7078–7087 Nucleic Acids Research, 2008, Vol. 36, No. 22 Published online 6 November 2008
doi:10.1093/nar/gkn868

Common physical basis of macromolecule-binding sites in proteins

Yao Chi Chen[1,2] and Carmay Lim[1,2,*]

[1]Institute of Biomedical Sciences, Academia Sinica, Taipei 115 and [2]Department of Chemistry, National Tsing Hua University, Hsinchu 300 Taiwan

Received July 10, 2008; Revised September 18, 2008; Accepted October 18, 2008

ABSTRACT

Protein–DNA/RNA/protein interactions play critical roles in many biological functions. Previous studies have focused on the different features characterizing the different macromolecule-binding sites and approaches to detect these sites. However, no common unique signature of these sites had been reported. Thus, this work aims to provide a ‘common’ principle dictating the location of the different macromolecule-binding sites founded upon fundamental principles of binding thermodynamics. To achieve this aim, a comprehensive set of structurally nonhomologous DNA-, RNA-, obligate protein- and nonobligate protein-binding proteins, both free and bound to their respective macromolecules, was created and a novel strategy for detecting clusters of residues with electrostatic or steric strain given the protein structure was developed. The results show that regardless of the macromolecule type, the binding strength and conformational changes upon binding, macromolecule-binding sites are energetically less stable than nonmacromolecule-binding sites. They also reveal new energetic features distinguishing DNA- from RNA-binding sites and obligate protein- from nonobligate protein-binding sites in both free/bound protein structures.

INTRODUCTION

Protein–macromolecule interactions play critical roles in many biological functions, including gene transcription and translation, signal transduction, enzyme regulation and immune response. Since protein–macromolecule interactions are central to various processes in a living cell, a detailed understanding of protein–macromolecule interactions is critical. Such an understanding has benefited from the increasing number of 3D structures of protein–macromolecule complexes that are being solved (1,2). These solved complexes in turn have spurred research efforts toward characterizing and detecting protein–macromolecule binding sites. The predicted macromolecule-binding site(s) of a given protein enable specific residues to be mutated and their effects on binding analyzed, thus aiding functional annotation of new structures from structural genomics projects. They also help to reduce conformational search in docking a macromolecule to its target protein, thus reducing the number of false positives (3). In the following, we summarize for each macromolecule ligand (DNA/RNA/protein), the known characteristics of the respective macromolecule-binding site on the protein and the key approaches used to detect the site(s).

In binding DNA, proteins achieve binding (i) ‘affinity’ through favorable charge–charge interactions between positively charged arginine and lysine side chains and the negatively charged DNA phosphate backbone and (ii) ‘specificity’ through directional hydrogen bonding and van der Waals (vdW) interactions (4–6). Hence, positively charged residues are enriched, whereas negatively charged residues are depleted in DNA-binding sites. Given the 3D structure of a DNA-binding protein (DBP), the DNA-binding site has been identified using mostly electrostatic potentials in conjunction with other parameters such as surface accessibility, the protein surface shape, and amino acid (aa) conservation (7–9) as well as neural network (10) and support vector machine (11). It has also been identified using support vector machine (12) given only the 1D sequence of a DBP.

In binding RNA, proteins employ a recognition strategy similar to DBPs to bind ‘double’-stranded RNA; in addition, they employ cavities to accommodate unstacked ‘single’-stranded RNA bases (13–21). Given only the RNA-binding protein (RBP) sequence, its RNA-binding site has been identified using machine learning approaches such as support vector machines (12), a neural network classifier (22) and a Naïve Bayesian classifier (23). If the

[*]To whom correspondence should be addressed. Tel: +886 2 2652 3031; Fax: +886 2 2788 7641; Email: carmay@gate.sinica.edu.tw

structure of the RBP is available, the RNA-binding site(s) can been identified using residue and residue pairing preferences at the protein–RNA interface in conjunction with the relative residue conservation (24) or the latter combined with electrostatic energies and the protein surface shape (25). Interestingly, although RNA/DNA-binding sites are more evolutionary conserved than the rest of the protein, considering only conservation led to many false positives and is thus a poor predictor of these sites.

Other than binding DNA/RNA, proteins may also interact with one another through various recognition strategies. These protein–protein interactions can generally be divided into 'obligate' interactions of protomers that cannot exist on their own *in vivo*; e.g. the Arc repressor homodimer, human cathepsin D heterodimer and multisubunit enzymes and 'nonobligate' interactions of protomers that can exist independently; e.g. intracellular signaling complexes and antibody–antigen, receptor–ligand/hormone and enzyme–inhibitor/substrate complexes (26). Compared to 'nonobligate' interfaces, obligate interfaces are predominantly nonpolar and larger with more contacts and conserved residues (27–29). In general, compared to 'noninterface' protein surfaces, protein–protein interfaces are enriched with nonpolar and aromatic residues as well as arginine but depleted in the other charged residues (30,31). The interface residues, which form vdW and electrostatic interactions between complementary surfaces (32), are more conserved (33) and solvent accessible (34), but less flexible (35) than noninterface surface residues. Given the 3D protein structure, protein-binding sites have been located based on their shape, electrostatics and hydrogen-bonding complementarities (36,37). They have been identified using linear regression (38), scoring function (34,39–43), support vector machine (44–46), neural network (33,47,48) or Bayesian networks (49) and the combination of parameters such as side chain energy, solvation potential, residue propensity and conservation, hydrophobicity, accessible surface area and different structural indexes.

Whereas previous studies have revealed different characteristics for the different macromolecule-binding sites, we aim in this work to provide a 'common' physical basis for DNA-, RNA-, obligate protein- and nonobligate protein-binding sites. Although DNA/RNA/protein-binding sites are more evolutionary conserved than the rest of the protein surface, conservation cannot serve as the common signature for such sites as it may arise not only for binding macromolecules, but also for structural purposes (see above). Considering fundamental principles of binding thermodynamics, however, could provide a common physical basis for the different macromolecule-binding sites: aa residues involved in binding a given type of macromolecule should make a net favorable enthalpic and/or entropic contribution to the binding free energy. In the absence of their binding partner and 'solvent', these residues possess suboptimal hydrogen-bonding interactions and packing, thus their 'gas-phase' electrostatic and vdW interactions should be less favorable than those of residues not involved in any binding interactions. However, no systematic studies dissecting the individual energetic contributions of a comprehensive set of DBPs, RBPs, obligate and nonobligate proteins have addressed the following questions (to the best of our knowledge): (i) Regardless of the binding macromolecule type and the conformational changes accompanying binding, is the binding site generally less energetically stable than the nonmacromolecule-binding regions? (ii) If so, does unfavorable electrostatic or vdW energy dictate the observed higher energy of a macromolecule-binding site compared to the nonmacromolecule-binding regions? (iii) In particular, how do DNA-binding sites differ energetically from RNA-binding sites considering that these two types of sites share in common positively charged aa side chains interacting with negatively charged DNA/RNA phosphate backbone. (iv) Along a similar vein, how do obligate protein-binding sites differ energetically from nonobligate protein-binding sites considering that residues in both types of sites interact similarly with a protein?

Herein, we address the above questions by first collecting structurally nonhomologous protein structures, both free and bound to DNA, RNA, obligate proteins and nonobligate proteins. Given the free or bound protein 3D structure, we computed the 'relative gas-phase' electrostatic or vdW energy of each residue, which in turn was used to assign an electrostatic and a vdW rank to each residue and its surrounding, as described in the next section. The results reveal a common physical basis for the different macromolecule-binding sites, consistent with thermodynamics considerations. They also reveal key features distinguishing the different DNA-, RNA-, obligate protein- and nonobligate protein-binding sites that can provide useful guidelines in developing methods to detect different functional sites in proteins binding to more than one macromolecule. Notably, these findings were found to be independent of the conformational changes accompanying macromolecule binding.

MATERIALS AND METHODS

Data set of protein–macromolecule complexes

The structurally nonhomologous macromolecule-binding protein complexes were obtained as follows: For the DNA/RNA-binding protein data sets, all available ≤3-Å X-ray structures of proteins bound to DNA/RNA (but 'not hybrid' DNA/RNA) were obtained from the Protein Data Bank (PDB) (50). For the obligate and nonobligate protein data sets, ≤3-Å X-ray structures of proteins bound to an obligate/nonobligate protein including antigens (51) were obtained from the PPI-Pred server (46) and the PDB. These DNA/RNA/obligate protein/nonobligate protein-binding chains were then grouped according to their CATH codes (52). For each group of protein structures with the same CATH code, the structure with the best resolution was selected as the representative one. This yielded 76 DNA-binding, 72 RNA-binding, 88 obligate protein-binding and 77 nonobligate-protein-binding representative protein structures, whose PDB entries are listed in Supplementary Table S1.

Data set of ‘free’ macromolecule-binding proteins

The free protein structures corresponding to the above complex structures were obtained as follows: For each protein sequence in the bound data set, the SAS database (53) was searched for sequences sharing ≥90% sequence identity according to pairwise sequence alignments using CLUSTALW (54). Next, the PDB was searched for ≤3-Å X-ray free structures of the homologous proteins. If the free structures were available, the C^{α} root-mean-square deviation (RMSD) from the corresponding bound structure was computed using the SSAP program (55); the one with the largest C^{α} RMSD was selected as the representative free structure to assess effects of conformational change upon macromolecule binding. This yielded 41 DNA-binding, 24 RNA-binding, 14 obligate protein-binding and 38 nonobligate-protein-binding free protein structures, whose PDB entries and C^{α} RMSD are listed in Supplementary Table S1.

Definition of true macromolecule-binding residues

For each representative protein–macromolecule complex, the same criterion was used to assign the residues involved in binding the respective DNA/RNA/protein. An aa residue was considered to be a macromolecule-binding residue if it contains one or more nonhydrogen atoms within vdW contact or hydrogen-bonding distance to the nonhydrogen atom of its binding partner directly or indirectly via a bridging water molecule. For a given complex structure, the HBPLUS (56) program was used to compute all possible hydrogen bonds and vdW contacts, which are defined by a donor atom to an acceptor atom distance of 3.5 and between 3.5 and 4.0 Å, respectively.

Definition of nonmacromolecule-binding residues

Since a given protein may not only bind to its cognate macromolecule, but also to other ligands, nonmacromolecule-binding residues were defined as those not involved in binding any macromolecule. Residues binding other macromolecules were determined by examining homologous structures of a given protein, which were obtained by searching the SAS database (53) and the PDB for ≤3-Å X-ray complex structures of homologous proteins sharing ≥90% sequence identity with the given protein. These homologous complex structures were used to determine the residues involved in binding DNA/RNA/protein according to the above criteria. Thus, for each protein in a data set, residues ‘not’ assigned as binding DNA/RNA/protein according to both the representative and homologous complex structures were defined as nonmacromolecule-binding, whereas residues assigned as binding according to the ‘representative’ complex structure only were defined as binding to the macromolecule. Supplementary Tables S2A and B illustrate the assignment of true DNA-binding and nonmacromolecule-binding residues, respectively, in transcription initiation factor TFIID (PDB entry 1nh2-C) containing a 50-residue protein (aa 228–231; 241–286) bound to DNA.

Assignment of protonation states of ionizable residues

For a given protein, all Asp/Glu residues were deprotonated, while Arg/Lys residues were protonated. His residues were protonated if both side chain nitrogen atoms were within hydrogen-bonding distance to any aa acceptor atom. Otherwise, they were assumed to be neutral and the side chain nitrogen that is within hydrogen-bonding distance of an acceptor atom in the protein was protonated.

Energy decomposition of a given protein structure

For each l–aa protein, its structure was energy minimized with heavy constraints on all nonhydrogen atoms using the AMBER (57) program to relieve bad contacts. Based on the energy-minimized structure, the gas-phase electrostatic or vdW energy contributed by residue i in the ‘folded’ state, $E^{ele}{}_{i}$ or $E^{vdW}{}_{i}$, relative to that in a ‘reference’ state ($E'^{ele}{}_{i}$ or $E'^{vdW}{}_{i}$) was computed, where the reference state for residue i was defined as CH_3NH–aa_i–$COCH_3$. The change in the gas-phase electrostatic or vdW energy from the ‘reference’ state to the ‘folded’ state is given by:

$$\Delta E^{ele/vdW}{}_{i} = E^{ele/vdW}{}_{i} - E'^{ele/vdW}{}_{i} \qquad (1)$$

The gas-phase electrostatic and vdW energies were computed with the all-hydrogen-atom AMBER force field (58) with $\varepsilon = 1$ and no cutoffs using the AMBER (57) program.

Knowing $\Delta E^{ele/vdW}{}_{i}$, the average electrostatic or vdW energy contributions of aa i and its neighbors, $<\Delta E^{ele/vdW}>_i$, was computed from:

$$<\Delta E^{ele/vdW}>_i = \Sigma \Delta E^{ele/vdW}{}_{j}/N^{aa}{}_{i} \qquad (2)$$

where the summation in Equation (2) is over $N^{aa}{}_{i}$ residues, which include aa i and all residues j whose C^{α} atoms are within 10 Å of the C^{α} atom of aa i.

Electrostatic and vdW energy ranking of ‘each’ residue in a protein

Each residue of a l–aa protein was assigned an electrostatic rank ($\mathrm{Rank}^{ele}{}_{i}$), a vdW rank ($\mathrm{Rank}^{vdW}{}_{i}$) and a combined electrostatic and vdW rank ($\mathrm{Rank}^{ele+vdW}{}_{i}$) based on its $<\Delta E^{ele}>_i$, $<\Delta E^{vdW}>_i$ and $<\Delta E^{ele+vdW}>_i = <\Delta E^{ele}>_i + <\Delta E^{vdW}>_i$ energies, respectively. The l $<\Delta E^{x}>_i$ ($x = ele$, vdW, or $ele + vdW$) energies were ordered from the most negative to the least negative/most positive. These were used to rank the l aa residues from 1 to 10 such that residues with the top 10% most negative $<\Delta E^{x}>_i$ energies were ranked 1, residues with the next top 10% most negative $<\Delta E^{x}>_i$ values were ranked 2, etc. When $l/10$ is not an integer, each rank except the largest one (i.e. 1, 2, ..., 9) is associated with an integral $l/10$ residues, while the largest rank of 10 corresponds to the remaining residues in the l–aa protein. Supplementary Table S3A illustrates the electrostatic and/or vdW ranking of each residue in transcription initiation factor TFIID (1nh2-C).

Electrostatic and vdW energy ranking of 'macromolecule'-binding and 'nonmacromolecule'-binding residues in a protein

For a protein p in a given data set, knowing Rank_i^x ($x = ele$, vdW or $ele + vdW$) of each residue i, the respective mean ranking of the N_p^{+m} residues involved in binding a given type of macromolecule m (denoted by superscript '$+m$') and N_p^- residues *not* known to be involved in binding any macromolecule (denoted by superscript '$-$') were computed from:

$$\langle \mathrm{Rank}^x \rangle_p^{+m} = \sum_i \mathrm{Rank}_i^x / N_p^{+m} \tag{3a}$$

$$\langle \mathrm{Rank}^x \rangle_p^{-} = \sum_j \mathrm{Rank}_j^x / N_p^{-} \tag{3b}$$

where the summation in Equation (3a) or (3b) is over the N_p^{+m} and N_p^- residues in protein p, respectively (see Supplementary Table S3B).

Electrostatic and vdW energy ranking of 'different' macromolecule–binding sites

For each data set of proteins that bind a given type of macromolecule m (m = DNA, RNA, obligate/nonobligate protein), its average rank was computed as:

$$\langle \mathrm{Rank}^x \rangle^{+m} = \sum_p \langle \mathrm{Rank}^x \rangle_p^{+m} / N_m \tag{4a}$$

$$\langle \mathrm{Rank}^x \rangle^{-} = \sum_p \langle Rank^x \rangle_p^{-} / N_m \tag{4b}$$

where the summation in Equation (4a) or (4b) is over the N_m proteins in the data set. In addition, the number of macromolecule-binding or nonmacromolecule-binding residues with a given combination of electrostatic and vdW energy ranking, $N^{+m/-}[\mathrm{Rank}^{ele}, \mathrm{Rank}^{vdW}]$, was counted. Since the number of residues with the largest electrostatic or vdW rank is generally greater than the number of residues with other lower ranks, this number was normalized as $n^{+m/-}[\mathrm{Rank}^{ele}, \mathrm{Rank}^{vdW}] = N^{+m/-}[\mathrm{Rank}^{ele}, \mathrm{Rank}^{vdW}]/(N^{ele}[\mathrm{Rank}^{ele}] \times N^{vdW}[\mathrm{Rank}^{vdW}])$, where $N^{ele}[\mathrm{Rank}^{ele}]$ and $N^{vdW}[\mathrm{Rank}^{vdW}]$ are the numbers of residues in the dataset corresponding to a given electrostatic and vdW rank, respectively. For each data set, the frequency of a given combination of electrostatic and vdW energy ranking was calculated as

$$\begin{aligned} &\nu^{+m}[\mathrm{Rank}^{ele}, \mathrm{Rank}^{vdW}] \\ &= n^{+m}[\mathrm{Rank}^{ele}, \mathrm{Rank}^{vdW}] / \sum n^{+m}[\mathrm{Rank}^{ele}, \mathrm{Rank}^{vdW}] \end{aligned} \tag{5a}$$

for the macromolecule-binding residues, and

$$\begin{aligned} &\nu^{-}[\mathrm{Rank}^{ele}, \mathrm{Rank}^{vdW}] \\ &= n^{-}[\mathrm{Rank}^{ele}, \mathrm{Rank}^{vdW}] / \sum n^{-}[\mathrm{Rank}^{ele}, \mathrm{Rank}^{vdW}] \end{aligned} \tag{5b}$$

for the nonmacromolecule-binding residues, where the summation in Equation (5a) or (5b) is over the different rank combinations.

RESULTS

In each free/bound data set containing DBPs, RBPs, obligate proteins or nonobligate proteins, the protein's 3D structure was used to compute the electrostatic rank ($\mathrm{Rank}^{ele}{}_i$), vdW rank ($\mathrm{Rank}^{vdW}{}_i$) and combined electrostatic and vdW rank ($\mathrm{Rank}^{ele+vdW}{}_i$) of residue i and its surrounding, as described in 'Materials and methods' section. Each residue's electrostatic and/or vdW rank is an integer number, ranging from 1 to 10. A high $\mathrm{Rank}^{ele}{}_i$ or $\mathrm{Rank}^{vdW}{}_i$ rank means that residue i and its surrounding have relatively high electrostatic or vdW energy in the protein, respectively. Averaging the $\mathrm{Rank}^x{}_i$ values ($x = ele$, vdW or $ele + vdW$) of macromolecule-binding and nonmacromolecule-binding residues in a given protein p yields $<\mathrm{Rank}^x>^{+m}{}_p$ and $<\mathrm{Rank}^x>^{-}{}_p$, respectively [see Equation (3) and Supplementary Table S3B]. Averaging $<\mathrm{Rank}^x>^{+m}{}_p$ and $<\mathrm{Rank}^x>^{-}{}_p$ over the number of macromolecule m-binding proteins in the given dataset (N_m) yields $<\mathrm{Rank}^x>^{+m}$ and $<\mathrm{Rank}^x>^{-}$, respectively [see Equation (4)]. For a given macromolecule m, the $<\mathrm{Rank}^x>^{+m}$ values of all macromolecule-binding residues and the respective $<\mathrm{Rank}^x>^{-}$ values of all 'non'macromolecule-binding residues, as well as the difference between $<\mathrm{Rank}^x>^{+m}$ and $<\mathrm{Rank}^x>^{-}$ (denoted by $\Delta^x{}_m$) were computed from both the free and bound protein structures (see Table 1).

Electrostatic versus vdW energy ranking distributions in protein−macromolecule complexes

To determine if the distribution of electrostatic and vdW energy ranks of residues binding a given macromolecule differs from that of residues not known to bind any macromolecule, $\nu^{+m}[\mathrm{Rank}^{ele}, \mathrm{Rank}^{vdW}]$ and $\nu^{-}[\mathrm{Rank}^{ele}, \mathrm{Rank}^{vdW}]$ were computed for each bound data set, as described in 'Materials and methods' section. The $\nu^{+m}[\mathrm{Rank}^{ele}, \mathrm{Rank}^{vdW}]$ minus the random frequency, which is equal to 0.01, shows that the distribution of electrostatic and vdW energy ranks of macromolecule-binding residues is not uniform and differs from that of residues 'not' known to bind to any macromolecule (Figure 1). The binding site for a given macromolecule is characterized by unfavorable electrostatic interactions and/or steric clashes, as evidenced by the frequency of high electrostatic and vdW energy ranks, which is greater than that in nonmacromolecule-binding regions. For example, the $\nu^{+}[10,10]$ of residues binding DNA (0.049), RNA (0.034), obligate proteins (0.047) and nonobligate proteins (0.050) are greater than the $\nu^{-}[10,10]$ of nonmacromolecule-binding residues in DBPs (0.017), RBPs (0.017), obligate (0.016) and nonobligate (0.017) proteins, respectively. In contrast, all the $\nu^{-}[\mathrm{Rank}^{ele}, \mathrm{Rank}^{vdW}]$ values are very close to the random frequency.

Common feature of macromolecule-binding sites

To verify if macromolecule-binding sites are indeed energetically less stable than nonmacromolecule-binding sites regardless of the ligand type and the conformational changes accompanying binding, the electrostatic and/or vdW ranks of all residues binding a given

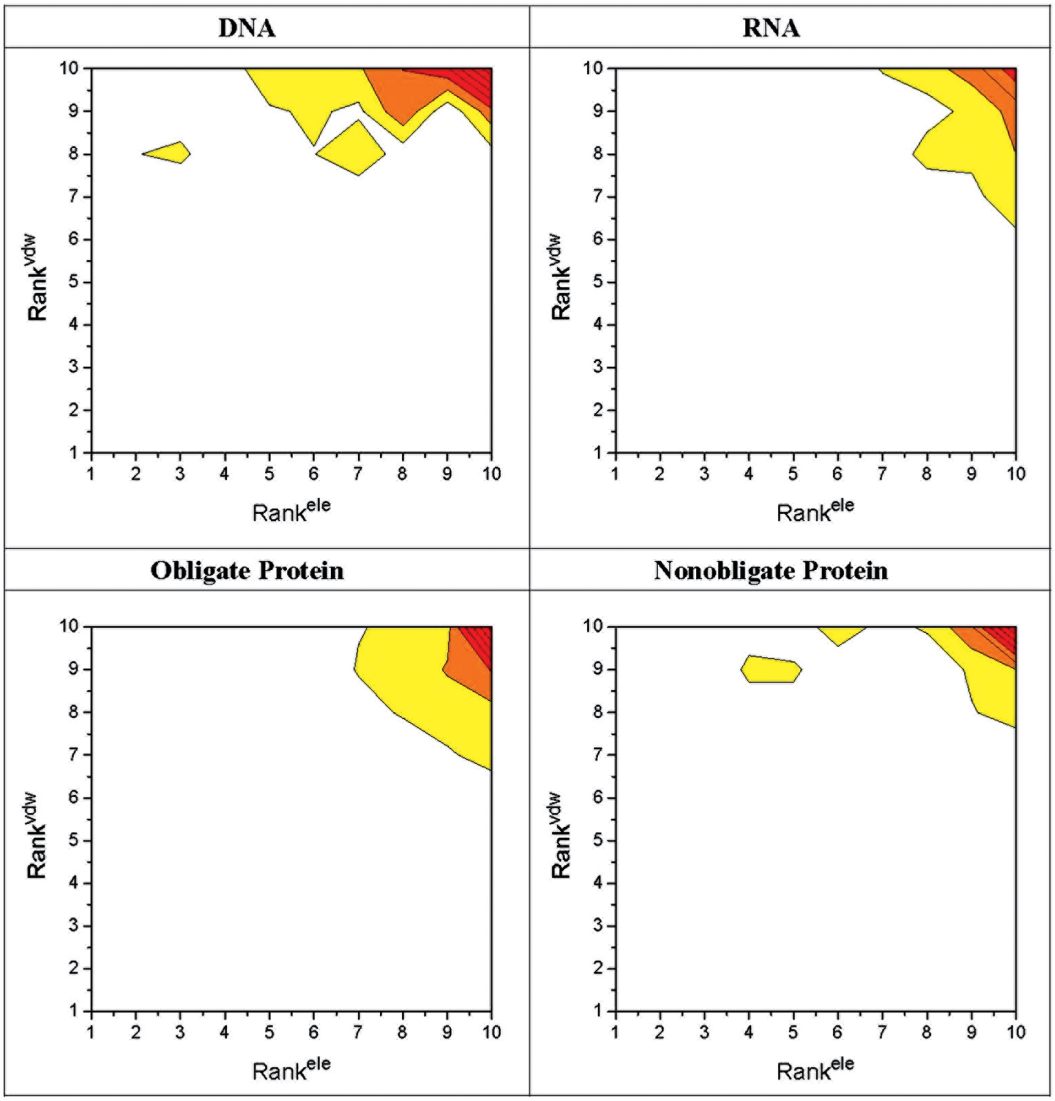

Figure 1. Electrostatic versus vdW energy ranking distributions in protein−macromolecule interactions. In each contour map, the x and y axes denote, respectively, the electrostatic and vdW energy ranks, which range from 1 to 10, of residues binding a given type of macromolecule. The ν^{+m}[Rank^{ele}, Rank^{vdW}] minus the random frequency, 0.01, are color-coded white for $0 < \nu^{+m} \leq 0.005$, yellow for $0.005 < \nu^{+m} \leq 0.010$, orange for $0.010 < \nu^{+m} \leq 0.015$ and red for $\nu^{+m} > 0.015$.

macromolecule m, <Rank^{x}>$^{+m}$ (x = ele, vdW, or ele + vdW) and the respective values of nonmacromolecule-binding residues, <Rank^{x}>$^{-}$, were derived from the bound and free protein structures (Table 1). Regardless of the conformational changes accompanying binding, all four types of macromolecule-binding sites generally possess electrostatic and/or vdW interactions that are more unfavorable than nonmacromolecule-binding regions; i.e. they possess electrostatic and/or vdW strain. For a given macromolecule m, the <Rank^{x}>$^{+m}$ of all macromolecule-binding residues in the bound/free protein structures is greater than the respective <Rank^{x}>$^{-}$ of all nonmacromolecule-binding residues (Table 1). For example, the electrostatic rank of DNA-binding residues in the DNA-bound structures (<Rank^{ele}>$^{+DNA}$ = 6.89) is greater than the respective value of 'non'macromolecule-binding residues (<Rank^{ele}>$^{-}$ = 5.34). Furthermore, the difference between <Rank^{x}>$^{+m}$ and <Rank^{x}>$^{-}$ in each data set ($\Delta^{x}{}_{m}$) is positive with magnitude > 0.2.

To further verify that macromolecule-binding residues are destabilized relative to their nonmacromolecule-binding counterparts regardless of the absence/presence of the macromolecule and its type, the Mann–Whitney U-test was used to test the null hypothesis that <Rank^{x}>$^{+m}{}_{p}$ is equal to or less than the respective <Rank^{x}>$^{-}{}_{p}$ for each protein in the free/bound dataset, and the results are summarized in Table 2 and Supplementary Table S4. The mean electrostatic/vdW ranks of macromolecule-binding residues, <Rank^{x}>$^{+m}{}_{p}$, derived from the 'bound' protein structures are significantly greater than the respective <Rank^{x}>$^{-}{}_{p}$ of nonmacromolecule-binding residues: the P-values in Table 2 are

Table 1. Energy ranking of macromolecule-binding versus nonmacromolecule-binding residues

Macromolecule, m[a]	N_m[b]	N^{+m}[c]	x	$<\mathrm{Rank}^x>^{+m}$	$<\mathrm{Rank}^x>^-$	Δ^x_m[d]
+DNA	76	11%	*ele*	6.89	5.34	1.55
			vdW	6.00	5.56	0.44
			Ele + vdW	6.88	5.33	1.55
−DNA	41	10%	*ele*	6.63	5.43	1.21
			vdW	6.40	5.45	0.95
			Ele + vdW	6.73	5.41	1.32
+RNA	72	20%	*ele*	6.18	5.55	0.63
			vdW	6.18	5.35	0.83
			Ele + vdW	6.24	5.51	0.73
−RNA	24	10%	*ele*	5.71	5.47	0.23
			vdW	6.60	5.43	1.17
			Ele + vdW	5.89	5.43	0.46
+Obligate protein	88	16%	*ele*	6.04	5.44	0.60
			vdW	6.49	5.34	1.14
			Ele + vdW	6.20	5.39	0.81
−Obligate protein	14	13%	*ele*	5.71	5.43	0.28
			vdW	6.44	5.27	1.17
			Ele + vdW	5.85	5.39	0.46
+Nonobligate protein	77	9%	*ele*	6.34	5.56	0.78
			vdW	6.13	5.50	0.63
			Ele + vdW	6.39	5.53	0.86
−Nonobligate protein	38	9%	*ele*	6.33	5.58	0.75
			vdW	6.29	5.52	0.77
			Ele + vdW	6.36	5.56	0.80

[a] The plus and minus sign indicate protein structures solved in the presence and absence of the macromolecule, respectively.
[b] The number of free or bound proteins in the dataset.
[c] The percentage of residues in the dataset that bind macromolecule *m*.
[d] $\Delta^x_m = <\mathrm{Rank}^x>^{+m} - <\mathrm{Rank}^x>^-$, where $<\mathrm{Rank}^x>^{+m}$ and $<\mathrm{Rank}^x>^-$ are computed according to Equation (4a) and (4b), respectively.

all <0.05, rejecting the null hypothesis with a 95% confidence level. The $<\mathrm{Rank}^x>^{+m}_p$ values derived from the 'free' protein structures are also significantly greater than the respective $<\mathrm{Rank}^x>^-_p$ ($P < 0.05$), except in two cases. Based on the 'free' structures of 24 RBPs and 14 obligate proteins, the mean electrostatic rank of RNA/obligate protein-binding residues (5.71) is 'not' significantly greater than the respective rank of nonmacromolecule residues (5.47/5.43), as the null hypothesis is rejected by $P > 0.05$, but the corresponding mean vdW $<\mathrm{Rank}^{vdW}>^{+m}$ ranks are significantly greater than the respective $<\mathrm{Rank}^{vdW}>^-$p ranks (see Table 1).

Difference between electrostatic or vdW strain in a given type of macromolecule-binding site

Although electrostatic and/or steric strain appears to be a common feature among the four different types of macromolecule-binding sites, it is not clear if electrostatic strain or vdW strain dictates the observed higher energy of a macromolecule-binding site compared to nonmacromolecule-binding regions. Thus, the differences between $<\mathrm{Rank}^x>^{+m}_p$ and $<\mathrm{Rank}^x>^-_p$ ($x = ele$ or vdW) were compared for each protein in the free/bound data set and the Mann–Whitney test was used to test the null hypothesis that $\Delta^{ele}_m = \Delta^{vdW}_m$. On the basis of the 'free' protein structures, steric strain is dominant in both RNA- and obligate protein-binding sites: the Δ^{vdW}_{RNA} (1.17) is greater than Δ^{ele}_{RNA} (0.23) in RBPs; likewise, the $\Delta^{vdW}_{obligate}$ (1.17) is greater than $\Delta^{ele}_{obligate}$ (0.28) in obligate proteins (Table 1). For these two free data sets, the null hypothesis, $\Delta^{\boldsymbol{ele}}_m \geq \Delta^{\boldsymbol{vdW}}_m$, is rejected with a 95% confidence level, as the computed *P*-values are <0.05 (Table 2). In contrast, neither electrostatic nor steric strain dictates the observed higher energy of DNA- and nonobligate protein-binding sites compared to nonmacromolecule-binding regions in the free structures: the null hypothesis, Δ^{ele}_{DNA} (1.21) ~ Δ^{vdW}_{DNA} (0.95) and $\Delta^{ele}_{nonobligate}$(0.75) ~ $\Delta^{vdW}_{nonobligate}$(0.77), is not rejected ($P > 0.05$ in Table 2). Interestingly, in the 'active' DNA-bound conformation, electrostatic strain becomes dominant in DNA-binding sites, as the null hypothesis, $\Delta^{\boldsymbol{ele}}_m \leq \Delta^{\boldsymbol{vdW}}_m$, is rejected by a *P*-value = 0.0001. On the other hand, in the 'active' RNA-bound conformation, steric strain is no longer dominant in RNA-binding sites, as the null hypothesis, Δ^{vdW}_{RNA} (0.83) ~ Δ^{ele}_{RNA} (0.63), is no longer rejected ($P > 0.05$ in Table 2).

Energetic difference between DNA- and RNA-binding sites

To determine the energetic difference between DNA- and RNA-binding sites, the $<\mathrm{Rank}^x>^{+DNA}$ ($x = ele$ or vdW) values derived from the free/bound data sets were compared with the respective $<\mathrm{Rank}^x>^{+RNA}$ values, and the Mann−Whitney test was used to verify the observed trends. The results in Tables 1 and 3 show that DNA and RNA-binding sites possess different electrostatic ranks but similar vdW ranks, regardless of the conformational changes that occur upon binding. The mean 'electrostatic' rank of DNA-binding residues derived from the bound/'free' structures (6.89/'6.63') is greater than that of RNA-binding residues (6.18/'5.71'). To evaluate if this difference is statistically significant, the Mann−Whitney U-test was used to test the null hypothesis that the $<\mathrm{Rank}^{ele}>^{+DNA}_p$ derived from the bound/free structures is equal to or less than the $<\mathrm{Rank}^{ele}>^{+RNA}_p$; the resulting *P*-values of <0.05 (see Table 3) rejected the null hypothesis with a 95% confidence level. On the other hand, the mean *vdW* rank of DNA-binding residues derived from the bound/'free' structures (6.00/'6.40') is similar to that of RNA-binding residues (6.18/'6.60'), as the null hypothesis, $<\mathrm{Rank}^{vdW}>^{+DNA}_p = <\mathrm{Rank}^{vdW}>^{+RNA}_p$, is accepted by *P*-values >0.05 (Table 3).

Energetic difference between obligate and nonobligate protein-binding sites

To determine the energetic difference between obligate- and nonobligate protein-binding sites, the $<\mathrm{Rank}^x>^{+obligate}$ ($x = ele$ or vdW) values derived from the free/bound data sets were compared with the respective $<\mathrm{Rank}^x>^{+nonobligate}$ values, and the Mann−Whitney test was again used to verify the observed trends. Although obligate and nonobligate protein-binding sites do not possess statistically different electrostatic/vdW ranks, they exhibit different electrostatic and vdW rank differences,

Table 2. *P*-values from Mann–Whitney U-tests to test if macromolecule-binding sites are energetically less stable than nonmacromolecule-binding sites and if electrostatic or vdW strain dictates the macromolecule-binding site[a]

Null Hypothesis	Dataset	m = DNA	m = RNA	m = obligate	m = nonobligate
$<\text{Rank}^{ele}>^{+\boldsymbol{m}} \leq <\text{Rank}^{ele}>^{-}$	Bound	0	0.0010	0	0.0010
	Free	0	**0.4023**	**0.2751**	0.0225
$<\text{Rank}^{vdW}>^{+\boldsymbol{m}} \leq <\text{Rank}^{vdW}>^{-}$	Bound	0.0005	0	0	0.0005
	Free	0	0.0006	0.0001	0.0007
$\Delta^{\boldsymbol{ele}}{}_{m} = \Delta^{\boldsymbol{vdW}}{}_{m}$	Bound	0.0002	**0.7432**	0.0259	**0.6580**
	Free	**0.3910**	0.0392	0.0482	**0.6551**
$\Delta^{\boldsymbol{ele}}{}_{m} \leq \Delta^{\boldsymbol{vdW}}{}_{m}$	Bound	0.0001	**0.6284**	**0.9871**	**0.3290**
	Free	**0.1955**	**0.9804**	**0.9759**	**0.6725**
$\Delta^{\boldsymbol{ele}}{}_{m} \geq \Delta^{\boldsymbol{vdW}}{}_{m}$	Bound	**0.9999**	**0.3716**	0.0129	**0.6710**
	Free	**0.8045**	0.0196	0.0241	**0.3275**

[a] *P*-values $>$ 0.05 are highlighted in bold.

Table 3. The *P*-values from Mann–Whitney U-tests to assess the energetic difference between similar macromolecule-binding sites[a]

Null hypothesis[b]	m = DNA, m' = RNA		m = obligate, m' = nonobligate	
	Bound[c]	Free[d]	Bound[c]	Free[d]
$<\text{Rank}^{ele}>^{+\boldsymbol{m}} = <\text{Rank}^{ele}>^{+\boldsymbol{m'}}$	0.0047	0.0074	**0.2760**	**0.2396**
$<\text{Rank}^{ele}>^{+\boldsymbol{m}} \leq <\text{Rank}^{ele}>^{+\boldsymbol{m'}}$	0.0024	0.0037	**0.8620**	**0.8802**
$<\text{Rank}^{ele}>^{+\boldsymbol{m}} \geq <\text{Rank}^{ele}>^{+\boldsymbol{m'}}$	**0.9976**	**0.9963**	**0.1380**	**0.1198**
$<\text{Rank}^{vdW}>^{+\boldsymbol{m}} = <\text{Rank}^{vdW}>^{+\boldsymbol{m'}}$	**0.4602**	**0.3589**	**0.1385**	**0.8689**
$<\text{Rank}^{vdW}>^{+\boldsymbol{m}} \leq <\text{Rank}^{vdW}>^{+\boldsymbol{m'}}$	**0.7699**	**0.8206**	**0.0693**	**0.4345**
$<\text{Rank}^{vdW}>^{+\boldsymbol{m}} \geq <\text{Rank}^{vdW}>^{+\boldsymbol{m'}}$	**0.2301**	**0.1794**	**0.9307**	**0.5655**
$<\delta^{ele-vdW}>^{+\boldsymbol{m}} = <\delta^{ele-vdW}>^{+\boldsymbol{m'}}$	0.0045	0.0267	0.0115	0.0465
$<\delta^{ele-vdW}>^{+\boldsymbol{m}} \leq <\delta^{ele-vdW}>^{+\boldsymbol{m'}}$	0.0023	0.0134	**0.9942**	**0.9767**
$<\delta^{ele-vdW}>^{+\boldsymbol{m}} \geq <\delta^{ele-vdW}>^{+\boldsymbol{m'}}$	**0.9977**	**0.9866**	0.0058	0.0233

[a] *P*-values $>$ 0.05 are highlighted in bold.
[b] $<\delta^{\text{ele-vdW}}>^{+\boldsymbol{m}} = \sum_{p} \langle \text{Rank}^{ele} \rangle_{p}^{+m} - \langle \text{Rank}^{vdW} \rangle_{p}^{+m} / N_m$, where the summation is over the N_m proteins in the dataset.
[c] *P*-values derived from protein structure solved with the macromolecule.
[d] *P*-values derived from protein structure solved without the macromolecule.

regardless of the conformational changes that occur upon binding. The mean electrostatic and vdW ranks of ‘obligate’ protein-binding residues derived from the bound/‘free’ data sets ($<\text{Rank}^{ele}>^{+obligate}$ = 6.04/‘5.71’ and $<\text{Rank}^{vdW}>^{+obligate}$ = 6.49/‘6.44’) do not differ significantly from those of ‘nonobligate’ protein-binding residues ($<\text{Rank}^{ele}>^{+nonobligate}$ = 6.34/‘6.33’ and $<\text{Rank}^{vdW}>^{+nonobligate}$ = 6.13/‘6.29’), as the null hypothesis, $<\text{Rank}^{x}>^{+obligate}{}_{p} = <\text{Rank}^{x}>^{+nonobligate}{}_{p}$ ($x = ele$ or vdW), is accepted by $P > 0.05$ (see Table 3). On the other hand, the mean difference between the electrostatic and vdW ranks of ‘obligate’ protein-binding residues, $<\delta^{ele-vdW}>^{+obligate}$, derived from the bound (−0.45) or free (−0.73) structures differs in sign from that of ‘nonobligate’ protein-binding residues, $<\delta^{ele-vdW}>^{+nonobligate}$, derived from the bound (0.21) or free (0.04) structures. The null hypothesis that $<\delta^{ele-vdW}>^{+obligate}$ is equal to or greater than $<\delta^{ele-vdW}>^{+nonobligate}$ is rejected by *P*-values $<$0.05 (Table 3). Thus, the mean electrostatic rank is less than the respective vdW rank for ‘obligate’ protein-binding residues, but the former is greater than the latter for ‘nonobligate’ protein-binding residues in the free/bound structures.

Most proteins possess electrostatic or steric strain in their macromolecule-binding sites

To determine if every protein has a binding site for a given type of macromolecule that is less electrostatically and/or sterically stable than nonmacromolecule-binding regions, the percentage of proteins in the respective data set with $<\text{Rank}^{x}>^{+m}{}_{p}$ greater than $<\text{Rank}^{x}>^{-}{}_{p}$ ($x = ele$ or vdW) was computed from the free/bound data sets. Regardless of the conformational changes that occur upon binding, over 81% of DBPs, RBPs, obligate and nonobligate proteins have binding sites with more electrostatic or steric strain than nonmacromolecule-binding regions (see Table 4). For example, out of the 41 free DBPs, 31 or 75.6% exhibit $<\text{Rank}^{ele}>^{+DNA}{}_{p}$ of DNA-binding residues greater than that of nonmacromolecule-binding residues. Among the 10 DBPs with $<\text{Rank}^{ele}>^{+DNA}{}_{p}$ equal to or less than the respective $<\text{Rank}^{ele}>^{-}{}_{p}$, eight DBPs exhibit $<\text{Rank}^{vdw}>^{+DNA}{}_{p}$ greater than $<\text{Rank}^{vdw}>^{-}{}_{p}$. Hence, 95% (39/41) of the DBPs possess DNA-binding sites with more electrostatic and/or steric strain than the respective nonmacromolecule-binding regions. The results derived from the free/bound structures also show that

Table 4. Number of proteins whose binding sites for a given macromolecule is electrostatically or sterically strained

Dataset[a]	N_m[b]	N_m^{ele}[c]	N_m^{vdW}[d]	N_m^{*}[e]
+DNA	76	62 (81.6%)	50 (65.8%)	70 (92.1%)
−DNA	41	31 (75.6%)	33 (80.5%)	39 (95.1%)
+RNA	72	44 (61.1%)	47 (65.3%)	59 (81.9%)
−RNA	24	12 (50.0%)	19 (79.2%)	22 (91.7%)
+Obligate protein	88	64 (72.7%)	76 (86.4%)	85 (96.6%)
−Obligate protein	14	6 (42.9%)	13 (92.9%)	13 (92.9%)
+Nonobligate protein	77	48 (62.3%)	52 (67.5%)	64 (83.1%)
−Nonobligate protein	38	24 (63.2%)	27 (71.1%)	31 (81.6%)

[a] The plus and minus sign indicate protein structures solved in the presence and absence of the macromolecule, respectively.
[b] The number of protein structures in the dataset.
[c] The number and (percentage) of proteins with $<\text{Rank}^{ele}>^{+m}_{p} > <\text{Rank}^{ele}>^{-}_{p}$.
[d] The number and (percentage) of proteins with $<\text{Rank}^{vdW}>^{+m}_{p} > <\text{Rank}^{vdW}>^{-}_{p}$.
[e] The number and (percentage) of proteins with $<\text{Rank}^{ele}>^{+m}_{p} > <\text{Rank}^{ele}>^{-}_{p}$ and/or $<\text{Rank}^{vdW}>^{+m}_{p} > <\text{Rank}^{vdW}>^{-}_{p}$.

between two similar macromolecules, more DBPs and obligate proteins have destabilized binding sites, as compared to RBPs and nonobligate proteins, respectively.

DISCUSSION

This work presents a general strategy for detecting electrostatic and vdW strain in a cluster of aa residues given the protein structure without any empirical or adjustable parameters. This strategy was applied to eight structurally nonhomologous data sets containing ≤3-Å X-ray structures of DBPs, RBPs, obligate and nonbligate proteins, free and bound to their cognate macromolecule. The results reveal a common physical basis for DNA-, RNA-, obligate protein- and nonobligate protein-binding sites; i.e. they have in common electrostatic and/or steric strain (Figure 1 and Table 1). This feature appears independent of (i) the type of macromolecule-binding partner, (ii) the strength of the protein−macromolecule interaction (59) and (iii) conformational changes upon macromolecule binding (see Supplementary Table S1). The finding that interacting residues possess strain in conjunction with residue conservation and other structural features can help in functional annotation, as shown for the prediction of catalytic residues (60) and DNA/RNA-binding residues (9,25). We are currently applying the findings herein to predict the hot spots for macromolecule-binding sites. It could also help in narrowing the conformational search in protein−macromolecule docking, and in designing therapeutic agents in cases where mutations of interacting residues result in pathogenesis.

Comparison with previous work

The finding that macromolecule-binding residues generally possess more electrostatic and/or steric strain compared to nonmacromolecule-binding residues is in accord with previous studies. Experimental studies have found substrate-binding residues in several enzymes such as barnase (61), *Escherichia coli* ribonuclease H1 (62), and T4 lysozyme (63) to be suboptimally stable, as their mutations increased stability but reduced activity. Likewise, mutations of interface residues in proteins such as retinoic acid-binding protein (64) and barstar (65) yielded more stable proteins. For these five proteins, Elcock (66) found that the functional residues known to be suboptimally stable were among the most electrostatically unstable residues identified by the change in the electrostatic free energy of a side chain upon transfer from aqueous solution to the protein. Elcock (66) further analyzed 216 protein–protein 'complex' structures and found that the top 10% most electrostatically destabilizing charged residues are more likely to be conserved (and thus important for function) than to be variable. Ota *et al.* (60) found that catalytic residues in 49 representative enzymes destabilize the protein structure more than noncatalytic residues. Liang and co-workers (67) showed that residues at the interfaces of eight 'nonobligate' heterodimeric protein 'complexes' (PDB entries 1ppf, 1cho, 1fss, 1brs, 2sic, 2ptc, 2sni and 1mlc) have higher sidechain energies than the other surface residues.

In contrast to the above findings, Dessailly *et al.* (68) found very poor overlap between destabilizing regions and protein/nucleic acid-binding sites: none of the 11 protein–protein or five protein−DNA complexes studied possess >40% binding residues (defined as residues with two or more nonhydrogen atoms within 6 Å of a ligand nonhydrogen atom) that are found in destabilizing regions; i.e. the reported sensitivity is ≤40%. This maybe due to the different energy functions used to define destabilizing residues: Dessailly *et al.* (68) used the electrostatic energy computed with a dielectric constant of eight and the solvation free energy, whereas we employed relative 'gas-phase' electrostatic and vdW energies [see Equation (1)] to define destabilizing residues. Thus, in DNA-binding sites for example, the electrostatic interactions of positively charged atoms among themselves would be 'unfavorable', but those with water molecules would be favorable. Consequently, the favorable solvation free energy would cancel in part the unfavorable gas-phase electrostatic energy, as shown in our previous work (9), which may partly account for the observed discrepancy.

Energetic features distinguishing binding sites of similar macromolecules

The results herein have also revealed novel features distinguishing DNA- from RNA-binding sites, and obligate protein- from nonobligate protein-binding sites that are independent of the conformational changes upon binding. Although both DNA- and RNA-binding sites have in common positively charged aa side chains interacting with negatively charged DNA/RNA phosphate groups, the mean electrostatic rank of DNA-binding residues is significantly greater than that of RNA-binding sites. This difference may reflect the fact that DBPs bind mostly double-stranded DNA, but RBPs bind not only double-stranded RNA but also single-stranded RNA. However, there were insufficient protein structures containing purely double-stranded or single-stranded RNA for statistical analyses (see Supplementary Table S1), as most RBPs bind to RNA molecules containing double-stranded and/or single-stranded regions and loops/bulges. Although both obligate and nonobligate protein-binding sites interact with another protein, the mean vdW rank is 'greater' than the mean electrostatic rank in obligate protein-binding sites, whereas it is 'less' than the mean electrostatic rank in nonobligate protein-binding sites. The new energetic features distinguishing obligate protein and nonobligate protein interfaces complement the structural characteristics found in previous works, which showed that obligate interfaces are more nonpolar and larger with more contacts than nonobligate interfaces (28,29).

Analysis of macromolecule-binding sites with no apparent electrostatic and/or steric strain

Although most macromolecule-binding sites possess electrostatic and/or steric strain, certain sites seem to possess negligible strain. One possible reason why some proteins possess apparently stable binding sites for a given type of macromolecule (with $\langle \mathrm{Rank}^{x}\rangle^{+m}_{p}$ less than or equal to $\langle \mathrm{Rank}^{x}\rangle^{-}_{p}$) is because their nonmacromolecule-binding regions may comprise of residues that are energetically unstable in the absence of metal or other cofactors that play a role in stabilizing the protein structure. Furthermore, the assignment of nonmacromolecule-binding residues depends on the availability of highly homologous structures (see 'Materials and methods' section). Another possible reason is the finding that most DBPs and RBPs as well as obligate and nonobligate proteins with relatively stable binding sites exist as multimers, whose electrostatic and vdW interactions were not considered in the present analyses. For example, the $\langle \mathrm{Rank}^{x}\rangle^{+RNA}_{p}$ in the ribosome proteins, 1vq8-1, 1vq8-2, 1vq8-D, 1vq8-E, 1vq8-J, 1vq8-X, are less than or equal to $\langle \mathrm{Rank}^{x}\rangle^{-}_{p}$, probably because each of these proteins is part of the large ribosomal subunit of Haloarcula Marismortui, which comprises 30 protein chains.

SUPPLEMENTARY DATA

Supplementary data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Dr H. Yuan for helpful discussion, and A. Lin and C. Hua for technical help.

FUNDING

The National Science Council, Taiwan (NSC 95-2113-M-001-038-MY5). Funding for open access charge: NSC 95-2113-M-001-038-MY5.

Conflict of interest statement. None declared.

REFERENCES

1. Brenner,S.E. and Levitt,M. (2000) Expectations from structural genomics. *Protein Sci.*, **9**, 197–200.
2. Lesley,S.A., Kuhn,P., Godzik,A., Deacon,A.M., Mathews,I., Kreusch,A., Spraggon,G., Klock,H.E., McMullan,D., Shin,T. *et al.* (2002) Structural genomics of the Thermotoga maritima proteome implemented in a high-throughput structure determination pipeline. *Proc. Natl Acad. Sci. USA*, **99**, 11664–11669.
3. Janin,J., Henrick,K., Moult,J., Eyck,L.T., Sternberg,M.J.E., Vajda,S., Vakser,I. and Wodak,S.J. (2003) CAPRI: A Critical Assessment of PRedicted Interactions. *Proteins: Struct. Funct. Genet.*, **52**, 2–9.
4. Ohlendorf,D.H. and Matthew,J.B. (1985) Electrostatics and flexibility in protein-DNA interactions. *Adv. Biophys.*, **20**, 137–151.
5. Jones,S., van,H.P., Berman,H.M. and Thornton,J.M. (1999) Protein-DNA interactions: a structural analysis. *J. Mol. Biol.*, **287**, 877–896.
6. Luscombe,N.M., Laskowski,R.A. and Thornton,J.M. (2001) Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res.*, **29**, 2860–2874.
7. Jones,S., Shanahan,H.P., Berman,H.M. and Thornton,J.M. (2003) Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. *Nucleic Acids Res.*, **31**, 7189–7198.
8. Tsuchiya,Y., Kinoshita,K. and Nakamura,H. (2004) Structure-based prediction of DNA-binding sites on proteins using the empirical preference of electrostatic potential and the shape of molecular surfaces. *Proteins: Struct. Funct. Genet.*, **55**, 885–894.
9. Chen,Y.C., Wu,C.Y. and Lim,C. (2007) Predicting DNA-binding sites on proteins from electrostatic stabilization upon mutation to Asp/Glu and evolutionary conservation. *Proteins: Struct. Funct. Bioinf.*, **67**, 671–680.
10. Ahmad,S., Gromiha,M.M. and Sarai,A. (2004) Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics*, **20**, 477–486.
11. Kuznetsov,I.B., Gou,Z., Li,R. and Hwang,S. (2006) Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins. *Proteins: Struct. Funct. Bioinf.*, **64**, 19–27.
12. Wang,L. and Brown,S.J. (2006) BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucleic Acids Res.*, **34**, W243–W248.
13. Cusack,S. (1999) RNA-protein complexes. *Curr. Opin. Struct. Biol.*, **9**, 66–73.
14. Draper,D.E. (1999) Themes in RNA-protein recognition. *J. Mol. Biol.*, **293**, 255–270.
15. Jones,S., Daley,D.T., Luscombe,N.M., Berman,H.M. and Thornton,J.M. (2001) Protein-RNA interactions: a structural analysis. *Nucleic Acids Res.*, **29**, 943–954.
16. Treger,M. and Westhof,E. (2001) Statistical analysis of atomic contacts at RNA-protein interfaces. *J. Mol. Recogn.*, **14**, 199–214.
17. Kim,H., Jeong,E., Lee,S.W. and Han,K. (2003) Computational analysis of hydrogen bonds in protein-RNA complexes for interaction patterns. *FEBS Lett.*, **552**, 231–239.
18. Chirgadze Iu,N. and Larionova,E.A. (2005) Principal role of large polar residue clusters of RNA-binding proteins in the formation of complexes with RNA. *Mol. Biol. (Mosk)*, **39**, 1017–1031.
19. Varani,G. (2005) How proteins and RNA recognize each other. *FEBS J.*, **272**, 2087–2087.
20. Morozova,N., Allers,J., Myers,J. and Shamoo,Y. (2006) Protein-RNA interactions: exploring binding patterns with a three-dimensional superposition analysis of high resolution structures. *Bioinformatics*, **22**, 2746–2752.
21. Ellis,J.J., Broom,M. and Jones,S. (2007) Protein-RNA interactions: structural analysis and functional classes. *Proteins: Struct. Funct. Bioinf.*, **66**, 903–911.
22. Jeong,E., Chung,I. and Miyano,S. (2004) A neural network method for identification of RNA-interacting residues in protein. *Genome Inform. Ser. Workshop Genome Inform.*, **15**, 105–116.
23. Terribilini,M., Lee,J.H., Yan,C., Jernigan,R.L., Honavar,V. and Dobbs,D. (2006) Prediction of RNA binding sites in proteins from amino acid sequence. *Rna*, **12**, 1450–1462.

24. Kim,O.T.P., Yura,K. and Go,N. (2006) Amino acid residue doublet propensity in the protein-RNA interface and its application to RNA interface prediction. *Nucleic Acids Res.*, **34**, 6450–6460.
25. Chen,Y.C. and Lim,C. (2008) Predicting RNA-binding sites from the protein structure based on electrostatics, evolution and geometry. *Nucleic Acids Res.*, **36**, e29.
26. Nooren,I.M.A. and Thornton,J.M. (2003) Diversity of protein-protein interactions. *EMBO J.*, **22**, 3486–3492.
27. Glaser,F., Steinberg,D.M., Vakser,I.A. and Ben-Tal,N. (2001) Residue frequencies and pairing preferences at protein-protein interfaces. *Proteins: Struct. Funct. Genet.*, **43**, 89–102.
28. Mintseris,J. and Weng,Z. (2005) Structure, function, and evolution of transient and obligate protein-protein interactions. *Proc. Natl Acad. Sci. USA*, **102**, 10930–10935.
29. De,S., Krishnadev,O., Srinivasan,N. and Rekha,N. (2005) Interaction preferences across protein-protein interfaces of obligatory and non-obligatory components are different. *BMC Struct. Biol.*, **5**, 15.
30. Jones,S. and Thornton,J.M. (1996) Principles of protein-protein interactions. *Proc. Natl Acad. Sci. USA*, **93**, 13.
31. Lo Conte,L., Chothia,C. and Janin,J. (1999) The atomic structure of protein-protein recognition sites. *J. Mol. Biol.*, **285**, 2177–2198.
32. Noskov,S. and Lim,C. (2001) Free energy decomposition of protein-protein interactions. *Biophys. J.*, **81**, 737–750.
33. Zhou,H.-X. and Shan,Y. (2001) Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins: Struct. Funct. Genet.*, **44**, 336–343.
34. Jones,S. and Thornton,J.M. (1997) Prediction of protein-protein interaction sites using patch analysis. *J. Mol. Biol.*, **272**, 133–143.
35. Cole,C. and Warwicker,J. (2002) Side-chain conformational entropy at protein-protein interfaces. *Protein Sci.*, **11**, 2860–2870.
36. Gilson,M.K. and Honig,B. (1988) Calculation of the total electrostatic energy of macromolecular system: solvation energy, binding energies and conformational analysis. *Proteins: Struct. Func. Genet.*, **4**, 7–18.
37. Gabb,H.A., Jackson,R.M. and Sternberg,M.J.E. (1997) Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J. Mol. Biol.*, **272**, 106–120.
38. Kufareva,I., Budagyan,L., Raush,E., Totrov,M. and Abagyan,R. (2007) PIER: Protein Interface Recognition for Structural Proteomics. *Proteins: Struct. Funct. Bioinf.*, **67**, 400–417.
39. Neuvirth,H., Raz,R. and Schreiber,G. (2004) ProMate: a structure based prediction program to identify the location of protein-protein binding sites. *J. Mol. Biol.*, **228**, 181–199.
40. Burgoyne,R. and Jackson,R.M. (2006) Predicting protein interaction sites: binding hot-spots in protein-protein and protein-ligand interfaces. *Bioinformatics*, **22**, 1335–1342.
41. Hoskins,J., Lovell,S. and Blundell,T.L. (2006) An algorithm for predicting protein-protein interaction sites: abnormally exposed amino acid residues and secondary structure elements. *Protein Sci.*, **15**, 1017–1029.
42. Liang,S., Zhang,C., Liu,S. and Zhou,Y. (2006) Protein binding site prediction using an empirical scoring function. *Nucleic Acids Res.*, **34**, 3698.
43. Murakami,Y. and Jones,S. (2006) SHARP2: protein-protein interaction predictions using patch analysis. *Bioinformatics*, **22**, 1794–1795.
44. Koike,A. and Takagi,T. (2004) Prediction of protein-protein interaction sites using support vector machines. *Prot. Eng. Des. Sel.*, **17**, 165–173.
45. Bordner,A.J. and Abagyan,R. (2005) Statistical analysis and prediction of protein-protein interfaces. *Proteins: Struct. Func. Bioinf.*, **60**, 353–366.
46. Bradford,J.R. and Westhead,D.R. (2005) Improved prediction of protein-protein binding sites using a support vector machines approach. *Bioinformatics*, **21**, 1487–1494.
47. Fariselli,P., Pazos,F., Valencia,A. and Casadio,R. (2002) Prediction of protein-protein interaction sites in heterocomplexes with neural networks. *Eur. J. Biochem.*, **269**, 1356–1361.
48. Chen,H. and Zhou,H.-X. (2005) Prediction of interface residues in protein-protein complexes by a consensus neural network method: test against NMR data. *Proteins: Struct. Funct. Bioinf.*, **61**, 21–35.
49. Bradford,J.R., Needham,C.J., Bulpitt,A.J. and Westhead,D.R. (2006) Insights into protein-potein interfaces using a Bayesian network prediction method. *J. Mol. Biol.*, **362**, 365–386.
50. Berman,H.M., Battistuz,T., Bhat,T.N., Bluhm,W.F., Bourne,P.E., Burkhardt,K., Iype,L., Jain,S., Fagan,P., Marvin,J. *et al.* (2002) The Protein Data Bank. *Acta Crystallogr. D*, **58**, 899–907.
51. Ponomarenko,J.V. and Bourne,P.E. (2007) Antibody-protein interactions: benchmark datasets and prediction tools evaluation. *BMC Struct. Biol.*, **7**, 64.
52. Pearl,F., Todd,A., Sillitoe,I., Dibley,M., Redfern,O., Lewis,T., Bennett,C., Marsden,R., Grant,A., Lee,D. *et al.* (2005) The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic Acids Res.*, **33**, D247–D251.
53. Milburn,D. (1998) Sequences annotated by structure: a tool to facilitate the use of structural information in sequence analysis. *Prot. Eng. Des. Sel.*, **11**, 855–859.
54. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignments through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
55. Taylor,W.R. and Orengo,C.A. (1989) Protein structure alignment. *J. Mol. Biol.*, **208**, 1–22.
56. McDonald,I.K. and Thornton,J.M. (1994) Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.*, **238**, 777–793.
57. Case,D.A., Darden,T., Cheatham,T.E. III, Simmerling,C., Wang,J., Duke,R.E., Luo,R., Merz,K.M., Pearlman,D.A. and Crowley,M. (2006) AMBER 9. *University of California, San Francisco*.
58. Duan,Y., Wu,C., Chowdhury,S., Lee,M.C., Xiong,G., Zhang,W., Yang,R., Cieplak,P., Luo,R., Lee,T. *et al.* (2003) A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comp. Chem.*, **24**, 1999–2012.
59. Brooijmans,N., Sharp,K.A. and Kuntz,I.D. (2002) Stability of macromolecular complexes. *Proteins: Struct. Funct. Genet.*, **48**, 645–653.
60. Ota,M., Kinoshita,K. and Nishikawa,K. (2003) Prediction of catalytic residues in enzymes based on known tertiary structure, stability profile, and sequence conservation. *J. Mol. Biol.*, **327**, 1053–1064.
61. Meiering,E.M., Serrano,L. and Fersht,A.R. (1992) Effect of active site residues in barnase on activity and stability. *J. Mol. Biol.*, **225**, 585–589.
62. Kanaya,S., Oobatake,M. and Liu,Y. (1996) Thermal stability of E-Coli Ribonuclease H1 and its active site mutants in the presence and absence of the Mg^{2+} ion: Proposal of a new catalytic role for Glu 48. *J. Biol. Chem.*, **271**, 32729–32736.
63. Shoichet,B.K., Baase,W.A., Kuroki,R. and Matthews,B.W. (1995) A relationship between protein stability and protein function. *Proc. Natl Acad. Sci. USA*, **92**, 452–456.
64. Zhang,J., Liu,Z.P., Jones,T.A., Gierasch,L.M. and Sambrook,J.F. (1992) Mutating the charged residues in the binding pocket of cellular retinoic acid-binding protein simultaneously reduces its binding affinity to retinoic acid and increases its thermostability. *Proteins*, **13**, 87–99.
65. Schreiber,G., Buckle,A.M. and Fersht,A.R. (1994) Stability and function: two constraints in the evolution of barstar and other proteins. *Structure*, **2**, 945–951.
66. Elcock,A.H. (2001) Prediction of functionally important residues based solely on the computed energetics of protein structure. *J. Mol. Biol.*, **312**, 885–896.
67. Liang,S., Zhang,J., Zhang,S. and Guo,H. (2004) Prediction of the interaction site on the surface of an isolated protein structure by analysis of side chain energy scores. *Proteins: Struct. Funct. Bioinf.*, **57**, 548–557.
68. Dessailly,B.H., Lensink,M.F. and Wodak,S.J. (2007) Relating destabilizing regions to known functional sites in proteins. *BMC Bioinformatics*, **8**, 141.