in the previous section. After making this initial investment of computing time, the energy of a specific configuration of the protein-ligand system can be rapidly calculated simply by adding up the appropriate terms from the precomputed energy matrices.

Probabilistic description of protein conformation

We represent the protein/ligand system as a probabilistic ensemble of the low energy backbone, side chain, and ligand conformations/positions (Figure 4). The probabilities are set by a mean field calculation,[19] which iteratively updates the probability of each side chain and ligand conformation based on its intrinsic energy plus its probability weighted interaction energies.

The probabilistic ensemble allows us to model conformational changes and thermal fluctuations. Mutating design positions to different amino acids may shift the protein conformation, and this shift will be described as a change in the probabilities of various conformations. A single rigid structure would be described by setting the probability of a single conformation to 1 and everything else to 0. However, in order to accurately calculate ligand binding affinity, it is important to model the protein's thermal fluctuations by allowing probabilities besides 0 and 1. The probabilistic model is more realistic, because proteins spend very little time in their global minimum potential energy conformation. Furthermore, if the protein can adopt 1000 conformations and only 1 of them binds ligand, then this decreases its binding affinity by 1000-fold. We can model these sorts of entropic effects using a probabilistic model.

Once the probabilities from the mean field calculation have converged, we can calculate the free energy of the system by adding the probability weighted average energy, and the energy due to the conformational entropy.


Probabilistic ensemble of loop conformations

Probabilistic ensemble of sidechain conformations

Figure 4. Probabilistic model of protein structure.

Evolving binding sites

An important point is that, experimentally, we can only control the amino acid sequence. The structure of the protein is determined by the energetics. Thus, we separate our sequence and structural optimization.

Using the procedure described above, we can calculate the free energy of several different states, given the amino acid sequence at design positions: the protein and ligand free in solution, the protein bound to various different ligands, and the unfolded protein. From these energies, we can calculate the stability of the protein, and its affinity and specificity for various ligands. Based on these energies, we can assign a score to each sequence. The score can also include structural criteria, such as the predicted geometry of catalytic residues.

We use a genetic algorithm[20] to evolve sequences that optimize our chosen scoring function. The genetic algorithm starts with a population of random sequences, and then alternates between rounds of selection and recombination/mutation (Figure 24 on p. 85).

Chapter 2: Potential energy functions for protein design

This chapter has been adapted from: Boas FE and Harbury PB. (2007) "Potential energy functions for protein design." *Current Opinion in Structural Biology*. 17: 199-204.

Summary

Different potential energy functions have been used in protein dynamics simulations, protein design calculations, and protein structure prediction. Clearly, the same physics applies in all three cases, so the variation in potential energy functions reflects differences in how the calculations are performed. With improvements in computer power and algorithms, the same potential energy function should be applicable to all three problems. Recently improved models of polarization, the hydrophobic effect, and hydrogen bonding may be applicable to both molecular mechanics and protein design.

Introduction

Computational protein design algorithms use models of protein energetics to engineer protein sequences with new functions. This is similar to more established branches of engineering, such as circuit simulation or stability analysis of buildings, where accurate computer models are used to evaluate designs before they are built. Protein design provides a rigorous test of the energetic model that is used, because the design algorithm must pick functional sequences out of an astronomically large space of non-functional sequences.

As with any calculation, there is a tradeoff between accuracy and speed when modeling or designing proteins. For example, simulation of a one-second dissociation event using a molecular dynamics calculation with explicit water would require 10 million years on a typical desktop computer. Protein design algorithms use several strategies to speed up the process. First, protein design algorithms do not simulate kinetics, but rather calculate the energies of a small number of target states (these energies are used as a surrogate for the free energies of conformational neighborhoods). Many fast algorithms exist for optimizing the structure of each target state. Second, protein design calculations do not explicitly model water, but rather use a continuum representation of water. Finally, protein design algorithms generally use less computationally intensive energy functions than molecular mechanics calculations.

Previous reviews have described potential energy functions (PEFs) used for molecular mechanics simulations,[23,24] protein design,[25,26] and protein structure prediction.[27] In this review, we compare these energy functions (Figure 5). We also describe advances in the molecular mechanics field that could be used in the next generation of design algorithms.

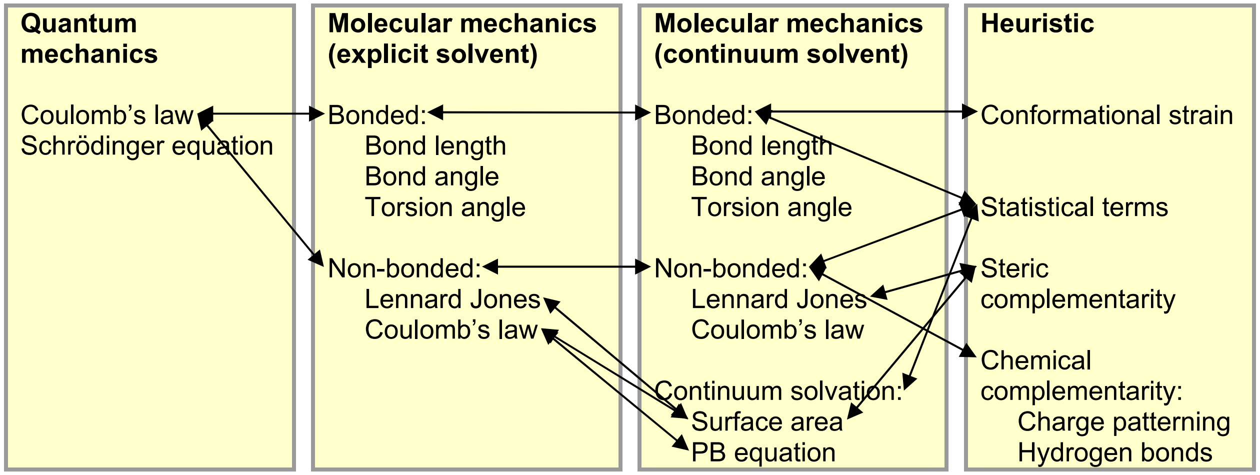


Figure 5. Proteins can be modeled at different levels of detail.

Potential energy functions for evaluating protein conformations range from quantum mechanics, which is accurate but very slow, to more heuristic energy functions that include statistical terms. In between are molecular mechanics potential energy functions, which are the most thoroughly tested models of molecular energetics. Currently, the protein design field uses heuristic energy functions, but the trend is towards using more physically based potential energy functions.

Potential energy functions

Overview

Molecular mechanics potential energy functions (MM-PEFs) incorporate two types of terms: “bonded” and “non-bonded” (Figure 6). The bonded terms apply to sets of 2 to 4 atoms that are covalently linked, and they serve to constrain bond lengths and angles near their equilibrium values. The bonded terms also include a torsional potential that models the periodic energy barriers encountered during bond rotation. The non-bonded terms consist of the Lennard-Jones function (which includes van der Waals attraction, and repulsion due to orbital overlap), and Coulomb’s law. The parameters for the bonded and non-bonded terms of an MM-PEF are derived from quantum calculations, and from thermodynamic, crystallographic, and spectroscopic data on a wide range of systems.[23,24] MM-PEF’s have been used predominately to simulate protein folding and dynamics, and are also used to refine X-ray crystal structures.

An alternative type of potential energy function is the knowledge-based, or statistical, energy function[27,28] (Figure 7). This type of energy function derives from the database of known protein structures. The probabilities that residues appear in specific configurations (such as rotamer conformations, or buried vs. surface environments), or the probabilities that pairs of residues appear together in a defined relative geometry is calculated. These probabilities are converted into an effective potential energy using the Boltzmann equation: $\Delta G = -RT \ln(p_{obs}/p_{exp})$, where p_{obs} is the probability of seeing a particular structural element, and p_{exp} is the expected probability of seeing that structural element based on chance.[29-31] The advantage of a knowledge-based energy function is that it can model any behavior seen in known protein crystal structures, even if no good physical understanding of the behavior exists. The disadvantage is that these energy functions are phenomenological and can't predict new behaviors absent from the training set.

Design potentials include a combination of MM-PEF, knowledge-based, and other terms. In contrast to MM-PEFs, which have become fairly standardized, design potentials vary enormously between labs. The various terms are typically calibrated and weighted to optimize performance for one type of prediction, such as experimental binding energy,[12,32] or to produce native-like sequences when redesigning natural proteins.[7] By way of illustration, we describe the potential energy functions used in two recent landmark protein design papers. In the first example, Looger *et al.* redesigned various bacterial periplasmic binding proteins to bind trinitrotoluene, lactate, and serotonin.[2] Their energy function included a Lennard-Jones term (using CHARMM22 parameters[14]) with the repulsive component scaled down to 35%, a Coulombic term with a distance-dependent dielectric constant of 8.0*r* and partial charges from CHARMM22, an explicit hydrogen bonding term derived from the DREIDING MM-PEF,[33] a surface area-based solvation term, a knowledge-based rotamer term,[34] and a term requiring all hydrogen bond donors and acceptors to be satisfied. In a subsequent paper, Dwyer *et al.* designed *de novo* triosephosphate isomerase activity into ribose binding protein,[3] using a more accurate electrostatics model that included multiple geometry-dependent dielectric constants.[35] In the second example, Kuhlman *et al.* designed a 93-residue protein with a new α/β fold.[7] Their energy function included an LJ term (with radii fit to match the distribution of distances seen in the PDB, and well depths from CHARMM19), a Lazaridis-Karplus empirical solvation term,[36] a knowledge-based hydrogen bonding term,[37] a knowledge-based rotamer term, and a knowledge-based pairwise residue interaction term. The scaling factors for each term were adjusted to optimize recovery of native sequences when redesigning a training set of 30 proteins.

Why are MM-PEFs and design PEFs so different, and why do the latter include so many *ad hoc* terms? The basic answer is that design PEFs must compensate for an incomplete simulation of protein behavior: many degrees of freedom are either ignored, modeled implicitly, or sampled at low resolution. We examine this question term-by-term in the following sections.

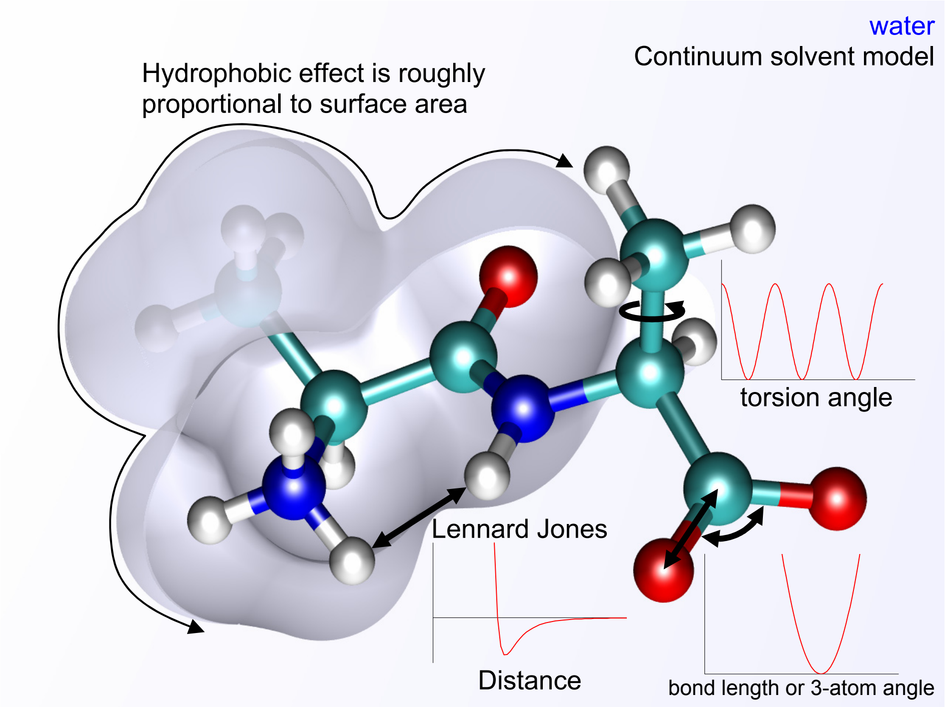

water
Continuum solvent model
Hydrophobic effect is roughly proportional to surface area
torsion angle
Lennard Jones
Distance
bond length or 3-atom angle

Figure 6. Molecular mechanics potential energy function with continuum solvent.

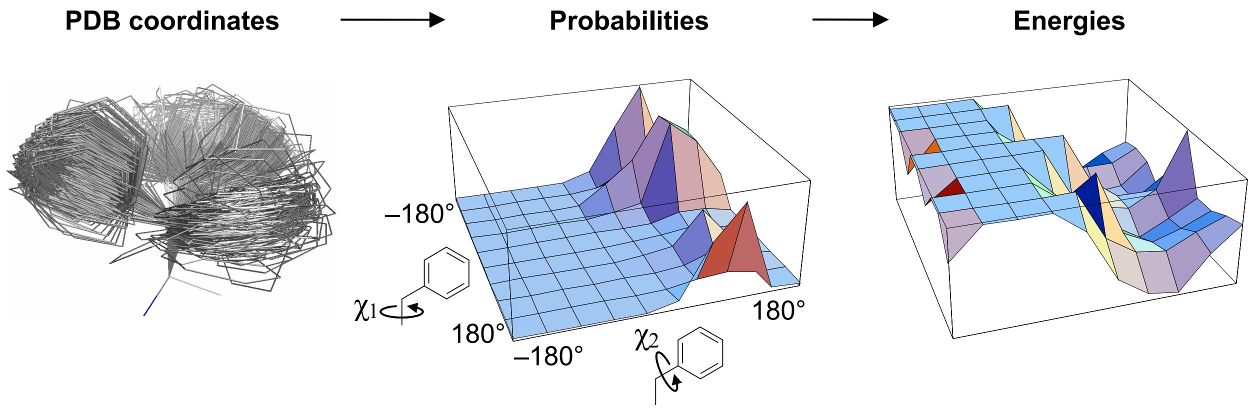


Figure 7. Knowledge-based potential energy function.

Bonded terms

Although it is straightforward to directly use the bonded portion of MM-PEFs to determine the relative energies of different rotamer geometries, design potentials have tended to use fixed rotamer coordinates and knowledge-based rotamer potentials. MM-PEF bonded energies vary greatly with small changes in bond lengths and angles. Thus, these energies are not meaningful unless the structures have first been locally energy minimized (perhaps with dihedral angle restraints).

Lennard-Jones

The Lennard-Jones (LJ) function includes a weakly attractive component at long distances (the van der Waals energy), and a strongly repulsive component at short distances. The repulsive component is sensitive to small atomic displacements: the LJ energy of a protein crystal structure can decrease by hundreds of kcal/mol upon local energy minimization, despite imperceptible changes in the atomic coordinates.

The discrete rotamer sampling used for protein design calculations inevitably leads to small atomic overlaps, producing large unfavorable Lennard Jones energies. In many cases, the overlaps could be eliminated by local minimization, but such minimization cannot be readily incorporated into combinatorial sequence design algorithms. Instead, the functional form of the LJ interaction is almost always softened so that overlaps are less energetically unfavorable. For example, the LJ radii can be scaled down,[38] the repulsive component of the LJ energy can be scaled down,[2] or the LJ function can be linearly extrapolated below a cutoff distance.[7]

Softening the LJ function is based on a presumption that protein cores are reasonably fluid and thus can always rearrange to accommodate small overlaps. However, this modification always leads to qualitative and quantitative errors in interaction energies. For example, modern MM-PEFs model hydrogen bonds as a combination of an electrostatic interaction and an LJ interaction. When overlaps are allowed, atoms can approach more closely, producing artificially favorable hydrogen bond energies. In general, changing the LJ parameters in any way will destroy the delicate balance engineered into an MM-PEF. Use of unmodified LJ functions for protein design will require either very high resolution discrete sampling, or some form of continuous optimization.

Solvation

Computing the energy of a protein embedded in explicit solvent molecules is time consuming, because the energy must be averaged over many solvent configurations. To speed up calculations, solvent can instead be modeled as a smooth continuous material with a characteristic dielectric constant and surface tension. The solvation energy of such protein continuum-solvent systems is generally separated into two components. The first component is the hydrophobic effect, which accounts for the interfacial free energy of the uncharged protein and the continuum solvent. The second component is the solvent polarization energy, which accounts for the interaction of partial charges in the protein with dipoles and ion clouds induced in the solvent. Charged atoms closer to the protein's surface have more favorable solvation energies and smaller apparent charge-charge interactions.

Both the LJ function and Coulomb's law are pairwise factorable, meaning that the total energy can be expressed as a sum of interactions between pairs of atoms without regard to the position of any other atom in the system. This is important, because the total energy can then be determined by summing precalculated pairwise interaction energies (required for most rapid structural optimization procedures). Solvation energies, on the other hand, are not inherently pairwise factorable. The interaction between two charges depends on the positions of other atoms, because the other atoms displace solvent and salt.

Hydrophobic effect

The continuum hydrophobic effect has traditionally been modeled as being proportional to the solvent accessible surface area of a solute.[39] Pairwise-factorable approximations of surface area have been developed for use in design calculations.[40] Although widely applied, the surface area-based model has clear limitations. For example, hydrophobic solutes in water can interact favorably when they are separated by a single layer of water molecules.[41] This type of interaction is completely absent from a surface-area based energy. Wagoner and Baker have developed a model[42] of the hydrophobic effect that captures such complex wetting phenomena, and produces energies that are closer to explicit solvent simulations than are surface-area based energies. Their energy function includes a term proportional to surface area, a term proportional to volume, and a solute-solvent van der Waals term. Adapting this improved model for protein design work will require either the development of a pairwise-factorable approximation, or the use of a design algorithm that does not require precalculated energies.

Solvent polarization

Solvent polarization is very difficult to simulate quickly and accurately. Consequently, many different empirical models that subsume polarization energies have been used in protein design efforts.[34-36,43] These models commonly include a solvation energy for charged atoms based on accessible surface area, and a Coulomb's law term with a distance-dependent dielectric constant. The surface area models disregard the non-zero contributions of fully buried charges to the polarization energy. The distance-dependent dielectric constant scales down Coulomb's law to account for screening of charge-charge interactions by water. However, it ignores the fact that screening depends on the local environment of each charge.

A more physical approach is to solve the Poisson-Boltzmann (PB) differential equation[44] that describes the relationship between fixed charge and the electric potential in a continuum dielectric environment. Water is assigned a dielectric constant of 80, the protein interior is typically assigned a dielectric constant between 1 and 20, and the molecular surface defines the boundary between protein and solvent. Values of the electric potential on a spatial grid can be obtained using a finite-difference algorithm. Marshall *et al.*[45] describe a pairwise-factorable approximation to the PB equation based on summing precalculated energies for single residues and for pairs of residues. However, this treatment does not take into account rotamer-conformation dependent changes in the protein-solvent boundary, or that solutions to the PB equation are not truly superimposable.

Alternatively, the generalized Born equation[15] provides a fast approximate solution to the Poisson-Boltzmann equation, and it has been used for protein design.[46] Recent improvements to the generalized Born functional form[47,48] yield solvation energies that are comparable to those derived from finite-difference calculations.[49]

Explicit water

Continuum solvent models break down when water molecules are tightly bound to proteins. However, it may be possible to incorporate a handful of explicit water molecules in a continuum solvent calculation. Schymkowitz *et al.* developed a method for predicting positions of tightly bound water molecules in proteins.[50] Jiang *et al.* show how to incorporate water molecules into amino acid rotamers.[51]

Hydrogen bonds

In an MM-PEF, hydrogen bonds are typically modeled as dipole-dipole interactions. The optimal geometry for a dipole-dipole interaction, for example between the C=O and N-H dipoles in the protein backbone, places all four atoms in a straight line. However, the charge distribution around the carbonyl oxygen adopts a trigonal sp^2 arrangement, which is not spherically symmetrical. The sp^2 lone-pair geometry should favor a bent hydrogen bond. Morozov *et al.* showed that the bent geometry is indeed preferred according to quantum calculations and crystal structures in the PDB.[52] Using the PDB statistics, they developed a knowledge-based hydrogen bonding energy function[37,53] and used it to design a new protein.[7]

Solute polarization and quantum effects

A widely recognized limitation of MM-PEFs is that they assume fixed atomic charges, and do not model environment-dependent rearrangement of charge on a solute. Recently developed polarizable force fields address this limitation by allowing the electric field to induce dipoles at each atom.[54,55] Importantly, solute polarization breaks down the pairwise-factorability property of traditional MM-PEFs. MM-PEFs also do not model chemical realities such as bond formation, partial covalent character of hydrogen bonds, and lone pairs. One possible compromise is to model key parts of the protein using quantum mechanics, and the rest of the protein using molecular mechanics.[56,57]

Reference states

Protein design potentials frequently use implicit reference states. The MM-PEF can only tell the energy difference between different conformations of the same sequence. To compare different sequences, we must subtract the energy of each sequence in an alternative undesired conformation, such as the unfolded or unbound states. These undesired conformations are typically treated implicitly by subtracting a fixed reference energy for each amino acid.

The unfolded and unbound states can also be modeled explicitly. For example, the unfolded state can be modeled using fixed reference energies for each amino acid, plus a random walk model of long range electrostatics.[58] The unbound state can be modeled explicitly using the same structural optimization algorithm used on the bound state.

Modeling the correct reference states is critical to calculating the binding energy of a complex. For example, the binding energy due to a salt bridge or hydrogen bond is the interaction energy of the charges in the bound conformation, relative to their interaction energies with water in the unbound conformation. A typical salt bridge might have a Coulomb interaction energy of 50 kcal/mol, but this is almost completely canceled out by the charges interacting with water in the unbound state. Thus, accurate calculations of the energies of both the bound and unbound structures are needed to calculate accurate binding energies. In many cases, salt bridges are actually destabilizing relative to a hydrophobic interaction[59]: the charges would prefer to interact with water than with each other.

Search algorithms

Three major algorithms have been used to search through sequence and conformational space in protein design. Many variations and hybrid algorithms are possible, but here we describe a typically implementation of each algorithm, and briefly discuss the advantages and disadvantages of each.

The dead-end elimination (DEE) algorithm[34,60] starts with a set of rotamers at each position in the protein, and a precalculated matrix of interaction energies between these rotamers. The algorithm uses a series of filters to eliminate rotamers that provably can not be present in the global energy minimum. Typically, a large fraction of the rotamers can be eliminated, and the remaining rotamers are searched by exhaustive enumeration or Monte Carlo search. The advantage of DEE is that very large sequence and structural spaces can be searched comprehensively. Sequence and structural space are typically searched simultaneously, which requires the use of an implicit reference state. In other words, undesired conformations such as the unfolded or unbound states are typically not modeled explicitly, but rather are treated using fixed reference energies for each amino acid. Thus, for example, dead end elimination does not distinguish between intramolecular and intermolecular interactions, and will propose mutations that stabilize the protein without improving its interaction with the ligand.[61] Reference states could be included if DEE were only used for structural optimization of single sequences, with another procedure used for sequence optimization. This is typically not done because it is much faster to optimize sequence and structure simultaneously.

The mean field algorithm[19,62] also uses a rotamer-based picture with precomputed energy matrices. However, rather than finding a single low energy structure, mean field treats the protein as a probabilistic ensemble. Each rotamer is assigned a probability, and these probabilities are updated iteratively to match the Boltzmann distribution. The final probabilities can be used to calculate the protein's conformational entropy. The mean field algorithm is typically used to optimize the structure of a single sequence, and the sequence optimization is typically done using a genetic algorithm. The advantage of this approach is that undesired conformations such as the bound, unbound, and unfolded states can be modeled explicitly. The use of multiple states allows for stability, affinity, and specificity to be explicitly calculated and optimized.

Monte Carlo methods[63-65] typically start with a single protein structure, and use a set of moves to perturb this structure. If the new structure has a lower energy, then it is accepted. If the new structure has higher energy, then it is accepted with probability $e^{-\Delta E/RT}$, where ΔE is the energy change, and T is the temperature, which is slowly annealed to 0. The advantage of this approach is that there is no need to precompute large energy matrices. Thus, it is CPU-intensive rather than memory-intensive, which better matches today's distributed computing systems. Furthermore, the energy function can include non-pairwise additive terms such as polarization. The Monte Carlo moves can include randomly switching from one rotamer conformation to another, but they can also include non-rotameric moves.

The Baker lab has developed a clever strategy for including backbone flexibility in protein design[7,66]. They alternate between sequence design on a fixed backbone, and structural optimization for a designed sequence.

Conclusions and future directions

The techniques described above have been used to design proteins with a wide variety of new functions. Clark *et al.*[60] optimized the recombining site of an antibody to increase the ligand affinity, and Lazar *et al.*[10] optimized the Fc region of an antibody to bind more tightly to the Fc receptor. Ashworth *et al.*[64] redesigned an endonuclease to recognize and cut a heterologous DNA sequence. Kuhlman *et al.*[65]

designed a protein that reversibly switches between two distinct protein folds with a change in pH or cobalt concentration.

These examples illustrate the diverse range of useful functions already accessible by protein design. As potential energy functions, search algorithms, and computational power continue to improve, protein design should become a standard and general research tool.

Acknowledgements

This work was supported by the National Institutes of Health (GM068126-01 to P.B.H.). F.E.B. was supported by a training grant from the National Institute of General Medical Sciences (5T32 GM07365-28).

Chapter 3: Physics-based design of protein-ligand Binding

This chapter has been adapted from:

Boas FE and Harbury PB. (2008) “Physics-based design of protein-ligand binding.” *Journal of Molecular Biology*. In press.

Summary

While the molecular-mechanics field has standardized on a few potential energy functions, computational protein design efforts are based on potentials that are unique to individual labs. Here we show that a standard molecular-mechanics potential energy function without any modifications can be used to engineer protein-ligand binding. A molecular-mechanics potential is used to reconstruct the coordinates of various binding sites with an average root mean square error of 0.61 Å, and to reproduce known ligand-induced side-chain conformational shifts. Within a series of 34 mutants, the calculation can always distinguish weak (K_d > 1 mM) and tight (K_d < 10 µM) binding sequences. Starting from partial coordinates of the ribose binding protein lacking the ligand and the ten primary contact residues, the molecular-mechanics potential is used to redesign a ribose binding site. Out of a search space of 2×10^{12} sequences, the calculation selects a point mutant of the native protein as the top solution (experimental K_d = 17 µM), and the native protein as the second best solution (experimental K_d = 210 nM). The quality of the predictions depends on the accuracy

of the generalized Born electrostatics model, treatment of protonation equilibria, high resolution rotamer sampling, a final local energy minimization step, and explicit modeling of the bound, unbound, and unfolded states. The application of unmodified molecular-mechanics potentials to protein design links two fields in a mutually beneficial way. Design provides a new avenue to test molecular-mechanics energy functions, and future improvements in these energy functions will presumably lead to more accurate design results.

Introduction

Computer-aided design of a ligand binding site is similar to solving a 3D jigsaw puzzle: it involves fitting together the right pieces (amino acid mutations) to create a properly shaped and functionalized pocket for a ligand. The inputs to the design procedure are the crystal structure of a scaffold protein, a ligand structure, and a set of amino-acid positions that will be mutated to create the binding site. The orientations of candidate jigsaw-puzzle pieces are determined by modeling the conformations that the ligand and surrounding amino acids can adopt, so as to identify the lowest energy arrangement. The design procedure searches through thousands of candidate sequences for one that optimizes the computed binding free energy of the ligand with the protein. The whole process depends heavily on the potential energy function (PEF), a mathematical expression embodying the physical laws that govern the protein-ligand and solvent system.

Over the past 30 years, potential energy functions have played a central role in the molecular-mechanics field. This field has converged on a small set of standard PEF's that have been extensively tested.[67] Identifying and correcting the limitations of these energy models is an area of active research.[54,55,68] The modern molecular-mechanics potential energy functions (MM-PEF's) treat proteins as a collection of atoms with partial charges and van der Waals parameters, connected by springs that maintain bond lengths and angles. The parameters are derived from quantum calculations and from experimental data on a wide range of systems.[23] MM-PEF's have been used to calculate binding constants[69-73], protein folding kinetics[74], protonation equilibria[75], and active site coordinates[71,76,77].

Perhaps surprisingly, standard MM-PEF's are not used for protein design.[78] The reason is that computing energies using MM-PEF's requires significant computer time and is very sensitive to detailed atom positions, necessitating fine conformational sampling. When thousands of different sequences must be evaluated, the computation time per sequence becomes critical. In order to accelerate calculations, design algorithms typically use simplified PEFs with various *ad hoc* energy terms[2,3,7-12,60,64,65,76,79,80] (heuristic potential energy functions are also often used to predict binding constants[81,82] and to predict active site coordinates[83]). Water is treated in a simplified way, for example by inserting a distance dependent dielectric constant into Coulomb's law, and by applying a surface-area based solvation energy.[2,3] The van der Waals interaction is frequently smoothed so that it is less sensitive to spatial position, and thus can be optimized with coarse sampling.[2,3,7] Rather than explicitly modeling reference states, such as the unfolded state, the reference states are typically treated implicitly by modifying the PEF.[2,3,7] Statistical terms derived by counting how frequently different residues and functional groups interact in crystal structures, are included as well.[2,3,7] Relative weights for the various energy terms are adjusted empirically so as to match experimental data.[7,12] Similar approximations were used in the early days of molecular-mechanics calculations, but were replaced as better models and increased computational power became available.

There are several motivations for trying to identify a single, standardized energy function that is practically useful for protein design. First, design results from different labs could be compared, and those results would collectively address where the energy model had failed and how to improve it. Second, the practice of computational protein design would be simplified if PEF development were not required. Finally, a PEF that had been broadly validated might be expected to generalize better to new design problems than would a customized PEF.

One reasonable choice for a universal energy function would be an MM-PEF. MM-PEF's are the most broadly tested PEF's,[67] and a direct correspondence exists between them and more rigorous quantum-mechanical treatments of matter.[23] A large group of scientists work on MM-PEF's, and the advances they make would be directly applicable to design. Here, we test whether protein-ligand binding sites can be successfully designed based on a standard MM-PEF that does not include any heuristic corrections. We first describe how we directly apply an MM-PEF to the protein design problem, and then detail various tests applied to the ribose binding protein.

Results

Design scheme

Using the genetic algorithm,[20] we search through thousands of sequences to find one sequence that maximizes the calculated protein-ligand dissociation energy without destabilizing the protein by more than 5 kcal/mol. To evaluate dissociation and unfolding energies, the bound, unbound, and unfolded states are modeled, and their calculated energies are differenced. For each state, we use a mean field rotamer-repacking algorithm to find the atomic coordinates that minimize the energy. As part of the rotamer repacking, titratable residues are allowed to protonate or deprotonate depending on the local energetics. Good structural sampling is achieved by using extremely large rotamer libraries (≥ 5449 rotamers per position), and several thousand ligand poses that sample the translational, rotational, and internal degrees of freedom of the ligand. The optimal structure generated by rotamer repacking is then subjected to gradient-based energy minimization. The energies of each state are evaluated with the unmodified CHARMM22 molecular-mechanics potential energy function[14] and the generalized Born solvation formalism[15] developed by Lee et al.[16] The design procedure is outlined in Figure 8. To evaluate the approach, we apply three tests: structural prediction, energetic prediction, and prediction of a binding site sequence.

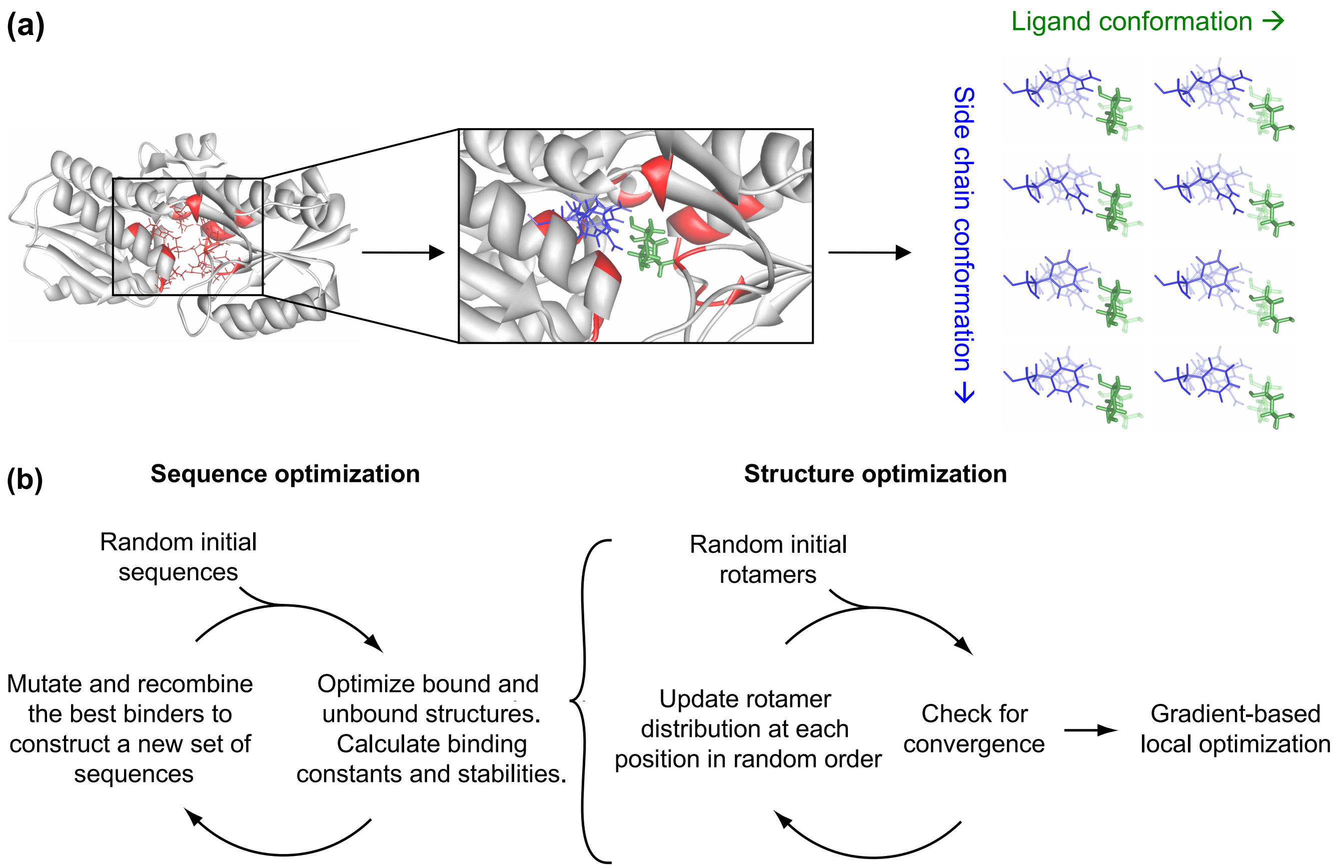

(a)
Ligand conformation →
Side chain conformation →
(b)
Sequence optimization
Structure optimization
Random initial sequences
Optimize bound and unbound structures. Calculate binding constants and stabilities.
Mutate and recombine the best binders to construct a new set of sequences
Random initial rotamers
Check for convergence
Update rotamer distribution at each position in random order
Gradient-based local optimization

Figure 8. Simplified schematic of the protein design algorithm.

(a) Setting up a design calculation. The design calculation is based on a scaffold protein (gray) with a known crystal structure, and a set of design positions (red). Possible ligand poses (green) and side chain conformations (blue) for each amino acid at each position are constructed. The right panel shows multiple side chain rotamers modeled at one design position, and two alternative ligand poses. Interaction energies between the possible ligand poses and the possible side chain conformations are precomputed. (b) Running a design calculation. The design procedure involves separate sequence optimization (to find sequences that bind ribose) and structural optimization (to determine the binding constant and stability of each sequence). In the RBP-ribose redesign, we search a space of 2×10^{12} sequences and an average of 5×10^{28} conformations per sequence.

Structural prediction

For structural predictions, we started with crystal structures and discarded the coordinates of the ligand and all contacting side chains. These coordinates were then predicted in the context of the rest of the protein. We first explored the effect of sampling resolution by predicting the structure of ribose binding protein (RBP) bound to ribose using four rotamer libraries of increasing size (Figure 9). With fewer than 5449 rotamers per position, the calculated energy of the predicted structure is less favorable than the calculated energy of the crystal structure, indicating that the crystal structure conformation is missed due to inadequate sampling resolution. At 5449 rotamers per position, the predicted structure has the same energy as the energy-minimized crystal structure, and the coordinates differ by a root mean square (RMS) error of 0.148 Å. This level of accuracy exceeds the experimental error in the crystallographic coordinates. This apparently surprising result likely occurs because the fixed portion of the crystallographic coordinates constrains the possible solutions at the modeled positions. However, this constraint alone is not sufficient to specify the binding site sequence and geometry (see below).

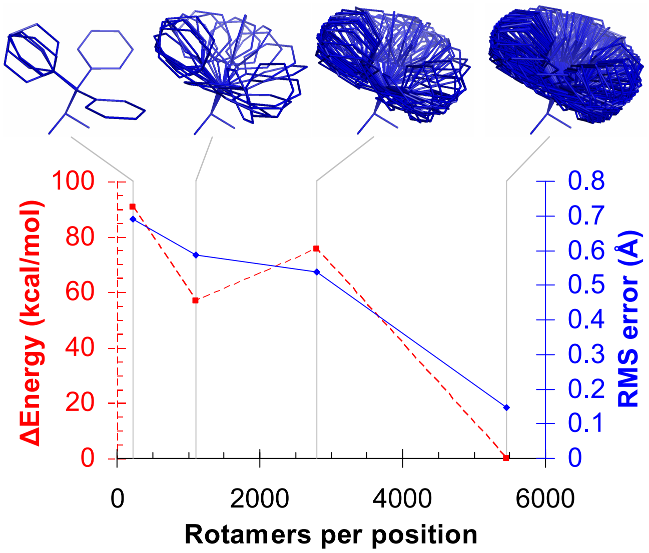


Figure 9. Higher rotamer resolution improves structural predictions for the RBP binding site (PDB code: 2DRI).

Δ Energy is the difference in potential energy between the calculated structure and the crystal structure, after both have been subjected to local energy minimization. RMS error is the root-mean-square deviation between the calculated and crystallographic coordinates of the repacked atoms, comprising the ligand and ten active site side chains. The phenylalanine rotamers from each rotamer library are shown to illustrate the sampling resolution. The lowest resolution rotamer library shown is the Richardson penultimate rotamer library[17] with protonation states added for His, Asp, and Glu. The other rotamer libraries were derived by clustering side chain conformations in high resolution crystal structures from the Protein Data Bank (see p. 75).

Using high resolution rotamer libraries (either 5449 or 6028 rotamers per position), side chains in the binding sites of 5 different structures were predicted with an average RMS error of 0.61 Å (Figure 10 & Figure 11). The number of predicted residues ranged from 9 to 23. The error was generally larger for surface residues, and when more positions were predicted.

For the RBP-ribose calculations, we restricted the ligand poses to be within 1.8 Å RMS of the native pose, resulting in the 4639 poses shown in Figure 10. For the

ABP-arabinose calculations, the ligand poses were restricted to be within 1.0 Å RMS of the native pose, resulting in the 4111 poses shown in Figure 10. Although we would have preferred to do the calculations without this filter, it was necessary to reduce the number of ligand poses to a manageable number (the precalculated interaction energy matrices had to be smaller than 2 GB to fit into memory).

We explicitly model the bound and unbound states, providing predictions of side-chain conformational shifts upon binding. The predicted changes match the crystal structures in 70% of the residues with the largest conformational shifts (Figure 12). Single-state design algorithms ignore such conformational shifts, in contrast to a multi-state design framework.[9] Note that we did not predict the *backbone* shift upon binding (4.1 Å RMS for RBP and 0.8 Å RMS for VEGF) because the bound and unbound backbone coordinates were used as inputs to the calculation.

The calculation predicts that one aspartic acid and one glutamic acid in the binding site of ABP are protonated (Table 1). If these residues are not allowed to protonate, the structural prediction is degraded (Figure 13).

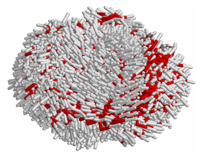
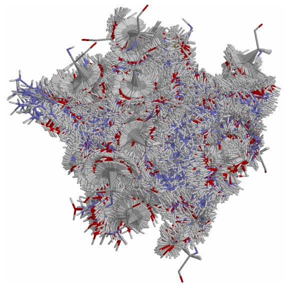
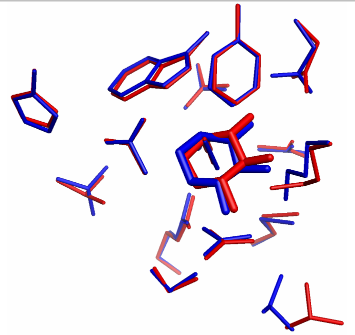
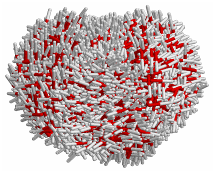
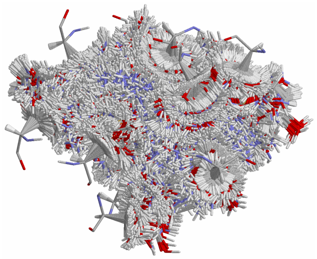
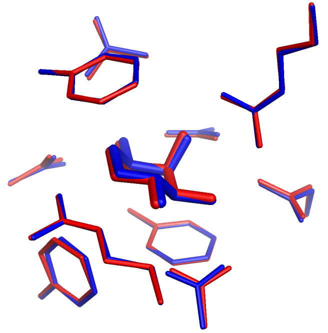
	Ligand poses	Side chain rotamers	crystal structure / predicted structure	RMS error
ABP-arabinose				0.677 Å
RBP-ribose				0.148 Å

Figure 10. Prediction of binding site coordinates.

Starting from crystal structures stripped of the ligand and the contacting residues, the active site was reconstructed by finding the lowest energy arrangement of the ligand and side chains. For ABP-arabinose (PBD code: 6ABP), the coordinates of the arabinose and 15 contacting residues (10, 14, 16, 17, 64, 89, 90, 108, 145, 147, 151, 204, 205, 232, 259) were predicted using 6028 rotamers per position and 4111 ligand poses. For RBP-ribose (PDB code: 2DRI), the coordinates of ribose and 10 contacting residues (13, 15, 16, 89, 90, 141, 164, 190, 215, 235) were predicted using 5449 rotamers per position, and 4639 ligand poses.

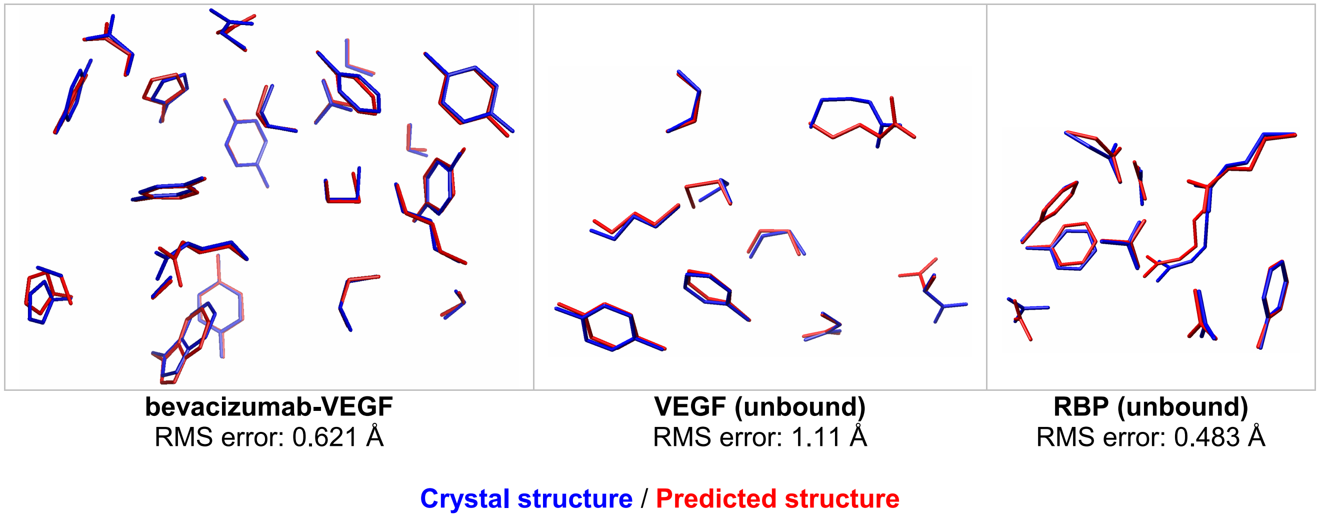


Figure 11. Prediction of binding site coordinates for bevacizumab-VEGF (1BJ1), unbound VEGF (2VPF), and unbound RBP (1URP). For bevacizumab-VEGF, the following 23 residues were repacked, using 6028 rotamers per position: V17, V21, W48, W79, W81, W82, W83, W91, W93, H28, H30, H31, H32, H54, H55, H99, H101, H102, H103, H105, H106, H107, H108. V and W are VEGF chains, H and L are antibody heavy and light chains. For unbound VEGF and RBP, the same set of residues were predicted as the bound structure.

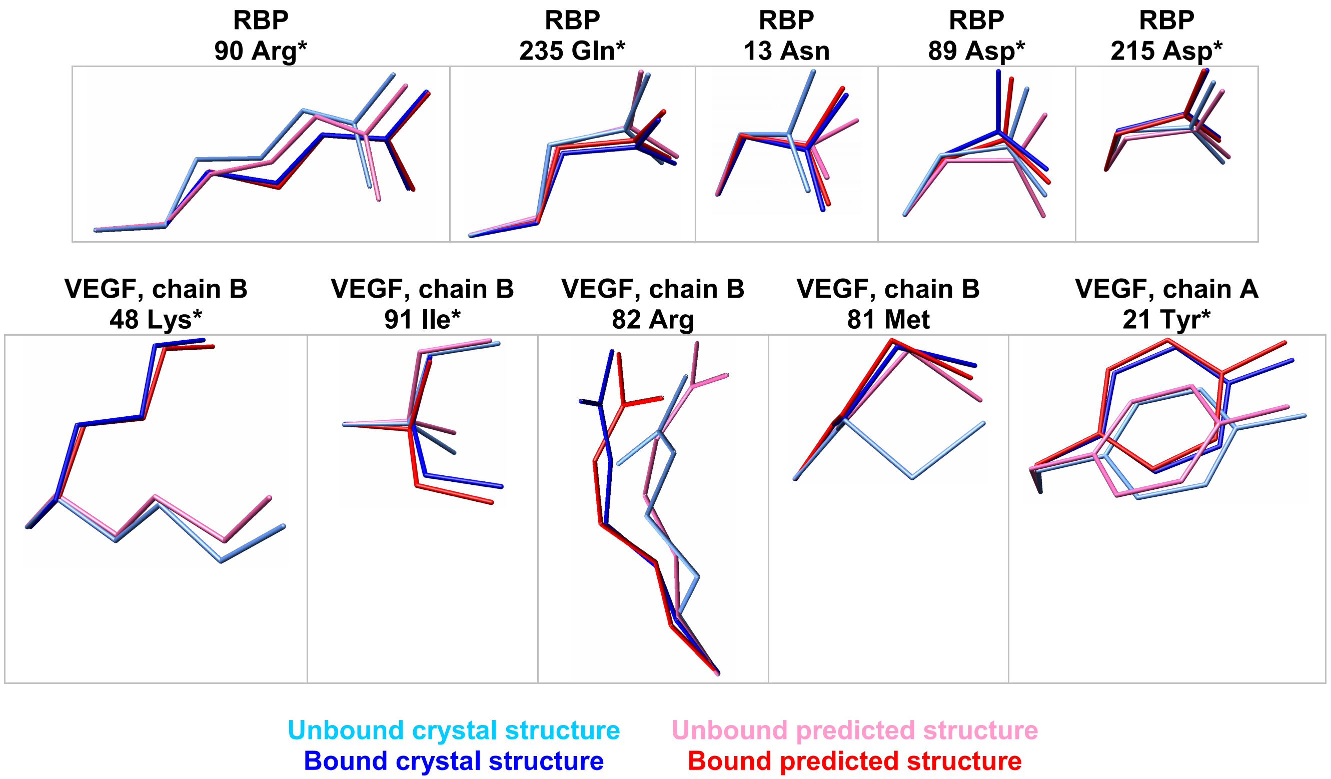


Figure 12. Prediction of side chain conformational shifts in RBP upon binding ribose, or VEGF upon binding bevacizumab. The five largest experimentally observed conformational shifts are shown for each protein. The residues were superimposed by aligning the backbone amide nitrogen, alpha carbon, and carbonyl carbon. * denotes correct predictions, where the unbound/bound predictions are closest to the unbound/bound crystallographic coordinates, respectively.

ABP-arabinose (6ABP)

	Protonation state	
Residue	bound	unbound
14	GUP	GUP
89	APP	APP
90	ASP	ASP
259	HSD	HSD

bevacizumab-VEGF (1BJ1, 2VPF)

	Protonation state	
Residue	bound	unbound
W93	GLU	GLU
H101	HSD	HSD
H107	HSD	HSD

RBP-ribose (2DRI, 1URP)

	Protonation state	
Residue	bound	unbound
89	ASP	ASP
215	ASP	ASP

Table 1. Predicted protonation states.

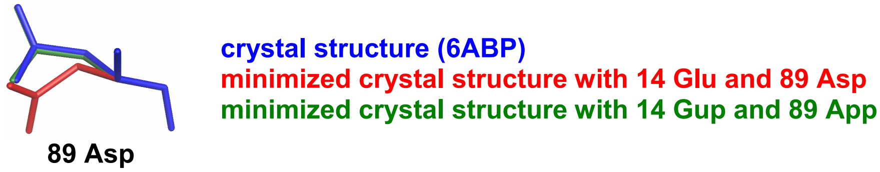

crystal structure (6ABP)
minimized crystal structure with 14 Glu and 89 Asp
minimized crystal structure with 14 Gup and 89 App
89 Asp

Figure 13. In ABP-arabinose, 14 Glu and 89 Asp must be protonated to maintain the crystal structure coordinates under local minimization. If they are deprotonated, then the coordinates for 89 Asp shift out of position.

Energetic prediction

To test if the energy function can properly rank the binding affinities of different binding site sequences, we first computed ligand binding energies for the native ABP and RBP sequences and for 1000 scrambled sequences. As expected, none of the scrambled sequences have better predicted stability and dissociation energy than the native (Figure 14a).

Next, we calculated the relative binding energies of 34 mutants of ABP for which dissociation energies have been measured. Two sequences were predicted to destabilize the protein by more than 10 kcal/mol relative to native ABP, and presumably adopt alternative backbone conformations. The binding energies of the remaining sequences are predicted with a correlation coefficient of r^2=0.57 (Figure 14b, Table 2). The predictions were performed without any adjustable parameters. As each calculation required about 1 minute of CPU time on a Pentium processor, the approach is fast enough for design applications. The data set includes single, double, and triple point mutants of wild type ABP, and covers a wide range of mutation types (hydrophobic to hydrophobic, hydrophobic to polar/charged, polar/charged to hydrophobic, and polar/charged to polar/charged).

Within the data set, the calculation can always distinguish weak (K_d > 1 mM) and tight (K_d < 10 μM) binding sequences. However, the absolute dissociation energies are not predicted correctly. One important possible source of error is that there is no published crystal structure of unbound ABP. We model the unbound protein backbone conformation based on the crystal structure of bound ABP. In reality, the unbound protein likely exists in an open conformation with better solvated binding-site residues.[84] Our incorrect unbound state might explain the 21.2 kcal/mol offset in calculated dissociation energies. The slope of the regression line is greater than one, which is likely due to modes of structural relaxation (such as backbone motions) that were not modeled. The resulting clashes will exaggerate any energy differences between sequences. Another possibility is that we are not adequately modeling entropy losses upon binding.[85]

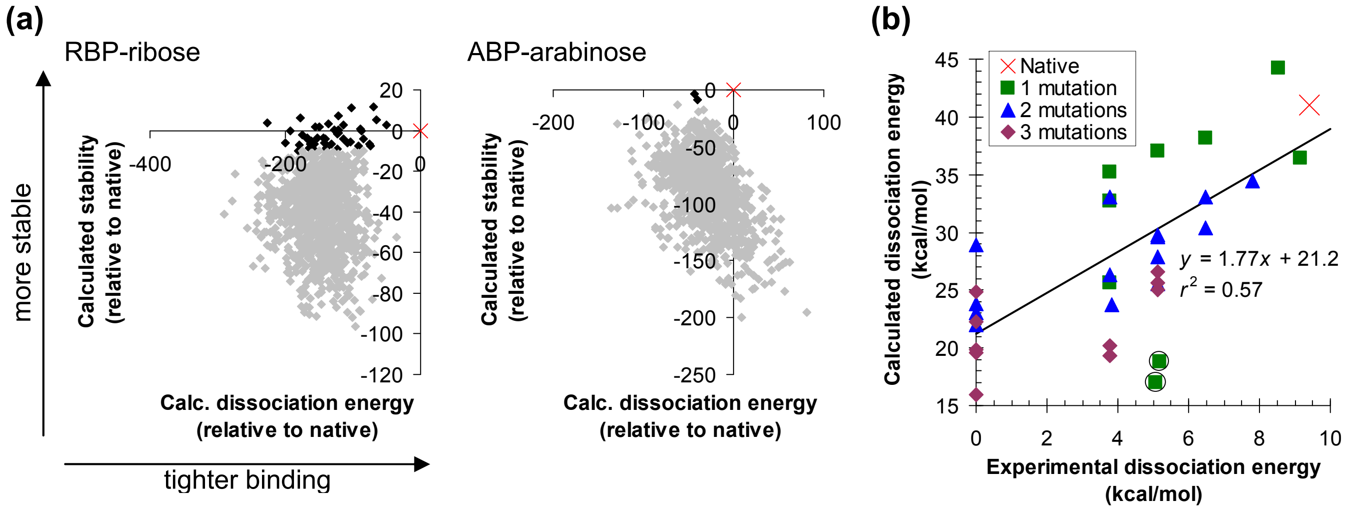


Figure 14. Predicting dissociation energies.

(a) Calculated stability and dissociation energy distinguish the native sequence (×) from 1000 scrambled sequences (♦) for ABP and RBP. Sequences predicted to be more then 10 kcal/mol destabilized relative to the native are shown in gray. (b) Predicting relative dissociation energies of mutants. The graph shows data on mutants of ABP binding to arabinose. Experimental data are from reference [86] and from measurements reported in Table 2. An experimental dissociation energy of zero means that there was no detectable binding. Calculations were performed using 6ABP as the scaffold structure for both the bound and unbound states, with 6028 rotamers per position. Coordinates of the fifteen primary ligand contacts and of residues 20 and 235 were optimized. The circled points are predicted to be destabilized by more than 10 kcal/mol relative to the native.

Experimental		**Calculated**		**Sequence**																
Dissoc. energy		**Dissoc. energy**	**Stability vs. native**																	
(kcal/mol)	**Source**	**(kcal/mol)**	**(kcal/mol)**	**10**	**14**	**16**	**17**	**20**	**64**	**89**	**90**	**108**	**145**	**147**	**151**	**204**	**205**	**232**	**235**	**259**
9.40	2	40.98	0.00	LYS	GLU	TRP	PHE	GLU	CYS	ASP	ASP	MET	LEU	THR	ARG	MET	ASN	ASN	ASP	HIS
9.15	3	36.45	1.64	LYS	GLU	TRP	PHE	GLU	CYS	ASP	ASP	MET	LEU	**SER**	ARG	MET	ASN	ASN	ASP	HIS
8.53	2	44.22	-6.62	LYS	GLU	TRP	PHE	GLU	CYS	ASP	ASP	**LEU**	LEU	THR	ARG	MET	ASN	ASN	ASP	HIS
7.81	3	34.47	-0.16	LYS	GLU	TRP	PHE	GLU	CYS	ASP	ASP	MET	**VAL**	**SER**	ARG	MET	ASN	ASN	ASP	HIS
6.47	3	38.16	0.36	LYS	GLU	TRP	PHE	GLU	CYS	ASP	ASP	MET	**VAL**	THR	ARG	MET	ASN	ASN	ASP	HIS
6.47	3	33.07	-5.50	LYS	GLU	TRP	PHE	GLU	CYS	ASP	ASP	MET	**VAL**	**ALA**	ARG	MET	ASN	ASN	ASP	HIS
6.47	3	30.43	1.54	LYS	GLU	TRP	PHE	GLU	CYS	ASP	ASP	MET	**ALA**	**SER**	ARG	MET	ASN	ASN	ASP	HIS
5.18	1	18.80	-17.56	LYS	GLU	TRP	**TRP**	GLU	CYS	ASP	ASP	MET	LEU	THR	ARG	MET	ASN	ASN	ASP	HIS
5.13	3	37.01	1.16	**ASN**	GLU	TRP	PHE	GLU	CYS	ASP	ASP	MET	LEU	THR	ARG	MET	ASN	ASN	ASP	HIS
5.13	3	29.75	0.66	**ASN**	GLU	TRP	PHE	GLU	CYS	ASP	ASP	MET	LEU	**SER**	ARG	MET	ASN	ASN	ASP	HIS
5.13	3	29.63	-0.38	**VAL**	GLU	TRP	PHE	GLU	CYS	ASP	ASP	MET	LEU	**SER**	ARG	MET	ASN	ASN	ASP	HIS
5.13	3	27.86	-1.87	LYS	GLU	TRP	PHE	GLU	CYS	ASP	ASP	MET	**ALA**	**ALA**	ARG	MET	ASN	ASN	ASP	HIS
5.13	3	26.59	1.08	**VAL**	GLU	TRP	PHE	GLU	CYS	ASP	ASP	MET	**VAL**	**SER**	ARG	MET	ASN	ASN	ASP	HIS
5.13	3	25.67	0.76	**ASN**	GLU	TRP	PHE	GLU	CYS	ASP	ASP	MET	**VAL**	**SER**	ARG	MET	ASN	ASN	ASP	HIS
5.13	3	25.50	3.85	**GLN**	GLU	TRP	PHE	GLU	CYS	ASP	ASP	MET	LEU	**SER**	ARG	MET	ASN	ASN	ASP	HIS
5.13	3	25.07	3.44	**GLN**	GLU	TRP	PHE	GLU	CYS	ASP	ASP	MET	**VAL**	**SER**	ARG	MET	ASN	ASN	ASP	HIS
5.07	1	17.00	-10.95	LYS	**ILE**	TRP	PHE	GLU	CYS	ASP	ASP	MET	LEU	THR	ARG	MET	ASN	ASN	ASP	HIS
3.83	3	23.77	5.02	LYS	GLU	TRP	PHE	GLU	CYS	ASP	**ALA**	MET	LEU	**SER**	ARG	MET	ASN	ASN	ASP	HIS
3.79	3	35.24	1.19	**VAL**	GLU	TRP	PHE	GLU	CYS	ASP	ASP	MET	LEU	THR	ARG	MET	ASN	ASN	ASP	HIS
3.79	3	33.08	-1.52	LYS	GLU	TRP	PHE	GLU	CYS	ASP	ASP	MET	**ASP**	**ALA**	ARG	MET	ASN	ASN	ASP	HIS
3.79	3	32.72	2.31	LYS	GLU	TRP	PHE	GLU	CYS	ASP	**ALA**	MET	LEU	THR	ARG	MET	ASN	ASN	ASP	HIS
3.79	3	26.34	7.93	LYS	GLU	TRP	PHE	GLU	CYS	ASP	ASP	MET	**ASP**	**SER**	ARG	MET	ASN	ASN	ASP	HIS
3.79	3	25.60	6.73	**GLN**	GLU	TRP	PHE	GLU	CYS	ASP	ASP	MET	LEU	THR	ARG	MET	ASN	ASN	ASP	HIS
3.79	3	20.21	-2.70	LYS	GLU	TRP	PHE	GLU	CYS	ASP	**ALA**	MET	**ASP**	**SER**	ARG	MET	ASN	ASN	ASP	HIS
3.79	3	19.29	11.30	**ASN**	GLU	TRP	PHE	GLU	CYS	ASP	**ALA**	MET	LEU	**SER**	ARG	MET	ASN	ASN	ASP	HIS
< 3.22	3	28.95	8.43	LYS	GLU	TRP	PHE	GLU	CYS	ASP	**ALA**	MET	**VAL**	THR	ARG	MET	ASN	ASN	ASP	HIS
< 3.22	3	24.88	1.13	LYS	GLU	TRP	PHE	GLU	CYS	ASP	**ALA**	MET	**VAL**	**ALA**	ARG	MET	ASN	ASN	ASP	HIS
< 3.22	3	23.83	7.24	**VAL**	GLU	TRP	PHE	GLU	CYS	ASP	**ALA**	MET	LEU	THR	ARG	MET	ASN	ASN	ASP	HIS
< 3.22	3	23.04	12.50	**GLN**	GLU	TRP	PHE	GLU	CYS	ASP	**ALA**	MET	LEU	THR	ARG	MET	ASN	ASN	ASP	HIS
< 3.22	3	22.29	7.09	**VAL**	GLU	TRP	PHE	GLU	CYS	ASP	**ALA**	MET	LEU	**SER**	ARG	MET	ASN	ASN	ASP	HIS
< 3.22	3	22.00	11.12	**ASN**	GLU	TRP	PHE	GLU	CYS	ASP	**ALA**	MET	LEU	THR	ARG	MET	ASN	ASN	ASP	HIS
< 3.22	3	19.86	6.98	LYS	GLU	TRP	PHE	GLU	CYS	ASP	**ALA**	MET	**ALA**	**ALA**	ARG	MET	ASN	ASN	ASP	HIS
< 3.22	3	19.58	10.52	LYS	GLU	TRP	PHE	GLU	CYS	ASP	**ALA**	MET	**ALA**	**SER**	ARG	MET	ASN	ASN	ASP	HIS
< 3.22	3	15.93	12.81	LYS	GLU	TRP	PHE	GLU	CYS	ASP	**ALA**	MET	**VAL**	**SER**	ARG	MET	ASN	ASN	ASP	HIS

Table 2. Predicted and calculated arabinose dissociation energy of ABP mutants.

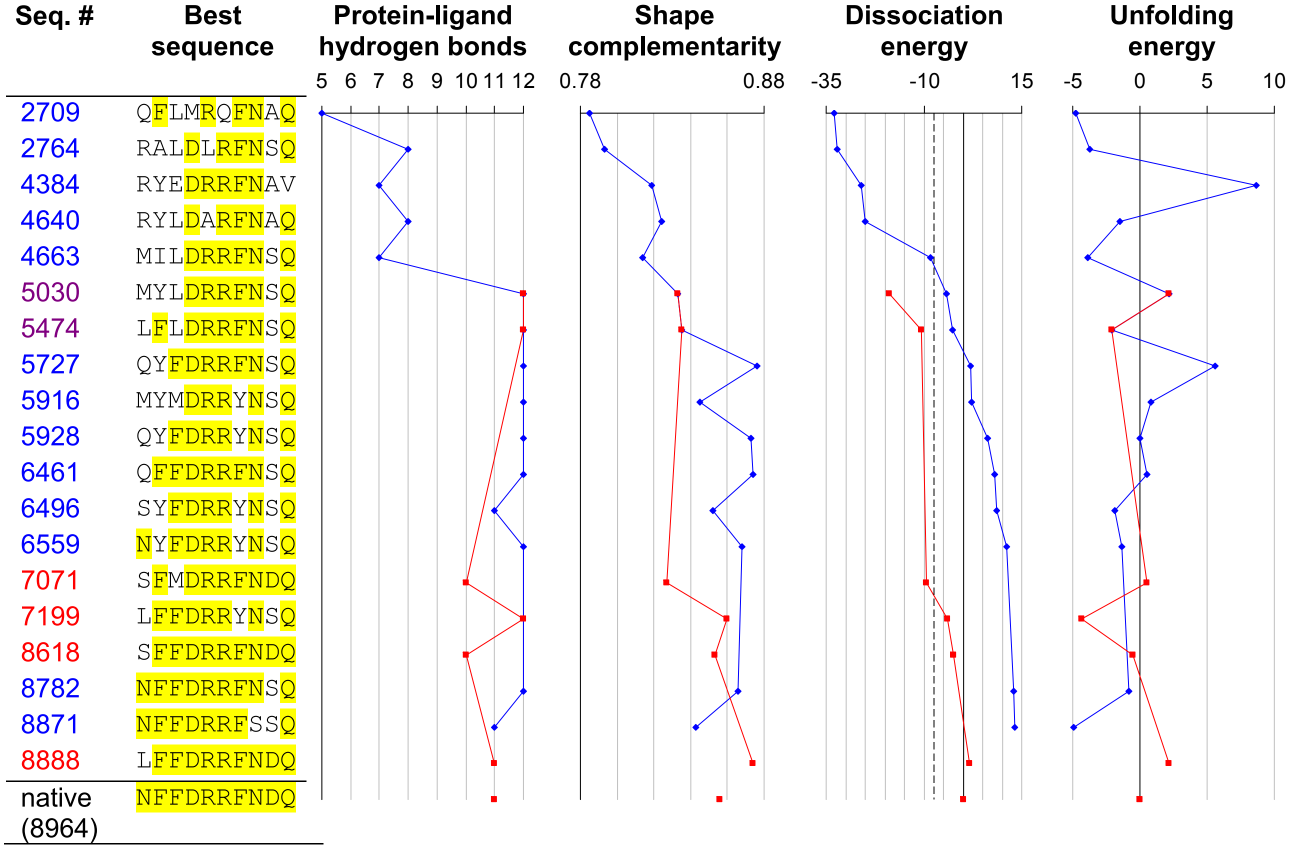
Top line shows the native sequence, and mutations are bolded. Data sources: 1. present work; 2. reference [87]; 3. reference [86].

Binding site design

The final and most stringent test of the molecular mechanics energy model was a redesign of the binding site in RBP (Figure 15). We discarded the ligand coordinates, and the sequence and coordinates of the 10 residues contacting the ligand. The total size of the sequence space searched was $17^{10} = 2.0 \times 10^{12}$ (Gly, Pro, and Cys were not allowed). The calculation was initiated from a population of random sequences. After evaluation of 8888 sequences, the energy function identified a point mutant (N13L) of native RBP as the tightest binding sequence. After 8964 sequences, it picked native RBP as the second tightest binding sequence. Evaluation of an additional 8879 sequences did not yield any further improvement. The entire process was repeated with a different random initial sequence population, and the same optimal sequences were selected. During the course of the design, first stability was achieved, then hydrogen bonding, and finally shape complementarity. The same pattern has been seen experimentally in the affinity maturation of antibodies against lysozyme.[88]

We experimentally tested the three top sequences from four different RBP-ribose redesign calculations to determine which aspects of the design algorithm were essential (Table 3). Decreasing the rotamer resolution (row a), omitting the final continuous minimization step (row b), or using a less accurate electrostatics model (row c) produces sequences that bind very weakly. Only when we use a high resolution rotamer library, a final continuous minimization step, and accurate electrostatics, does the design algorithm predict sequences that bind well (row d).

Prior to adding the stability requirement to the design calculation, all of our designed proteins expressed at very low concentrations in *E. coli*, probably because of proteolysis. After adding the stability requirement, the calculation predicts the top redesigned sequence (N13L point mutant) to be 1.5 kcal/mol more stable than the native RBP. Experimentally, this sequence is 1.2 kcal/mol more stable than the native (3.7 vs 2.5 kcal/mol, measured from urea denaturation curves[89]). We have not measured unfolding free energies for the remaining proteins.



Seq. #	Best sequence
2709	QFLMRQFNAQ
2764	RALDLRFNSQ
4384	RYEDRRFNAV
4640	RYLDARFNAQ
4663	MILDRRFNSQ
5030	MYLDRRFNSQ
5474	LFLDRRFNSQ
5727	QYFDRRFNSQ
5916	MYMDRRYNSQ
5928	QYFDRRYNSQ
6461	QFFDRRFNSQ
6496	SYFDRRYNSQ
6559	NYFDRRYNSQ
7071	SFMDRRFNDQ
7199	LFFDRRYNSQ
8618	SFFDRRFNDQ
8782	NFFDRRFNSQ
8871	NFFDRRFSSQ
8888	LFFDRRFNDQ
native (8964)	NFFDRRFNDQ

Figure 15. Redesigning the ribose binding site in RBP.

Positions identical to the native are highlighted in yellow. The figure shows the best sequence as a function of the number of sequences considered, using either the mean field dissociation energy as the criterion (blue trajectories) or alternatively the dissociation energy calculated using minimized structures (red trajectories). All sequences with a mean field dissociation energy greater than 30 kcal/mol (corresponding to -

7.5 kcal/mol relative to the native sequence, dashed line) were locally energy minimized to generate the red trajectory. Sequence 8871 is the top sequence when ranked by mean field dissociation energy (corresponding to Table 3b), and sequence 8888 is the top sequence when ranked by minimized dissociation energy (corresponding to Table 3d). The native sequence was found out of a possible 2×10^{12} sequences after 8964 sequence evaluations. Dissociation and unfolding energies are reported in kcal/mol, relative to the native sequence. The number of protein-ligand hydrogen bonds was determined using bndlst.[90] Shape complementarity (which ranges from 0 for perfectly non-complementary surfaces to 1 for perfectly complementary surfaces) was calculated using sc.[91] Backbone coordinates for the bound state are from 2DRI, and backbone coordinates for the unbound state are from 1URP.

Design calc.	Rotamers per position	Local minimization	Solvent treatment	Rank	# of residues identical to native	K_d (experimental)	Sequence (10 primary contacts)
(a)	2800	yes	Lee	1	3	210 ± 80 mM*	NIMLMMFNAN
				2	4	8.8 ± 0.4 mM*	NFMLNMFNAN
				3	4	83 ± 32 mM*	NFMLMMFNAN
(b)	5449	no	Lee	1	8	12 ± 0.9 mM*	NFFDRRFSSQ
				2	9	48 ± 13 mM*	NFFDRRFNSQ
				3	8	84 ± 10 mM*	NMFDRRFNSQ
(c)	5449	yes	Qiu	1	6	99 ± 2 mM*	NYYDRRYNAQ
				2	6	84 ± 2 mM*	NYMDRRYNSQ
				3	7	13 ± 1 mM*	NYFDRRYNAQ
(d)	5449	yes	Lee	1	9	19 ± 8 μM†	LFFDRRFNDQ
				2	10	0.30 ± 0.07 μM†	NFFDRRFNDQ
				3	9	80 ± 2 μM†	NTFDRRFNDQ
Native					10	0.30 ± 0.07 μM†	NFFDRRFNDQ

Table 3. High resolution rotamer library, gradient-based local minimization, and an accurate solvation model are required to successfully redesign the ribose binding site in RBP.

Multiple design calculations (a–d) were performed using different sampling resolutions and solvent models. The top three sequences from each calculation and their experimentally measured binding constants are shown. Parts of the sequence identical to the native sequence are highlighted in yellow. (a) Design calculation using a lower resolution rotamer library. (b) Design calculation without a gradient-based local minimization step. (c) Design calculation using a less accurate generalized Born solvent treatment.[92] (d) Design calculation using a high resolution rotamer library, gradient-based local minimization, and an accurate generalized Born solvation model.[16] Sequences are ranked by calculated dissociation energy, allowing 5 kcal/mol destabilization relative to the native sequence for 5449 rotamers / position, and 20 kcal/mol destabilization for 2800 rotamers / position. The native sequence was not within the top 100 sequences for design calculations *A*, *B*, or *C*. * K_d measured using the solid phase radioligand binding assay. † K_d measured using the centrifugal concentrator assay. The reported error is the standard deviation of 3 measurements.

Discussion

This paper reports the first successful redesign of an entire binding site based on an unmodified molecular-mechanics potential energy function. This is a stringent test of the energy function, because the native sequence and a point mutant are distinguished from 2.0×10^{12} alternative sequences. Good hydrogen bonds and steric complementarity were picked out directly by the energy function, without energy terms or selection criteria that specifically required these features. Given that the underlying physics is the same for the design of new proteins and for the simulation of known proteins, it is satisfying to see that the same energy models can be used as well.

We tested a number of simplifications commonly used in protein design calculations, and found that they all resulted in less successful predictions. For example, low sampling resolution or an inaccurate solvation model led to sequences that lacked critical hydrogen bonds. Scaling down the electrostatic energy (which is frequently done to compensate for a crude electrostatics model) reduced the accuracy of the energetic predictions. Eliminating the unfolded state resulted in unstable designed proteins. Softening the van der Waals interaction allowed atoms to pack together more closely, making hydrogen bonds and salt bridges appear artificially strong (Figure 28), and resulting in the burial of charged and polar functional groups (Table 10).

An important conclusion from this work is that MM-PEF's must be paired with an accurate continuum solvent model and with protonation equilibria in order to correctly redesign a polar binding site. Individual polar protein-ligand interactions can exhibit energies up to 100 kcal/mol (the Coulomb energy between unit charges separated by 3.3 Å). These energies are almost exactly counterbalanced by interactions with water in the unbound protein. Thus, small errors in the solvation energy grossly alter the design predictions. Finite difference algorithms are generally considered the most accurate methods to solve the Poisson-Boltzmann differential equation that defines the continuum solvent model, but they are currently too slow for protein design. Very accurate generalized Born approaches have been developed over the last few years,[16] and produce solvation energies that differ from the finite difference result by only 2% (Figure 20). We have shown that this level of accuracy is both necessary and sufficient for protein design calculations.

The results in this paper suggest that the protein design and molecular-mechanics fields can work together on the same potential energy functions, and that future developments in MM-PEFs will be immediately applicable to protein design (although ad hoc terms may still be necessary for modeling aggregated and misfolded states). Currently, there are active efforts to develop polarizable potential energy functions that more accurately reproduce the physical characteristics of small molecules,[54,55,68] and hybrid quantum mechanical / molecular mechanical potential energy functions that model charge transfer and changes in covalent bonding.[93,94] It will be exciting to see how these improved energy models will impact the protein design problem.

Materials and methods

Calculations

Protein structures were predicted using a rotamer-based mean field algorithm.[19] The energy was calculated as the sum of the CHARMM22 molecular-mechanics energy,[14] a generalized Born surface-area solvation energy[15,16] using a microscopic surface tension[95] of 0.0072 kcal/mol/Å^2, and a deprotonation energy.[96] The most probable mean field structure was then locally minimized using the L-BFGS optimization algorithm[97] in TINKER[98] to obtain a final structure and energy. The unfolded protein energy was calculated by assuming that the protein backbone adopts an ensemble of random walk conformations in water (see ref. [58,99] and p. 81). The stability of the protein was calculated as the energy difference between the unfolded protein and the folded unbound protein, and the dissociation energy was calculated as the energy difference between the uncomplexed and the complexed protein-ligand system. All calculations were performed at 25°C, pH 7.0, 100 mM monovalent salt. Ribose binding proteins were designed using a genetic algorithm[20] that optimized the calculated ribose dissociation energy, given a 5 kcal/mol limit on protein destabilization. The genetic algorithm was initialized with a population of random sequences. Calculations were performed using CNSsolve[100], TINKER[98], and custom code written in C++, and run on a Pentium-based Linux cluster.

Protein purification and constructs

RBP without a periplasmic signal peptide was cloned into the NcoI/XhoI sites of pET28a (EMD Biosciences), generating a derivative with a C-terminal His_6 tag. Mutants were made by Kunkel mutagenesis[101] or by QuikChange (Stratagene). Protein was expressed in BL21 DE3 *E. coli* cells (Novagen) with 1 mM IPTG for 5 hr at 37°C. Cells were lysed with lysozyme and sonication in the presence of 1 mM phenylmethylsulfonyl fluoride. Protein was purified by immobilized metal affinity chromatography, followed by gel filtration chromatography in 20 mM potassium phosphate pH 7.0, 100 mM NaCl. The purified protein was then concentrated, and its final concentration determined by absorbance.[102]

Centrifugal concentrator radioligand binding assay

Proteins were diluted into 1 ml of 20 mM potassium phosphate pH 7.0, 100 mM NaCl, and 0.5 µCi 1-^{3}H(N)-D-ribose (Moravek). After equilibration for 30 minutes, the samples were placed in centrifugal concentrators (Amicon Ultra, 5 kDa MWCO), and centrifuged until at least 500 µl of filtrate had crossed the membrane. Any filtrate in excess of 500 µl was returned to the retentate, and the quantity of radioligand in the filtrate and retentate were measured by scintillation counting. Dissociation constants were calculated as $K_d = \frac{2P}{r-1} - \frac{2L}{r+1}$, where r is the ratio of retentate to filtrate radioligand, P is the initial protein concentration, and L is the initial radioligand concentration. We chose conditions where $P > K_d$ and r fell between 1.2 and 20. The analysis depends on the assumption that water and the ligand cross the membrane at equal rates. This assumption was tested by centrifuging a ribose solution across the membrane in the absence of protein; the specific activities of the retentate and filtrate were identical to within 4%.

Solid phase radioligand binding assay

A solid phase radioligand binding assay was used to detect binding with K_d's in the high millimolar range. Nickel-NTA agarose slurry (Qiagen) was washed and resuspended in buffer (20 mM potassium phosphate pH 7.0 and 100 mM NaCl) to form a 50% (v/v) slurry. Twenty microliters of the slurry were mixed with 5 nmol of His_6-tagged protein and 1.0 µCi of radioligand in a final buffer volume of 50 µl. Following a 30 minute equilibration, the mixture was transferred to 0.45 µm centrifugal filter units (Millipore #UFC30HV0S) and centrifuged at 12000×g for 2 minutes to remove unbound ligand. The resin was washed three times by addition of 500 µl of 50 % ethanol and centrifugation at 12000×*g* for 2 minutes. The bound ligand was eluted with 250 µl guanidinium HCl, and quantified by scintillation counting. Radioligand eluted from a no-protein control was included to account for non-specific binding to the resin, and a control of 0.5 µCi radioligand was used to determine counting efficiency. Dissociation constants were calculated as

$K_d = \frac{L - Lr - Pr + Pr^2}{r}$, where r is the fraction of protein bound to radioligand, P is the initial protein concentration, and L is the initial ligand concentration.

Acknowledgments

We gratefully thank J.D. King, R.S. Fenn and S.R. Pfeffer for help with protein purification. We thank R.B. Altman, R.L. Baldwin, A.T. Brunger, R.S. Fenn, J.J. Havranek, D. Herschlag, M. Levitt, L.L. Looger, W.L. Martin, V.S. Pande, E.L. Raffauf, T.J. Wandless, R.M. Weisinger, and S.J. Wrenn for many helpful suggestions. This work was supported by NIH grants GM068126-01 and 1-DP1-0D000429 to P.B.H. F.E.B. was partially supported by a training grant from the National Institute of General Medical Sciences (5T32 GM07365-28).

Chapter 4: The protein design algorithm

This chapter has been adapted from supporting information for:

Boas FE and Harbury PB. (2008) "Physics-based design of protein-ligand binding." *Journal of Molecular Biology*. In press.

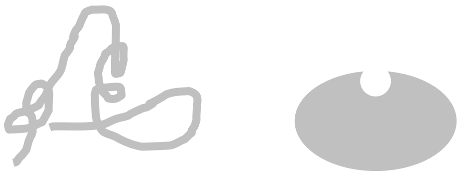
Potential energy function

Overview

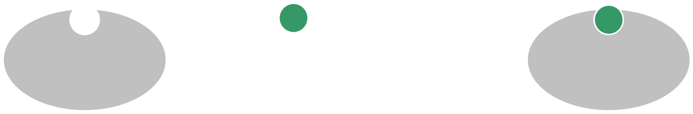
The potential energy of a specific protein conformation can be partitioned into different categories. On one extreme, in a quantum calculation, all of the energy is electrostatic. On the other extreme, in an intuitive sense, the energy can be thought of in terms of hydrogen bonds, salt bridges, and steric complementarity. In between, there are molecular mechanics models that treat the protein as a collection of atoms with partial charges and van der Waals parameters, connected by springs to maintain bond lengths and angles.[103]

What is the right type of model to use for protein design? Currently, most protein design algorithms use statistical terms, derived by, for example, counting how frequently different types of hydrogen bonds and salt bridges are seen in crystal structures. The advantage of this approach is that the geometry of the interaction does not have to be exactly correct to get a reasonable energy, and it can include empirically observed phenomena that otherwise might not be modeled correctly. The disadvantage is that you can’t model cases that are missing in your training set. With more detailed sampling in conformational space, we believe it will be more accurate to directly calculate the strengths of salt bridges and hydrogen bonds from Coulomb’s law and continuum electrostatics. Thus, in this paper, we have avoided statistical terms, and base all of our calculations on molecular mechanics with continuum solvent.

We calculate protein stability and protein-ligand dissociation energy as a difference between states:



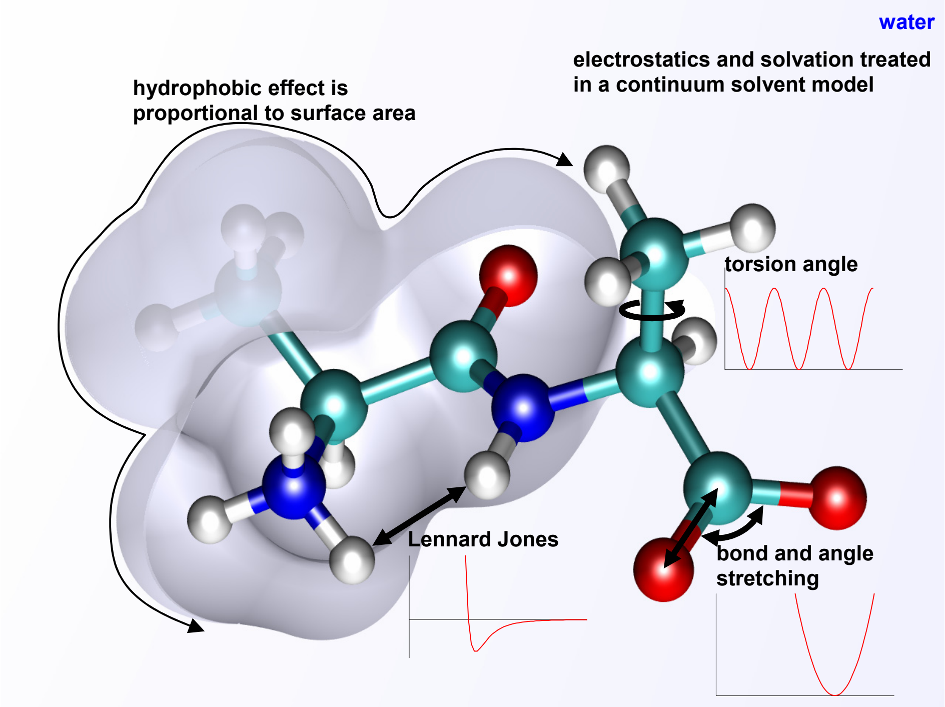
Stability = unfolded protein energy – protein energy



Dissociation energy = protein energy + ligand energy – protein·ligand complex energy

Thus, for example, a buried salt bridge might have a Coulomb interaction energy of 100 kcal/mol, but the dissociation energy will be much less than this, because in the undocked state, those charges will have similarly favorable interactions with water. These calculations generally have a lot of large terms that almost cancel each other out, so it is important to do the calculations very carefully.

Our potential energy function (Figure 16) allows us to model effects that are typically ignored in protein design (Figure 17).


water
electrostatics and solvation treated in a continuum solvent model
hydrophobic effect is proportional to surface area
torsion angle
Lennard Jones
bond and angle stretching

potential energy =	**molecular mechanics +**	**generalized Born +**	**surface area +**	**protonation energy**
	(bond length + angle + torsion + LJ + Coulomb)	(solvent polarization)	("hydrophobic" effect)	(pH effect)

$$= \sum_{\text{bonds}} k_b(b-b_0)^2 + \sum_{\text{angles}} (k_{UB}(S-S_0)^2 + k_\theta(\theta-\theta_0)^2) + \sum_{\text{dihedrals}} k_\chi(1+\cos(n\chi-\delta))$$

$$+ \sum_{\text{impropers}} k_\phi(\phi-\phi_0)^2 + \sum_{\substack{\text{nonbonded}\\ i<j}} E_{VDW}\left(\left(\frac{r_{min}}{r}\right)^{12} - 2\left(\frac{r_{min}}{r}\right)^{6}\right)$$

$$+332*\sum_{\substack{\text{nonbonded}\\ i<j}} \frac{q_i q_j}{\varepsilon_{in} r} + 332*\sum_{i,j}^{N} \mathrm{GB}(q_i,q_j,a_i,a_j,r,\varepsilon_{in},\varepsilon_{out},\kappa)$$

$$+k_{SASA}\mathrm{SASA} + \sum_{\substack{\text{deprotonated}\\ \text{amino acids}}} U_{deprot.}$$

Figure 16. Potential energy function.

All parameters[14] were from CHARMM22 except for k_{SASA} and U_{deprot}. For the generalized-Born solvation energy, a water radius of 1.4 Å was used to define the molecular surface. Distance is in angstroms, charge is in elementary charge units, and energy is in kcal/mol. “332” is the Coulomb electrostatic constant for these units.

Variables

k_b	spring constant for bond length
b	bond length
b_0	equilibrium bond length
k_{UB}	Urey-Bradley constant for atoms separated by two bonds
S	distance between atoms separated by two bonds
S_0	equilibrium distance
k_θ	spring constant for bond angle
θ	bond angle
θ_0	equilibrium bond angle
k_χ, n, δ	Fourier series terms for periodic barrier to rotation around bonds
χ	torsion angle
k_φ	spring constant for torsion angle to restrain planar groups
φ	torsion angle
φ_0	equilibrium torsion angle
E_{VDW}	van der Waals energy
r	inter-atom distance
r_{min}	minimum-energy inter-atom distance
q_i, q_j	charge on atoms i and j
ε_{in}	protein and ligand dielectric constant = 1.0
ε_{out}	water dielectric constant = 78.4
GB()	generalized-Born solvation energy
a_i, a_j	generalized-Born radii of atoms i and j
κ	inverse Debye-Hückel length (salt screening length)
k_{SASA}	microscopic surface tension of water[95] = 0.0072 kcal/mol/Å^2
$SASA$	solvent-accessible surface area (the area traced out by the center of a spherical probe touching the protein’s VDW surface); calculated using a water probe radius of 1.4 Å
$U_{deprot.}$	deprotonation energy (from a thermodynamic cycle based on the pK$_A$’s of free amino acids)

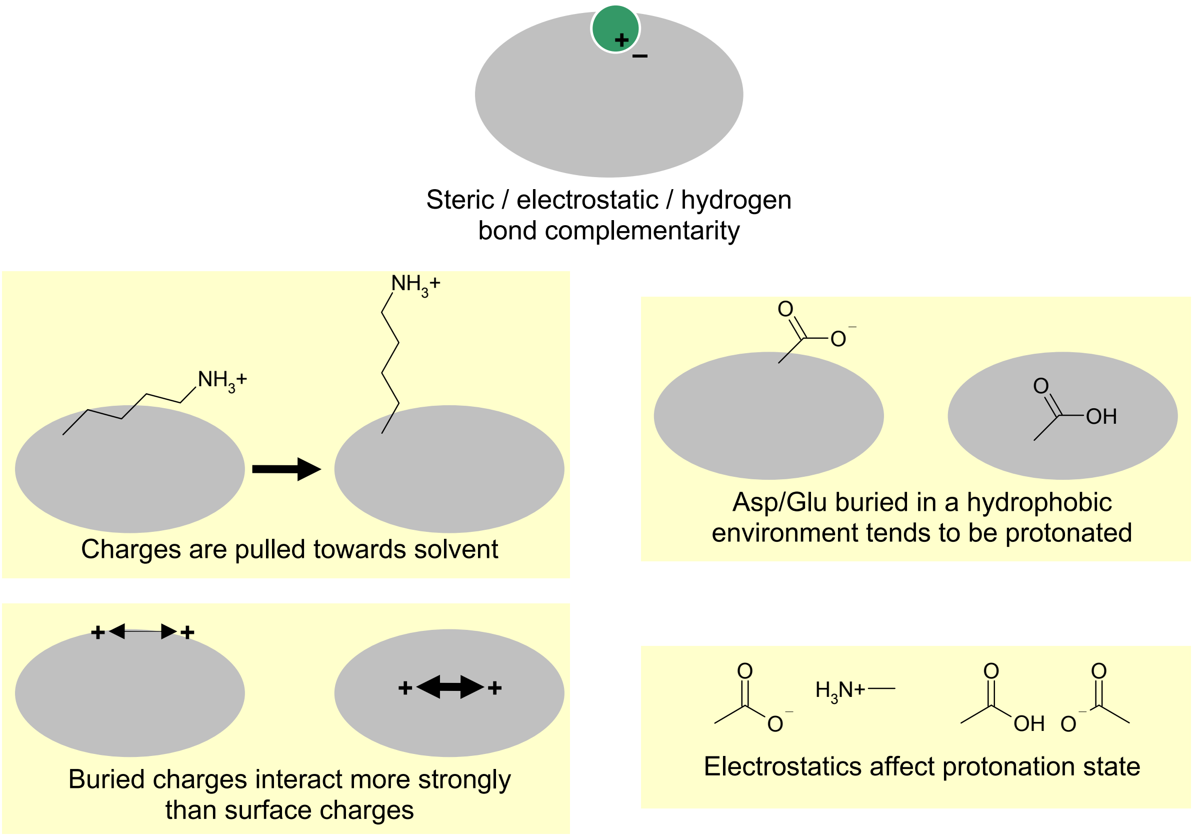

+
–
Steric / electrostatic / hydrogen bond complementarity
NH_3+
NH_3+
Charges are pulled towards solvent
O
O⁻
O
OH
Asp/Glu buried in a hydrophobic environment tends to be protonated
+ +
+ +
Buried charges interact more strongly than surface charges
O
O⁻
H_3N+
O
OH
O
O⁻
Electrostatics affect protonation state

Figure 17. Examples of protein behaviors treated by our model. Factors typically ignored in design calculations are highlighted.

Notes

van der Waals energy: r_{min} for AB interaction is the arithmetic mean of r_{min} for AA and BB interactions. E_{VDW} for AB interaction is the geometric mean of E_{VDW} for AA and BB interactions. Bonded and 1,3 atoms (atoms separated by two bonds) are excluded from this sum.

Coulomb electrostatics: Bonded and 1,3 atoms are excluded from this sum.

Generalized Born solvation energy: All pairs of atoms are included in this sum (including self). Each non-self pair occurs twice in the sum.

Capping: The VDW energy was capped at 2000 kcal/mol/atom pair, and the total electrostatic energy (Coulomb plus generalized Born) was capped at ±1000 kcal/mol/atom pair to prevent floating point overflow of Boltzman weights. In well-packed structures, no interaction energies exceeded the caps.

Hydrogen bonds: These are treated as a combination of electrostatics and van der Waals interactions.

Distance cutoff: None.

Critical parameters

van der Waals: These should not be modified from the values in CHARMM22. In the design literature, van der Waals parameters are frequently stretched or scaled so as not to penalize small steric clashes resulting from limited sampling resolution. However, we've found that this has the side effect of making hydrogen bonds and salt bridges appear stronger than they actually are (see Figure 28 and Table 10).

Internal dielectric constant: We found that an internal dielectric constant of 1.0, which is the default for CHARMM22, produced the most accurate energies. In the design literature, the internal dielectric constant is frequently set to values between 4 and 20, to account for rearrangements in the rest of the protein, or to scale down the Coulomb energy in the absence of a good solvent model. We've found that this is unnecessary (and actually harmful) when the relevant residues are modeled explicitly, and a good solvent model is used.

Water

Our model treats water as a continuum dielectric with salt and surface tension. In a vacuum, the partial charges on the protein atoms interact through Coulomb's law. When we put the protein in water, there is an additional solvation energy term. Part of the solvation energy is roughly proportional to the surface area: VDW interactions with solvent, and the entropy and enthalpy of rearrangment of water molecules at a surface (the hydrophobic effect). The rest of the solvation energy is due to partial charges in the protein interacting with induced surface charges and ion clouds in the solvent ($\Delta G_{polarization}$). Charged atoms closer to the protein's surface have more favorable solvation energy and smaller charge-charge interactions. These energies are calculated using the generalized Born equation.

Generalized-Born energy

Atomic partial charges in a protein reorient water dipoles, inducing surface charges that interact favorably with the partial charges in the protein, and that screen

Coulombic interactions within the protein. Salt forms a counter-ion atmosphere around the protein that neutralizes charge over the Debye-Hückel length. We calculated the interaction energy of the protein with these induced solvent charges using the generalized-Born equation,[15] which provides an approximate solution to the Poisson-Boltzmann differential equation.[44]

The generalized-Born approach requires the calculation of generalized-Born radii for each atom (Figure 18). The manuscript compares two numerical approaches for obtaining the radii. In the first approach, generalized-Born radii are computed on the basis of an r^{-4}-weighted spatial integral (Figure 19):

$$a_i = 4\pi\left(\int_{solvent} \frac{1}{r^4}\, dV\right)^{-1}.$$

Here r is the distance from the atom center to each volume element in the integrand. The $1/r^4$ in this equation comes from the fact that the energy of a charge–induced dipole interaction (partial charge in the protein interacting with water) is $1/r^4$. Alternatively, more accurate radii are obtained from an empirical sum of r^{-4}- and r^{-5}-weighted spatial integrals:[16]

$$a_i = 4\pi\left(-\int_{solvent} \frac{1}{r^4}\, dV + P\sqrt{4\pi \int_{solvent} \frac{1}{r^5}\, dV}\right)^{-1},$$

where *P*=3.0. The integrals were performed on a rectangular grid (0.5 Å resolution) with the dielectric boundary defined as the molecular surface. Grid points were assigned to solvent if they were contained within a solvent sphere (1.4 Å) centered on a grid point outside the solvent-accessible volume of the protein. For design calculations, the molecular surface was initialized using the crystal structure of the scaffold protein, and was iteratively updated using an average of the currently optimal structures. Final energy evaluations on minimized structures used the exact molecular surface. Formulas in the Appendix give values for the spatial integrals from the grid boundary to infinity. A simpler alternative for integrating the solvent on a grid might be to analytically integrate outside the spherical atom, then subtract the protein regions on the grid outside the atom.

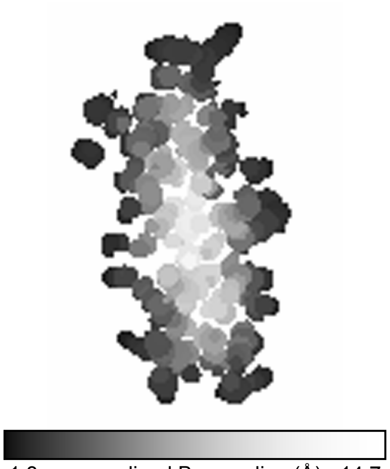


Figure 18. Slice through ribose binding protein, showing generalized Born radii. The radii correlate with atom burial.

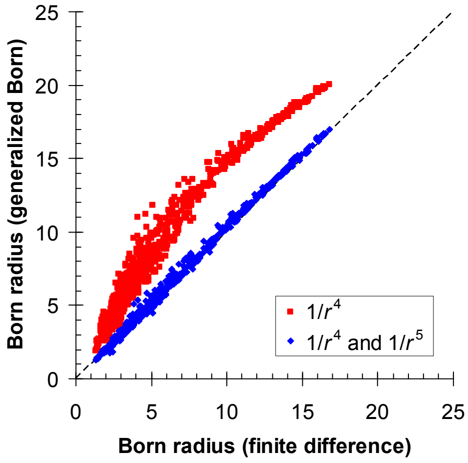


Figure 19. Comparison of generalized Born radii for protein tyrosine phosphatase 1B calculated using an integral formula (y-axis) with radii calculated using a finite-difference approach (x-axis).

Similar results were reported in [16].

After calculating generalized Born radii for each atom, we can calculate the solvation energy using the generalized Born equation:

$$\Delta G_{polarization} = \sum_{i,j} \mathrm{GB}(q_i, q_j, a_i, a_j, r, \varepsilon_{in}, \varepsilon_{out}, \kappa) = -\frac{1}{2}(\frac{1}{\varepsilon_{in}} - \frac{1}{\varepsilon_{out}}) \sum_{i,j} \frac{q_i q_j}{\sqrt{r_{ij}^2 + a_i a_j e^{-r_{ij}^2/(4a_i a_j)}}}$$

This equation gives exact answers for the limiting cases of very close and very distant charges, and interpolates between these two extremes. In the limit of $a_1 = a_2 >> r$, the generalized Born equation calculates a solvation energy of:

$$\Delta G_{polarization} = -\frac{q^2}{2a}(\frac{1}{\varepsilon_{in}} - \frac{1}{\varepsilon_{out}}),$$

which is called the Born equation. And in the limit of $r >> a$, the generalized Born equation plus the Coulomb term gives an interaction energy of $\frac{q_1 q_2}{\varepsilon_{out} r}$, which is also correct.

The generalized Born equation can be modified to handle salt as well:[104]

$$\sum_{i,j} \mathrm{GB}(q_i, q_j, a_i, a_j, r, \varepsilon_{in}, \varepsilon_{out}, \kappa) = -\frac{1}{2} \sum_{i,j} (\frac{1}{\varepsilon_{in}} - \frac{e^{-\kappa \sqrt{r_{ij}^2 + a_i a_j e^{-r_{ij}^2/(4a_i a_j)}}}}{\varepsilon_{out}}) \frac{q_i q_j}{\sqrt{r_{ij}^2 + a_i a_j e^{-r_{ij}^2/(4a_i a_j)}}},$$

with $\kappa = \dfrac{\sqrt{I/\varepsilon_{out}}}{0.343}$ at 25°C.

Variables:

κ	inverse Debye-Hückel length in Å^{-1}
I	ionic strength in mol/l

A salt concentration of 100 mM was used for the calculations reported here (Figure 20).

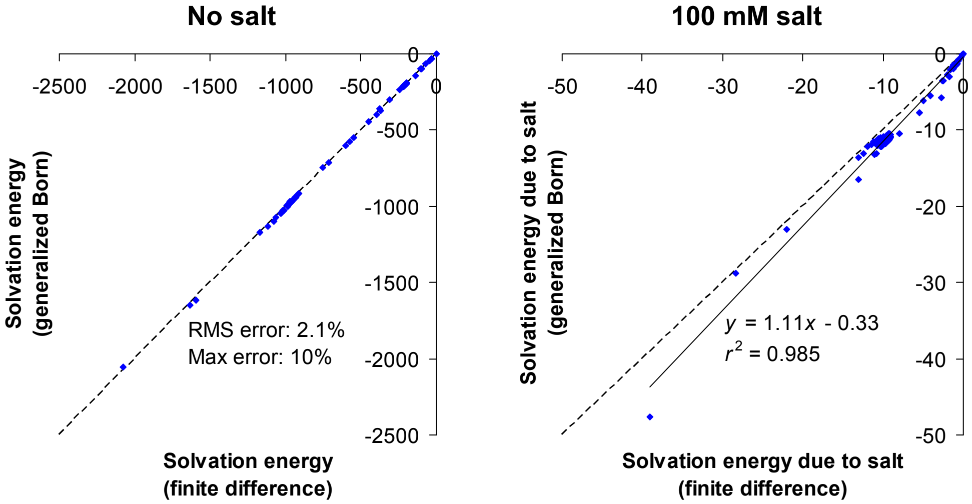


Figure 20. Comparison of solvent polarization energies for a set of small molecules, peptides, and proteins calculated using the generalized-Born approach (y-axis) with values calculated using a finite-difference approach (x-axis).

Pairwise approximation of solvent accessible surface area (SASA)

Following Street and Mayo,[40] we approximated the total SASA as the sum of accessible surface areas for each amino acid within the context of the fixed structural elements of the design, less the probability weighted sum of the pairwise surface areas buried by each variable structural element of the design (for example a rotamer or a ligand pose). The pairwise surface areas are scaled to correct for over-counting, which occurs when multiple variable structural elements simultaneously bury one surface patch. The scaling factors were determined by a linear regression that optimized agreement between the pairwise approximation and the exact solvent accessible surface areas of 100,000 random conformations of the protein with random sequences present at the design positions. Optimal values of the scaling factors are highly under-constrained, due to correlations between the various area terms. To address this issue, we used a singular value decomposition[105] to perform the linear regression. Any scaling factors greater than 100 or less than –100 were set to 0, and the regression was repeated without them.

$$\text{SASA (linear regression form)} = \sum_{\substack{i \in \text{variable} \\ \text{position}}} t_i A_i - \sum_{\substack{i \in \text{variable} \\ \text{position}}} s_i \sum_{j \neq i} A_{i,j} - \sum_{\substack{i \in \text{fixed} \\ \text{position}}} s_i \sum_{\substack{j \in \text{variable} \\ \text{position}}} A_{i,j} + C$$

Here, variable positions included the repacked residues and the ligand. The fixed positions were the residues in the protein whose identity and conformation were held fixed during the design. This linear regression form can be rearranged into a pairwise factorable form.

SASA (pairwise form) =

C	additive constant
$+\sum_{\substack{i\in\text{variable}\\ \text{position}}}(t_iA_i - s_i\sum_{\substack{j\in\text{fixed}\\ \text{position}}}A_{i,j} - \sum_{\substack{j\in\text{fixed}\\ \text{position}}}s_jA_{j,i})$	SASA of rotamers and ligand poses less the pairwise area buried at interfaces with fixed structural elements
$-\sum_{\substack{i\in\text{variable}\\ \text{position}}}\sum_{\substack{j\in\text{variable}\\ \text{position}, j<i}}(s_iA_{i,j} + s_jA_{j,i})$	pairwise area buried at interfaces between variable structural elements

Variables:

A_i the accessible surface area of a rotamer, pose or fixed conformation at position i within the context of the fixed structural elements of the design.

t_i scaling factors for accessible surface areas of rotamers or poses

$A_{i,j}$ The portion of A_i buried by the variable rotamer or pose at position j within the context of the fixed structural elements of the design.

s_i scaling factors for pairwise buried areas

The interfacial solvation energy is the product of the SASA and a microscopic surface tension of 7.2 cal/mol/Å^2 [95]. The "hydrophobic effect" driving aggregation of hydrophobic solutes in water increases in proportion to solute surface area with a slope[39] of 24 cal/mol/Å^2. This slope is reconciled with the 7.2 cal/mol/Å^2 microscopic surface tension by adding the van der Waals interaction energy between explicitly modeled hydrophobic solutes, which evaluates to roughly 17 cal/mol/Å^2 for CHARMM22.

Deprotonation energy

The structural calculations reported here modeled the pH- and environment-dependent titration of histidine and the acidic amino acids. The doubly protonated and two singly protonated states of histidine, and the protonated and deprotonated states of aspartate and glutamate were modeled as independent rotamers. Because molecular-mechanics potentials do not treat changes in covalent bonding, the energy difference between protonated and deprotonated rotamers was computed using a thermodynamic cycle (Figure 21). For example, the deprotonation energy for an aspartate residue within a protein (labeled A in Figure 21) was determined indirectly by summing two transfer free energies (B and D) and the experimentally measured free energy for deprotonation of acetylated asparate amide in free solution (C). Free energies for the small-molecule aspartate derivatives were obtained by building a complete set of aspartate side-chain rotamers onto each member of an amino-acid backbone ensemble, evaluating the energy of each configuration, and computing the free energy as:

$U(\text{solution}) = -RT\ \ln(\text{partition sum}).$

Then:

$B = U(\text{AspH, solution}) - E(\text{AspH, protein})$

$C = -2.3RT{*}(\text{pH} - \text{pK}_a)$

$D = E(\text{Asp}^{-}\text{, protein}) - U(\text{Asp}^{-}\text{, solution})$

where U is free energy and E is potential energy. Adding these together:

$A = B + C + D = [E(\text{Asp}^{-}\text{, protein}) - E(\text{AspH, protein})] + [-U(\text{Asp}^{-}\text{, solution}) + U(\text{AspH, solution}) - 2.3RT{*}(\text{pH} - \text{pK}_a)]$

We denote the terms within the right bracket above, –*U*(Asp$^-$, solution) + *U*(AspH, solution) – 2.3*RT**(pH – pK$_a$), as the deprotonation energy. It is added to the self-energy of each deprotonated rotamer to establish the appropriate energy relationship between the deprotonated and protonated forms of the amino acid (Table 4). The deprotonation energy is pH dependent, and all of the calculations reported here were performed at pH 7.0.

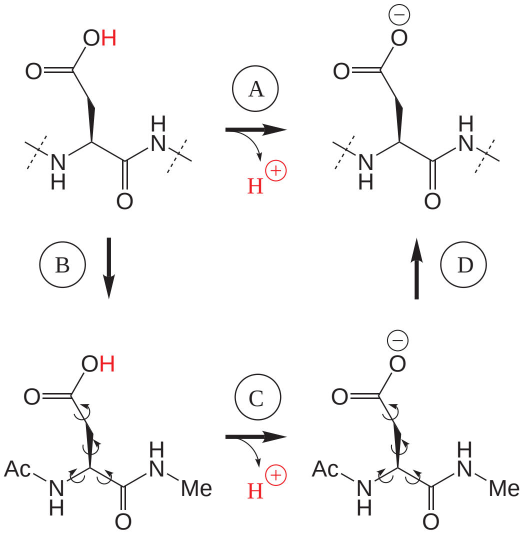


Figure 21. Thermodynamic cycle used to evaluate the deprotonation energy for aspartate (A).

The dashed lines in the top structures represent bonds to the complete polypeptide chain of the protein, which is not shown. The bottom structures depict N-acetyl, N'-methyl aspartate α-amide in its protonated and deprotonated forms. The rotational arrows on the structures at the bottom indicate that they are modeled as a structural ensemble, whereas the structures at the top are single rotamers. The deprotonation energy is calculated as the sum of two transfer energies (B and D) and the experimentally-measured free energy for protonation of the acetyl-aspartate amide (C).

Amino acid	Deprotonation energy
HSP	0
HSD	–23.19 – 1.36 (pH – 6.74)
HSE	–2.53 – 1.36 (pH – 6.14)
ASP	37.21 – 1.36 (pH – 3.71)
APP	0
GLU	41.64 – 1.36 (pH – 4.15)
GUP	0

Table 4. Deprotonation energies for the titratable amino acids in the 6028-member rotamer library.
Experimental pKa values for free amino acids are from ref. [106,107]. We did not include protonation states for CYS, TYR, LYS, or ARG because of a lack of published CHARMM22 parameters for those amino acids. 1.36 = *RT* ln 10 at *T* = 25°C.

Discrete sampling

Protein scaffold coordinates

Hydrogen coordinates were added to scaffold crystal structures using Reduce.[108]

Selection of design positions

For ABP, all side chains where the van der Waals spheres were within 1 Å of the ligand van der Waals spheres in any of four crystal structures (8ABP, 6ABP, 1ABE, 5ABP) were selected as design positions. For RBP, hydrogen bonding and hydrophobic contacts determined by the program HBPLUS[109] were selected as design positions. The resulting positions are listed in the caption to Figure 10.

For Avastin-VEGF, the repacked residues were hand picked, because with our current level of computer power, we were unable to model all interface residues at high resolution. Starting with the 6 Fab positions where mutations have been reported to improve the affinity, we then added side chains (except ALA, GLY, PRO) in Fab and VEGF contacting side chains at these 6 positions, and also included positions that showed a high conformational variability among different crystal structures (1BJ1, 1CZ8, 1FLT, 1KAT, 1QTY, 1TZH, 1TZI, 1VPP, 2VPF). The resulting positions are listed in the caption to Figure 11.

Rotamer library

A detailed rotamer library (including polar and non-polar hydrogens) was created by clustering the side chain conformations seen in high-resolution crystal structures (Table 5). Starting with the 18528 structures in Protein Data Bank Release #101 (July 2002), we removed theoretical models, structures with resolution > 1.9 Å, structures with a CAVEAT record, and structures with ≤ 10% of atoms in one of the 20 natural amino acids. This resulted in a list of 7312 structures. Hydrogens were added to each structure using Reduce[108] from the Richardson lab. The side chain conformations for each amino acid were then clustered at the resolution listed in Table 5. The clustering process involved selecting the conformation with the most close neighbors, discarding all neighbors (defined by an RMS cutoff), and repeating until a predetermined fraction of the conformations had been covered. Finally, each rotamer was locally minimized with a constraint of ± 1° on each dihedral angle. No rotamer in the library corresponds to any of the crystallographic cooordinates of ABP, RBP, or Avastin-VEGF.

For repacking calculations, rotamers were placed at each variable position of the protein scaffold, and energy minimized using dihedral restraints and no electrostatics. The energy minimization slightly adjusted bond lengths and angles to match the equilibrium values in CHARMM22. Rotamers with energies more than 15 kcal/mol over the lowest energy rotamer of the same amino acid at the same position were eliminated.

Amino acid	**Number of rotamers**	**Neighbor RMS cutoff (Å)**	**Close neighbor RMS cutoff (Å)**	**Coverage**
ALA	3	0.5	0.3	0.999
APP	141	0.5	0.3	0.999
ARG	974	1.0	0.4	0.98
ASN	132	0.5	0.3	0.999
ASP	62	0.5	0.3	0.999
CYS	29	0.5	0.3	0.999
CYX	8	0.5	0.3	0.999
GLN	758	0.5	0.3	0.999
GLU	412	0.5	0.3	0.999
GLY	1	0.5	0.3	0.999
GUP	649	0.5	0.3	0.999
HSD	233	0.5	0.3	0.999
HSE	255	0.5	0.3	0.999
HSP	245	0.5	0.3	0.999
ILE	215	0.5	0.3	0.999
LEU	325	0.5	0.3	0.999
LYS	400	1.0	0.4	0.98
MET	181	0.8	0.4	0.99
PHE	193	0.5	0.3	0.999
PRO	8	0.5	0.3	0.999
SER	32	0.5	0.3	0.999
THR	64	0.5	0.3	0.999
TRP	238	0.6	0.3	0.99
TYR	414	0.5	0.3	0.999
VAL	56	0.5	0.3	0.999
Total	6028			

Table 5. The highest resolution rotamer library with 6028 rotamers.

APP = protonated Asp, GUP = protonated Glu, HSP = doubly protonated His, HSD = His protonated on the delta nitrogen, HSE = His protonated on the epsilon nitrogen, CYX = disulfide-bonded cysteine.

Ligand poses

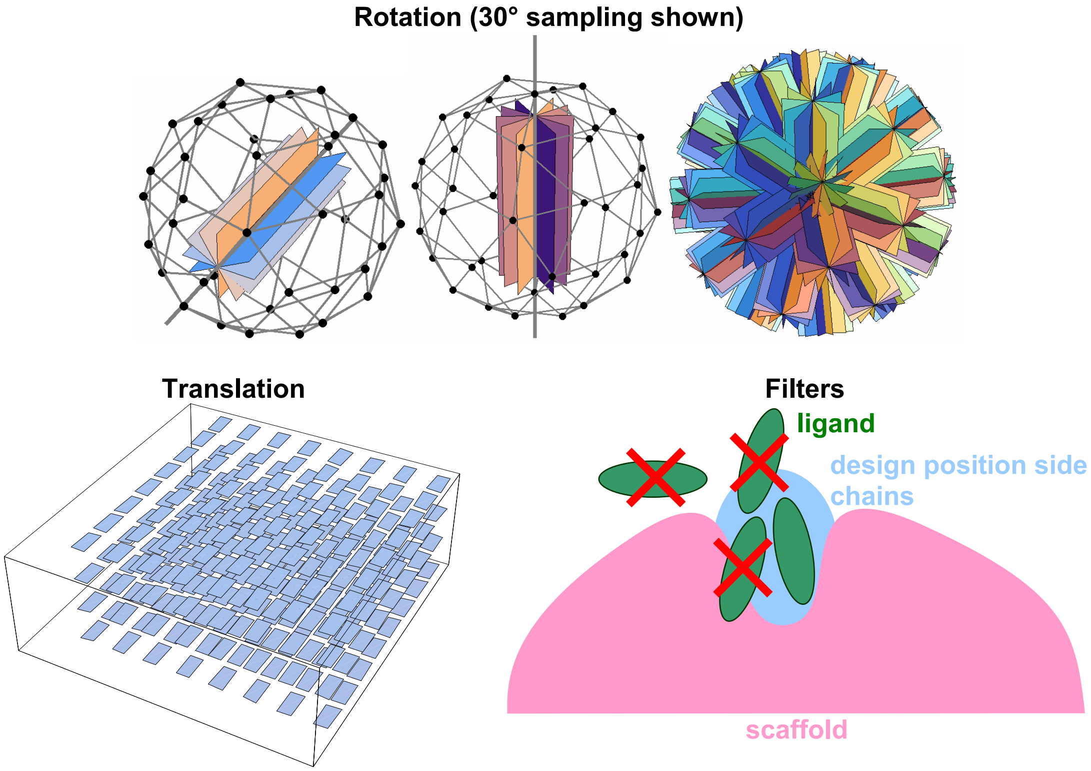


Figure 22. Ligand sampling and filters.
Ligand poses were identified by generating conformers of the ligand, and then exploring rotational and translational degrees of freedom. A series of filters was applied to identify poses that overlapped well with the-side chain regions of the design positions but not with the fixed portions of the scaffold, and that exhibited energies within 10 kcal/mol of the isolated ligand.

A series of 26 ribose and 19 arabinose conformational isomers were generated to sample the internal degrees of freedom of the two sugars. The crystal structure coordinates were not included. The 19 arabinose rotamers were generated by starting with the two chair flip conformations of the α and β anomers of the pyranose. Each of these 4 ring conformations adopts 3^4 hydroxyl rotamers, for a total of 324 rotamers. We did not include furanose, aldehyde, or boat conformations. We calculated the CHARMM22 energy of each conformation using TINKER, including a GBSA energy term.[92] Finally, we applied a 6 kcal/mol cutoff above the lowest energy conformation, and then clustered the remaining conformations at 0.5 Å resolution. The clustering process involved selecting the lowest energy conformation, discarding all conformations within 0.5 Å RMS of this conformation, and repeating until no conformations were left. The 26 ribose rotamers were generated the same way, except that an 8 kcal/mol energy cutoff was applied.

These isomers were then rotated in 10° increments along axes defined by a triangulated icosahedron, producing 6516 rotational orientations. Using a fast Fourier transform algorithm,[110] the internal/rotational ensemble was translated along a 0.5 Å grid to find poses that overlapped well with the side-chain regions of the design positions but not with fixed regions of the scaffold. (Figure 22). The energies of poses in this subset, excluding the electrostatic energy, were evaluated. Poses with energies exceeding the energy of the isolated ligand by more than 10 kcal/mol were discarded. The remaining poses were clustered at 0.5 Å resolution to generate the set of poses used for repacking and design calculations.

Structural optimization

We optimize the bound, unbound, and unfolded states separately. The advantage of this multi-state framework[9] is that we can predict conformational changes upon binding, and also optimize for the desired combination of stability, affinity, and specificity. In contrast, algorithms that optimize a single target structure, such as dead end elimination,[34,111] do not distinguish between intramolecular and intermolecular interactions, and thus will propose mutations that stabilize the protein without improving its interaction with the ligand.[61]

We break the protein/ligand system into several parts, identify the low energy conformations of each part, then precompute all the intrinsic energies and interaction energy matrices. This allows us to quickly recalculate the binding energy for different amino acid sequences and conformations — all the energy terms have already been precomputed.

Rotamer probabilities were either initialized randomly, or set to 0's and 1's to match a single structure generated by simulated annealing or by the FASTER procedure.[112] Using a mean-field algorithm, the probabilities were then adjusted iteratively to minimize the free energy of the system.[19] New probabilities for all rotamers were first computed using the mean-field energy of each rotamer and the Boltzmann equation:

$$p_{new} = \frac{\exp(-(E_{self} + E_{interaction}) / RT)}{Z}.$$

Here, Z normalizes the probabilities at a single position so that they sum to one. To prevent oscillating probabilities that do not converge, we updated probabilities with the geometric mean of the old and new values:

$$p_{updated} = \begin{cases} 0, \text{ if } p_{old} < m \text{ and } p_{new} < m \\ rp_{new}, \text{ if } p_{old} < m \\ rp_{old}, \text{ if } p_{new} < m \\ \sqrt{p_{old}\, p_{new}}, \text{ otherwise} \end{cases}$$

where r is a random number between 0 and 0.5, and m is the smallest positive single-precision floating point number ($\sim 1.18\times 10^{-38}$). $p_{updated}$ must be normalized after this procedure. Alternatively, we updated one position at a time in random order, without any probability averaging. The repacking procedure was repeated 10 to 1000 times, using different initial rotamer probabilities. Two-thirds of the repacking runs used the single site update method, and the remainder were run using the simultaneous update method.

The mean field calculation is good at jumping over local barriers that stymie gradient-based minimization or molecular dynamics, because there’s no barrier for sidechains flipping to completely different rotamer configurations. However, there *is* a barrier for multiple simultaneous rotamer changes, so the calculation must be repeated from different starting probabilities.

The most probable structure from the lowest energy mean-field solution was subjected to a final local minimization step. Thus, we discretely sampled a rough energy landscape to identify the lowest-lying energy well, and locally minimized to get to its bottom (Figure 23). The calculated side-chain conformational entropy for different sequences typically varied by less than one kcal/mol, which is small relative to the other energy terms. Hu and Kuhlman also observed that side-chain conformational entropy makes small contributions in their design calculations.[113] However, it is important to note that we did not include entropy changes outside the binding site in our calculation.

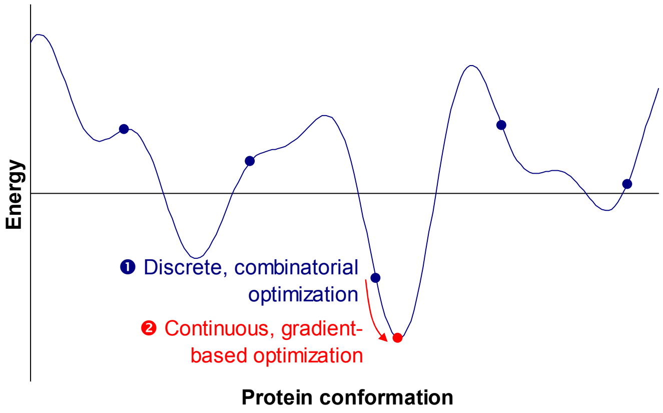

Energy
❶ Discrete, combinatorial optimization
❷ Continuous, gradient-based optimization
Protein conformation

Figure 23. Discrete then continuous optimization of protein structure.

Unfolded state

The intrinsic unfolded-state chemical potential for each amino acid was determined by placing a complete rotamer set at the middle position of an ALA-ALA-ALA tripeptide library comprising multiple peptide backbone conformations with no termini (similar to the approach in [99]). The energy of each configuration was calculated, and the intrinsic unfolded-state chemical potential (Table 6) was evaluated as RT ln(partition sum).

Inter-residue electrostatic interaction energies in the unfolded state were calculated following [58], assuming that the distance distribution between residues is determined by a random walk. The total unfolded-state energy was summed as:

Unfolded state energy =

$$\overbrace{\sum_i \mu(aa_i)}^{\text{Intrinsic unfolded-state chemical potential}} \quad \overbrace{+332*\sum_{i<j}\frac{q_i q_j(\sqrt{6/\pi}-\kappa d\exp(\kappa^2 d^2/6)\,\mathrm{erfc}(\kappa d/\sqrt{6}))}{\varepsilon_{out} d}}^{\text{Inter-residue electrostatic interaction (Gaussian chain model)}}$$

with $d = b_{eff}\sqrt{i-j}+s$.

Variables:

$\mu(aa_i)$	intrinsic chemical potential of am. acid at position i	b_{eff}	effective bond length = 7.5 Å
q_i	charge of the amino acid at position i	s	distance offset = 5 Å
d	RMS inter-residue distance	κ	inverse Debye-Hückel length in $Å^{-1}$

Amino acid	Intrinsic μ (kcal/mol)
ALA	1.39
ARG	–272.99
ASN	–78.70
ASP	–110.07
CYS	1.82
GLN	–57.51
GLU	–86.71
GLY	–8.67
HIS	–44.33
ILE	6.12
LEU	–12.49
LYS	–62.29
MET	–1.62
PHE	6.09
PRO	25.48
SER	5.38
THR	–15.83
TRP	7.68
TYR	–10.22
VAL	1.17

Table 6. Intrinsic unfolded-state chemical potentials for the amino acids in the 6028-member rotamer library.

Sequence optimization (genetic algorithm)

For sequence design, a random population of sequences was initially chosen. Putative energies and structures for each sequence were calculated as described above. The population was then ranked by computed ligand affinity, with a limit on allowable protein destabilization (10 kcal/mol in the initial generations, and 5 kcal/mol in the final generations). The top ranked sequences were mutated and recombined to generate a child population. This evolutionary procedure was iterated until functional improvements ceased to occur. (See Figure 24) We started with a high mutation rate (0.25 mutation probability per position) and low selection stringency (tournament selection where the best of 4 randomly picked sequences is a parent for the next generation). As the population converged, we decreased the mutation rate to 0.15 and increased the selection stringency to tournament selections with 5 – 8 sequences. See Table 7 for details.

Calc phase	Generations	Seqs/gen*	Tournament	Mutation	Destab. (kcal/mol)
1	23	200	4	0.25	10
2	21	200	8	0.2	10
3	21	200	5	0.2	5
4	21	200	5	0.15	5

* The initial generation of calculation phases 1 – 3 had between 175 and 224 sequences, depending on how many top sequences were included from the previous phase. The initial generation of calculation phase 4 had 844 sequences, which included all point mutants of the top 3 sequences, double mutants of the top sequence, and random recombinants of the top sequences.

Table 7. Genetic algorithm parameters.

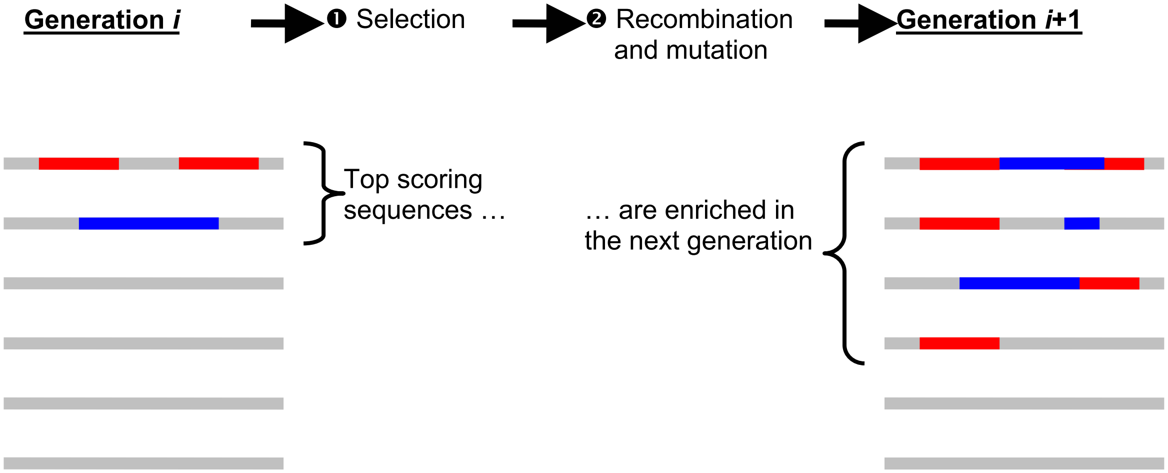

Generation *i*
❶ Selection
❷ Recombination and mutation
Generation *i*+1
Top scoring sequences …
… are enriched in the next generation

Figure 24. Genetic algorithm.

Appendix: Integrals

To calculate generalized Born radii, we integrated r^{-4} or r^{-5} outside the rectangular region $x_1 < x < x_2$, $y_1 < y < y_2$, $z_1 < z < z_2$ using these formulas:

$$\iiint\limits_{\substack{\text{outside}\\\text{rectangular}\\\text{region}}} \frac{dx\,dy\,dz}{(x^2+y^2+z^2)^2} = \sum_{i=1}^{3}\sum_{j=1}^{2}\sum_{k=1}^{2}\left((-1)^{j-k+1} \begin{pmatrix} g_4(x_j,y_k,z_1,z_2) & \text{if } i=1 \\ g_4(x_j,z_k,y_1,y_2) & \text{if } i=2 \\ g_4(y_j,z_k,x_1,x_2) & \text{if } i=3 \end{pmatrix} \right),$$

$$\text{with } g_4(a,b,c_1,c_2) = \frac{\sqrt{a^2+b^2}\left(\tan^{-1}\left(\frac{c_1}{\sqrt{a^2+b^2}}\right) - \tan^{-1}\left(\frac{c_2}{\sqrt{a^2+b^2}}\right)\right)}{2ab}$$

$$\iiint\limits_{\substack{\text{outside}\\\text{rectangular}\\\text{region}}} \frac{dx\,dy\,dz}{(x^2+y^2+z^2)^{5/2}}$$

$$= \frac{1}{6}\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}\left((-1)^{i+j+k} \frac{\sqrt{x_i^2+y_j^2+z_k^2}}{x_i y_j z_k} \right)$$

$$+ \frac{1}{6}\sum_{h=1}^{3}\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}\left((-1)^{i+j+k} \begin{pmatrix} g_5(x_i,y_j,z_k) & \text{if } h=1 \\ g_5(x_i,z_j,y_k) & \text{if } h=2 \\ g_5(y_i,z_j,x_k) & \text{if } h=3 \end{pmatrix} \right),$$

$$\text{with } g_5(a,b,c) = \frac{\tan^{-1}\left(\frac{ab}{c\sqrt{a^2+b^2+c^2}}\right)}{c^2}.$$

Chapter 5: Physics-based design of new binding proteins

This chapter reports on unpublished work. Several collaborators have helped with further characterizing these designed proteins, although the results are not yet final and are not reported here. Pavel Strop solved a unliganded crystal structure of one of the designed proteins, and it agrees well with the prediction. Rebecca Fenn is currently working on solving a liganded crystal structure. She and Jan Lipfert collected small angle X-ray scattering data on one of the designed proteins, which shows that it undergoes the same conformational change upon binding the target ligand as the native ribose binding protein.

Summary

Using a standard molecular mechanics potential energy function, we redesigned ribose binding protein to bind a series of ligands: L-arabinose, D-xylose, indole-3-acetic acid, and estradiol. The resulting proteins have 5 – 10 mutations from the native, are stable, the predicted structures have good hydrogen bonds and shape complementarity, and they use motifs similar to natural binding proteins. All of the designed proteins bind to their target ligands with measurable but weak affinity. The affinity was improved by random mutagenesis and screening. Combined with our earlier results, this is the first time a single model has been used to predict structures, binding constants, and to design new small-molecule binding sites. Using a standard model should improve the

generality of protein design, which could enable the creation of custom proteins for a wide variety of applications, including sensors, enzymes, and protein therapeutics.

Introduction

There are many well-established experimental techniques for creating new binding sites in proteins: phage display, antibodies, and gene shuffling. These techniques start with large random libraries of proteins and select or screen for sequences that bind to the desired target. They are limited by the library size and the availability of appropriate selections and screens. For example, randomizing 12 residues in a protein yields a sequence size of 10^{15}, but phage display libraries generally contain fewer than 10^{10} different sequences.[114] Devising selections can be difficult, especially for small molecules that can not be attached to solid support without disrupting a large fraction of the ligand's available binding surface area. Furthermore, selections for catalysis are limited by the accuracy and synthetic accessibility of a transition state analog.

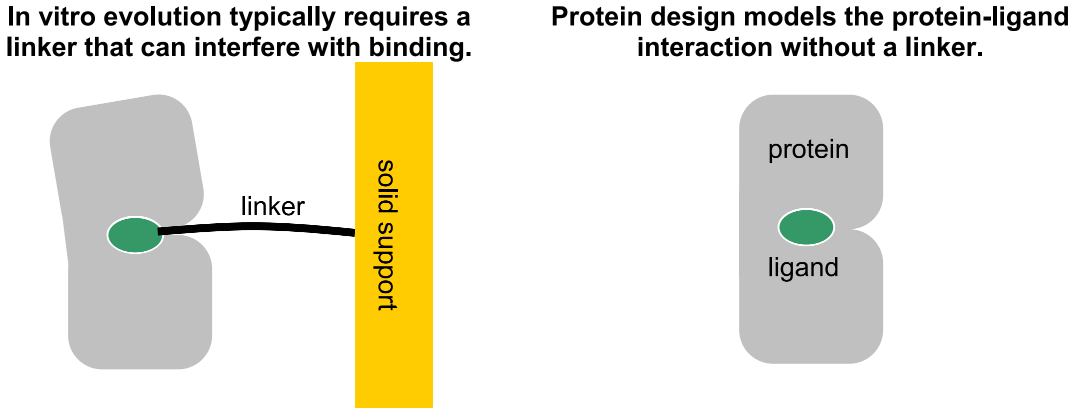
In the long term, we anticipate that a computational technique for engineering protein-ligand binding can address some of these limitations. For example, with modern computers, the sequence search algorithms can effectively access a larger sequence space than a phage display library. The computational techniques are also not limited by experimental constraints such as linkers (Figure 25), and they can directly model an unstable transition state rather than using a stable transition state analog.

In the short term, these design calculations provide perhaps the most rigorous test of the current models of protein structure and energetics.

We previously described a protein design algorithm that uses a standard molecular mechanics potential energy function with an accurate continuum solvent model.[62] The design algorithm takes the structure of a scaffold protein, and the structure of a small molecule, and designs a set of mutations needed to create a binding site in the scaffold. We only consider mutations at a limited number of “design positions”; the rest of the protein simply serves as a rigid structure for constraining the conformational flexibility of the designed binding site.

In this paper, we use this algorithm to switch the ligand specificity of ribose binding protein (RBP). High resolution crystal structures have been solved for both bound and unbound RBP.[115,116] The binding site is lined with sidechain and not backbone atoms, which may facilitate its use as a scaffold. This test system for protein design was pioneered by Hellinga,[2] who designed trinitrotoluene, lactate, and serotonin binding sites in various bacterial periplasmic binding proteins, including ribose binding protein. They showed that these designed proteins could be used as sensors, and could be incorporated into signaling pathways that drive gene expression in response to trinitritoluene or lactate. Their landmark paper used a molecular mechanics potential energy function (CHARMM22) that was modified by scaling the van der Waals repulsion energy, using a distance dependent dielectric constant, explicit hydrogen bond term, and various other modifications.

In contrast, we test whether an unmodified molecular mechanics potential energy function (CHARMM22) can be used for a similar set of binding site design problems. Using a standard model should improve the generality of protein design,[78] which could enable the creation of custom proteins for a wide variety of applications, including sensors, enzymes, and protein therapeutics.


In vitro evolution typically requires a linker that can interfere with binding.

Protein design models the protein-ligand interaction without a linker.

Figure 25. In vitro evolution vs computational protein design.

Results

We picked the 10 primary ribose contacts in RBP as the core set of design positions, and computationally redesigned the protein to bind L-arabinose, D-xylose, indole-3-acetic acid, and estradiol (Figure 26). Additional design positions were picked as needed in subsequent iterations of the design calculation (Table 8). D-xylose differs from the native ligand, ribose, at a single stereocenter, and L-arabinose differs at 2 sterocenters. Indole-3-acetic acid is the major plant growth hormone, and the ligand parameters can be copied from tryptophan. Estradiol, the major estrogen in mammals, was picked as a prototypical hydrophobic ligand.

Ligand parameters (Table 9) were validated by calculating the free energy of the α and β anomer of each sugar and comparing it to the experimental value. (Figure 27).

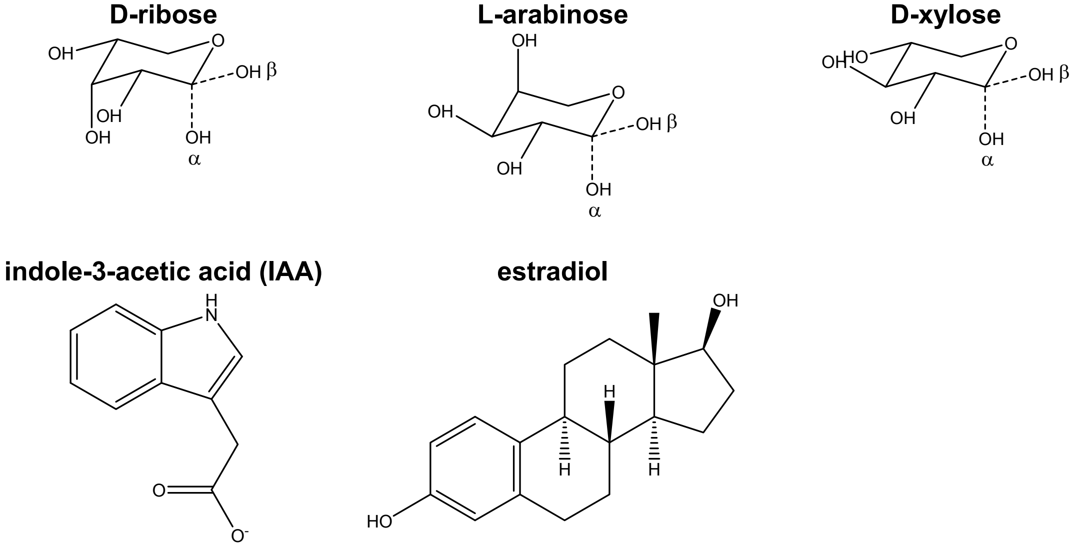


Figure 26. Target ligands.

	9	13	15	16	89	90	103	105	141	164	190	215	235
RBP (native)	SER	ASN	PHE	PHE	ASP	ARG	SER	ASN	ARG	PHE	ASN	ASP	GLN
RBP→arabinose 2	GLN	ASN	MET	TYR	VAL	MET	GLN		MET	PHE	ASN	SER	VAL
RBP→xylose 1		ASN	PHE	PHE	GLN	GLN			MET	PHE	ASN	SER	MET
RBP→xylose 2		MET	TYR	PHE	GLN	HIS			MET	PHE	ASN	SER	GLN
RBP→estradiol 4	SER	ASN	VAL	MET	ALA	ASN	ASN		MET	PHE	ASN	SER	ILE
RBP→IAA 1		ARG	THR	MET	VAL	MET	HIS	TYR	MET	PHE	ASN	ALA	SER
RBP→IAA 2		ARG	THR	MET	ALA	MET	HIS	TYR	MET	PHE	ASN	SER	SER
RBP→IAA 3		ARG	THR	MET	VAL	ASN	HIS	TYR	MET	PHE	ASN	ALA	SER
RBP→IAA 101A-F11		ARG	SER	MET	GLY	CYS	HIS	TYR	MET	PHE	ASN	ALA	SER
RBP→IAA 95A-C1		ARG	SER	MET	ILE	CYS	HIS	TYR	MET	PHE	ASN	ALA	SER

Table 8. Sequences of RBP redesigned to bind other ligands.
Sequence is only shown at positions being designed. Positively charged amino acids are colored blue, negatively charged amino acids are colored red, polar amino acids are colored blue, and nonpolar amino acids are colored black.

Ligand	Atomic partial charges	Other energy terms
D-ribose	CHARMM22	CHARMM22
L-arabinose	MM3/PM5	CHARMM22
D-xylose	CHARMM22	CHARMM22
IAA	CHARMM22	CHARMM22
estradiol	Pullman	Tripos force field

Table 9. Ligand parameters.

Pullman charges[117] were calculated using Sybyl (Tripos, St. Louis, MO). MM3/PM5 charges were calculted using CaChe (Fujitsu, Newton, MA). CHARMM22[14] energies were calculated using TINKER[98]. Tripos force field[118] energies were calculated using Sybyl.

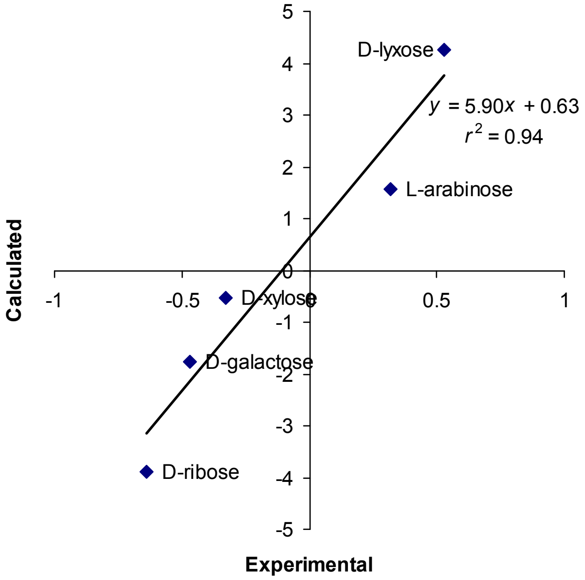


Figure 27. Experimental[119,120] and calculated β-pyranose energy – α-pyranose energy (kcal/mol).

Effect of softening the van der Waals energy

The van der Waals energy is frequently softened so as not to penalize the small steric clashes resulting from limited sampling resolution. A side effect of this is to make hydrogen bonds appear stronger than they actually are (Figure 28). This encourages the design algorithm to bury charges and polar residues at a designed hydrophobic interface (Table 10). Therefore, an unmodified VDW energy was used to design the proteins described in this paper.

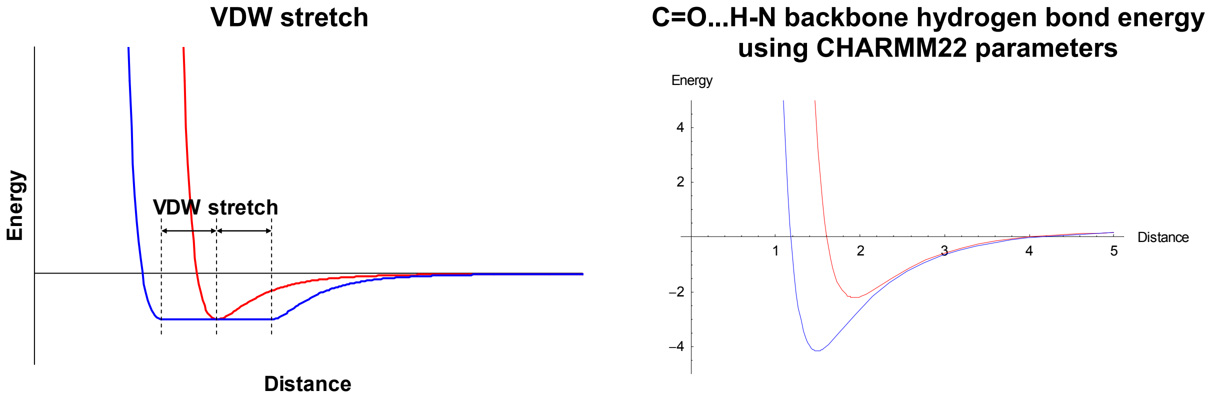


Figure 28. The Lennard-Jones potential is frequently softened in design calculations to compensate for low sampling resolution. However, this has the side effect of making hydrogen bonds appear artificially strong. The figure shows the energy of a C=O...H-N backbone hydrogen bond energy (Lennard-Jones plus Coulomb energy using CHARMM22 parameters). The red line uses the standard Lennard-Jones energy term (total energy has a minimum of –2.2 kcal/mol at 1.9 Å). The blue line uses a van der Waals function where the minimum energy has been expanded by ± 0.3 Å (total energy has a minimum of –4.1 kcal/mol at 1.5 Å).

	13	15	16	89	90	141	164	190	215	235	Average hydrophobicity
Designed estradiol binding site in RBP (VDW stretch = 0.3)	ARG	GLU	MET	SER	ALA	MET	PHE	ASN	SER	SER	-0.55
	ARG	GLU	MET	TYR	SER	MET	TYR	ASN	GLN	ASN	-1.81
	ARG	GLU	LEU	ALA	LEU	ILE	PHE	ASN	ASN	SER	0.09
	ARG	GLU	THR	ALA	ASN	MET	ASN	GLU	ALA	ALA	-1.19
	ARG	GLU	THR	ALA	ASN	MET	ASN	ASN	ASN	VAL	-1.48
	HIS	GLU	LEU	MET	ASN	GLU	PHE	ASN	GLU	THR	-1.29
	ARG	GLU	MET	TYR	SER	MET	TYR	ASP	GLN	ASN	-1.81
	ARG	ASP	ALA	MET	ASN	MET	ASN	ASN	ASN	THR	-1.71
	ARG	GLU	THR	ALA	ASN	MET	ASN	GLU	ASN	ALA	-1.72
	ARG	PHE	LEU	ALA	THR	MET	PHE	ASN	SER	VAL	0.78
Designed estradiol binding site in RBP (VDW stretch = 0.0)	ASN	ALA	MET	ALA	ASN	MET	PHE	ASN	ALA	ALA	0.33
	ASN	VAL	MET	ALA	ASN	MET	PHE	ASN	ALA	ALA	0.57
	ASN	VAL	MET	ALA	ASN	MET	PHE	ASN	ALA	SER	0.31
	ASN	SER	MET	ALA	ASN	MET	PHE	ASN	ALA	ALA	0.07
	ASN	VAL	MET	SER	ASN	MET	PHE	ASN	ALA	ALA	0.31
	MET	VAL	MET	ALA	ALA	MET	PHE	ASN	ALA	ALA	1.64
	ASN	VAL	MET	ALA	ASN	MET	PHE	ASN	SER	ALA	0.31
	ASN	ILE	MET	ALA	ASN	MET	PHE	ASN	ALA	ALA	0.6
	SER	VAL	MET	ALA	ASN	MET	PHE	ASN	ALA	ALA	0.84
	ASN	VAL	LEU	ALA	ASN	MET	PHE	ASN	ALA	ALA	0.76

Table 10. Designed estradiol binding site in RBP is more polar when VDW stretch = 0.3 Å.

The average hydrophobicity [121] of the designs with VDW stretch = 0.3 Å is –1.07, the average hydrophobicity of the designs with VDW stretch = 0.0 Å is 0.57, and the average hydrophobicity of the human estrogen receptor binding site (PDB code: 1A52) is 1.75. Even without the VDW stretch, the designs are still more polar than the human estrogen receptor. Most of the remaining polar residues are retained from the native sequence, so this is presumably due to limitations imposed by the scaffold protein. 2800 rotamers were modeled at each design position.

Structures of designed receptors

We examine the shape complementarity and hydrogen bonding of the designed binding proteins in Figure 29 and Table 11. All of the designed proteins have good shape complementarity and hydrogen bonding, comparable to natural binding proteins.

The designed estradiol receptor has a hydrogen bond to one of the hydroxyls. There is a conserved phenylalanine and two methionines seen at comparable positions in both the human estrogen receptor and the designed estradiol binding protein, which is remarkable given that these binding sites are hosted on proteins with completely different folds. The phenylalanines interact with the estradiol via favorable electrostatic π-π interactions,[122] and the methionines interact via favorable hydrophobic interactions.

The designed indole acetic acid binding protein has an arginine forming a salt bridge with the carboxylic acid in the ligand, and all of the hydrogen bond donors and acceptors in the ligand are satisfied.

Importantly, these binding motifs were picked out directly from an unmodified molecular mechanics potential energy function, and not by explicitly asking the design algorithm for particular types of interactions.

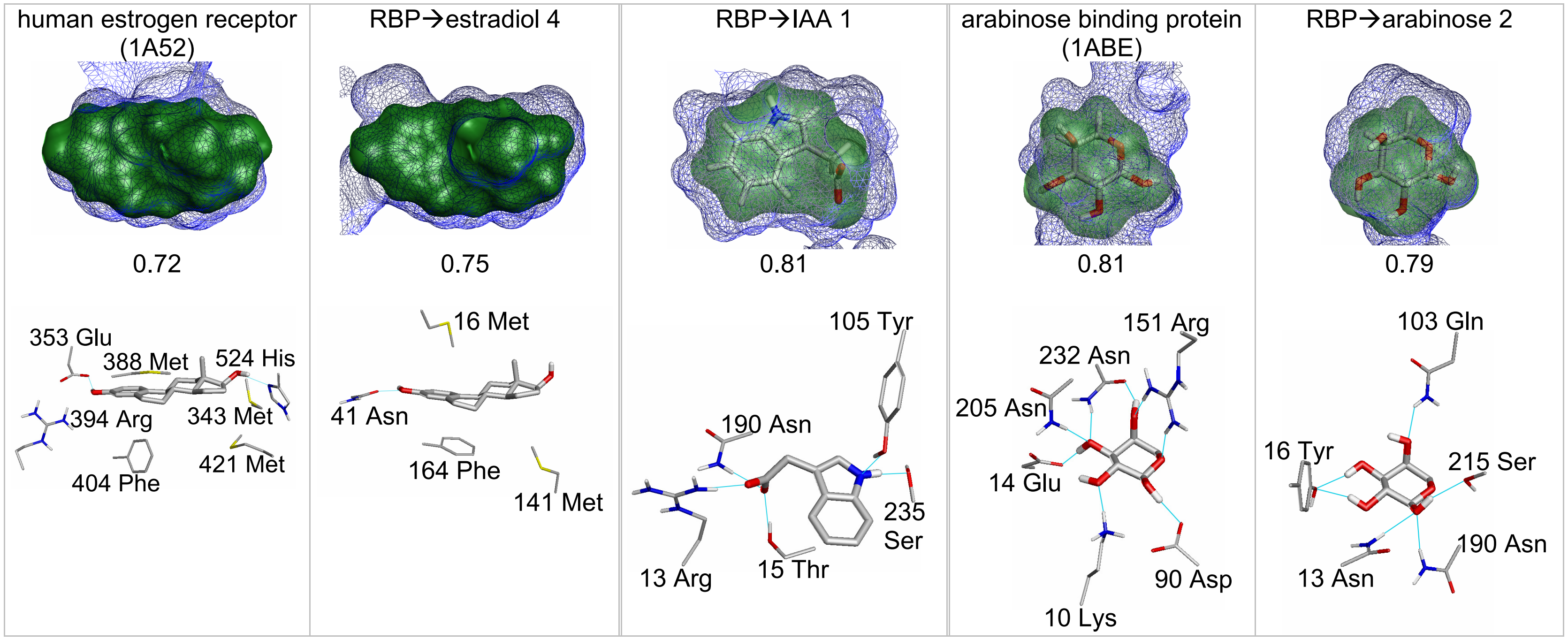


Figure 29. Structures of designed and natural binding proteins.

Top row: ligand (solid green) and protein binding pocket (blue mesh). The number is the shape complementarity [91], which ranges from 0 for no complementarity to 1 for perfect complementarity. Bottom row: Hydrogen bonds and other key protein-ligand interactions. Crystal structures are shown for the natural binding proteins, and predicted structures shown for the designed proteins.

Ligand	**Protein**	***K_d***	**Protein stability (kcal/mol)**	**Protein-ligand hydrogen bonds**	**Shape complementarity**
ribose	RBP (2DRI)	210 nM†	2.5	11	0.86
L-arabinose	RBP	790 mM*	2.5		
	ABP (1ABE)	190 nM*		8	0.81
	AraC (2ARC)			6	0.77
	RBP→arabinose 2	250 mM*		6	0.79
D-xylose	RBP	700 mM*	2.5		
	RBP→xylose 1	160 mM*	4.2	5	0.82
	RBP→xylose 2	270 mM*		5	0.80
estradiol	RBP	60 mM*	2.5		
	Human estrogen receptor (1A52)	10 pM‡		2	0.72
	IgG – estradiol (1JGL)	2 nM‡		4	0.87
	RBP→estradiol 4	46 mM*	2.0	1	0.75
IAA	RBP	32 mM*	2.5		
	RBP→IAA 1	11 mM*	2.5	5	0.81
	RBP→IAA 2	14 mM*	1.0	5	0.79
	RBP→IAA 3	16 mM*	1.9	5	0.82
	RBP→IAA 101A-F11	1.4 mM*	2.0	5	0.69
	RBP→IAA 95A-C1	1.1 mM*	4.4	3	0.73

Table 11. Properties of designed and natural binding proteins.

Designed and selected proteins are highlighted. K_d was determined as follows: * solid phase radioligand binding assay, † centrifugal concentrator assay, ‡ published value. Stability was measured by extrapolating urea denaturation curves to 0 urea concentration. Hydrogen bonds and shape complementarity were calculated using predicted structures for designed proteins, and the crystal structures for native proteins.

Experimental characterization of designed receptors

The measured dissociation constants for the designed proteins are shown in Table 11. The native has very low affinity for the target ligands, and the designed proteins all improve on this affinity, although the K_d's are still in the millimolar range. The designed proteins have expression levels and stabilities comparable to the native, despite having 5 – 10 mutations from the native. In contrast, if we remove the stability requirements from the design calculation, the resulting designed proteins have low expression levels and little secondary structure as measured by circular dichroism.

Since the designed interactions are so weak, they might be due to a non-specific effect, such as destabilization of the protein, or a simple change in the size of the binding pocket. To address this possibility, we constructed a library of RBP variants using mutagenic PCR,[123] and also by QuikChange mutagenesis with degenerate codons (N N G/C) to randomize positions in the binding site. We sequenced 12 random clones from the library, and they had an average of 3.1 mutations/clone, with only a single sequence containing no mutations. We then screened 48 library members for binding to xylose and arabinose. The tightest binder from both screens was the native sequence, indicating that the improved binding affinity of the designed sequences is not due to a non-specific effect.

Experimental screen

Given the good shape complementarity and hydrogen bonding in the predicted binding site structures, the weak affinity of the designed interactions is surprising. To test the possiblity that the designed sequences are close to a more optimal solution, but missed it because of errors in the potential energy function, or limitations in the structural sampling, we constructed a library of variants of the RBP→IAA design. Part of the library was generated using mutagenic PCR starting from the 3 designed sequences. The rest of the library was generated using QuikChange mutagenesis with degenerate oligos designed to match the amino acid frequencies seen in the top 72 sequences from the RBP→IAA design, including a low mutation rate to other amino acids (Table 12). We screened 279 sequences from the library, and the best two sequences, 95A-C1 and 101A-F11, have dissociation constants of 1.1 mM and 1.4 mM respectively (Table 8, Table 11). In the next round of selection, the 3 designed sequences and the top 4 sequences from the screen were shuffled,[124] followed by mutagenic PCR.[125] 186 sequences were screened from the second round, and no further improvement was seen in binding affinity. Both of our top hits contain mutations to cysteine, which were not allowed in the design calculation to prevent disulfide bond formation.

Thus, the only way the screen was able to improve the affinity of the designed binding proteins was by going outside the parameters of the original design problem. This suggests that the design calculation may have done the best job possible, given the constraints of the scaffold and the mutations it was allowed to make. To examine this hypothesis further, we took the top two sequences from the screen and plugged them back into the calculation to determine their predicted affinities. 101A-F11 is predicted to bind tighter than the designed sequences, which is correct. 95A-C1 is predicted to bind less well than the designed sequences, which is incorrect. Thus, 101A-F11 was missed by the design algorithm because of the sequence restrictions on the design algorithm, and 95A-C1 was missed because of problems with the sampling or potential energy function.

	A3			B2		C2		D	E	F	G	H
	13	15	16	89	90	103	105	141	164	190	215	235
ALA	0%	0%	0%	33%	7%	0%	0%	0%	0%	0%	61%	46%
ARG	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
ASN	0%	0%	0%	0%	8%	18%	0%	0%	0%	100%	3%	0%
ASP	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%
GLN	0%	0%	0%	0%	3%	0%	0%	0%	0%	0%	0%	0%
GLU	0%	0%	0%	0%	0%	3%	0%	0%	0%	0%	0%	0%
HIS	0%	0%	0%	0%	0%	32%	0%	0%	0%	0%	0%	0%
ILE	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
LEU	0%	0%	0%	0%	3%	0%	0%	0%	0%	0%	0%	0%
LYS	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
MET	0%	0%	100%	1%	32%	0%	0%	100%	0%	0%	0%	0%
PHE	0%	0%	0%	0%	6%	0%	25%	0%	100%	0%	0%	0%
SER	0%	28%	0%	0%	0%	0%	0%	0%	0%	0%	36%	53%
THR	0%	72%	0%	13%	0%	0%	0%	0%	0%	0%	0%	1%
TRP	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
TYR	0%	0%	0%	0%	25%	0%	75%	0%	0%	0%	0%	0%
VAL	0%	0%	0%	51%	17%	47%	0%	0%	0%	0%	0%	0%
exp(–Σ *p* ln *p*)	1.0	1.8	1.0	3.0	5.8	3.1	1.8	1.0	1.0	1.0	2.2	2.1

Table 12. RBP-IAA library.
We generated a library of RBP variants based on amino acid frequencies in the top 72 sequences from the RBP→IAA design, plus a low frequency of mutation to other amino acids. In the first round, 10% of the oligos had degenerate N N G/C codons. In the second round, 10 – 25% of the oligos had degenerate N N G/C codons. Letters A – H indicate mutagenic oligos

Discussion

We redesigned RBP to bind a series of other ligands, using a standard molecular mechanics potential energy function. The resulting proteins have 5 – 10 mutations from the native, are stable, and the predicted structures have good hydrogen bonds and shape complementarity, and use similar motifs seen in natural binding proteins. All of the designed proteins bound to their target ligands with measurable but very weak affinity, in the millimolar range.

Furthermore, we show that protein design can be used to design libraries for screening. Essentially, the design algorithm picks out a promising region of sequence space, vastly reducing the number of sequences that must be screened experimentally.

Why do the designed binding proteins have such poor affinity for their target ligands? Several aspects of the design algorithm need improvement: the energy function, structural sampling, and scaffold selection.

Current molecular mechanics potential energy functions have several known limitations. They mispredict hydrogen bond geometries [126], ignore protein polarization, do not model lone pairs, and do not model quantum effects. Furthermore, continuum solvent models do not properly treat tightly bound water molecules. Many groups are working to address these limitations, but this is a challenging problem, because fixing one problem can often have unintended side effects. Thus, changes to the potential energy function must be tested against a wide range of experimental data and quantum calculations.

Structural sampling is also a problem, due to the huge space of potential protein conformations. Currently, we use a fixed backbone and only model rotamer flexibility for sidechains directly contacting the ligand. However, positions far from a binding site can often affect binding,[127] so it may be important to include additional design positions. More sampling will be possible with increases in computer power, but there is also room for clever sampling strategies. For example, Baker includes backbone flexibility by alternating between sequence design on a fixed backbone, and structural optimization for a designed sequence.[7] However, greater structural sampling also requires a more accurate energy function, as there is a wider range of conformations to be evaluated. In other words, limited structural sampling can constrain a poor energy function from straying too far from reality.

Scaffold selection is perhaps the least examined step in protein design, but it is important to choose a scaffold that is compatible with the ligand. Presumably, it will be easier to redesign a protein to bind a ligand that is similar to the natural ligand. Some protein folds can host a wide range of binding sites, such as antibodies binding different antigens, or alpha/beta barrel proteins which host a wide range of enzyme active sites.[1] Even these natural scaffolds have limitations: antibodies, for example, do not easily bind to certain targets.[128] Beyond these observations, there are very few general rules for picking the right scaffold.

Materials and methods

Characterization of designed proteins

Designed proteins were constructed, expressed, purified, and binding constants were measured as decribed earlier.[62] For the solid phase radioligand binding assay, the wash solution was chosen to optimize the ratio of ligand eluted from Ni-NTA resin + protein and ligand eluted from Ni-NTA resin alone. Xylose binding assays used water for the wash. IAA, ribose, and arabinose binding assays used 50% (v/v) ethanol +

50% water for the wash. Estradiol binding assays used ethanol for the wash. Protein stability was calculated from urea melting curves measured using the circular dichroism signal at 220 nm by linearly extrapolating the measured stability back to 0 urea concentration.[129]

Library screening

The libraries were transfected into BL21 DE3 *E coli*, and clones were expressed in 1.3 ml culture in 96-well blocks using Airpore tape (Qiagen). Cultures were shaked at 300 rpm for 5 hours at 37°C, induced with 1 mM IPTG, and shaked for 5 hours more. Protein was purified using Qiagen Ni-NTA resin using the manufacturer's protocol. For native RBP, this yields 1 nmol protein / well. Binding was measured using a solid phase radioligand assay,[62] assumuing native levels of expression. This effectively penalizes poorly expressed proteins by raising their apparent K_d.

Chapter 6: Conclusion

Computer simulation tools play an important role in established branches of engineering, such as designing aircraft, bridges, buildings, and circuits. In many cases, the computer can tell us how good our design is before we even build it. Our goal is to produce tools that will move protein engineering towards the same level of sophistication. Specifically, we've developed an algorithm to engineer a new ligand binding site into a protein, or remodel an existing one to fit a different ligand. This algorithm could eventually be used to design custom sensors, enzymes, and protein therapeutics.

The design algorithm has two major components: a calculation to determine the ligand's binding affinity for a given amino acid sequence, and a genetic algorithm that "evolves" the amino acid sequence to optimize the calculated binding affinity. To calculate the energy of a given molecular conformation, we use a standard molecular mechanics potential with an accurate continuum solvent model. To model conformational changes and thermal fluctuations, we represent the protein / ligand system as a probabilistic ensemble of different backbone, side chain, and ligand conformations. To ensure stability and specificity, we compare the free energy of the bound state with several competing states, such as the unbound state or the protein bound to a related molecule.

Using this algorithm, we were able to predict binding constants, active site structures, and to design new small molecule binding proteins (Figure 30). This is the first successful redesign of an entire binding site based on an unmodified molecular-mechanics potential energy function. It is also the first time a single model has been used to predict structures, binding constants, and to design new small-molecule binding sites.

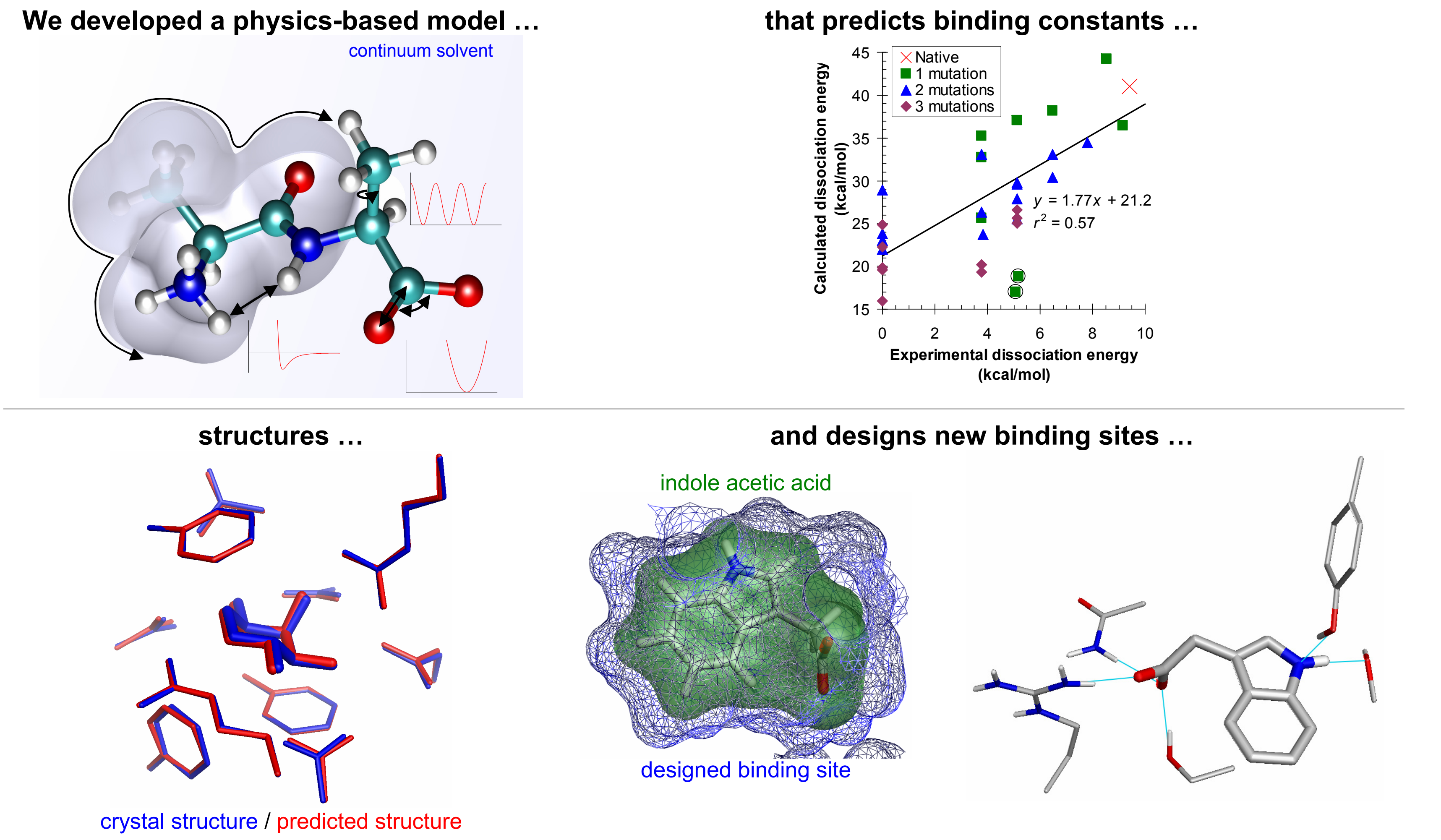

We developed a physics-based model …
continuum solvent
that predicts binding constants …
45
40
35
30
25
20
15
0
2
4
6
8
10
Native
1 mutation
2 mutations
3 mutations
y = 1.77x + 21.2
r^2 = 0.57
Calculated dissociation energy (kcal/mol)
Experimental dissociation energy (kcal/mol)
structures …
crystal structure / predicted structure
and designs new binding sites …
indole acetic acid
designed binding site

Figure 30. Summary figure.

Designing for specificity

In this thesis, the genetic algorithm was used to select for affinity and stability. Other types of selections are also possible. For example, we can select for specificity, hydrogen bonding, geometry of catalytic residues, etc. Figure 31 shows that we can select out mutants of arabinose binding protein (ABP) that have various combinations of predicted affinity and specificity for arabinose and galactose. Native ABP binds to both arabinose and galactose. Figure 32 shows that predicted specificity for arabinose is achieved by creating a steric clash with the extra CH_2OH group in galactose.

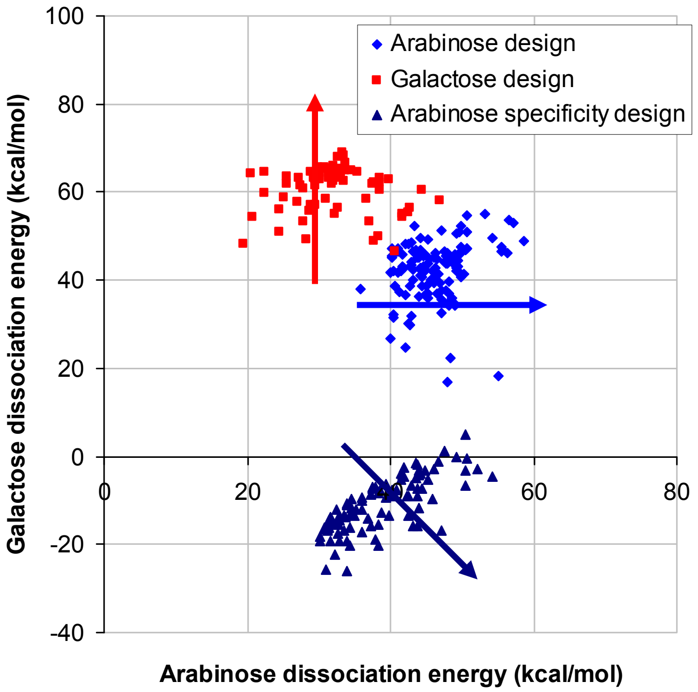


Figure 31. Designing for specificity in arabinose binding protein. Each point represents a single sequence.

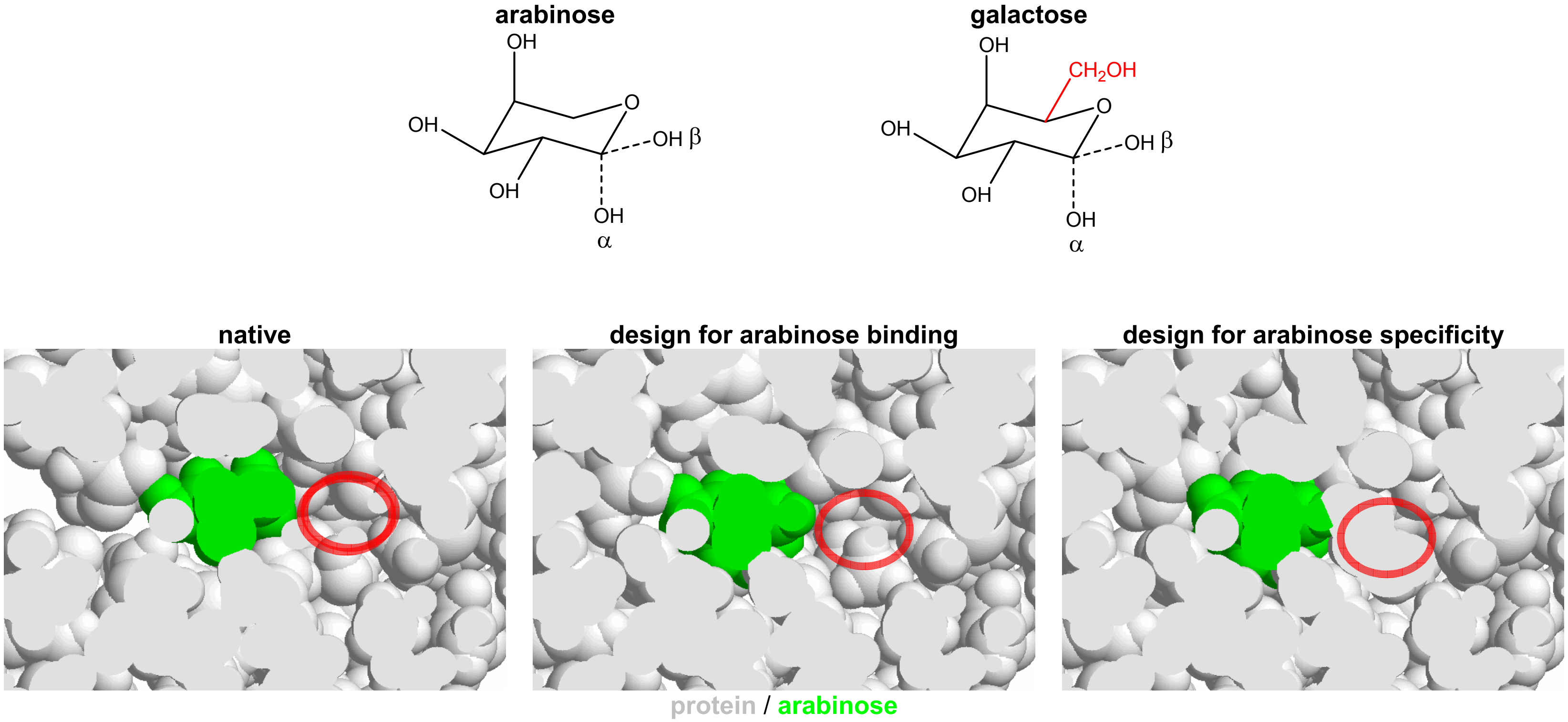


Figure 32. Structural determinants of specificity.

Space for galactose CH_2OH is seen in the native and arabinose binding design (both of which bind galactose), but not in the specificity design.

Energetic vs structural predictions

We have found that structures are easier to predict than energies. This can be understood as follows. If the bound state is a deep well in the energy landscape, then errors in the energy function will affect the well depth (dissociation energy) much more than the well position (bound structure). See Figure 33.

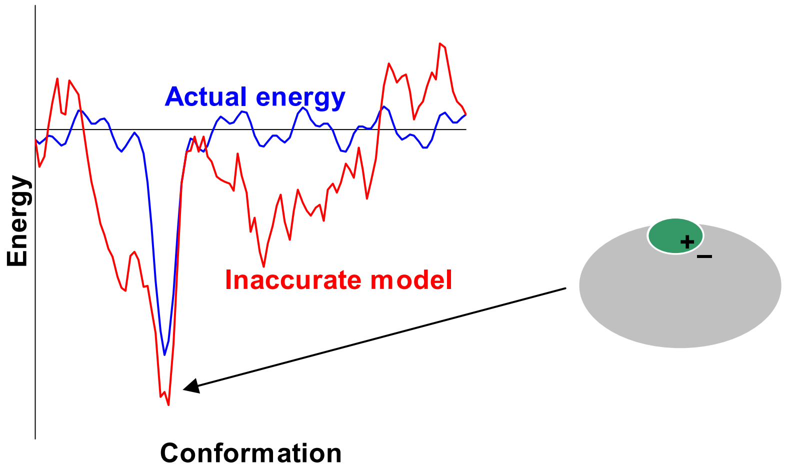


Figure 33. Energetic vs structural predictions in an inaccurate energy model.

Comparison to more established branches of engineering

Computational protein design is a young field. Can we take any clues from more established branches of engineering? For example, in electronic circuit design, if you connect a bunch of transistors in a random fashion, it will be very difficult to predict the behavior of the system without doing detailed computer simulations. Yet, this is our approach to protein design: generate a bunch of random sequences and try to predict their behavior. Of course, circuits are not designed that way. Instead, circuits are built up from a set of modular components, with well defined rules for how these modules can be connected to each other. For example, logic inputs have to be at 0 or 5 volts, you can't switch them too fast, there's a maximum amount of current you should try to draw from certain outputs, and so on. If you follow these rules, then you can understand the behavior of the system just by thinking about it. However, if you violate the rules, then it's harder to predict what will happen, and you have to look inside each module to figure out what will happen.

Similarly, it might be possible to design a set of modular components for use in protein engineering. For example, specific protein-protein interaction motifs might be displayed side-by-side in a combinatorial fashion to create a larger set of interaction motifs.

Alternatively, there might be a set of rules for identifying amino acid sequences whose behaviour will be difficult to predict. This set of rules would be used to screen which sequences are run through a mean field calculation.

Application: Sensors

Binding is perhaps the simplest function that a protein can perform. However, with appropriate modifications to the scoring function that the genetic algorithm uses to select "good" sequences, the binding site design algorithm can be extended to engineer proteins with more sophisticated functions.

For example, if the binding event can be transduced into a detectable signal, this would produce a biosensor (Figure 34). One strategy involves adding a prosthetic fluorophore that occludes a binding site.[130] Then, when the ligand binds, the fluorophore will swing out into solvent, changing its fluorescence. A second strategy involves attaching two fluorophores or fusing two fluorescent proteins to a protein that undergoes a conformational change upon binding.[131-134] Then, ligand binding can be detected as a change in fluorescence resonance energy transfer (FRET). A third strategy involves taking an allosteric enzyme that produces a colored product, and engineering a binding site that stabilizes the enzyme's active conformation.[135,136]

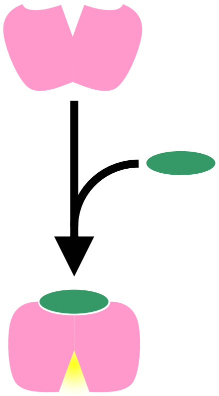


Figure 34. Biosensor. Binding induces a conformational change that results in a change in fluorescence or enzymatic activity.

There already are many systems for detecting small molecules from complex mixtures, including mass spectrometry, antibody-based assays, enzyme-coated electrodes, and arrays of materials whose electrical or physical properties change when molecules are adsorbed. Computationally engineered protein biosensors have some potential advantages over these competing technologies. Importantly, the signal readout is directly coupled to binding, does not require additional reagents or expensive equipment, and the analyte does not need to be labeled. Furthermore, once a suitable sensor scaffold is identified, it can be engineered to detect a wide range of different molecules, enabling the creation of small-molecule microarrays that could detect a panel of biomarkers for medical diagnostic purposes (Table 13).

Application		Molecule
Subcellular fluorescence imaging of signaling molecules		IP_3, cAMP, leukotrienes
Detecting bacteria	Gram negative bacteria	lipid A (conserved portion of LPS)
	Staph. aureus	toluene
	Klebsiella pneumoniae	methyl ethyl ketone
	Pseudomonas aeruginosa	o-aminoacetophenone
	Bacillis	dipicolinic acid
	Bacterial vaginitis	putrescine
Blood / urine tests	Metabolic molecules	≈ 2500 metabolites in humans
	Hormones	
	Therapeutic drugs with a low therapeutic index	theophylline, digoxin, phenytoin, cyclosporine, methotrexate
	Illicit drugs	
	Toxins	
Cancer screening	Lung cancer	in breath: toluidine, acetophenone, benzothiazole
	Pheochromocytoma or neuroblastoma	in urine: vanillylmandelic acid
	Carcinoid tumors	in urine: 5-hydroxyindoleacetic acid
Fertility test	In axillary secretions during ovulation:	dehydroxyepiandrosterone sulfate

Table 13. Sensor applications.

Application: Custom enzymes

Modeling enzymes computationally is much more difficult than modeling non-covalent binding, because making and breaking covalent bonds needs to be treated quantum mechanically. However, simply binding the substrates in the correct orientation required for reactivity and hence stabilizing the transition state, can significantly speed up a reaction. For example, when the entire catalytic triad (Asp, His, Ser) of a serine protease is mutated, the mutant enzyme still produces 3 orders of magnitude of rate enhancement.[137] Published catalytic antibodies provide up to 8 orders of magnitude rate enhancement, although 4 orders of magnitude is more typical. Thus, even if we ignore covalent chemistry, we can still get significant catalysis with non-covalent stabilization of the transition state (Figure 35).

Furthermore, if we do know the desired geometry of the catalytic residues, we can ask the genetic algorithm to rank sequences based on both non-covalent transition state stabilization and proper predicted orientation of the catalytic residues.

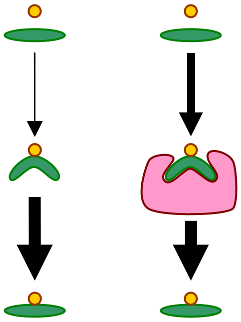


Figure 35. Binding to the transition state of a reaction catalyzes that reaction.

Custom enzymes could be used for chemical synthesis and pharmaceutical manufacturing. The extraordinary specificity of enzyme-catalyzed reactions stands in stark contrast to inorganic catalysts, which are typically very promiscuous. Thus, laboratory synthesis of complex molecules typically involves multiple steps, and multiple protecting groups that have to be added and removed at various points during the reaction. Having a customizable toolbox of enzymes that performs a desired set of reactions specifically could enable one-pot multi-step synthesis without protecting groups.

Other possible applications for custom enzymes include: custom proteases and restriction enzymes for molecular biology experiments, enzymes to degrade toxins and biofilms, and enzymes to remove antigens from transplanted cells.

Application: Therapeutic proteins

Design of better protein therapeutics will probably be the most significant commercial application of computational protein design. Therapeutic antibodies have been developed for a wide range of indications, from anti-cancer to anti-inflammatory. They currently have $5.1 billion in annual sales, and 30 antibodies are in late stage clinical trials.

Antibodies are the most widely used custom binding proteins, but they have several known limitations. Human and mouse antibodies are unable to bind to deep grooves,[128] and other targets have proven elusive as well.[138] Many antibodies are unstable or aggregation-prone.[139] Non-human antibodies are immunogenic in humans.

Therapeutic antibodies are glycosylated and thus more expensive to manufacture. Finally, the large size of whole antibodies may limit their tissue distribution.

Many of these problems can be addressed by moving to a small, stable scaffold protein that can be expressed in *E. coli*. Many such scaffolds have been proposed, including A domains, fibronectin, PDZ domains, ankyrin repeat proteins, and protein A.[140-142] Computational protein design, followed by experimental screening or selection experiments, could be used to engineer new binding sites in these scaffolds. The design framework described in this thesis can be used to select for affinity, stability, and specificity. Furthermore, computational techniques are available for predicting aggregation[143] and immunogenicity.[144,145]

References

1. Hegyi, H. & Gerstein, M. (1999). The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol.* **288**, 147-64.
2. Looger, L. L., Dwyer, M. A., Smith, J. J. & Hellinga, H. W. (2003). Computational design of receptor and sensor proteins with novel functions. *Nature* **423**, 185-90.
3. Dwyer, M. A., Looger, L. L. & Hellinga, H. W. (2004). Computational design of a biologically active enzyme. *Science* **304**, 1967-71.
4. Dwyer, M. A., Looger, L. L. & Hellinga, H. W. (2008). Retraction. *Science* **319**, 569.
5. Dahiyat, B. I. & Mayo, S. L. (1997). De novo protein design: fully automated sequence selection. *Science* **278**, 82-7.
6. Harbury, P. B., Plecs, J. J., Tidor, B., Alber, T. & Kim, P. S. (1998). High-resolution protein design with backbone freedom. *Science* **282**, 1462-7.
7. Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L. & Baker, D. (2003). Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364-8.
8. Dahiyat, B. I. & Mayo, S. L. (1997). De novo protein design: Fully automated sequence selection. *Science* **278**, 82-87.
9. Havranek, J. J. & Harbury, P. B. (2003). Automated design of specificity in molecular recognition. *Nat. Struct. Biol.* **10**, 45-52.
10. Lazar, G. A., Dang, W., Karki, S., Vafa, O., Peng, J. S., Hyun, L., Chan, C., Chung, H. S., Eivazi, A., Yoder, S. C., Vielmetter, J., Carmichael, D. F., Hayes, R. J. & Dahiyat, B. I. (2006). Engineered antibody Fc variants with enhanced effector function. *Proc. Natl. Acad. Sci. USA* **103**, 4005-10.
11. Steed, P. M., Tansey, M. G., Zalevsky, J., Zhukovsky, E. A., Desjarlais, J. R., Szymkowski, D. E., Abbott, C., Carmichael, D., Chan, C., Cherry, L., Cheung, P., Chirino, A. J., Chung, H. H., Doberstein, S. K., Eivazi, A., Filikov, A. V., Gao, S. X., Hubert, R. S., Hwang, M., Hyun, L., Kashi, S., Kim, A., Kim, E., Kung, J., Martinez, S. P., Muchhal, U. S., Nguyen, D. H., O'Brien, C., O'Keefe, D., Singer, K., Vafa, O., Vielmetter, J., Yoder, S. C. & Dahiyat, B. I. (2003). Inactivation of TNF signaling by rationally designed dominant-negative TNF variants. *Science* **301**, 1895-8.
12. Kortemme, T., Joachimiak, L. A., Bullock, A. N., Schuler, A. D., Stoddard, B. L. & Baker, D. (2004). Computational redesign of protein-protein interaction specificity. *Nat. Struct. Mol. Biol.* **11**, 371-9.
13. Jiang, L., Althoff, E. A., Clemente, F. R., Doyle, L., Rothlisberger, D., Zanghellini, A., Gallaher, J. L., Betker, J. L., Tanaka, F., Barbas, C. F., 3rd, Hilvert, D., Houk, K. N., Stoddard, B. L. & Baker, D. (2008). De novo computational design of retro-aldol enzymes. *Science* **319**, 1387-91.
14. MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D.,

Kuchnir, L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., Reiher, W. E., Roux, B., Schlenkrich, M., Smith, J. C., Stote, R., Straub, J., Watanabe, M., Wiorkiewicz-Kuczera, J., Yin, D. & Karplus, M. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **102**, 3586-3616.

15. Bashford, D. & Case, D. A. (2000). Generalized born models of macromolecular solvation effects. *Annu. Rev. Phys. Chem.* **51**, 129-152.

16. Lee, M. S., Salsbury, F. R. & Brooks, C. L. (2002). Novel generalized Born methods. *J. Chem. Phys.* **116**, 10606-10614.

17. Lovell, S. C., Word, J. M., Richardson, J. S. & Richardson, D. C. (2000). The penultimate rotamer library. *Proteins: Structure Function and Genetics* **40**, 389-408.

18. Katchalskikatzir, E., Shariv, I., Eisenstein, M., Friesem, A. A., Aflalo, C. & Vakser, I. A. (1992). Molecular-Surface Recognition - Determination of Geometric Fit between Proteins and Their Ligands by Correlation Techniques. *Proceedings of the National Academy of Sciences of the United States of America* **89**, 2195-2199.

19. Koehl, P. & Delarue, M. (1996). Mean-field minimization methods for biological macromolecules. *Curr. Opin. Struct. Biol.* **6**, 222-6.

20. Forrest, S. (1993). Genetic algorithms: principles of natural selection applied to computation. *Science* **261**, 872-8.

21. Kortemme, T., Joachimiak, L. A., Bullock, A. N., Schuler, A. D., Stoddard, B. L. & Baker, D. (2004). Computational redesign of protein-protein interaction specificity. *Nature structural & molecular biology* **11**, 371-9.

22. Kauzmann, W. (1959). Some factors in the interpretation of protein denaturation. *Adv. Protein Chem.* **14**, 1-63.

23. Mackerell, A. D., Jr. (2004). Empirical force fields for biological macromolecules: overview and issues. *J. Comput. Chem.* **25**, 1584-604.

24. Jorgensen, W. L. & Tirado-Rives, J. (2005). Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. *Proc. Natl. Acad. Sci. USA* **102**, 6665-70.

25. Gordon, D. B., Marshall, S. A. & Mayo, S. L. (1999). Energy functions for protein design. *Curr. Opin. Struct. Biol.* **9**, 509-13.

26. Pokala, N. & Handel, T. M. (2001). Review: protein design--where we were, where we are, where we're going. *J Struct Biol* **134**, 269-81.

27. Lazaridis, T. & Karplus, M. (2000). Effective energy functions for protein structure prediction. *Curr. Opin. Struct. Biol.* **10**, 139-45.

28. Mohanty, D., Dominy, B. N., Kolinski, A., Brooks, C. L., 3rd & Skolnick, J. (1999). Correlation between knowledge-based and detailed atomic potentials: application to the unfolding of the GCN4 leucine zipper. *Proteins* **35**, 447-52.

29. Ben-Naim, A. (1997). Statistical potentials extracted from protein structures: Are these meaningful potentials? *J. Chem. Phys.* **107**, 3698-3706.

30. Simons, K. T., Kooperberg, C., Huang, E. & Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using

simulated annealing and Bayesian scoring functions. *Journal of molecular biology* **268**, 209-25.

31. Dehouck, Y., Gilis, D. & Rooman, M. (2006). A new generation of statistical potentials for proteins. *Biophys J.* **90**, 4010-7.
32. Kortemme, T. & Baker, D. (2002). A simple physical model for binding energy hot spots in protein-protein complexes. *Proc. Natl. Acad. Sci. USA* **99**, 14116-21.
33. Dahiyat, B. I., Gordon, D. B. & Mayo, S. L. (1997). Automated design of the surface positions of protein helices. *Protein Sci* **6**, 1333-7.
34. Looger, L. L. & Hellinga, H. W. (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**, 429-45.
35. Wisz, M. S. & Hellinga, H. W. (2003). An empirical model for electrostatic interactions in proteins incorporating multiple geometry-dependent dielectric constants. *Proteins* **51**, 360-77.
36. Lazaridis, T. & Karplus, M. (1999). Effective energy function for proteins in solution. *Proteins* **35**, 133-52.
37. Kortemme, T., Morozov, A. V. & Baker, D. (2003). An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *J. Mol. Biol.* **326**, 1239-59.
38. Dahiyat, B. I. & Mayo, S. L. (1997). Probing the role of packing specificity in protein design. *Proc. Natl. Acad. Sci. USA* **94**, 10172-7.
39. Chothia, C. (1974). Hydrophobic bonding and accessible surface area in proteins. *Nature* **248**, 338-9.
40. Street, A. G. & Mayo, S. L. (1998). Pairwise calculation of protein solvent-accessible surface areas. *Fold. Des.* **3**, 253-8.
41. Choudhury, N. & Pettitt, B. M. (2005). On the mechanism of hydrophobic association of nanoscopic solutes. *J. Am. Chem. Soc.* **127**, 3556-67.
42. Wagoner, J. A. & Baker, N. A. (2006). Assessing implicit models for nonpolar mean solvation forces: the importance of dispersion and volume terms. *Proc. Natl. Acad. Sci. USA* **103**, 8331-6.
43. Eisenberg, D. & McLachlan, A. D. (1986). Solvation energy in protein folding and binding. *Nature* **319**, 199-203.
44. Honig, B., Sharp, K. & Yang, A. S. (1993). Macroscopic Models of Aqueous Solutions: Biological and Chemical Applications. *J. Phys. Chem.* **97**, 1101-1109.
45. Marshall, S. A., Vizcarra, C. L. & Mayo, S. L. (2005). One- and two-body decomposable Poisson-Boltzmann methods for protein design calculations. *Protein Sci.* **14**, 1293-304.
46. Pokala, N. & Handel, T. M. (2005). Energy functions for protein design: adjustment with protein-protein complex affinities, models for the unfolded state, and negative design of solubility and specificity. *J. Mol. Biol.* **347**, 203-27.

47. Lee, M. S., Feig, M., Salsbury, F. R., Jr. & Brooks, C. L., 3rd. (2003). New analytic approximation to the standard molecular volume definition and its application to generalized Born calculations. *J. Comput. Chem.* **24**, 1348-56.

48. Yu, Z., Jacobson, M. P. & Friesner, R. A. (2006). What role do surfaces play in GB models? A new-generation of surface-generalized born model based on a novel gaussian surface for biomolecules. *J. Comput. Chem.* **27**, 72-89.

49. Feig, M., Onufriev, A., Lee, M. S., Im, W., Case, D. A. & Brooks, C. L., 3rd. (2004). Performance comparison of generalized born and Poisson methods in the calculation of electrostatic solvation energies for protein structures. *J. Comput. Chem.* **25**, 265-84.

50. Schymkowitz, J. W., Rousseau, F., Martins, I. C., Ferkinghoff-Borg, J., Stricher, F. & Serrano, L. (2005). Prediction of water and metal binding sites and their affinities by using the Fold-X force field. *Proc. Natl. Acad. Sci. USA* **102**, 10147-52.

51. Jiang, L., Kuhlman, B., Kortemme, T. & Baker, D. (2005). A "solvated rotamer" approach to modeling water-mediated hydrogen bonds at protein-protein interfaces. *Proteins* **58**, 893-904.

52. Morozov, A. V., Kortemme, T., Tsemekhman, K. & Baker, D. (2004). Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 6946-51.

53. Morozov, A. V. & Kortemme, T. (2005). Potential functions for hydrogen bonds in protein structure prediction and design. *Adv. Protein. Chem.* **72**, 1-38.

54. Friesner, R. A. (2006). Modeling polarization in proteins and protein-ligand complexes: Methods and preliminary results. *Adv. Protein Chem.* **72**, 79-104.

55. Maple, J. R., Cao, Y. X., Damm, W. G., Halgren, T. A., Kaminski, G. A., Zhang, L. Y. & Friesner, R. A. (2005). A polarizable force field and continuum solvation methodology for modeling of protein-ligand interactions. *Journal of Chemical Theory and Computation* **1**, 694-715.

56. Friesner, R. A. (2005). Ab initio quantum chemistry: methodology and applications. *Proc. Natl. Acad. Sci. USA* **102**, 6648-53.

57. Cho, A. E., Guallar, V., Berne, B. J. & Friesner, R. (2005). Importance of accurate charges in molecular docking: quantum mechanical/molecular mechanical (QM/MM) approach. *J. Comput. Chem.* **26**, 915-31.

58. Zhou, H. X. (2003). Direct test of the Gaussian-chain model for treating residual charge-charge interactions in the unfolded state of proteins. *J. Am. Chem. Soc.* **125**, 2060-2061.

59. Waldburger, C. D., Schildbach, J. F. & Sauer, R. T. (1995). Are buried salt bridges important for protein stability and conformational specificity? *Nat. Struct. Biol.* **2**, 122-8.

60. Clark, L. A., Boriack-Sjodin, P. A., Eldredge, J., Fitch, C., Friedman, B., Hanf, K. J., Jarpe, M., Liparoto, S. F., Li, Y., Lugovskoy, A., Miller, S., Rushe, M., Sherman, W., Simon, K. & Van Vlijmen, H. (2006). Affinity enhancement of an in vivo matured therapeutic antibody using structure-based computational design. *Protein Sci.* **15**, 949-60.

61. Shifman, J. M. & Mayo, S. L. (2003). Exploring the origins of binding specificity through the computational redesign of calmodulin. *Proc. Natl. Acad. Sci. USA* **100**, 13274-9.

62. Boas, F. E. & Harbury, P. B. (2008). Design of protein-ligand binding based on the molecular-mechanics energy model. *J. Mol. Biol.* **In press.**

63. Kuhlman, B. & Baker, D. (2000). Native protein sequences are close to optimal for their structures. *Proc. Natl. Acad. Sci. USA* **97**, 10383-8.

64. Ashworth, J., Havranek, J. J., Duarte, C. M., Sussman, D., Monnat, R. J., Jr., Stoddard, B. L. & Baker, D. (2006). Computational redesign of endonuclease DNA binding and cleavage specificity. *Nature* **441**, 656-9.

65. Ambroggio, X. I. & Kuhlman, B. (2006). Computational design of a single amino acid sequence that can switch between two distinct protein folds. *J. Am. Chem. Soc.* **128**, 1154-61.

66. Saunders, C. T. & Baker, D. (2005). Recapitulation of protein family divergence using flexible backbone protein design. *J. Mol. Biol.* **346**, 631-44.

67. Snow, C. D., Sorin, E. J., Rhee, Y. M. & Pande, V. S. (2005). How well can simulation predict protein folding kinetics and thermodynamics? *Annu. Rev. Biophys. Biomol. Struct.* **34**, 43-69.

68. Ren, P. Y. & Ponder, J. W. (2002). Consistent treatment of inter- and intramolecular polarization in molecular mechanics calculations. *J. Comput. Chem.* **23**, 1497-1506.

69. Kuhn, B. & Kollman, P. A. (2000). Binding of a diverse set of ligands to avidin and streptavidin: an accurate quantitative prediction of their relative affinities by a combination of molecular mechanics and continuum solvent models. *J Med Chem* **43**, 3786-91.

70. Huo, S., Massova, I. & Kollman, P. A. (2002). Computational alanine scanning of the 1:1 human growth hormone-receptor complex. *J Comput Chem* **23**, 15-27.

71. Mobley, D. L., Graves, A. P., Chodera, J. D., McReynolds, A. C., Shoichet, B. K. & Dill, K. A. (2007). Predicting absolute ligand binding free energies to a simple model site. *J Mol Biol* **371**, 1118-34.

72. Wang, J., Kang, X., Kuntz, I. D. & Kollman, P. A. (2005). Hierarchical database screenings for HIV-1 reverse transcriptase using a pharmacophore model, rigid docking, solvation docking, and MM-PB/SA. *J Med Chem* **48**, 2432-44.

73. Kollman, P. A., Massova, I., Reyes, C., Kuhn, B., Huo, S., Chong, L., Lee, M., Lee, T., Duan, Y., Wang, W., Donini, O., Cieplak, P., Srinivasan, J., Case, D. A. & Cheatham, T. E., 3rd. (2000). Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc Chem Res* **33**, 889-97.

74. Snow, C. D., Nguyen, H., Pande, V. S. & Gruebele, M. (2002). Absolute comparison of simulated and experimental protein-folding dynamics. *Nature* **420**, 102-6.

75. Barth, P., Alber, T. & Harbury, P. B. (2007). Accurate, conformation-dependent predictions of solvent effects on protein ionization constants. *Proc Natl Acad Sci U S A* **104**, 4898-903.
76. Chakrabarti, R., Klibanov, A. M. & Friesner, R. A. (2005). Computational prediction of native protein ligand-binding and enzyme active site sequences. *Proc Natl Acad Sci U S A* **102**, 10153-8.
77. Chakrabarti, R., Klibanov, A. M. & Friesner, R. A. (2005). Sequence optimization and designability of enzyme active sites. *Proc Natl Acad Sci U S A* **102**, 12035-40.
78. Boas, F. E. & Harbury, P. B. (2007). Potential energy functions for protein design. *Curr. Opin. Struct. Biol.* **17**, 199-204.
79. Bolon, D. N. & Mayo, S. L. (2001). Enzyme-like proteins by computational design. *Proc. Natl. Acad. Sci. USA* **98**, 14274-9.
80. Zanghellini, A., Jiang, L., Wollacott, A. M., Cheng, G., Meiler, J., Althoff, E. A., Rothlisberger, D. & Baker, D. (2006). New algorithms and an in silico benchmark for computational enzyme design. *Protein Sci* **15**, 2785-94.
81. Friesner, R. A., Murphy, R. B., Repasky, M. P., Frye, L. L., Greenwood, J. R., Halgren, T. A., Sanschagrin, P. C. & Mainz, D. T. (2006). Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J Med Chem* **49**, 6177-96.
82. Jain, A. N. (2006). Scoring functions for protein-ligand docking. *Curr Protein Pept Sci* **7**, 407-20.
83. Lassila, J. K., Privett, H. K., Allen, B. D. & Mayo, S. L. (2006). Combinatorial methods for small-molecule placement in computational enzyme design. *Proc Natl Acad Sci U S A* **103**, 16710-5.
84. Quiocho, F. A. & Ledvina, P. S. (1996). Atomic structure and specificity of bacterial periplasmic receptors for active transport and chemotaxis: variation of common themes. *Mol. Microbiol.* **20**, 17-25.
85. Gilson, M. K. & Zhou, H. X. (2007). Calculation of protein-ligand binding affinities. *Annu Rev Biophys Biomol Struct* **36**, 21-42.
86. Declerck, N. & Abelson, J. (1994). Novel substrate specificity engineered in the arabinose binding protein. *Protein Eng.* **7**, 997-1004.
87. Vermersch, P. S., Lemon, D. D., Tesmer, J. J. & Quiocho, F. A. (1991). Sugar-binding and crystallographic studies of an arabinose-binding protein mutant (Met108Leu) that exhibits enhanced affinity and altered specificity. *Biochemistry* **30**, 6861-6.
88. Li, Y., Li, H., Yang, F., Smith-Gill, S. J. & Mariuzza, R. A. (2003). X-ray snapshots of the maturation of an antibody response to a protein antigen. *Nat. Struct. Biol.* **10**, 482-8.
89. Fersht, A. (1999). *Structure and mechanism in protein science: A guide to enzyme catalysis and protein folding*, W.H. Freeman and Company, New York.
90. Word, J. M., Richardson, J. S. & Richardson, D. C. (2000). *bndlst version 1.6 (http://kinemage.biochem.duke.edu/).*

91. Lawrence, M. C. & Colman, P. M. (1993). Shape complementarity at protein/protein interfaces. *J. Mol. Biol.* **234**, 946-50.

92. Qiu, D., Shenkin, P. S., Hollinger, F. P. & Still, W. C. (1997). The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii. *J. Phys. Chem. A* **101**, 3005-3014.

93. Monard, G. & Merz, K. M. (1999). Combined quantum mechanical/molecular mechanical methodologies applied to biomolecular systems. *Accounts Chem. Res.* **32**, 904-911.

94. Liu, H., Elstner, M., Kaxiras, E., Frauenheim, T., Hermans, J. & Yang, W. (2001). Quantum mechanics simulation of protein dynamics on long timescale. *Proteins* **44**, 484-9.

95. Still, W. C., Tempczyk, A., Hawley, R. C. & Hendrickson, T. (1990). Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics. *J. Am. Chem. Soc.* **112**, 6127-6129.

96. Lim, C., Bashford, D. & Karplus, M. (1991). Absolute pKa Calculations with Continuum Dielectric Methods. *J. Phys. Chem.* **95**, 5610-5620.

97. Nocedal, J. & Wright, S. J. (1999). *Numerical Optimization*. Springer series in operations research, Springer, New York.

98. Ponder, J. W. (2004). *TINKER version 4.2 (http://dasher.wustl.edu/tinker/)*.

99. Slovic, A. M., Kono, H., Lear, J. D., Saven, J. G. & DeGrado, W. F. (2004). Computational design of water-soluble analogues of the potassium channel KcsA. *Proc. Natl. Acad. Sci. USA* **101**, 1828-33.

100. Brunger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallographica Section D-Biological Crystallography* **54**, 905-921.

101. Kunkel, T. A., Bebenek, K. & McClary, J. (1991). Efficient site-directed mutagenesis using uracil-containing DNA. *Methods Enzymol.* **204**, 125-39.

102. Pace, C. N., Vajdos, F., Fee, L., Grimsley, G. & Gray, T. (1995). How to measure and predict the molar absorption coefficient of a protein. *Protein Sci.* **4**, 2411-23.

103. Leach, A. R. (2001). *Molecular Modelling : Pinciples and Applications*. 2nd edit, Prentice Hall, Harlow, England ; New York.

104. Srinivasan, J., Trevathan, M. W., Beroza, P. & Case, D. A. (1999). Application of a pairwise generalized Born model to proteins and nucleic acids: inclusion of salt effects. *Theoretical Chemistry Accounts* **101**, 426-434.

105. Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1996). *Numerical Recipes in C: The Art of Scientific Computing*. 2nd edit, Cambridge University Press, Cambridge; New York.

106. Lide, D. R. (2000). *CRC Handbook of Chemistry and Physics*. 81st edit, CRC Press, Boca Raton; New York; Washington D.C.

107. Creighton, T. E. (1993). *Proteins: Structures and Molecular Properties*. 2nd edit, W.H. Freeman and Company, New York.

108. Word, J. M., Lovell, S. C., Richardson, J. S. & Richardson, D. C. (1999). Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.* **285**, 1735-47.
109. McDonald, I. K. & Thornton, J. M. (1994). Satisfying hydrogen bonding potential in proteins. *J Mol Biol* **238**, 777-93.
110. Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A. A., Aflalo, C. & Vakser, I. A. (1992). Molecular-Surface Recognition: Determination of Geometric Fit between Proteins and Their Ligands by Correlation Techniques. *Proc. Natl. Acad. Sci. USA* **89**, 2195-2199.
111. Desmet, J., Demaeyer, M., Hazes, B. & Lasters, I. (1992). The Dead-End Elimination Theorem and Its Use in Protein Side-Chain Positioning. *Nature* **356**, 539-542.
112. Desmet, J., Spriet, J. & Lasters, I. (2002). Fast and accurate side-chain topology and energy refinement (FASTER) as a new method for protein structure optimization. *Proteins* **48**, 31-43.
113. Hu, X. & Kuhlman, B. (2006). Protein design simulations suggest that side-chain conformational entropy is not a strong determinant of amino acid environmental preferences. *Proteins* **62**, 739-48.
114. Lin, H. & Cornish, V. W. (2002). Screening and selection methods for large-scale analysis of protein function. *Angew. Chem. Int. Ed. Engl.* **41**, 4402-25.
115. Mowbray, S. L. & Cole, L. B. (1992). 1.7 A X-ray structure of the periplasmic ribose receptor from Escherichia coli. *J. Mol. Biol.* **225**, 155-75.
116. Bjorkman, A. J. & Mowbray, S. L. (1998). Multiple open forms of ribose-binding protein trace the path of its conformational change. *J. Mol. Biol.* **279**, 651-64.
117. Berthod, H., Giessner*, C. & Pullman, A. (1967). Sur les roles respectifs des electrons sigma et pi dans les proprietes des derives halogenes des molecules conjuguees: Application a letude de luracile et du fluorouracile. *Theoretica Chimica Acta* **8**, 212-&.
118. Matthew Clark, R. D. C. I. I. I. N. V. O. (1989). Validation of the general purpose tripos 5.2 force field. *Journal of Computational Chemistry* **10**, 982-1012.
119. Loudon, G. M. (1995). *Organic Chemistry*. 3rd edit.
120. Matsuo, K. & Gekko, K. (2004). Vacuum-ultraviolet circular dichroism study of saccharides by synchrotron radiation spectrophotometry. *Carbohydr Res* **339**, 591-7.
121. Kyte, J. & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105-32.
122. Hunter, C. A., Singh, J. & Thornton, J. M. (1991). Pi-pi interactions: the geometry and energetics of phenylalanine-phenylalanine interactions in proteins. *J. Mol. Biol.* **218**, 837-46.
123. Cadwell, R. C. & Joyce, G. F. (1992). Randomization of genes by PCR mutagenesis. *PCR methods and applications* **2**, 28-33.

124. Zhao, H., Giver, L., Shao, Z., Affholter, J. A. & Arnold, F. H. (1998). Molecular evolution by staggered extension process (StEP) in vitro recombination. *Nature biotechnology* **16**, 258-61.

125. Cadwell, R. C. & Joyce, G. F. (1992). Randomization of genes by PCR mutagenesis. *PCR Methods Appl.* **2**, 28-33.

126. Morozov, A. V., Kortemme, T., Tsemekhman, K. & Baker, D. (2004). Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations. *Proc. Natl. Acad. Sci. USA* **101**, 6946-51.

127. Lockless, S. W. & Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* **286**, 295-9.

128. Desmyter, A., Transue, T. R., Ghahroudi, M. A., Thi, M. H., Poortmans, F., Hamers, R., Muyldermans, S. & Wyns, L. (1996). Crystal structure of a camel single-domain VH antibody fragment in complex with lysozyme. *Nat. Struct. Biol.* **3**, 803-11.

129. Fersht, A. (1999). *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*, W.H. Freeman, New York.

130. Morii, T., Sugimoto, K., Makino, K., Otsuka, M., Imoto, K. & Mori, Y. (2002). A new fluorescent biosensor for inositol trisphosphate. *J. Am. Chem. Soc.* **124**, 1138-9.

131. Fehr, M., Frommer, W. B. & Lalonde, S. (2002). Visualization of maltose uptake in living yeast cells by fluorescent nanosensors. *Proc. Natl. Acad. Sci. USA* **99**, 9846-51.

132. Miyawaki, A. & Tsien, R. Y. (2000). Monitoring protein conformations and interactions by fluorescence resonance energy transfer between mutants of green fluorescent protein. *Methods Enzymol.* **327**, 472-500.

133. Muddana, S. S. & Peterson, B. R. (2003). Fluorescent cellular sensors of steroid receptor ligands. *Chembiochem* **4**, 848-55.

134. Bacskai, B. J., Hochner, B., Mahaut-Smith, M., Adams, S. R., Kaang, B. K., Kandel, E. R. & Tsien, R. Y. (1993). Spatially resolved dynamics of cAMP and protein kinase A subunits in Aplysia sensory neurons. *Science* **260**, 222-6.

135. Villaverde, A. (2003). Allosteric enzymes as biosensors for molecular diagnosis. *FEBS Letters* **554**, 169-172.

136. Ha, J. H., Butler, J. S., Mitrea, D. M. & Loh, S. N. (2006). Modular enzyme design: Regulation by mutually exclusive protein folding. *Journal of Molecular Biology* **357**, 1058-1062.

137. Carter, P. & Wells, J. A. (1988). Dissecting the catalytic triad of a serine protease. *Nature* **332**, 564-8.

138. Kwong, P. D., Doyle, M. L., Casper, D. J., Cicala, C., Leavitt, S. A., Majeed, S., Steenbeke, T. D., Venturi, M., Chaiken, I., Fung, M., Katinger, H., Parren, P. W., Robinson, J., Van Ryk, D., Wang, L., Burton, D. R., Freire, E., Wyatt, R., Sodroski, J., Hendrickson, W. A. & Arthos, J. (2002). HIV-1 evades antibody-mediated neutralization through conformational masking of receptor-binding sites. *Nature* **420**, 678-82.

139. Jespers, L., Schon, O., Famm, K. & Winter, G. (2004). Aggregation-resistant domain antibodies selected on phage by heat denaturation. *Nat. Biotechnol.* **22**, 1161-5.

140. Silverman, J., Liu, Q., Bakker, A., To, W., Duguay, A., Alba, B. M., Smith, R., Rivas, A., Li, P., Le, H., Whitehorn, E., Moore, K. W., Swimmer, C., Perlroth, V., Vogt, M., Kolkman, J. & Stemmer, W. P. (2005). Multivalent avimer proteins evolved by exon shuffling of a family of human receptor domains. *Nat. Biotechnol.* **23**, 1556-61.

141. Binz, H. K., Amstutz, P. & Pluckthun, A. (2005). Engineering novel binding proteins from nonimmunoglobulin domains. *Nat. Biotechnol.* **23**, 1257-68.

142. Binz, H. K. & Pluckthun, A. (2005). Engineered proteins as specific binding reagents. *Curr. Opin. Biotechnol.* **16**, 459-69.

143. Fernandez-Escamilla, A. M., Rousseau, F., Schymkowitz, J. & Serrano, L. (2004). Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol.* **22**, 1302-6.

144. Brusic, V., Rudy, G. & Harrison, L. C. (1998). MHCPEP, a database of MHC-binding peptides: update 1997. *Nucleic Acids Res.* **26**, 368-71.

145. Kolaskar, A. S. & Tongaonkar, P. C. (1990). A semi-empirical method for prediction of antigenic determinants on protein antigens. *FEBS Lett.* **276**, 172-4.