A simple physical model for binding energy hot spots in protein–protein complexes

Tanja Kortemme and David Baker*

Howard Hughes Medical Institute and Department of Biochemistry, University of Washington, Seattle, WA 98195

Protein–protein recognition plays a central role in most biological processes. Although the structures of many protein–protein complexes have been solved in molecular detail, general rules describing affinity and selectivity of protein–protein interactions do not accurately account for the extremely diverse nature of the interfaces. We investigate the extent to which a simple physical model can account for the wide range of experimentally measured free energy changes brought about by alanine mutation at protein–protein interfaces. The model successfully predicts the results of alanine scanning experiments on globular proteins (743 mutations) and 19 protein–protein interfaces (233 mutations) with average unsigned errors of 0.81 kcal/mol and 1.06 kcal/mol, respectively. The results test our understanding of the dominant contributions to the free energy of protein–protein interactions, can guide experiments aimed at the design of protein interaction inhibitors, and provide a stepping-stone to important applications such as interface redesign.

Protein–protein interactions are essential to many processes within living cells and organisms. The biological function of a protein can be seen as defined by the context of its interactions in the cell (1). Delineating the complex interaction networks revealed by recent large-scale studies (2) will require tools to rationally alter and interfere with protein interactions, which in turn calls for a predictive description of the physical basis of affinity and specificity in protein interfaces. A significant amount of experimental work has addressed these questions. Particularly notable is the characterization of protein interaction hot spots: by systematically replacing protein residues by alanine (alanine scanning) and measuring the effect on binding, Wells and coworkers (3) demonstrated that only a small set of "hot spot" residues at the interface contribute significantly to the binding free energy of human growth hormone to its receptor. Many subsequent studies suggest that the presence of a few hot spots may be a general characteristic of most protein–protein interfaces (4).

Based on sequence and structural analysis, several general rules have been proposed to explain the features of interface hot spots. Although these rules appear useful for the analysis of specific interfaces, they break down when applied to a larger set of protein–protein complexes, underlining the extreme variation in size, shape, amino acid character, and solvent content of protein–protein interfaces (5). This leads to several puzzling aspects and conflicting observations associated with binding energy hot spots as identified by alanine scanning mutagenesis. First, it is not obvious from looking at structural contacts which residues are important for binding; in some cases, interactions in the center of an interface have been found to be energetically more important than those in the periphery (4), but there is no general correlation between the surface accessibility and the contribution of a residue to the binding energy (3, 4). Second, polar residues (Arg, Gln, His, Asp, and Asn) were found to be generally conserved in interfaces, and it was suggested that conserved polar residues constitute hot spots (6). However, many interaction hot spots involve hydrophobic or large aromatic residues, and it is unclear whether buried polar interactions are energetically net stabilizing or merely facilitating specificity (7). Third, some residues without significant contacts in the interface apparently contribute substantially to the free energy of

binding when assayed by alanine scanning mutagenesis, perhaps because of destabilization of the unbound proteins (8).

A quantitative model for binding energies that can be applied over a wide variety of naturally occurring protein–protein interfaces would enhance our current understanding of molecular recognition and highlight areas that require further investigation. Here we develop such a model based on an all-atom rotamer description of the side chains together with an energy function dominated by Lennard Jones interactions, solvation interactions, and hydrogen bonding. The relative success and speed of the model make it feasible to generate hot spot predictions for a wide range of protein complexes with known structures, and to guide future experiments aimed at modulating protein–protein interactions.

Materials and Methods

Datasets. Datasets for single mutations were taken from the ProTherm database (www.rtc.riken.go.jp/jouhou/protherm/protherm.html). Mutational data for protein complexes were taken from the Alanine Scanning Energetics database (ASEdb; ref. 9 and references therein; http://mullinslab.ucsf.edu/~kurt/hotspot/index.php) and additional reports (see Table 4, which is published as supporting information on the PNAS web site, www.pnas.org).

Atomic Coordinates and Preparation of Structures. Atomic coordinates were taken from structures solved by x-ray crystallography. Polar hydrogens were added to all structures, using CHARMM 19 standard bond lengths and angles. For rotatable bonds in polar hydrogen containing side chains, several rotamers reflecting different hydrogen positions were created, including a 180° flip of Asn and Gln amide groups and the two His imidazole tautomers (assumed to be uncharged). Global optimization of the hydrogen bonding network was performed for each structure by using a simple Metropolis Monte Carlo procedure as described (10), with the energy function given in Eq. **1** and described below.

The Free Energy Function. The free energy function is given in Eq. **1**. The Lennard-Jones potential, solvation term, and backbone-dependent amino acid probabilities are as described (10, 11). Energies of side chain–backbone and side chain–side chain hydrogen bonds were determined using an empirical function (T.K., A. Morozov, and D.B., unpublished work) taking into account (*i*) the distance between the hydrogen (H) and the acceptor (A) atoms, (*ii*) the angle at the hydrogen atom (D-H---A; D, donor atom), and (*iii*) the angle at the acceptor atom (H---A-AB; AB, heavy atom bound to the acceptor atom). The distance and angular-dependent terms of the hydrogen bonding potential were derived from hydrogen bond geometries observed in high-resolution (2.0 Å or better) protein crystal structures. Only hydrogen bonds with proton positions given by the chemistry of the donor group were considered for the derivation of the parameters of the potential. Coulomb electrostatics used CHARMM 19 partial charges (12) and a linear distance-dependent dielectric constant. Details and parameterization are available as *Supporting Text* and Tables 5–7, which are published as supporting information on the PNAS web site. Hydrogen bonding and Coulomb interactions were divided into three environment classes, dependent on the extent of burial of both participating

---

*To whom correspondence should be addressed. E-mail: dabaker@u.washington.edu.

residues (class 1, exposed–exposed and exposed–intermediate; class 2, exposed–buried and intermediate–intermediate; class 3, intermediate–buried and buried–buried). The extent of burial was defined by the number of Cβ atoms within a sphere of 8 Å radius of the Cβ atom of the residue of interest: exposed 0–8, intermediate 9–14, and buried >14.

Parameterizing the Energy Function on Monomeric Proteins. The relative contributions of the different terms of the free energy function were parameterized on the ProTherm dataset of X → Ala mutations by minimizing the sum of the squared differences of calculated and observed differences in stability ($\Delta G_{calc}(i) - \Delta G_{obs}(i))^2$ over all mutations i, using a conjugate-gradient-based optimization method (see Tables 5 and 6 for all weights). The Coulomb term had a negligible contribution and was excluded. The weights for the side chain–side chain hydrogen bonds showed the expected dependency on burial, with exposed hydrogen bonds contributing little energy. The amino acid type dependent reference energies (10) cancel out in the analysis of binding energy changes in interfaces, as the unbound partners are used as the reference state in this case (Eq. **2**).

Computational Alanine Scan on Interfaces. Binding free energy changes upon alanine mutation are calculated using Eqs. **1** and **2**. A schematic outline of the procedure can be found in Fig. 4, which is published as supporting information on the PNAS web site.

For modeling side-chain conformational changes on complex formation and mutation, side chains were represented as rotamers on a fixed backbone template. Rotamers were taken from a backbone dependent library by Dunbrack (13), with additional rotamers added by rotations around the χ_1 and χ_2 angles by 5–20° and extra rotamers for χ_3 and χ_4 angles as described by Dahiyat and Mayo (14). The x-ray coordinates of the native side chains at each position were included in the library for both the complex and the isolated partners. All residues having at least one side-chain atom within a sphere of 5 Å radius of the site of mutation were repacked. All other amino acid side chains were left in the conformations observed in the parent crystal structures. Energies were computed for each rotamer with the constant part of the molecule (the template backbone and all unchanged side chains), and for all pairwise rotamer–rotamer combinations by using the free energy function described in Eq. **1**. Global optimization of side-chain conformations was performed using a Monte Carlo simulated annealing procedure (10) in which a move consisted of the replacement of a randomly picked side-chain rotamer at a single position by another rotamer from the library.

Hot Spot Classification. Neutral residues and hot spots are defined as residues showing a change in the binding free energy by less or more than 1 kcal/mol when replaced by alanine ($\Delta\Delta G_{bind}$), respectively. (Alanine substitutions with experimentally measured stabilizing effects were rare and not larger than −0.9 kcal/mol; these were included in the neutral category.) A correctly identified hot spot thus means a residue with a predicted and observed $\Delta\Delta G_{bind}$ value larger than or equal to 1 kcal/mol; a correctly identified neutral residue has both predicted and observed $\Delta\Delta G_{bind}$ values less than 1 kcal/mol. Interface residues were defined as residues with a side chain having at least one atom within a sphere with 4 Å radius of an atom belonging to the other partner in the complex.

Results

A Simple Physical Model. An effective model for the free energy of interactions at protein–protein interfaces should include the following dominant physical considerations: (*i*) shape complementarity including the detailed packing interactions of interface atoms, (*ii*) polar interactions involving ion pairs and hydrogen bonds, and (*iii*) the interactions of protein atoms with the solvent, including a penalty for the desolvation of buried polar groups. Furthermore, the effects of mutations on both the protein–protein complex and

the unbound partners must be assessed. To model these contributions, our computational method uses an atomic representation of the protein (including all heavy atoms, as well as polar hydrogens) and a free energy function consisting of a linear combination of the attractive part of a Lennard-Jones potential (E_{LJattr}), a linear distance-dependent repulsive term (E_{LJrep}), an orientation-dependent side chain–backbone and side chain–side chain hydrogen bond potential ($E_{HB(sc-bb)}$ and $E_{HB(sc-sc)}$; T.K., A. Morozov, and D.B., unpublished work), Coulomb electrostatics (E_{Coul}), and an implicit solvation model (G_{sol}) (11):

$$\Delta G = W_{attr}E_{LJattr} + W_{rep}E_{LJrep} + W_{HB(sc-bb)}E_{HB(sc-bb)}$$
$$+ W_{HB(sc-sc)}E_{HB(sc-sc)} + W_{Coul}E_{Coul} + W_{sol}G_{sol}$$
$$+ W_{\phi/\psi}E_{\phi/\psi}(aa) + \sum_{aa=1}^{20} n_{aa}E_{aa}^{ref}, \qquad [1]$$

where W are the relative weights of the different energy terms (for details on the parameterization see *Materials and Methods* and Tables 5 and 6), $E_{\phi/\psi}(aa)$ is an amino acid type (aa)-dependent backbone torsion angle propensity, and E_{aa}^{ref} is an amino acid type-dependent reference energy which approximates the interactions made in the unfolded state ensemble (ref. 10; n_{aa} is the number of amino acids of a certain type); the last two terms were included to model changes in protein stability on mutation, as described below, but do not contribute to free energy changes of protein–protein interactions.

The effects of alanine replacement were computed both for the protein complex and for the corresponding uncomplexed partners to yield the change in binding energy $\Delta\Delta G_{bind}$:

$$\Delta\Delta G_{bind} = \Delta G_{bind}^{MUT} - \Delta G_{bind}^{WT}$$
$$= (\Delta G_{complex}^{MUT} - \Delta G_{partner\,A}^{MUT} - \Delta G_{partner\,B}^{MUT})$$
$$- (\Delta G_{complex}^{WT} - \Delta G_{partner\,A}^{WT} - \Delta G_{partner\,B}^{WT}), \qquad [2]$$

where $\Delta G_{complex}$, $\Delta G_{partner\,A}$, and $\Delta G_{partner\,B}$ are the stabilities of the complex and the unbound partners, and WT and MUT describe wild-type and mutant proteins.

In contrast to binding energy calculations using molecular dynamics simulations, conformational changes are restricted to side chain movements: amino acid side chains are modeled as rotamers on a fixed polypeptide backbone, as has been used successfully in protein design methods (ref. 15 and references therein).

Model Parameterization and Performance on Monomeric Proteins. The relative weights of the energy terms (W) and amino acid dependent reference energies (E_{aa}^{ref}) were parameterized using a dataset of stability changes measured in 743 single mutations of type X → Ala in monomeric proteins taken from the PROTHERM database (16). The overall correlation coefficient for predicted versus observed stability changes is 0.75 over the entire dataset with an expected slope and intercept of 1.0 and 0.0, respectively (data not shown). Cross validation by splitting the set into a training and test set yielded essentially identical results, with an average unsigned error of 0.81 kcal/mol.

Our energy function also does reasonably well in predicting changes in protein stability for monomeric proteins brought about by nonalanine mutation. Application of the method to 1,584 mutations of type $X \to Y$ (where Y is an amino acid smaller or of same size as X), allowing side chain conformational rearrangements, yields a correlation between experimental and predicted stability changes of 0.70. The method could be useful for classification of amino acid changes in proteins caused by single nucleotide polymorphisms into neutral and deleterious classes (17, 18).

Table 1. Predicted hot spot and neutral residues for all 19 protein complexes studied experimentally by alanine scanning mutagenesis

Neutral residues		Hot spot residues		Hot spot residues in the interface		Protein complex				
Total	Fraction correct	Total	Fraction correct	Total	Fraction correct	PDB code	Partner A	Partner B	No. mutations in partner A	No. mutations in partner B
50	0.78	15	0.47	11	0.64	1a22	HGH	hGBbp	29	36
3	1.00	5	0.20	1	1.00	1ahw	TF-Fab	Fab 5G9	8	—
22	0.55	6	0.83	6	0.83	1a4y	Angiogenin	RNase Inh.	14	14
3	0.67	11	0.82	9	1.00	1brs	Barnase	Barstar	8	6
20	0.95	10	0.50	9	0.56	1bxi	Im9	E9 DNase	30	—
7	0.71	1	1.00	1	1.00	1cbw	BPTI	Chymotrypsin	8	—
3	1.00	11	0.55	10	0.60	1dfj	Rnase	RNase Inh.	—	14
0	0.00	5	1.00	5	1.00	1dn2	IgG	Peptide	3	2
39	0.90	4	1.00	4	1.00	1dan	FactorVII	TF	—	43
5	0.60	4	0.75	3	1.00	1f47	ZipA	FTSZ fragment	—	9
1	1.00	2	0.00	2	0.00	1fc2	Protein A	IgG	3	—
3	1.00	5	0.80	5	0.80	1fcc	Protein G	IgG	8	—
45	0.93	4	0.50	2	0.50	1gc1	CD4	gp120	49	—
1	1.00	8	0.62	8	0.62	1jck	SEC3	TCR Vβ	9	—
13	0.92	18	0.78	17	0.82	1jrh	Interferon γ	A6	12	19
0	0.00	1	1.00	1	1.00	1nmb	Neurominid.	NC10	—	1
18	0.89	10	0.80	8	1.00	1vfb	D1.3	HEL	16	12
0	0.00	1	1.00	1	1.00	2ptc	BPTI	Trypsin	1	—
8	0.88	18	0.83	17	0.88	3hfm	HEL	HYHEL-10	13	13
241	0.84	139	0.69	120	0.79	ALL				

For a table containing data on all the individual mutations see Table 8, which is published as supporting information on the PNAS web site. For hot spot classification see *Materials and Methods.*

Application of the Model to Computational Alanine Scanning on Protein Interfaces. The parameterized free energy function was applied to a database of 19 complexes with known crystal structures and experimentally measured changes in binding energy on alanine mutagenesis (see *Materials and Methods* and Table 4). The weights obtained from the monomeric protein training set described above were not reparameterized on the interface dataset because we wanted to test whether a general energy function not parameterized on protein interfaces would be able to explain the hot spot phenomenon. However, the hydrogen bonding contribution was scaled so that the maximal effect of removing one of the partners in a buried hydrogen bond with ideal geometry was −4.5 kcal/mol, as previously estimated experimentally (ref. 19; this term includes a penalty for leaving an unpaired donor or acceptor). This modification was motivated by the fact that buried polar interactions are abundant in protein interfaces, but underrepresented in the monomeric protein mutation data set, and thus might not be properly modeled by fitting the hydrogen bonding contribution on the monomeric set.

Overall Performance on Interfaces. Table 1 shows the results of the simplest static model for all 19 complexes, leaving all side chains apart from the mutated residue in their crystal structure conformations in both the bound and unbound proteins. The overall correlation between observed and calculated changes in binding energy is slightly worse than for the monomeric proteins, with an average unsigned error of 1.06 kcal/mol for the 233 mutations in the interface area (the average unsigned error for all 380 mutations in the experimental dataset, including residues not making significant contacts in the interface, is 0.83 kcal/mol). In 84% of the cases a small effect of alanine replacement on the binding energy ($\Delta\Delta G_{bind} \geq 1$ kcal/mol) is correctly predicted, whereas 69% of hot spots are identified (Table 1). If only residues in the interface are considered, the fraction of correctly predicted hot spots increases to 79%.

Dominant Contributions to Interface Energetics. How important are the different terms of our simple physical model for characterizing

protein interaction hot spots? A major contributor to protein stability in monomeric proteins is likely the precise packing of hydrophobic residues in the protein core. In contrast to protein cores, protein–protein interfaces are considerably more polar, with an average residue composition intermediate between that of protein cores and surface (5). Table 2 analyzes the relative importance of the terms describing polar interactions in our model: hydrogen bonding, Coulomb electrostatics, and solvation. Clearly the hydrogen bonding term contributes significantly to the correct prediction of hot spots. Both backbone–side chain and side chain–side chain hydrogen bonds play a significant role (backbone–backbone hydrogen bonds are also expected to play a role, but are not probed by alanine scanning). The increase of the contribution of the hydrogen bonding term relative to the monomeric protein training set and the environment dependency of the hydrogen

Table 2. Contribution of polar interactions to hot spot prediction

Model	Fraction hot spots correct		Fraction neutral residues correct	
	All	Interface	All	Interface
Complete model (see Table 1)	0.69	0.79	0.84	0.68
Exclusion of all hb terms	0.40	0.47	0.94	0.88
Exclusion of bb-sc hb term	0.63	0.72	0.87	0.73
Exclusion of sc-sc hb term	0.50	0.58	0.91	0.81
Hb terms not environment dependent	0.70	0.81	0.76	0.51
Weights of hb from monomeric proteins	0.60	0.70	0.88	0.75
Coulomb electrostatics instead of hb terms	0.49	0.56	0.91	0.81
Exclusion of solvation term	0.83	0.95	0.71	0.40

For each model calculation, the relative contribution of certain interactions to the free energy function is altered. hb, hydrogen bonding; bb-sc, backbone–side chain; sc-sc, side chain–side chain.
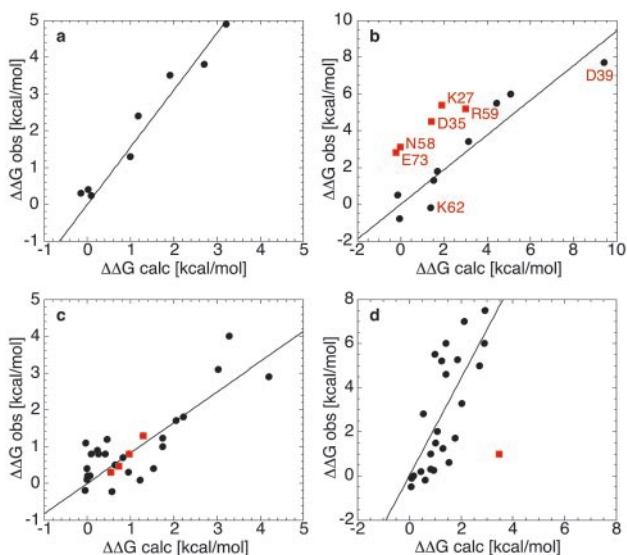
Fig. 1. Predicted versus observed changes in binding free energy brought about by alanine replacement for four selected protein complexes. Lines reflect linear fit with a fixed zero intercept. (a) Protein G bound to an IgG Fc fragment (1fcc). Linear fit yields a correlation coefficient of 0.97. (b) Barnase–barstar (1brs). Residues marked in red are reported to form water-mediated hydrogen bonds in the interface, and show underpredictions in five of seven cases (red squares). If these residues are excluded, a linear fit yields a correlation coefficient of 0.96. (c) Hen egg-white lysozyme bound to the antibody D1.3 (1vfb). Residues reported to form water-mediated hydrogen bonds in the complex are indicated by red squares. A linear fit to all mutations yields a correlation coefficient of 0.82. (d) Hen egg-white lysozyme bound to the antibody HYHEL-10 (3hfm). A linear fit to all mutations excluding a residue making a water mediated hydrogen bond (red square) yields a correlation coefficient of 0.76.
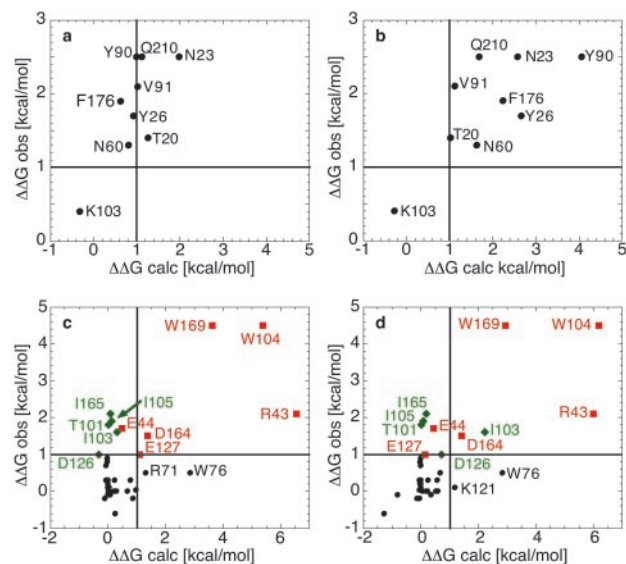

Fig. 2. Predicted versus observed changed in binding free energy brought about by alanine replacement for two selected protein complexes, illustrating the effect of including side-chain rearrangements in the vicinity of the mutation. Solid lines at 1 kcal/mol indicate the qualitative hot spot classification (see *Materials and Methods*; correctly predicted hot spots are in the upper right quadrant). (a and b) Staphylococcal enterotoxin C3 bound to the T cell receptor β chain (1jck). a and b are excluding or including side-chain conformational changes, respectively. (c and d) Human growth hormone bound to its receptor (hGHbp; 1a22); for clarity only mutations on the hGHbp are shown. Hot spots (red squares) and residues exerting indirect effects on the binding free energy (green diamonds) are marked as identified by the authors (ref. 21; Pro-106 is not shown; a Pro to Ala mutation might cause a structural change in the backbone, which is not accounted for in the current method). c and d are excluding or including side-chain conformational changes, respectively.

bonding term add to the accuracy of hot spot prediction: in particular, dropping the environment dependency of the hydrogen bonding strength results in a large increase in false positives. The results in Table 2 also suggest that the environment-dependent, database-derived hydrogen bonding term is a better description of polar interactions in interfaces than Coulomb electrostatics with a distance-dependent dielectric: the fraction of correctly predicted hot spots in interface positions by using just a Coulomb term (adjusted to have a roughly equal magnitude and environment dependency as the hydrogen bonding term) compared with using just the hydrogen bonding term decreases from 0.79 to 0.56 (Table 2). Lastly, the exclusion of the implicit solvation term also decreases the performance: a large fraction of neutral residues are predicted to be hot spots.

Overview of Predictions for Individual Interfaces. The accuracy of the predictions for different interfaces was quite variable, with the fraction of correctly predicted interface hot spot residues ranging from 0.50 to 1.0 (Table 1). Examples of particularly good predictions are shown in Fig. 1 for the protein G B1 domain bound to an IgG fragment (1fcc, Fig. 1a) and for the barnase–barstar complex (1brs, Fig. 1b). It should be noted that both these complexes are highly polar, indicating that our simple model of polar interactions is reasonably accurate. In the protein G-IgG case, the high correlation between the experimental and predicted $\Delta\Delta G_{bind}$ values may also reflect the pronounced knobs into holes packing of this interface (20) that allows only minimal conformational change on mutation. Reasonable predictions are also obtained for antibody–antigen complexes (Fig. 1 c and d). For a number of interfaces the predicted $\Delta\Delta G_{bind}$ values are quite well correlated with the experimental values but significantly smaller in magnitude (Fig. 1 a and d, for example).

Particular challenges for interface modeling are first, the treat-

ment of interactions involving specific water molecules, and second, the treatment of conformational rearrangements. In the following two sections we provide examples of successes and failures of the model in both areas.

Interfaces Containing Explicit Water Molecules. Water molecules are not explicitly represented in our model; thus, a hydrogen bond mediated by a conformationally restricted water molecule (presumably stabilizing the interface) will be missed. Alanine replacement of such a hydrogen bonding side chain will be predicted to be neutral instead of destabilizing; for example, in the barnase–barstar interface the effect of alanine replacement on binding is underpredicted in five of seven water bridge-forming residues (Fig. 1b). Conversely, loss of hydrogen bonds on alanine mutation can be compensated for by stable inclusion of additional water molecules reducing the loss in binding energy. This has been observed experimentally for a lysozyme–antibody complex (1vfb). The reasonably accurate predictions for such mutations in the 1vfb interface (Fig. 1c) suggest that the combination of the environment dependence of the hydrogen bonding term and the implicit solvation model to some extent captures this effect.

Modeling of Conformational Rearrangements. Flexibility is a hallmark of many protein–protein interfaces. Side-chain rearrangements, which can play an important role in protein–protein interaction energetics (21), can in principle be modeled using rotamer repacking methods (10, 14). Although for most interfaces simultaneous optimization of rotameric side-chain conformations did not change the predictions significantly, an example of improved predictions is the complex between staphylococcal enterotoxin C3 and the T cell receptor β chain (1jck): by incorporating side-chain flexibility, all hot spots are identified correctly (Fig. 2 a and b). Interestingly, this improvement is mainly due to rearrangements to

Table 3. Hot spot prediction for the mdm2–p53 interface

P53			mdm2		
Residue	$\Delta\Delta G_{bind}$ predicted, kcal/mol	No. of allowed substitutions (peptide)	Residue	DDG$_{bind}$ predicted, kcal/mol	Contacts p53 residues*
E17	−0.17	19	I 30	0.97	**W23**
T 18	−0.02	13	L 33	0.29	**W23**
F 19	3.85	1	I 37	0.90	**F19, W23**
S 20	0.15	7	M 38	0.51	S20
D 21	0.00	6	**Y 43**	1.23	**F19**
L 22	1.06	3	**Q 48**	1.39	T18, **F19**
W 23	4.54	1	**V 69**	1.21	**F19, L22**
K 24	−0.02	15	K 70	0.06	E17
L 25	0.00	14	L 75	0.82	**L26**
L 26	2.04	1	**Y 76**	1.37	**L26**

All residues in the p53 peptide fragment and mdm2 residues contacting the peptide were computationally replaced by alanine. The number of allowed substitutions was taken from ref. 21 and reflects the number of different amino acids that could be individually substituted at each position in a phage display selected peptide similar to the p53 fragment without significant loss in binding affinity. Predicted hot spot residues are shown in bold type.
*Residues in bold type are p53 hot spots (see left half).

energetically more favorable rotamers when repacking the *native* complex structure, leading to a larger predicted effect on alanine mutation that is closer to the experimentally observed change. This effect might be due to the relatively low resolution (3.5 Å) of the 1jck structure.

The interface between human growth hormone and its receptor is perhaps the most dramatic example of interface plasticity in our dataset (21, 22). The two main experimentally determined hot spots in this interface, two tryptophan residues, are identified correctly (Fig. 2c). Several residues in the interface have an effect on binding without forming significant contacts to the other molecule in the complex, presumably because they are important in positioning the tryptophan residues in the complex (21), and their $\Delta\Delta G_{bind}$ values are underpredicted by our model. Although the inclusion of side-chain rearrangements slightly improves the prediction for two of these residues and the overall correlation between observed and calculated $\Delta\Delta G_{bind}$ values, prediction for one hot spot residue was worse (Fig. 2d). Interestingly, the inclusion of side-chain rearrangements correctly predicts an alanine mutation that actually stabilizes the interface (Fig. 2d). Nevertheless, even with rotamer repacking the dynamical nature of the human growth hormone–receptor interface can clearly not be reproduced accurately by using our model.

Prediction of Binding Energy Hot Spots Yet to be Characterized. Our method requires the availability of a structure for the protein complex to be analyzed, but can then be applied to protein complexes with uncharacterized binding energetics. Particularly interesting are complexes involving large protein families for which there is evolutionary and/or experimental information about amino acid preferences at each site: do the sequence positions intolerant to amino acid changes correspond to hot spots? As an example, the results of computational alanine scanning mutagenesis on the interface between mouse double minute 2 (mdm2) and a fragment of p53 (23) are given in Table 3. All positions in the p53 fragment that experiments have shown do not tolerate amino acid substitutions without loss of binding affinity (24) are identified computationally as binding energy hot spots (Table 3, left). All predicted hot spot residues in mdm2 contact predicted hot spot residues in p53 (Table 3, right). Interestingly, there is a strong correlation (0.96) between results using the molecular mechanics Poisson Boltzmann surface area (MM-PBSA) approach to predict binding energy changes at this interface (25) and the simple model presented here (Fig. 3).

Discussion

Evaluation of the Simple Model for Hot Spot Prediction. Despite the vast diversity of known protein–protein interfaces, the simple model described here shows considerable success in the qualitative prediction of binding energy hot spots experimentally determined by alanine scanning mutagenesis. Remarkably, 79% of all interface hot spots can be predicted using the simple free energy function dominated by packing interactions, hydrogen bonds, and an implicit solvation model (Table 1), ignoring changes in backbone conformation or effects on the dynamics of the interface. These results suggest that the model captures much of the key physics underlying protein–protein interactions, which is encouraging given the issues and contradictory observations outlined in the introduction.

What are the origins of this encouraging performance? Our treatment of hydrogen bonding clearly yields better agreement with experimental data than a description of polar interactions by using Coulomb electrostatics with a linear distance-dependent dielectric constant (Table 2). There are two likely reasons for this. First, the hydrogen bond model incorporates the significant orientation dependence of the hydrogen bond, which is ignored in the Coulomb description. Second, the Coulomb model is likely to introduce considerable noise in cases where shifts in ionization constants occur: the assumption that acidic or basic residues largely buried in an interface are charged might not be warranted. Modeling of the free energy contributions of hydrogen bonds and electrostatic interactions is complicated because the enthalpic gains are offset by the cost of desolvating polar groups and the loss in side chain
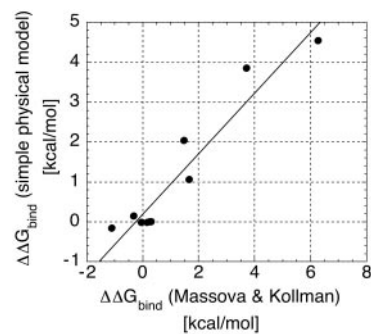

Fig. 3. Comparison of changes in binding energy for the interaction of mdm2 and a p53 fragment (1ycq) brought about by alanine scanning calculated by the simple model or by the MM-PBSA approach (25).

Kortemme and Baker

conformational entropy. The implicit solvation term in our model opposes burial of polar groups and prevents an overestimation of the magnitude of electrostatic and hydrogen-bonding interactions that has been identified previously as a problem in energy functions used in protein design (26). While inclusion of an explicit measure of side-chain conformation entropy changes did not improve model performance (data not shown), to some extent the environment-dependent hydrogen bonding term incorporates the differences in the entropic cost of freezing exposed and buried side chains implicitly.

The model makes more precise some of the general arguments about protein–protein interface energetics. For example, two different explanations have been given for the observation that the largest effects observed in alanine scanning experiments are frequently in the center rather than the periphery of protein–protein interfaces. First, the peripheral residues serve as an O-ring to exclude solvent from the center (4), where a lowered effective dielectric constant in a "dryer" environment strengthens electrostatic and hydrogen bonding interactions. Second, the residues in the core and periphery make equivalent contributions to stability, but an interaction deleted by alanine mutagenesis in the periphery can be replaced by a water molecule in the periphery and hence causes less loss in stability (27). The evaluation of the contributions of the individual terms in the model (Table 2) suggests that both effects are operative: the environment-dependency of the hydrogen bonding term corresponds to different effective dielectrics in buried versus exposed environments and the replacement of interactions in the periphery by water molecules is modeled by a favorable solvation contribution for exposed polar groups.

A number of aspects of protein interactions cannot be predicted well, reflecting the simplifications of the model: (*i*) In many cases, although a residue can be qualitatively identified as a hot spot, the magnitude of electrostatic effects cannot be captured; this is particularly noticeable for specific amino acid types (such as aspartic acid and glutamic acid). This is a clear case where more accurate treatments of electrostatics in proteins accounting for induced polarization effects and shifts in ionization constants (28) are necessary. (*ii*) Mutational effects of replacing residues forming water-mediated hydrogen bonds across the interface (Fig. 1*b*) are often underpredicted, which reflects the fact that specifically bound water molecules are not taken into account and solvation effects are treated only implicitly. An explicit inclusion of defined water molecules in the interface could yield a significant improvement (29), and can be incorporated into the current rotamer approach by modeling water moieties as side chain extensions. (*iii*) Only limited side-chain conformational changes are taken into account. While in some cases the incorporation of these side-chain rearrangements improves performance (Fig. 2), it is surprising that the very simplified static approach yields successful qualitative predictions in most cases. Much more time intensive molecular dynamics methods using a molecular mechanics-Poisson-Boltzmann surface area

(MM-PBSA)-based energy function (30) also could not fully reproduce the alanine scanning results for protein complexes with a high degree of plasticity, such as the human growth hormone–receptor interface.

Despite significant progress in specific cases (28, 30), general computational methods to predict the molecular determinants of protein interfaces have remained elusive (8). To our knowledge, this study is the first attempt at modeling alanine scanning results for a large set of interfaces. In addition to the details of the representation—notably the explicit hydrogen bond term—an advantage of our simple method is its speed; computational alanine scanning required between less than 2 min (20 positions in the mdm2-p53 interface) and several hours (65 mutations including considerable side-chain rearrangements in the 1a22 interface) on an Intel 800-MHz processor, making larger-scale applications feasible. To further test the method and perhaps guide future experiments, hot spot predictions for protein–protein complexes of known structure are available from the authors on request (kortemme@u.washington.edu; dabaker@u.washington.edu).

Applications of the Simple Model for Binding Energy Hot Spots. Protein–protein interfaces have been particularly challenging targets for inhibitor development because of the often large size and nonsequential nature of the binding site. Hot spot predictions using the model described here could provide a starting point to narrow down the large interface area: a small molecule drug mimicking a significant number of interactions made by hot spot residues could block formation of the protein–protein complex (3).

A model for the free energy of protein–protein interactions is necessary for any design approaches aimed at redesign of interface specificity or creation of new interfaces. All terms in our free energy function are pairwise additive, and optimal interface sequences can be rapidly obtained using a simple Monte Carlo procedure (10). We have very recently used such an approach to create a functional endonuclease with a new DNA cleavage specificity by computationally optimizing a domain–domain interface generated by fusing domains from distantly related endonucleases. The x-ray structure of this enzyme showed that the amino acid side-chain conformations were essentially identical to those predicted by the optimization procedure (31), indicating further that the simple physical model developed in this paper provides a reasonable description of protein–protein interaction free energies.

Note Added in Proof. A related model for free energy changes upon mutation in proteins and protein–protein complexes has recently been published by Serrano and coworkers (32).

1. Eisenberg, D., Marcotte, E. M., Xenarios, I. & Yeates, T. O. (2000) *Nature* **405,** 823–826.
2. Zhu, H. & Snyder, M. (2002) *Curr. Opin. Cell Biol.* **14,** 173–179.
3. Clackson, T. & Wells, J. A. (1995) *Science* **267,** 383–386.
4. Bogan, A. A. & Thorn, K. S. (1998) *J. Mol. Biol.* **280,** 1–9.
5. Conte, L. L., Chothia, C. & Janin, J. (1999) *J. Mol. Biol.* **285,** 2177–2198.
6. Hu, Z., Ma, B., Wolfson, H. & Nussinov, R. (2000) *Proteins* **39,** 331–342.
7. Hendsch, Z. S. & Tidor, B. (1994) *Protein Sci.* **3,** 211–226.
8. DeLano, W. L. (2002) *Curr. Opin. Struct. Biol.* **12,** 14–20.
9. Thorn, K. S. & Bogan, A. A. (2001) *Bioinformatics* **17,** 284–285.
10. Kuhlman, B. & Baker, D. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 10383–10388.
11. Lazaridis, T. & Karplus, M. (1999) *Proteins* **35,** 133–152.
12. Neria, E., Fischer, S. & Karplus, M. (1996) *J. Chem. Phys.* **105,** 1902–1921.
13. Dunbrack, R. L., Jr., & Cohen, F. E. (1997) *Protein Sci.* **6,** 1661–1681.
14. Dahiyat, B. I. & Mayo, S. L. (1997) *Science* **278,** 82–87.
15. Pokala, N. & Handel, T. M. (2001) *J. Struct. Biol.* **134,** 269–281.
16. Gromiha, M. M., Uedaira, H., An, J., Selvaraj, S., Prabakaran, P. & Sarai, A. (2002) *Nucleic Acids Res.* **30,** 301–302.
17. Wang, Z. & Moult, J. (2001) *Hum. Mutat.* **17,** 262–270.
18. Ng, P. C. & Henikoff, S. (2001) *Genome Res.* **11,** 863–874.
19. Fersht, A. R., Shi, J. P., Knill-Jones, J., Lowe, D. M., Wilkinson, A. J., Blow, D. M., Brick, P., Carter, P., Waye, M. M. & Winter, G. (1985) *Nature* **314,** 235–238.
20. Sloan, D. J. & Hellinga, H. W. (1999) *Protein Sci.* **8,** 1643–1648.
21. Clackson, T., Ultsch, M. H., Wells, J. A. & de Vos, A. M. (1998) *J. Mol. Biol.* **277,** 1111–1128.
22. Atwell, S., Ultsch, M., De Vos, A. M. & Wells, J. A. (1997) *Science* **278,** 1125–1128.
23. Kussie, P. H., Gorina, S., Marechal, V., Elenbaas, B., Moreau, J., Levine, A. J. & Pavletich, N. P. (1996) *Science* **274,** 948–953.
24. Bottger, A., Bottger, V., Garcia-Echeverria, C., Chene, P., Hochkeppel, H. K., Sampson, W., Ang, K., Howard, S. F., Picksley, S. M. & Lane, D. P. (1997) *J. Mol. Biol.* **269,** 744–756.
25. Massova, I. & Kollman, P. A. (1999) *J. Am. Chem. Soc.* **121,** 8133–8143.
26. Marshall, S. A., Morgan, C. S. & Mayo, S. L. (2002) *J. Mol. Biol.* **316,** 189–199.
27. Janin, J. (1999) *Struct. Fold Des.* **7,** R277–R279.
28. Sharp, K. A. (1998) *Proteins* **33,** 39–48.
29. Covell, D. G. & Wallquist, A. (1997) *J. Mol. Biol.* **269,** 281–297.
30. Kollman, P. A., Massova, I., Reyes, C., Kuhn, B., Huo, S., Chong, L., Lee, M., Lee, T., Duan, Y., Wang, W., *et al.* (2000) *Acc. Chem. Res.* **33,** 889–897.
31. Chevalier, B. S., Kortemme, T., Chadsey, M. S., Baker, D., Monnat, R. J. & Stoddard, B. L. (2002) *Mol. Cell*, in press.
32. Guerois, R., Nielsen, J. E. & Serrano, L. (2002) *J. Mol. Biol.* **320,** 369–387.

BIOPHYSICS