J|A|C|S
ARTICLES
Published on Web 07/03/2003

Role of Water Mediated Interactions in Protein−Protein Recognition Landscapes

Garegin A. Papoian,*,† Johan Ulander,† and Peter G. Wolynes

Contribution from the Department of Chemistry & Biochemistry, University of California at San Diego, 9500 Gilman Dr., La Jolla, California 92093-0371

Received February 17, 2003; E-mail: gpapoian@ucsd.edu

Abstract: The energy landscape picture of protein folding and binding is employed to optimize a number of pair potentials for direct and water-mediated interactions in protein complex interfaces. We find that water-mediated interactions greatly complement direct interactions in discriminating against various types of trap interactions that model those present in the cell. We highlight the context dependent nature of knowledge-based binding potentials, as contrasted with the situation for autonomous folding. By performing a Principal Component Analysis (PCA) of the corresponding interaction matrixes, we rationalize the strength of the recognition signal for each combination of the contact type and reference trap states using the differential in the idealized "canonical" amino acid compositions of native and trap layers. The comparison of direct and water-mediated contact potential matrixes emphasizes the importance of partial solvation in stabilizing charged groups in the protein interfaces. Specific water-mediated interresidue interactions are expected to influence significantly the kinetics as well as thermodynamics of protein association.

Introduction

The interplay between water and proteins has been a subject of intense theoretical and experimental studies for many decades.[1,2] It has been long understood that water plays a crucial role in determining the structure and dynamics of biological macromolecules such as proteins or DNA. For example, the so-called "hydrophobic force" is considered to be the major ingredient in protein folding as well as in other biological processes, such as large scale macromolecular assembly and biomolecular recognition. Detailed understanding of the effects exerted by water on biological interactions remains elusive, however. One reason for this is that water takes on numerous different roles. Among other roles, water participates in many specific interactions, screens efficiently Coulombic interactions, mediates proton transfer,[3] and even is used as a structural component in protein secondary structure.[4]

When trying to understand the underlying interactions between building blocks of biological molecules such as proteins, water degrees of freedom are, however, often averaged out, leading to effective (many-body) forces that describe the behavior of the reduced system. In this regard, water would seem to be merely as a complicating factor rather than an integral part of the system under study. This extreme view is based on a basically sound physical underpinning, namely the fast relaxation times of most of the water degrees of freedom compared to protein motion which involve activated steps over dihedral angle potential barriers. Yet, on the other extreme, it has long been known that there often do exist a few water molecules tightly incorporated into the protein framework, which are better regarded as part of the protein structure. Motions of these bound waters are extremely slow.[3] Between these two extreme regimes there exists a gray area—water molecules residing transiently near a protein surface exhibit a behavior which is intermediate between the nearly frozen structural waters and bulk water.[5,6]

Some generic features of medium-range protein−protein interactions mediated by crystallographically characterized (Figure 1a) as well as by more *transiently* residing water molecules may be approximately described by a generic double well potential, the first well specifying the direct contacts and the second well specifying water-separated contacts. Such a potential was recently used to study the effect of the water-mediated interactions on the protein hydrophobic collapse.[7]

The water in the interface immediately outside the protein is often referred to as the hydration shell. The hydration shell is not a rigid entity and is described best by statistical means.[6] A hydration level of less than 0.4 g of water per gram of protein, not enough to fully cover the protein surface, is sufficient for "full" activation of dynamics and functionality of many globular proteins.[8] This suggests that in some regions of the protein water plays a very specific and important role.

† Garegin A. Papoian and Johan Ulander have contributed equally to this work.

(1) Bastolla, U.; Farwer, J.; Knapp, E. W.; Vendruscolo, M. *Annu. Rev. Biophys. Biomolec. Struct.* **1993**, *22*, 67−97.
(2) Wyman, J. *Adv. Protein. Chem.* **1964**, *19*, 223−286.
(3) Gottschalk, M.; Dencher, N. A.; Halle, B. *J. Mol. Biol.* **2001**, *311*, 605−621.
(4) Marino, M.; Braun, L.; Cossart, P.; Ghosh, P. *Mol. Cell.* **1999**, *4*, 1063−1072.
(5) Makarov, V.; Pettitt, B. M.; Feig, M. *Acc. Chem. Res.* **2002**, *35*, 376−384.
(6) Timasheff, S. N. *Biochemistry* **2002**, *41*, 13 473−13 482.
(7) Cheung, M. S.; Garcia, A. E.; Onuchic, J. N. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 685−690.

10.1021/ja034729u CCC: $25.00

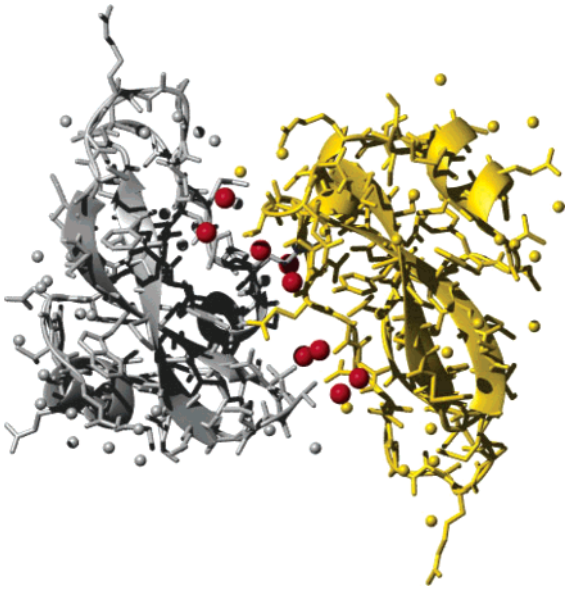

Figure 1. Schematic representation of a representative database protein complex (PDB code 1aap) with two chains (depicted in gray and gold) and interface crystallographic waters (red) as well as other crystallographic waters that are not shared between two chains (given in the corresponding chain's color).

We believe that the binding interface between two associated protein chains may be among such important regions. We conjecture that water-mediated interactions are actually specific in character, facilitating biomolecular recognition. To investigate this possibility, we derive in the current work a number of direct and water-mediated residue-level knowledge-based protein association contact potentials. We assume that specific water-mediated interactions at least to a certain degree may be projected on to pairwise additive interactions at an amino acid residue resolution level. Water mediated contacts, as opposed to direct contacts between amino acid residues, may thus provide an additional layer of recognition in protein binding and perhaps even in protein folding processes.

The importance of a heterogeneous environment for protein−ligand thermodynamics has long been recognized.[2,6] Effective interactions between for example two residues will by necessity be modulated by the presence of, e.g., other residues in a given molecule as well as solvent molecules. A biological cell constitutes a very heterogeneous environment and specificity of protein−protein interactions is of primary importance in protein folding as well as protein−protein association. A protein must be able to find its targets but also must be able to let go of false "decoys". This imposes special requirements on the composition of proteins in the cell but can also be used to rationalize constraints on the optimization of effective potentials from a bioinformatic perspective. The results from our study strongly suggest that water-mediated contacts may indeed be highly specific and do actually complement the direct contacts in the task of recognizing a particular binding partner out of many competing proteins.

The outline of the paper is as follows. We proceed first to describe an energy landscape theory of coupling protein association with protein conformational motion. Afterward, we highlight the problem of optimizing binding potentials taking into consideration the effect of the highly heterogeneous cellular environment. Finally, after providing computational details, we derive and rationalize coarse-grained direct and water-mediated protein association potentials.

Theory and Computational Details

Energy Landscape Theory of Binding and Folding. At first, protein association may appear as a deceivingly simple process. One might think that two proteins, complementary with each other in shape and electrostatics, approach as rigid bodies and dock into the native protein complex. This is the celebrated "lock-and-key" paradigm of protein association, hypothesized by Fischer more than 100 years ago.[9,10] The simplicity of this model comes from a presumed low-dimensionality of the search problem: only six rotational-translation degrees of freedom are lost during the association process. For instance, by constructing a grid in this space (given accurate interaction potentials), the protein docking problem can apparently be solved in a brute force manner.

Starting from the seminal works of Koshland and co-workers, there has, however, been a realization that often protein monomers adjust their conformation, at least to a limited extent, during the association event.[11] In this so-called "induced-fit" model of binding, a relatively small number of protein residues (usually near the binding site) exhibit conformational plasticity, adjusting side-chain and backbone conformations to better fit the partner protein. This clearly represents a large expansion of the search space for binding, where perhaps tens of degrees of freedom must be explicitly considered.

In the past decade or so, a number of researchers have further come to realization that many binding events involve partially or completely unfolded partner proteins, that organize only upon binding.[12−19] The dimensionality of the search space becomes huge in this case, presenting at least the same level of difficulty as the protein folding problem.[20,21] Indeed, there have been several recent papers investigating the importance of funneling in binding energy landscapes.[22−25]

Thus, given the varying degrees of flexibility that partner proteins might have before associating with each other, it would seem beneficial to have an encompassing theory that describes the coupling of the binding events with protein conformational mobility. Such an analytical model, based on the energy landscape theory, has recently been put forward by Papoian and Wolynes.[26] We envision that prior to binding both partner proteins are individually flexible to a particular degree. However, when they encounter each other, the information encoded in their energy landscapes allows them to form the native interface, with additional ordering of their corresponding internal degrees of freedom occurring concurrently.[26]

We partition the phase space of interacting proteins A and B into two equilibria (Figure 2): first (I) the monomers A and B associate

(8) Bizzarri, A. R.; Cannistraro, S. *J. Phys. Chem. B* **2002**, *106*, 6617−6633.
(9) Fischer, E. *Ber. Dtsch. Chem. Ges.* **1890**, *23*, 2611.
(10) Fischer, E. *Ber. Dtsch. Chem. Ges.* **1894**, *27*, 2985.
(11) Koshland, D. E., Jr. *Angew. Chem., Int. Ed. Engl.* **1994**, *33*, 2375−2378.
(12) Wright, P. E.; Dyson, H. J. *J. Mol. Biol.* **1999**, *293*, 321−331.
(13) Dyson, H. J.; Wright, P. E. *Curr. Opin. Struct. Biol.* **2002**, *12*, 54−60.
(14) Sauer, R. T. *Nature* **1990**, *347*, 514−515.
(15) Spolar, R. S.; Record, M. T., Jr. *Science* **1994**, *263*, 777−784.
(16) Johnson, N. P.; Lindstrom, J.; Baase, W. A.; von Hippel, P. H. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 4840−4844.
(17) Johnson, C. R.; Morin, P. E.; Arrowsmith, C. H.; Freire, E. *Biochemistry* **1995**, *34*, 5309−5316.
(18) Zitzewitz, J. A.; Bilsel, O.; Luo, J.; Jones, B. E.; Matthews, C. R. *Biochemistry* **1995**, *34*, 12 812−12 819.
(19) Ozawa, T.; Nogami, S.; Sato, M.; Ohya, Y.; Y., U. *Anal. Chem.* **2000**, *72*, 5151−5157.
(20) Bryngelson, J. D.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 7524−7528.
(21) Wolynes, P. G.; Eaton, W. A. *Phys. World* **1999**, *12*, 39−44.
(22) Tsai, C. J.; Kumar, S.; Ma, B.; Nussinov, R. *Prot. Sci.* **1999**, *8*, 1181−1190.
(23) Verkhivker, G. M.; Bouzida, D.; Gehlhar, D. K.; Rejto, P. A.; Freer, S. T.; Rose, P. W. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 5148−5153.
(24) Miller, D. W.; Dill, K. A. *Protein Sci.* **1997**, *6*, 2166−2179.
(25) Zhang, C.; Chen, J.; DeLisi, C. *Proteins: Struct. Funct. Genet.* **1999**, *34*, 255−267.
(26) Papoian, G. A.; Wolynes, P. G. *Biopolymers* **2003**, *68*, 333−349.

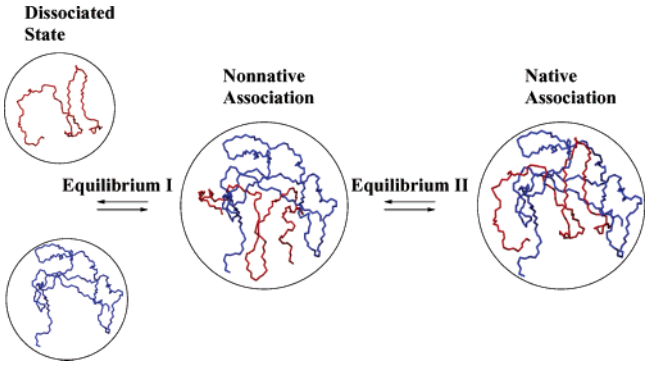

Figure 2. Two equilibria that describe the association of protein monomers. Monomers form a nonnative association complex (equilibrium I) which under right thermodynamic and kinetic conditions might subsequently order into the native association complex (equilibrium II).

into a disordered AB complex, next (II) the non-native AB complex undergoes a phase transition, where internal ordering of A and B is coupled to the formation of the native binding interface. According to the minimally frustrated random energy model (MFREM) of protein association,[26] the free energy change during this ordering process (i.e., equilibrium II) may be written as

$$\delta F = \delta E_a + \delta E_b + \delta E_i - \left(- S^0 T - \frac{\Delta\epsilon_a^2}{2k_B T} - \frac{\Delta\epsilon_b^2}{2k_B T} - \frac{\Delta\epsilon_i^2}{2k_B T} \right) \quad (1)$$

δE_a, δE_b represent the energy gaps between the native *folding* energies for monomers A and B by themselves and the average energies of their corresponding disordered states. Similarly, δE_i measures the energy gap between the native interface contacts and the average energy of representative disordered interfaces. The nonnative states are also characterized by another energy parameter, the disordered ensemble energy variance, which represents the ruggedness of underlying energy landscapes. $\Delta\epsilon_a^2$ and $\Delta\epsilon_b^2$ characterize the ruggedness of misfolded conformations for monomers A and B, whereas $\Delta\epsilon_i^2$ is a similar measure for disordered interfaces. Finally, S^0 is the combined configurational entropy of the disordered phase (i.e., it includes both folding and binding components).

According to the MFREM theory, larger binding energy gap, δE_i, facilitates the formation of the ordered native interface, while larger configurational entropy, S^0, as well as greater heterogeneity of nonnative energies, $\Delta\epsilon_i^2$, favor the interface disorder. Thus, as far as the binding part of the funnel is concerned, the corresponding association potential should be optimized in such a way so to enhance the binding energy gap in units of the binding ruggedness.

In the arguments above, we have focused only on equilibrium II, i.e., the order−disorder transition between the nonnatively and natively associated protein complexes. In this context, analogous to the similar strategy when designing protein folding potentials,[27] maximization of the binding gap in units of the binding energy variance is well justified. If the first equilibrium is also taken into account, however, then certain protein association trap states might escape through dissociation into the medium, thus altering the kinetics of trapping. In the current work, we discard the latter possibility: a detailed study of the interplay between trap state dissociation and protein complex configurational dynamics will be published elsewhere.[28]

Ambiguous Nature of Trap States in Protein Association vs Protein Folding. For protein folding prediction as well as protein threading studies, a fully atomistic description of the force field remains computationally cumbersome even today. Therefore, coarse-grained pairwise additive contact potentials fill an important niche in biocomputational studies. A common way to obtain these potentials is to exploit the wealth of information contained in various protein structural databases, i.e., the frequency of specific residue−residue pairings may be used to determine contact free energies. A well-known potential of Miyazawa and Jernigan belongs to a class of such potentials that are derived by assuming a Boltzmann distribution of contact probabilities in the structural database with an ideal-gaslike reference state. Effective interactions for each contact type are then constructed by computing the potential of mean force from the relative contact probabilities.[29,30] Several other potential inversion schemes have also been suggested, including the perceptron optimization, Z-score optimization and a number of other related techniques.[31−35] These techniques can all be understood using energy landscape theory.[27,36] In this work, we use a strategy, similar to Z-score optimization to derive protein association potentials, as described in detail below.

The largest conceptual obstacle when considering the optimization of a protein association potential is the ambiguity in the treatment of the nature of the trap states (i.e., "decoys" or non-native configurations). In protein folding, nonnative states may be defined quite straightforwardly due to the constraints of the polymer chain: only a finite (but large) number of nonnative configurations are sterically allowed. Furthermore, in the molten globule ensemble an additional fraction of these configurations are readily discarded because competitive configurations are expected to be compact and low in energy. Non-native states in binding, on the other hand, are not determined by the binding partner protein alone but are highly context dependent, i.e., good discrimination in one environment of the cell may turn out to be ineffective in another.

To mimic the conditions present in the biological cell, nonnative binding states need to be considered as all possible modes of probe protein association with *all* other proteins present in the cell, except the *natively* associated configurations. Because the cell is a very crowded multi-protein environment (estimated protein concentration 300 mg/ml[37]), a particular protein must recognize its specific target hidden in the heterogeneous soup of numerous other (trap) protein targets. There are kinetic constraints to such recognition (i.e., nonnative complexes should dissociate fast enough), as well thermodynamic constraints (the binding equilibrium with the native partner must outweigh the consolidated effect of the set of all nonnative equilibria). In this paper, we explore quantitatively certain elements of the latter, the thermodynamics of recognition in protein association. We will provide a more comprehensive treatment of the generic phase diagrams and kinetics specificity of proteomic networks in a forthcoming publication.

In addition to the randomness caused by the environment heterogeneity, we conjecture that there also exist *systematic* variations of protein association affinity, such as the systematic dependence of the binding affinity on the degree of the conformational flexibility of the partner proteins. As discussed earlier, it has been estimated that a sizable fraction of proteins are partially or fully unfolded in the eukaryotic cell.[12,13,26,38,39] These unfolded proteins are expected to expose more hydrophobic residues on their surface than do well-folded proteins, which in turn suggests that they would have stronger nonnative association affinity toward the probe protein. To take into account this *systematic* trend, we have chosen to consider three distinct models of trap states based on the degree of protein flexibility (see Figure 3). To

(27) Eastwood, M. P.; Hardin, C.; Wolynes, P. G. *J. Chem. Phys.* **2002**, *117*, 4602−4615.
(28) Papoian, G. A.; Ulander, J.; and Wolynes, P. G., in preparation.
(29) Miyazawa, S.; Jernigan, R. L. *Macromolecules* **1985**, *18*, 534−552.
(30) Miyazawa, S.; Jernigan, R. L. *J. Mol. Biol.* **1996**, *256*, 623−644.
(31) Zhang, L.; Skolnick, J. *Protein Sci.* **1998**, *7*, 112−122.
(32) Betancourt, M. R.; Thirumalai, D. *Protein Sci.* **1999**, *8*, 361−369.
(33) Vendruscolo, M.; Mirny, L. A.; Shakhnovich, E. I.; Domany, E. *Proteins: Struct. Funct. Genet.* **2000**, *41*, 192−201.
(34) Chang, I.; Cieplak, M.; Dima, R. I.; Maritan, A.; Banavar, J. R. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 14 350−14 355.
(35) Bastolla, U.; Farwer, J.; Knapp, E. W.; Vendruscolo, M. *Proteins: Struct. Funct. Genet.* **2001**, *44*, 79−96.
(36) Onuchic, J. N.; Luthey-Schulten, Z. A.; Wolynes, P. G. *Annu. Rev. Phys. Chem.* **1997**, *48*, 545−600.
(37) Frydman, J. *Annu. Rev. Biochem.* **2001**, *70*, 603−647.

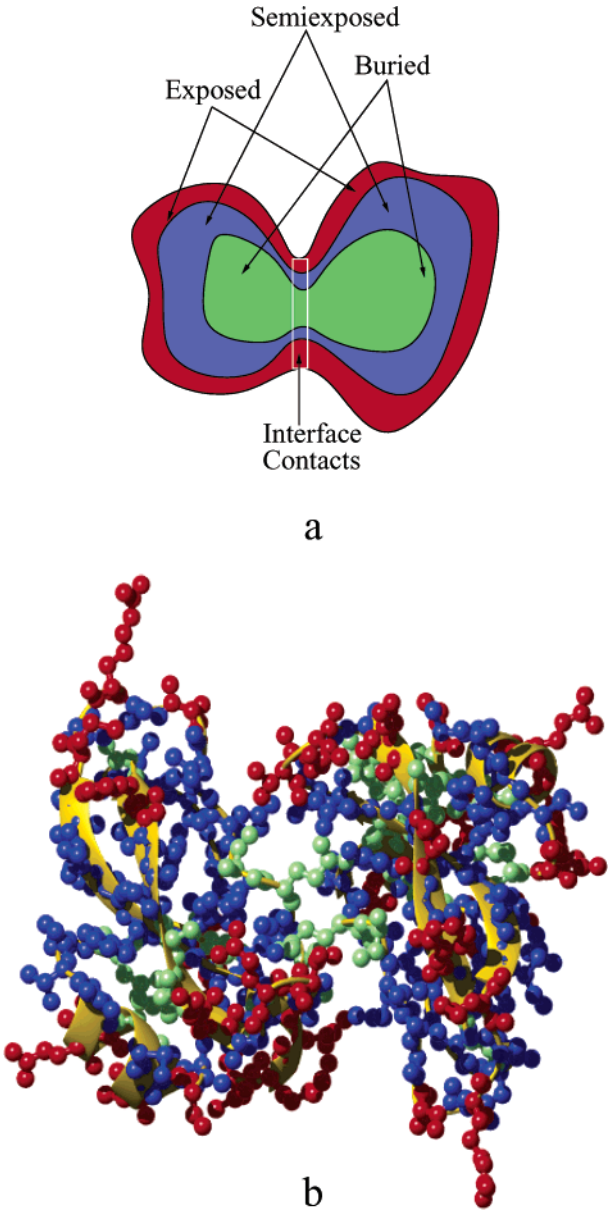

Figure 3. (A) Protein residues were partitioned into three layers: completely buried (green); semiexposed (blue); very exposed (red). Flexible binding reference states include all three layers (green, blue, and red). Semi-Flexible binding reference states include only semiexposed and exposed layer residues (blue and red). Rigid binding reference states include only the exposed layer residues (red). (B) A schematic representation of a representative database protein complex (PDB code 1aap) where the coloring scheme described above was applied using the surface accessible area (SAS) algorithm to partition the protein into layers. The surface accessibility thresholds were chosen in such a way, that, on average, for the database proteins, 40% of all residues were found in the buried (green) layer, 42% of all residues were found in the semiexposed (blue) and 18% of all residues were found in the exposed (red) layer.

model the effect of protein flexibility, one recognizes that deeper lying, core residues may participate (transiently) in binding only if the protein is sufficiently flexible. Thus, we have partitioned each protein into three layers based on the degree of residue burial (Figure 3). We then use various combinations of these layers to model corresponding trap types of different flexibility (additional details are given in Figure 3). The resulting three trap models, which we call Flexible, Semi-Flexible, and Rigid, refer to the degree of flexibility of the residues of partner proteins prior to binding, not to the ordering of surface water molecules.

Definition of Direct and Water-Mediated Contact Types. Several alternatives exist for the definition of direct contacts between pairs of residues in protein folding and protein association literature.[29,40,41] Among more elaborate schemes, the closest distance between all atoms of the given pair of residues is calculated,[41] and the residues are considered in contact if that distance is lower than some predefined threshold value (typically 2−4 Å). Similarly, the distance between positions of geometrical centers or centers of mass may be calculated,[29] and again compared against some threshold distance. Perhaps the most commonplace, yet the simplest practice is to measure the distance between amino acid C_β (C_α for Gly) atoms,[40] using typical threshold values between 6 and 8 Å. Given the coarse grained nature of the potential used in this work, we have used the latter definition for Direct contacts, choosing 6.5 Å as the customary *upper* threshold distance (i.e., only residue pairs having C_βs closer than 6.5 Å were considered to be in contact).

There is much less prior work on the defining criteria for water-mediated pair interactions in the context of coarse-grained protein folding or protein association studies. At first, this does not seem to be an easy task. One might choose to use only the crystallographic waters to define through-water interactions. Interface waters are, however, often too disordered to be resolved crystallographically. Furthermore, as discussed earlier, the interplay between proteins and waters may be exerted through water molecules with short residence times and that are hence invisible to conventional X-ray crystallography.

When considering possible alternatives for a definition of water-mediated interactions purely from the protein's crystal structure (but not taking into account crystallographic waters), we have found two significant constraints that helped us limit our choices. First, in order for two residues to interact through water, they both must have access to surface water. This, for instance, excludes the significant fraction of interactions in the center of protein interfaces that are "through-protein" in character. In practice, we have computed the degree of burial, i.e., surface accessible area (SAS), for each residue in each protein complex.[42] A residue pair was considered as a *candidate* water-mediated contact only if both residues in the pair had at least some exposure to water (i.e., if their surface was buried by less than 95%).

Second, two residues in a water-mediated interresidue contact must be further away than when in a direct contact, thus making 6.5 Å as a possible *lower* threshold value for such contacts (and still consistent with our definition of direct contacts). We have also considered 7.8 Å as the *lower* threshold distance for water mediated interactions. As for the *upper* threshold distance we have considered 9.5 and 10.8 Å, thus allowing on average from one to two water molecules to mediate the interactions. Overall we have examined three interval definitions of water-mediated contacts: from 6.5 to 9.5 Å, from 7.8 to 9.5 Å, and finally from 7.8 to 10.8 Å. We have found that our results are rather insensitive to the precise interval definition. The 7.8 Å to 9.5 Å interval definition did produce a slightly larger recognition signal compared to the other two intervals and was chosen for reporting our results.

The robustness of our results with respect to several different definitions of water-mediated contacts as well as the consequent physical interpretation of the obtained interaction matrixes (discussed in detail below), confirm, a posteriori, the soundness of the criteria outlined above.

In the next section, we explore the details of the optimization strategy used to derive the interaction matrixes for direct and water-mediated

(38) Romero, P.; Obradovic, Z.; Li, X.; Garner, E.; Brown, C. J.; Dunker, A. K. *Proteins: Struct. Funct. Genet.* **2001**, *42*, 38−48.
(39) Uversky, V. N. *Protein Sci.* **2002**, *11*, 739−756.
(40) Dima, R. I.; Settanni, G.; Micheletti, C.; Banavar, J. R.; Maritan, A. *J. Chem. Phys.* **2000**, *112*, 9151−9166.
(41) Moont, G.; Gabb, H. A.; Sternberg, M. J. E. *Proteins: Struct. Funct. Genet.* **1999**, *35*, 364−373.
(42) Eisenhaber, F.; Lijnzaad, P.; Argos, P.; Sander, C.; Scharf M, M. *J. Comput. Chem.* **1995**, *16*, 273−284.

contacts in the context of three different types of trap states (i.e., $2 \times 3 = 6$ contact potentials, overall).

Optimization of Potentials using Energy Landscapes. For a given protein complex interface, we define a pairwise additive Hamiltonian, $H = \sum_i^{210} \gamma_i \lambda_i$, where the index i runs for all unique combinations of amino acids (210 for a 20-letter code), γ_i indicates the strength of the corresponding interaction potential, and λ_i indicates the number of interresidue pairs of a particular type. Then, for a particular protein complex, we compute the differential between the averaged energy of the native configurations, $\langle H \rangle_n$, and the average energy of the nonnative ensemble, $\langle H \rangle_u$. We have used the native conformations of protein complexes, as found in the corresponding crystal structures, to calculate $\langle H \rangle_n$. To generate ensembles of non-native configurations, we have used a protein sequence permutation procedure often employed in the development of coarse-grained contact potentials.[43] For each protein complex, we have constructed 10 000 (to ensure statistical convergence) decoy sequences threaded on the same native structure. We have used this ensemble of non-native configurations to compute the mean energy $\langle H \rangle_u$, and the energy variance, $\langle (H - \langle H \rangle)^2 \rangle_u$, of the corresponding denaturated ensemble. Each decoy sequence was obtained from the native sequence by the following permutation scheme. The native interface direct and/or water-mediated contacts were randomly permuted with: 1. All other residues in the protein complex (to obtain the Flexible potentials); 2. Residues in the Semi-Flexible layer (to obtain the Semi-Flexible potentials); 3. Residues in the Exposed Surface layer (to obtain the Rigid potentials). For each run, these permutations were carried out to such a degree that no original native contacts were present in the interface. Notice, that this protocol conserves the amino acid composition of the original sequence, which in turn is important for picking up the strong compositional signals, as discussed below.

Nonthermal averaging of the denaturated ensemble configuration energies, i.e., $\langle H \rangle_u$, sets the origin of the energy axis. Consequently, $\langle H \rangle_n - \langle H \rangle_u$ is indicative of the (designed) energy gap between the native configuration and the nonnative ensemble of states. Similarly, the variance $\langle (H - \langle H \rangle)^2 \rangle_u$ is suggestive of the ruggedness of the nonnative ensemble energy landscape, i.e., it sets the absolute magnitude of the energy scale. The energy gap and the ruggedness, along with the configurational entropy, characterize gross features of the binding energy landscape (see the earlier discussion). To make that landscape smoother and more energetically downhill toward the native structure, the $z = (\langle H \rangle_n - \langle H \rangle_u)/\sqrt{\langle (H - \langle H \rangle)^2 \rangle_u}$ is to be maximized.[27] This optimization condition may be systematically improved using higher order terms in an elaborate cumulant expansion scheme described earlier.[27]

Next, we write the energy gap as $\langle \Delta H \rangle = \sum_i A_i \gamma_i$, where $A_i = \langle \lambda_i \rangle_n - \langle \lambda_i \rangle_u$, i.e., the difference between native and nonnative occupation frequencies for a particular contact type i. Similarly, $\langle \Delta H^2 \rangle = \sum i,j \; \gamma_i B_{ij} \gamma_j$, where $B_{ij} = \langle \lambda_i \lambda_j \rangle_u - \langle \lambda_i \rangle_u \langle \lambda_i \rangle_u$. The optimization of $z = \sum_i A_i \gamma_i / \sum_{i,j} \gamma_i B_{ij} \gamma_j$ leads to a system of linear equations $\bar{B}\bar{\gamma} = \mu \bar{A}$ (μ is an undetermined Lagrange multiplier which determines the energy scale), which must be solved for $\bar{\gamma}$. To avoid spurious noise when inverting $\bar{B}$, $\bar{B}$ is first diagonalized and its smallest eigenvalues are filtered out. As may be seen from the functional form of z, the potential derived from its extremization is defined up to a scale factor.

The $\bar{B}$ matrix and the $\bar{A}$ vector were defined above for a particular protein complex. When many such complexes are used for training the potential, the highly overdetermined set of equations may be solved in least-squares sense using Singular Value Decomposition (SVD). Alternatively, one may average the $\bar{B}$ matrix and the $\bar{A}$ vector over the set of training proteins prior to solving for $\bar{\gamma}$. We obtained nearly identical potentials with both methods. However, when a self-consistent low-temperature optimization is carried out,[27] then we would expect these two procedures to produce two related, but distinct potentials.

Selection of Training and Test Protein Complexes. A set of 276 protein complexes was randomly selected from the nonredundant database of more than 500 protein complexes compiled by Ben-Tal and co-workers.[44] Those 276 complexes were further randomly partitioned into 222 training and 54 test protein complexes. Protein complexes in the test set are unrelated to the ones in the training set, thus providing a relatively objective way of evaluating the recognition power of the derived potentials (all results presented in the Tables and Figures refer to the *test* set calculations).

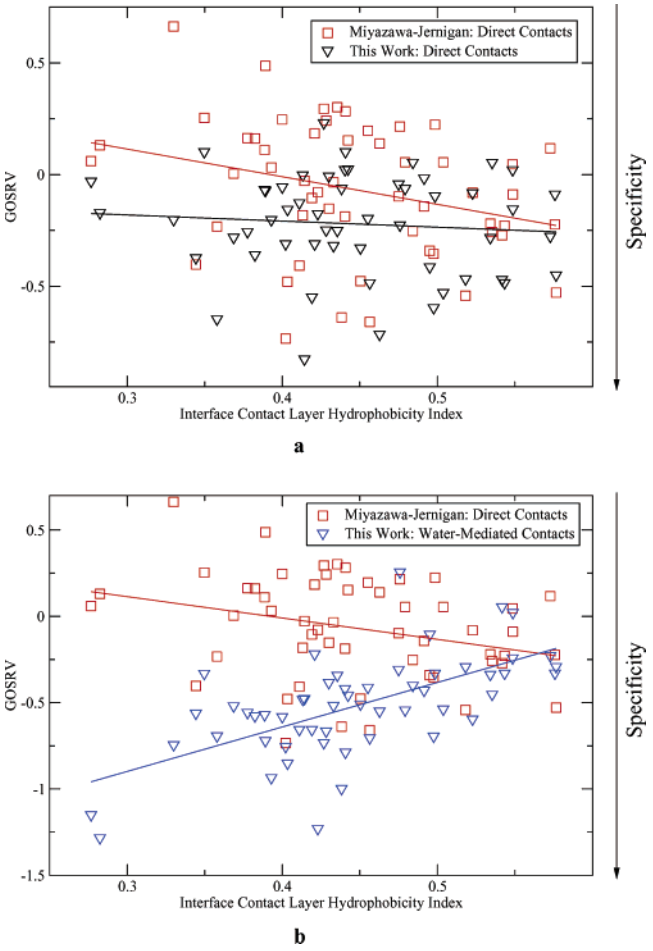

Figure 4. Degree of recognition of protein complex interfaces as a function of the interface contact layer hydrophobicity index, which was defined for each protein complex by averaging over residues involved in direct and water-mediated contacts using the Pacios' hydrophobicity scale.[45] Flexible binding trap model is used. Solid lines indicates the linear regression fit for the corresponding data. Red rectangles: Miyazawa-Jernigan potential applied to direct contacts. (a) Black triangles: the potential derived in this work applied to direct contacts. (b) Blue triangles: the potential derived in this work applied to water-mediated contacts.

Results and Discussion

Comparative Analysis of Direct and Water-Mediated Pair Potentials. One intriguing question is whether already available protein folding contact potentials can be applied, as is, to the binding problem. Using the popular Miyazawa-Jernigan (MJ) potential[30] we have computed (when discriminating against Flexible trap states) the ratio of the binding energy gap to the square root of the non-native energy variance (GOSRV) for a collection of protein complexes as a function of the interface contact residues hydrophobicity index (Figure 4). As is evident

(43) Ramanathan, S.; Shakhnovich, E. I. *Phys. Rev. E* **1994**, *50*, 1303−1312.

(44) Glaser, F.; Steinberg, D. M.; Vakser, I. A.; Ben-Tal, N. *Proteins: Struct. Funct. Genet.* **2001**, *43*, 89−102.

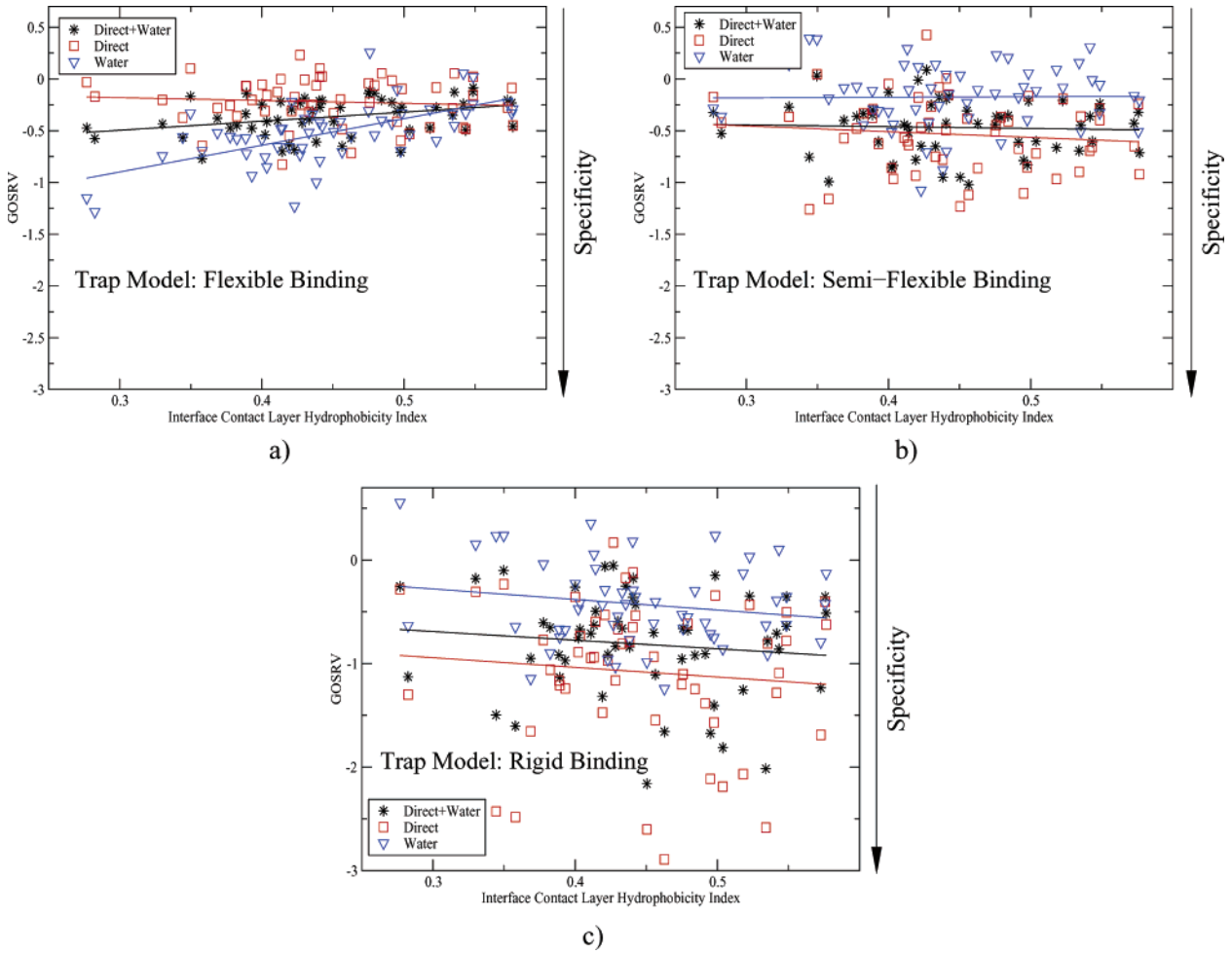

Figure 5. Degree of recognition of protein complex interfaces as a function of the interface contact layer hydrophobicity index (the same as in Figure 4) using the potentials derived in the current work. Solid lines indicates the linear regression fit for the corresponding data. (a) Calculated with direct, water-mediated and combined contact potentials discriminating against the Flexible binding trap model; (b) Calculated with direct, water-mediated and combined contact potentials discriminating against the Semi-Flexible binding trap model; (c) Calculated with direct, water-mediated and combined contact potentials discriminating against the Rigid binding trap model.

from Figure 4, the MJ potential discriminates well for the hydrophobic interfaces, giving a funneled binding landscape, however, hydrophilic interfaces are largely antifunneled with this potential (the Miyazawa-Jernigan potential performs noticeably better when Semi-Flexible and Rigid trap states are considered). This trend may be rationalized by recognizing the similarity of hydrophobic interfaces to protein interiors (discussed below), thus a protein folding potential might be expected to do well for hydrophobic interfaces. But are the hydrophilic interfaces not funneled in real life or is the folding-based potential inadequate? To see if it is possible to discriminate more uniformly for the cases of both hydrophobic and hydrophilic interfaces, we have derived direct contact potentials with three different trap states using the optimization scheme described earlier. In addition, to examine the possibility of additional recognition in hydrophilic interfaces, we have also optimized sets of interaction potentials for water-mediated interface contacts (see the Theory section for details). All parameters were derived from a nonredundant set of 222 protein complexes and independently tested on a set of 54 unrelated protein complexes. The recognition results for the test proteins using the potentials optimized in the current work are compiled in Figure 5.

To facilitate the subsequent discussion we first carry out Principal Component Analysis of the pair-potential interaction matrixes,[46] both for the MJ interaction matrix and the potential derived here. Through the Principal Component vectors one can decompose each element of the interaction matrix as the following sum: $\Gamma_{ij} = \lambda_A{}^*A_i{}^*A_j + \lambda_B{}^*B_i{}^*B_j + \lambda_C{}^*C_i{}^*C_j + ...$, where $\lambda_A,\lambda_B,\lambda_C$, ... are eigenvalues of the interaction matrix Γ and A_i, B_i, C_i, ... are the corresponding eigenvector elements. When the Miyazawa-Jernigan interaction matrix is reconstructed in this way, the truncation of the eigenvector expansion at the first term already produces a matrix that is 86% correlated with the original matrix, whereas the first two eigenvectors produce a matrix 99% correlated with the original matrix.[46] The two basis vectors A and B that thus span the 20 × 20 MJ-matrix correlate to a large degree with a typical hydrophobicity index. Eigenvector A's large magnitude elements correspond to more hydrophobic (H) residues, whereas eigenvector B's large magnitude elements correspond to more polar (P) residues. Thus, the MJ potential manifests itself largely as an HP scheme, an approach often used in protein folding studies.[47]

(45) Pacios, L. F. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1427−1435.
(46) Tang, H. L. C.; Wingreen, N. S. *Phys. Rev. Lett.* **1997**, *79*, 765−768.
(47) Dill, K. A. *Protein Sci.* **1999**, *8*, 1166−1180.

Because each contact layer as well as the reference states, contains a unique 20-letter vector of amino acid populations, these population vectors may be correlated with the corresponding interaction matrix eigenvectors ("H" and "P" vectors in the case of the MJ potential). We may think of these eigenvectors as representing idealized "canonical" amino acids that associate quantitatively a physical property with a particular energy value. This physical property is a combination of basic and composite physicochemical descriptors such as charge, hydrophobicity, side chain size, and more subtle attributes such as polarizability and side chain branching. Although we optimize two-body potentials, the resulting potentials incorporate highly many-body effects (e.g., the hydrophobic effect) in some effective way, which in turn translates into specific combinations of traits for various "canonical" amino acids. In a discussion below, we briefly characterize some of the dominant eigenvectors obtained for the interaction matrixes derived in this work.

The "canonical" amino acid representation provides a convenient description for coarse-graining the protein into larger regions (e.g., layers). In the case of various protein layers, the differential in populations of "canonical" amino acids in the native and reference state layers determines the strength of the compositional recognition signal. The signal from the MJ potential in protein folding is dominantly compositional, mostly determined by the abundance of "H" residues in the protein interior and depletion of such residues on the protein surface and not as their specific pairing (as mentioned earlier, the first two eigenvectors of the MJ interaction matrix produce a matrix 99% correlated with the original matrix[46]). Similarly, the performance of the MJ potential for binding interfaces depends on the contrast between the relative abundances of the "H" and "P" residues in the contact layer as opposed to the abundances in the trap states. The abundance of "P" residues in hydrophilic interfaces renders the native interaction energy destabilized compared to more "H"-rich Flexible trap states, rationalizing our earlier observation that the MJ potential performs relatively well only for hydrophobic interfaces when the Flexible trap states are discriminated against (Figure 4). However, when the Rigid trap state model is considered, the reference state are much more "P" in character, which significantly enhances the MJ potential recognition signal (not shown).

Although the direct contact potentials derived in this work are correlated with the MJ potential, PCA of the direct contact interaction matrix reveals that approximately 10 eigenvectors are needed for reconstructing the original interaction matrix to a high degree of accuracy, i.e., it is not reducible to a plain HP scheme. To understand better the compositional dependence of recognition for direct and water-mediated potentials (discussed below), we calculated the correlation coefficients between the vectors of amino acid populations in these layers and the first three leading eigenvectors (named A, B, and C for convenience) for each particular interaction matrix (i.e., "H" and "P" in the MJ matrix are replaced by "A", "B", and "C"; Figure 6). We have alluded earlier to the fact that these "canonical" amino acids may be described by a combination of physical traits. The dominant eigenvector A strongly correlates with the conventional hydrophobicity regardless of the particular combination of contact and reference states. The second dominant eigenvector B also correlates with the hydrophobicity, but to a somewhat lesser degree. In the case of Flexible and Semi-Flexible trap

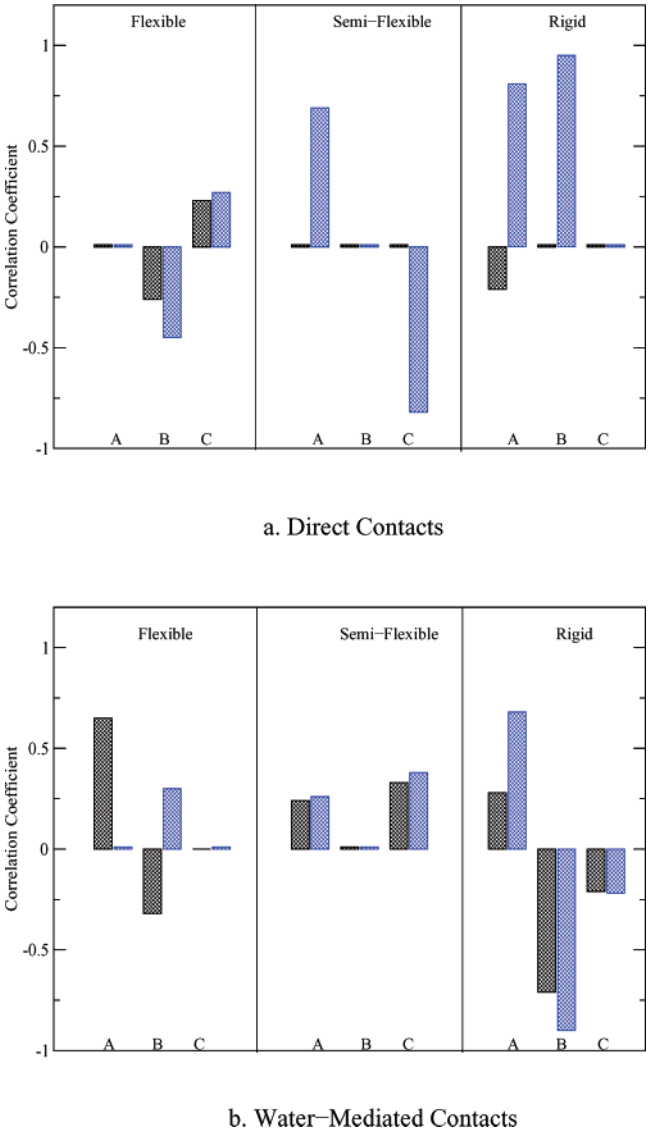

Figure 6. Correlation coefficients between the various 20-letter amino acid population vectors (see below) and the leading Principal Component Vectors (named for convenience "A", "B", and "C") of the relevant interaction matrixes. Gray histograms refer to the correlations calculated with the contact layer amino acid composition vectors (direct contacts are given in the top plot and water-mediated contacts are given in the bottom plot). Blue histograms refer to the correlations calculated with the reference layer amino acid composition vectors for Flexible, Semi-Flexible, and Rigid binding trap models.

states, the "canonical" amino acids A and B have reversed polarity on the HP-scale with respect to each other, while in the case of Rigid trap states they are of the same polarity. The third dominant eigenvector C as well as the remaining ones are not correlated with the HP-scheme and carry complementary information.

As discussed above, one expects that if a certain "canonical" amino acid is abundant in the contact layer and is depleted in the reference state, then a strong recognition signal should result. For instance, when native water-mediated contacts are compared with the Semi-Flexible trap states, the presence of approximately the same amount of "canonical" A and C amino acids renders recognition challenging (middle panel in Figure 6b). This is also consistent with almost the same amino acid population vectors (correlation coefficient 0.98) for the water-mediated native contact layer and the Semi-Flexible trap states. This suggests

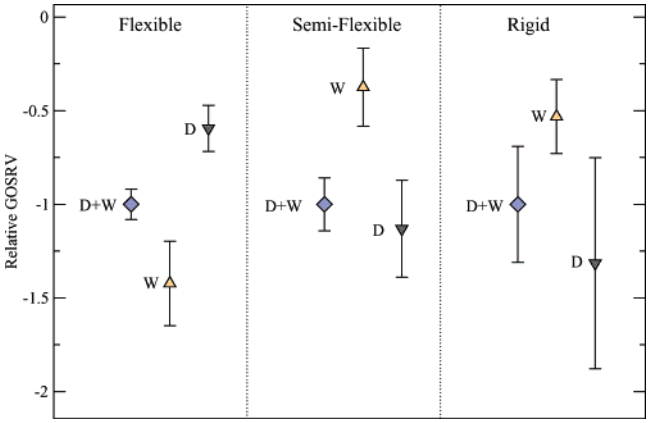

Figure 7. Relative gap over square root variance (GOSRV) ratio averages and corresponding standard deviations calculated for the test set proteins using the potentials derived in this paper with three models of trap states (Flexible, Semi-Flexible, and Rigid). The relative GOSRV values were obtained in each region by scaling the corresponding absolute values in such a way that the combined potential (sum of direct and water potentials) GOSVR equals −1 (see text for details). D (direct potential), W (water potential), D+W (a combination of direct and water potentials).

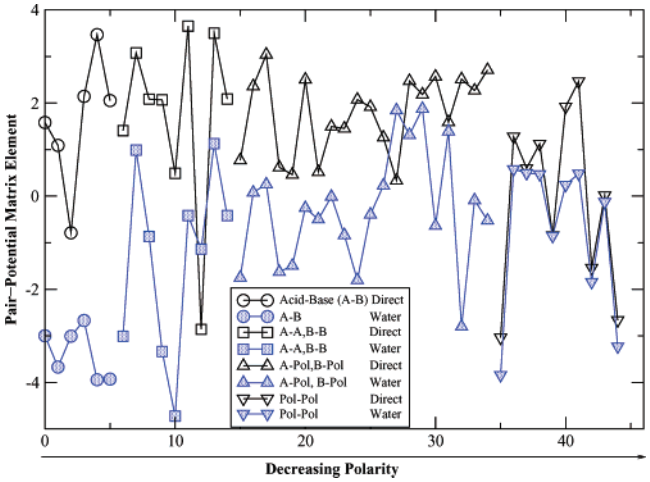

Figure 8. Direct (black) and water-mediated (blue) pair-potential matrix elements among pairs of non-hydrophobic residues arranged in the order of decreasing polarity. A−acidic, B−basic, Pol−polar. These potentials were optimized against the Semi-Flexible binding trap states.

that there should be no strong dependence of the recognition signal as a function of interface hydrophobicity index, consistent with data in Figure 5a.

Similar analysis for the other five panels in Figure 6 allows one to rationalize the strengths of the corresponding recognition signals (Figure 7). For instance, the water-mediated contacts contain a very strong recognition signal when discriminating against the Flexible states (left panel in Figure 7), because the "canonical" amino acid composition is very different in the contact and reference state layers (left panel in Figure 6b). Conversely, for the direct contacts the discrimination is high against Semi-Flexible and Rigid trap states (middle and right panels in Figure 7), as these reference states provide the highest compositional contrast with the direct contact layer population vector (middle and right panels in Figure 6a). Overall, in terms of the stability of recognition against all different types of trap states (which in turn model classes of trap proteins in the cell), the combination of direct and water-mediated potentials performs the best. As evident from the analysis of the quality of recognition for various contact/reference-state pairs (Figure 6), there does not exist an "absolute" interaction matrix for the given contact potential (direct or water-mediated), but the choice of the reference state model significantly influences the optimal choice of the interaction potential (although the resulting matrixes are still noticeably correlated).

As we have discussed earlier, when the interface is hydrophobic, the MJ potential provides strong discrimination of the native interface from the Flexible trap states. The same interaction potential does not yield good discrimination for hydrophilic interfaces (Figure 4). However, the direct contact potential derived in this work discriminates against Flexible trap states no matter what hydrophobicity of the interface (Figure 5a), which in turn is rationalized by the similarity of the populations of the "canonical" amino acids in the contact and reference state layers (left panel in Figures 6a). On the other hand, when the recognition against Flexible trap states is examined for the water-mediated potential, a strong dependence of recognition on the interface hydrophobicity index is observed (5a). For this reference state, water-mediated native contacts are recognized strongest for the most hydrophilic interfaces, which is indicative of a strong primary composition signal. Interestingly enough, the combination of direct and water-separated potentials improves the stability not only against three different type of trap states as mentioned previously, but also against hydrophobicity index variability of interface contacts (Figures 5).

Physical Interpretation of Water-Mediated Interactions. The differences in direct contact and water-mediated contact interaction matrixes may be rationalized by analyzing the physical nature of forces distinguishing these interactions. When a water molecule is expelled from a water-mediated contact to form a direct contact several factors contribute to the positive or negative free energy change. Among the most important factors are the following: (1) Hydrophobic free energy change; (2) Backbone and side chain entropy loss; (3) Desolvation penalty for charged and highly polar residues; (4) Free energy change due to the interresidue electrostatic interactions. The entropy and desolvation terms act to counterbalance the first effect which is usually stabilizing. The outcome of the pair electrostatics contributions depends largely on the specific nature of the interacting residues charge distributions. As an example, when two oppositely charged groups (acid−base pair) are brought to form a direct contact, the overall free energy change depends strongly on the exact compensation scheme between the favorable electrostatic interactions and a large desolvation penalty. The relative values of the corresponding direct and water-mediated matrix elements (Figure 8) suggest that the desolvation penalty is very large, thus, incomplete desolvation realized in a water-mediated contact may be an optimal solution. In addition, as the total charge of two interacting residues diminish (i.e., charged-charged to charge−polar to polar−polar), one expects the contribution of the desolvation penalty to become less critical, which is consistent with the growing similarity between the values of the direct and water-mediated potential matrix elements as a function of decreasing total charge (Figure 8).

When two carboxylic (or two basic) groups are near to each other, the ionization of the first functional group makes the ionization of the second similar functional group more difficult; therefore, such pairs are not expected necessarily to be in

dianionic (dicationic) states (i.e., they are probably better regarded as charged-polar or polar−polar pairs). Among the corresponding acidic-acidic/basic-basic contacts in the interresidue interaction matrix with Semi-Flexible reference states, only the His−His direct contact has a negative value, whereas for water-mediated contacts six pairings are stabilizing, indicating that complete desolvation is not energetically favorable in the presence of charged groups.

The differences between direct and water-mediated interactions for polar−polar partners are small (Figure 8). The hydrophobic interactions are however much more stabilizing for the direct contacts than for the water-mediated ones, (data not shown), in agreement with common sense intuition.

Conclusions

In summary, we have used the energy landscape theory of protein folding/binding to derive several sets of direct and water-mediated contact potentials for protein native interface recognition. Our results clearly show that water-mediated contacts may carry significant information content complementary to direct contact information content. The combination of these two pair potentials provides smooth discrimination against a variety of trap states that mimic the obstacles to be avoided in dense cellular environment. The PCA analysis of the corresponding interaction matrixes indicates that the differential in "canonical" amino acid composition between the contact layer and the reference states determines the magnitude of recognition for the given potential. We find that both direct and water-mediated potentials derived in this work go well beyond the familiar two letter HP code and carry additional information. The comparative analysis of the direct and water-mediated contact matrixes can be rationalized largely on the grounds of a significant desolvation penalty for charged groups. The kinetic consequences of the interresidue specific water-mediated interactions on both the binding and folding processes is under study.

Acknowledgment. Garegin A. Papoian gratefully acknowledges the National Institute of Health for its generous support of this work through a National Institute of Health Postdoctoral Fellowship Award. Johan Ulander thanks the Swedish Research Council and San Diego Supercomputing Center for providing postdoctoral fellowships. The effort of Peter G. Wolynes in concepts of protein folding is supported through NIH Grant No. 5R01GM44557. We have made an extensive use of C++ Biochemical Algorithms Library (BALL) in the computational part of our study.[48]

Supporting Information Available: Six tables with the optimized potentials used in this work, a table listing the set of 222 training proteins and a table listing 54 test proteins are provided as Supporting Information. This material is available free of charge via the Internet at http://pubs.acs.org.

JA034729U

(48) Kohlbacher, O.; Lenhof, H. P. *Bioinformatics* **2000**, *16*, 815−824.