Journal of Computer-Aided Molecular Design, **14:** 23–37, 2000.
KLUWER/ESCOM

Statistical relationships among docking scores for different protein binding sites

Ryan T. Koehler[1] & Hugo O. Villar*,[2]
Computational Chemistry Laboratory, Telik, Inc., 750 Gateway Blvd, South San Francisco, CA 94080, U.S.A.

Received 8 December 1998; Accepted 20 April 1999

Key words: affinity correlations, binding site similarities, DOCK, ligand promiscuity, molecular recognition

Summary

This report describes the existence of statistical relationships among scores computed with the DOCK program for a library of small molecules and a panel of protein binding sites. Multivariate relationships are observed in docking scores computed for a constant set of ligands in different binding sites of proteins that are dissimilar in structure and function. The structural basis for the correlations found among scores is analyzed in terms of size, shape and charge characteristics of the binding sites considered. Interestingly, these results parallel a growing body of evidence demonstrating the promiscuous behavior of small molecules in their interactions with macromolecules that could have an impact in future efforts in drug design.

Introduction

Docking methodologies seek to predict the ways in which compounds may fit into a macromolecular binding site. Typically these methods associate a score with each predicted intermolecular complex to quantify the goodness of the fit between the docking partners. Docking techniques have been at the core of the structure-based drug design paradigm for quite some time, and therefore numerous methodologies have been devised for this purpose [1, 2].

DOCK is one of the earliest, most widely used ligand-target matching programs, and the DOCK methodology has been exploited extensively [3–5]. In its original form, the matching algorithm checked only for steric complementarity between ligands and binding sites, yielding scores based on the extent of ligand-protein contacts [3, 6]. More recent implementations have addressed the importance of different terms in the scoring functions. In the most widely used version, a molecular force field may be employed to evaluate the energetics of predicted intermolecular complexes [7], or to optimize a given complex [8], or both. Other schemes, such as scores reflecting hydrophobic interactions [9], have also been investigated. While extensions allowing for ligand flexibility have been described [10], DOCK 3.5 treats both ligand and target molecules as rigid. (The ability to handle ligand flexibility has been added in DOCK 4.0: see http://www.cmpharm.ucsf.edu/kuntz/kuntz.html.)

Recent work has demonstrated how DOCK may be used for more than virtual screening of three dimensional compound databases. For instance, Briem and Kuntz [11] have utilized DOCK scores to classify small organic molecules based on their scores obtained for a set of unrelated proteins. Their approach aimed at simulating affinity fingerprint methodology [12] using DOCK as a computational surrogate in lieu of experimentally measured protein binding values. One of our goals is to complement this work from the perspective of protein binding site structures.

In this report, the existence of multicollinearities in force field scores generated by DOCK for a library of small molecules and seven of the protein binding sites employed by Briem and Kuntz is shown. Multicollinearities indicate the presence of relationships in the data such that one variable may be represented as a linear combination of others. In other words, they allow that scores for one protein binding site

*To whom correspondence should be addressed. E-mail: [1]koehler@telik.com and [2]hugo@telik.com.

Table 1. Proteins employed to generate ligand docking scores[a]

PDB code	Protein / complexed ligand(s)	Class	Resolution (Å)
1acl [14]	Acetylcholinesterase / decamethonium	Esterase	2.8
1dwc [15]	α-Thrombin / modified hirudin and argobatan	Serine protease	3.0
1eed [16]	Endothiopepsin / PD125754	Aspartate protease	2.0
1pop [17]	Papain / leupeptin	Thiol protease	2.1
2tsc [18]	Thymidylate synthase / dUMP and an anti-folate	Methyltransferase	1.97
3cla [19]	Chloramphenicol acetyltransferase / chloramphenicol	Acetyltransferase	1.75
3dfr [20]	Dihydrofolate reductase / NADPH and methotrexate	Oxidoreductase	1.7

[a]Adopted from Table 1 of Briem and Kuntz [11].

may be predicted from scores for the same ligands docked into different binding sites. Such statistical relationships in docking scores may become useful tools in the design of libraries for screening if they are manifestations of an underlying commonality in binding sites which may not be evident when protein structure or function are compared. Analogous statistical relationships have been observed in experimentally determined protein affinity fingerprint data [12]. In order to determine how multicollinearities might be exploited for structural design, it is necessary to gain a clear understanding of their nature and investigate if they are a manifestation of cryptic similarities among protein binding sites. DOCK scores are not directly analogous to binding affinities and should not be confused with them, but to a first approximation, a simple DOCK-based model is a reasonable system for structural investigations of ligand binding with the advantage that predicted complexes may be fully scrutinized. To this end, we first analyze pairwise correlations in the DOCK score data sets used, then show the existence of statistically significant multicollinearities for the same set of proteins. Next, correlates between simple properties of the small molecules and proteins and their associated scores are scrutinized to reconcile observed statistical trends with structural parameters. Finally, we compare the different protein binding sites seeking to understand the origin of these multivariate correlations in the data.

Methods

The DOCK 3.5 (UCSF, San Francisco, CA) suite of programs was used for docking and directly related processing steps. Insight II (version 97.0; Molecular Simulations Inc., San Diego, CA) was used for visualization and generation of dot surfaces [13]; 1.4 Å probe and 15 dots/Å^2. Regression analysis and other statistics were computed with S-plus for Windows (version 4.0, MathSoft Inc., Cambridge, MA).

X-ray structures of the proteins employed by Briem and Kuntz [11] in the context of DOCK-generated fingerprints (Table 1) were obtained from the PDB. Binding sites were selected for docking using standard procedures. Specifically, sites were identified via X-ray ligands, ligand and water atoms were removed, hydrogen atoms were added to proteins (DOCK suite program *addprh*), then spheres representing the negative image of the binding sites were computed (*sphgen*; maximum radius 5 Å). Spatially contiguous sphere clusters were then generated (*cluster*; 10% maximum sphere overlap) and graphically inspected to identify the clusters most similarly situated to X-ray ligands. Scoring grids encompassing the sphere clusters (5 Å margin, 0.5 Å resolution) were computed with AMBER force field [21] parameters (as implemented in DOCK; *vdw.parms.amb.mindock*) with a distant dependent dielectric of $\delta = 4r$ and cutoff of 10 Å.

For docking, parameter settings as suggested in the DOCK manual were used (url: http://www.cmpharm.ucsf.edu/kuntz/dock.html). Specifically, minimum matching nodes = 4, distance tolerance = 1.0 Å, maximum 'bumps' = 2, ligand bin size = 0.4 Å, receptor bin size = 0.8 Å, ligand bin overlap = 0.1 Å, receptor bin overlap = 0.2 Å. Trilinear interpolation and minimization were enabled with default settings. Enabling minimization directs DOCK to energy minimize ligand-sphere matchings in the context of the protein before a final score (and ligand orientation) is determined [6]. Test runs revealed that minimization leads to the identification of more favorably scoring binding modes than is otherwise the case. It also allows atoms to move away from their matching part-

Table 2. Statistical summary of DOCK scores and score components[a]

Protein	1acl	1dwc	1eed	1pop	2tsc	3cla	3dfr
Full scores (kcal mol^{-1})							
mean	−28.04	−24.64	−27.30	−21.62	−26.72	−18.06	−21.69
s.d.	5.25	4.10	8.41	3.48	5.79	3.03	5.24
min.	−43.26	−42.64	−55.01	−32.30	−57.50	−28.70	−39.56
max.	−8.12	−8.40	−7.15	−8.17	−7.49	−7.22	−6.08
van der Waals score components (kcal mol^{-1})							
mean	−24.50	−20.70	−19.91	−19.14	−21.62	−17.06	−19.53
s.d.	5.36	4.11	4.34	3.74	4.78	3.13	4.44
min.	−40.48	−33.20	−30.45	−28.91	−35.27	−27.82	−33.28
max.	−4.86	−6.35	−5.97	−7.08	−3.27	−6.44	−7.36
Electrostatic score components (kcal mol^{-1})							
mean	−3.53	−3.93	−7.37	−2.47	−5.09	−0.99	−2.15
s.d.	3.06	3.31	7.33	2.75	5.20	1.44	3.79
min.	−16.13	−16.37	−31.02	−14.55	−30.28	−7.04	−20.89
max.	4.67	1.44	6.93	0.88	1.70	2.66	9.48
Ratio of van der Waals to full score components							
mean	0.87	0.85	0.77	0.89	0.82	0.95	0.92
s.d.	0.11	0.13	0.20	0.12	0.15	0.08	0.18
min.	0.41	0.36	0.31	0.35	0.17	0.56	0.36
max.	1.21	1.06	1.40	1.06	1.08	1.20	2.53

[a] N = 692; s.d., standard deviation; min., minimum score; max., maximum score.

ners, reducing artifacts associated with the somewhat arbitrary locations of binding site spheres.

For our model ligand library, a structurally diverse subset of 939 compounds from the cmC3D database of bioactive compounds was chosen (version 94.1; MDL Information Systems, Inc., San Leandro, CA). We limited the number of compounds from any particular pharmacological or structural class to avoid biasing results in the subsequent analysis. Pharmacological class records in the database guided selection. Structural 3D models were extracted directly from the database then prepared for docking with SYBYL (version 6.1; Tripos Inc., St. Louis, MO) using scripts distributed with DOCK. Atoms in acidic and basic groups were explicitly set to be ionized assuming pH 7 (via SYBYL atom type), then hydrogens were added and atomic partial charges computed [22] within SYBYL.

DOCK output was processed (*scoreopt*) to obtain the total energy score, the electrostatic and van der Waals score components, and the number of any ligand atoms falling outside of the scoring grid for each ligand-protein pair. Prior to statistical analysis, some of the compounds were removed from consideration. Ligands that DOCK was unable to match, or that matched with unfavorable (positive) scores, to one or more of the panel proteins were dropped. Also, ligands with more than 10% of constituent atoms falling outside of any protein scoring grid were dropped. Atoms outside of a grid do not contribute to docking scores, and visual inspection revealed that such atoms tend to protrude from the binding sites into solvent regions; ignoring atoms distant from the protein surface generally has a minor consequence on ligand score. We avoided using expanded grids that encompass larger portions of the target proteins, as this permitted some ligand models to dock into different pockets than the chief sites we sought to compare. The 10% threshold represents a compromise, as eliminating all ligands with any atom falling outside of any protein grid (428 of 939) would dramatically reduce the size and heterogeneity of our model library. In total, 692 favorably docking ligand models that encompass the same struc-

Table 3. Correlations among DOCK scores, score components, and ligand attributes[a]

Protein	1acl	1dwc	1eed	1pop	2tsc	3cla	3dfr
Full scores							
1dwc	0.606						
1eed	0.625	0.617					
1pop	0.653	0.692	0.422				
2tsc	0.505	0.605	0.280	0.765			
3cla	0.427	0.535	0.161	0.725	0.594		
3dfr	0.443	0.331	0.037	0.461	0.526	0.445	
Atom count[b]	−0.344	−0.436	−0.567	−0.514	−0.353	−0.360	0.102
Net charge[b]	−0.299	−0.258	−0.810	0.068	0.236	0.277	0.334
van der Waals score components							
1dwc	0.572						
1eed	0.716	0.675					
1pop	0.731	0.719	0.806				
2tsc	0.706	0.651	0.781	0.781			
3cla	0.550	0.599	0.683	0.658	0.625		
3dfr	0.684	0.531	0.680	0.642	0.540	0.515	
Atom count[b]	−0.370	−0.396	−0.609	−0.576	−0.596	−0.412	−0.227
Net charge[b]	−0.034	−0.179	−0.071	−0.190	−0.059	0.036	−0.040
Electrostatic score components							
1dwc	0.577						
1eed	0.659	0.444					
1pop	0.370	0.599	0.003				
2tsc	0.398	0.724	0.073	0.708			
3cla	0.047	0.316	−0.283	0.479	0.519		
3dfr	0.237	0.459	−0.210	0.645	0.680	0.526	
Atom count[b]	0.060	−0.049	−0.288	0.133	0.155	0.132	0.409
Net charge[b]	−0.451	−0.097	−0.886	0.344	0.317	0.503	0.509

[a]Values are correlation coefficients, r. N = 692.
[b]Atom count and net charge refer to ligand characteristics as represented in the DOCK database.

tural and pharmacological diversity as the originally chosen set were retained to investigate multicollinearities. These model ligands range in size from 6 to 74 atoms (37.12 mean, 12.39 standard deviation) and have associated (formal) charges ranging from −3 to +2 (+0.33 mean, 0.76 standard deviation). All data (DOCK parameters, ligand coordinates, scores) may be obtained from the authors upon request.

Results and discussion

Analysis of DOCK scores

Table 1 describes the protein binding sites explored with DOCK and a summary of the scores associated with these proteins is given in Table 2. For each protein, statistics describing the scores and the separated steric (van der Waals, vdW) and electrostatic components of the scores are listed. The relative importance of the two score components is characterized by the ratio of the vdW components to the full scores; the larger the ratio, the more significant is the steric component. A ratio of 1.0 indicates no electrostatic contribution to the score, while ratios greater than 1.0 result when a somewhat unfavorable (positive) electrostatic com-

Table 4. Parameters of stepwise linear regression models derived from full DOCK scores[a]

	1acl	1dwc	1eed	1pop	2tsc	3cla	3dfr
1acl		–	0.858	0.166	−0.063	–	0.450
		–	*16.603*	*10.348*	*−1.504*	–	*10.484*
1dwc	–		1.017	0.128	0.278	0.157	−0.222
	–		*15.015*	*5.773*	*5.330*	*5.325*	*−9.373*
1eed	0.298	0.214		–	−0.045	−0.088	–
	18.412	*16.979*		–	*−1.839*	*−7.507*	–
1pop	0.547	0.250	–		0.989	0.558	–
	9.404	*4.690*	–		*16.442*	*17.471*	–
2tsc	−0.052	0.148	–	0.222		–	0.300
	−1.538	5.892	–	13.923		–	*8.536*
3cla	–	0.251	−0.589	0.387	–		0.195
	–	*5.449*	*−6.968*	*14.131*	–		*3.015*
3dfr	0.289	–	−0.434	−0.029	0.237	0.051	
	10.416	–	*−9.496*	*−1.969*	*7.488*	*3.047*	
Intercept	−3.202	−4.894	1.750	−1.522	3.679	−3.422	−3.608
	−4.029	*−7.756*	*1.201*	*−3.367*	*3.981*	*−6.626*	*−3.451*
r	0.796	0.802	0.772	0.880	0.799	0.760	0.641
F	298.0	310.1	254.0	472.3	242.8	234.8	119.8
s.d.	3.18	2.46	5.36	1.66	3.49	1.98	4.04

[a]Parameters associated with each model are arranged by column. Coefficients for each panel protein are listed from top to bottom, with the associated *t* value directly below these in *italics*. Coefficients (and *t* values) corresponding to protein score variables dropped during the stepwise derivation of the regression models are denoted by '–' entries. At the bottom, intercepts, intercept *t* values, correlation coefficients (*r*), F statistics, and the standard deviation of the regression (s.d.) are listed for each model.

ponent is accompanied by a favorable (negative) vdW component.

vdW terms generally dominate the scores with electrostatic components constituting, on average, only ca. 5% (3cla) to 23% (1eed) of the total scores. For individual ligand-protein complex scores the relative magnitudes of score components can deviate significantly from average values as evidenced by the minimum and maximum ratios at the bottom of Table 2.

Pairwise correlations

Pairwise correlation coefficients between the docking scores computed for each protein are given in Table 3, together with their correlation to atom counts and net charges of models in the ligand library. Pairwise correlation between the separated score components and ligand characteristics are also reported. Atom count is used as a general indicator of ligand size and net charge is used as a gross indicator of ligand polarity, though some compounds may have zero net charge even though multiple ionized polar groups are present (e.g. zwitterions).

Without exception positive correlations are found between the score distributions. Atom counts tend to be negatively correlated with scores, except in the case of 3dfr. This means that, on average, larger structures are associated with more favorable (negative) scores, as the total number of possible protein contacts increases. This behavior, in light of the dominant role of steric relative to electrostatic components, likely explains the positive correlations observed between scores. For most proteins, correlations between full scores and ligand net charge are modest, except for 1eed, which has the most polar binding site (see below). The correlations between vdW score terms associated with different proteins follow analogous trends to those of full scores. Correlations between electrostatic terms are more disparate across proteins.

A number of correlations reported in Table 3 are significantly larger than analogous values reported by Briem and Kuntz [11]. Several factors likely account for these differences. First, spheres employed in the current study were constrained to form a spatially contiguous clusters so that all ligands sampled one clearly bounded binding site of each protein. Spheres employed for the previous study had no such constraint

Table 5. Parameters of stepwise linear regression models derived from separated DOCK scores[a]

	1acl	1dwc	1eed	1pop	2tsc	3cla	3dfr
1acl.ele		–	1.288	0.116	−0.160	–	−0.414
		–	*18.171*	*4.333*	*−3.347*	–	*5.309*
1acl.vdw		–	0.252	0.108	–	–	0.480
		–	*4.816*	*6.081*	–	–	*11.195*
1dwc.ele	–		1.056	0.154	0.594	0.121	–
	–		*14.486*	*4.695*	*11.014*	*2.895*	–
1dwc.vdw	–		0.532	0.103	0.081	0.174	–
	–		*8.232*	*4.357*	*1.794*	*6.196*	–
1eed.ele	0.287	0.234		–	–	−0.124	−0.220
	16.366	*15.288*		–	–	*−10.220*	*−7.061*
1eed.vdw	0.112	0.118		0.125	0.302	0.157	0.186
	1.968	*2.781*		4.708	6.180	5.109	2.997
1pop.ele	0.393	0.192	–		0.813	0.345	–
	5.879	*2.998*	–		*11.263*	*8.230*	–
1pop.vdw	0.531	0.370	0.373		0.666	0.338	–
	7.141	*6.036*	*4.156*		*9.789*	*8.462*	–
2tsc.ele	–	0.151	–	0.172		0.049	0.339
	–	*4.582*	–	*8.272*		*2.163*	*8.530*
2tsc.vdw	0.159	0.108	0.203	0.269		–	−0.119
	3.521	*2.770*	*3.444*	*11.276*		–	*−2.118*
3cla.ele	–	0.368	−1.248	0.359	0.458		0.385
	–	*3.783*	*−8.699*	*6.448*	*4.271*		2.635
3cla.vdw	−0.081	0.304	−0.377	0.265	0.135		0.127
	−1.462	*5.995*	*−4.812*	*8.581*	*2.266*		*1.713*
3dfr.ele	0.251	0.056	−0.668	–	0.289	–	
	5.546	*1.426*	*−11.170*	–	*6.764*	–	
3dfr.vdw	0.361	–	0.150	–	–	0.038	
	9.552	–	*2.853*	–	–	*1.752*	
Intercept	−2.911	−4.239	3.930	−1.748	0.878	−3.445	−4.677
	−3.497	*−6.484*	*3.296*	*−3.944*	*1.053*	*−7.284*	*−4.620*
r	0.807	0.809	0.870	0.892	0.867	0.811	0.694
F	160	143	212	296	230	164	79.3
s.d.	3.12	2.43	4.18	1.58	2.90	1.78	3.80

[a] This table is arranged as Table 5, except here panel protein scores are broken into separated van der Waals (vdW) and electrostatic (ele) components.

(Dr. H. Briem kindly supplied sphere coordinates allowing visual comparison), and the relatively large RMSD values reported in the earlier study ([11] Table 6) reflect the low spatial cohesion obtained with the docked ligands. Second, we report correlations among minimized scores while Briem and Kuntz do not. Tests show that employing DOCK with minimization consistently increases correlations among scores. Finally, use of different ligand sets would be expected to yield discrepancies in scores, and possible correlations among these.

Score multicollinearities

Statistical models for each of the seven panel proteins were derived from stepwise multivariate regression [23] using the scores of the remaining six panel members as independent variables. This yields an equation to predict the scores of target protein, *T*, given the scores for the six remaining proteins, P_i, as input:

$$T_{\text{calc}} = c_1 P_1 + c_2 P_2 \cdots c_6 P_6 + k.$$

Here T_{calc} is the calculated score for protein *T*, parameters c_1, $c_2 \cdots c_6$ are coefficients for scores of the other six proteins, and k is a constant. This process

Table 6. Characteristics of the differently delineated protein binding sites

Characteristic	1acl	1dwc	1eed	1pop	2tsc	3cla	3dfr
Qualitative descriptions[a]							
General topography	pocket	complex	cleft	valley	pocket	valley	pocket
Polarity	weak	medium	high	weak	medium	weak	weak
Polar residues[b]	Y^{118}, S^{119}, Y^{331}	H^{72}, D^{228}, E^{231},Y^{331}	D^{33}, D^{35}, D^{80} D^{117}, D^{213}	V^{133}, A^{160}	R^{21}, K^{48}, R^{126}, D^{169}	H^{144}, N^{146}, Y^{168}	S^{48}, T^{45}, R^{57}, A^{97}
Nonpolar residues[b]	W^{81}, F^{327}, F^{328}, Y^{331}	A^{229}, E^{231}, V^{254}	Y^{78}, I^{120}, L^{123}	G^{66}, H^{159}, D^{158}	L^{143}, L^{172}, F^{176}, V^{226}	V^{162}, F^{165}, Y^{168}, I^{172}	L^{19}, F^{30}, F^{40}, L^{54}
Binding sites delineated at 20% contact frequency							
Surface area [c]	442	300	343	228	416	311	354
Constricted[d]	68.3	64.7	66.6	61.2	69.3	66.3	65.6
Semi-constricted[d]	29.0	28.2	29.3	26.2	27.8	20.5	32.1
Unconstricted[d]	2.7	7.1	4.1	12.6	2.9	13.1	2.3
Average E.P.[e]	−0.008	−0.014	−0.041	0.000	0.004	0.005	0.004
E.P. variability[e]	0.023	0.024	0.032	0.017	0.029	0.010	0.018
Contacted atoms[f]	100	68	73	45	90	62	67
Polar atoms [g]	31.0	38.2	34.2	28.9	37.8	32.3	26.9
Semi-polar atoms[g]	29.0	48.5	34.2	48.9	31.1	32.3	31.3
Non-polar atoms[g]	40.0	13.2	31.5	22.2	31.1	35.5	41.8
Binding sites delineated at 40% contact frequency							
Surface area[c]	170	168	180	131	143	120	164
Constricted[d]	52.9	64.1	60.9	57.9	51.4	58.9	60.6
Semi-constricted[d]	45.3	29.8	36.0	34.0	43.9	21.3	38.5
Unconstricted[d]	1.8	6.1	3.1	8.1	4.8	19.8	0.9
Average E.P.[e]	−0.003	−0.015	−0.049	−0.001	−0.001	0.004	0.001
E.P. variability[e]	0.023	0.021	0.031	0.015	0.021	0.005	0.012
Contacted atoms[f]	48	44	40	28	40	23	38
Polar atoms[g]	33.3	36.4	40.0	25.0	25.0	34.8	21.1
Semi-polar atoms[g]	20.8	52.3	27.5	53.6	30.0	17.4	26.3
Non-polar atoms[g]	45.8	11.4	32.5	21.4	45.0	47.8	52.6
Binding sites delineated at 60% contact frequency							
Surface area [c]	28	75	51	43	21	53	63
Constricted[d]	25.1	55.2	49.6	42.9	26.0	46.8	58.1
Semi-constricted[d]	74.9	38.9	48.3	48.8	63.8	22.0	41.9
Unconstricted[d]	0.0	5.9	2.1	8.3	10.3	31.1	0.0
Average E.P.[e]	−0.003	−0.011	−0.060	−0.003	0.001	0.004	0.001
E.P. variability[e]	0.020	0.015	0.025	0.018	0.015	0.004	0.008
Contacted atoms[f]	16	26	14	12	13	18	15
Polar atoms[g]	18.8	42.3	57.1	41.7	15.4	33.3	20.0

was repeated, treating each protein as a target in turn and using the scores of the remaining proteins to derive a statistical model for the protein. Because of the stepwise regression procedure, where models are simplified by removal of variables with low statistical significance, not all models incorporate terms for all six possible panel proteins. Similarly, scores that are highly correlated in a simple pair-wise manner with

Table 6 (continued)

Characteristic	1acl	1dwc	1eed	1pop	2tsc	3cla	3dfr
Semi-polar atoms[g]	6.3	53.8	14.3	58.3	38.5	16.7	6.7
Non-polar atoms[g]	75.0	3.8	28.6	0.0	46.2	50.0	73.3

[a]Qualitative descriptions are based on visual inspection of binding sites delineated with contact frequencies of ca. 10% or more; see text for explanation of contact frequency.
[b]Residues with frequently contacted atoms classified as polar or nonpolar (i.e. hetero or carbon) are denoted by one letter code and listed in order of sequence numbering in the source PDB file. A single residue may display both polar and nonpolar atoms in a given site.
[c]$Å^2$, computed by counting surface dots ($15/Å^2$).
[d]Percentage of total delineated surface area classified with different degrees of steric constriction as determined by MaxAccess (MA, see text) at surface dot positions: Constricted, MA < 2.0 Å; Semi-constricted, MA 2.0–4.0 Å; Unconstricted, MA > 4.0 Å.
[e]E.P., electrostatic potential evaluated at surface dots with variability determined as the standard deviation.
[f]Number of protein atoms contacted at the indicated frequency thresholds; hydrogens attached to carbon are ignored.
[g]Percentage of contacted protein atoms classified as polar (hetero atoms and attached hydrogens), nonpolar (carbons not attached to any hetero atoms), and semi-polar (remaining atoms).

the target protein scores (Table 3) may have low significance because their information is accounted for by other model terms.

Table 4 summarizes all models. The correlation coefficients and large F statistic values indicate that significant correlation is present in all models. Further, the small relative magnitudes of the standard deviations of regression indicate that most scores are accurately fitted by the derived regression equations. Below each coefficient in Table 4 is a *t* value (ratio of the coefficient to the standard error of estimate for the coefficient) which indicates the significance of each coefficient in the model.

Decomposing the scores into the vdW and electrostatic components results in improved statistical models (Table 5). More interestingly, though, these new models frequently scale the separate steric and electrostatic terms of single proteins differently to optimally represent target protein scores. This is seen by comparing the magnitudes of regression equation coefficients associated with separated components for a given protein, and it suggests that some panel proteins primarily encode either steric or electrostatic determinants for target protein sites. For example, the 1acl model emphasizes the electrostatic term from 1eed and the vdW terms of both 1pop and 3dfr, while the electrostatic components of 2tsc and 3cla are dropped during the stepwise regression procedure.

A similar analysis can be performed for the 1eed model where some of the separated score components of panel proteins have associated coefficients differing in magnitude by several fold. This is consistent with the marked polar character (anionic, see below) of the 1eed site. For this protein in particular, it is evident that when electrostatics and vdW score components are separated, panel proteins with a grossly similar electrostatic response (e.g., 1eed, 1acl, and 1dwc all favor cationic ligands) contribute most to the electrostatic components of the 1eed model. The same is true for the vdW components, with panel proteins most similar or different from 1eed in terms of steric restrictions (e.g. 1dwc and 3dfr) factoring prominently in the model.

All models were cross-validated to ascertain robustness. To do this, the scores were randomly divided into two halves, new models were derived using scores for each half, then each model was used to predict the *other* half of scores not used in its derivation. In no case did the cross-validated correlation coefficients, i.e. predicted scores vs. observed scores, differ by more than 0.03 from the *r* values listed in Table 4, indicating stable models.

Analysis of the binding sites

Because of the statistical nature of the models, binding site features contacted frequently by bound ligands should encode a majority of the information necessary to understand the origin of the observed statistical relationships among different proteins. Thus, we will focus our comparison of the different proteins around the binding site areas contacted by a majority of the ligands.

We compare binding sites graphically and via atom-based or surface properties. In all cases, it is important to unambiguously determine exactly which atoms or surface regions are relevant for the comparison of properties. Here, binding site delineation follows directly from the positions of docked ligands. To choose protein atoms for representing the binding sites, the number of ligands positioning any atom within a specified distance of each protein atom is tallied. The number of contacts are then normalized by the number of ligands, generating a contact frequency for each protein atom. For example, a value of 0.5 indicates that half of the docked ligands position at least one atom in contact with the associated protein atom at a specified cut-off distance. For choosing surface regions, a slightly different, grid-based procedure is used to qualify surface dots. Contact frequencies are first tabulated and normalized at points on grids (identical to those used by DOCK), then contact frequencies for individual surface dots are computed from grid values via trilinear interpolation (as *per* DOCK). In this way two parameters determine which subsets of atoms or protein surface should be used to represent binding sites: a cut-off for the contact distance and a threshold for contact frequency.

Instead of arbitrarily selecting a single cut-off distance and contact frequency to compare binding sites, we varied both parameters over physically meaningful ranges and compared the resulting sets of atoms and surface points. For atom sets, cut-off distances ranging from 3.0 to 5.0 Å in steps of 0.5 Å were investigated with contact frequencies ranging from 5% to 70% in steps of 5%. This distance range approximates the sum of two vdW radii plus a margin. Atom sets defined with a constant contact frequency but different cut-off distances are very similar to each other within the ranges studied. Therefore, a single cut-off distance of 4.0 Å was chosen for discussion. Conversely, the contact frequency threshold strongly influences atom subsets ascribed to binding sites, so several different threshold values are considered for each protein site. An analysis of surface set sensitivity to cut-off distance and contact frequency was also undertaken. Again, we examined contact frequencies ranging from 5% to 70% in steps of 5%, but this time cut-off distances ranging from 1.5 Å and 3.0 Å in steps of 0.5 Å were used. This distance range corresponds to a bloated vdW radius, as protein surface this close to ligand atoms would normally be considered in contact with such atoms. Results were similar to those obtained with protein atom sets, and we likewise illustrate results obtained with a single cut-off distance of 2.5 Å and several different contact frequency thresholds.

Panel protein sites are illustrated in Figure 1 where surfaces are color-coded to indicate the degree of ligand contact frequency associated with different surface regions. Regions colored red are those most frequently contacted by the docked ligands, followed by yellow, green, then blue.

Computed characteristics of the binding sites delineated with 20%, 40%, and 60% contact frequency thresholds are listed in Table 6, together with qualitative descriptions and, for reference, lists of prominent residues contributing frequently contacted atoms to each of the sites. MaxAccess [24] parameters that describe the maximal size of a spherical probe that can contact a given point on a molecular surface are also reported. For each contact frequency, MA values were computed and classified as constricted (MA < 2 Å; includes reentrant surface area), semi-constricted (MA 2 to 4 Å), or unconstricted (MA > 4 Å). The fractions of each delineated binding site surface falling into each of these classes are listed in Table 6. The larger the unconstricted fraction, the smaller the steric congestion of the site.

Shapes of the ligand-delineated binding sites range from nearly fully enclosed pockets (1acl, 2tsc), to relatively deep, roughly linear (1eed, 3dfr) or rugous (1dwc) clefts, to shallow, substantially exposed (1pop, 3cla) valleys on the protein surfaces. These site attributes are reflected in the fractions of surface areas classified by MA as constricted, semi-constricted, or unconstricted. The small amounts of surface area delineated in some sites, particularly at 60% contact frequency, lower the confidence of associated property statistics. Binding site shape and size is also reflected to some extent by the total amount of surface area and the numbers of contacted atoms delineated in each protein. For example, the largely flat and exposed 1pop site has the smallest surface area and the fewest atoms when a 20% contact frequency threshold is applied, while the deep, enclosed pocket of 1acl has the largest contacted surface area and the most contacted protein atoms. This stems from the fact that only a limited amount of protein surface is available for ligand contact in a flat site like 1pop, regardless of how a ligand is oriented relative to the protein. For the 1acl site the opposite is usually true, with all but the smallest ligands in contact with multiple faces of the enclosed binding pocket. Along these same lines, surface areas and atom counts agree with the classification of 2tsc as

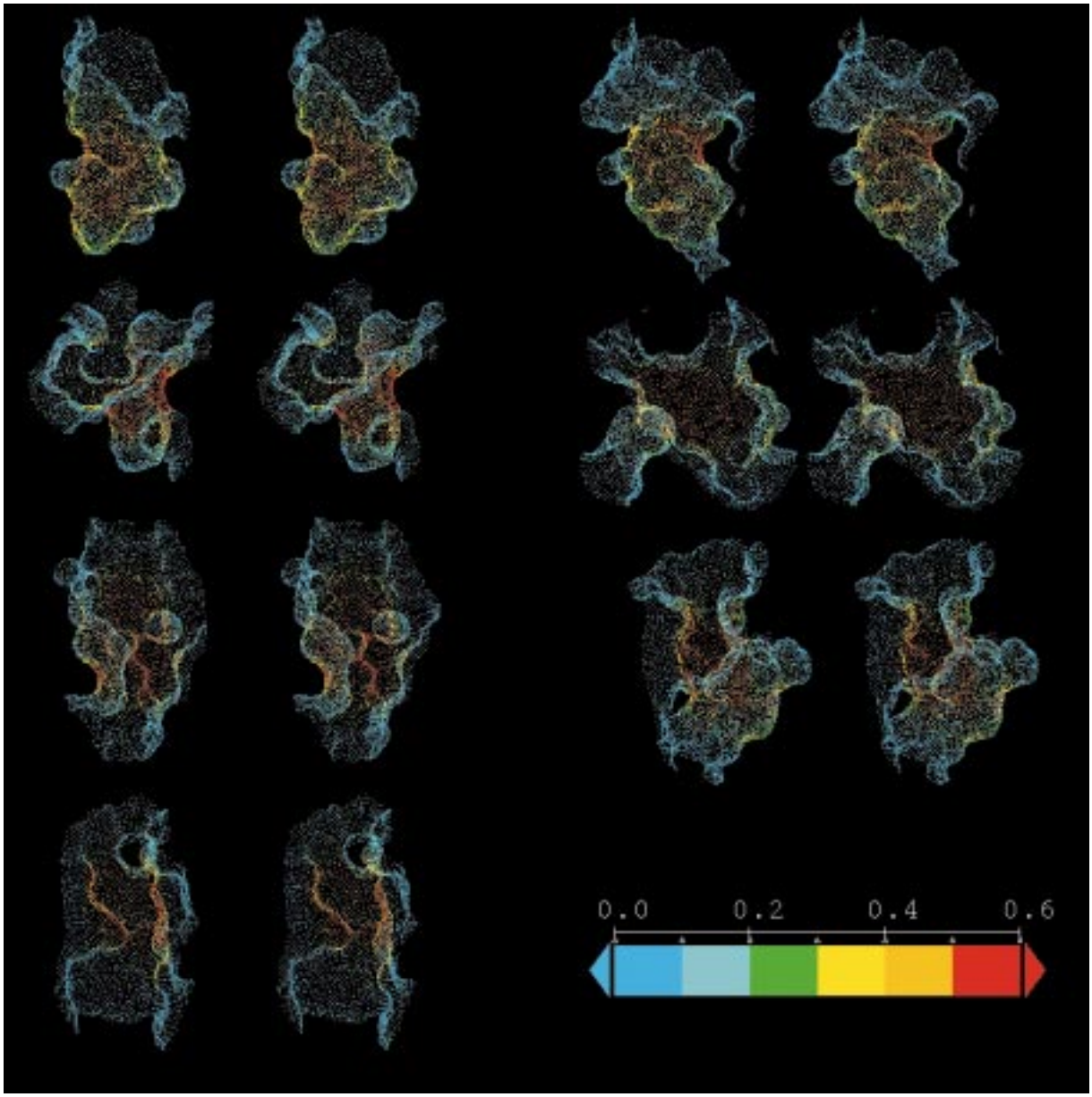

Figure 1 Delineated panel protein binding site surfaces colored to indicate ligand contact frequency. For each protein, a stereo pair is provided with color as indicated in the scale bar. Left side, top to bottom: 1acl, 1dwc, 1eed, 1pop; Right side, top to bottom: 2tsc, 3cla, 3dfr. All sites are shown at the same scale (the color bar is ca. 30 Å across) with the largest openings approximately directed upward. Only surface with contact frequency ≥ 0.05 is shown.

largely enclosed, 3cla as largely exposed, and 1dwc, 1eed and 3dfr as intermediate.

Electrostatic potentials (EP) for each protein, computed on grids identical to those used by DOCK for scoring, are displayed on the protein surfaces in Figure 2, where red denotes surface regions with negative potential and blue denotes regions with positive potential. Statistics describing EP values evaluated on the surface dots for each of the multiply delineated sites are listed in Table 6. Fractions of contacted protein atoms classified as polar, semi-polar, or nonpolar are also listed in Table 6. Polar atoms are defined here as heteroatoms and hydrogens attached to them, nonpolar atoms are defined as carbon atoms not attached

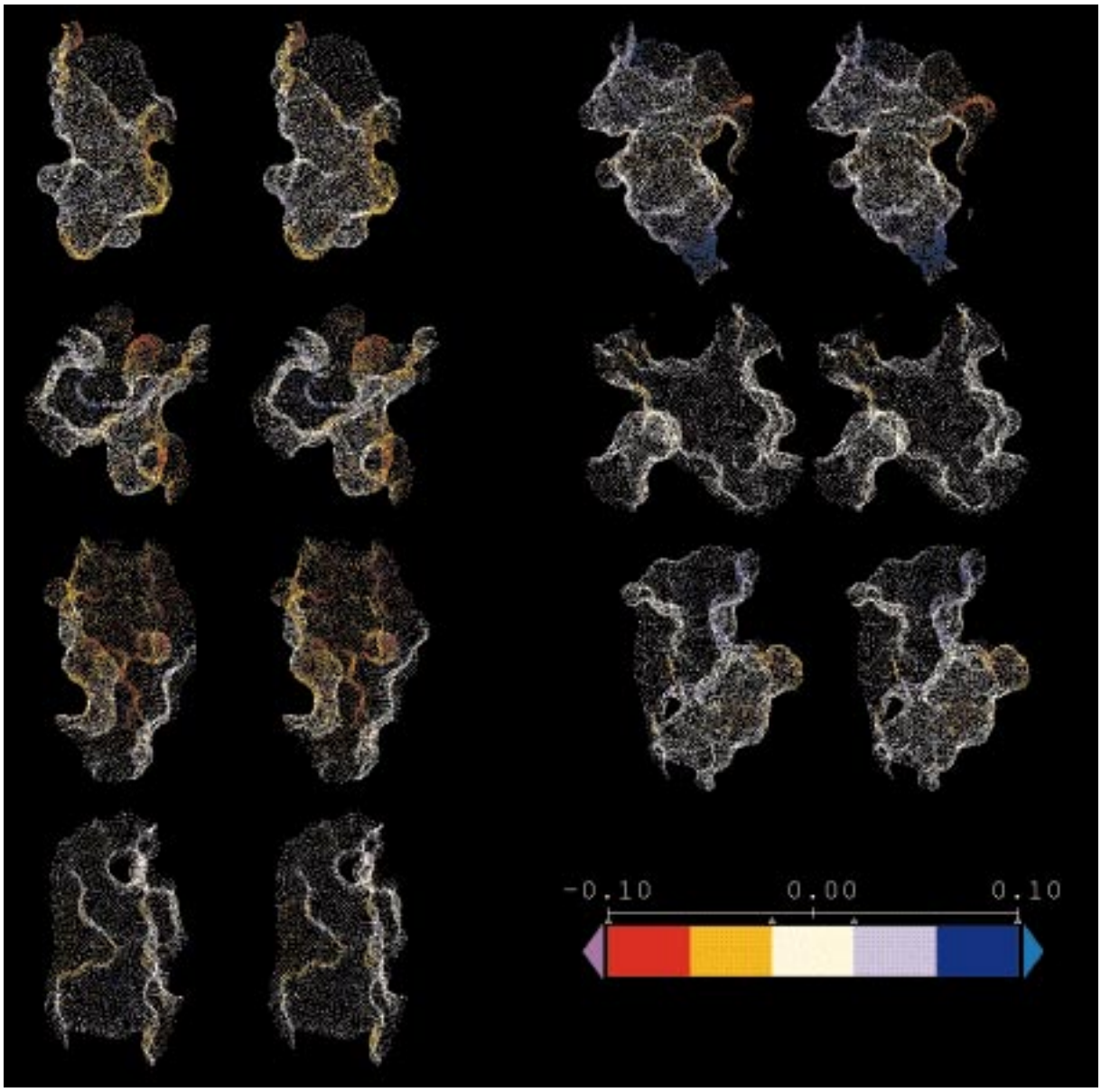

Figure 2 Binding site electrostatic potential. Except for the color scheme indicated by the scale bar, this figure is identical to Figure 1.

to hetero atoms, and semi-polar atoms are defined as those not classified as polar or nonpolar.

Protein 1eed is clearly the most polar of any panel member, with a highly anionic site, followed by the 1dwc site (also anionic) and 2tsc (mixed anionic and cationic). The site of 1acl exhibits a weak excess of anionic character, while the sites of 1pop and 3dfr exhibit weak excesses of cationic character. These characteristics are visually evident in Figure 2 and are consistent with EP statistics listed in Table 6. They are also somewhat reflected in the fractions of frequently contacted protein atoms classified as polar, nonpolar, or semi-polar. 1dwc and 1eed have relatively large fractions of polar and semi-polar atoms, while 1acl, 2tsc, 3cla, 3dfr have smaller fractions of atoms in these classes; the 1pop site includes multiple peptide backbone atoms, which are classified as polar or semi-polar in our scheme.

Table 7. Scores for compounds used to exemplify the 1acl statistical model[a]

	veralipride			ropinirole		
	full	ele.	vdW	full	ele.	vdW
1eed	−42.085	−13.513	−28.553	−33.934	−13.25	−20.685
1pop	−24.156	−1.364	−22.753	−23.354	−3.417	−19.939
2tsc	−27.933	−4.246	−23.674	−27.949	−3.298	−24.653
3dfr	−14.195	4.853	−19.017	−22.394	−0.07	−22.362
1acl actual	−10.461	−5.472	−4.864	−31.032	−5.331	−25.699
1acl predicted	−31.598			−31.102		

[a]Scores and separated electrostatic (ele) and vdW components (in kcal mol^{-1}) are listed for proteins factoring into the 1acl statistical model as well as the actual and predicted 1acl scores.

Rationalizing the multicollinearities

Statistical models for predicting scores of 1acl and 1pop will be used to exemplify our findings, because these show the strongest multicollinearities and differ in the panel proteins they draw upon. For both regression models, target proteins are compared and contrasted to panel proteins included in their respective equations, keeping in mind the different magnitudes of panel protein coefficients (Tables 4, 5).

Protein 1acl has an enclosed pocket-like site and exhibits a slight preference for positively charged ligands. The model for this protein draws most heavily upon 1pop followed by 1eed and 3dfr, while 2tsc factors only slightly (Table 4). Protein sites 1acl and 1pop have a similar fairly enclosed shape. Both sites are nonpolar and the weak correlation between electrostatic score components of 1acl and 1pop is compensated by a strong correlation between vdW score components. Despite fairly high correlation between 1acl and 1eed vdW score components, the electrostatic score term of 1eed in the 1acl model derived from separated components is more than 2.5-fold larger than the vdW score and considerably more statistically significant (i.e. ratio of t values is ∼8:1; Table 5), suggesting the 1eed score primarily encodes anionic character for the 1acl site. The constricted, modestly basic 3dfr site may discriminate against large structures that could be favorably accommodated by the less constricted 1pop and 1eed sites, and the weak charge preference of 3dfr may attenuate the strong bias toward cationic ligands encoded by 1eed. A small coefficient and t value associated with 2tsc indicate that this protein poorly encodes ligand biases for the 1acl model. In the 1acl model derived from separated score components (Table 5), the electrostatic term contributed by 2tsc is dropped during the stepwise derivation but the coefficient for the vdW component is relatively large (i.e. about a third that for the vdW component of 1pop, Table 5). The binding sites of 1acl and 2tsc share a degree of steric similarity, but are quite different in terms of electrostatics.

Missing from the regression model for 1acl are proteins 1dwc or 3cla (Table 4). The complex rugous surface of the 1dwc site, and its mixed electrostatic character, likely render this protein an ineffective surrogate for the 1acl target site. The site of 3cla is flat, exposed, and nonpolar, and thus quite dissimilar to 1acl in terms of ligand discrimination tendencies.

To a certain extent the correlations appear to reflect some coarse similarities among binding sites. For instance, large size compounds that cannot be accommodated by 1acl, also cannot be accommodated by some of the proteins factoring into the multicollinearities, such as 1pop, a dominant contributor to the regression equation (Table 4). The slight preference for cationic ligands shown by 1acl is represented by 1eed. Another way of expressing these results is that effectively what these proteins in the regression equation are doing is screening out large ligands and anionic ligands, which leaves the subset of compounds that 1acl accommodates favorably in its binding site.

The other model we consider is for the 1pop site which is a relatively exposed, semi-cylindrical valley, largely devoid of polar centers. The regression equation for 1pop draws most heavily upon 3cla followed by 2tsc, 1acl, 1dwc, and 3dfr in decreasing order of coefficient magnitude (Table 4). Similar to 1pop, the 3cla site is rather unrestricted sterically and devoid of strong polar centers. Accordingly, scores for 3cla are a reasonable predictor of scores for 1pop and so this protein is prominent in the 1pop model. However, the 1pop site is smaller and not quite as flat and exposed

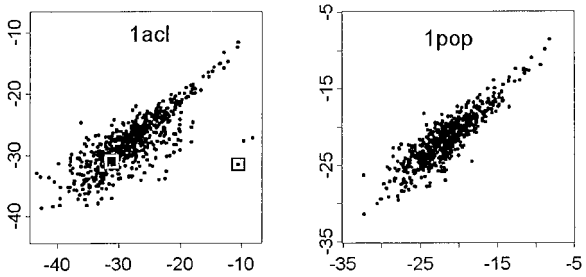

Figure 3. Plots for the 1acl and 1pop statistical models derived using full DOCK scores. Actual scores (kcal mol^{-1}) are indicated on the horizontal axes and predicted scores on the vertical axes. Boxes denote data points associated with example compounds discussed for the 1acl model (see text).

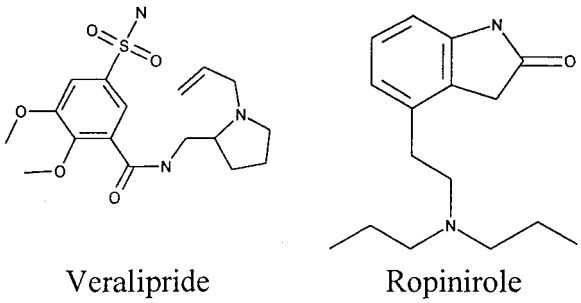

Figure 4. Compounds used to exemplify the 1acl statistical model. Structures and names were taken directly from the CMC3D database.

as that of 3cla, and the more constricted (though still relatively large) sites of 2tsc and 1acl may encode steric limitations not encoded in the scores for 3cla. The 1pop site shows a weak preference for negative charge on ligands, and the slightly greater weighting associated with 2tsc scores relative to 1acl scores may be translating this information; on average, 2tsc favors negative ligands while 1acl shows a slight bias toward positive ligands. 1dwc has a weak coefficient and its contribution to the 1pop equation is not obvious.

While 3dfr is retained by the stepwise model derivation procedure (indicating a statistically significant contribution to the model), its coefficient and associated *t* values are very small. When separated score components are used to derive a 1pop model, both 3dfr terms are dropped. The restricted 3dfr site probably cannot encode the tolerance for larger ligands that may dock at the 1pop site. Despite strongly correlated vdW score components between 1pop and 1eed, the preference for cationic ligands exhibited by the very polar 1eed site probably precludes its incorporation into the 1pop model when unseparated scores are considered.

As with the 1acl model, coarse similarities among binding sites appear to drive the regression model for 1pop, as proteins retained in the model tend to express some of the same preferences or tendencies to exclude certain types of ligands as does 1pop. These coarse similarities between binding sites, manifest statistically as library-wide biases for or against particular ligand characteristics, appear to suffice for panel proteins to productively contribute to the regression models.

Ligand protein interaction models

Figure 3 shows plots of predicted vs. actual scores for 1acl and 1pop. In most instances, both models predict scores reasonably well, though some poorly predicted outliers are seen for both systems. Comparison of outliers against accurately predicted examples reveals that poor model prediction is usually associated with atypical docking modes for either the target binding site, some of the panel sites figuring significantly in the model, or both. To illustrate this, we show two examples for the 1acl model. Structures of these example ligands are shown in Figure 4, and the corresponding data points are boxed in Figure 3. Veralipride has the largest residual in the 1acl model, while scores for ligand ropinirole fit the regression equation very well, giving this compound a small residual.

Figure 5 shows poorly predicted veralipride (top row) and correctly predicted ropinirole (bottom row) in docked complexes with proteins relevant to the 1acl statistical model. Proteins are arranged with the target site (1acl) on the left and panel members in the equation for 1acl shown, in decreasing order of regression coefficient magnitude (1pop, 1eed, 3dfr, 2tsc), progressively towards the right. As in Figure 1, protein surfaces are colored by contact frequency. Scores and separated components for the compounds are listed in Table 7.

Veralipride does not fit well in the 1acl site, with the docked orientation positioning the phenyl ring and pendant groups largely away from the protein surface in a region contacted infrequently by most other ligands. However, the same compound interacts well with the other panel sites, showing particularly good shape complementarity with all of the remaining proteins except for 3dfr. Veralipride is poorly predicted because it adopts an atypical (poor) binding orientation with target 1acl, while at the same time

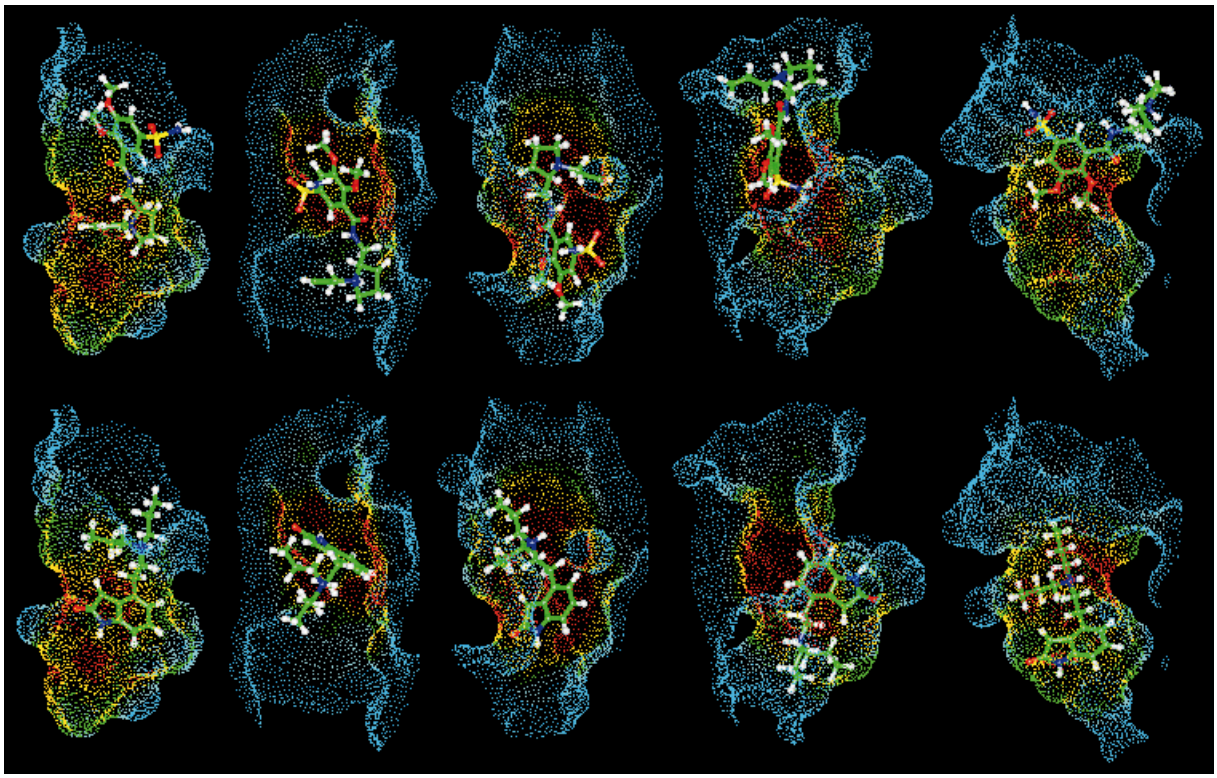

Figure 5. Example compounds docked into proteins relevant to the 1acl regression equation. The top row shows docked complexes of veralipride, which the model equation predicted poorly, and the bottom row shows complexes of ropinirole, which was predicted accurately. The 1acl target site is shown at the left, with the sites of 1pop, 1eed, 3dfr, and 2tsc arranged progressively to the right. The protein surfaces are colored by contact frequency as in Figure 1.

docking favorably to other proteins prominent in the 1acl model. In contrast, the score for ropinirole is correctly predicted because it interacts with the regions of relevant proteins that are frequently contacted by many other docked compounds.

As the example illustrates, outliers tend to be compounds that dock to one or more relevant proteins in regions that are infrequently contacted by most other library compounds and thus poorly represented by the derived regression equations. Or, in other words, atypical scores associated with atypical docking modes are in general poorly fit by the regression equations. Similar atypical patterns of docking are observed for outliers associated with each of the other protein models.

Conclusions

Real protein-ligand binding processes are, of course, far more complex than the simple model described by DOCK, and the computed scores used in the present study are not directly analogous to binding affinities. However, DOCK has been shown capable of predicting ligand binding complexes [25, 26] and thus, to a first approximation, is a reasonable model for structural investigations of ligand binding. In this case, automated docking procedures have been shown to exhibit a behavior that parallels the existence of multicollinearities observed in experimentally determined protein binding affinities [12, 27].

From our study, we conclude that some general characteristics of binding sites limit the types of ligands favorably accommodated by them. Some of these limitations may be shared, at least in part, by sites on completely unrelated proteins. General characteristics and binding determinants displayed by each protein site appear to be encoded in the form of coarse, library-wide biases towards or against ligands with certain shapes and sizes or ionized groups. Thus, while different binding sites inevitably accommodate individual ligands differently, repeatedly contacted protein features akin to ‘fuzzy recognition templates’ [28]

may modulate ligand affinities in a statistical sense, encoding underlying similarities and differences between binding sites and facilitating the existence of multicollinearities in affinities. It is tempting to speculate that a similar phenomenon may underlie the analogous correlations observed in actual experimental protein affinity data.

While detailed structural information is invaluable for optimizing the potency and selectivity of ligands directed towards any specific target, there is increasing realization that protein-ligand binding is far less specific [29, 30] than suggested by the classic lock and key analogy. The present findings suggest a number of avenues that could be explored in the design of libraries for screening, based on the fact that unrelated proteins may share certain biases in steric or electronic properties of their ligands. It is possible to envision protein classification schemes based on similarities in small molecule binding patterns. Such classification might facilitate design of compounds exhibiting specificities for (or against) a group of proteins rather than a single target. Greater understanding of structural underpinnings giving rise to these correlations should aid the development of more effective methods for binding site comparison, library design and drug discovery at large.

Acknowledgements

We acknowledge Lee Kozar and the CMGL at Stanford, as part of the work was performed while R.T.K. was a visiting scientist at the CMGL under the Spectrum program. We also thank Dr. Steve Dixon for helpful discussions and Dr. Harel Weinstein for his comments.

References

1. Lengauer, T. and Rarey, M., Curr. Opin. Struct. Biol., 3 (1996) 402.
2. Gschwend, D.A., Good, A.C. and Kuntz, I.D., J. Mol. Recognit., 2 (1996) 175.
3. Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R. and Ferrin, T.E., J. Mol. Biol., 161 (1982) 269.
4. Kuntz, I.D., Science, 257 (1992) 1078.
5. Kuntz, I.D., Meng, E.C. and Shoichet, B.K., Acc. Chem. Res., 27 (1994) 117.
6. Shoichet, B.K., Bodian, D.L. and Kuntz, I.D., J. Comput. Chem., 13 (1992) 380.
7. Meng, E.C., Shoichet, B.K. and Kuntz, I.D., J. Comput. Chem., 13 (1992) 505.
8. Meng, E.C., Gschwend, D.A., Blaney, J.M. and Kuntz, I.D., Proteins, 17 (1993) 266.
9. Meng, E.C., Kuntz, I.D., Abraham, D.J. and Kellogg, G.E., J. Comput.-Aided Mol. Design, 8 (1994) 299.
10. Lorber, D.M. and Shoichet, B.K., Protein Sci., 4 (1988) 938.
11. Briem, H. and Kuntz, I.D., J. Med. Chem., 39 (1996) 3401.
12. Kauvar, M.L., Higgins, D.L., Villar, H.O., Sportsman, J.R., Engqvist-Goldstein, A., Bukar, R., Bauer, K.E., Dilley, H. and Rocke, D.M., Chem. Biol., 2 (1995) 107.
13. Connolly, M.L., Science, 221 (1983) 709.
14. Sussman, J.L., Harel, M., Frolow, F., Oefner, C., Goldman, A., Toker, L. and Silman, I., Protein Sci., 253 (1991) 872.
15. Banner, D.W. and Hadvary, P., J. Biol. Chem., 266 (1991) 20085.
16. Cooper, J., Quail, W., Frazao, C., Foundling, S.I., Blundell, T.L., Humblet, C., Lunney, E.A., Lowther, W.T. and Dunn, B.M., Biochemistry, 31 (1992) 8142.
17. Kamphuis, I.G., Kalk, K.H., Swarte, M.B. and Drenth, J., J. Mol. Biol., 179 (1984) 233.
18. Montfort, W.R., Perry, K.M., Fauman, E.B., Finer-Moore, J.S., Maley, G.F., Hardy, L., Maley, F. and Stroud, R.M., Biochemistry, 29 (1990) 6964.
19. Leslie, A.G., J. Mol. Biol., 213 (1990) 167.
20. Filman, D.J., Bolin, J.T., Matthews, D.A. and Kraut, J., J. Biol. Chem., 257 (1982) 13663.
21. Weiner, S.J., Kollman, P.A., Case, D.A., Singh, U.C., Ghio, C., Alagona, G., Profeta, S., Jr. and Weiner, P., J. Am. Chem. Soc., 106 (1984) 765.
22. Gasteiger, J. and Marsili, M., Tetrahedron, 36 (1980) 3219.
23. Miller, A.J., Subset Selection in Regression. Monographs on Statistics and Applied Probability 40, Chapman and Hall, London, 1990.
24. Kuhn, L.A., Siani, M.A., Pique, M.E., Fisher, C.L., Getzoff, E.D. and Tainer, J.A., J. Mol. Biol., 228 (1992) 13.
25. Shoichet, B.K., Stroud, R.M., Santi, D.V., Kuntz, I.D. and Perry, K.M., Science, 259 (1993) 1445.
26. DesJarlais, R.L., Sheridan, R.P., Seibel, G.L., Dixon, J.S., Kuntz, I.D. and Venkataraghavan, R., J. Med. Chem., 31 (1988) 722.
27. Dixon, S.L. and Villar, H.O., J. Chem. Information Comput. Sci., 38 (1998) 1192.
28. Moodie, S.L., Mitchell, J.B. and Thornton, J.M., J. Mol. Biol., 263 (1996) 486.
29. La Bella, F.S., Biochem. Pharmacol., 42 (1991) S1.
30. Romesberg, F.E., Spiller, B., Schultz, P.G. and Stevens, R.C., Science, 279 (1998) 1929.