BIOINFORMATICS

Vol. 18 no. 7 2002
Pages 939–948

Ligand-induced changes in the binding sites of proteins

Xavier Fradera[1], Xavier de la Cruz[2,3], Carlos H. T. P. Silva[2], Jose Luis Gelpí[2], F. Javier Luque[1,*] and Modesto Orozco[2,*]

[1]Departament de Fisicoquímica, Facultat de Farmàcia, Universitat de Barcelona, Avgda Diagonal s/n, Barcelona 08028, Spain, [2]Departament de Bioquímica i Biología Molecular, Facultat de Química, Universitat de Barcelona, Martí i Franquès 1, Barcelona 08028, Spain and [3]Institut Catala de Recerca i Estudis Avançats (ICREA), Universitat de Barcelona, Martí i Franquès 1, Barcelona 08028, Spain

Received on June 20, 2001; revised on October 1, 2001; accepted on October 4, 2001

ABSTRACT

Classical molecular interaction potentials, in conjunction with other theoretical techniques, are used to analyze the dependence of the binding sites of representative proteins on the bound ligand. It is found that the ligand bound introduces in general small structural perturbations at the binding site of the protein. However, such small structural changes can lead to important alterations in the recognition pattern of the protein. The impact of these findings in docking procedures is discussed.

Contact: modesto@luz.bq.ub.es; javier@far1.far.ub.es

INTRODUCTION

Structural genomic is the next frontier in massive genome research projects (Burley *et al.*, 1999; Burley, 2000; Skolnick *et al.*, 2000). The final goal of structural genomics is to obtain the complete structure of the proteome of all the species of interest for humans. Knowledge of this massive amount of structural information on proteins is expected to allow us to gain insight into their biological function and their interactions with other macromolecules. Furthermore, it will also allow researchers to develop new molecules able to interact with them and to control their functionality. The aim of this structure-based drug design project is twofold. First, detailed knowledge of the structural features of proteins will facilitate the design of new drugs able to interact strongly with a target protein. Second, the new designed drugs will not be able to establish secondary interactions with proteins other than the target one. As a result, the design of more powerful, specific drugs should be greatly enhanced (Goodsell and Olson, 1990; Gschwend *et al.*, 1996; Kuntz, 1992; Kuntz *et al.*, 1994; Lengauer and Rarey, 1996; Walters *et al.*, 1998; Morris *et al.*, 1998; Liu and Wang, 1999; Farber, 1999; Fradera *et al.*, 2000; Gelpí *et al.*, 2001).

Docking programs (Goodsell and Olson, 1990; Kuntz, 1992; Kuntz *et al.*, 1994; Morris *et al.*, 1998; Walters *et al.*, 1998; Rarey *et al.*, 1996; Fradera *et al.*, 2000) exploit the structural information of the recognition site of a protein to define a reactivity map, which is then used to predict different binding modes of a given drug. Systematic search, optimization routines, molecular dynamics or Monte Carlo techniques are used to refine the binding mode of the drug. Finally, the goodness of the final model is determined with the help of scoring functions (Meng *et al.*, 1993; Goodsell and Olson, 1990; Ewing and Kuntz, 1997; Knegtel and Grootenhuis, 1998; Morris *et al.*, 1998; Liu and Wang, 1999; Ha *et al.*, 2000; Gohlke *et al.*, 2000).

Docking programs are computationally very efficient, which allows the screening of large databases of compounds searching for new 'hits' able to interact with the target protein. *In silico* screening (Kuntz, 1992; Blaney and Dizo, 1993; Kuntz *et al.*, 1994; Walters *et al.*, 1998; Farber, 1999) actually complements *high-throughput screening* methods in the discovery of new lead compound in the post-genomic era. However, the success of *in silico* screening and generally of docking techniques greatly depends on the knowledge of fine structural details of the recognition site. This means that docking strategies are often unable to detect the binding of a drug to a protein, whose structure has been determined bound to a different ligand. This suggests that, at least for some proteins, multiple sets of ligand–protein coordinates should be considered to account for the range of configurational space accessible in the binding site (Knegtel *et al.*, 1997; Apostolakis *et al.*, 1998; Carlson *et al.*, 1999).

In this paper we present a systematic study of binding sites of different proteins. These proteins were chosen due

*To whom correspondence should be addressed.

to the availability of different high-resolution crystal structures of the protein with different ligands. Comparison of the binding sites allowed us to quantify the magnitude of the ligand-induced changes in recognition properties, as well as to develop strategies to select the most suitable binding site conformation for docking studies.

METHODS

Structure selection

The Protein data bank (Headley *et al.*, 1998) was explored and 60 structures of 8 different proteins were selected for the study (see Table 1). For dimeric proteins only one of the monomers was randomly selected. In order to avoid artifacts due to errors in the resolution of the ligand–protein complex, only proteins with a resolution around or below 2 Å were considered. Each family of ligand–protein complexes was defined based on a common protein sequence. When the complex involved a mutant protein, it was rejected unless the mutation(s) was (were) far from the binding site, and modeling of the aminoacid substitution(s) was straightforward. Hydrogen atoms were added using standard protonation states for the residues. All residues that have at least 1 atom at less than 4 Å of any atom of any of the ligands in the set of ligand–protein complexes were used to define the binding site (with the exception of 1dbs ions were not considered as specific ligands of the protein). The selected complexes were then oriented along a common reference system obtained by superposition of the backbone skeleton of the protein binding sites. Finally, ligand, ions and crystallographic waters were removed.

Geometric calculations

Heavy atom Root-Mean Square deviation (RMSd) was computed to quantify the geometrical changes at the binding site induced upon ligand binding. Separate studies were performed including all heavy atoms and only backbone or the side chains (the backbone was used as reference for fitting in all the cases). A complementary analysis was performed by means of the displacement histograms. To this end, each pair of protein binding sites (the problem and the reference one) was superposed. Then, for each atom of the binding site whose accessibility change upon ligand binding, the distance between the reference and the problem protein was calculated. Finally, the atoms were grouped according to their displacement with respect to the reference position.

Cavity calculations

Cavities at the binding site were computed using the SURFNET protocol as developed by Laskowski (1995) after removing the ligands from the binding sites. The program provides up to 25 cavities for each protein complex ranging from the largest to the smallest one. The cavity corresponding to the binding site was chosen as that containing the largest number of atoms used to define the binding site (see Structure section for definition).

The shape of the binding site cavity was compared numerically using regular grids that were defined identically in all the proteins of the family. Each grid point was assigned to 1 (inside the cavity) or 0 (outside the cavity) depending on its accessibility according to Laskowski's method (Laskowski, 1995). To reduce statistical noise in proteins with very exposed binding sites, all the points located at more than 4 Å from any atom of the binding site were set to 0. The 3D matrices defined by the grids were then used to define accessible volumes from Equation (1).

$$\mathrm{Vol}_i = \alpha \sum_{k=1}^{N} \delta_{ik} \tag{1}$$

where i stands for a cavity, α is the volume of the grid element, N is the total number of points in the grid, and δ is a delta function equal to 1 if the point is inside the cavity and 0 otherwise.

cMIP calculations

Classical Molecular Interaction Potential (cMIP; Gelpí *et al.*, 2001) calculations were performed to quantify the ability of empty binding sites to interact with ligands. For this purpose, the interaction energy between the protein and three prototypical groups [an sp^3 aliphatic carbon, a positive oxygen ($q = 0.3e^-$), and a negative oxygen ($q = -0.3e$)] placed in a grid (spacing 0.5 Å) which covers all the binding site were computed. The interaction energy is determined (Gelpí *et al.*, 2001) as the addition of electrostatic and van der Waals interactions (equation 2).

$$V(r) = V_{\mathrm{ele}}(r) + V_{\mathrm{vW}}(r) \tag{2}$$

where V_{ele} and V_{vW} are the electrostatic and van der Waals potentials.

The electrostatic contribution was calculated from the solvent-screened potential determined by solving the Poisson equation (equation 3) with the standard procedure (see Gilson and Honig, 1988; Gilson *et al.*, 1988; Orozco and Luque, 2000). To capture the effect of the entire protein and solvent on the electrostatic potential at the binding site, a focusing strategy was used. To this end, the protein was initially enclosed in a box containing at least 40% empty space, and the Poisson equation is then solved numerically using a grid spacing of 1 Å. Then, a box (centered at the center of mass of all the ligands) containing all the residues of the binding site (see above) is built up, whose size is subsequently scaled by a factor of 2, and finally each axis is enlarged ±3 Å. This procedure allows us to define a very conservative box containing all

Table 1. Proteins considered in this study

Protein family	Structure (PDB code)	Ligand	Resolution (Å)	Reference
Lysozyme	1rex	−	1.5	Muraki *et al.* (1996)
	1bb5	+	1.8	Headley *et al.* (1998)
	1lzr	+	1.5	Song *et al.* (1994)
	1lzs	+	1.6	Song *et al.* (1994)
Dethiobiotin synthetase	1byi	−	1.0	Sandalova *et al.* (1999)
	1dbs	+	1.8	Alexeev *et al.* (1994)
	1bs1	+	1.8	Kaeck *et al.* (1998)
	1dad	+	1.6	Huang *et al.* (1995)
	1daf	+	1.7	Huang *et al.* (1995)
	1dag	+	1.6	Huang *et al.* (1995)
	1dah	+	1.6	Huang *et al.* (1995)
	1dam	+	1.8	Kaeck *et al.* (1998)
Cyt P450-CAM	1phc	−	1.6	Poulos *et al.* (1986)
	1pha	+	1.6	Raag *et al.* (1993)
	1phb	+	1.6	Raag *et al.* (1993)
	1phd	+	1.6	Poulos and Howard (1987)
	1phe	+	1.6	Poulos and Howard (1987)
	1phf	+	1.6	Poulos and Howard (1987)
	1phg	+	1.6	Poulos and Howard (1987)
	1cp4	+	1.9	Raag *et al.* (1990)
	2cpp	+	1.6	Poulos *et al.* (1987)
	3cpp	+	1.9	Raag and Poulos (1989a)
	5cp4	+	1.7	Vidakovic *et al.* (1998)
	6cpp	+	1.9	Raag and Poulos (1991)
	7cpp	+	2.0	Raag and Poulos (1989b)
Papain	1cvz	+	1.7	Tsuge *et al.* (1999)
	1pe6	+	2.1	Yamamoto *et al.* (1991)
	1pip	+	1.7	Yamamoto *et al.* (1992)
	1pop	+	2.1	Schroeder *et al.* (1993)
	1ppp	+	1.9	Kim *et al.* (1992)
Trypsin	1bty	+	1.5	Katz *et al.* (1995)
	1tng	+	1.8	Kurinov and Harrison (1994a)
	1tnh	+	1.8	Kurinov and Harrison (1994b)
	1tni	+	1.9	Kurinov and Harrison (1994b)
	1tnj	+	1.8	Kurinov and Harrison (1994b)
	1tnk	+	1.8	Kurinov and Harrison (1994b)
	1tnl	+	1.9	Kurinov and Harrison (1994b)
D-xylose-isomerase	1xib	−	1.6	Carrell *et al.* (1994)
	1xic	+	1.6	Carrell *et al.* (1994)
	1xid	+	1.7	Carrell *et al.* (1994)
	1xie	+	1.7	Carrell *et al.* (1994)
	1xif	+	1.6	Carrell *et al.* (1994)
	1xig	+	1.7	Carrell *et al.* (1994)
	1xih	+	1.7	Carrell *et al.* (1994)
	1xii	+	1.7	Carrell *et al.* (1994)
	1xij	+	1.7	Carrell *et al.* (1994)
	8xia	+	1.9	Carrell *et al.* (1989)
	9xia	+	1.9	Carrell *et al.* (1989)
Chymotrypsin	2gch	−	1.9	Cohen *et al.* (1981)
	2gmt	+	1.8	Kreutter *et al.* (1994)
	3gch	+	1.9	Stoddard *et al.* (1990)
	4gch	+	1.9	Stoddard *et al.* (1990)
	7gch	+	1.8	Brady *et al.* (1990)
Thymidine kinase	1e2k	−	1.7	Vogt *et al.* (2000)
	1e2m	+	2.2	Wurth *et al.* (2001)

(*continued . . .*)

Table 1 *continued . . .*

Protein family	Structure (PDB code)	Ligand	Resolution (Å)	Reference
	1qhi	+	1.9	Bennet *et al.* (1999)
	1kim	+	2.1	Champness *et al.* (1998)
	1ki8	+	2.2	Champness *et al.* (1998)
	1vtk	+	2.7	Wild *et al.* (1997)
	2vtk	+	2.8	Wild *et al.* (1997)

The presence (+) or absence (−) of ligand, the resolution (in Å), the pdb code and the key references are noted.

the region of interest around the binding site. The Poisson equation is solved using a grid spacing of 0.5 Å, and the potential computed previously by using the initial box.

$$\nabla \cdot [\varepsilon(r_i) \cdot \nabla \cdot V_{\mathrm{ele}}(r_i)] = -4\pi\rho(r_i) \tag{3}$$

where ε is the dielectric constant (2 inside the protein and 80 outside), V_{ele} is the electrostatic potential, ρ is the charge density, and i stands for a grid position.

The van der Waals term in the binding site box was computed using parameters adopted from the AMBER-98 force field for the sp^3 aliphatic carbon and water oxygen and the AMBER force field for the residues in the protein (Cornell *et al.*, 1995), and (4), where z stands for the probe atom considered, L stands for all the residues of the protein, ε and R are van der Waals parameters.

$$V_{\mathrm{vW}}^{z}(r_i) = \sum_{i=1}^{L} (\varepsilon_1 \varepsilon_z)^{1/2} \left[\left(\frac{R_z + R_l}{r_1 - r_i} \right)^{12} - 2 \left(\frac{R_z + R_l}{r_1 - r_i} \right)^{6} \right]. \tag{4}$$

Statistical analysis

The absolute and relative change in the volume of the binding site cavity induced upon ligand binding were computed from (5) and (6), where Vol_i is defined as noted in (1), and P and Q denote two different ligand–protein complexes of the same protein. The differences in shapes of the binding site cavities were quantified by using the similarity index η defined in (7), where δ_{ik} (P) and γ_{ik} (Q) are delta functions for the two proteins compared. These functions are 1 if grid point k is within the cavity i, and 0 otherwise.

$$\Delta \mathrm{Vol}_i^{\mathrm{P-Q}} = \mathrm{Vol}_i^{\mathrm{P}} - \mathrm{Vol}_i^{\mathrm{Q}} \tag{5}$$

$$\Delta r \mathrm{Vol}_i^{\mathrm{P-Q}} = 2 \frac{\mathrm{Vol}_i^{\mathrm{P-Q}}}{\mathrm{Vol}_i^{\mathrm{P}} + \mathrm{Vol}_i^{\mathrm{Q}}} \tag{6}$$

$$\eta_i^{\mathrm{P/Q}} = \frac{\sum_{k=1}^{N} \delta_{ik}\gamma_{ik}}{\left(\left(\sum_{k=1}^{N} \delta_{ik} \right) \left(\sum_{k=1}^{N} \gamma_{ik} \right) \right)^{1/2}}. \tag{7}$$

The cMIP for the three different probes was compared using non-parametrical Spearman's test. Accordingly, the correlation coefficient (r(P, Q)) between two binding site grids (of the same size, centered at the same position, and computed after superposition of the residues at the binding site) is defined in (8). To reduce the noise in the calculation of Spearman's matrices, points with very small (in absolute value) interaction energies ($|E| < 0.01$ kcal mol^{-1}), and points with very unfavorable interaction energies ($E > 5$ kcal mol^{-1}) for the two proteins that were compared were eliminated from their original grids.

$$r(\mathrm{P}, \mathrm{Q}) = \frac{\sum_{k=1}^{N} (R_k - \bar{R})(S_k - \bar{S})}{\left(\sum_{k=1}^{N} (R_k - \bar{R})^2 \right)^{1/2} \left(\sum_{k=1}^{N} (S_k - \bar{S})^2 \right)^{1/2}} \tag{8}$$

where R_k and S_k are the cMIP ranks of grid point k for proteins P and Q.

The Spearman coefficients for all the pairs of structures (and for each probe) define a cross correlation matrix R_{PQ}, which indicates the degree of similarity between all the pairs of protein structures in a given family. The cross-correlation matrix can define three different scenarios: (i) very similar binding sites, which would yield to a matrix with all elements close to 1, (ii) very different binding sites, which would yield to a matrix with all elements close to 0, and (iii) binding sites which can be grouped in several classes, thus leading to matrices with elements close to 1 and others not far from 0. Principal Component Analysis (PCA) was used to examine the cross-correlation matrix. To this end, we first standardized the cross-correlation matrices in such a way that all the values are centered in 0 and display a variance of 1. The resulting matrices are then diagonalized to obtain the principal components. The analysis of the first and second principal components (in all the cases these two components explain more than 95% of the total variance) allows us to cluster the binding sites according to their similarity in terms of reactivity. The study was performed for the cMIP grids defined with the three probes, but only the results obtained for the positive probe are displayed (other PCA analyses are available upon request to the authors).

RESULTS AND DISCUSSION

The general macromolecular structure of the proteins studied is not largely altered by the ligand. With regard to the binding site, backbone-RMSd (reference structures are 1rex, 1dbs, 1phc, 1cvz, 1bty, 1xib, 2gch and 1e2k) are generally small (see Table 2). Significant deviations are found in proteins like lysozyme, and specially dethiobiotin synthase, where no negligible backbone movements occur upon ligand binding. Other proteins like D-xylose isomerase have a very rigid backbone. As expected, the local RMSd at the binding site (see Table 2) increases if side chains are considered, indicating that most of the structural rearrangement induced by the ligand involves side chain movements (see Table 2). However, most all-heavy atoms RMSd at the binding site are still below 1 Å, and only in one case (1dbs versus 1byi) the difference is greater than 2 Å (see Table 2). This suggests that the structure of the binding site is generally preserved irrespective of the nature of the ligand, and only small side chain movements are necessary to accommodate different ligands.

More detailed information about geometrical changes comes from the deviation frequency of 'contact' atoms, i.e. the frequency in which atoms whose solvent accessibility change upon removal of the ligand deviate owing to different ligand binding (the same reference structures noted above were used. As noted in Figure 1 more than 90% of the contact atoms deviate less than 1 Å from the reference position, and less than 3% of the atoms deviate more than 2 Å from the reference position (due to its large RMSd 1byi was excluded from this analysis). This means that with some exceptions (see below) the structure of the binding site is not dramatically altered by the bound ligand, even for those atoms which interact directly with the ligand.

To clarify whether or not small geometrical changes can affect the ability of the binding site to bind different ligands, we examined the binding site cavities (see Section **Methods**). Table 3 shows the absolute and relative change in volume, and the similarity index η for the different systems (the same set of reference structures noted above was used). As suggested by Laskowski (1995) the binding site defines the larger cavity in the studied set of proteins. The volume of the binding site cavity ranges between 900 and 5000 Å^3 (see Table 3), but there are several proteins for which the size of the binding pocket is similar despite the fact that they bind different substrates. This indicates that, at least, for the reduced set of proteins considered here the volume of the binding site is not a major discriminant factor in ligand binding.

The binding of different ligands introduces remarkable changes in the accessible volume of the binding site (see Table 3), which were not obvious from the small geometrical changes induced upon ligand binding (see above). The changes in volume represent in general

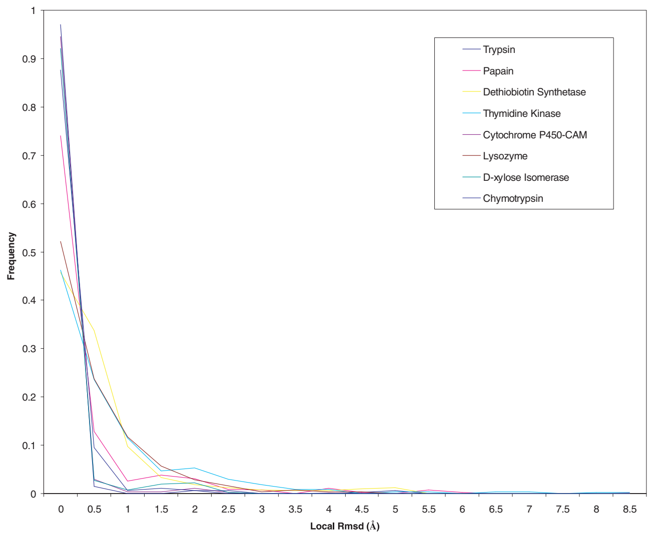

Fig. 1. Frequency plot representing the population of contact atoms (see text for definition) as a function of the deviation (RMSd in Å) with respect to the reference geometry.

around 20–40% of the total volume, but there is strong variability between proteins. Thus, the relative volume change is below 10% for D-xylose isomerase, while it is larger than 60% for thymidine kinase.

Analysis of the similarity index (η) provides complementary, more precise information, since it accounts not only for the total volume of the binding site cavity, but also for its shape (see equation 7). Similarity indexes around 40–60% (see Table 3) indicate that, in general, the binding site cavity is quite flexible to fit the bound ligand. Once again, there is strong variability, since similarity indexes around 90% are found for D-xylose isomerase, while values below 30% are detected for thymidine kinase. It is clear then that volume calculations are much more sensible to changes in binding cavities than simple RMSd calculations.

In summary, as noted in the RMSd analysis the geometry of the binding site of the proteins examined here seems to be quite insensitive to the binding of different ligands. However, these small geometrical changes can introduce important modifications in the volume and shape of the binding site cavity. To determine the impact of these subtle geometrical changes on the molecular recognition properties of the binding site, we analyzed the cMIPs for three prototypical probes: a positive group O$^+$($q = 0.3$e), a negative one ($q = 0.3$e), and a van der Waals particle (see Section **Methods**). The cMIPs for the different proteins were compared to derive cross-correlation matrices using Spearman's test. The non-parametric nature of the Spearman's index, and the removal of regions of steric collapse,

Table 2. RMSd in Å between the different structures (only active site residues are considered) of the eight families studied and the corresponding reference structures (see text, and first column in table). RMSds are computed considering only the backbone, all the heavy atoms and the side-chain groups

Protein family (Reference structure)	Structure (PDB code)	RMSD (Å)			Structure (PDB code)	RMSD (Å)		
		Back	All	Side		Back	All	Side
Lysozyme	1bb5	0.61	0.95	1.12				
(1rex)					1lzr	0.54	0.80	0.97
	1lzs	0.77	1.10	1.35				
Dethiobiotin	1bs1	0.61	0.96	1.25				
synthetase					1dah	0.54	0.92	1.22
(1dbs)	1dad	0.40	0.74	1.01				
					1dam	0.62	0.84	1.04
	1daf	0.61	0.98	1.29				
					1byi	2.01	2.61	3.19
	1dag	0.51	0.99	1.33				
Oxidoreductase	1pha	0.14	1.48	2.10	1cp4	0.13	0.19	0.23
Cyt P450-CAM	1phb	0.16	1.42	2.03	2cpp	0.16	0.46	0.64
(1phc)	1phd	0.19	0.21	0.23	3cpp	0.16	0.47	0.64
	1phe	0.44	0.60	0.73	5cp4	0.16	0.30	0.40
	1phf	0.19	0.28	0.34	6cpp	0.12	0.44	0.62
	1phg	0.19	0.46	0.63	7cpp	0.13	0.46	0.64
Papain	1pe6	0.16	0.60	0.86	1pop	0.20	0.82	1.19
(1cvz)	1pip	0.45	0.78	1.08	1ppp	0.51	0.95	1.32
Trypsin	1tng	0.19	0.88	1.27	1tnj	0.15	0.43	1.22
(1bty)	1tnh	0.15	0.84	1.22	1tnk	0.15	0.86	1.24
	1tni	0.15	0.86	1.24	1tnl	0.14	0.82	1.19
D-xylose-	1xic	0.07	0.31	0.40	1xih	0.14	0.65	0.85
isomerase	1xid	0.07	0.13	0.16	1xii	0.06	0.19	0.25
(1xib)	1xie	0.09	0.23	0.29	1xij	0.14	0.37	0.48
	1xif	0.08	0.22	0.28	8xia	0.09	0.49	0.64
	1xig	0.12	0.17	0.21	9xia	0.10	0.64	0.64
Chymotrypsin	2gmt	0.25	0.73	1.08	4gch	0.23	0.31	0.39
(2gch)	3gch	0.28	0.35	0.43	7gch	0.28	0.43	0.57
Thymidine	1e2m	0.37	0.52	0.68	1ki8	0.38	1.22	1.61
kinase	1qhi	0.48	1.62	2.10	1vtk	0.69	1.51	1.97
(1e2k)	1kim	0.29	1.57	2.22	2vtk	0.51	1.34	1.76

or irrelevant for binding from the cMIP calculation makes the test very robust to detect correlations between binding sites.

Analysis of cross-correlation factors are shown in Tables 4a–h (in order to reduce the length of the paper these tables are removed from the printed version, and are available as pdf files in http://www.bq.ub.es/recmol/docs/Table4.pdf. Cross-correlation factors between 0.6 and 0.8 in most cases (coefficients below 0.03 were detected in control calculations where random distributions of binding sites were used). Extreme values from 0.20 to 0.98 are detected depending on the protein and the bound ligand. As expected, proteins where the fine geometrical details of the binding site are more dependent on the ligand show also the lowest cross-correlation factors. Interestingly, the results obtained with the neutral and the two charged probes (O^+ and O^-) are not too different, but in general the cross-correlation coefficients are smaller for the neutral probe. These results, combined with those obtained from the analysis of cavities, strongly suggest that the binding sites are in general more flexible in terms of shape and steric properties than in terms of the electrostatic distribution.

PCA was used to cluster the structures of proteins within each family, and to analyze which ligand(s) induce(s) the most dramatic changes in the structure of the protein. PCA can also help us to find representative structures of the protein (those placed near the center of the clusters), which can be useful for multiple-structure docking purposes.

Figure 2 displays a projection of the different structures

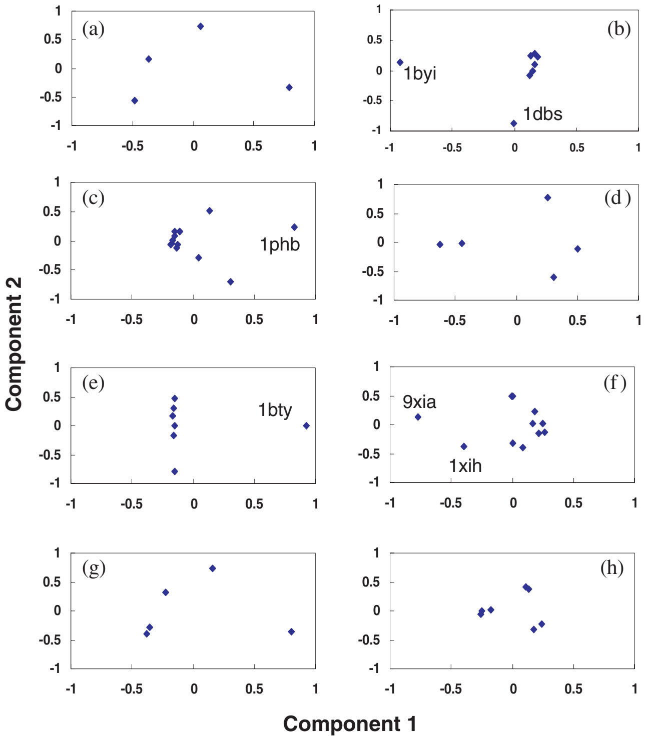

Fig. 2. PCA plot representing the projection of the different structures of each family along the two first principal components obtained after diagonalization of the standarized cross-correlation matrix corresponding the cMIP (probe = O^{+}).

of each family in the first two principal components obtained by diagonalization of the standarized cross-correlation matrices derived from the cMIP with a positive probe (see Section **Methods**). Two general situations are found: (i) families of structures dispersed in a quasi-random way (lysozyme, thymidine kinase, papsin and chymotrypsin), and (ii) families where most structures are found in one cluster, and only a few outliers are detected (dethiobiotin synthetase, trypsin, D-xylose isomerase, and Cyt P450-CAM). It is worth noting that in any case our definition of a cluster is very conservative due to the limited number of points introduced in the study.

Analysis of protein families that are not clearly clustered shows a diversity of situations (note that due to the standarization of cross-correlation matrix, PCA plots of different families cannot be compared). For instance, lysozyme shows a large flexibility at the binding site as noted in no-standarized cross-correlation matrix in Table 4a (http://www.bq.ub.es/recmol/docs/Table4.pdf), and a fully random distribution is detected in the PCA plot (see Figure 2). On the contrary, three subfamilies can be detected for thymidine kinase, corresponding to structures solved with different ligands (non-nucleotidic inhibitors

Table 3. Average total (Δ vol in $Å^3$) relative volume (rel Δ vol) change and volume similarity index (η) for each family of proteins. Standard deviations are reported in parentheses

	Average Δ vol	Average rel. Δ vol	η
Lysozyme	71.34 (58.1)	0.071 (0.059)	0.752 (0.015)
Dethiobiotin synthetase	1070.08 (362.8)	0.312 (0.121)	0.545 (0.073)
Oxidoreductase Cytochrome P450-CAM	674.26 (437.27)	0.198 (0.143)	0.745 (0.105)
Papain	104.83 (44.97)	0.106 (0.041)	0.724 (0.016)
Trypsin	233.81 (41.68)	0.210 (0.033)	0.709 (0.012)
D-xylose isomerase	78.80 (58.57)	0.016 (0.012)	0.900 (0.041)
Chymotrypsin	362.24 (170.91)	0.177 (0.088)	0.720 (0.048)
Thymidine kinase	738.23 (587.56)	0.426 (0.407)	0.449 (0.128)

+ SO_4^{2-}, nucleotides + SO_4^{2-} and nucleotides + ADP). It is worth noting that for these families of proteins the recognition properties of the binding site are not directly related to the presence or absence of ligands. That is the case of lysozyme, where the similarity indexes between the unbound protein (1rex) and any of the bound forms (for instance 1lzr) are similar to those obtained when bound forms are compared (for instance 1bb5 and 1lzs).

Protein families where a majority of structures appear clustered can be interpreted as proteins that have a preferred configuration of the binding site, but that can adapt its binding site configuration under some conditions. For instance, the unique binding site configuration of 1bty (trypsin) is due to rotation of the side chain of one Gln^{192}. A different orientation of the side chain of one Phe^{96} and one Tyr^{193} is the reason for the unique characteristics of 1phb in the Cyt P450-CAM family. Small side chain movements of different polar residues like His, Glu and Asp are responsible for the moderate outlier characteristics of 9xia and 1xih in the D-xylose isomerase family. Finally, a different backbone arrangement in positions 10–13 (see also Table 2) and changes in the orientation of a Pro^{210} and one Glu^{115} are likely responsible for the differential characteristics of 1byi with respect to the other structures of the dethiobiotin synthetase family.

A detailed analysis (in Figure 2 and Table 4) shows that only one of the outliers (1byi for dethiobiotin synthase) corresponds to an unbound protein. In all the other cases (Cyt P450-CAM, trypsin and D-xylose isomerase) the larger differences in recognition properties appears in binding sites bound to ligands. This finding suggests that most of the proteins studied here do not follow a two-step 'induced fitting' mechanism implying two conformational states for the 'unbound' and 'bound' forms. On the contrary, results support a mechanism in which the binding sites show a certain degree of flexibility, which help them to fit different molecules, either unstructured solvent (for the unbound form) or specific ligands. However, caution is necessary since for 3 proteins (thymidine kinase, chymotrypsyn, and trypsin) the unbound form of the enzyme is not reported in PDB. It can be suggested that for these proteins the structure and flexibility of the protein in its free and bound forms may be very different.

Overall, our studies suggest that the structures of binding sites are preserved upon ligand binding. However, the structural conservation does not imply a similar conservation in binding properties. Rather, small side chain (and in some cases backbone) movements alter the volume and recognition at the binding site. The changes are not necessarily larger when unbound and ligand-bound structures are compared relative to the comparison between pairs of ligand–protein complexes. Interestingly, the proteins are less flexible in terms of conservation of the electrostatic distribution than in preservation of the steric properties, which agrees with the fact that electrostatic properties are generally the main reason for differential binding in proteins. Finally, the results also suggest that multiple-structure docking is preferred. The cross-correlation found between the recognition properties of binding sites of the same protein bound to different ligands is mediocre, even when the structures were solved by the same group under the same experimental conditions. This suggests that docking strategies performed with a single protein structure can yield to erroneous results even if the structure of a ligand–protein complex is available. cMIP calculations coupled to the analysis of Spearman's cross correlation matrices and PCA can help to select representative structures for proteins for docking purposes. Whether or not structures generated from molecular dynamics simulations can be used to complement the ensemble of protein structures for binding will be the issue of a future work.

ACKNOWLEDGEMENTS

We thank Dr A.Valencia, A.Muñoz, D.A.de Juan and O.Graña for suggestions on the manuscript. This work has been supported by the Spanish Ministry of Science and Technology (PB98-1222 and PM99-0046), and the Centre de Supercomputació de Catalunya (CESCA; Mol. Recog. Project).

REFERENCES

Alexeev,D., Baxter,R.L. and Sawyer,L. (1994) Mechanistic implications and family from the structure of dethiobiotin synthetase. *Structure (London)*, **2**, 1061–1072.

Apostolakis,J., Plückthum,A. and Caflisch,A. (1998) Docking small ligands in flexible binding sites. *J. Comput. Chem.*, **19**, 21–37.

Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Blaney,J.M. and Dizo,J.S. (1993) A good ligand is hard to find: automatic docking methods. *Perspect Drug. Discov. Des.*, **1**, 301–319.

Brady,K., Wei,A.Z., Ringe,D. and Abeles,R.H. (1990) Structure of chymotrypsin-trifluoromethyl ketone inhibitor complexes: comparison of slowly and rapidly equilibrating inhibitors. *Biochemistry*, **29**, 7600–7607.

Burley,S.K. (2000) An overview of structural genomics. *Nature Struct. Biol.*, **7**, 932–934.

Burley,S.K., Almo,S.C., Bonanno,J.B., Capel,M., Chance,M.R., Gaasterland,T., Lin,D., Sali,A., Studier,F.W. and Swaminathan,S. (1999) Strutural genomics: beyond the human genome project. *Nat. Genet.*, **23**, 151–157.

Carlson,H., Masukawa,K.M. and McCammon,J.A. (1999) Method for including the dynamic fluctuations of a protein in computer-aided drug design. *J. Phys. Chem.* A, **103**, 10 213–10 219.

Carrell,H.L., Glusker,J.P., Burger,V., Manfre,F., Tritsch,D. and Biellmann,J.F. (1989) X-ray analysis of D-xylose isomerase at 1.9 angstroms: native enzyme in complex with substrate and with a mechanism-designed inactivator. *Proc. Natl Acad. Sci. USA*, **86**, 4440–4444.

Carrell,H.L., Hoier,H. and Glusker,J.P. (1994) Modes of binding substrates and their analogues to the enzyme D-xylose isomerase. *Acta Crystallogr. D Biol. Crystallogr.*, **50**, 113–123.

Cohen,G.H., Silverton,E.W. and Davies,D.R. (1981) Refined crystal structure of gamma-chymotrypsin at 1.9 A resolution. Comparison with other pancreatic serine proteases. *J. Mol. Biol.*, **148**, 449–479.

Cornell,W.D., Cieplak,P., Bayly,C.I., Gould,I.R., Merz,K.M., Ferguson,D.M., Spellmeyer,D.C., Fox,T., Caldwell,J.W. and Kollman,P.A. (1995) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.*, **117**, 5179–5197.

Ewing,T. and Kuntz,I.D. (1997) Critical evaluation of search algorithm for flexible docking. *J. Comput. Chem.*, **18**, 1175–1189.

Farber,G.K. (1999) New approaches to rational drug design. *Pharmacol. Ther.*, **84**, 327–332.

Fradera,X., Knegttel,R.M.A. and Mestres,J. (2000) Similarity-driven flexible ligand docking. *Proteins*, **40**, 623–636.

Gelpí,J.L., Kalko,S., de la Cruz,X., Barril,X., Cirera,J., Luque,F.J. and Orozco,M. (2001) Use of classical molecular interaction potentials in molecular dynamics simulations of proteins. *Proteins*, **45**, 428–437.

Gilson,M. and Honig,B. (1988) Calculation of the total electrostatic energy of a macromolecular system. Solvation energies, binding energies and conformational analysis. *Proteins*, **4**, 7–18.

Gilson,M., Sharp,K. and Honig,B. (1988) Calculating the electrostatic potential of molecules in solution. *J. Comput. Chem.*, **9**, 327–335.

Goodsell,D.S. and Olson,A.J. (1990) Automated docking of substrates to proteins by simulated annealing. *Proteins*, **8**, 195–202.

Gohlke,H., Hendlich,M. and Klebe,G. (2000) Knowledge-based scoring functions to predict protein–ligand interactions. *J. Mol. Biol.*, **295**, 337–356.

Gschwend,D.A., Good,A.C. and Kuntz,I.D. (1996) Molecular docking towards drug discovery. *J. Mol. Recognit.*, **8**, 175–186.

Ha,S., Andreani,R., Robbins,A. and Muegge,I. (2000) Evaluation of docking/scoring approaches: A comparatice study based on MMP3 inhibitors. *J. Comput. Aided Mol. Des.*, **14**, 435–448.

Headley,A.G., Roe,S.M. and Pearl,L.H. (1998) Protein data bank 1bb5, unpublished.

Huang,W., Jia,J., Gibson,K.J., Taylor,W.S., Rendina,A.R., Schneider,G. and Lindqvist,Y. (1995) Mechanism of an ATP-dependent carboxylase, dethiobiotin synthetase, based on studies of complexes with substrates and a intermediate. *Biochemistry*, **34**, 10 985–10 995.

Kaeck,H., Sandmark,J., Gibson,K.J., Schneider,G. and Lindqvist,Y. (1998) Crystal structure of two quaternary complexes of dethiobiotin synthetase, enzyme-MgADP-ALF3-diaminopelargonic acid and enzyme-MgADP-dethiobiotin-phosphate; implications for catalysis. *Protein Sci.*, **7**, 2560–2566.

Katz,B.A., Finer-Moore,J., Mortezaei,R., Rich,S.H. and Stroud,R.M. (1995) Episelection: novel Ki ∼nanomolar inhibitors of serine proteases selected by binding or chemistry on an enzyme surface. *Biochemistry*, **34**, 8264–8280.

Kim,M.J., Yamamoto,D., Matsumoto,K., Inoue,M., Ishida,T., Mizuno,H., Sumiya,S. and Kitamura,K. (1992) Crystal structure of papain-E64-c complex. Binding diversity of E64-c to papain S2 and S3 subsites. *Biochem. J.*, **287**, 797–803.

Knegtel,R.M.A. and Grootenhuis,P.D.J. (1998) Binding affinities and non-bonded interaction energies. *Perpect. Drug Discov. Des.*, **9**, 99–114.

Knegtel,R.M.A., Kuntz,I.D. and Oshiro,C.M. (1997) Molecular docking to ensembles of protein structures. *J. Mol. Biol.*, **266**, 424–440.

Kreutter,K., Steinmetz,A.C., Liang,T.C., Prorok,M., Abeles,R.H. and Ringe,D. (1994) Three-dimensional structure of chymotrypsin inactivated with (2S)-*N*-acetyl-L-alanyl-L-phenylalanyl alpha-chloroethane: implications for the mechanism of inactivation of serine proteases by chloroketones. *Biochemistry*, **33**, 13 792–13 800.

Kuntz,I.D. (1992) Structure-based strategies for drug design and discovery. *Science*, **257**, 1078–1082.

Kuntz,I.D., Meng,E.C. and Shoichet,B.K. (1994) Structure-based molecular design. *Acc. Chem. Res.*, **27**, 117–123.

Laskowski,R.A. (1995) SURFNET: a program for visualizing molecular surfaces, cavities and intermolecular interactions. *J. Mol. Graph.*, **13**, 323–330.

Liu,M. and Wang,S. (1999) MCDOCK: a Monte Carlo simulation approach to the molecular docking problem. *J. Comput. Aided Mol. Des.*, **13**, 435–451.

Lengauer,T. and Rarey,M. (1996) Computational methods for biomolecular docking. *Curr. Opin. Struct. Biol.*, **6**, 402–406.

Meng,E.C., Shoichet,B.K. and Kuntz,I.D. (1993) Orientational sampling and rigid-body minimization in molecular docking. *Proteins*, **17**, 266–278.

Morris,G.M., Goodsell,D.S., Halliday,R.S., Huey,R., Hart,W.E., Belew,R.K. and Olson,A.J. (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.*, **19**, 1639–1662.

Muraki,M., Harata,K., Sugita,N. and Sato,K. (1996) Origin of carbohydrate recognition specificity human lysozyme revealed by affinity labeling. *Biochemistry*, **35**, 13 562–13 567.x

Orozco,M. and Luque,F.J. (2000) Theoretical methods for the description of the solvent effect in biomolecular systems. *Chem. Rev.*, **100**, 4187–4225.

Poulos,T.L., Finzel,B.C. and Howard,A.J. (1986) Crystal structure of substrate-free *Pseudomonas putida* cytochrome P450. *Biochemistry*, **25**, 5314–5322.

Poulos,T.L. and Howard,A.J. (1987) Crystal structures of metyrapone- and phenylimidazole-inhibited complexes of cytochrome P450cam. *Biochemistry*, **26**, 8165–8174.

Poulos,T.L., Finzel,B.C. and Howard,A.J. (1987) High-resolution crystal structure of cytochrome P450cam. *J. Mol. Biol.*, **195**, 687–700.

Raag,R. and Poulos,T.L. (1989a) Crystal structure of the carbon monoxy-substrate-cytochrome P450/cam ternary complex. *Biochemistry*, **28**, 7586–7592.

Raag,R. and Poulos,T.L. (1989b) The structural basis for substrate-induced changes in redox potential and spin equilibrium in cytochrome P-450(cam). *Biochemistry*, **28**, 917–922.

Raag,R. and Poulos,T.L. (1991) Crystal structures of cytochrome P-450/cam complexed with camphane, thiocamphor, and adamantane: factors controlling P-450 substrate hydroxylation. *Biochemistry*, **30**, 2674–2684.

Raag,R., Swanson,B.A., Poulos,T.L. and Ortiz De Montellano,P.R. (1990) Formation, crystal structure, and rearrangement of a cytochrome P450cam-iron–phenyl complex. *Biochemistry*, **29**, 8119–8126.

Raag,R., Li,H., Jones,B.C. and Poulos,T.L. (1993) Inhibitor-induced conformational change in cytochrome P450cam. *Biochemistry*, **32**, 4571–4578.

Rarey,M., Kramer,B., Lengauer,T. and Klebe,G. (1996) A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.*, **261**, 470–489.

Sandalova,T., Schneider,G., Kack,H. and Lindqvist,Y. (1999) Structure of dethiobiotin synthetase at 0.97 A resolution. *Acta Crystallogr. Sect.* D, **55**, 610–624.

Schroeder,E., Phillips,C., Garman,E., Harlos,K. and Crawford,C. (1993) X-ray crystallographic structure of a papain-leupeptin complex. *FEBS Lett.*, **315**, 38–42.

Skolnick,J., Fetrow,J.S. and Kolinski,A. (2000) Structural genomics and its importance for gene function analysis. *Nat. Biotechnol.*, **18**, 283–287.

Song,H., Inaka,K., Maenaka,K. and Matsushima,M. (1994) Structural changes of the active site cleft and different saccharide binding modes in lysozyme co-crystallized with hexa-n-acetyl-chitohexaose at pH 4.0. *J. Mol. Biol.*, **244**, 522–540.

Stoddard,B.L., Bruhnke,J., Porter,N., Ringe,D. and Petsko,G.A. (1990) Structure and activity of two photoreversible cinnamates bound to chymotrypsin. *Biochemistry*, **29**, 4871–4879.

Tsuge,H., Nishimura,T., Tada,Y., Asao,T., Turk,D., Turk,V. and Katunuma,N. (1999) Inhibition mechanism of cathepsin l-specific inhibitors based on the crystal structure of papain-CLIK148 complex. *Biochem. Biophys. Res. Commun.*, **266**, 411–416.

Vidakovic,M., Sligar,S.G., Li,H. and Poulos,T.L. (1998) Understanding the role of the essential Asp251 in cytochrome P450cam using site-directed mutagenesis, crystallography, and kinetic solvent isotope effect. *Biochemistry*, **37**, 9211–9219.

Walters,W.P., Stahl,M.T. and Murcko,M.A. (1998) Virtual screening an overview. *Drug Discov. Today*, **3**, 160–178.

Yamamoto,D., Matsumoto,K., Ohishi,H., Ishida,T., Inoue,M., Kitamura,K. and Mizuno,H. (1991) Refined x-ray structure of papain(dot)E-64-c complex at 2.1-Angstroms resolution. *J. Biol. Chem.*, **266**, 14 771–14 777.

Yamamoto,A., Tomoo,K., Doi,M., Ohishi,H., Inoue,M., Ishida,T., Yamamoto,D., Tsuboi,S., Okamoto,H. and Okada,Y. (1992) Crystal structure of papain-succinyl-Gln-Val-Val-Ala-Ala-p-nitroanilide complex at 1.7 Angstroms resolution: noncovalent binding mode of a common sequence of endogenous thiol protease inhibitors. *Biochemistry*, **31**, 11 305–11 309.