doi:10.1016/S0022-2836(03)00021-4 J. Mol. Biol. (2003) 326, 1239–1259

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

An Orientation-dependent Hydrogen Bonding Potential Improves Prediction of Specificity and Structure for Proteins and Protein–Protein Complexes

Tanja Kortemme[a], Alexandre V. Morozov[a,b] and David Baker[a]*

[a]*Howard Hughes Medical Institute and Department of Biochemistry J-567 Health Sciences Box 357350, University of Washington, Seattle WA 98195-7350, USA*

[b]*Department of Physics University of Washington Box 351560, Seattle WA 98195-1560, USA*

**Corresponding author*

Hydrogen bonding is a key contributor to the specificity of intramolecular and intermolecular interactions in biological systems. Here, we develop an orientation-dependent hydrogen bonding potential based on the geometric characteristics of hydrogen bonds in high-resolution protein crystal structures, and evaluate it using four tests related to the prediction and design of protein structures and protein–protein complexes. The new potential is superior to the widely used Coulomb model of hydrogen bonding in prediction of the sequences of proteins and protein–protein interfaces from their structures, and improves discrimination of correctly docked protein–protein complexes from large sets of alternative structures.

Keywords: hydrogen bond; electrostatics; protein docking; protein design; free energy function

Introduction

Hydrogen bonding interactions are abundant in proteins and protein–protein complexes.[1,2] Despite their ubiquitous nature, the relative importance of hydrogen bonds for protein stability and protein–protein recognition has been somewhat controversial.[3,4] Most hydrogen bond donors and acceptors are satisfied in non-surface-accessible parts of proteins.[1,5] However, replacement of buried salt-bridge networks with hydrophobic residues can lead to protein stabilization.[6] There may be no net gain in free energy for hydrogen bond formation in folding and binding, as the formation of hydrogen bonds between protein atoms results in the loss of hydrogen bonds formed with water.[7,8] Thus hydrogen bonds might primarily provide specificity rather than stability to proteins and protein–protein interfaces.[9,10]

An accurate energetic description of hydrogen bonding interactions is required for understanding the role of hydrogen bonds in both intramolecular and intermolecular interactions. However, the physical nature of hydrogen bonds is complex. *Ab initio* calculations decompose the total energy of a hydrogen bond into several components: electrostatics, polarization, exchange repulsion, charge-transfer and coupling contributions,[11,12] and calculation of these terms from first principles is not straightforward for biological macromolecules. Phenomenologically, a hydrogen bond is formed when a positively polarized hydrogen atom (bound to an electronegative donor atom) penetrates the van der Waals sphere of an acceptor atom to interact with its lone pair electrons (or polarizable π-electrons in the case of aromatic rings). This partial covalent character implies a directionality of the hydrogen bond. The observed orientation dependence of hydrogen bonds in crystal structures of small molecules,[13–18] and proteins[1,19–21] generally supports an orientation of the hydrogen towards the lone electron pairs of the acceptor atom. However, the location of the lone pair cannot be simply assumed based on the hybridization of the acceptor, as the hybridization state of the acceptor atom itself is perturbed by hydrogen bond formation, leading to a distortion of the original hybridization by mixing with the 1s orbital of the hydrogen.[22,23] This highlights the potential "environment dependency" of hydrogen bonding interactions. Additional problems are posed by polarization effects causing non-additivity in hydrogen bond energetics.

Whereas earlier molecular mechanics potentials included explicit hydrogen bonding terms,[24,25] current force fields generally attempt to model the specifics of hydrogen bonds by a combination of Coulomb and Lennard–Jones interactions with

Abbreviation used: rmsd, root-mean-square deviation.

E-mail address of the corresponding author: dabaker@u.washington.edu

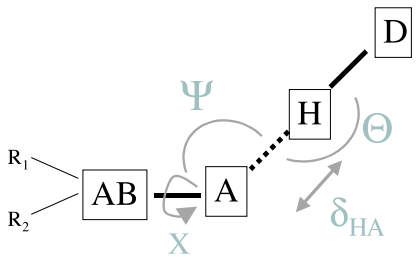

Figure 1. Schematic representation of the parameters used to describe hydrogen bond geometry. δ_{HA}, distance between the hydrogen and acceptor atoms; Θ, angle at the hydrogen atom; Ψ, angle at the acceptor atom; *X*, the dihedral angle given by rotation around the acceptor–acceptor base bond in the case of an sp^2 hybridized acceptor. A, acceptor; D, donor; H, hydrogen; AB, acceptor base; R_1, R_2, atoms bound to the acceptor base.

refined atomic charges,[26–29] although explicit hydrogen bonding has been used in potentials applied successfully to protein design.[30,31] In the absence of feasible first principle methods, our approach to the improvement of current hydrogen bonding potentials relies on chemical intuition and the vast information available in the protein structure database. This approach is conceptually similar to previous studies of hydrogen bonds which use information available in databases of small molecule crystal structures,[15,18] but determines the relevant parameters for proteins (while the physical principles governing interactions should be transferable between different classes of molecules, the details might not be).

We derive a hydrogen bond energy function based on geometrical parameters of hydrogen bonds observed in high-resolution protein crystal structures. Subsequently, we evaluate the new hydrogen bonding potential and compare it to a purely electrostatic representation of polar interactions using four different tests: the recovery of the native amino acid sequence based on the structure of proteins (test 1) and protein–protein complexes (test 2), the discrimination of misfolded from native or near-native protein structures (test 3) and the identification of correct relative orientations of protein partners in protein–protein complexes (test 4). The four tests are closely related to the protein design problem (test 1, for review see Pokala & Handel[32]), the protein–protein interface design problem (test 2), the decoy discrimination problem (test 3)[33,34] and the protein docking problem (test 4).[35] Our tests demonstrate the usefulness of the database-derived hydrogen bonding function, its superiority to simple effective distance-dependent Coulomb treatments of electrostatic interactions in our test cases, and highlight the importance of continued development of accurate descriptions of hydrogen bonding interactions in biological systems.

Results

Derivation of the hydrogen bonding function

Hydrogen bond geometries were derived from a set of 698 crystal structures with a resolution of better than 1.6 Å and *R*-factors of better than 0.25 (see Methods). Figure 1 illustrates the four geometrical parameters considered: (a) the distance δ_{HA} between the hydrogen and acceptor atoms, (b) the angle Θ at the hydrogen atom, (c) the angle Ψ at the acceptor atom and (d) the dihedral angle *X* corresponding to rotation around the acceptor–acceptor base bond in the case of an sp^2 hybridized acceptor (the distribution in the sp^3 case is flat).

This analysis requires the explicit placement of polar hydrogen atoms, which has been noted to be important for correct treatment of hydrogen bonding interactions.[36] As hydrogen atoms are generally not included in the coordinates derived from the crystal diffraction data, polar hydrogen atoms were added in cases where the position of the hydrogen atom was defined by the chemistry of the donor group (backbone amide protons, tryptophan indole, histidine imidazole, asparagine and glutamine amide groups and arginine guanido group). Standard bond lengths and angles were taken from the CHARMM19 force field parameters.[28] Polar hydrogen atoms associated with a rotatable bond (serine, threonine and tyrosine hydroxyl groups and lysine amino group) were not considered for the compilation of statistics as they could not be placed without making assumptions about the hydrogen bond geometry.

Figure 2 shows the distributions of δ_{HA}, Θ, Ψ and *X* obtained from the analysis of a total of 11,680 side-chain–side-chain and 89,537 backbone–backbone hydrogen bonds. Backbone hydrogen bonds were treated separately as their geometry was found to differ significantly from that of side-chain–side-chain hydrogen bonds, presumably due to steric constraints imposed by regular secondary structure elements. For the angular distributions of backbone–backbone hydrogen bonds, only occurrences with a proton-acceptor distance between 1.4 Å and 2.6 Å were considered. For side-chain–side-chain hydrogen bonds, angle distributions were collected in two different distance ranges (1.4–2.1 Å and 2.1–3.0 Å) to take into account a shift in hydrogen bond geometry at longer distances observed in high-resolution protein structures.[37] The distributions shown were corrected for the differences in volume elements of the bins (see Figure 2). For the dependence on the acceptor angle Ψ, separate statistics were collected for sp^2 and sp^3 hybridized acceptor atoms to take into account different electronic distributions around the acceptor atom. The distance distributions for both side-chain–side-chain and backbone–backbone hydrogen bonds show a maximum at around 2.0 Å, with slightly shorter distances observed for side-chain–side-chain hydrogen bonds with an sp^2 hybridized acceptor

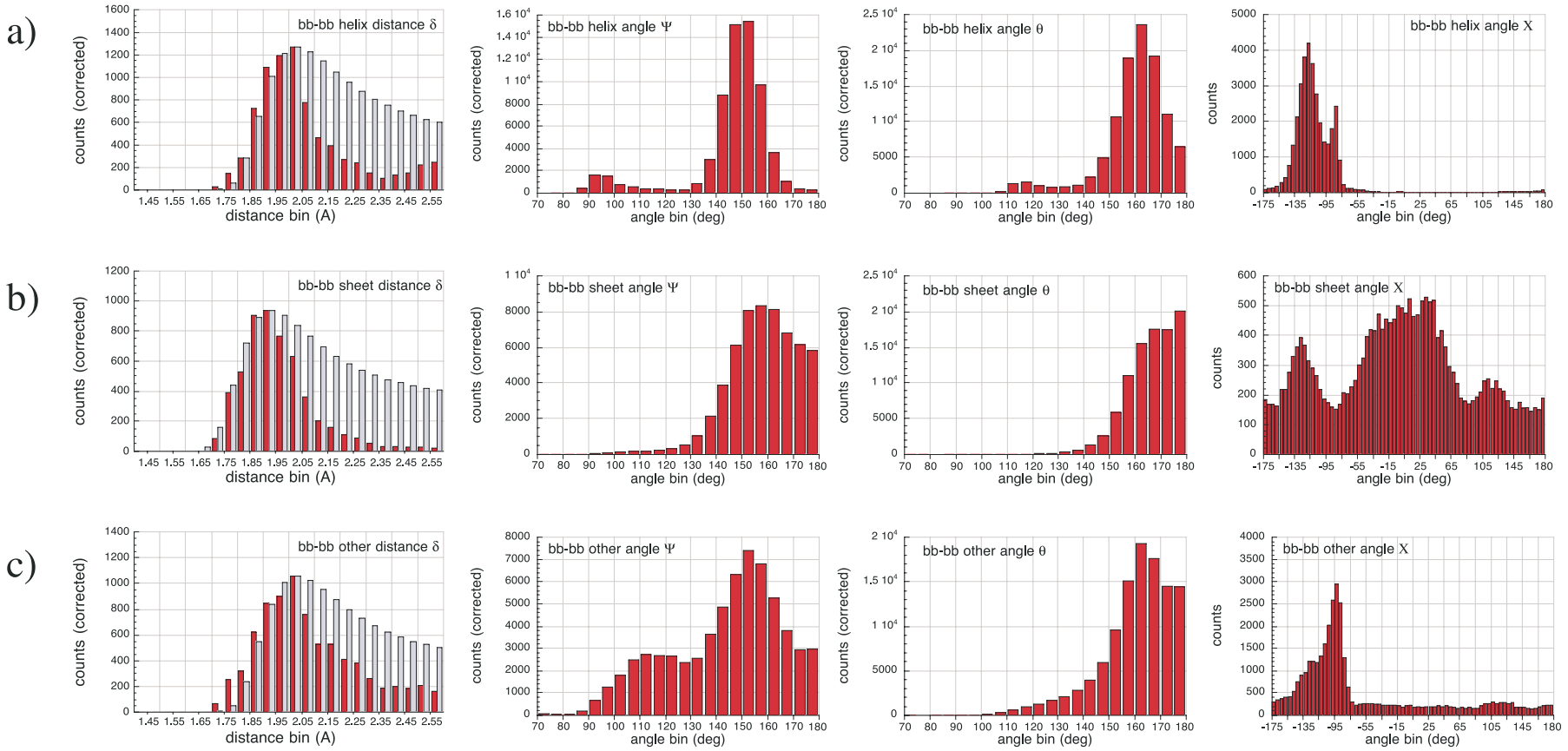

Figure 2 (*continued on next page*)

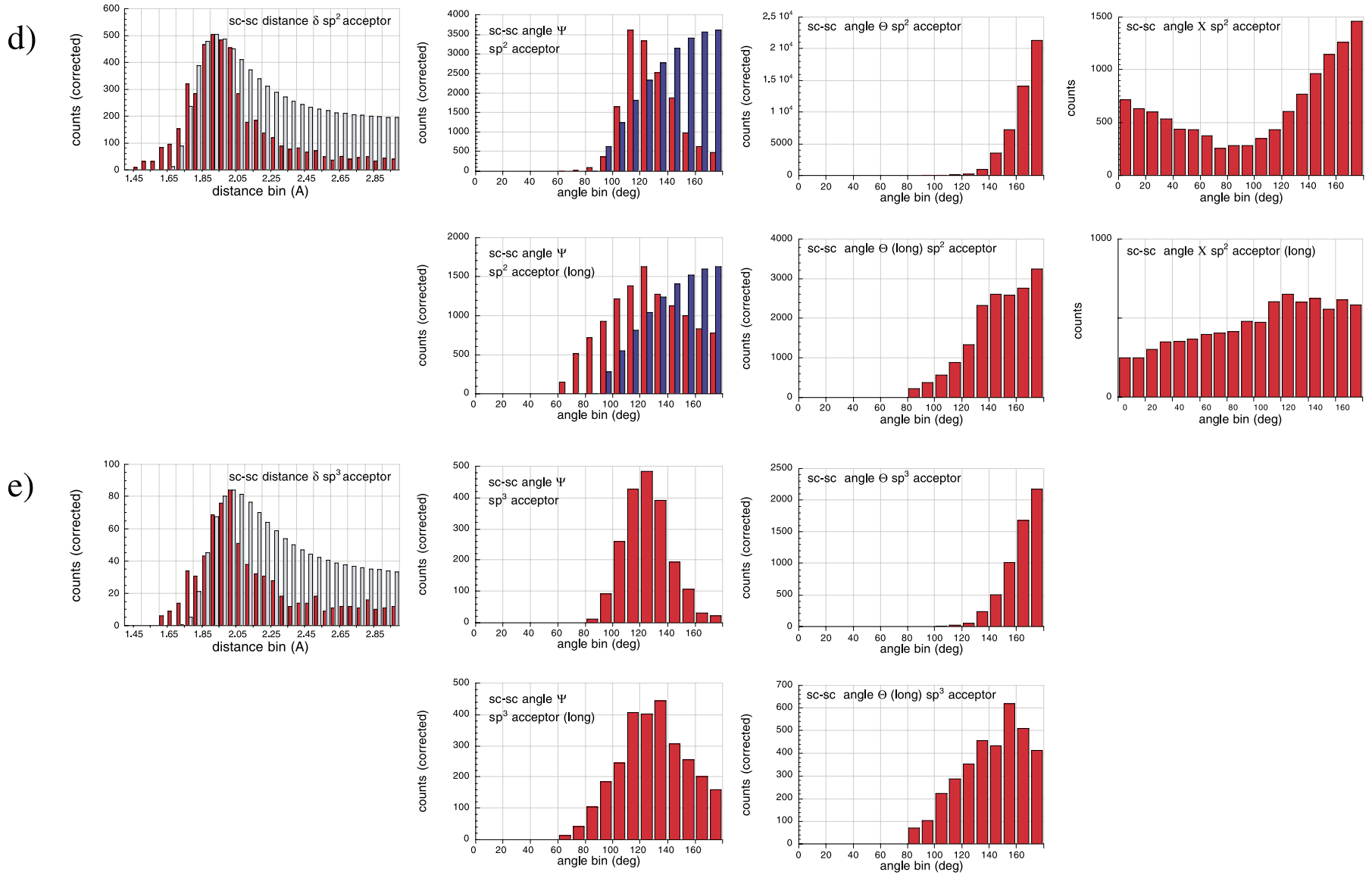

Figure 2 (*legend opposite*)

and backbone–backbone hydrogen bonds in β-sheets, and hardly any hydrogen bonds shorter than 1.6 Å. The angle at the hydrogen atom shows a clear preference for linearity for short side-chain–side-chain hydrogen bonds, whereas for backbone–backbone hydrogen bonds the distribution is shifted to lower angles, presumably caused by steric constraints of hydrogen bonds in regular secondary structure elements. Differences between side-chain–side-chain and backbone–backbone hydrogen bonds are also observed for the angle at the acceptor, with maxima at around 120° for side-chain–side-chain hydrogen bonds and shifted to higher angles for the backbone–backbone distributions. The differences between the different hybridization states at the acceptor for side-chain–side-chain hydrogen bonds are small, with a slightly sharper distribution (but not shifted to significantly lower angles) for the sp^3 case. A larger difference is seen comparing the geometries in short and long hydrogen bonds, with broader distributions found in the latter case. The dihedral angle shows fairly flat distributions except for a well-defined peak in the case of helical hydrogen bonds and other non-β backbone–backbone hydrogen bonds (mainly involving turns).

Our hydrogen bonding potential consists of a distance-dependent energy term ($E(\delta_{HA})$) and three angular dependent energy components ($E(\Theta)$ dependent on the angle at the hydrogen, $E(\Psi)$ dependent on the angle at the acceptor atom, and $E(X)$ dependent on the dihedral angle in hydrogen bonds involving an sp^2 hybridized acceptor) derived from the logarithm of the probability distributions found in the crystal structure analysis (see Methods). The total hydrogen bond energy (E_{HB}) was taken to be a linear combination of the four terms:

$$E_{HB} = W_{HB}[E(\delta_{HA}) + E(\Theta) + E(\Psi) + E(X)] \quad (1)$$

where W_{HB} is the relative weight of the hydrogen bonding term with respect to the other terms of the energy function (see Methods). The addition of the different energy terms assumes an independence of the geometric parameters, which appears to be a reasonable approximation in our dataset. An exception is the distance dependence of the angular distributions for side-chain–side-chain hydrogen bonds, which we approximate using two different distance ranges for collecting angular statistics. The observed acceptor angle for side-chain–side-chain hydrogen bonds differs significantly from the cosine-dependence with a maximum at linear angles used in other geometry-dependent hydrogen bonding potentials (see Figure 2(d)). Other differences are the replacement of the often used 10–12 potential for the distance-dependence by the more sharply peaked database-derived term (see Figure 2), as well as the inclusion of explicit hydrogen atoms for determining the hydrogen bonding geometry.

Testing of the hydrogen bonding function

It is not trivial to demonstrate that a new description of a contribution to macromolecular energetics is an improvement over previous descriptions because the individual components of the free energy cannot readily be measured independently in experiments. Here, we use a number of different tests to compare our new treatment to previous representations. The first two tests are based on the assumption that the substitution of the sequences of monomeric proteins and interfaces of protein–protein complexes with non-native amino acids on average produces an increase in free energy over the naturally occurring sequence. This assumption is consistent with extensive mutational data that show that amino acid replacements are far more often destabilizing rather than stabilizing. The third and fourth tests are based on the assumption that native protein structures and protein–protein interfaces are lower in free energy than the vast majority of non-native conformations.[38] While it is not necessary that the individual contributions to the free energy (such as the electrostatic component) all favor the native structure, it is plausible that given several

Figure 2. Distribution of geometric hydrogen bonding parameters obtained from 698 protein crystal structures. (a) Backbone–backbone hydrogen bonds occurring in α-helices (for secondary structure classification see Methods). (b) Backbone–backbone hydrogen bonds occurring in β-sheets. (c) All other backbone–backbone hydrogen bonds. (d) Side-chain–side-chain hydrogen bonds involving an sp^2 hybridized acceptor. (e) Side-chain–side-chain hydrogen bonds involving an sp^3 hybridized acceptor. The X distributions are not shown, as they are uniform. Raw counts were corrected for the different volume elements encompassed by the bins (angular correction: sin(angle) except for the X angle; distance correction: (distance)^2). Distributions shown are (indicated as label in each graph): hydrogen-acceptor distance δ_{HA}, red bars, light blue bars show the comparison with a standard 10–12 potential; angle Θ at the hydrogen; angle Ψ at the acceptor (red bars, blue bars in the plot for sp^2 hybridized acceptors show the comparison with the angular dependency of a dipole–dipole interaction assuming a 180° angle at the hydrogen and planarity of the hydrogen bond); dihedral angle X. Side-chain–side-chain angular distributions were collected for two different distance ranges as explained in the text. The Θ and Ψ angular distributions for helical backbone–backbone hydrogen bonds show side peaks at angles lower than 120° which most likely result from 3_{10} conformations at helix termini or other $i = i, i+3$ interactions. These conformations also cause the small peak in the distance distribution at distances larger than 2.4 Å. The Ψ angular distribution for backbone–backbone hydrogen bonds classified as other has a shoulder around 110° probably resulting from strained $i, i+2$ interactions that could be omitted in the potential.

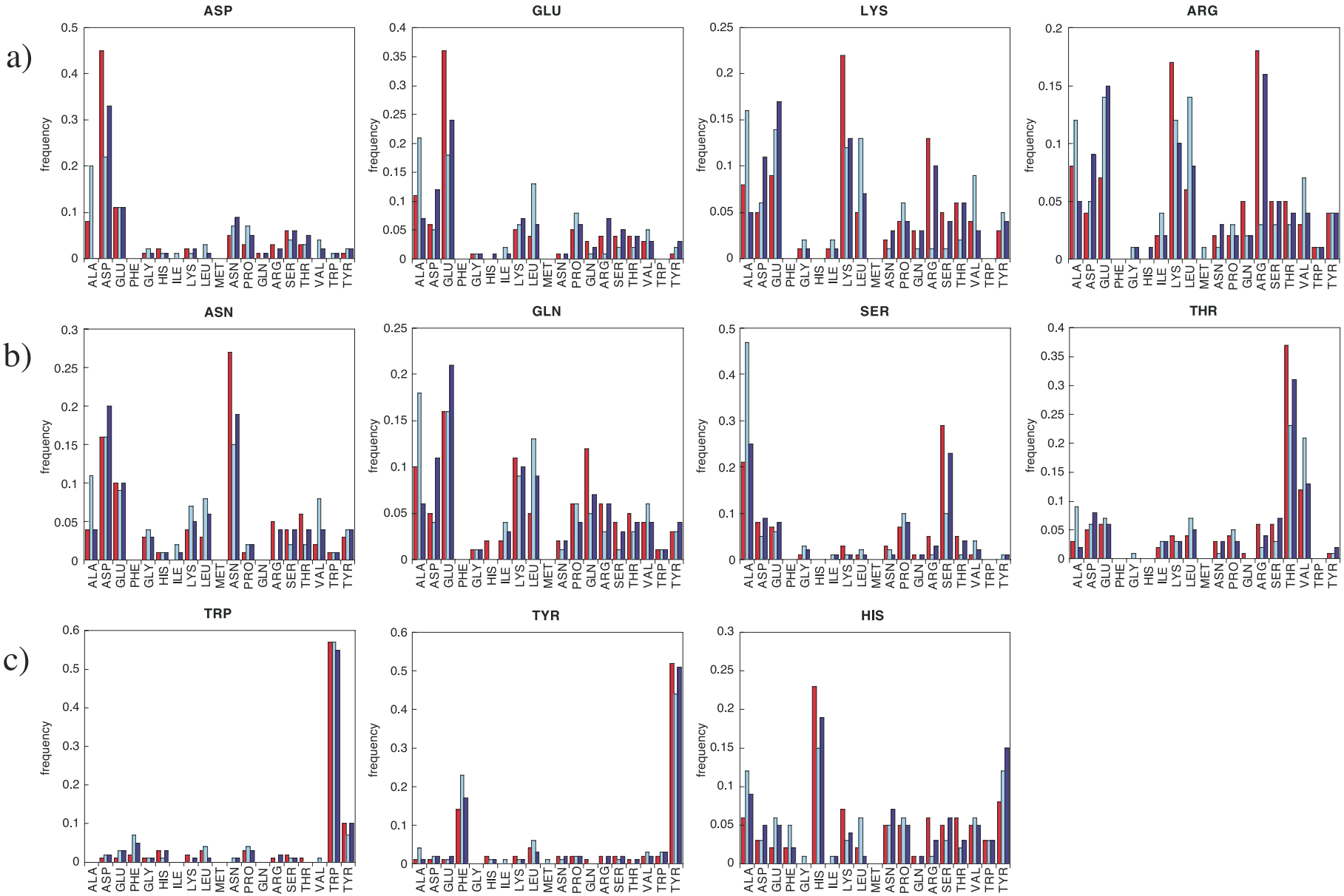

Figure 3 (*legend opposite*)

alternative models of electrostatic interactions, the one which most favors the native sequence and structure is the most accurate.

Prediction of amino acid identity in monomeric proteins (test 1)

The first test is to compare the recovery of the naturally occurring amino acids in protein design calculations using the new hydrogen bonding potential or a Coulomb treatment of electrostatics. In these calculations, side-chains at each sequence position in a set of proteins were substituted one-by-one by all amino acids in all the rotamer conformations in the Dunbrack backbone-dependent library (see Methods).[39] For each of a total of 7308 sequence positions, the free energy of all rotamers of all amino acids is determined, and the lowest free energy amino acid is selected. For each amino acid type, Figure 3 shows a substitution profile, depicting how often the native amino acid was found to be energetically most favorable, and how often each of the other 18 amino acids (cysteine residues were excluded because potential disulfide bonds were not modeled) was chosen to be the most favorable replacement. Three different energy functions were used in creating these substitution profiles to pinpoint the influence of the representation of hydrogen bonds and electrostatic interactions to the total free energy: (1) inclusion of all energy terms (red bars) (see Methods for the complete free energy function and parameterization); (2) exclusion of the hydrogen bonding contribution (light blue bars); and (3) exclusion of the hydrogen bonding term and representation of electrostatic and hydrogen bonding interactions instead by a Coulomb potential with a linear distance-dependent dielectric constant and a weight adjusted to approximately match the magnitude of the hydrogen bonding term used in (1). Clear differences between the three energy functions are observed for the charged (D, E, R, K) and polar (N, Q, T, S) amino acids (Figure 3(a) and (b)). In all cases, the inclusion of the new hydrogen bonding term is useful in discriminating the native amino acid type from others. For all amino acids except glutamine the native amino acid is predicted with the highest frequency (for a sequence position with a native glutamine residue, glutamate is picked with a higher probability). Representing hydrogen bonding interactions with a Coulomb term using a linear dielectric constant gives worse results for all amino acids, in some cases selecting a non-native amino acid type with the highest frequency. Particularly dramatic examples are glutamate, lysine, and serine, where alanine is preferred frequently if hydrogen bonds and Coulomb terms are excluded, an effect that can only partially be rescued by representing hydrogen bonding with a strong Coulomb term alone.

For the polar aromatic amino acids (W, Y, H; Figure 3(c)), the differences between the three energy functions are small. Presumably the selection is dominated by the packing interactions of the large aromatic ring in these cases, although using the hydrogen bonding potential improves the discrimination of tyrosine from phenylalanine, and histidine from tyrosine.

Although, as shown in Figure 3, the hydrogen bonding potential provides a significant improvement of the recognition of native amino acids for polar and charged residues, the overall prediction accuracy for these residue classes is worse than for the hydrophobic amino acids A, I, L, V, and F as well as G and P (for substitution profiles of the hydrophobic amino acids see the distributions for interfaces in Figure 4 which were similar to those for monomeric proteins). This is partially due to a limitation inherent in our test, as one assumption in our optimization procedure is that the native amino acid type is lowest in free energy at each sequence position. This assumption does not necessarily hold true for all sequence positions (as the choice of a certain amino acid type at a certain position might also have been optimized for function and solubility), and will be more valid for amino acids in the protein core that are presumably selected for stability. The positions of polar and charged residues in native structures are likely to be determined not only by energetic considerations tested by our procedure but also by functional and solubility constraints (this might explain the particularly poor predictions for histidine residues, which are often found in active sites due to their pK_a value in the experimental pH range; also, all histidine residues are modeled as neutral in our procedure). Moreover, in particular for the long polar amino acids, limited conformational sampling using the rotamer approach might make it difficult to reach optimal hydrogen bonding geometries required for selecting the native amino acid (hydrophobic packing might be less demanding). Also, since our free energy function does not contain explicit penalties for cavities (apart from the loss of favorable packing interactions) or unsatisfied hydrogen bonding donors or acceptors, replacement of a larger side-chain by a smaller residue may not be sufficiently penalized.

Figure 3. Recovery of native sequences in single domain proteins. For all sequence positions containing a specific amino acid type, the bars show for all amino acids except cysteine how often each amino acid type is found to be energetically most favorable. Substitution profiles are calculated with different energy functions: red bars, complete energy function; light blue bars, energy function without the hydrogen bonding term; dark blue bars, energy function without the hydrogen bonding term, but scaling the Coulomb term to be of similar magnitude. (a) Charged amino acids; (b) polar amino acids; (c) polar aromatic amino acids.

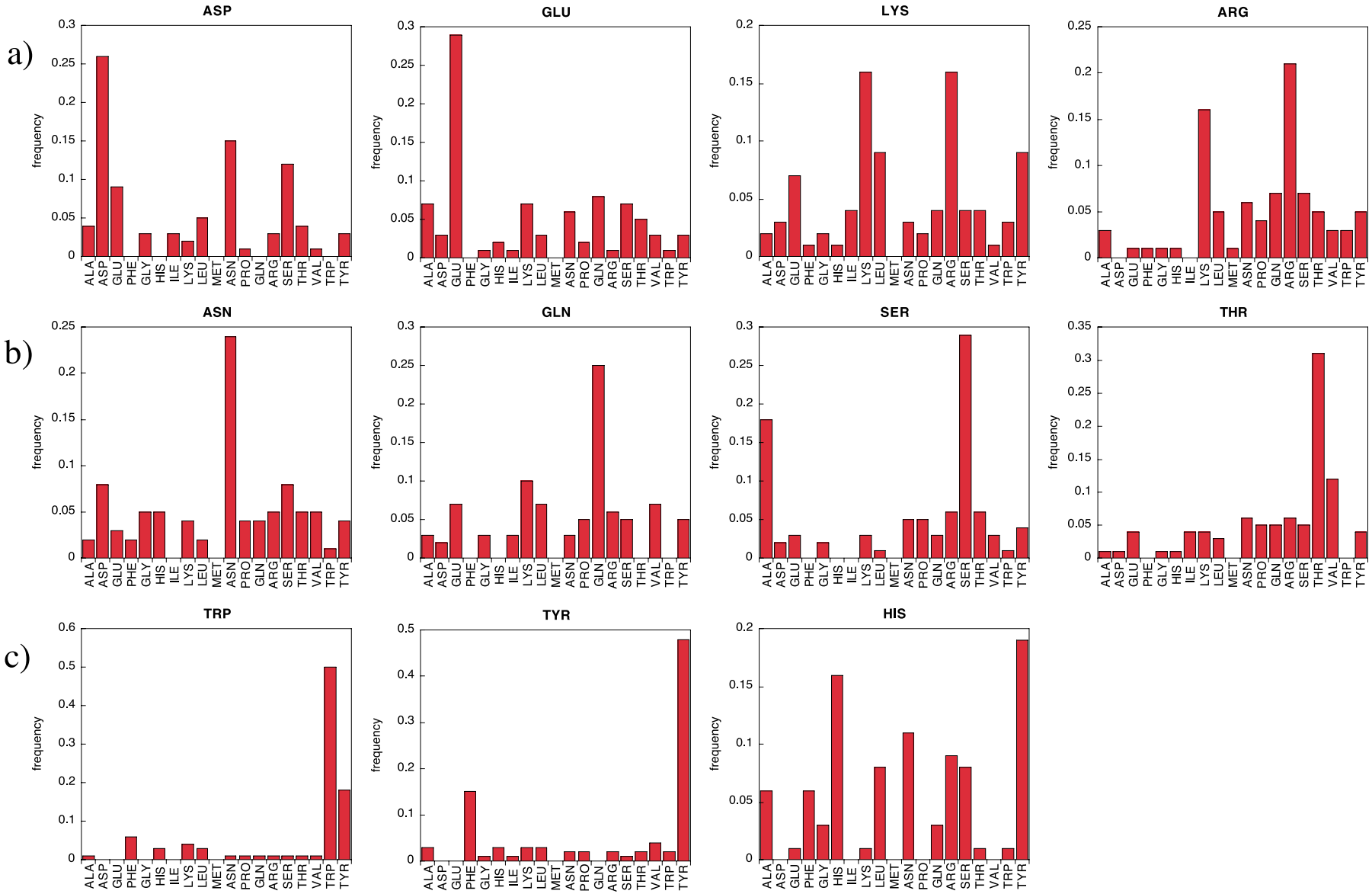

Figure 4 (*legend opposite*)

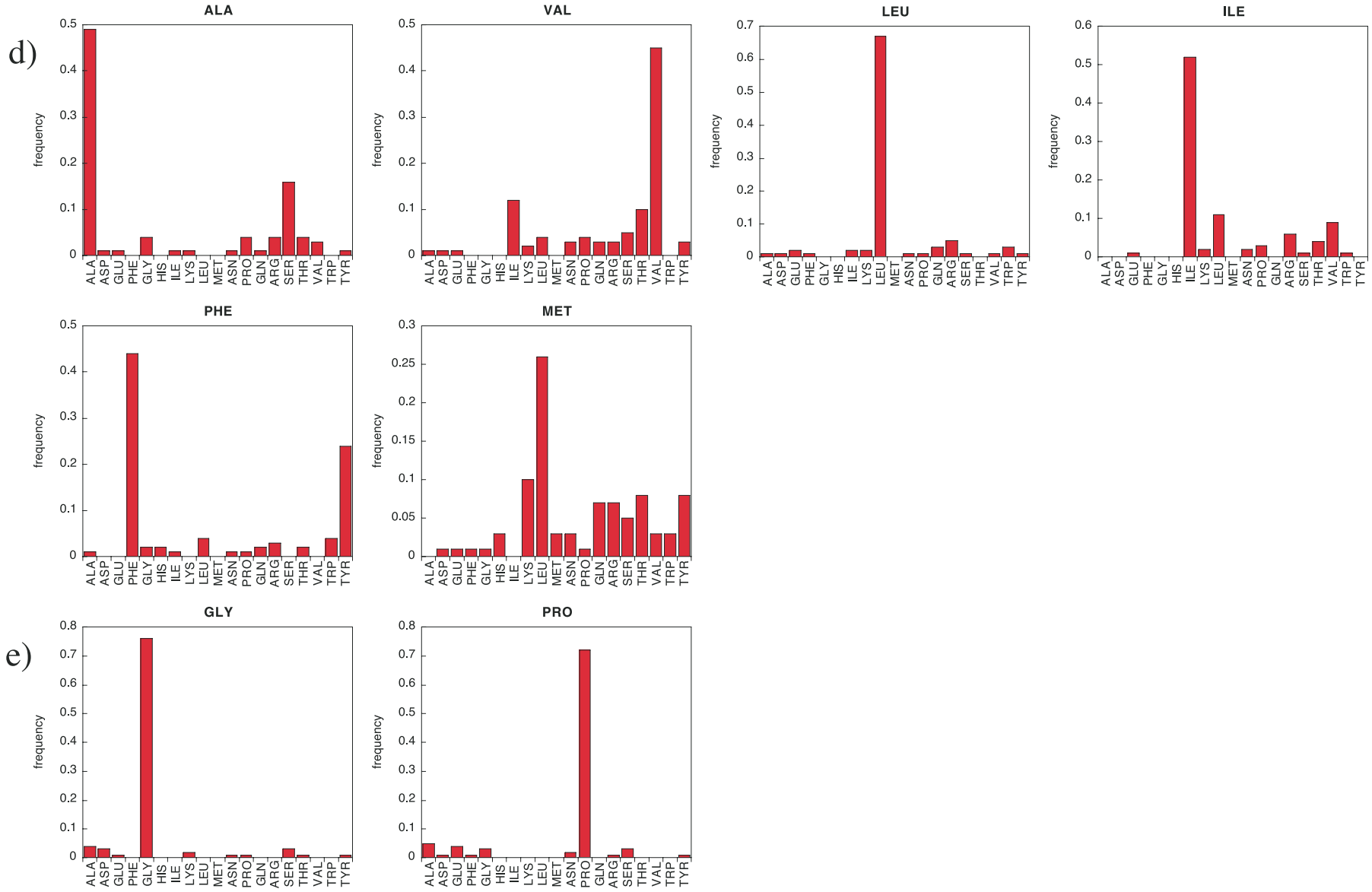

Figure 4. Application of the energy function to the prediction of sequences in protein–protein interfaces. For all sequence positions containing a specific amino acid type, the bars show for all amino acids except cysteine how often each amino acid type is found to be energetically most favorable. Substitution profiles are calculated with the complete energy function including the hydrogen bonding term. (a) Charged amino acids; (b) polar amino acids; (c) polar aromatic amino acids; (d) hydrophobic amino acids; (e) glycine and proline.

Table 1. Native (Zn) and native repacked (Znr) *Z*-scores for the monomeric single domain decoy set (SS-secondary structure class: α-helix, β-sheet) for the following energetic contributions: side-chain–backbone hydrogen bonds (HB sc–bb), side-chain–side-chain hydrogen bonds (HB sc–sc), backbone–backbone hydrogen bonds (HB bb–bb)

PDB code	SS	HB sc–bb		HB sc–sc		HB bb–bb		Combined HB score	
		Zn	Znr	Zn	Znr	Zn	Znr	Zn	Znr
1a32	α	1.07	0.25	5.98	0.39	3.04	3.04	4.59	3.12
1ail	α	−1.75	−0.97	5.79	1.18	8.32	8.32	8.22	7.56
1am3	α	1.79	2.15	0.84	−0.41	1.50	1.50	2.39	2.40
1cc5	α	0.29	3.59	−0.30	−0.30	−1.72	−1.72	−1.53	−0.16
1cei	α	0.17	1.54	6.50	−0.51	4.69	4.69	5.80	4.89
1hyp	α	2.56	1.39	−0.28	−0.28	2.35	2.35	3.30	2.85
1lfb	α	−1.11	−0.26	1.09	0.26	0.73	0.73	0.45	0.59
1mzm	α	0.44	−0.38	−0.31	−0.31	2.79	2.79	2.79	2.51
1r69	α	2.22	1.94	3.75	1.85	1.40	1.40	3.36	2.58
1utg	αβ	0.55	−0.97	2.75	−0.41	4.18	4.18	4.80	3.54
1ctf	αβ	2.43	−1.37	−0.43	−0.43	5.12	5.12	6.01	4.33
1dol	αβ	−0.22	2.84	−0.42	−0.42	1.08	1.08	0.57	2.65
1orc	αβ	−0.93	−0.16	4.68	2.77	3.87	3.87	3.57	3.45
1pgx	αβ	0.19	2.21	1.80	−0.39	5.28	5.28	4.47	5.33
1ptq	αβ	5.77	4.62	2.47	1.05	−0.51	−0.51	3.18	2.19
1tif	αβ	1.05	0.79	7.03	−0.48	7.09	7.09	7.09	5.36
1vcc	αβ	2.42	2.85	4.00	−0.43	3.66	3.66	5.50	4.76
2fxb	αβ	−0.07	1.09	9.53	9.13	−0.11	−0.11	2.48	1.88
5icb	αβ	2.11	3.07	7.04	−0.41	3.80	3.80	5.61	5.03
1bq9	β	2.30	−1.26	3.69	−0.30	5.09	5.09	6.37	2.69
1csp	β	−1.25	−0.38	−0.26	−0.26	4.67	4.67	2.43	3.15
1msi	β	2.50	0.95	2.88	−0.29	2.15	2.15	3.82	2.40
1tuc	β	1.51	2.99	2.50	1.07	3.45	3.45	4.38	5.16
1vif	β	1.88	−1.10	3.31	0.30	3.42	3.42	4.47	2.21
5pti	β	−0.60	−0.25	13.31	−0.34	4.29	4.29	6.62	3.11
Mean		1.01	1.01	3.48	0.48	3.20	3.20	4.03	3.34
Stdev		1.66	1.72	3.46	1.99	2.33	2.33	2.18	1.63

Combined HB score, generalized linear model fit using the three hydrogen bond scores. Stdev, standard deviation from mean value. Successful discrimination is defined as a *Z*-score >1.

Prediction of amino acid identities in protein–protein complexes (test 2)

Electrostatic interactions are thought to be particularly relevant in molecular recognition. Experimental as well as theoretical studies have highlighted the role of electrostatic interactions in determining the rate of association of protein–protein complexes,[40] as well as for recognition specificity.[41] While the role of hydrogen bonds and charge–charge interactions for specificity appears well established, their importance for the stability of protein–protein complexes is somewhat under debate.[42] The energy function described above was applied unchanged to a different dataset of 50 binary protein–protein complexes. The rationale was twofold. First, we wanted to see whether the energy function is also useful for prediction of specificity in protein recognition by testing whether it performs well in recognizing the native amino acid sequence in protein interfaces. Second, as the energy function was parameterized on the monomeric dataset, we used this second set as an independent test of performance. The substitution profiles for a total of 2986 sequence positions located in protein interfaces (Figure 4) show that, as in the case of the monomeric protein set, for most positions the most strongly predicted amino acid is the naturally occurring residue. This suggests that the potential function should be generally applicable to the prediction of specificity in protein–protein interactions for whole families of protein sequences where a structure is available for at least one homologous protein–protein complex. If, given the structure of a protein–protein interface and the sequence of one partner, the sequences of potentially interacting partner can be predicted, the method can be used to computationally subdivide the sequences into interacting and non-interacting pairs as a guide for further investigations.

Decoy discrimination for monomeric single domain proteins (test 3)

Next, we investigated the difference in the hydrogen bond energy of native or native-repacked structures (native backbone coordinates with modeled side-chains from a rotamer library) and alternative conformations (decoys) generated using the ROSETTA *ab initio* structure prediction method.[43,44] We use the normalized energy gap (*Z*-score: energy difference between the native or near-native structure and the mean of the decoy distribution, divided by the standard deviation) to quantify the signal-to-noise ratio for decoy discrimination.[33] Table 1 shows the *Z*-scores for

Table 2. Low rmsd (Zlrms) *Z*-scores for the monomeric single domain decoy set (SS, secondary structure class: α-helix, β-sheet) for the following energetic contributions: Coulomb electrostatics with a linear distance-dependent dielectric constant (Coulomb), side-chain–backbone hydrogen bonds (HB sc–bb), side-chain–side-chain hydrogen bonds (HB sc–sc), backbone–backbone hydrogen bonds (HB bb–bb)

PDB code	SS	Coulomb (Zlrms)		HB sc–bb (Zlrms)		HB sc–sc (Zlrms)		HB bb–bb (Zlrms)		Combined HB score (Zlrms)		Combined HB + vdW score (Zlrms)	
		Dec.	PN.	Dec.	PN.	Dec.	PN.	Dec.	PN.	Dec.	PN.	Dec.	PN.
1a32	α	0.20	0.14	−0.49	−0.46	0.05	0.09	1.16	0.59	1.14	0.56	1.02	0.73
1am3	α	−0.39	−0.26	0.00	0.14	−0.26	−0.19	0.43	0.46	0.42	0.47	0.69	0.84
1bw6	α	−0.27	0.08	0.20	0.11	−0.04	−0.03	0.51	0.10	0.53	0.12	0.66	0.52
1gab	α	0.58	0.48	0.54	0.41	0.39	0.33	0.55	0.03	0.60	0.12	1.13	0.69
1kjs	α	−0.01	0.13	−0.16	−0.13	0.01	0.10	0.32	0.31	0.31	0.29	0.69	0.80
1mzm	α	−0.04	0.39	0.01	−0.17	−0.02	0.05	0.55	1.92	0.56	1.90	0.46	2.06
1nkl	α	−0.13	0.41	−0.17	−0.04	0.02	0.21	0.14	2.25	0.14	2.25	0.20	2.10
1nre	α	1.05	0.67	−0.81	−0.75	0.17	−0.02	1.27	1.81	1.25	1.80	0.90	1.67
1pou	α	0.23	0.47	−0.12	−0.08	0.40	0.15	0.22	1.69	0.23	1.71	0.37	1.55
1r69	α	0.80	1.13	0.41	0.68	0.13	0.07	0.02	1.83	0.06	1.91	0.98	2.26
1res	α	−0.46	−0.43	0.07	0.11	0.08	0.06	0.32	0.10	0.33	0.13	0.43	0.48
1uba	α	−0.39	−0.07	0.06	0.30	0.11	0.04	0.24	−0.17	0.24	−0.09	0.13	0.02
1uxd	α	0.26	0.31	−0.43	−0.45	−0.05	0.00	1.14	0.43	1.13	0.36	1.25	0.84
2ezh	α	0.15	−0.08	−0.26	0.45	0.15	0.00	0.53	1.61	0.52	1.66	0.48	1.38
2pdd	α	0.31	0.33	0.20	−0.15	0.43	1.09	0.55	0.37	0.59	0.44	0.89	1.19
1aa3	αβ	0.84	0.36	0.28	0.32	−0.06	0.02	−0.34	0.61	−0.32	0.68	0.42	0.85
1afi	αβ	0.84	0.62	−0.09	0.30	0.31	0.22	0.63	2.26	0.65	2.25	1.07	2.30
1ctf	αβ	0.53	1.20	0.15	−0.04	−0.14	−0.15	−0.13	2.75	−0.13	2.72	0.41	2.58
1pgx	αβ	0.77	1.22	0.27	0.98	−0.03	−0.26	0.74	2.84	0.76	2.85	0.59	1.92
2fow	αβ	0.35	0.15	−0.07	−0.07	−0.09	−0.26	−0.51	1.51	−0.52	1.48	−0.14	1.49
2ptl	αβ	0.52	0.73	−0.21	−0.03	−0.18	0.24	0.32	1.56	0.29	1.58	0.89	1.64
1sro	β	0.97	2.19	0.19	0.30	0.20	0.31	0.38	0.26	0.41	0.40	0.47	1.48
1vif	β	2.03	1.86	0.23	−0.03	0.13	0.22	1.55	1.52	1.57	1.47	1.58	1.29
Mean		0.38	0.52	−0.01	0.07	0.08	0.10	0.46	1.16	0.47	1.18	0.68	1.33
Stdev		0.58	0.64	0.31	0.38	0.18	0.27	0.49	0.93	0.48	0.90	0.39	0.66

Combined HB score: generalized linear model fit using the three hydrogen bond scores only. Combined HB + vdW score, generalized linear model fit using the three hydrogen bond scores and the van der Waals scores. Dec. and PN. denote *Z*-scores for the original *ab initio* decoy set and the perturbed-native set (see Methods). Stdev, standard deviation from mean value. Successful discrimination is defined as a *Z*-score >1.

the hydrogen bonding term of monomeric single-domain proteins, split into side-chain–side-chain, side-chain–backbone (using side-chain–side-chain statistics, see Methods) and backbone–backbone contributions. The backbone–backbone contribution is found to be the best discriminator of the native structure *versus* the decoys (Zn column, Table 1). It alone is capable of successful discrimination for 21 out of the 25 structures (a failure is defined as a *Z*-score <1). If all three hydrogen bonding terms are combined using a generalized linear model fit, 22 out of the 25 structures are successfully discriminated. Interestingly, one of the failures contains a heme cofactor (1cc5) not taken into account in either decoy creation or in the scoring of decoys and the native structures.

We also computed *Z*-scores for native structures with all side-chains repacked (native-repacked *z*-scores) using the same potential as for the decoys. Similar values for side-chain–backbone contributions are obtained when the decoys are compared to either the native structures (Zn column) or the native-repacked structures (Znr column).[45,46] However, there is a noticeable drop in the side-chain–side-chain *Z*-score for native *versus* native repacked structures, indicating that for some proteins the repacking procedure does not reproduce the precise side-chain–side-chain geometries observed in the native structure. This effect can be due to limitations in rotamer sampling in particular for long polar side-chains, or, alternatively, the energy function favors the exposure of surface side-chains to solvent over the formation of side-chain–side-chain hydrogen bonds.

A further, more challenging test is to be able to rank different decoy conformations by their closeness to the native structure, often represented as a correlation between the free energy score and the root-mean-square deviation (rmsd) of the backbone atoms to the experimentally determined native protein structure (the decoy discrimination problem).[10,47–49] Table 2 shows the *Z*-scores for discriminating near-native decoys from all other decoy conformations (Zlrms). Near-native decoys are defined as the lowest 5% of the rmsd distribution in the decoys; the cutoff rmsd value varies between 1.10 Å and 2.84 Å for different proteins in the set. Discrimination is poor in the set, with successful discrimination for only four out of 23 proteins both using all hydrogen bonding terms in combination or just considering the backbone–backbone contribution alone (Table 2, Dec. columns).

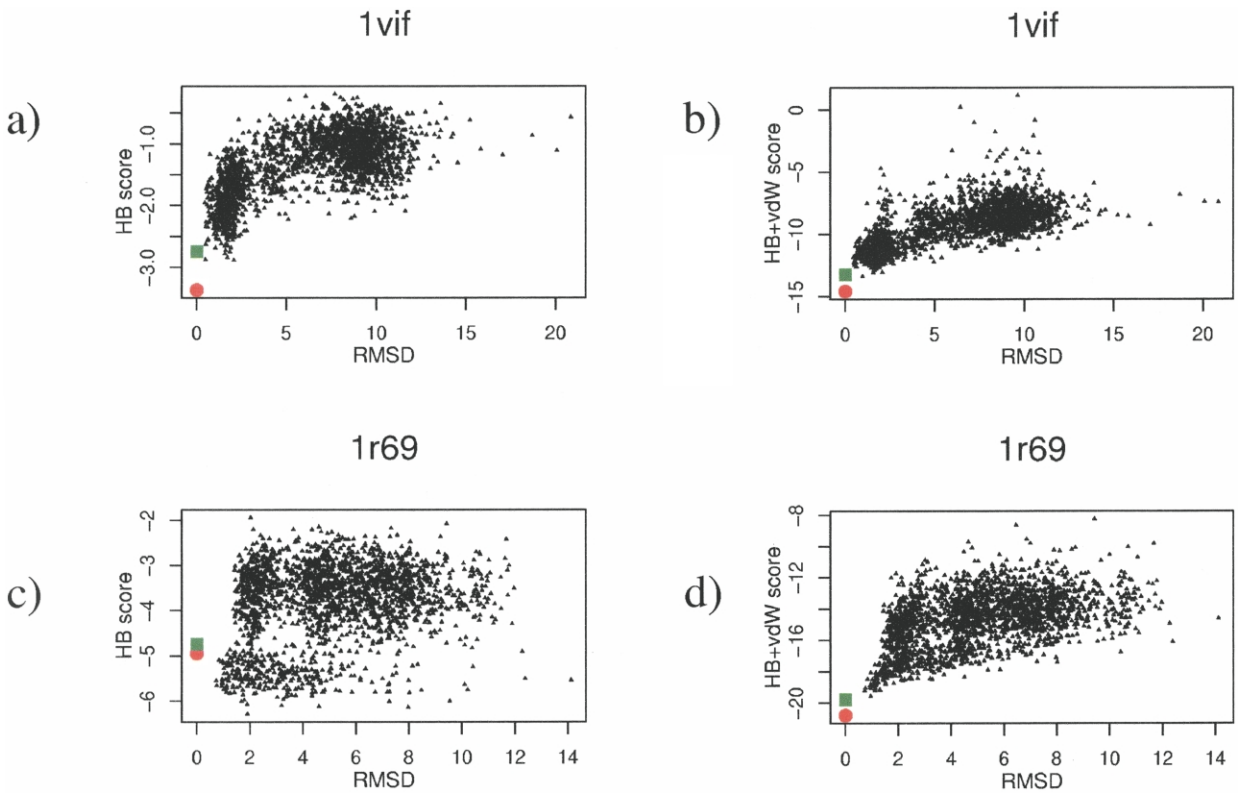

Figure 5. Scatter plots of the combined hydrogen bonding score alone (a) and (c) and in combination with the van der Waals score (b) and (d) *versus* decoy C^{α} rmsd from the native structure for selected monomeric proteins from the perturbed native data set. a) and b) Structure 1vif; c) and d) structure 1r69. Native structures are shown with red circles, native structures with the side-chains modeled using our rotamer repacking protocol are shown with green squares and decoys are shown with black triangles.

Even the addition of a van der Waals term (see Methods) does not improve discrimination significantly (five out of 23 structures are successfully discriminated). As for the recovery of native amino acid types in monomeric proteins (Figure 2), a representation of electrostatic interactions purely using a Coulomb term with a linear distance-dependent dielectric performs worse than the hydrogen bonding term, discriminating low rmsd decoys for only two out of 23 structures in the set (Table 2). It should, however, be noted that the overall *Z*-scores for discriminating near-native from other decoys are low and do not justify unequivocal conclusions comparing the performance of the different energy terms.

The hydrogen bonding potential used for decoy discrimination is relatively short-ranged. Thus, if there are few structures close enough to the native structure to detect native-like hydrogen bonding interactions, discrimination is expected to be poor. To test this hypothesis, we repeated the discrimination test for near-native structures with a different decoy set created by perturbations starting from the native structure, containing many structures with rmsd values to the native structure in the 1–3 Å range (Table 2, PN columns). While there is still no significant signal using a Coulomb term, side-chain–backbone or side-chain–side-chain hydrogen bonds alone, 12 out of 23 structures can now be successfully discriminated using the backbone–backbone hydrogen bonding term, with slight improvements by combining all hydrogen bonding scores and adding in van-der-Waals interactions (Table 2, PN columns).

In cases of successful discrimination of low rmsd decoys in the perturbed native set, scatter plots show a correlation between the hydrogen bonding term (alone and in combination with the van-der-Waals scores) with the rmsd to the native structure. As suggested by the data in Table 2, in most of these cases the hydrogen bonding term alone provides the main part of the signal; an example of this is the protein 1vif shown in Figure 5(a) and (b). However, in a few cases the addition of the van-der-Waals term significantly improves discrimination (Figure 5(c) and (d)).

Decoy discrimination in protein docking (test 4)

As mentioned above, electrostatic interactions have been shown to be important in protein–protein recognition. Therefore, as the fourth and final test of our hydrogen bond function, we assessed its ability to discriminate native and near-native binary protein–protein complex structures from docked arrangements with a range of rmsd values (rmsd values are obtained by computing the overall C^{α} rmsd in the complex) (the protein docking

Table 3. Native (Zn) and native repacked (Znr) *Z*-scores for the protein–protein complex decoy sets for the following energetic contributions: side-chain–backbone hydrogen bonds (HB sc–bb), side-chain–side-chain hydrogen bonds (HB sc–sc), backbone–backbone hydrogen bonds (HB bb–bb): (A) Antibody/antigen (ab) decoy set; (B) non-antibody/antigen (nab) decoy set

PDB code		HB sc–bb		HB sc–sc		HB bb–bb		Combined HB score	
		Zn	Znr	Zn	Znr	Zn	Znr	Zn	Znr
(A)									
1a2y	ab	4.07	3.08	3.99	2.96	0.92	0.92	2.47	1.98
1cz8	ab	0.16	−0.25	0.44	0.07	6.12	6.12	6.04	5.99
1dqj	ab	1.06	2.21	1.71	2.29	5.46	5.46	5.80	5.59
1e6j	ab	1.79	2.78	1.24	2.76	6.28	6.28	5.28	6.17
1egj	ab	1.74	0.12	1.76	0.29	−0.22	−0.22	0.72	0.12
1eo8	ab	−0.65	0.92	0.28	1.95	−0.19	−0.19	0.96	2.01
1fdl	ab	3.02	2.10	2.93	2.26	1.89	1.89	2.66	2.56
1fj1	ab	2.84	2.97	2.61	2.90	−0.13	−0.13	1.51	1.90
1g7h	ab	2.94	1.32	2.74	1.14	2.88	2.88	3.38	2.72
1ic4	ab	1.68	1.71	2.05	1.75	5.06	5.06	5.29	4.86
1jhl	ab	−1.58	−0.40	−1.52	−0.12	3.96	3.96	2.31	3.18
1jrh	ab	4.07	2.16	4.41	2.51	8.08	8.08	8.56	7.75
1mlc	ab	−0.55	2.57	0.11	2.93	2.40	2.40	2.33	3.40
1nca	ab	1.88	4.54	1.58	4.10	−0.23	−0.23	0.50	1.91
1nsn	ab	−0.77	0.13	−0.62	0.13	−0.16	−0.16	−0.36	−0.01
1osp	ab	1.93	1.32	1.69	1.44	8.46	8.46	7.82	7.96
1qfu	ab	−0.86	4.14	−0.30	4.81	−0.26	−0.26	0.01	2.11
1wej	ab	1.07	1.53	1.29	1.50	−0.20	−0.20	0.79	0.69
Mean		1.32	1.83	1.47	1.98	2.78	2.78	3.12	3.38
Stdev		1.72	1.41	1.56	1.37	3.11	3.11	2.64	2.38
(B)									
1ACB	nab	−0.85	−0.88	−0.99	−0.87	12.70	12.70	11.33	11.42
1AVZ	nab	1.07	1.89	1.35	2.41	−0.16	−0.16	1.05	1.96
1brs	nab	5.74	5.09	6.33	4.80	−0.32	−0.32	3.43	2.32
1CHO	nab	−0.51	−0.32	−0.59	−0.13	12.79	12.79	12.06	12.25
1MDA	nab	−1.27	−1.22	−1.40	−1.37	−0.35	−0.35	−1.03	−1.02
1PPF	nab	−1.09	−0.27	−1.20	−0.34	9.82	9.82	8.77	9.07
1SPB	nab	7.00	5.65	6.32	5.44	13.85	13.85	14.06	13.90
1UGH	nab	3.95	6.64	3.63	6.24	−0.27	−0.27	2.34	4.21
2PCC	nab	−0.93	−0.63	−0.97	−0.62	−0.32	−0.32	−0.87	−0.62
2PTC	nab	3.43	2.44	3.17	2.51	5.70	5.70	6.18	6.00
1CSE	nab	2.14	0.56	1.94	0.52	9.64	9.64	9.16	8.66
1FIN	nab	5.01	5.40	4.93	5.36	−0.20	−0.20	3.65	3.99
2BTF	nab	1.70	2.49	2.19	2.68	3.54	3.54	4.18	4.40
Mean		1.95	2.06	1.90	2.05	5.11	5.11	5.72	5.89
Stdev		2.86	2.81	2.84	2.72	5.86	5.86	4.78	4.64

Combined HB score, generalized linear model fit using the three hydrogen bond scores. Stdev, standard deviation from mean value. Successful discrimination is defined as a *Z*-score >1.

problem). We used the backbone conformations of the protein partners as observed in the native complex structure. However, the side-chain conformations in the native complex structure were discarded, and modeled from our standard rotamer library during all docking steps. All decoys were created by rigid-body motions of the two proteins *versus* each other, incorporating extensive conformational rearrangement of the side-chains in the complex interface.[46,50] Tables 3 and 4 show the results of a *Z*-score analysis for protein–protein complexes as described for the single-domain monomeric proteins above. The protein–protein complex decoy set was divided into antibody-antigen (18 structures) and other complexes (13 structures). The hydrogen bonding model alone successfully discriminates the native structure for 23 out of the 31 protein–protein complexes studied. All three hydrogen bonding terms contribute to decoy discrimination, with successful predictions in about two-thirds of all cases for each component separately (Table 3).

Lastly, we tested whether the hydrogen bonding function also helps in discriminating near-native from high rsmd decoys (Table 4). In contrast to the results obtained for the single domain monomeric proteins (Table 2), for protein–protein complexes good discrimination is achieved using a combination of the three hydrogen bonding terms in about two-thirds of all structures studied. Again all three hydrogen bond components are significant contributors (the combined HB score discriminates

Table 4. Low rmsd (Zlrms) *Z*-scores for the protein–protein complex decoy sets for the following energetic contributions: Coulomb electrostatics with a linear distance-dependent dielectric constant (Coulomb), side-chain–backbone hydrogen bonds (HB sc–bb), side-chain–side-chain hydrogen bonds (HB sc–sc), backbone–backbone hydrogen bonds (HB bb–bb): (A) antibody/antigen (ab) decoy set; (B) non-antibody/antigen (nab) decoy set

PDB code		Coulomb (linear model) Zlrms	HB sc–bb Zlrms	HB sc–sc Zlrms	HB bb–bb Zlrms	Combined HB score Zlrms	Combined HB + vdW score Zlrms
(A)							
1a2y	ab	0.41	1.57	1.57	−0.27	1.28	1.19
1cz8	ab	1.53	0.60	0.63	1.99	1.66	1.71
1dqj	ab	1.42	1.97	1.90	−0.21	1.50	2.36
1e6j	ab	0.73	1.12	1.40	1.04	1.76	1.96
1egj	ab	0.63	1.27	1.42	−0.22	1.42	1.55
1eo8	ab	1.27	−0.67	−0.25	−0.20	0.48	1.04
1fdl	ab	−0.27	1.00	1.12	0.44	1.20	0.85
1fj1	ab	0.34	2.66	2.74	−0.14	2.58	2.72
1g7h	ab	0.80	1.69	1.72	1.88	1.99	1.24
1ic4	ab	1.61	1.15	1.33	2.50	1.97	2.19
1jhl	ab	0.13	−0.44	−0.31	−0.24	−0.11	0.18
1jrh	ab	1.64	1.35	1.47	1.56	1.81	1.42
1mlc	ab	1.13	0.47	0.82	0.95	1.45	1.78
1nca	ab	1.59	1.66	1.67	−0.09	1.39	1.86
1nsn	ab	0.83	−0.04	0.04	−0.16	0.14	0.13
1osp	ab	0.39	0.16	0.25	−0.13	0.31	0.83
1qfu	ab	0.80	1.32	1.72	0.32	2.15	2.18
1wej	ab	0.67	−0.31	−0.05	−0.20	0.28	0.07
Mean		0.87	0.92	1.07	0.49	1.29	1.40
Stdev		0.56	0.91	0.85	0.92	0.75	0.76
(B)							
1ACB	nab	0.67	−0.24	−0.36	3.84	2.14	2.13
1AVZ	nab	1.12	−0.24	−0.02	−0.16	0.24	0.65
1brs	nab	1.42	3.29	3.31	0.27	2.53	2.53
1CHO	nab	1.12	−0.34	−0.33	4.20	3.39	3.19
1MDA	nab	0.00	0.72	0.65	−0.11	0.27	1.55
1PPF	nab	0.78	−0.69	−0.67	16.45	10.61	7.23
1SPB	nab	2.64	1.05	1.09	6.77	5.29	4.23
1UGH	nab	1.11	1.87	1.86	−0.02	1.48	1.52
2PCC	nab	0.50	1.20	1.10	−0.32	0.55	0.46
2PTC	nab	0.85	0.97	0.96	3.80	3.23	2.61
1CSE	nab	1.36	0.63	0.76	3.93	3.32	2.94
1FIN	nab	0.98	0.14	0.25	−0.22	0.29	1.23
2BTF	nab	0.60	0.54	1.04	0.61	1.79	1.88
Mean		1.01	0.68	0.74	3.00	2.70	2.47
Stdev		0.62	1.07	1.06	4.67	2.71	1.70

Combined HB score, generalized linear model fit using the three hydrogen bond scores only. Combined HB + vdW score, generalized linear model fit using the three hydrogen bond scores and the van der Waals scores. Stdev, standard deviation from mean value. Successful discrimination is defined as a Z-score >1. Stdev, standard deviation of mean value. Successful discrimination is defined as a *Z*-score >1. The mean *Z*-score for the combined HB scores and combined HB + vdW scores in (B) is lower than the score for the bb–bb hydrogen bonds alone, but maximizes the number of protein structures that can be successfully discriminated (lower standard deviation).

22 out of 31 structures, whereas backbone–backbone hydrogen bonds alone only discriminate 11 structures). As seen before, a Coulomb term alone is substantially less effective than the combined hydrogen bonding terms, only discriminating 13 structures.

The chemical character of protein–protein interfaces is a combination of that seen on protein surfaces and in protein cores.[2] Thus we wanted to test whether the inclusion of a van der Waals term into the free energy function, describing non-polar packing interactions in the interface, could further improve the discrimination of docking decoys. The encouraging results obtained just using the hydrogen bonding terms of our energy function can be slightly improved when including the van der Waals component, leading to successful discrimination in 24 out of the 31 complexes. The slight improvement is only seen for non-antibody complexes. This behavior might be due to the known poorer shape complementarity and larger solvation in antibody–antigen complexes,[51] which was our original justification for splitting the protein–protein complex set into the two different classes.

In 70% of the cases for non-antibody complexes and 50% for antibody/antigen complexes, a clear correlation between the combined hydrogen

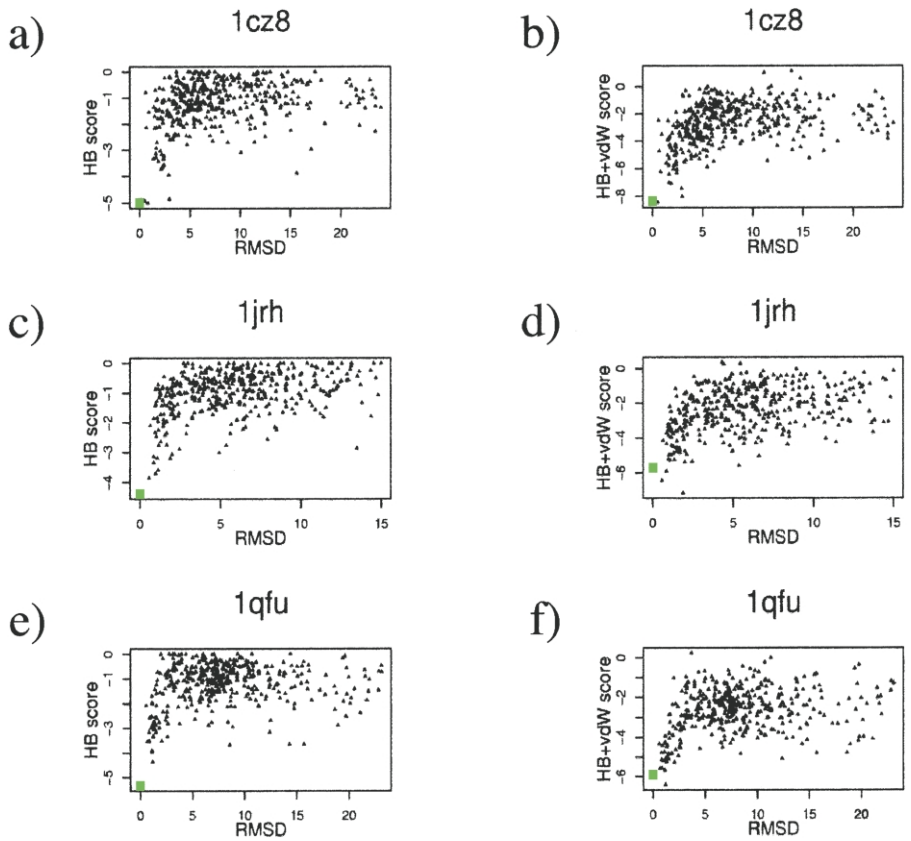

Figure 6. Scatter plots of the combined hydrogen bonding score alone (a), c) and e) and in combination with the van der Waals score (b), d) and f) *versus* decoy C^{α} rmsd from the native structure for selected antibody/antigen complexes. Native structures with the side-chains modeled using our rotamer repacking protocol are shown with green squares and decoys with black triangles.

bonding score (alone or in combination with the van der Waals scores *versus* score) and rmsd is apparent when approaching the native structure, starting from about 3 Å away. Examples are given in Figures 6 and 7 for antibody/antigen and non-antibody complexes.

Discussion

We have developed a simple orientation-dependent hydrogen bonding function, derived from the geometries of hydrogen bonds observed in high-resolution protein crystal structures (Figure 2). Several tests of this function have been performed: the prediction of amino acid sequences in monomeric proteins (1) and protein–protein interfaces (2); the discrimination of native and near-native structures from misfolded conformations for single domain monomeric proteins (3); and the application of the hydrogen bonding term to the protein–protein docking problem (using bound protein backbones, but repacked side-chains) (4). The hydrogen bonding term contributes significantly to the performance of the energy function in all four tests.

The prediction of amino acid sequences in monomeric structures for polar and charged amino acids is clearly improved by inclusion of the hydrogen bonding potential (Figure 3). It is particularly notable that this effect cannot be reproduced using a Coulomb potential of similar magnitude with a linear distance-dependent dielectric constant. This suggests that the Coulomb model of electrostatic interactions ignores some of the essential physical chemistry. Although the hydrogen bonding potential developed here is simple, it appears to capture the specifics of hydrogen bonding interactions reasonably well. Its directionality and explicit placement of polar hydrogen atoms are likely to be the major advantage over the Coulomb description of electrostatic interactions. The inclusion of polar hydrogen atoms in traditional treatments of electrostatic interactions has been suggested to introduce significant noise due to the sensitivity of the electrostatic energy to the precise locations of the protons.[10] Interestingly, molecular mechanics potentials originally contained explicit hydrogen bonding terms, which were replaced by electrostatic representations in later versions. In contrast, our results suggest a clear advantage of the inclusion of the explicit hydrogen bonding, but not based on a simple model such as dipole–dipole interactions (see Figure 2(d)).

The hydrogen bonding potential also performs well in discriminating native structures from

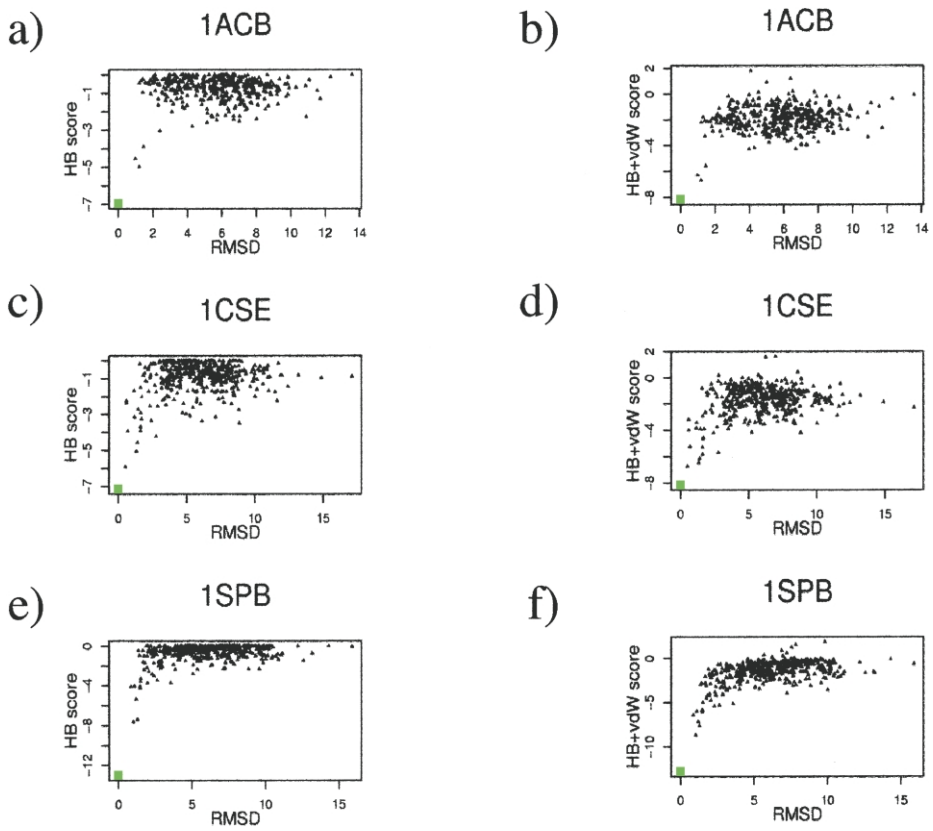

Figure 7. Scatter plots of the combined hydrogen bonding score alone (a), c) and e) and in combination with the van der Waals score (b), d) and f) *versus* decoy C^{α} rmsd from the native structure for selected non-antibody complexes. Native structures with the side-chains modeled using our rotamer repacking protocol are shown with green squares and decoys with black triangles.

misfolded decoys in both the monomeric and protein–protein complex decoy sets. However, it does not discriminate incorrect conformations from native-like structures in the case of the single-domain decoys generated by the ROSETTA *ab initio* method, and subjected to refinement and extensive side-chain repacking. The hydrogen bonding potential is clearly sensitive to distances of atoms between 1.5 Å and 2.5 Å, and there are very few decoys within the 1–3 Å rmsd range for most of the structures in the single domain decoy set. The manifestation of the specificity of hydrogen bonds suggests that the width of the "hydrogen bond funnel" around the native structure is narrow on the scale of these decoy sets. This is supported by the results on the perturbed-native decoy set: in cases where there are many decoys in the 1–3 Å range available, discrimination of native-like structures is possible for about half of the structures in the set (Table 2, Figure 5). Backbone hydrogen bonds are the best discriminator, while side-chain–side-chain and side-chain–backbone hydrogen bonds are not contributing significantly. Perhaps in less well packed globally misfolded decoy structures sufficient alternative side-chain conformations are available that local side-chain hydrogen bonding patterns can be optimized to a similar extent as in the native structure.

In contrast, both backbone and side-chain mediated hydrogen bonds contribute significantly to the discrimination of docking decoys. In many cases, we observe a good correlation between the hydrogen bond score (alone or in combination with the van der Waals score) and the rmsd to the native structure (Figures 6 and 7). This correlation starts to become apparent in the rmsd range of 2–3 Å, as suggested by the width of the hydrogen bonding funnel in the single-domain set. This result points to the applicability of this type of energy function to the minimization of docking decoys towards the native complex once decoys structurally close enough can be generated.[52] Our protein complex data set was generated using the backbones of the protein partners in their bound conformations (although the native side-chain conformations are eliminated in the creation of all docking decoys). A more stringent test for future work will be to create protein complex decoys starting from the conformations of separately crystallized (unbound) components.

It should be emphasized that the success in the hydrogen bonding potential in reproducing native sequences and discriminating misfolded from native and near-native conformations does not bear on the question of whether hydrogen bonds are contributing to the stability of proteins and protein interfaces. Our approach is based on the

assumption that the sequences and structures of native single domain structures and protein interfaces are on average electrostatically optimized[42] when compared to non-native sequences and alternative compact conformations. This does not require that hydrogen bonding interactions in native proteins and protein–protein complexes are more favorable than the hydrogen bonds the same groups make with water in the solvated unfolded or unbound ensembles.

The encouraging results predicting the identity of amino acid residues in protein interfaces (Figure 4), taken together with a previous study predicting binding energy hotspots in 19 protein–protein complexes with reasonable accuracy,[46] suggest that our energy function including the new hydrogen bonding term recapitulates determinants of both specificity and affinity in protein–protein interfaces. This suggests that a combination of the energy function and our side-chain repacking protocol will enhance the prediction of specificity in protein interactions and aid the design of protein–protein complexes. Given the available structure of a specific complex of one or more members of a large family of known protein interaction domains, the method should allow the generation of specificity profiles for all family members with significant overall sequence and assumed structural similarity (T.K. & D.B., unpublished results). With regard to the design aspect, we have applied this method to the redesign of specificity in a complex between a bacterial DNase and its inhibitor protein as well as to the design of a protein–protein interface to create a chimeric artificial endonuclease.[53]

Methods

Native protein structure datasets

Three different collections of protein structures solved by X-ray crystallography were used in this study. (1) The dataset used for compiling hydrogen bonding statistics contained 698 proteins with a resolution of 1.6 Å or better and a crystallographic *R* factor of 0.25 or better, taken from the Dunbrack culled pdb collection †. The list was additionally filtered to only include single-chain proteins. (2) The high-resolution dataset used for parameterizing the energy function was taken from Word et al.[54] and contained non-redundant structures (less than 30% sequence homology) with a resolution of 1.7 Å or better, a crystallographic *R* factor of 0.2 or better, and a Pro-Check overall *G*-factor of −0.6 or better.[55] An additional filter excluded structures with missing side-chain atoms and backbone sections and yielded a final set of 52 structures. (3) The protein interface dataset was generated from the non-redundant set of protein–protein interfaces compiled by Tsai *et al*.[56] Only heterodimeric protein–protein complexes were selected, and structures were additionally filtered to not contain significant portions of missing density, yielding a total of 50 structures. Ligands and ions contained in the structures were ignored in the analysis.

† http://www.fccc.edu/research/labs/dunbrack

Atomic coordinates and preparation of native structures

Atomic coordinates were taken from structures solved by X-ray crystallography. Polar hydrogen atoms 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 (see below), including a 180° flip of asparagine and glutamine amide groups and the two histidine imidazole tautomers (assumed to be uncharged). Global optimization of the hydrogen bonding network and replacement of missing atoms for amino acid side-chains was performed for each structure using a simple Metropolis Monte Carlo procedure as described previously[45] and the energy function described below. No other minimization of native structures was performed.

Definition of secondary structure for backbone–backbone hydrogen bonds

Secondary structure was defined by backbone torsion angles (helix: $-180° < \phi < -20°$, $-90° < \psi < -10°$; sheet: $-180° < \phi < -20°$, $180° > \psi > 20°$ or $-180° < \psi < -170°$). Classification of "helix" or "strand" required that at least two adjacent residues have the same secondary structure. For a hydrogen bond to be counted as occurring in a helix or strand, both residues were required to have the same secondary structure classification; all other backbone–backbone hydrogen bonds were summarized as "other".

The free energy function

The free energy function is a linear combination of van der Waals interactions (represented by the attractive part of a Lennard–Jones potential (E_{LJattr}) and a linear distance-dependent repulsive term (E_{LJrep})), orientation-dependent terms for side-chain–side-chain, side-chain–backbone and backbone–backbone hydrogen bonds ($E_{\mathrm{HB(sc-bb)}}$, $E_{\mathrm{HB(sc-sc)}}$ and $E_{\mathrm{HB(bb-bb)}}$) (see Results and below), an implicit solvation model (G_{sol}),[57] an energy derived from an amino acid and backbone-dependent rotamer probability ($E_{\mathrm{rot}}(\mathrm{aa}, \phi, \psi)$),[39] an energy derived from amino-acid type (aa) dependent backbone ϕ, ψ probabilities ($E_{\phi/\psi}(\mathrm{aa})$), and amino acid type dependent reference energies to approximate the interactions made in the unfolded state ensemble ($E^{\mathrm{ref}}_{\mathrm{aa}}$; n_{aa} is the number of amino acids of a certain type):

$$\begin{aligned}\Delta G = {} & W_{\mathrm{attr}}E_{\mathrm{LJattr}} + W_{\mathrm{rep}}E_{\mathrm{LJrep}} + W_{\mathrm{HB(sc-bb)}}E_{\mathrm{HB(sc-bb)}} \\ & + W_{\mathrm{HB(sc-sc)}}E_{\mathrm{HB(sc-sc)}} + W_{\mathrm{HB(bb-bb)}}E_{\mathrm{HB(bb-bb)}} \\ & + W_{\mathrm{sol}}G_{\mathrm{sol}} + W_{\phi/\psi}E_{\phi/\psi}(\mathrm{aa}) + W_{\mathrm{rot}}E_{\mathrm{rot}}(\mathrm{aa}, \phi, \psi) \\ & + \sum_{\mathrm{aa}=1}^{20} n_{\mathrm{aa}}E^{\mathrm{ref}}_{\mathrm{aa}} \end{aligned} \quad (2)$$

The Lennard–Jones potential, solvation term, and backbone-dependent amino acid and rotamer probabilities were as previously described.[45,57] The amino acid type dependent reference energies which approximate the free energy of the unfolded reference state[45] and the

weights W for the relative contributions of the different energy terms were obtained by fitting all parts of the scoring function to reproduce native sequences of naturally occurring proteins as described below. Energies for the different geometric parameters describing backbone–backbone and side-chain–side-chain hydrogen bonds $E(p)$ were obtained using:

$$E(p) = -\ln \frac{f_{\text{protein}}(p)}{f_{\text{random}}(p)} \quad (3)$$

where $f_{\text{protein}}(p)$ is the frequency at which a geometric parameter p is observed in a certain bin in the high-resolution crystal structure dataset, and $f_{\text{random}}(p)$ is a reference frequency value assuming equal distribution in all bins (for the distance distributions, the long-range cutoff was 2.6 Å for backbone–backbone hydrogen bonds and 3.0 Å for side-chain–side-chain hydrogen bonds). The kT prefactor is left out in equation (3) as it is included in the weight given in equation (1). The distributions were collected as described in Results. Energies for backbone–side-chain hydrogen bonds were taken from the side-chain–side-chain statistics. Hydrogen bonding energies with largely unfavorable energies for one or more of the component energy terms (resulting in a positive total hydrogen bonding energy) were set to zero. Coulomb electrostatics (to replace the hydrogen bonding term in our tests) used a linear distance-dependent dielectric constant. All parameters for the hydrogen bonding potential can be found in the Supplementary Material. Partial charges were taken from the CHARMM19 parameter set[28] with or without a correction for charged residues neutralizing charged groups but increasing their polarity;[57] no significant difference between the two charge sets was observed in our tests.

Parameterizing the energy function on monomeric proteins

The relative contributions of the different terms of the free energy function were parameterized on the high-resolution structure dataset as described previously.[45] Briefly, rotamers for all amino acids at all sequence positions in the data set with 52 crystal structures with a resolution of 1.7 Å or better were created (a total of 7308 sequence positions with an average of 684 rotamers per sequence position). The components of the energy function (attractive and repulsive van der Waals interactions, solvation, hydrogen bonding, Coulomb electrostatics, backbone-dependent amino acid type and rotamer probabilities and reference energies) were computed for all rotamers at each sequence position assuming a constant environment of all other amino acids in their native conformation. The weights on all terms were optimized using a conjugate gradient method to maximize the probability of the native amino acid type at each position. Different initial values yielded similar fitted parameters, showing convergence of the fit. The relative weights were 1.06 (attractive Lennard–Jones), 0.77 (repulsive Lennard–Jones), 0.72 (solvation), 0.42 (backbone–side-chain hydrogen bonding) 0.40 (side-chain–side-chain hydrogen bonding; the weight for backbone–backbone hydrogen bonding could not be determined using this procedure as the backbone stayed constant), 0.89 (backbone-dependent rotamer probability) and 0.86 (backbone-dependent amino acid type probability). These weights result in a free energy of about 3 kcal/mol for a hydrogen bond with ideal geometry. The rotamer library used was taken from Dunbrack[39] as described by Kuhlman & Baker.[45] Additional rotamers were included with small deviations (10–20°) of the χ_1 and χ_2 angles for buried residues, and extra angles for χ_3 and χ_4 angles as described by Mayo & co-workers.[58] For serine, threonine and tyrosine, hydrogen conformations were chosen according to those observed in neutron diffraction maps.[59] For threonine and serine hydroxyl groups, the three different staggered positions, for tyrosine hydroxyl groups hydrogen atoms were assumed to be in the plane of the aromatic ring. In all cases, small deviations ($\pm 20°$) were included additionally.

Decoy sets

The generation of the decoy sets is described in more detail elsewhere.[50,60,61] In brief, two sets of decoys were used: the first set contains decoys for 41 monomeric, single domain proteins with less than 90 amino acid residues, generated by the ROSETTA method for *ab initio* protein structure prediction. Approximately 2000 decoys were used for each structure. This set was further subdivided into (a) 25 proteins with an available high resolution crystal structure (used for discrimination of the native structure) and (b) 23 proteins where 10% of the decoys produced by ROSETTA had a C^{α} rmsd from the native structure of 4 Å or better (note that some proteins belong to both subsets). In addition, for each of these 23 structures, 300 decoys were created by peptide fragment-insertion starting from the native structure ("perturbed-native decoy set") using the ROSETTA method, followed by refinement on the centroid and full-atom level. The latter two decoy sets were used for discriminating low-rmsd from high-rmsd decoys. Finally, polar hydrogen atoms were added to all decoys, followed by side-chain repacking, and simultaneous optimization of the hydrogen bonding network and scoring using the energy function described above.

The second set contained docking decoys for 31 protein–protein complexes, with 18 antibody–antigen and 13 enzyme–enzyme inhibitor and other interfaces. For each structure, 400 decoys were used (the results were unchanged when a larger set of 2000 decoys per structure was tested). Decoys were created by rigid-body perturbations of the relative orientation of the two partners in the protein–protein complex.[50] Note that the backbone coordinates of the bound conformations of both partners were used. However, the side-chain conformational information contained in the crystal structure coordinates was eliminated by repacking all side-chains using the rotamer repacking protocol described previously[45] prior to rigid-body docking. As a last step, the interface of all docked decoys was repacked using a Monte-Carlo simulated annealing protocol and the energy function described above, and final scores were collected. For the repacking and scoring of decoys, the side-chain–side-chain hydrogen bonds were divided into three environment classes, dependent on the extent of burial of both participating 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 C^{β} atoms within a sphere of 10 Å radius of the C^{β} atom of the residue of interest: exposed 0–14, intermediate 15–20, buried >20). The relative contributions of the environment-dependent hydrogen

bonding terms were estimated from changes in protein stability upon point mutation, and were 0.18, 0.28 and 0.91.[46]

Z-score analysis and logistic regression

Three different *Z*-score measures were used, defined as follows:

$$Z_{\mathrm{ref}} = \frac{\langle E \rangle - E_{\mathrm{ref}}}{\sigma_{\mathrm{E}}} \tag{4}$$

where:

$$\langle E \rangle = \frac{1}{N} \sum_{i=1}^{N} E_i \tag{5}$$

is an average energy of N decoys:

$$\sigma_{\mathrm{E}}^2 = \frac{1}{N} \sum_{i=1}^{N} (E_i - \langle E \rangle)^2 \tag{6}$$

is the standard deviation of decoy energies, and E_{ref} is the reference energy which is either E_{nat} (energy of the native structure experimentally determined by X-ray diffraction or nuclear magnetic resonance spectroscopy) or $E_{\mathrm{nat\_rep}}$ (energy of the structure with the native polypeptide backbone but repacked side-chains using our rotamer repacking protocol). These *Z*-scores are referred to as Zn (native) and Znr (native-repacked) in the Tables.

The low rmsd (root mean standard deviation in C^{α} coordinates from the native structure) *Z*-scores (Z_{lrms}, discriminating near-native from non-native conformations) are defined as:

$$Z_{\mathrm{lrms}} = \frac{\langle E \rangle_{\mathrm{hi}} - \langle E \rangle_{\mathrm{lo}}}{\sigma_{\mathrm{E}}^{\mathrm{hi}}} \tag{7}$$

where the sum of the averages and the standard deviation are computed over decoys with high (hi) and low (lo) rmsd separately. Low rmsd (near-native) decoys are defined as the lowest 5% of the rmsd distribution.

Combined free energies (and their *Z*-scores) using different energy terms were obtained by logistic regression using a generalized linear model implemented in the *R* statistical software package.

Acknowledgements

We thank members of the Baker laboratory for many stimulating discussions, Jerry Tsai, Kira Misura, Jeff Gray and Stewart Moughon for help with creating the original decoy sets, Kira Misura and a reviewer for very helpful comments on the manuscript, and Keith Laidig for computing support. T.K. was supported by the Human Frontier Science Program and EMBO. This work was also supported by a grant from the NIH and the Howard Hughes Medical Institute.

References

1. Baker, E. N. & Hubbard, R. E. (1984). Hydrogen bonding in globular proteins. *Prog. Biophys. Mol. Biol.* **44**, 97–179.
2. Conte, L. L., Chothia, C. & Janin, J. (1999). The atomic structure of protein–protein recognition sites. *J. Mol. Biol.* **285**, 2177–2198.
3. Hendsch, Z. S. & Tidor, B. (1994). Do salt bridges stabilize proteins? A continuum electrostatic analysis. *Protein Sci.* **3**, 211–226.
4. Pace, C. N. (2001). Polar group burial contributes more to protein stability than nonpolar group burial. *Biochemistry*, **40**, 310–313.
5. McDonald, I. K. & Thornton, J. M. (1994). Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* **238**, 777–793.
6. Waldburger, C. D., Schildbach, J. F. & Sauer, R. T. (1995). Are buried salt bridges important for protein stability and conformational specificity? *Nature Struct. Biol.* **2**, 122–128.
7. Yang, A. S. & Honig, B. (1995). Free energy determinants of secondary structure formation: I. alpha-helices. *J. Mol. Biol.* **252**, 351–365.
8. Hendsch, Z. S., Jonsson, T., Sauer, R. T. & Tidor, B. (1996). Protein stabilization by removal of unsatisfied polar groups: computational approaches and experimental tests. *Biochemistry*, **35**, 7621–7625.
9. Lumb, K. J. & Kim, P. S. (1995). A buried polar interaction imparts structural uniqueness in a designed heterodimeric coiled coil. *Biochemistry*, **34**, 8642–8648.
10. Petrey, D. & Honig, B. (2000). Free energy determinants of tertiary structure and the evaluation of protein models. *Protein Sci.* **9**, 2181–2191.
11. Morokuma, K. (1977). Why do molecules interact? The origin of electron donor–acceptor complexes, hydrogen bonding and proton affinity. *Accts. Chem. Res.* **10**, 294–300.
12. Kollman, P. A. (1977). Noncovalent interactions. *Accts. Chem. Res.* **10**, 365–371.
13. Taylor, R., Kennard, O. & Versichel, W. (1983). Geometry of the N–H···O=C hydrogen bond. 1. Lone-pair directionality. *J. Am. Chem. Soc.* **105**, 5761–5766.
14. Taylor, R. & Kennard, O. (1984). Hydrogen-bond geometry in organic crystals. *Accts. Chem. Res.* **17**, 320–325.
15. Gavezzotti, A. & Filippini, G. (1994). Geometry of the intermolecular X–H···Y (X, Y = N, O) hydrogen bond and the calibration of empirical hydrogen-bond potentials. *J. Phys. Chem. ser. B*, **98**, 4831–4837.
16. Platts, J. A., Howard, S. T. & Bracke, B. R. F. (1996). Directionality of hydrogen bonds to sulfur and oxygen. *J. Am. Chem. Soc.* **118**, 2726–2733.
17. Lommerse, J. P. M., Price, S. L. & Taylor, R. (1997). Hydrogen bonding of carbonyl, ether, and ester oxygen atoms with alkanol hydroxyl groups. *J. Comput. Chem.* **18**, 757–774.
18. Grzybowski, B. A., Ishchenko, A. V., DeWitte, R. S., Whitesides, G. M. & Shakhnovich, E. I. (2000). Development of a knowledge-based potential for crystals of small organic molecules: calculation of energy surfaces for C=O···H–N hydrogen bonds. *J. Phys. Chem. ser. B*, **104**, 7293–7298.
19. Ippolito, J. A., Alexander, R. S. & Christianson, D. W. (1990). Hydrogen bond stereochemistry in protein structure and function. *J. Mol. Biol.* **215**, 457–471.

20. Stickle, D. F., Presta, L. G., Dill, K. A. & Rose, G. D. (1992). Hydrogen bonding in globular proteins. *J. Mol. Biol.* **226**, 1143–1159.
21. Fabiola, F., Bertram, R., Korostelev, A. & Chapman, M. S. (2002). An improved hydrogen bond potential: impact on medium resolution protein structures. *Protein Sci.* **11**, 1415–1423.
22. McGuire, R. F., Momany, F. A. & Scheraga, H. A. (1972). Energy parameters in polypeptides. V. An empirical hydrogen bond function based on molecular orbital calculations. *J. Phys. Chem.* **76**, 375–393.
23. Wiberg, K. B., Marquez, M. & Castejon, H. (1994). Lone pairs in carbonyl compounds and ethers. *J. Org. Chem.* **59**, 6817–6822.
24. Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983). CHARMM: a program for macromolecular energy, minimization and dynamics calculations. *J. Comput. Chem.* **4**, 187–217.
25. Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G. *et al.* (1984). A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.* **106**, 765–784.
26. Jorgensen, W. J. & Tirado-Rives, J. (1988). The OPLS potential function for proteins. Energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.* **110**, 1657–1666.
27. Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M. J., Ferguson, D. M. *et al.* (1995). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **117**, 5179–5197.
28. Neria, E., Fischer, S. & Karplus, M. (1996). Simulation of activation free energies in molecular systems. *J. Chem. Phys.* **105**, 1902–1921.
29. Buck, M. & Karplus, M. (2001). Hydrogen bond energetics: a simulation and statistical analysis of *N*-methyl acetamide (NMA), water and human lysozyme. *J. Phys. Chem. ser. B*, **105**, 11000–11015.
30. Mayo, S. L., Olafson, B. D. & Goddard, W. A. I. (1990). DREIDING: a generic force field for molecular simulations. *J. Phys. Chem.* **94**, 8897–8909.
31. Dahiyat, B. I., Gordon, D. B. & Mayo, S. L. (1997). Automated design of the surface positions of protein helices. *Protein Sci.* **6**, 1333–1337.
32. Pokala, N. & Handel, T. M. (2001). Review: protein design—where we were, where we are, where we're going. *J. Struct. Biol.* **134**, 269–281.
33. Hao, M. H. & Scheraga, H. A. (1999). Designing potential energy functions for protein folding. *Curr. Opin. Struct. Biol.* **9**, 184–188.
34. Osguthorpe, D. J. (2000). *Ab initio* protein folding. *Curr. Opin. Struct. Biol.* **10**, 146–152.
35. Sternberg, M. J., Gabb, H. A. & Jackson, R. M. (1998). Predictive docking of protein–protein and protein–DNA complexes. *Curr. Opin. Struct. Biol.* **8**, 250–256.
36. Brunger, A. T. & Karplus, M. (1988). Polar hydrogen positions in proteins: empirical energy placement and neutron diffraction comparison. *Proteins: Struct. Funct. Genet.* **4**, 148–156.
37. Lipsitz, R. S., Sharma, Y., Brooks, B. R. & Tjandra, N. (2002). Hydrogen bonding in high-resolution protein structures: a new method to assess NMR protein geometry. *J. Am. Chem. Soc.* **124**, 10621–10626.
38. Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, **181**, 223–230.
39. Dunbrack, R. L., Jr & Cohen, F. E. (1997). Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci.* **6**, 1661–1681.
40. Schreiber, G. (2002). Kinetic studies of protein–protein interactions. *Curr. Opin. Struct. Biol.* **12**, 41–47.
41. Sheinerman, F. B., Norel, R. & Honig, B. (2000). Electrostatic aspects of protein–protein interactions. *Curr. Opin. Struct. Biol.* **10**, 153–159.
42. Lee, L. P. & Tidor, B. (2001). Barstar is electrostatically optimized for tight binding to barnase. *Nature Struct. Biol.* **8**, 73–76.
43. Bonneau, R., Tsai, J., Ruczinski, I., Chivian, D., Rohl, C., Strauss, C. E. & Baker, D. (2001). Rosetta in CASP4: progress in *ab initio* protein structure prediction. *Proteins: Struct. Funct. Genet.* **45**, 119–126.
44. Simons, K. T., Ruczinski, I., Kooperberg, C., Fox, B. A., Bystroff, C. & Baker, D. (1999). Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins: Struct. Funct. Genet.* **34**, 82–95.
45. Kuhlman, B. & Baker, D. (2000). Native protein sequences are close to optimal for their structures. *Proc. Natl Acad. Sci. USA*, **97**, 10383–10388.
46. Kortemme, T. & Baker, D. (2002). A simple physical model for binding energy hot spots in protein–protein complexes. *Proc. Natl Acad. Sci. USA*, **99**, 14116–14121.
47. Park, B. H., Huang, E. S. & Levitt, M. (1997). Factors affecting the ability of energy functions to discriminate correct from incorrect folds. *J. Mol. Biol.* **266**, 831–846.
48. Gatchell, D. W., Dennis, S. & Vajda, S. (2000). Discrimination of near-native protein structures from misfolded models by empirical free energy functions. *Proteins: Struct. Funct. Genet.* **41**, 518–534.
49. Vorobjev, Y. N. & Hermans, J. (2001). Free energies of protein decoys provide insight into determinants of protein stability. *Protein Sci.* **10**, 2498–2506.
50. Gray, J. J., Moughon, S., Kortemme, T., Schueler-Furman, O., Misura, K. M. S., Morozov, A. V. & Baker, D. (2003). Protein–protein docking predictions for the CAPRI experiment. In press.
51. Lawrence, M. C. & Colman, P. M. (1993). Shape complementarity at protein/protein interfaces. *J. Mol. Biol.* **234**, 946–950.
52. Camacho, C. J. & Vajda, S. (2001). Protein docking along smooth association pathways. *Proc. Natl Acad. Sci. USA*, **98**, 10636–10641.
53. Chevalier, B. S., Kortemme, T., Chadsey, M. S., Baker, D., Monnat, R. J. J. & Stoddard, B. L. (2002). Design, activity and structure of E-DreI, a highly site-specific artifical endonuclease. *Mol. Cell*, **10**, 895–905.
54. Word, J. M., Lovell, S. C., LaBean, T. H., Taylor, H. C., Zalis, M. E., Presley, B. K. *et al.* (1999). Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. *J. Mol. Biol.* **285**, 1711–1733.
55. Laskowski, R. A., Rullmannn, J. A., MacArthur, M. W., Kaptein, R. & Thornton, J. M. (1996). AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR*, **8**, 477–486.
56. Tsai, C. J., Lin, S. L., Wolfson, H. J. & Nussinov, R. (1996). A dataset of protein–protein interfaces generated with a sequence-order-independent comparison technique. *J. Mol. Biol.* **260**, 604–620.
57. Lazaridis, T. & Karplus, M. (1999). Effective energy function for proteins in solution. *Proteins: Struct. Funct. Genet.* **35**, 133–152.

58. Dahiyat, B. I. & Mayo, S. L. (1997). *De novo* protein design: fully automated sequence selection. *Science*, **278**, 82–87.
59. Kossiakoff, A. A., Shpungin, J. & Sintchak, M. D. (1990). Hydroxyl hydrogen conformations in trypsin determined by the neutron diffraction solvent difference map method: relative importance of steric and electrostatic factors in defining hydrogen-bonding geometries. *Proc. Natl Acad. Sci. USA*, **87**, 4468–4472.
60. Tsai, J., Bonneau, R., Morozov, A. V., Kuhlman, B., Rohl, C. A. & Baker, D. (2003). An improved protein decoy set for testing energy functions for protein structure prediction. In press
61. Morozov, A. V., Kortemme, T. & Baker, D. (2003). Evaluation of models of electrostatic interactions in proteins. *J. Phys. Chem. B*. In press.

Edited by P. Wright

(Received 19 July 2002; received in revised form 17 December 2002; accepted 17 December 2002)

Supplementary Material comprising four Tables is available on Science Direct