Design of Protein-Ligand Binding Based on the Molecular-Mechanics Energy Model

Supplementary information

F. Edward Boas, Pehr B. Harbury

Potential energy function

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

$$= \sum_{\text{bonds}} k_b(b-b_0)^2 + \sum_{\text{angles}} (k_{UB}(S-S_0)^2 + k_\theta(\theta-\theta_0)^2) + \sum_{\text{dihedrals}} k_\chi(1+\cos(n\chi-\delta)) + \sum_{\text{impropers}} k_\phi(\phi-\phi_0)^2$$

$$+ \sum_{\substack{\text{nonbonded} \\ i<j}} E_{VDW}\left(\left(\frac{r_{min}}{r}\right)^{12} - 2\left(\frac{r_{min}}{r}\right)^{6}\right) + 332 * \sum_{\substack{\text{nonbonded} \\ i<j}} \frac{q_i q_j}{\varepsilon_{in} r} + 332 * \sum_{i,j}^{N} \text{GB}(q_i, q_j, a_i, a_j, r, \varepsilon_{in}, \varepsilon_{out}, \kappa)$$

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

Variables

k_b	spring constant for bond length	E_{VDW}	van der Waals energy
b	bond length	r	inter-atom distance
b_0	equilibrium bond length	r_{min}	minimum-energy inter-atom distance
k_{UB}	Urey-Bradley constant for atoms separated by two bonds	q_i, q_j	charge on atoms i and j
		ε_{in}	protein and ligand dielectric constant = 1.0
S	distance between atoms separated by two bonds	ε_{out}	water dielectric constant = 78.4
S_0	equilibrium distance	GB()	generalized-Born solvation energy
k_θ	spring constant for bond angle	a_i, a_j	generalized-Born radii of atoms i and j
θ	bond angle	κ	inverse Debye-Hückel length (salt screening length)
θ_0	equilibrium bond angle	k_{SASA}	microscopic surface tension of water[1] = 0.0072 kcal/mol/Å^2
k_χ, n, δ	Fourier series terms for periodic barrier to rotation around bonds	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
χ	torsion angle		
k_φ	spring constant for torsion angle to restrain planar groups	$U_{deprot.}$	deprotonation energy (from a thermodynamic cycle based on the pK_A's of free amino acids)
φ	torsion angle		
φ_0	equilibrium torsion angle		

All parameters[2] 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.

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.

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,[3] which provides an approximate solution to the Poisson-Boltzmann differential equation.[4]

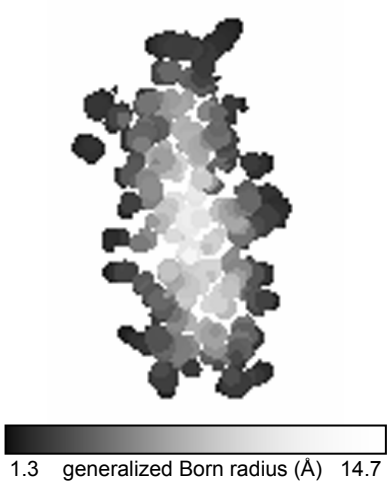
The generalized-Born approach requires the calculation of generalized-Born radii for each atom (Supplementary figure 1). The manuscript compares two numerical approaches for obtaining the radii. In the first approach, generalized-Born radii are computed on the basis of an r^{-4}-weighted spatial integral (Supplementary figure 2):

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

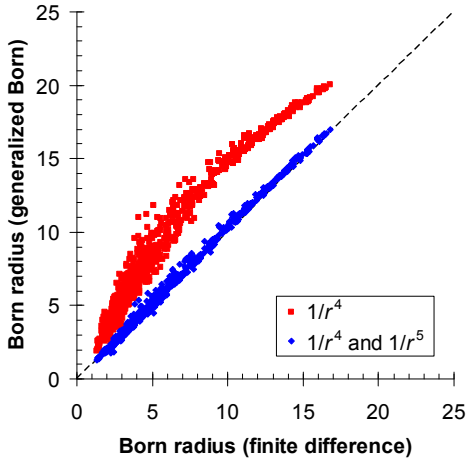
Here r is the distance from the atom center to each volume element in the integrand. Alternatively, more accurate radii are obtained from an empirical sum of r^{-4}- and r^{-5}-weighted spatial integrals:[5]

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

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.



Supplementary figure 1. Slice through ribose binding protein, showing generalized Born radii. The radii correlate with atom burial.



Supplementary figure 2. 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 [5].

Given generalized-Born radii, the polarization energy was evaluated as:[6]

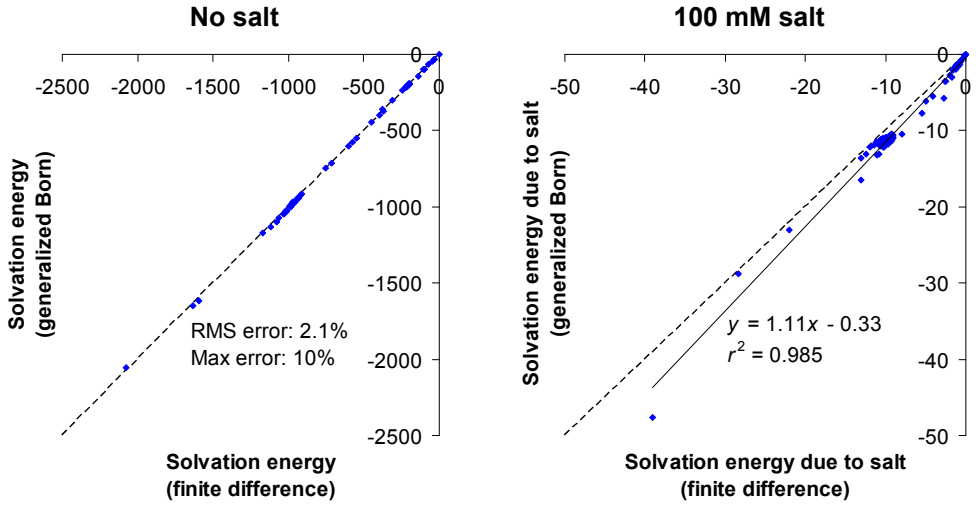
$$\sum_{i,j} \mathrm{GB}(q_i, q_j, a_i, a_j, r, \varepsilon_{in}, \varepsilon_{out}, \kappa) = -\frac{1}{2}\sum_{i,j}\left(\frac{1}{\varepsilon_{in}} - \frac{e^{-\kappa\sqrt{r_{ij}^2 + a_i a_j e^{-r_{ij}^2/(4a_i a_j)}}}}{\varepsilon_{out}}\right)\frac{q_i q_j}{\sqrt{r_{ij}^2 + a_i a_j e^{-r_{ij}^2/(4a_i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

with $\kappa = \frac{\sqrt{I/\varepsilon_{out}}}{0.343}$ 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 (Supplementary figure 3).



Supplementary figure 3. 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,[7] 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[8] 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) =

$$C + \sum_{\substack{i \in \text{variable} \\ \text{position}}} \left(t_i A_i - s_i \sum_{\substack{j \in \text{fixed} \\ \text{position}}} A_{i,j} - \sum_{\substack{j \in \text{fixed} \\ \text{position}}} s_j A_{j,i} \right) - \sum_{\substack{i \in \text{variable} \\ \text{position}}} \sum_{\substack{j \in \text{variable} \\ \text{position}, j<i}} \left(s_i A_{i,j} + s_j A_{j,i} \right)$$

C — additive constant

$t_i A_i$ — SASA of rotamers and ligand poses less the

$s_i \sum_{j \in \text{fixed position}} A_{i,j}$ — pairwise area buried at interfaces with fixed

$\sum_{j \in \text{fixed position}} s_j A_{j,i}$ — structural elements

$s_i A_{i,j} + s_j A_{j,i}$ — 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 [1]. The "hydrophobic effect" driving aggregation of hydrophobic solutes in water increases in proportion to solute surface area with a slope[9] 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 (Supplementary figure 4). For example, the deprotonation energy for an aspartate residue within a protein (labeled A in Supplementary figure 4) 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(solution) = $-RT$ ln(partition sum).

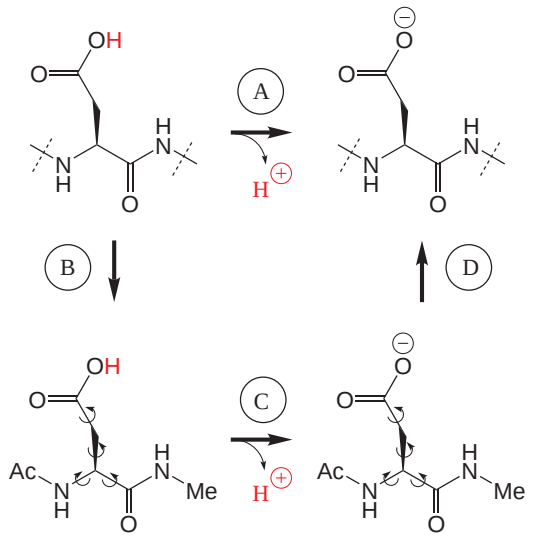
Then:

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

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, $-U(\text{Asp}^-, \text{solution}) + U(\text{AspH, solution}) - 2.3RT*(\text{pH} - \text{pK}_a)$, 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 (Supplementary table 1). The deprotonation energy is pH dependent, and all of the calculations reported here were performed at pH 7.0.



Supplementary figure 4. 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

Supplementary table 1. Deprotonation energies for the titratable amino acids in the 6028-member rotamer library (1.36 = RT ln 10 at T = 25°C). Experimental pKa values for free amino acids are from [10,11]. We did not include protonation states for CYS, TYR, LYS, or ARG because of a lack of published CHARMM22 parameters for those amino acids.

Discrete sampling

Protein scaffold coordinates

Hydrogen coordinates were added to scaffold crystal structures using Reduce.[12]

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[13] were selected as design positions. The resulting positions are listed in the caption to Figure 3 in the main paper.

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 4 in the main paper.

Rotamer library

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			

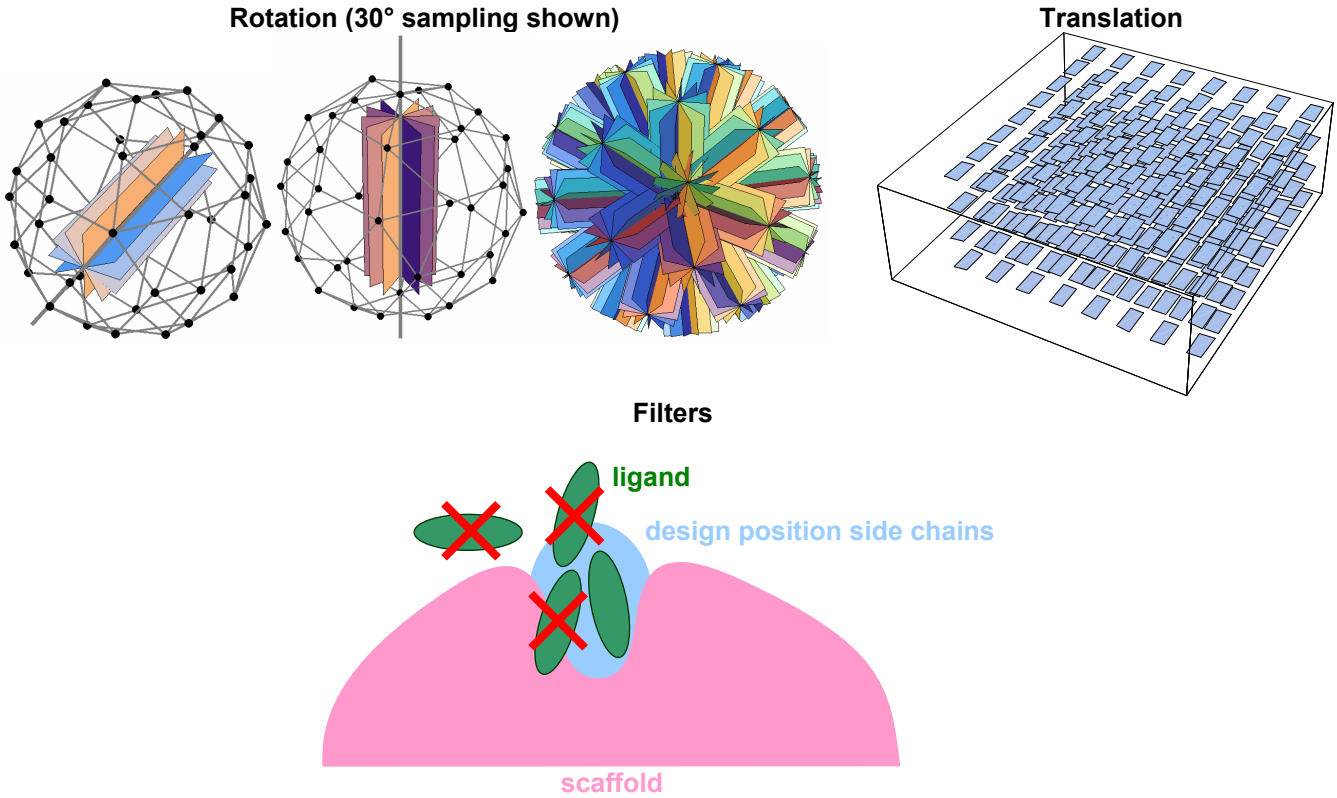
Supplementary table 2. The highest resolution rotamer library with 6028 rotamers. APP = protonated Asp, GUP = protonated Glu, HSP = doubly protonated His, HSD = His protonated on the delta nitrogen, HSE = His protonated on the epsilon nitrogen, CYX = disulfide-bonded cysteine.

A detailed rotamer library (including polar and non-polar hydrogens) was created by clustering the side chain conformations seen in high-resolution crystal structures (Supplementary table 2). 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[12] from the Richardson lab. The side chain conformations for each amino acid were then clustered at the resolution listed in Supplementary table 2. 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.

Ligand poses



Supplementary figure 5. 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.[14] 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,[15] 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. (Supplementary figure 5). 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.

Searching conformational space

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.[16] Using a mean-field algorithm, the probabilities were then adjusted iteratively to minimize the free energy of the system.[17] New probabilities for all rotamers were first computed using the mean-field energy of each rotamer and the Boltzmann equation:

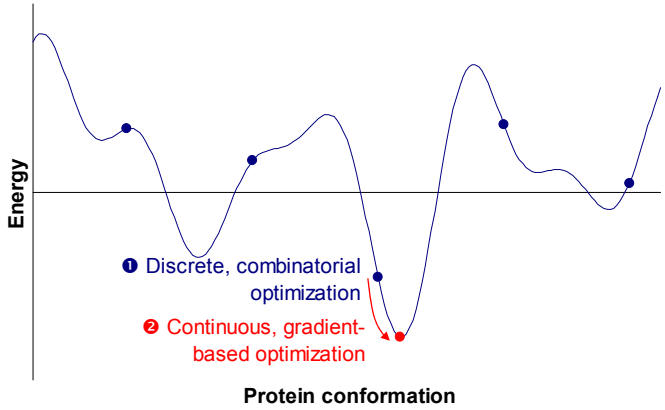
$$p_{new} = \frac{\exp(-(E_{self} + E_{interaction})/RT)}{Z}.$$

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 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 (Supplementary figure 6). 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.[18] However, it is important to note that we did not include entropy changes outside the binding site in our calculation.



Supplementary figure 6. 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 [19]). The energy of each configuration was calculated, and the intrinsic unfolded-state chemical potential (Supplementary table 3) was evaluated as RT ln(partition sum).

Inter-residue electrostatic interaction energies in the unfolded state were calculated following [20], assuming that the distance distribution between residues is determined by a random walk. The total unfolded-state energy was summed as:

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

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

Supplementary table 3. 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. 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 Supplementary table 4 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.

Supplementary table 4. Genetic algorithm parameters.

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:

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

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

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

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

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

Appendix: Supplementary tables and figures cited in the main paper

Stability data

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.

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[21]). We have not measured unfolding free energies for the remaining proteins.

ABP-arabinose (6ABP)

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

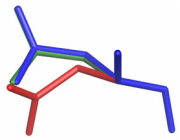
bevacizumab-VEGF (1BJ1, 2VPF)

	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

Supplementary table 5. Predicted protonation states.



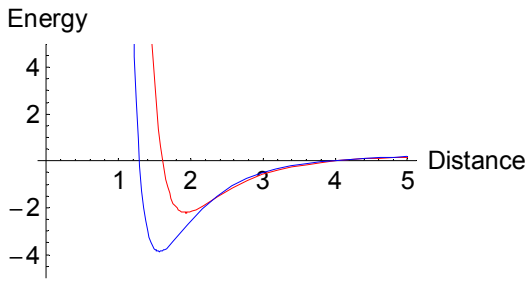
89 Asp

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

Supplementary figure 7. 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.

Experimental Dissoc. Energy (kcal/mol)	Source	Calculated Dissoc. Energy (kcal/mol)	Stability vs. Native (kcal/mol)	Sequence 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

Supplementary table 6. 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 [22]; 3. reference [23].



Supplementary figure 8. 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 radius that has been scaled to 90% (total energy has a minimum of –3.9 kcal/mol at 1.6 Å).

References

1. 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.
2. 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.
3. Bashford, D. & Case, D. A. (2000). Generalized born models of macromolecular solvation effects. *Annu. Rev. Phys. Chem.* **51**, 129-152.
4. Honig, B., Sharp, K. & Yang, A. S. (1993). Macroscopic Models of Aqueous Solutions: Biological and Chemical Applications. *J. Phys. Chem.* **97**, 1101-1109.
5. Lee, M. S., Salsbury, F. R. & Brooks, C. L. (2002). Novel generalized Born methods. *J. Chem. Phys.* **116**, 10606-10614.
6. 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.
7. Street, A. G. & Mayo, S. L. (1998). Pairwise calculation of protein solvent-accessible surface areas. *Fold. Des.* **3**, 253-8.
8. 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.
9. Chothia, C. (1974). Hydrophobic bonding and accessible surface area in proteins. *Nature* **248**, 338-9.
10. Lide, D. R. (2000). *CRC Handbook of Chemistry and Physics*. 81st edit, CRC Press, Boca Raton; New York; Washington D.C.
11. Creighton, T. E. (1993). *Proteins: Structures and Molecular Properties*. 2nd edit, W.H. Freeman and Company, New York.

12. 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.
13. McDonald, I. K. & Thornton, J. M. (1994). Satisfying hydrogen bonding potential in proteins. *J Mol Biol* **238**, 777-93.
14. 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.
15. 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.
16. 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.
17. Koehl, P. & Delarue, M. (1996). Mean-field minimization methods for biological macromolecules. *Curr. Opin. Struct. Biol.* **6**, 222-6.
18. 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.
19. 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.
20. 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.
21. Fersht, A. (1999). *Structure and mechanism in protein science: A guide to enzyme catalysis and protein folding*, W.H. Freeman and Company, New York.
22. 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.
23. Declerck, N. & Abelson, J. (1994). Novel substrate specificity engineered in the arabinose binding protein. *Protein Eng.* **7**, 997-1004.