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This chapter reports on unpublished work.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.

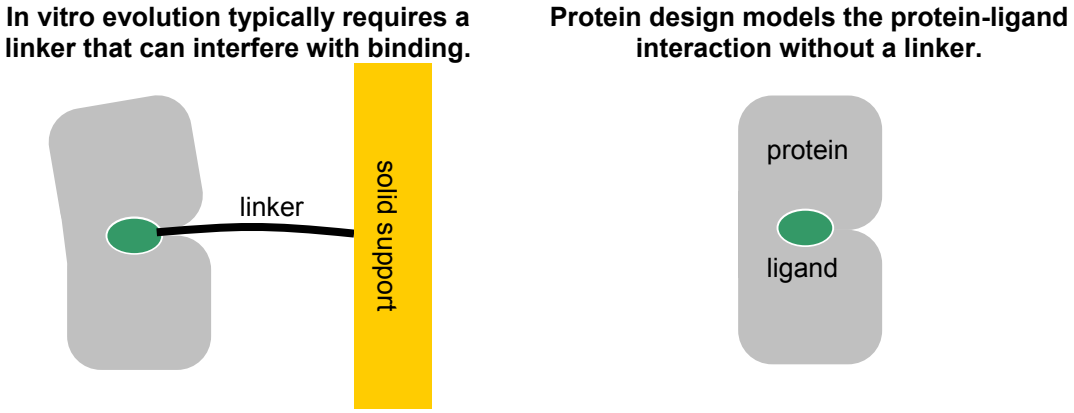

Figure 25. In vitro evolution vs computational protein design.

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).

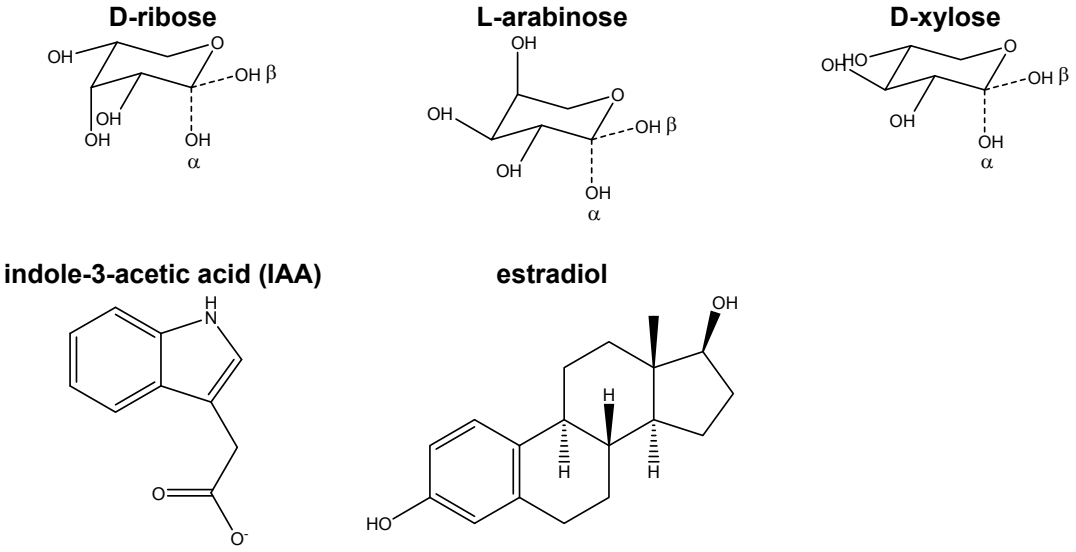

Figure 26. Target ligands.

	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.

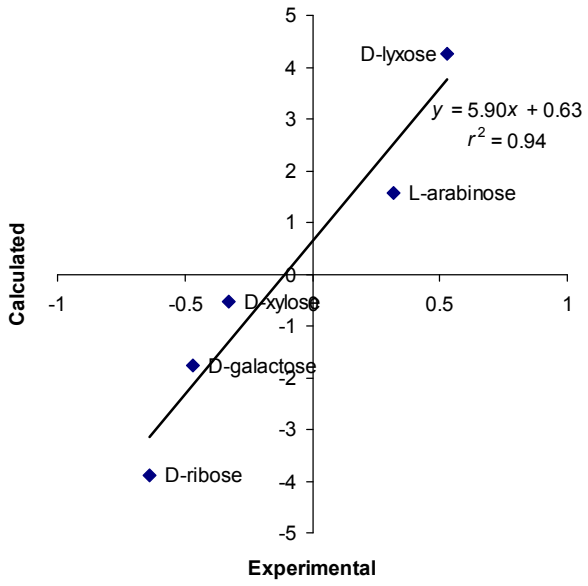

Figure 27. Experimental[119,120] and calculated β-pyranose energy – α-pyranose energy (kcal/mol).

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.

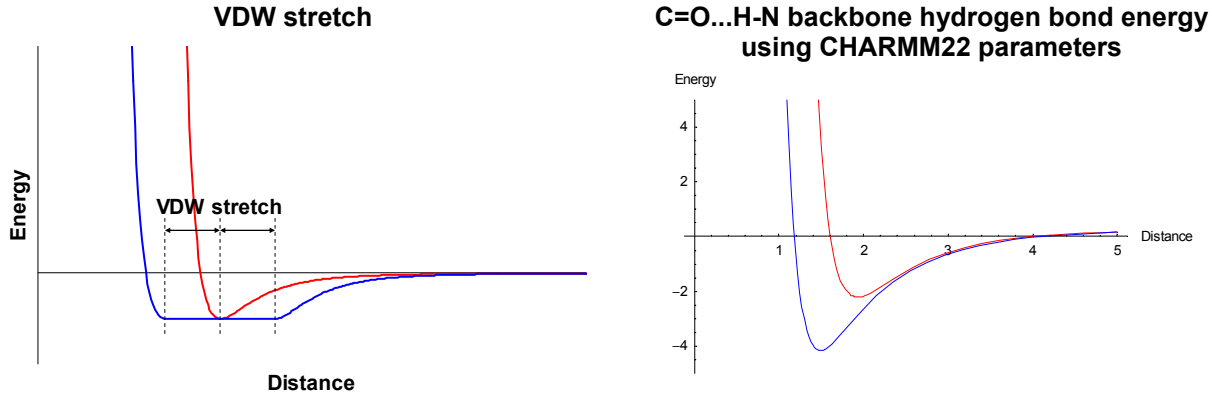

Figure 28. The Lennard-Jones potential is frequently softened in design calculations to compensate for low sampling resolution. However, this has the side effect of making hydrogen bonds appear artificially strong. The figure shows the energy of a C=O...H-N backbone hydrogen bond energy (Lennard-Jones 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.

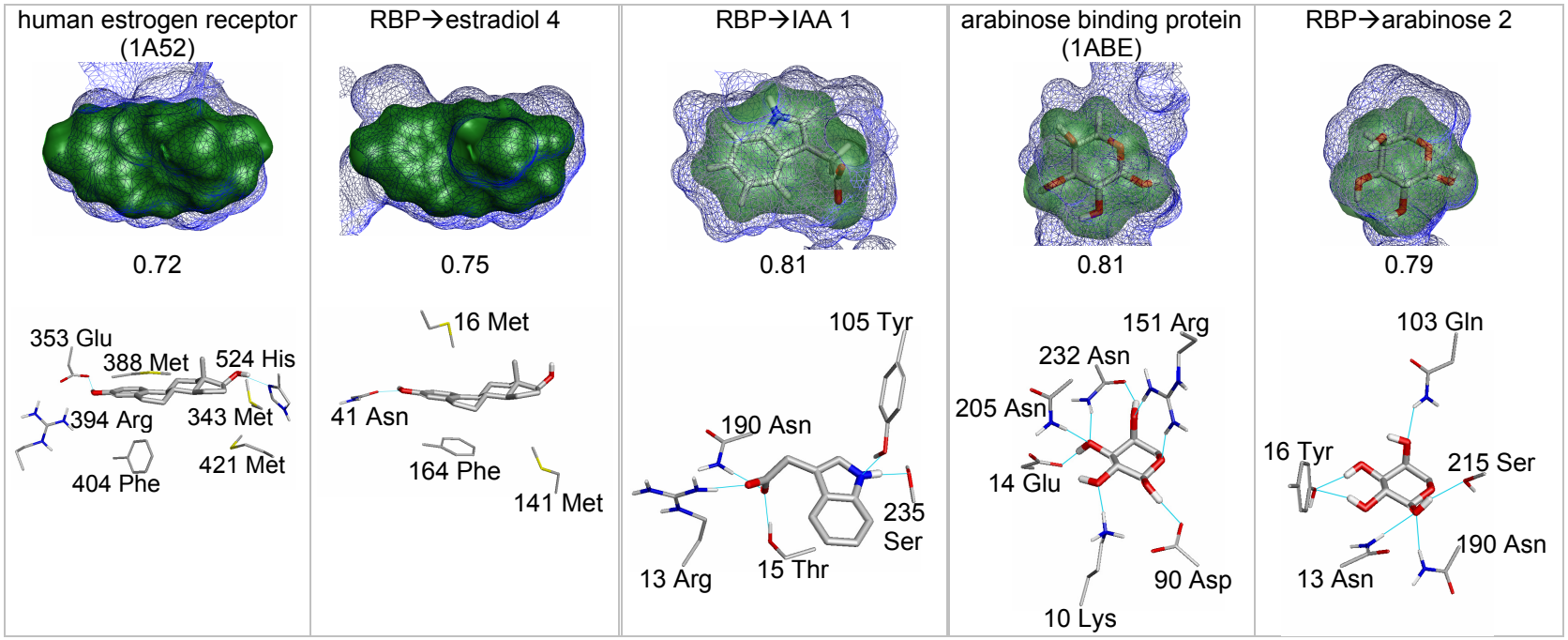

Figure 29. Structures of designed and natural binding proteins.

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.