doi:10.1016/j.jmb.2004.07.047 J. Mol. Biol. (2004) 342, 1053–1066

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Electrostatic Potential of Nucleotide-free Protein is Sufficient for Discrimination Between Adenine and Guanine-specific Binding Sites

Gautam Basu[1,2,3]*, Dakshanamurthy Sivanesan[1], Takeshi Kawabata[1] and Nobuhiro Go[1,2]

[1]*Bioinformatics Unit, Graduate School of Information Science Nara Institute of Science and Technology, 8916-5 Takayama Ikoma, Nara 630-0192, Japan*

[2]*CCSE, Japan Atomic Energy Research Institute, 8-1 Umemidai, Kizu-cho Souraku-gun, Kyoto-fu 619-0215, Japan*

[3]*Department of Biophysics Bose Institute, P-1/12 CIT Scheme VII M, Kolkata 700054 India*

**Corresponding author*

Despite sharing many common features, adenine-binding and guanine-binding sites in proteins often show a clear preference for the cognate over the non-cognate ligand. We have analyzed electrostatic potential (ESP) patterns at adenine and guanine-binding sites of a large number of non-redundant proteins where each binding site was first annotated as adenine/guanine-specific or non-specific from a survey of primary literature. We show that more than 90% of ESP variance at the binding sites is accounted for by only two principal component ESP vectors, each aligned to molecular dipoles of adenine and guanine. Projected on these principal component vectors, the adenine/guanine-specific and non-specific binding sites, including adenine-containing dinucleotides, show non-overlapping distributions. Adenine or guanine specificities of the binding sites also show high correlation with the corresponding electrostatic replacement (cognate by non-cognate ligand) energies. High correlation coefficients (0.94 for 35 adenine-binding sites and 1.0 for 20 guanine-binding sites) were obtained when adenine/guanine specificities were predicted using the replacement energies. Our results demonstrate that ligand-free protein ESP is an excellent indicator for discrimination between adenine and guanine-specific binding sites and that ESP of ligand-free protein can be used as a tool to annotate known and putative purine-binding sites in proteins as adenine or guanine-specific.

Keywords: electrostatic potential; adenine; guanine; ligand-discrimination; ligand binding

Introduction

Protein–protein and protein–ligand interactions play a central role in controlling biological functions. In order to precisely understand these interactions at the molecular level, it is important that the three-dimensional structures of proteins and ligands, including their interacting geometries, be known. Post-structural analysis, such as the statistical analysis of non-redundant structures in the RCSB Protein Data Bank (pdb)† is a simple yet robust way to decipher weak rules governing molecular interactions.

The implicit assumption underlying such statistical studies of molecular interactions is that interaction free energies, the driving force behind any molecular association, can be mapped onto simple and conspicuous structural (or sequence) features like hydrogen bonds or the local propensity of amino acid residues around the interaction site. This generally holds true when the interaction energy arises mainly from a handful of dominant interactions. However, if the interaction energy is composed of a large number of weak interactions, no single structural feature may be conspicuous in the interacting complex. In such cases, simple structural analyses will fail to pinpoint factors responsible for the molecular interactions.

The situation can be even more difficult if, instead of the factors responsible for ligand binding to a

Abbreviations used: ESP, electrostatic potential; A, adenine; G, guanine; PCA, principal component analysis; PC, principal component; pdb, Protein Data Bank.

E-mail addresses of the corresponding author: gautam@is.naist.jp; gautam@boseinst.ernet.in

† http://www.rcsb.org/pdb

protein, the factors responsible for discrimination between two very similar ligands by a protein are considered. A prime example of this problem is to understand the factors responsible for discrimination between adenine (A) and guanine (G) by nucleotide-binding proteins.[1] A variety of ligands that bind to proteins and play a key role in protein function contain A or G. Although A and G are very similar in shape and size (Figure 1), subtle differences between the two are exploited by proteins to often exhibit specificity towards one over the other. It is important to identify any underlying common energetic mechanisms for such discrimination. In addition, the identification of specific sequences or structural motifs that recognize and discriminate A/G are also essential for the prediction of binding sites and functions of unknown proteins that possibly utilize A/G nucleotides.[2,3] However, established nucleotide-binding motifs[4–8] typically focus on the non-purine part of the nucleotide, for example the phosphate group, and not on A/G. Known A-binding motifs,[9,10] on the other hand, are not universal and apply to only a subset of all A-binding sites.

Nobeli *et al.*[1] addressed the question of A/G discrimination by proteins by analyzing a large number of known structures of A and G-bound proteins. The empirical study, with particular emphasis placed on hydrogen bond networks, found different clustering of hydrogen bonds around the two rings. Yet, the clustering alone was not enough to explain molecular discrimination between A and G by proteins. Instead, the concept of fuzzy clustering, reflecting the variety of ways a protein may evolve to recognize the same molecular moiety, was invoked to explain the molecular discrimination.

Here, we re-examine the problem of discrimination between A and G, from a different point of view. Instead of focusing on some conspicuous structural feature that defines the binding site environment, and analyzing how it differs between the A and G-binding sites, the current focus is on deciphering any consistent bias in the overall effect of the entire protein to differential binding. Specifically we focus on the electrostatic potential (ESP) at the purine-binding sites and aim to correlate ESP patterns and A/G specificity of the binding sites. In doing so, we emphasize that the A/G specificity of a protein needs to be determined independently, irrespective of its occurrence as A-bound or G-bound in the pdb, if one desires to attribute any biological meaning to specific structural or other features of the binding site.

Electrostatic interaction in proteins is known to play a crucial role in defining the nature of binding sites,[11,12] in the kinetic[13] and thermodynamic[14] control of protein–protein interactions, and in the hot and cold adaptation of proteins.[15] Suitable alteration of electrostatic interactions can lead to novel engineered binding sites,[16] in addition to their usefulness in the prediction of binding sites in proteins,[17] especially when the bound ligand possesses a net charge, like DNA.[18] Although A and G are electrically neutral, they are characterized by substantial dipole moments. The large difference in dipole moments of A (μ_A) and G (μ_G) (see Figure 1) prompted us to investigate whether the electrostatic character of purine-binding sites, as defined by the ligand-free protein ESP, shows a clear correlation with the biological requirement of the binding site, i.e. the discrimination between A and G, across protein families. We show that there is indeed a very strong correlation between the two, establishing the existence of a strong electrostatic component in A/G discrimination by proteins.

Results and Discussion

Classification and A/G specificity annotation of purine-binding sites

Our aim is to determine if the electrostatic nature of purine-binding sites in proteins is sufficient for discrimination between two ligands, one containing A and the other containing G. The question implicitly assumes that the non-purine parts of the

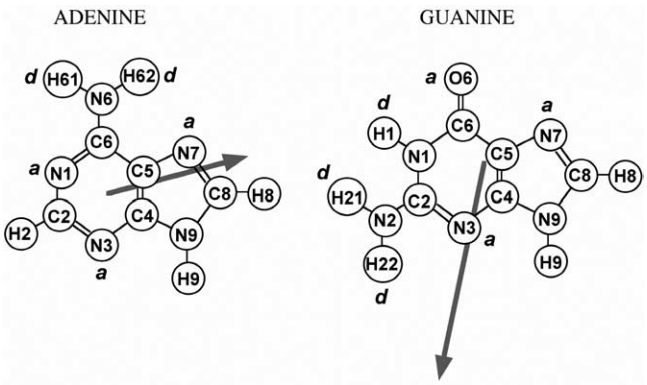

Figure 1. Structures of isolated adenine and guanine, annotated with atom names. Potential hydrogen bond acceptors and hydrogen bond donors are annotated with the letters *a* and *d*, respectively. The two arrows indicate the direction and magnitude of molecular dipoles.

ligands are identical. However, the dataset of protein–ligand complexes used in this work contains two kinds of ligands (i) ligands for which there exists (in nature) another ligand containing the complementary purine base (e.g. ATP has a counterpart: GTP); (ii) other A/G-containing ligands (e.g. dinucleotide FAD has no G counterpart). Accordingly, depending on the type of bound ligand, the A-bound and G-bound proteins in the dataset were divided into two classes (i) the ATP-set and the G-set; and (ii) the FAD-set (see Materials and Methods). A/G specificity and binding site ESP can be compared only in the first class, since for all bound ligands in the ATP-set and most bound ligands (for two entries, 1aa6 and 1dmr, the bound ligand is a G dinucleotide) in the G-set, there exists another ligand containing the complementary purine base. On the other hand, the electrostatic nature of purine-binding sites in the FAD-set may or may not be responsible for any observed A/G specificity, and therefore, a comparison of the electrostatic nature of these binding sites and A/G specificity will not be attempted. Although this sounds natural, it should be pointed out that our work differs from that of Nobeli *et al.*[1] in that in the latter no such division was attempted.

In addition, for the ATP or the G-set, the occurrence of A or G-bound proteins in the pdb by itself is not sufficient for drawing any conclusion about the A/G specificity of the binding site. Therefore, we examined the ligand-binding affinities of all ligand-bound proteins in the dataset as reported in the primary literature (except the FAD-set), focusing on the biological need, or any experimental evidence, for A/G specificity. A/G specificity annotations of purine-binding sites in the ATP and G-sets are shown in Tables 1 and 2. Use of literature-derived specificity annotation highlights another aspect of our approach not used by Nobeli *et al.*[1] Having classified and annotated our dataset, next we sought a correlation between the literature-derived specificity annotation and electrostatic nature of A/G-binding sites.

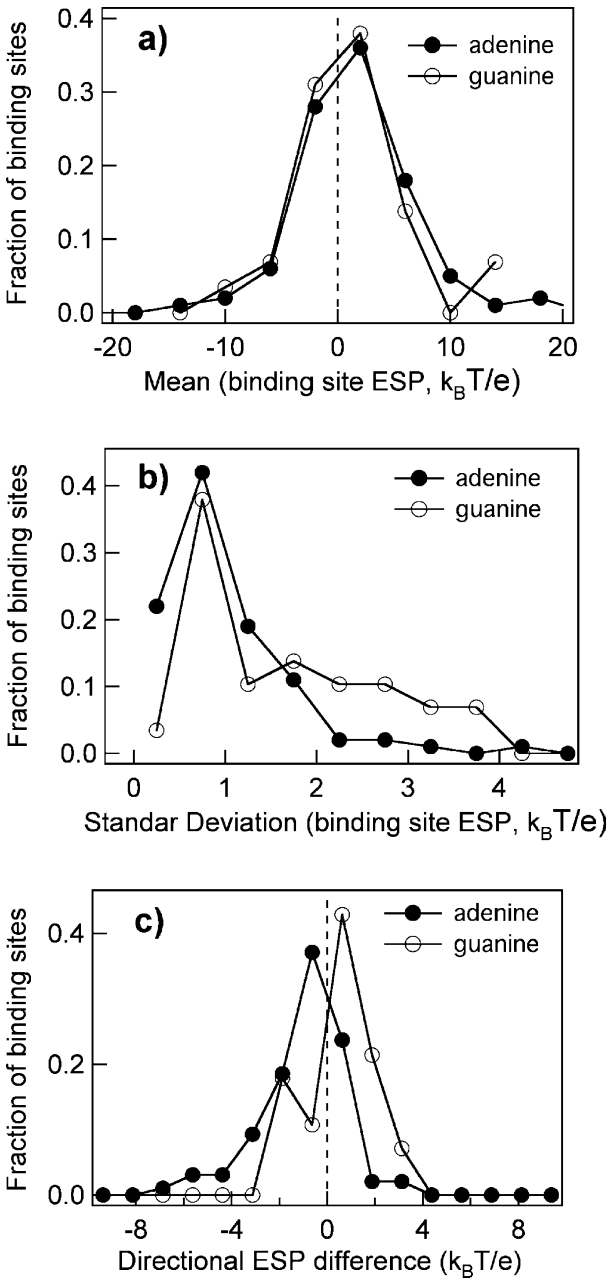

Figure 2. Overview of ESP distribution at adenine (filled circles) and guanine- (open circles) binding sites in proteins: (a) histogram of mean binding site ESP; (b) histogram of standard deviation of binding site ESP; and (c) histogram of directional ESP difference $-[\phi_{N6} - \phi_{N3}]$ for adenine-binding sites and $[\phi_{O6} - \phi_{N3}]$ for guanine-binding sites (see Figure 1 for atom annotations).

Overview of binding site ESP

Electrostatic potentials were calculated for all proteins considered in this work (without the bound A or G containing ligand). Subsequently, the magnitudes of ESP at the binding sites (protein-bound A and G atom positions) of each protein were calculated. Salient features of the resulting binding site ESP are shown in Figure 2 for A-bound (ATP and FAD-set) and G-bound (G-set) proteins. In Figure 2(a), the distributions of mean ESP at A and G-binding sites are shown. The distributions are similar for A and G-binding sites with a slight bias towards positive potentials. The near identical ESP distribution for A and G-binding sites is not surprising, since A and G are electrically neutral, and it is the relative spatial distribution of ESP among the ligand atoms, rather than the absolute value of ESP, that determines the electrostatic interactions of the binding site.

Standard deviations of binding site ESP were calculated for all binding sites. The distributions of standard deviations are shown in Figure 2(b). Unlike the distribution of mean potentials (Figure 2(a)), the distributions of standard deviations for A and G-binding sites are different. Figure 2(b) clearly shows that for the A-binding sites, this spread is

Table 1. The ATP-set

pdb/ligand	Protein name/Comments (Reference)
A. *Non-specific*	
1b0u/ATP	Histidine permease/low nucleotide specificity (*J. Biol. Chem.* (2000). **275**, 29407–29412)
1bg2/ADP	Kinesin motor domain/can utilize GTP (*Biochemistry* (1993). **32**, 4677–4684)
1kpf/AMP	Protein kinase inhibitor/adenosine nucleosides show highest affinity (*Science* (1997).**278**, 286–290)
6rnt/2AM	Ribonuclease T_1/AMP-binding site non-specific (*J. Mol. Biol. (*1992). **223**, 1013–1028)
1a49/ATP	Pyruvate kinase/GDP can bind (*Biochemistry* (1981). **20**, 6711–6720)
1ecj/AMP	Glutamine phosphoribosylpyrophosphate/AMP and GMP are feedback inhibitors (*Protein Sci.* (1998). **7**, 39–51)
1f52/ADP	Glutamine synthetase/GDP feed back inhibitor (*Biochemistry* (1994). **33**, 11184–11188)
1rkd/ADP	ribokinase/GTP activity in *S. typhimurium* (*Arch. Biochem. Biophys.* (1974). **164**, 560–570)
3tsl/TYA	Tyrosyl-tRNA synthetase/GTP competitive inhibitor in aminoacylation and PPi-exchange (*Eur. J. Biochem.*(1990). **193**, 783–788)
1cza/ADP	Hexokinase type I/GTP can bind (*Arch. Biochem. Biophys.* (1991). **291**, 59–68)
1glb/ADP	Glycerol kinase/can bind GMP (*Biochemistry* (1987). **26**, 1723–1727)
1kny/APC	Kanamycin nucleotidyl transferase/can utilize GTP (*Biochemistry* (1995). **34**, 13305–13311)
1nhk/CMP	Nucleoside diphosphate kinase/nucleotide non-specific (*J. Mol. Biol.* (1993). **234**, 1230–1247)
1pfk/ADP	Phosphofructokinase/GDP is also an inhibitor (*J. Mol. Biol.* (1997). **267**, 476–480)
3gap/CMP	cAMP receptor protein/cGMP produces non-functional binding (*J. Biol. Chem.* (1995). **270**, 21679–21683)
1bcp/ATP	Pertussis toxin/GTP can induce subunit dissociation (*J. Biol. Chem.* (1986). **261**, 4324–4327)
1frp/AMP	Fructose-2,6- bisphosphatase/GTP inhibitor (*Biochem. J.* (1997). **328**, 751–756)
1zin/AP5	Adenylate kinase/GTP can be phosphoryl group donor (*Eur. J. Biochem.* (1980). **103**, 481–491)
8gpb/AMP	Glycogen phosphorylase/GMP activator (*J. Mol. Biol.* (2001). **307**, 707–720)
B. *Non-purine sites*	
1rpg/CPA	Ribonuclease A/uridine binds at catalytic site, adenine binds at a sub-site (*Protein Sci.* (1994). **3**, 2322–2339)
1son/AMP	Adenylosuccinate synthetase/AMP binds the IMP-binding pocket (*J. Biol. Chem.* (2002). **43**, 40536–40543)
C. *Adenine-preferred sites*	
1dad/ADP	Dethiobiotin synthetase/GTP 10–20% effective as ATP (*Methods Enzymol.* (1979). **62**, 326–338)
1qb7/ADE	Adenine phosphoribosyltransferase synthetase/guanylate inhibitor (*Biochim. Biophys. Acta*, (1972). **268**, 70–76)
1mxb/ADP	*S*-Adenosylmethionine synthetase/GTP poor inhibitor (*J. Biol. Chem.* (1980). **255**, 9082–9092)
1eqo/APC	6-Hydroxymethyl-7,8-dihydropterin pyrophokinase/GTP shows weaker binding (*Biochim. Biophys. Acta*, (2000). **1478**, 289–299)
1b8a/ATP	Aspartyl-tRNA synthetase/can use GTP (much weaker) instead of ATP (*FEBS Letters* (1996). **394**, 66–70)
1d2a/ADE	Low molecular weight phosphatase/adenine strongest activator (*J. Biol. Chem.* (1995). **270**, 18491–18499)
4at1/ATP	Aspartate carbamoyl transferase/ATP: natural activator; CTP: natural inhibitor; GTP: weaker inhibitor than CTP (*Protein Sci.* (2000). **9**, 953–963)
D. *Adenine-specific sites*	
1byq/ADP	Heat shock protein 90/adenine specific (*Eur. J. Biochem.* (2003). **270**, 2421–2428)
1efv/AMP	Electron transfer flavoprotein/AMP specific (*Biochemistry* (1999). **38**, 1977–1989)
3r1r/ATP	Ribonucleotide reductase protein R1/adenine-specific (*Structure* (1997). **5**, 1077–1092)
1a0i/ATP	DNA ligase/ specific ATP-dependendence (*Cell* (1996). **85**, 607–615)
1ayl/ATP	Phosphoenolpyruvate carboxykinase/ATP-specific in bacteria, GTP-specific in animals (*J. Mol. Biol.* (2002). **316**, 257–264)
1cjt/DAD	Adenylate cyclase/ATP is a specific substrate (*Science* (1999). **285**, 756–760)
1mrj/AND	Trichosanthin/hydrolysis of the *N*-glycosidic bond of A in GAGA (*Proteins: Struct. Funct. Genet.* (2000). **39**, 37–46)
1mud/ADE	Adenine glycosylase/excises adenines from mispairs (*Nature Struct. Biol.* (1998). **5**, 1058–1063)
16pk/BIS	Phosphoglycerate kinase/ATP-specific in *T. brucei* (*Mol. Biochem. Parasitol.* (1993). **60**, 265–272)

For the following we could not find any information in the literature pertaining to guanine-binding or guanine-utilization of the adenine-bound binding sites: 1mjh (ATP/hypothetical protein Mj0577); 1aon (ADP/GroEL-GroES); 1der (ATP/GroEL-GroES); 1bg0 (ADP/arginine kinase); 2gnk (ATP/Glnk); 2uag (ADP/MurD); 1amu (AMP/gramicidin synthetase); 2src (ANP/tyrosine-protein kinase Src); 1nsy (AMP/Nh3-dependent Nad^+ synthetase).

much more limited than in the G-binding sites with a slightly lower mean value. Therefore, the main difference between the A and G-binding sites lies in the distribution of zero-mean binding site ESP.

When the relative values of ESP between specific ligand atoms were considered, distinct trends were observed. Isolated A and G ligands primarily differ in the disposition of H-bond donor/acceptor atoms in the six-membered ring (Figure 1), and so we first examined the directional variation of ESP across the six-membered ring for A and G (up → down, with respect to Figure 1). For the G-binding sites, the difference in ESP at the O6 position and the N3 position was calculated. The ESP difference showed a tendency (about 72% of the cases) in ESP to decrease from atom positions O6 to N3, as evident in the distribution of $[\phi_{O6}-\phi_{N3}]$ in Figure 2(c). For the A-binding sites, on the contrary, a majority of sites (72%) exhibited an increase in ESP from atom positions N6 to N3, as evident in the distribution of $[\phi_{N6}-\phi_{N3}]$ in Figure 2(c). This directional variation of ESP shows: (1) the existence of characteristic ESP patterns at the binding sites; and (2) different ESP patterns for A and G-binding sites. A more thorough analysis of the ESP patterns at the A and G-binding sites follows.

Principal component analysis of the ESP distribution

For a meaningful analysis of binding site ESP distributions, we first normalized the binding site

Table 2. The G-set

pdb/ligand	Protein name/Comments (Reference)
A. *Non-specific*	
1c3x/8IG	Purine nucleoside phosphorylase/adenosine can bind (*J. Mol. Biol.* (1999). **294**, 1239–1255)
1d6a/GUN	Pokeweed antiviral protein/can deadenylate and deguanylate rRNA (*Protein Sci.* (1999). **8**, 2399–2405)
1waj/5GP	DNA polymerase (bacteriophage Rb69)/biological significance of bound 5GP unclear (*Cell* (1997). **89**, 1087–1099)
3rhn/5GP	Histidine triad nucleotide-binding protein/known binders: adenine, 8Br-AMP and GMP (*Nature Struct. Biol.* (1997). **4**, 231–238)
1day/GNP	Protein kinase Ck2/phosphoryl donors: GTP or ATP (*Nature Struct. Biol.* (1999). **6**, 1100–1103)
1ecb/5GP	Glutamine phosphoribosylpyrophosphate amidotransferase/inhibitors: GMP AMP (*Protein Sci.* (1998). **7**, 39–51)
1nue/GDP	Nucleoside diphosphate kinase/both ADP and GDP can bind (*J. Bioenerg. Biomembr.* (2000). **32**, 237–246)
1a8r/GTP	GTP cyclohydrolase I/ATP competitive inhibitor (*Biochem. Clin. Aspects Pteridines* (1984). **3**, 77–92)
1ch6/GTP	Glutamate dehydrogenase/ATP can bind GTP-binding site (*J. Mol. Biol.* (2001). **307**, 707–720)
B. *Non-purine sites*	
1tlc/DGP	Thymidylate synthase/dUMP natural substrate (*Proc. Natl Acad. Sci. USA* (1995). **92**, 3493–3497)
1mre/GDP	IgG Jel 103 Fab fragment/binding site accomodates IDP, GDP, IMP (*J. Mol. Biol.* (1994). **243**, 283–297)
1qhi/BPG	Thymidine kinase/guanine bound to thyamidine-binding site (*FEBS Letters* (1999). **443**, 121–125)
1rnc/CPG	Ribonuclease A/guanine bound to pyrimidine-binding site (*Protein Sci.*(2000). **9**, 1217–1225)
C. *Guanine-specific sites*	
1ckm/GTP	mRNA capping enzyme/GTP specific (*Cell* (1997). **89**, 545–553)
1ej1/M7G	RNA 5′ cap-binding protein/recognizes G-specific caps (*Cell* (1997). **89**, 951–961)
1fsz/GDP	Cell-division protein Ftsz/GTPase: guanine-specific (*Nature* (1992). **359**, 251–254)
1rge/2GP	Ribonuclease Sa/guanine-specific hydrolysis of single stranded RNA (*Acta Crystallog. sect. D* (2002). **58**, 1307–1313)
1v39/MDG	Dc26 mutant of vaccinia Vp39/recognizes G-specific caps (*Proc. Natl Acad. Sci. USA* (1999). **96**, 7149–7154)
1c4k/GTP	Ornithine decarboxylase/guanine-specific activation (*Acta Crystallog. sect D* (1999). **55**, 1978–1985)
1cip/GNP	Guanine nucleotide-binding protein/guanine-specific G-protein (*J. Biol. Chem.* (1999). **274**, 16669–16672)
1dek/DGP	Deoxynucleoside monophosphate kinase/dAMP is not recognized (*EMBO J.* (1996). **15**, 3487–3497)
1gky/5GP	Guanylate kinase/guanine-specific (*J. Mol. Biol.* (1992). **224**, 1127–1141)
1qf5/GDP	Adenylosuccinate synthetase/guanine-specific (*J. Biol. Chem.* (1994). **39**, 24046–24049)
2ng1/GDP	Ng Gtpase fragment of recognition protein Ffh/guanine-specific (*Nature Struct. Biol.* (1999). **6**, 793–801)

The following two proteins (four G-binding sites), are bound to guanine dinucleotides and therefore the G-specificity of the sites were not considered: 1aa6 (formate dehydrogenase; ligand: molybdopterin guanine dinucleotide) and 1dmr (DMSO reductase; ligand: molybdopterin guanine dinucleotide).

ESP distributions by calculating zero-mean (for each binding site, the mean ESP is subtracted from the binding site ESP) ESP values, since the relative variation of ESP within a binding site is our main focus. In addition, being electrically neutral, the energy (sum of charge times potential) of placing A or G in a space defined by some particular ESP is independent of the mean potential. Intensity plots of the zero-mean ESP at the binding sites are shown in Figure 3 for the A-set, FAD-set and G-set. The binding site ESP, organized according to *k*-means clustering, shows clear patterns across the dataset. Although it is tempting to derive consensus templates from the clusters, the boundaries of ESP clusters in Figure 3 are arbitrary and not necessarily disjointed. Instead, further analysis was performed following a more natural way of extracting essential information from the ESP patterns. The ESP pattern of each binding site (Figure 3) is represented by a vector, spanning N_A-dimensional space for A and N_G-dimensional space for G, where $N_A(=15)$ and $N_G(=16)$ are the total number of atoms in A and G, respectively. Often, the essential information content in multi-dimensional space can be represented by a much lower dimensional space spanned by only a few orthogonal vectors, each a unique linear combination of the basis vectors spanning the original multidimensional space. Using principal component analysis (PCA) we first removed any redundancy in the dimensionality in the ESP data, similar to what is often performed to identify functionally relevant collective motions in biological macromolecules.[19]

Four sets of PCA were performed. The first three correspond to the three groups shown in Figure 3 and the fourth was performed on all the three sets combined (combo-set), by first reducing the dimensionality to 13 using a united-atom approach as mentioned in Materials and Methods. In terms of reducing the dimensionality of the A/G binding site ESP distributions, PCA proved to be very effective, since the total variance, accounted for by the first two PC axes, was 93% for all the four sets, indicating that only two PC vectors essentially represent the entire ESP distribution. The first two PC vectors, associated with the highest variances and corresponding to each of these four PCA analyses, are shown in Figure 4. The PC vectors corresponding to the ATP-set, the G-set and the combo-set are very similar in their overall ESP distribution, despite different fractional contributions to the total variance. This is remarkable and implies the robustness of the PC1/PC2 vectors in describing an arbitrary set of purine-binding sites, irrespective of the makeup of the set (pure A, pure G or mixed A/G). In contrast, the FAD-set stands out in that the ESP distributions are different from the other three sets. The FAD-set is, therefore, fundamentally different from the ATP or the G-set. As we will show later, this difference arises due to secondary role played by the A-binding site in the overall ligand binding in the FAD-set.

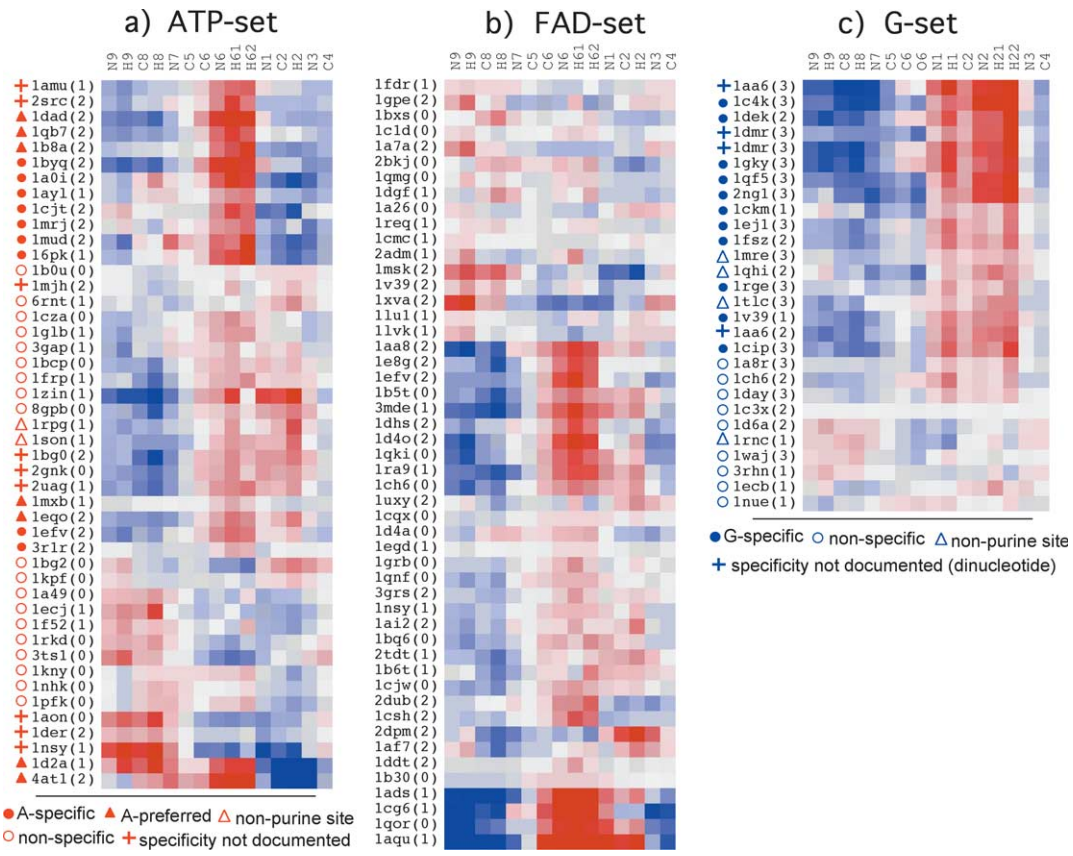

Figure 3. Intensity plots of binding site zero-mean electrostatic potential (blue, positive; red, negative) for (a) ATP-set (ligands: ATP, ADP, AMP, cAMP and A; see Table 1); (b) FAD-set (ligands: all A-containing nucleotides except those included in the ATP-set); and (c) G-set (ligands: G-containing nucleotides; see Table 2). Each binding site is annotated by a pdb code followed by a number indicating how many discriminatory hydrogen bonding atoms in the ligand (N1 and N6 for adenine; N1, N2 and O6 for guanine) are hydrogen bonded[1] in the crystal structure. The biological specificity of binding sites in the ATP and G-sets, obtained from experimental data in primary literature are indicated by appropriate symbols (see Tables 1 and 2). The overall patterns of zero-mean ESP (not the absolute intensities) were relatively insensitive to using a different partial charge set or to a different solvent probe radius (data not shown).

Binding site ESP for the ATP and the G-sets are projected on the combo-set PC1–PC2 plane in Figure 5. The two sets show distinctly different distributions. Along the PC1 axis, the ATP-set is distributed evenly around zero, while the G-set strongly prefers negative values. Along the PC2 axis, the ATP-set prefers positive values, while the G-set prefers negative values. Although the ATP and G-sets, as two separate sets, show some overlap, the A/G-specific sites clearly show a distinctly different preference than the non-specific sites. The non-specific sites are distributed along the diagonal, whereas the A-specific sites prefer the bottom-right corner and the G-specific sites prefer the top-left corner with non-overlapping distributions. In other words, binding site ESP patterns are strongly correlated with A/G specificity. When binding site ESP of the FAD-set is projected on the combo-set PC1–PC2 plane (Figure 5, inset), the resulting distribution is very similar to the non-specific sites. In essence, what we have shown here is that patterns of ligand-free protein ESP at purine binding sites are sufficient for discrimination between A and G-specific sites. Further, the ESP patterns of A/G-specific sites are different from non-specific sites.

Simple cognate and non-cognate energies

In the above discussion, ESP distribution is treated as a simple pattern associated with A/G-binding sites, without any associated physical meaning. In reality, ESP and energies of electrostatic interaction are related. However, without any information about the ESP of the free ligands and ligand-bound protein, ESP distribution of ligand-free protein can only provide indirect and approximate estimates of ligand–protein electrostatic interaction. We define (see Materials and Methods) simple cognate (E_{AA} and E_{GG}) and non-cognate (E_{AG} and E_{GA}) energies from ESP distribution of ligand-free protein. They represent the electrostatic energy of placing isolated and non-interacting ligand point charges at the binding site characterized by the ligand-free protein ESP. The relationship between simple energies and more realistic electrostatic components of binding free energies[20] will be discussed later.

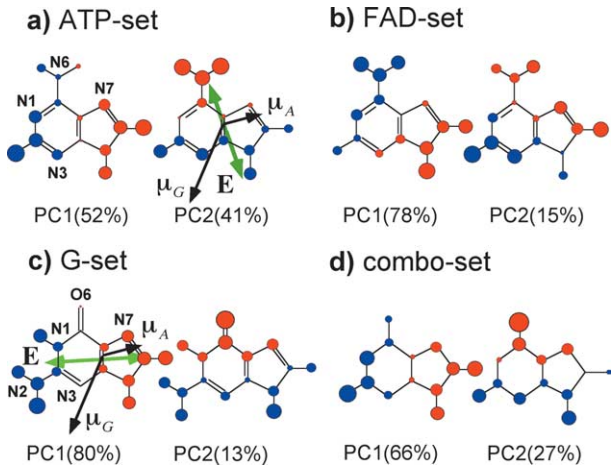

Figure 4. The electrostatic potential distributions of the first two PC vectors (unit vectors along positive direction), projected on A or G atom positions, from PCA of (a) ATP-set (all-atom); (b) FAD-set (all-atom); (c) G-set (all-atom); and (d) united-atom combo-set (ATP, FAD and G-sets combined). The magnitudes of positive (blue) and negative (red) potentials are indicated by the relative size of the atoms. Numbers within parentheses correspond to the fractional contribution of a particular PC axis towards total variance. For two special cases (see the text), ATP-set PC2 and G-set PC1, the relative orientations of the A and G-dipoles (black arrows) and the electric field vectors (green arrows) associated with the PC vectors are shown. The electric field vectors are given by gradients (at the geometric centers of the binding sites) of the electrostatic potential of unit PC vectors along both positive and negative directions along the PC axes.

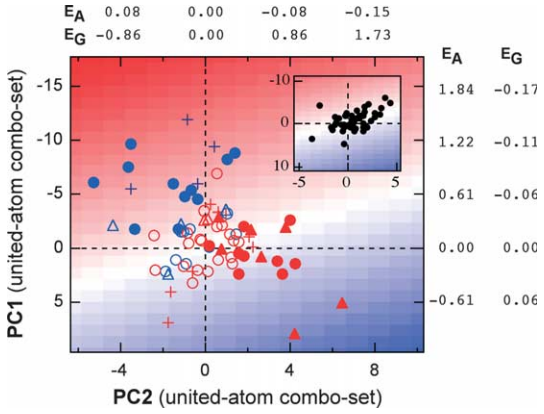

Figure 5. Binding site electrostatic potential distributions for the entire dataset projected on the combo-set (Figure 4(d)) PC1–PC2 plane. The ATP-set (red) and the G-set (blue) are annotated according to their A/G specificity (see the legend to Figure 3 for explanations for the symbols). The FAD-set (black) is shown in the inset. Energies (in k_BT) of placing A (E_A) or G (E_G) partial charges in a space defined by the ESP of the combo-set PC vectors (equation (2); with $\Phi_{ij}^{A} = \Phi_{ij}^{G} = \Phi_{ij}^{PC}$, $E_A = E_{GA}$ and $E_G = E_{AG}$) are shown along the two axes. The PC1–PC2 plane is color-coded by the difference in interaction energies ($E_G - E_A$, in k_BT). A positive energy (blue) indicates replacement of A (by G) as unfavorable while negative energy (red) indicates replacement of G (by A) as unfavorable.

In Figure 6, histograms of cognate and non-cognate energies for the ATP and FAD-sets are shown. The cognate and non-cognate energy distributions overlap for both the FAD-set and the non-specific sites in the ATP-set. On the other hand, the two distributions are very different for the A-specific (and A-preferred) sites in the ATP-set, all non-cognate energies are positive while all cognate energies are negative. The situation is very similar for the G-set as well (Figure 7) where the cognate/non-cognate energy distributions overlap for the non-specific sites and separate for the G-specific sites. The picture that emerges from the energy distributions, in terms of their correlation with A/G specificities, is consistent with what was already assessed from the PCA of ESP patterns.

Histograms of replacement (cognate by non-cognate ligand) energies ΔE (equation (3)), the difference between cognate and the corresponding non-cognate energy, are shown in Figure 8 for the ATP, FAD and G-sets. For the ATP-set (Figure 8(a)), the non-specific and A-specific sites exhibit very different distributions. The mean $\Delta E_{A\to G}$ value for the A-specific sites is higher than the non-specific sites, indicating that replacement of A (cognate) by G (non-cognate) for A-specific sites is electrostatically more unfavorable than similar ligand replacements in the non-specific sites. The correlation between experimentally established A specificity and simple replacement energy is remarkable. The rest of the A-bound sites in the dataset, the FAD-set, for which we argued that ligand specificity, if any, might not necessarily arise from the purine-binding sites, exhibit a $\Delta E_{A\to G}$ distribution (Figure 8(a)) very similar to the non-specific sites in the ATP-set. Thus, the $\Delta E_{A\to G}$ distributions of the FAD-set and the non-specific sites in the ATP-set are overlapping. Similar to A-specific and non-specific sites in the ATP-set, the G-specific and non-specific sites in the G-set are also distinct from each other in terms of their $\Delta E_{G\to A}$ distributions (Figure 8(b)). The average value of $\Delta E_{G\to A}$ for the G-specific sites is higher than the non-specific sites in the G-set, implying that replacement of G (cognate) by A (non-cognate) at G-specific sites is electrostatically more unfavorable than similar ligand replacements in the non-specific sites. Therefore, both A-specific and G-specific sites show a clear correlation between electrostatic cost for ligand replacement (cognate by non-cognate) and their respective A/G specificities. The non-specific sites and members of the FAD-set are near neutral to ligand exchange.

Electrostatic basis for discrimination between A and G has also recently been reported by Rockey & Elcock[21] who considered the distribution of average docking energies between two ligands (ADP and GDP) and a set of receptors (ADP-binding and GDP-binding). Energy distributions of cognate and

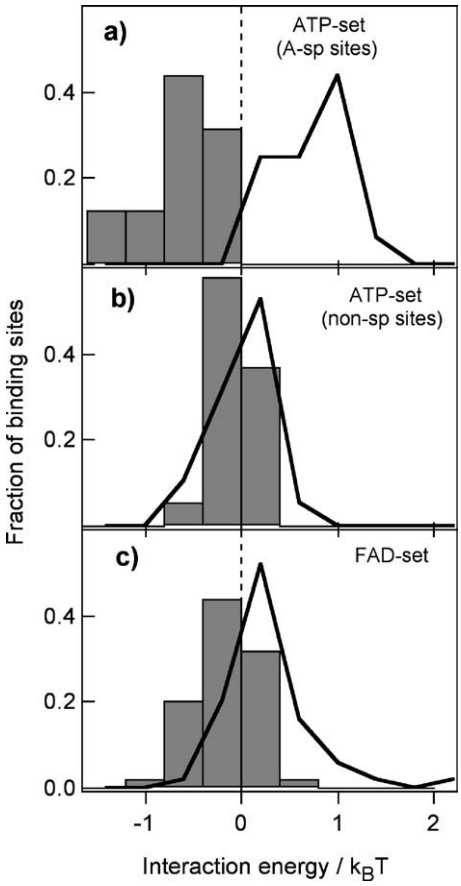

Figure 6. Histograms of cognate (shaded bars; E_{AA}) and non-cognate (continuous line; E_{AG}) energies for: (a) A-specific (and A-preferred) sites in the ATP-set ($\langle E_{AA}\rangle = -0.62$; $\langle E_{AG}\rangle = 0.75$); (b) non-specific sites in the ATP-set ($\langle E_{AA}\rangle = -0.05$; $\langle E_{AG}\rangle = 0.01$); and (c) the FAD-set ($\langle E_{AA}\rangle = -0.16$; $\langle E_{AG}\rangle = 0.27$). The specificity annotation of the ATP-set is literature-derived and shown in Figure 3 and Table 1.

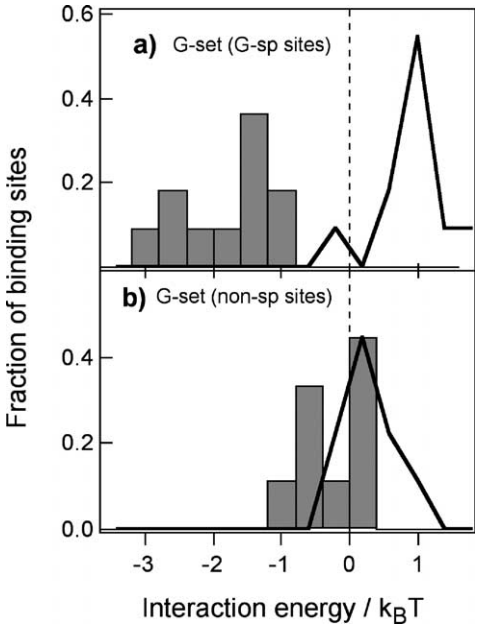

Figure 7. Histograms of cognate (shaded bars; E_{GG}) and non-cognate (continuous line; E_{GA}) energies for: (a) G-specific sites in the G-set ($\langle E_{GG}\rangle = -1.75$; $\langle E_{GA}\rangle = 1.01$); (b) non-specific sites in the G-set ($\langle E_{GG}\rangle = -0.33$; $\langle E_{GA}\rangle = 0.22$). See Figure 3 and Table 2 for the G-specificity annotation of the G-set.

non-cognate complexes were not much different when the total (electrostatic and non-electrostatic) ligand–protein interaction energies were considered. However, when only the electrostatic interaction energy of the purine ring moiety was considered, the cognate and non-cognate energy distributions separated. Although the essential message of our work, that the basis of A/G discrimination by proteins is electrostatic, and that of Rockey & Elcock[21] are similar, there are differences. One difference is in the methodology itself. While Rockey & Elcock[21] restricted their work to only ADP and GDP, and used averaged docking energies based on molecular mechanics and simple solvation parameters, we consider a wide variety of A/G-containing ligands and our interaction energies are derived from ligand-free protein ESP. In fact, even without any explicit energy calculations, we showed that a suitable analysis of ligand-free ESP at A/G-binding sites is enough for discerning A/G specificity. As a result of using a variety of ligands, we were also able to provide some new

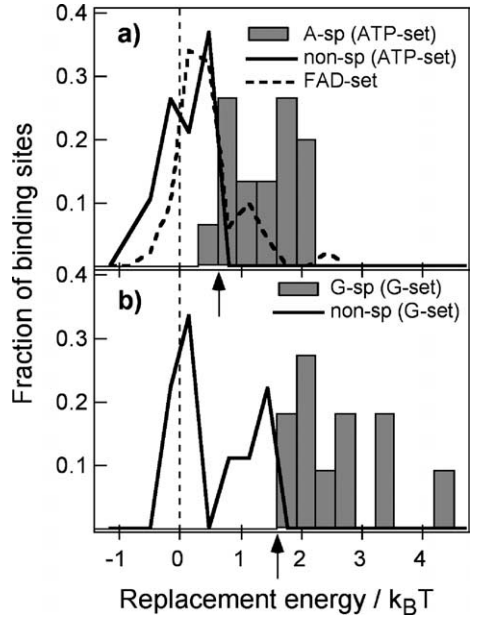

Figure 8. Histograms of ligand replacement energy (cognate by non-cognate) for: (a) the ATP and FAD-sets (replacement energy $= E_{AG} - E_{AA}$); and (b) the G-set (replacement energy $= E_{GA} - E_{GG}$). The A/G specificity information is literature-derived and shown in Figure 3 and Tables 1 and 2. The two arrows indicate threshold values for replacement energy corresponding to the best separation between A/G-specific and non-specific sets.

insights, the near resemblance of the FAD-set with the non-specific sites. Nevertheless, despite the differences in our approaches, the similarity of our conclusion and that of Rockey & Elcock[21] only reinforces our methodology and results. The merit of our approach of separating specific from non-specific sites before analyzing the ESP data, is also reflected in Rockey & Elcock's observation that separation of cognate and non-cognate energy distributions for ADP (as the ligand) is more pronounced when only ADP-specific receptors are used.

Simple replacement energy and the electrostatic component of binding free energy differences

Simple replacement energies, being based on only ligand-free protein ESP, do not reflect the entire electrostatic cost of ligand replacement. The electrostatic component of binding free energy difference $\Delta\Delta G$ (equation (A2); G in $\Delta\Delta G$ and ΔG stands for the electrostatic component of free energy throughout the text) captures ligand replacement more comprehensively, since it includes appropriate solvation and desolvation terms (for the free protein and ligand) and proper solvent screening (in the context of the bound state). Although simple replacement energy is adequate for arriving at the main conclusion of this work, it is still important to ask whether the correlation between ΔE and A/G specificity also implies a similar correlation between $\Delta\Delta G$ and A/G specificity. In Appendix A we relate ΔE and $\Delta\Delta G$ in equation (A5). Simply put, $\Delta\Delta G$ is equal to ΔE multiplied by a factor λ (>1), that captures solvent screening for ligand–protein interaction in the context of the bound state, plus a term that represents solvation changes in A/G upon ligand replacement. Because the A/G solvation term is positive for A→G replacement (see Appendix A), it is straightforward to conclude that $\langle \Delta E^{A-sp}_{A\to G}\rangle > \langle \Delta E^{non-sp}_{A\to G}\rangle$, obtained from Figure 8, also implies $\langle \Delta\Delta G^{A-sp}_{A\to G}\rangle > \langle \Delta\Delta G^{non-sp}_{A\to G}\rangle$. The solvation term is negative for G→A replacement, therefore the exact balance between the solvation term and $\lambda(\Delta E^{G-sp}_{G\to A} - \Delta E^{non-sp}_{G\to A})$ will decide if $\langle \Delta E^{G-sp}_{G\to A}\rangle > \langle \Delta E^{non-sp}_{G\to A}\rangle$, obtained from Figure 8, also implies $\langle \Delta\Delta G^{G-sp}_{G\to A}\rangle > \langle \Delta\Delta G^{non-sp}_{G\to A}\rangle$. In Appendix A we show this to be true from simple estimates of solvation changes in A/G and λ.

Dipolar nature of electrostatic interaction

Given a set of ligand partial charges, a variety of binding site ESP distributions could potentially yield the same interaction energy. However, the A/G-specific binding sites are not only distinct in terms of replacement energies, but the ESP patterns that give rise to this energetic difference are also conserved. One reason for this is the inherent nature of ESP to have a smooth spatial variation in the protein exterior. In addition, it is under the functional constraint to be A/G discriminatory. Thus, not only two PC vectors capture the essence of ESP distributions, the direction of electric field ($\boldsymbol{E}$) associated with each PC vector, given by the gradient of the ESP distribution, is roughly aligned with $\boldsymbol{\mu}_A$ and $\boldsymbol{\mu}_G$ (see PC2 in Figure 4(a) and PC1 in Figure 4(c)). Near alignment of $\boldsymbol{E}$ with $\boldsymbol{\mu}$ implies maximum A/G recognition or discrimination under a simple dipolar approximation (energy$=\boldsymbol{\mu}\cdot\boldsymbol{E}$); maximum recognition corresponds to $\boldsymbol{E}/\boldsymbol{\mu}$ angle $\sim$180° and maximum discrimination corresponds to $\boldsymbol{E}/\boldsymbol{\mu}$ angle $\sim$0°. In fact, the key conclusions of this work can be completely reproduced using simple dipolar energies (data not shown).

Energies (in k_BT) of placing A (E_A) or G (E_G) partial charges in a space defined by the ESP of the combo-set PC vectors can be calculated using equation (2). This energy is shown along the two axes in Figure 5. The difference ($E_A - E_G$), or ($E_G - E_A$), represents the contribution to the simple replacement energies from the PC axes and is the basis of color-coding of the PC1–PC2 plane. Similar to the energy distributions in Figure 8, the A/G-specific sites are strongly biased towards positive replacement energies while the non-specific sites and the FAD-set are distributed along a direction characterized by zero replacement energy. This implies a strong correlation between the actual energy (Figure 8) and the energy contributions from only PC1 and PC2 (Figure 5), underlining the near complete representation of the electrostatic component of the binding site energetics by only PC1 and PC2.

Compared to A-specific sites, G-specific sites show a higher affinity for cognate ligand ($E_{GG} < E_{AA}$), although the non-cognate ligand affinities are identical ($E_{AG} \sim E_{GA}$) (Figures 6 and 7), giving rise to somewhat higher replacement energies for the G-set compared to the ATP and FAD-sets (Figure 8). The origin of this difference can be understood from a simple picture of dipolar interaction, where E_{AG} ($=|\mathbf{E}_A||\boldsymbol{\mu}_G|\cos\theta$) and E_{GA} ($=|\mathbf{E}_G||\boldsymbol{\mu}_A|\cos\theta$) take the most positive (most discrimination) values when $\boldsymbol{E}$ is aligned to non-cognate $\boldsymbol{\mu}$. In order to yield comparable non-cognate ligand-affinities (given comparable magnitudes of $\boldsymbol{E}_G$ and $\boldsymbol{E}_A$), $\boldsymbol{E}_G$ is under more constraint to be aligned with $\boldsymbol{\mu}_A$ than $\boldsymbol{E}_A$ is with $\boldsymbol{\mu}_G$. This is because $|\boldsymbol{\mu}_A| \sim |\boldsymbol{\mu}_G|/2$. The nature of the ATP-set PC2 axis and the G-set PC1 axis (see Figure 4), the two main discriminatory axes identified earlier, demonstrates this fact. The G-set PC1 axis exhibits a much larger fractional contribution to total variance (80%) than the ATP-set PC2 axis (41%) and $\boldsymbol{E}$ is more aligned to the non-cognate $\boldsymbol{\mu}$ ($\mathbf{E}^{PC1}_{G-set}/\boldsymbol{\mu}_A$ angle $\sim$15°; $\mathbf{E}^{PC2}_{ATP-set}/\boldsymbol{\mu}_G$ angle $\sim$40°). The differential angular constraints, arising from comparable non-cognate ligand discrimination, implicitly translate into different cognate ligand recognition energies for A-specific and G-specific sites: $\mathbf{E}^{PC1}_{G-set}$ favors $\boldsymbol{\mu}_G$ ($\mathbf{E}^{PC1}_{G-set}/\boldsymbol{\mu}_G$ angle $>$90°) while $\mathbf{E}^{PC2}_{ATP-set}$ is A-insensitive ($\mathbf{E}^{PC2}_{ATP-set}/\boldsymbol{\mu}_A$ angle $\sim$90°). Thus, G-specific sites are more

optimized for cognate ligand-recognition than A-specific sites. In other words, a larger magnitude of the G dipole and the particular orientation of the ESP gradients of PC1 and PC2 in A/G-specific sites in proteins give rise to the larger replacement energy for the G-set. To draw any biologically meaningful conclusion from the fact that the G-specific sites show larger replacement energies than the A-specific sites, replacement energies need to be first corrected to yield more realistic binding free energy differences by adding the desolvation and solvent screening terms, an accurate estimate of which is beyond the scope of this work.

A/G-specific hydrogen bonds

The nature of protein hydrogen bonds is known to be dominantly electrostatic.[22] Consistent with this notion, observed hydrogen bonds at the A/G-binding sites in the dataset were found to correlate well with local (around hydrogen bonding donor/acceptor sites; see Figure 1) patterns of ESP distribution (data not shown). However, Nobeli *et al.*,[1] who used the same dataset as us, failed to detect clear correlation between protein–ligand hydrogen bonds and A/G discrimination. For each entry in Figure 3, we show the number of discriminatory hydrogen bonds (N1 and N7 for A and N1, N2 and O6 for G) as reported by Nobeli *et al.*[1] If one classifies the entire dataset into A-binding and G-binding sites, as was done by Nobeli *et al.*,[1] clearly there is no correlation between the observed hydrogen bonds and the two groups. However, if the A/G specificities of binding sites are used to classify the dataset, hydrogen bonds and annotated A/G-specific and non-specific sites are found to be moderately correlated. This validates our strategy of classification and functional annotation of the A/G-binding sites and demonstrates the importance of a close examination of the functional properties of proteins from independent experimental data before deriving any biologically meaningful conclusion from simple structural analyses.

The observed correlation between A/G specificity and the number of discriminatory hydrogen bonds in the X-ray structure can be used to predict A/G specificity by formulating a simple rule: (1) A-sites with two discriminatory hydrogen bonds are A-specific; and, (2) G-sites with three discriminatory hydrogen bonds are G-specific. An examination of Figure 3 shows that, based on these simple rules, four (out of 16) A-specific and A-preferred sites are wrongly predicted to be non-specific while all (19) non-specific A-sites are predicted correctly for the ATP-set. For the G-sites, four (out of 11) G-specific sites are wrongly predicted to be non-specific while three (out of nine) non-specific G-sites are wrongly predicted to be G-specific. This yields a prediction accuracy (correlation coefficient; equation (4)) of 0.79 for the A-set and 0.30 for the G-set. We will compare this prediction accuracy with an ESP-based prediction in the next section.

Implications for prediction of A/G specificity of known and putative purine-binding sites

A nucleotide-binding protein with known structure may or may not be associated with a bound nucleotide in the X-ray (or NMR) structure. When the protein is already bound to a nucleotide, as were all members of the database used here, based on the ESP potential of the purine-base binding site, one can predict the binding to be A/G-specific or non-specific. The underlying philosophy for such a prediction is similar to quantitative structure–activity relationships (QSAR)-based methods like CoMFA,[23] which involves statistical analysis of a set of descriptors or properties for a series of biologically active ligands in order to predict the activity of additional ligands. In our approach, by analyzing ESP-based replacement energy-A/G specificity patterns for a large number of binding sites, we identify a threshold value for the appropriate replacement energy that separates the A/G-specific and non-specific sites and use it to predict the properties of new and unknown binding sites.

For example, for A-binding sites in the ATP-set, a threshold value of 0.65 k_BT for $\Delta\mathbf{E}_{A\rightarrow G}$ correctly predicts all A-specific, A-preferred and non-specific sites (Figure 8(a)). The only exception is *S*-adenosylmethionine synthetase (1mxb), an A-preferred site that is predicted to be non-specific. Similarly, for G-binding sites in the G-set, a threshold value of 1.75 k_BT for $\Delta\mathbf{E}_{A\rightarrow G}$ correctly predicts all G-specific and non-specific sites (Figure 8(b)). The near-perfect prediction accuracy (0.94 for the ATP-set and 1.0 for the G-set) may be an artifact of the small size of the dataset, but clearly it is better than hydrogen bond-based prediction on the same dataset.

The reason why analysis of binding site ESP is superior to simple counting of hydrogen bonds[10,24] for detecting A/G specificity is that A/G specificity arises from the electrostatic effect of the protein as felt by the entire purine base (binding site ESP) and not necessarily as felt by only a few atoms in the purine base involved in hydrogen bonds. To assess the global *versus* local contribution of the protein towards the binding site ESP, we recalculated ESP at five select A-specific binding sites by turning off charges of all amino acid residues for which at least one heavy atom was not within 4 Å (6 Å for one case) from the A-ring. The resulting ESP patterns were very similar to the original full-charge ESP patterns and yielded replacement energies almost identical to the full-charge calculation. Thus, the dominant contribution to binding site ESP comes from residues immediately surrounding the binding site. However, a combined effect of all neighboring residues (typically ~10), rather than specific contributions from one or two select neighboring residues, contributed to the overall ESP pattern.

Next, we predict A/G specificities of binding sites in the dataset with unknown specificities using the threshold values of replacement energies used before. Of the nine A-binding sites with unknown specificity (footnote to Table 1), 1amu (gramicidin

synthetase), 2src (tyrosine-protein kinase C-Src) and 1nsy (NH3-dependent NAD^+ synthetase) are predicted to be A-specific/A-preferred. Similarly, of the 50 members of the FAD-set, A-binding sites of 11 are predicted to be A-specific/A-preferred: 1aa8 (D-amino acid oxidase), 1e8g (vanillyl-alcohol oxidase), 1efv (electron transfer flavoprotein), 1dhs (deoxyhypusine synthetase), 1d4o (NADP(H) binding domain III of transhydrogenase), 1ads (aldose reductase), 2dub (2-enoyl-CoA hydratase), 1csh (citrate synthetase), 1msk (methionine synthetase), 1cg6 (5′-deoxy-5′-methylthioadenosine phosphorylase) and 1aqu (estrogen sulfotransferase). For the G-set, one non-purine binding site, that in 1qhi (thymidine kinase), is predicted to be G-specific while two proteins bound to guanine dinucleotides, 1aa6 (formate dehydrogenase) and 1dmr (DMSO reductase), are predicted to be G-specific.

When the nucleotide-binding site is unknown in a nucleotide binding protein of known structure, a typical strategy is to predict several putative binding sites using a suitable docking program. As has been shown from docking studies with A/G-containing ligands,[3,21] the docking energy often cannot differentiate between A/G binding sites. Under such a situation one can envisage constructing a suitable score that can provide additional information that simple docking energies lack, as was attempted recently by Zhao *et al*.[3] Our result, that ligand-free ESP is enough for identification of a purine-binding site as A/G-specific or non-specific, implies that a suitable score (for example ΔE of equation (3)), based on the ligand-free ESP of the protein, can be used to annotate putative purine-binding sites (from docking studies) as A/G-specific or not. It requires only a single calculation for the determination of the ligand-free protein ESP. Of course, this does not improve the quality of the docked complex, but it allows one to predict the specificity of the putative purine-binding site.

Conclusion

We have shown that key features of ESP distributions at A/G-binding sites are conserved across protein families with fundamental implications on the origin of cognate/non-cognate ligand (A/G) discrimination by proteins. The strong correlation between experimental A/G specificity of the binding sites and simple A/G replacement energies, despite a total neglect of the non-electrostatic effects in the latter, clearly demonstrates that A/G discrimination by proteins is dominantly electrostatic in nature, although it appears fuzzy in terms of conserved sequence or structural motifs in the neighborhood of the binding site.[1] Additional support for electrostatic control comes from the fact that the A/G replacement energies of the FAD-set are distinctly different from the A/G-specific sites, the former overlapping only with non-specific sites. The binding site ESP for the dataset can be well expressed as a linear combination of two ESP distributions with gradients ($\boldsymbol{E}$) roughly aligned with $\boldsymbol{\mu}_A$ or $\boldsymbol{\mu}_G$, emphasizing the dipolar nature of ligand–protein interaction as the main mechanism of A/G-discrimination. From a bioinformatics viewpoint, functional characterization of unknown binding sites in proteins is a continuing challenge.[1–3,17,21,25–27] For pre-identified purine-binding sites, either from X-ray structure or from docking studies, our results suggest that one can use a score based on simple replacement energies (equation (3)) for annotating the sites as non-specific or A/G-specific. Continuing work in our laboratory is directed towards realizing this goal by refining our calculations on a larger set of proteins.

Materials and Methods

The dataset

A non-redundant set of protein–ligand complexes was chosen, identical (for the sake of comparison and consistency) to that used by Nobeli *et al*.[1] The dataset was divided into three sets: (1) ATP-set (bound ligands: ATP, ADP, AMP, cAMP and A); (2) FAD-set (bound ligands: dinucleotides, CoA and its derivatives, *S*-adenosyl-L-homocysteine and *S*-adenosylmethionine); and (3) G-set (all G-containing ligands, mostly mono-nucleotides). All ligands for the ATP and the G-sets are shown in Tables 1 and 2.

Electrostatic potential at binding sites

ESP (ϕ) was calculated for the non-redundant set of protein–ligand complexes by numerically solving the linearized Poisson–Boltzmann equation (Delphi module of Insight II package; Biosym Inc.). All ligands and heteroatoms were removed and hydrogen atoms were added before the calculation. The protein was placed at the center of a cubic box with 70% of the box-edge occupying the protein's longest Cartesian dimension. The following parameters were used: partial charges: AMBER;[28] protein and solvent dielectric constants: 2 and 80; ionic strength: 0.145 M; ionic radius: 2.0 Å; solvent probe radius: 1.4 Å; grid resolution: 0.6 Å/grid point. Binding site ESP values (at each protein-bound A and G atom sites) were obtained by linear interpolation of ESP of surrounding grid points.

Principal component analysis

PCA was performed in multi-dimensional ESP-space by diagonalization of the variance–covariance matrix:

$$\sigma_{jk} = \frac{1}{N-1}\sum_i (\phi_{ij} - \bar{\phi}_j)(\phi_{ik} - \bar{\phi}_k)$$

where the indices j and k run over purine atoms, and the index i corresponds to binding sites. The ESP-space, equal to the total number of ligand atoms, is 15-dimensional for the ATP and FAD-sets and 16-dimensional for the G-set. In order to perform PCA on all the three sets combined (combo-set), the dimensionality of the ESP-space was reduced to 13 using united-atom potentials ($\boldsymbol{\Phi}$) for both the ATP and FAD-sets [$\boldsymbol{\Phi}^A_{N6} = (\phi^A_{N6} + \phi^A_{H61} + \phi^A_{H62})/3$; $\boldsymbol{\Phi}^A_{H61} = \boldsymbol{\Phi}^A_{H62} = 0$], and the G-set [$\boldsymbol{\Phi}^G_{N1} = (\phi^G_{N1} + \phi^G_{H1})/2$; $\boldsymbol{\Phi}^G_{H1} = 0$; $\phi^G_{N2} = (\phi^G_{N2} + \phi^G_{H21} + \phi^G_{H22})/3$; $\boldsymbol{\Phi}^G_{H21} = \boldsymbol{\Phi}^G_{H22} = 0$].

United-atom potentials ($\mathbf{\Phi}$) were also used for calculating non-cognate energies as explained below.

Cognate and non-cognate energies

Cognate energies, E_{AA} (energy of placing A at A-binding site) and E_{GG} (energy of placing G at G-binding site), were calculated as:

$$E_{AA}(i) = \sum_{j=1}^{15} \phi_{ij}^{A} q_j^{A}, \quad E_{GG}(i) = \sum_{j=1}^{16} \phi_{ij}^{G} q_j^{G} \qquad (1)$$

where q_j^{A}/q_j^{G} is the partial charge of the jth A/G-atom and $\phi_{ij}^{A}/\phi_{ij}^{G}$ is the ESP at the jth A/G atom position in the ith A/G-bound protein.

Strictly speaking, the corresponding non-cognate interaction energies, E_{AG} and E_{GA}, representing the placement of G at the A-binding site and A at the G-binding site, can be calculated by exchanging q_j^{A} and q_j^{G} in equation (1). However, the number of atoms are non-identical in A and G, and therefore there is always an uncertainty in optimal placement of the G moiety in cognate A-binding sites (and *vice versa*). In addition, ESP at some non-cognate atom positions (e.g. H61, H62 in A and H21, H22, H1 in G) may be associated with unrealistic values because in the cognate-ligand–protein complex those sites are too close to the protein surface or are buried. Hence, we adopted an *ad hoc* united-atom approach that represents parts of A or G in terms of united atoms. In the united-atom scheme, partial charges are represented by Q (instead of q) with the q set of charges being identical to Q set of charges except for: $Q_{N6}^{A} = q_{N6}^{A} + q_{H61}^{A} + q_{H62}^{A}$; $Q_{H61}^{A} = Q_{H62}^{A} = 0$; $Q_{N2}^{G} = q_{N2}^{G} + q_{H21}^{G} + q_{H22}^{G}$; $Q_{H21}^{G} = Q_{H22}^{G} = 0$; $Q_{N1}^{G} = q_{N1}^{G} + q_{H1}^{G}$; $Q_{H1}^{G} = 0$. This *ad hoc* scheme, of using Q instead of q and $\mathbf{\Phi}$ instead of ϕ, is strictly correct (in terms of yielding correct energies) if $\phi_{N6}^{A} \approx \phi_{H61}^{A} \approx \phi_{H62}^{A}$, $\phi_{N2}^{G} \approx \phi_{H21}^{G} \approx \phi_{H22}^{G}$ and $\phi_{N1}^{G} \approx \phi_{H1}^{G}$. This was found to hold true for the ESP of the dataset. The non-cognate interaction energies can now be defined as:

$$E_{AG}(i) = \sum_{j=1}^{13} \mathbf{\Phi}_{ij}^{A} Q_j^{G}, \quad E_{GA}(i) = \sum_{j=1}^{13} \mathbf{\Phi}_{ij}^{G} Q_j^{A} \qquad (2)$$

Cognate by non-cognate replacement energies ($\Delta E_{A\to G}$ and $\Delta E_{G\to A}$) are defined as:

$$\Delta E_{A\to G} = E_{AG} - E_{AA}, \quad \Delta E_{G\to A} = E_{GA} - E_{GG} \qquad (3)$$

They represent the total energy of removal of a cognate ligand and placement of a non-cognate ligand at a binding site.

Prediction accuracy

Prediction accuracies were judged from correlation coefficients (cc) constructed from true positives (TP), true negatives (TN), false positives (FP), false negatives (FN), actual positives (AP = TP + FN), predicted positives (PP = TP + FP), actual negatives (AN = FP + TN) and predicted negatives (PN = TN + FN) as:

$$\mathrm{cc} = \frac{[(\mathrm{TP})(\mathrm{TN}) - (\mathrm{FP})(\mathrm{FN})]}{\sqrt{[(\mathrm{AN})(\mathrm{PP})(\mathrm{AP})(\mathrm{PN})]}} \qquad (4)$$

Acknowledgements

We acknowledge financial support from the special coordination fund promoting science and technology, MEXT, Japan. We are also grateful to Dr I. Nobeli for kindly providing relevant hydrogen bonding and accessible surface area data for the proteins studied and thank Robert Wright for a careful reading of the manuscript.

References

1. Nobeli, I., Laskowski, R. A., Valdar, W. S. & Thornton, J. M. (2001). On the molecular discrimination between adenine and guanine by proteins. *Nucl. Acids Res.* **29**, 4294–4309.
2. Campbell, S. J., Gold, N. D., Jackson, R. M. & Westhead, D. R. (2003). Ligand binding: functional site location, similarity and docking. *Curr. Opin. Struct. Biol.* **13**, 389–395.
3. Zhao, S., Morris, G. M., Olson, A. J. & Goodsell, D. S. (2001). Recognition templates for predicting adenylate-binding sites in proteins. *J. Mol. Biol.* **314**, 1245–1255.
4. Walker, J. E., Saraste, M., Runswick, M. J. & Gay, N. J. (1982). Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *EMBO J.* **1**, 945–951.
5. Saraste, M., Sibbald, P. R. & Wittinghofer, A. (1990). The P-loop—a common motif in ATP- and GTP-binding proteins. *Trends Biochem. Sci.* **15**, 430–434.
6. Schulz, G. E. (1992). Binding of nucleotides by proteins. *Curr. Opin. Struct. Biol.* **2**, 61–67.
7. Traut, T. W. (1994). The functions and consensus motifs of nine types of peptide segments that form different types of nucleotide-binding sites. *Eur. J. Biochem.* **222**, 9–19.
8. Kinoshita, K., Sadanami, K., Kidera, A. & Go, N. (1999). Structural motif of phosphate-binding site common to various protein superfamilies: all-against-all structural comparison of protein–mononucleotide complexes. *Protein Eng.* **12**, 11–14.
9. Kobayashi, N. & Go, N. (1997). ATP binding proteins with different folds share a common ATP-binding structural motif. *Nature Struct. Biol.* **4**, 6–7.
10. Denessiouk, K. A. & Johnson, M. S. (2000). When fold is not important: a common structural framework for adenine and AMP binding in 12 unrelated protein families. *Proteins: Struct. Funct. Genet.* **38**, 310–326.
11. Gilson, M. K. & Honig, B. (1987). Calculation of electrostatic potentials in an enzyme active site. *Nature*, **330**, 84–86.
12. Sternberg, M. J. E., Hayes, F. R. F., Russel, A. J., Thomas, P. G. & Fersht, A. R. (1987). Prediction of electrostatic effects of engineering of protein charges. *Nature*, **330**, 86–88.
13. Selzer, T., Albeck, S. & Schreiber, G. (2000). Rational design of faster associating and tighter binding protein complexes. *Nature Struct. Biol.* **7**, 537–541.
14. Sheinerman, F. B. & Honig, B. (2002). On the role of electrostatic interactions in the design of protein–protein interfaces. *J. Mol. Biol.* **318**, 161–177.
15. Kumar, S. & Nussinov, R. (2004). Different roles of electrostatics in heat and in cold: adaptation by citrate synthase. *Chembiochem*, **5**, 280–290.
16. Looger, L. L., Dwyer, M. A., Smith, J. J. &

Hellinga, H. W. (2003). Computational design of receptor and sensor proteins with novel functions. *Nature*, **423**, 185–190.
17. Kinoshita, K. & Nakamura, H. (2003). Protein informatics towards function identification. *Curr. Opin. Struct. Biol.* **13**, 396–400.
18. Jones, S., Shanahan, H. P., Berman, H. M. & Thornton, J. M. (2003). Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. *Nucl. Acids Res.* **31**, 7189–7198.
19. Kiato, A. & Go, N. (1999). Investigating protein dynamics in collective coordinate space. *Curr. Opin. Struct. Biol.* **9**, 164–169.
20. Gilson, M. K. & Honig, B. (1988). Calculation of the total electrostatic energy of a macromolecular system: solvation energies, binding energies, and conformational analysis. *Proteins: Struct. Funct. Genet.* **4**, 7–18.
21. Rockey, W. M. & Elcock, A. H. (2002). Progress toward virtual screening for drug side effects. *Proteins: Struct. Funct. Genet.* **48**, 664–671.
22. Jeffrey, G. A. (1997). *An Introduction to Hydrogen Bonding*. Oxford University Press, New York.
23. Cramer, R. D., III., Patterson, D. E. & Bunce, J. D. (1988). Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **110**, 5959–5967.
24. Cappello, V., Tramontano, A. & Koch, U. (2002). Classification of proteins based on the properties of the ligand-binding site: the case of adenine-binding proteins. *Proteins: Struct. Funct. Genet.* **47**, 106–115.
25. Stawiski, E. W., Gregoret, L. M. & Mandel-Gutfreund, Y. (2003). Annotating nucleic acid-binding function based on protein structure. *J. Mol. Biol.* **326**, 1065–1079.
26. Thornton, J. M., Todd, A. E., Milburn, D., Borkakoti, N. & Orengo, C. A. (2000). From structure to function: approaches and limitations. *Nature Struct. Biol.* **7**, 991–994.
27. Yao, H., Kristensen, D. M., Mihalek, I., Sowa, M. E., Shaw, C., Kimmel, M. *et al*. (2003). An accurate, sensitive, and scalable method to identify functional sites in protein structures. *J. Mol. Biol.* **326**, 255–261.
28. Weiner, S. J., Kollman, P. A., Nguyen, D. T. & Case, D. A. (1986). An all atom force field for simulation of proteins and nucleic acids. *J. Comput. Chem.* **7**, 230–252.

Appendix: Relationship between simple electrostatic replacement energies and electrostatic component of binding free energy differences

Here, we examine the relationship between simple replacement energies ($\Delta E_{A\to G}$ and $\Delta E_{G\to A}$; equation (3)) and the corresponding electrostatic component of binding free energy differences ($\Delta\Delta G_{A\to G}$ and $\Delta\Delta G_{G\to A}$). The aim is not to derive exact values of $\Delta\Delta G$ from ΔE. Instead we want to show that the difference in ΔE between A/G-specific and non-specific sites (as depicted in Figure 8) implicitly suggests the existence of a similar difference in $\Delta\Delta G$ between the two sets. In addition to the fact that all energies referred to in this section are electrostatic-only, we implicitly assume that the binding site is identical for the cognate and the non-cognate ligands (no major conformational change of the protein, especially at the binding site) and ignore any entropic contribution to differential binding.

Free energies of binding G or A to a common binding site in a protein P, ΔG_G and ΔG_A, are given by:

$$\begin{aligned}\Delta G_G &= G_{G-P} - [G_G + G_P]\\ \Delta G_A &= G_{A-P} - [G_A + G_P]\end{aligned} \tag{A1}$$

where the subscripts P, A, G, A–P and G–P stand for ligand-free protein, adenine, guanine, adenine–protein complex and guanine–protein complex, respectively. Assuming that the cognate ligand is A, the free energy cost of replacing the cognate by the non-cognate ligand ($\Delta\Delta G_{A\to G}$) is given by:

$$\Delta\Delta G_{A\to G} = \Delta G_G - \Delta G_A \tag{A2}$$

$$\Delta\Delta G_{A\to G} = G_{G-P} - G_{A-P} + (G_A - G_G)$$

G_{G-P} and G_{A-P} in equation (A2) can be approximated to be composed of three components:

$$G_{G-P} \approx G^*_P + G^*_G + G^*_{G-P},\quad G_{A-P} \approx G^*_P + G^*_A + G^*_{A-P} \tag{A3}$$

where G^*_P is the energy of assembling protein point charges in a low-dielectric cavity (surrounded by water) defined by the shape of the protein–ligand complex (PL cavity), G^*_A/G^*_G is the energy of assembling ligand (A/G) point charges in the PL cavity, and G^*_{A-P}/G^*_{G-P} is the protein–ligand interaction energy (cognate: G^*_{A-P}, non-cognate: G^*_{G-p}) between protein and ligand point charges, embedded in the PL-cavity.

The protein–ligand interaction energy (G^*_{A-P}/G^*_{G-P}) is almost identical to simple energy (E_{AA}/E_{AG}; equation (2)) except for the extent of solvent screening. The interaction is more screened in E_{AA}/E_{AG} (where ESP of ligand-free protein, implying that ligand charges lie in a high dielectric medium resembling water, is employed) than in G^*_{A-P}/G^*_{G-P}. A factor λ (>1), accounting for the differential screening effect, can be used to relate the two as:

$$(G^*_{G-P} - G^*_{G-P}) = \lambda(E_{AG} - E_{AA}) = \lambda\Delta E_{A\to G} \tag{A4}$$

Equations (A2)–(A4) can be combined to yield (G^*_P cancels out):

$$\Delta\Delta G_{A\to G} \approx \lambda(\Delta E_{A\to G}) + [(G^*_G - G_G) + (G_A - G^*_A)] \tag{A5}$$

Equation (A5) is the central equation that relates simple replacement energies $\Delta E_{A\to G}$ and the corresponding binding free energy difference $\Delta\Delta G_{A\to G}$. We will now estimate some numerical values for λ and $[(G^*_G - G_G) + (G_A - G^*_A)]$.

The factor λ can be considered as the ratio of effective dielectric constants for screening of protein–ligand interaction in E_{AA}/E_{AG} and G^*_{A-P}/G^*_{G-P}. A numerical estimate for the value for λ is about 2–8, corresponding to an effective dielectric constant of 10–20 (for G^*_{A-P}/G^*_{G-P}) and 40–80 (for

E_{AA}/E_{AG}). The term $[(G^*_G - G_G) + (G_A - G^*_A)]$ corresponds to solvation changes in A (from protein-bound to free state) and G (from free state to protein-bound) upon ligand replacement, and was estimated from two sets of electrostatic calculations. In the first set, the Poisson–Boltzmann equation was solved for free ligands (A/G) with a high ($\sim$80) external dielectric constant (mimicking free ligand in water). In the other set, the external dielectric constant was set to be intermediary between protein and water (mimicking protein-bound ligand). This estimate is 1–2 kcal/mol when the effective dielectric constant for the protein-bound ligand case is in the range 6–20.

When considering A$\rightarrow$G replacement, since both $[(G^*_G - G_G) + (G_A - G^*_A)]$ and λ are positive in equation (A5), $\langle \Delta E^{A-sp}_{A\rightarrow G} \rangle > \langle \Delta E^{non-sp}_{A\rightarrow G} \rangle$ (obtained from Figure 8(a)) implies $\langle \Delta\Delta G^{A-sp}_{A\rightarrow G} \rangle > \langle \Delta\Delta G^{non-sp}_{A\rightarrow G} \rangle$. For the difference binding free energy for the reverse ligand-replacement, $\Delta\Delta G_{G\rightarrow A}$, however, the sign of the ligand solvation term in equation (A5) will also be reversed to -1–2 kcal/mol. Therefore, $\langle \Delta E^{G-sp}_{G\rightarrow A} \rangle > \langle \Delta E^{non-sp}_{G\rightarrow A} \rangle$ (obtained from Figure 8(b)) will imply $\langle \Delta\Delta G^{G-sp}_{G\rightarrow A} \rangle > \langle \Delta\Delta G^{non-sp}_{G\rightarrow A} \rangle$ only if $\lambda(\Delta E^{G-sp}_{G\rightarrow A} - \Delta E^{non-sp}_{G\rightarrow A})$ is greater than 1–2 kcal/mol. This holds true from simple estimates of λ (2–8) and $\langle \Delta E^{G-sp}_{G\rightarrow A} \rangle - \langle \Delta E^{non-sp}_{G\rightarrow A} \rangle$ ($\sim$1.3 kcal/mol; see Figure 8(b)).

Edited by B. Honig

(Received 18 March 2004; received in revised form 2 June 2004; accepted 8 July 2004)