Articles

A Novel Strategy for Improving Ligand Selectivity in Receptor-Based Drug Design

Manuel Pastor and Gabriele Cruciani*

Department of Chemistry, University of Perugia, Via Elce di Sotto, 8, 06100 Perugia, Italy

Received June 1, 1995®

A major desirable characteristic of many drugs is their ability to interact specifically with only one variety of the target receptor among many others. It is remarkable that, even when accurate three dimensional structures for the target biomolecules are available, there is no well-established methodology to describe their differences and use them for the design of selectively-interacting compounds. This work presents a novel method that uses multivariate GRID descriptors and principal component analysis (PCA) with the aim of revealing the most relevant structural and physicochemical differences between biomacromolecules related to receptor selectivity. The methodology is described through an example involving the study of bacterial (*Escherichia coli*) and recombinant human varieties of the dihydrofolate reductase (EC 1.5.1.3, DHFR) enzyme. This analysis easily unveils the most important regions on these biomolecules which should be taken into consideration for the design of selectively interacting compounds.

Introduction

There are a large number of situations in which the ability of drugs to interact selectively with one and only one biomolecule is by far their most important characteristic. A typical example is the design of antibacterial and antiprotozoal agents, where the drugs should be toxic for the infectious agent but innocuous to human beings. This property is also relevant in the area of antineoplastic agents and in many other fields.

In the search for new, selectively interacting compounds, the X-ray three-dimensional structures of biomolecules have often been regarded as a valuable source of information. The number of such structures available is growing steadily, and the prospects for the future are optimistic.[1,2]

However, the intermolecular interaction is a complex phenomenon, and at first glance these structures do not give sufficient useful information for the design of selective compounds. The exploitation of the information contained in the X-ray structure, in order to design better inhibitors, therefore continues to be an elusive goal.[3]

The dihydrofolate reductase (5,6,7,8-tetrahydrofolate: $NADP^+$ oxidoreductase, EC 1.5.1.3, DHFR) is an example which illustrates the problem. This is a small, stable, and crystallographically well-behaved enzyme, well suited for high-resolution crystallographic studies.[4] From the point of view of its properties it is also very interesting. It catalyzes the reduction of dihydrofolate to tetrahydrofolate, which is required as a coenzyme for the synthesis of thymidylate and other important metabolites. The blockade of DHFR produces a depletion of thymidylate and subsequently leads to the cessation of cell proliferation.[5] Inhibitors of this target have been used in therapy as antibacterials (trimethoprim, TMP), antimalarials (pyrimethamine) and antineoplastic agents (methotrexate, MTX). Some of these inhibitors exhibit a high degree of species selectivity and, for instance, the TMP IC_{50} values against *Escherichia coli* and human DHFR differ by more than 5 orders of magnitude.[5]

After the publication in the late 1970s of the three-dimensional structure of *E. coli* dihydrofolate reductase,[6] there was a lot of expectation about the use of this information for the design of more specific chemotherapeutic and anticancer drugs.[6] Since that time, crystal structures of a variety of substrates, cofactor, and inhibitory complexes, both binary and ternary, have been determined for DHFR from different species,[7–13] and a lot of work has been devoted to the understanding of the enzymatic mechanism and of ligand and inhibitor binding. The enzyme has become a classical system in which many drug design techniques have been tested, ranging from classical QSAR to the most sophisticated 3D QSAR methodologies. Despite the information gained, a quantitative understanding of the ligand–protein specificity in this system continues to remain elusive,[14] and contradictory results are not rare.[7,10] Until now none of these studies has led to better and more specific inhibitors.

In the present work, we have used DHFR as an example to illustrate a novel methodology which we have developed for the description of the differences between two biomolecules. The basis of this methodology has already been applied to the study of specificity in DNA–drug interaction.[15] As a generally applicable technique it needs (a) to be objective, (b) needs to be relevant, and (c) should unveil the different docking between the ligands and the targets and the potential regions on the target enzymes with highly selective interactions.

The objective of this methodology is to multivariately characterize the ligand–macromolecule interactions in order to identify the most selective chemical groups which could be incorporated in the new ligand and also

® Abstract published in *Advance ACS Abstracts*, October 15, 1995.

0022-2623/95/1838-4637$09.00/0

Table 1. Available Dihydrofolate Reductase (EC 1.5.1.3) Structures in the Brookhaven Data Bank

entry	species	complexed	mutant form	resolution (Å)
3dfr	*Lactobacillus casei*	MTX·NADPH		1.7
4dfr[a]	*Escherichia coli*	MTX		1.7
5dfr	*Escherichia coli*			2.3
6dfr	*Escherichia coli*	$NADP^+$		2.4
7dfr	*Escherichia coli*	folate·$NADP^+$		2.5
8dfr	chicken liver			1.7
1dhf[b]	recombinant human	folate		2.3
1dhi	*Escherichia coli*	MTX	Asp-27-Ser	1.9
1dhj	*Escherichia coli*	MTX	Asp-27-Ser, Phe-137-Ser	1.8
2dhf	recombinant human	5-deazafolate		2.3
1dr1	chicken liver	biopterin·$NADP^+$		2.2
1dr2	chicken liver	oxidized thio-$NADP^+$		2.3
1dr3	chicken liver	biopterin·oxidized thio-$NADP^+$		2.3
1dr4	chicken liver	biopterin·$NADP^+$	Cys-11[c]	2.4
1dr5	chicken liver	$NADP^+$	Cys-11[c]	2.4
1dr6	chicken liver	biopterin·$NADP^+$	Cys-11[d]	2.4
1dr7	chicken liver	$NADP^+$	Cys-11[d]	2.4
1dra	*Escherichia coli*		Asp-27-Glu	1.9
1drb	*Escherichia coli*		Asp-27-Cys	1.9
1drf	recombinant human	folate		2.0
2drc	*Escherichia coli*		Trp-22-Phe	1.9
3drc	*Escherichia coli*			1.9

[a] Structure selected to represent bacterial DHFR. [b] Structure selected to represent human DHFR. [c] Cys-11 modified with methyl mercury. [d] Cys-11 covalently bound to mercuriobenzoate.

to exploit the most relevant regions of the target macromolecules for selective interactions.

Briefly, the methodology involves five major steps: (1) Obtain adequate three dimensional structures for each of the target macromolecules. (2) Produce a superposition of the regions in which ligands are know to interact (binding site). (3) Obtain the multivariate characterization of the binding site by the energies of interaction between the target macromolecules and different small chemical groups, from the GRID force field. (4) Rationalize the results from GRID by means of principal component analysis (PCA). (5) Carry out graphical analysis and chemical interpretation of the results of the PCA, aiming at the design of new compounds.

Methods

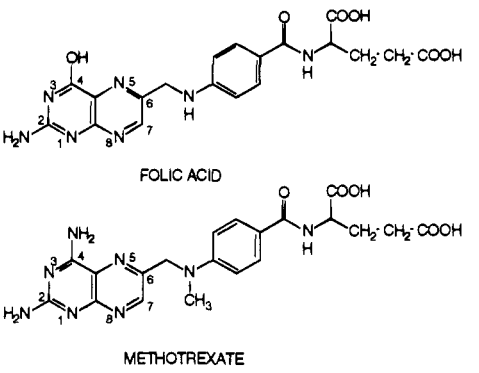
Selection of the DHFR Structures. At the time this work was done there were 22 available structures of DHFR in the Brookhaven National Laboratories Protein Databank (PDB). The details concerning each structure are shown in Table 1.

In this study we look for a comparison between the bacterial and human varieties of DHFR. Therefore we preselected only two structures, one for *E. coli* and the other for recombinant human DHFR, as similar as possible. Multiple comparison would also be possible and may be more informative, but with the risk of introducing more noise, thus leading to much more confusing results.

The use of binary complexes with substrate or inhibitors would be preferable because it can be expected that the conformations of the dihydrofolate binding site in complexed structures are more similar to the conformation occurring during ligand interaction. Another reason supporting this preference is that the ligand molecule present in the structure can be helpful for assessing the correct superposition of both enzymes and the extent of the substrate binding site. The ternary complexes were not taken into account because the influence of the cofactor binding over the binding site conformation is not clear.[5,7] Moreover, the mutant forms were rejected from the analysis in order to produce more consistent results.

Obviously, binary complexes of bacterial and human DHFR with the same ligand would have been the optimum choice. Unfortunately, under the above constraints such structures are not available. The best alternatives were the complex of bacterial (*E. coli*) DHFR with the inhibitor MTX and the

Chart 1



complex of recombinant human DHRF with the substrate derivative folate (Chart 1). Both ligands are known to interact at the same place[6] and exhibit important structural similarities, so the conformational differences coming from the differences of ligands can be expected to be minimal. Among the available structures for these complexes, we chose the PDB entries with higher resolution, 4dfr[13] and 1dhf[4] as representatives of bacterial (*E. coli*) and recombinant human DHFR, respectively. The 4dfr entry describes an asymmetric unit containing two molecules; the molecule designated as B was used because it is more complete and less perturbed by intermolecular contacts. These will be referred to as ecDHFR and rhDHFR in the rest of this paper.

The selected structures were downloaded from the PDB database and imported into the SYBYL[16] molecular modeling program where minor modifications were performed. In particular, water and ions were removed from both proteins.

Superposition of the Substrate Binding Sites. In this paper we focus our attention in the substrate binding site, a deep hydrophobic pocket which bisects the protein and which is distinguished by the presence of a strictly conserved acidic residue (Asp-27 in the ecDHFR and Glu-30 in rhDHFR). In all known complexes the substrate or inhibitor is held in the active site by a network of hydrogen bonds in which the acidic residue plays a central role. The acidic residue also makes different interactions with some residues in the binding site which differ from one ligand to another.

Therefore, in the context of this work the goal of the superimposition is not to compare the whole protein but only the substrate binding site where ligands are known to interact. As a test, if the operation is correct, the ligands MTX or folate included in the structures will be superimposed as well.

Table 2. Set of Key Active Site Amino Acids (See Refs 5, 7, and 10) for Bacterial and Human DHFR

ecDHFR[a]	rhDHFR[a]	element of secondary structure
Ile-5	Ile-7	βA
(Asp-27)	(Glu-30)	αB
(Leu-28)	(Phe-31)	αB
Phe-31	Phe-34	αB
Ile-50	Ile-60	αC
Leu-54	Leu-67	loop αB–αC
Ile-94	Val-115	βE
Thr-113	Thr-136	βF

[a] The residues in parentheses were not included in the set used for the superposition of human and bacterial DHFR (see text).

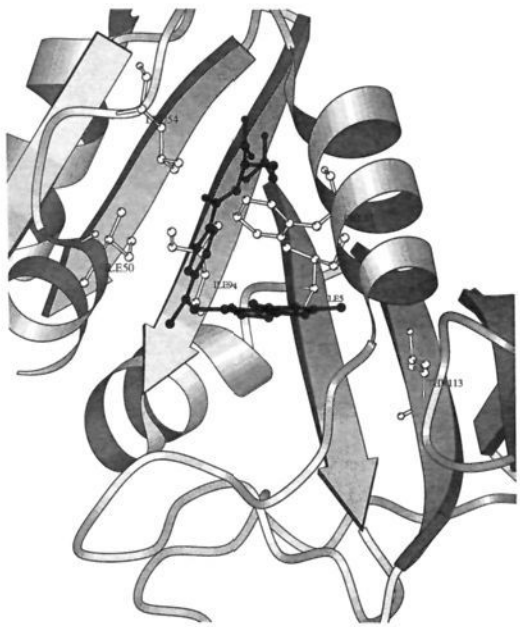


Figure 1. Schematic representation of the backbone and the main secondary structure elements of *E. coli* DHFR arround the binding site, showing a bound methotrexate molecule (in dark gray) and the side chains of the six residues (in white) used to superimpose it with recombinant human DHFR.

In such a well-characterized system it was not difficult to find previous studies in which the substrate–protein and inhibitor–protein were discussed in great detail, even comparing the interspecie's selectivity.[4–10,13] These studies gave different sets of conserved residues that interact with substrates or inhibitors and represent the parts of the binding site involved in the ligand recognition. From them we focused our attention on the set of eight key active site amino acids mentioned in refs 5, 7, 10 and shown in Table 2. This set has been used by different authors to compare the active site of bacterial DHFR with either chicken liver DHFR[5,7] or recombinant human DHFR.[10] The set includes residues that are known to interact with MTX[13] and TMP; these residues were originally selected to study the relative geometric relationships between different species DHFR active sites. It is especially well suited for our purposes, because the residues were chosen in such a way that they properly represent the different secondary structural elements that form the active site (see Figure 1 and Table 2).

For the superimposition purpose only the α-carbons of the aforementioned residues were initially considered. However, it should be noted that the polar amino acid, the most relevant residue for the emplacement of the substrate in the active site, is different in bacterial and vertebrate DHFR, being aspartic (Asp-29,ec) for *E. coli* and glutamic (Glu-30,rh) for humans. Although in both cases it is functionally equivalent, the length of the side chain is longer for the human variety of the enzyme and it follows that superimposing the α-carbons might lead to wrong results. The binding sites were finally superimposed using only the backbone positions of six residues, removing from the set the Asp-27,ec (Glu-30,rh) and also the neighbor Leu-28,ec (Phe-31,rh). Other possible approaches, *i.e.* superimposing the carboxyl groups of the polar residues, were tested, leading to quite similar results.

The α-carbons of these six residues were fitted by the least-squares method resulting in the superimposition of the substrate binding site and of the ligands, shown in Figure 2. The rms deviation was 0.03 Å for the 6 α-carbons and the largest deviation of their backbone positions was only 0.57 Å between Leu-54,ec and Leu-67,rh.

Multivariate Characterization of the Substrate Binding Site. The GRID Force Field. The next step is the description of the ligand interactions with the substrate binding sites. The GRID program[17–19] was used to calculate the energetic interactions of both enzymes (targets) with a large number of different small chemical groups (probes).

Basically GRID is a computational procedure for detecting energetically favorable binding sites on molecules. The energies are calculated as the electrostatic, hydrogen bond and Lennard-Jones interactions of chemically selective probes with the chosen targets. The method of images[17] is used by GRID in order to account for the dielectric influence of solvent water, since unmodulated electrostatic interactions would give misleading results in the absence of explicit water molecules. The program works by defining a three-dimensional grid of points

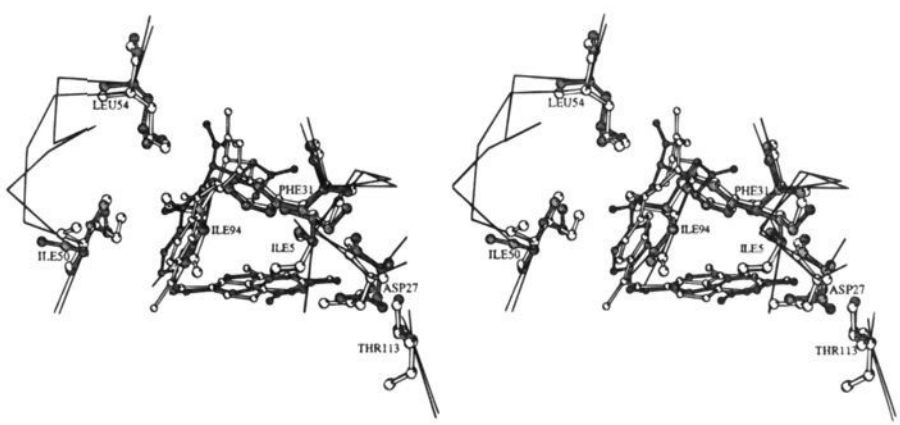


Figure 2. Stereodrawing of the binding sites of *E. coli* DHFR and recombinant human DHFR after the superimposition operation described in the text. The picture shows the side chains of the six pairs of residues used for the superimposition, listed in Table 2, plus those of the acidic residues Asp-27,ec and Glu-30,rh. The side chains of *E. coli* DHFR were represented by open bonds and atoms while the side chains of recombinant human DHFR were representd by darkened bonds and atoms. The ligands were represented with thiner bonds and atoms. Open thin bonds represent the methotrexate molecule bound to *E. coli* DHFR cristal structure. Darkened thin bonds represent the folate molecule bound to recombinant human DHFR.

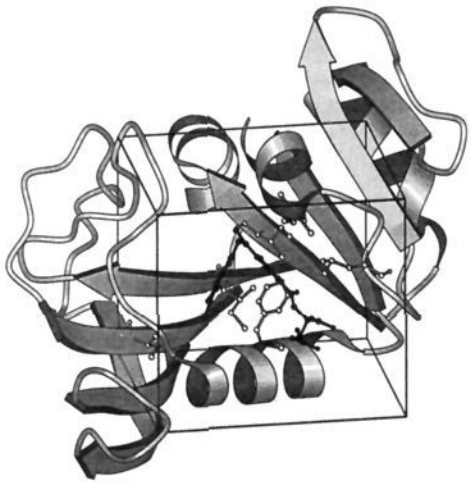


Figure 3. View of the 18 × 18 × 20 Å cage enclosing the binding site in which the GRID runs were carried out superimposed over a schematic view of *E. coli* DHFR.

that contains the chosen substrate binding site. At each node of the grid, the energy between probe and target is calculated as indicated in eq 1

$$E_{xyz} = \sum E_{EL} + \sum E_{HB} + \sum E_{LJ} \qquad (1)$$

where E_{EL} is the appropriately modulated electrostatic energy, E_{HB} the hydrogen-bonding energy, and E_{LJ} the Lennard-Jones potential energy between the constituent atoms of the probe and all the atoms of the target. The same calculation is repeated for each node in the network and for each probe considered. The results of these calculations are a collection of three-dimensional matrices, one for each probe–target interaction. A detailed description of the GRID program, the force field parameters, and details of calculations are beyond the scope of this work and can be found elsewhere.[17-19]

The global charge of the enzyme target was 0 for the rhDHFR and −9 for the ecDHFR. Since uncharged systems are preferable for the analysis, potassium counterions where added to the 4drf structure. To perform this task a GRID map for potassium cations was prepared, and the utility programs MINIM and FILMAP, which are included in version 12.0 of the GRID program,[20] were then used in tandem. The first program MINIM finds and lists all the energy minima in the grid map for potassium counterions. Then, the program FILMAP used a simulated annealing procedure to postprocess the list of minima in order to select the minima location subset that gives the most favorable overall interaction energy.

The selection of an adequate grid size and location is crucial for the success of the analysis. In particular it should enclose all the positions around the substrate binding site in which atoms of a potential ligand could be found. A GRID run was therefore performed using a simple C3 probe, and the SYBYL software was used to visualize the GRID surface defined by interaction energies of +0.2 kcal/mol. The cage size was then selected in such a way that it fully enclosed the part of this surface immersed in the hydrophobic binding pocket, resulting in the 18 × 18 × 20 Å cage shown in Figure 3. The main GRID runs were then carried out with a grid spacing of 1 Å.

Nonbonded interaction energies between each protein and the 41 probes shown in Table 3 were calculated. The list of probes includes 32 monoatomic and 9 polyatomic chemical groups and is quite comprehensive, since we were looking for an exhaustive description of any possible ligand–target interaction.

Matrix Generation and Pretreatments. The three-dimensional matrices obtained from GRID were rearranged as one-dimensional vectors. One such vector is obtained for each probe–target interaction, and the vectors were used to build a two-dimensional **X** matrix, in which the rows are the probe–target interactions (the objects) and the columns are the variables that describe energetically these interactions. This matrix contains 82 rows (41 probes × 2 targets) and 6480 columns (18 × 18 × 20 nodes). The process used to obtain the **X** matrix is illustrated in Figure 4

Table 3. Table of Probes Used in GRID for the Binding Site Analysis

no.	code	description
1	C1=	aromatic CH group
2	C3	methyl group
3	N:	sp^3 nitrogen with lone pair
4	N:=	sp^2 nitrogen with lone pair
5	N-:	anionic nitrogen of tetrazole
6	N:#	sp nitrogen with lone pair
7	N1	amide NH group
8	N1#	sp NH group eg. acetylene
9	N1=	sp^2 cationic NH group
10	NH=	sp^2 NH group with lone pair
11	N1:	sp^3 NH group with lone pair
12	N1+	sp^3 NH cation
13	N2	amide NH_2 group
14	N2=	sp^2 cationic NH_2 group
15	N2:	sp^3 NH_2 group with lone pair
16	N2+	sp^3 cationic NH_2 group
17	N3+	sp^3 cationic NH_3 group
18	NM3	trimethylammonium cation
19	OC2	ether oxygen aton
20	OES	ester oxygen atom
21	O	carbonyl oxygen atom
22	O::	carboxy oxygen atom
23	ON	nitro oxygen atom
24	O=	phosphate oxygen atom
25	O-	anionic phenolate oxygen atom
26	OS	oxygen of sulfone or sulfoxide
27	O1	aliphatic hydroxyl group
28	OH	phenolic hydroxyl group
29	F	fluorine atom
30	CL	chlorine atom
31	BR	bromine atom
32	I	iodine atom
33	COO-	ionized alkyl carboxyl group (multiatom)
34	AR. COO-	ionized aryl carboxyl group (multiatom)
35	CONH2	alkyl amide (multiatom)
36	AR. COHN2	aryl amide (multiatom)
37	CONHR	alkyl *N*-alkylamide R CONHR (multiatom)
38	AR. COHN2	aryl *N*-alkylamide AR.CONHR (multiatom)
39	AMIDINE	alkylamidine $R.C(NH_2)_2$ (multiatom)
40	AR. AMIDINE	arylamidine $AR.C(NH_2)_2$ (multiatom)
41	M-DIAMINE	*m*-diaminobenzene (multiatom)

[a] Probes 1–32 are single-atom probes. Probes 33–41 are multiatom probes.

Any positive interaction energy present in the **X** matrix was then set to 0 kcal/mol. The use of only the negative, favorable interaction energies has the advantage of removing part of the information related with steric interactions. In fact, as steric interactions usually map small protein shape differences, the removal of the positive interaction energies focus the work only on favorable ligand–enzyme interactions.

All the data were centered by subtracting from each column the column average. Autoscaling was not applied, since all the data come from the same source (GRID probe–target interaction energies) and all the data are expressed in the same units (kilocalories per mole). In this context, autoscaling might introduce noise in the model, increasing the importance of variables with small variance.

Principal Components Analysis. All the information describing the probe–target interaction is contained in the **X** matrix. However, such information is hidden in the rows and in so many columns that no useful information can easily be extracted. In order to simplify this matrix and to obtain an informative picture of the data structure we applied PCA. A

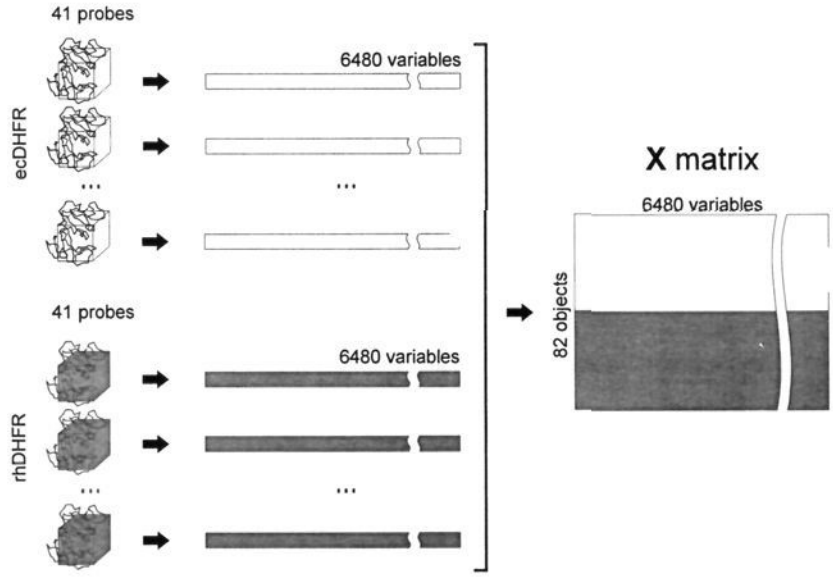


Figure 4. Procedure used for building the **X** matrix. The analysis of the interaction energies of the 41 probes described in Table 3 with the 2 target molecules (ecDHFR and rhDHFR) produced 82 three-dimensional matrices. Then, they were unfolded to obtain 82 one-dimensional vectors, from which the two-dimensional **X** matrix was built.

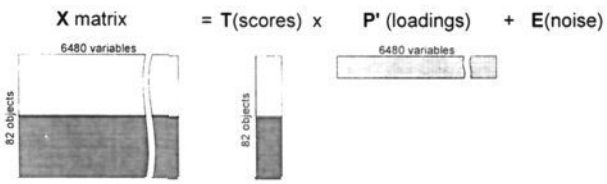


Figure 5. PCA was used for the chemometrical analysis of the matrix **X**. The PCA model provides an approximation of a data matrix **X** in terms of the product of two small matrices; the matrix of objects (**T**, scores matrix) and the matrix of variables (**P′**, loading matrix). The matrix **E** contains the variance not explained by this model, that can be regarded as noise and has the same metric as the **X** matrix.

detailed discussion of the theoretical background of PCA has been reviewed elsewhere.[21,22] Briefly, PCA provides an approximation of a data matrix **X** in terms of the product of two small matrices **T** (score matrix) and **P′** (loading matrix), as shown in Figure 5.

The score matrix gives a simplified picture of the objects (probe−target interactions), represented by only a few, uncorrelated new variables (the so-called principal components, PCs) that explain most of the variation contained in the original matrix. Score plots (plots of the objects in the PC's space) are used to reveal the essential data patterns of the objects, and in the context of this work, they can display clusters of objects according to the different kind of targets (macromolecules) and probes (ligand chemical groups) involved. On the other hand, the loading matrix reveals the relation between the original variables and the new PCs. Loading plots are useful to discover the relation between the original variables and the PCs and, in the context of this work, provide an interpretation in terms of the binding site regions that contain the variables most related to each PC.

The previously described **X** matrix was analyzed by PCA as implemented in the GOLPE[23] program. The variance explained for the first components is reported in Figure 6.

It is appropriate to stress that statistical significance and chemical significance are two different concepts. In fact, from a pure statistical meaning, all the components reported in Figure 6 are significant. However, from a chemical point of view the number of useful PCs is practically limited by our ability to interpret their meaning. In the context of this work, only the two first PCs reflect the general variance patterns of the set. The subsequent ones are devoted only to explain

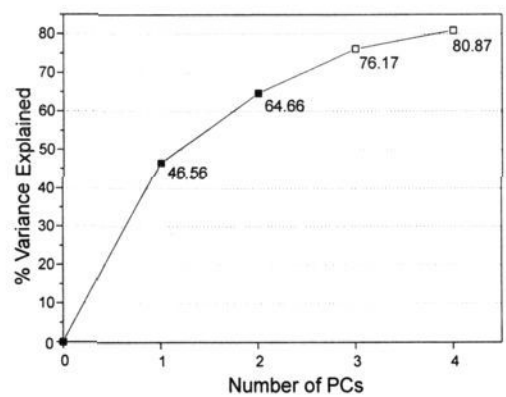


Figure 6. Acumulative percentage of variance explained by the PCA model as a function of the number of PC's considered.

singular points and small deviations from the general behavior, the interpretation of which is worthless if not impossible.

Results

Score Plots. The score plot for the PC model of the **X** matrix is shown in Figure 7. In this plot the points represents single probe-target interactions (objects). The interactions with human DHFR are represented as filled points (◆) while the interactions with bacterial DHFR appear as open points (◇).

The figure clearly shows that PC 1 distinguishes between the two target proteins, clustering the objects into two groups, while PC 2 ranks the probes. This ability of the first PC to discern between the probe−target interactions involving ecDHRF and rhDHFR can be exploited in several ways. First, variables with high PC 1 loadings will delimit regions on the binding site where the probes show an extremely different behavior in their interaction with the human and bacterial enzymes. As only negative (favorable) interactions are considered, these regions will reveal positions where a chemical group can bind loosely with one of the targets and tightly with the other. Also the scores of this first PC are useful, because the score absolute values are related to the ability of the represented groups to

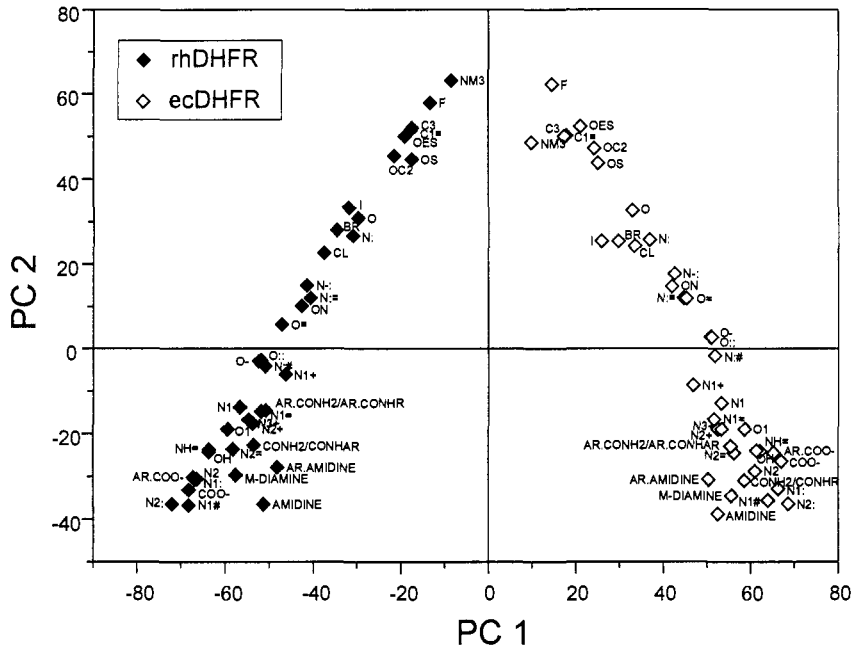


Figure 7. PC 1 *vs* PC 2 score plot (t_1 *vs* t_2). The points in this plot represent the *objects* of the **X** matrix: the interactions of a given probe with a given variety of DHFR. Filled points (◆) represent probe–rhDHFR interactions. Open (◇) points represent probe–ecDHFR interactions.

establish selective interactions. Therefore, for the design of selective compounds it would be preferable to insert the chemical groups with higher absolute PC 1 score values.

On the other hand, PC 2 is related with nonselective ligand–target interactions. This PC ranks the probes according to their ability to interact chemically with common regions of the binding site. Regions with high PC 2 scores highlight areas where the probes interact with the same strength for *both targets*. It should be noticed that because of the favorable binding results on negative energies there is an inverse relationship between the strength of the binding and the PC 2 scores. Accordingly the probes with less ability to interact with common regions of the targets are in the top part of the score plot in Figure 7, and the probes that interact in a stronger way (including all the multiatom probes) are in the bottom part. Moreover, it can be seen how the points spread from top to bottom, showing that the probes which interact strongly with common regions of the targets are also the most interesting from the point of view of selectivity.

To summarize all these findings: selective probes are placed at the bottom right and bottom left areas of Figure 7 and have high absolute PC 1 values and negative PC 2 values. For example, when looking for substituents to include in a novel compound, groups such as sulfone and sulfoxide (represented by probe OS) cannot be regarded as good choices for increasing selectivity, because of the low absolute values for PC 1 shown in Figure 7. Neither can they be expected to increase the nonspecific affinity of the compound for the enzyme, because of the positive values of PC 2 scores in their interaction with both targets. Instead, groups represented by probes at the bottom left and right areas of Figure 7 (*e.g.* primary amines represented by probe N2:), when properly placed on the receptor site, can potentially increase the selectivity of the interaction toward the human or bacterial variety of the enzyme. Moreover, these groups exhibit negative PC 2 scores and can also be placed on common parts of the receptor to increase the nonspecific affinity of the compound for the enzyme.

Two-Dimensional Loading Plots. Figure 8 shows a loading plot of the PCA model. In this plot the points represent the contribution to the PCs of each position in the lattice where probe–target interactions were computed (variables).

On the basis of the previous discussion, the understanding of the loading plot is straightforward. The horizontal axis represents PC 1 loadings and, in general, the greater the horizontal spread of a point, the more relevant is this position in the lattice for the discrimination between the two proteins. The vertical axis represents PC 2 loadings and the points in the top part represent positions where the probes interact in a similar way with bacterial and human enzyme.

According to this interpretation we can distinguish three types of positions on the binding site. (a) Low absolute values for PC 1 loadings and low PC 2 loadings: positions where the probes only establish weak, nonselective interactions. Most of the binding site positions fall into this category. (b) Low absolute values for PC 1 loadings and high PC 2 loadings: positions where the probes interact strongly with both targets. They are not interesting from the point of view of the selectivity but might be exploited to increase the affinity for the target. (c) High absolute values for PC 1 loadings and intermediate PC 2 loadings: positions where the probes establish strong selective interactions. Adequate groups located in these positions would induce or increase the selectivity of a ligand.

It is appropriate to point out that there is no unique criterion on which we can distinguish the three types of variables. In Figure 8, arbitrary boundary levels were defined at −0.03 and +0.03 for PC 1 and at 0.07 for PC 2 just for illustrative purposes.

Loading Contour Maps. These maps are three-dimensional plots representing the important 3D regions highlighted by the statistical model. Loading

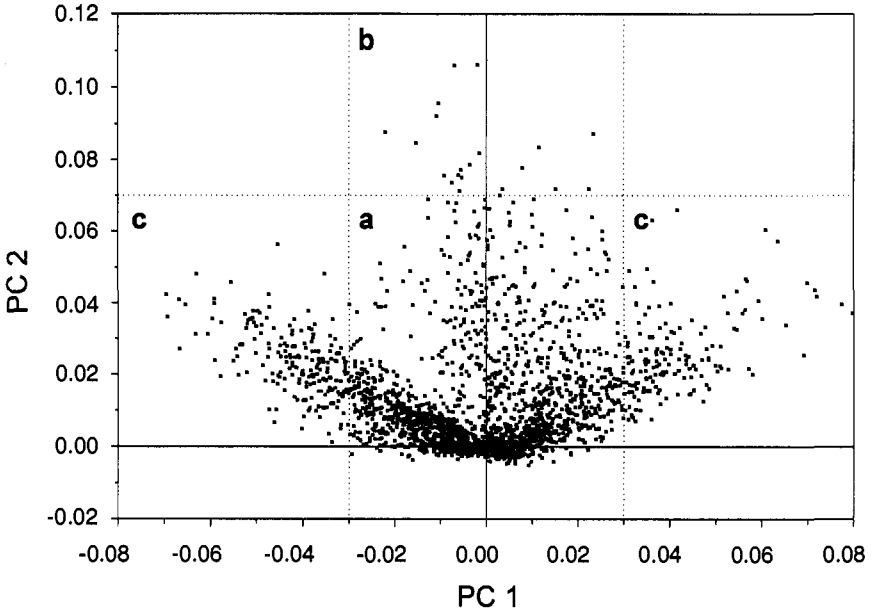


Figure 8. PC1 *vs* PC 2 loading plot (p_1 *vs* p_2). The points in this plot represent the *variables* of the **X** matrix: the positions in the grid space. The dotted lines define three areas (a, b, and c) in the plot, where the three types of positions referenced in the text as a, b, and c can be found.

Table 4. Main Regions in the Contour Plot of PC 1 Loadings Involved in Selective Ligand Recognition

name[a]	location	neighboring residues in ecDHFR	neighboring residues in rhDHFR
bottom	loop βA–αB	Ala-19, Met-20	Asp-21, Leu-22
middle	αB	Leu-28	Phe-31
top	loop αC–βC	Ile-50, Gly-51, Arg-52, Pro-53	Ile-60, Pro-61, Glu-62, Lys-63, Asn-64, Arg-65, Pro-66

[a] The name makes reference to the position in the binding site when oriented as in Figure 1.

contour maps are useful for identifying the regions in the active site which interact strongly or selectively with ligands. Hence, by selecting the appropriate PC and contour level it is possible to display the regions of the binding site most relevant for selective binding and common affinity. These contour maps have an obvious interest for the design of novel compounds, as they identify positions in the space where the localization of an appropriate chemical group would lead to an increase of the desired properties.

Areas Involved in Ligand Discrimination. Selectivity Regions. The regions involved in ligand discrimination were displayed by means of contour maps of PC 1 loadings. From the inspection of the loading plot (Figure 8), contour levels of 0.03 and −0.03 were arbitrarily chosen. Lower levels would include too many variables and produce too confused contoured regions, while a higher level might result in the elimination of variables containing useful information.

In order to simplify the picture, we have focused our attention on the three most interesting regions of the plot. They will be referred to as bottom, middle, and top regions throughout this work, according to their position in the binding site when it is oriented as in Figure 1. Residues and elements of secondary structure associated to each region are listed in Table 4.

As a consequence of the negative sign assigned to favorable interaction energies, the loading signs are somewhat inverted. Areas with negative loadings express the fact that in these areas the interactions are favorable with targets in the positive part of the score plot, and vice versa. This explains why the positive contoured areas (in black in Figures 9–11) highlight regions where the probes interact selectively with rhDHFR, while the negative areas (in gray in Figures 9–11) highlight regions where the probes interact selectively with ecDHFR.

a. Bottom Region. The bottom region, represented in Figure 9, is located in the lower part of the substrate binding site, in the vicinity of the loop connecting βA to αB. The importance of this loop for ligand selectivity has been known for a long time, and it is surprising how PCA highlights this region without any external information. According to some authors the residues in this loop are responsible for the different site width on which depends the species selective interaction of TMP.[12] Also, it is known that cofactor binding dramatically increases the species selectivity of TMP. As this loop is a region of major importance for cofactor binding, the cooperative effect would be explained as a consequence of the conformational changes in the loop accompanying the NADPH binding.[12]

The coordinates of the α-carbons on this loop exhibit large differences between the bacterial and human enzyme. These differences produce the large gray area on the left of Figure 9. Despite its size, this area may be not too relevant, because the positions of the backbone atoms in the loop are highly sensitive to conformational changes and also because it is placed in the outermost part of the binding site. Much more interesting are the positions contoured in black, in the top right corner of Figure 9. This region encloses positions where the probes can interact by hydrogen bond with the carboxylic side chain of Asp-21,rh, while there is no such residue in the bacterial enzyme.

It is remarkable that, even when the position of some side chains exhibit large differences (*e.g.* Met-20,ec and Leu-22,rh), no contoured area can be observed in their

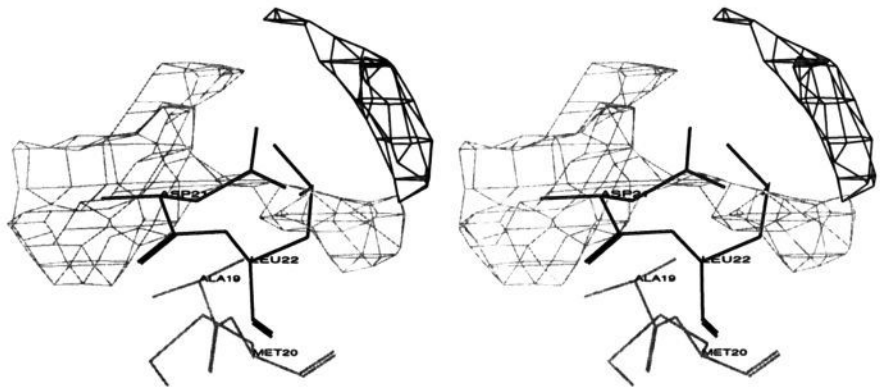


Figure 9. Bottom region of the DHFR binding site. Contour map (in stereoview) of the PC 1 loading contoured at −0.03 (gray) and +0.03 (black). The regions represent the positions at which the probes would interact most selectively with the enzymes. Color scheme: negative loading (gray thin lines), positive loading (black thin lines), residues of ecDHFR (gray thick lines), residues of rhDHFR (black thick lines). For interpretation, see text.

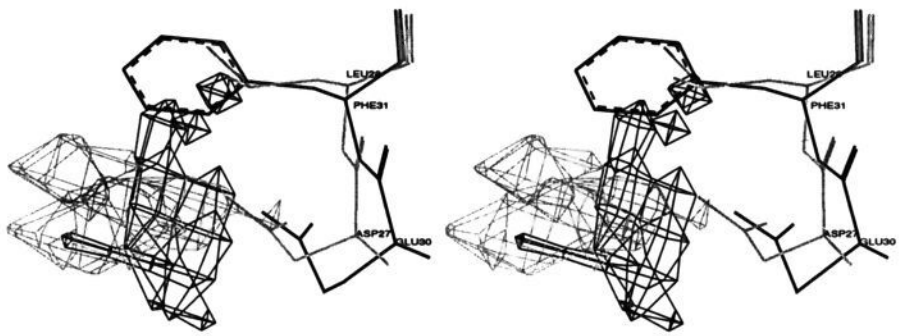


Figure 10. Middle region of the DHFR binding site. Contour map (in stereoview) of the PC 1 loading contoured at −0.03 (gray) and +0.03 (black). The regions represent the positions at which the probes would interact most selectively with the enzymes. Color scheme: negative loading (gray thin lines), positive loading (black thin lines), residues of ecDHFR (gray thick lines), residues of rhDHFR (black thick lines). For interpretation, see text.

vicinity. This accounts for the fact that these residues were not involved in strong favorable probe−target interactions, although the importance of these residues is not negligible and could play a role through steric contacts. From the point of view of favorable interactions, the differences in the positions of their carbonyl groups seems more relevant, and actually these differences generate the small gray area represented on the right part of Figure 9.

As it concerns the design of selective compounds, the most relevant region is the above-mentioned area neighboring the Asp-21,rh. The introduction of groups binding Asp-21,rh would lead to more selective human inhibitors. However, only selective bacterial inhibitors are of therapeutic interest.

It should be further emphasized that the aforementioned regions, plotted in Figure 9, were obtained by contouring the loadings for the first PC and *not* the energies of interaction. Therefore, their meaning is directly bound to the selectivity of the ligand interaction between the enzymes, and no subjective comparison was required.

b. Middle Region. The middle region is placed in the αB, deeply buried in the hydrophobic pocket and near the acidic residues Asp-27,ec and Glu-30,rh. As can be seen in Figure 10, the contoured areas are produced by the substitution of Leu-28,ec by Phe-31,-rh. Both residues exhibit hydrophobic side chains, and the main differences came from the different size and orientation of the areas accessible to favorable probe−target interactions. The contoured regions highlight with great detail the differences between the two side chains; the two-lobuled gray region on the left of Figure 10 delimits the positions where the probes can favorably interact with the aliphatic side chain of Leu-28,ec but not with the phenyl ring of Phe-31,rh. The black region on the right of Figure 10 contours positions favorable for the interaction with Phe-31,rh but not with Leu-28,-ec. However, we should bear in mind that this method does not consider the conformational freedom of the side chains, and small changes (*e.g.* a rotation of the Phe-31,rh phenyl ring) would change the picture completely.

The main interest of this region comes from its proximity to the aforementioned acidic residues, which are known to be bound by all the ligands and inhibitors. When designing new inhibitors it would be easy to find substitutions that fall into this area, and exploit their different interactions with Phe-31,rh and Leu-28,ec to improve the inhibitor selectivity. However, no strong interactions can be expected and the effects of conformational changes should be carefully examined.

c. Top Region. The top region is placed by the loop that connects the αC to the βC, adjacent to the insertion point III[12] where the human enzyme includes three extra residues. It is remarkable that even when the number of residues is different and the differences in the position of the α-carbons are large, the probe−target interactions are not as different as might have been.

The side chains of the three human extra residues, represented on the left in Figure 11, exhibit side chains shorter than the equivalent residues in the bacterial enzyme. Therefore, the Lys-63,rh and Asn-64,rh side

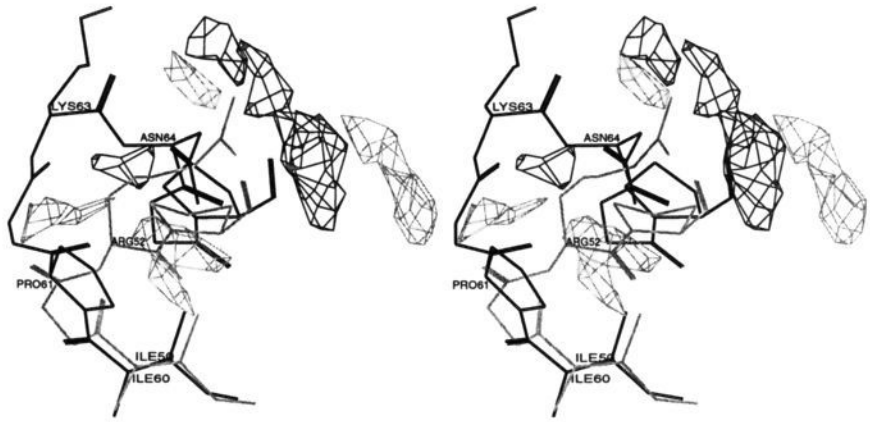


Figure 11. Top region of the DHFR binding site. Contour map (in stereoview) of the PC 1 loading contoured at −0.03 (gray) and +0.03 (black). The regions represent the positions at which the probes would interact most selectively with the enzymes. Color scheme: negative loading (gray thin lines), positive loading (black thin lines), residues of ecDHFR (gray thick lines), residues of rhDHFR (black thick lines). For interpretation, see text.

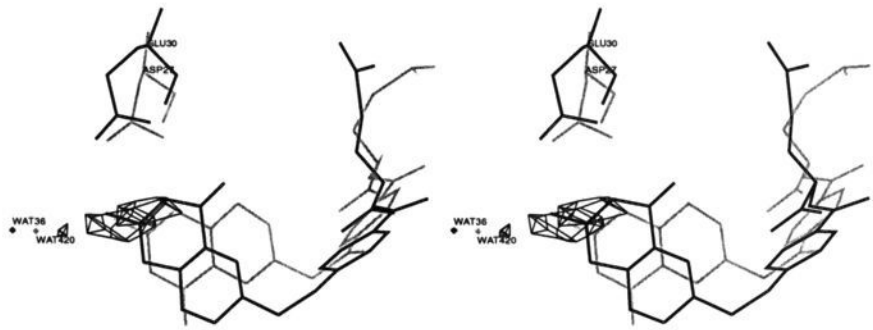


Figure 12. Contour map (in stereoview) of the PC 2 loading contoured at +0.07. Color scheme: positive loading (black thin lines), MTX and residues of ecDHFR (gray thick lines), folate and residues of rhDHFR (black thick lines). For interpretation, see text. Notice that the ligands represented in this figure (MTX in gray, and folate in black) as well as the water molecules were not considered for the GRID analysis of the binding site.

chains are not far from the position occupied by Arg-52,ec side chain. A set of significant contours may be observed at the upper right corner of Figure 11. The elongated gray area on the right edge and the smaller gray one on the top center represent areas where the probes can interact with the bacterial residue Arg-52,-ec and not with the human enzyme. On the other hand, the large black area between these two seems to correspond to areas where probes can interact with the side chain and the backbone carbonyl group of Asn-64,-rh and not with any residue of the bacterial enzyme. Taking into account the flexibility of the side chains, it appears that, by slight conformational changes, the Arg-52,ec can be placed in positions equivalent to the Asn-64,rh.

Another interesting contoured area is produced by the different orientation of the Ile-50,ec and Ile-60,rh backbone carbonyl groups that appear as a consequence of the Pro-61,rh tight turn. The two gray areas shown in the center of Figure 11 mark positions where the probes can establish hydrogen bonds with the lone pairs of the Ile-50,ec backbone carbonyl group, while no similar interaction can be established with the human enzyme. The reduced conformational freedom of the backbone positions, compared with the residue side chains, makes this area a promising target for groups aiming at selective binding with the bacterial enzyme.

Areas Involved in Ligand Recognition. As previously stated, PC 2 explains the different abilities of the probes to favourably interact with common parts of the enzyme. Therefore, the contour map of the PC 2 loadings would highlight these common parts where ligands can bind both the bacterial and human varieties of DHFR.

Figure 12 represents a contour map of PC 2 loadings at the arbitrary cutoff level of +0.7. The ligands present in the crystal structures (not used in the GRID calculations) were included in the figure to show how the volume contoured contains some heteroatoms, common to both ligands. These atoms (the 4-N and 3-amino group for MTX and the 3-N and 2-amino group for folate) appear in the crystal bonded to the acidic residues (Asp-27,ec and Glu-30,rh) which hold the substrate in the right orientation. The contoured area also contains positions in the vicinity of the conserved water molecules Wat-420,ec and Wat-36,rh, which are also known to play an important role in the hydrogen bond network. Bearing in mind that neither the water nor the ligands were considered in the analysis, these results are in surprisingly good agreement with experimental observations.

From the point of view of ligand design, the strong interactions present in these positions should be used to produce high-affinity ligands, while being aware that the affinity is being increased for both targets and that no selectivity can be obtained from substituents placed here. An interesting possibility would be to place in the positions occupied by fixed water molecules some chemical groups that can make more efficient hydrogen bonds with the surrounding residues.

Discussion

The ligand–protein interaction is a phenomenon extremely difficult to describe. Nowadays, no approach provides a complete picture of all the forces involved. The reported method is aimed only at giving partial, but useful, answers to a few particular questions: what substituent would improve the selectivity of a given compound and where to place this substituent? This method is intended as a tool that would provide some clues about favorable structural features in newly designed ligands and it by no means pretends to replace deeper, but harder, ligand–protein insights obtained through a detailed rationalization of the structural data and site-directed mutagenesis.

Actually, the method can be applied only to compare two or more target molecules for which three-dimensional structures are available. An even more restrictive requisite is that the targets should be similar enough to permit a rational superimposition of the binding sites. The importance of this step for the success of the whole process is critical, and a poor superimposition would lead to nonsense results. However, the design of selective compounds is difficult only when the proteins are very similar, and in this case there is no trouble in superimposing the targets.

It is important to be aware of the simplifications introduced in the method:

(a) The method considers the targets as static entities and does not consider the conformational freedom of the backbone and the side chains. However, the effect of the orientation in the protein–probe interactions is accounted for to a certain extent by GRID,[17] and it is also possible to introduce some chemical expertise in the rationalization of the results, rejecting the areas in which any apparent diversity of the targets comes only from slight conformational differences.

(b) In the example all the water was removed from the target protein structures. However, very often water molecules play an important role in the enzyme and might be considered constitutive of the protein structure.[17] In such cases, the final results and the success of our method may depend on keeping certain water molecules in the protein structure. The decision of which water molecules should be kept depends mainly on external hints, such as the high-occupancy and low-temperature factor in the crystallographic refinement or previous knowledge about its function in the protein.

(c) Only enthalpy is considered, but entropy is also known to be a determinant for the understanding of the protein–ligand interactions. However, this method is intended mainly to identify strong selective interactions and in this context the enthalpy is more important that entropy by far.

Nevertheless, the methodology described here is the first one that uses chemometric methods in order to deal with the problem of selectivity in the design of novel compounds. From the chemometric point of view it is based on PCA, which is one of the best known techniques in this area and has been extensively reviewed from a general point of view[21] and in its application to chemical problems.[22] Also, the whole method can be carried out in a very simple way, with the help of readily available software,[20,23] and it is not expensive in terms of computer time or computer resource requirements.

The DHFR has been chosen as the subject of this method only for testing purposes. Despite the great advantage of using such a well-known system, the binding site of this enzyme is mainly hydrophobic which makes the example a difficult test for a method that favors strong (electrostatic and hydrogen bond) interactions. Therefore it is surprising that the GRID/PCA method gives results consistent with others reported in the literature, thus confirming its general validity.

Conclusions

Despite its limitations, the GRID/PCA method has proved useful for extracting relevant information from three-dimensional structures. In particular, it gave interesting results identifying the areas in the binding site where selective interactions can be achieved.

The detailed analysis of all the possible residue–ligand interactions is a difficult task, and when comparing two or more targets, the evaluation of the quantitative importance of different effects is sometimes subjective. This method permits us to focus our attention on well-defined regions and provides objective information about their relative importance. Moreover, it has been developed for practical purposes and the results can be directly applicable for ligand design.

On the basis of this method, some interesting areas for selective inhibitor binding have been identified.

Experimental Section

Proteins were manipulated and displayed using SYBYL programs.[16] All the probe–target energy interactions were calculated using the GRID program.[20] For data pretreatment and principal components analysis, the GOLPE program[23] (version 2.1) was used. All calculations and displays were performed on UNIX workstations.

Acknowledgment. M.P. acknowledges the Consejo Social of the Universidad de Alcala, Alcala de Henares, Spain for a grant supporting his contribution to this work. G.C. acknowledges the Italian Government (MURST) and the National Research Council (CNR) for grants. The authors want also to thank Dr. Per J. Kraulis for his software Molscript v1.4,[24] used for the generation of Figures 1, 2 and 3 in this article.

References

(1) Kuntz, I. D. Structure-Based Strategies for Drug Design and Discovery. *Science* **1992**, *257*, 1078–1082.
(2) Greer, J.; Erickson, J. W.; Baldwind, J. J.; Varney, M. D. Application of the Three-Dimensional Structures of Protein Target Molecules in Structure-Based Drug Design. *J. Med. Chem.* **1994**, *37*, 1036–1054.
(3) Reich, S. H.; Webber, S. F. Structure-based drug design (SBDD): every structure tells a story. *Perspect. Drug Discovery Des.* **1993**, *1*, 371–391.
(4) Davies, J. F. II; Delcamp, T. J.; Prendergast, N. J.; Ashford, V. A.; Freisheim, J. H.; Kraut, J. Crystal Structures of Recombinant Human Dihydrofolate Reductase Complexed with Folate and 5-Deazafolate. *Biochemistry* **1990**, *29*, 9467–9479.
(5) Freisheim, J. H.; Matthews, D. A. In *Folate Antagonists as Therapeutic Agents Vol. 1*; Sirotnak, F. M., Burchall, J. J., Ensminger, W. D., Montgomery, J. A., Eds.; Academic Press, Orlando, FL, 1984; pp 69–131.
(6) Matthews, D. A.; Alden, R. A.; Bolin, J. T.; Freer, S. T.; Hamlin, R.; Xuong, N.; Kraut, J.; Poe, M.; Williams, M.; Hoogsteen H. Dihydrofolate Reductase, X-ray Structure of the Binary Complex with Methotrexate. *Science* **1977**, *197*, 452–455.
(7) Oefner, C.; D'Arcy, A.; Winkler, F. K. Crystal Structure of Human Dihydrofolate Reductase Complexed with Folate. *Eur. J. Biochem.* **1988**, *174*, 377–385.
(8) Bystroff C.; Oatley, S. J.; Kraut J. Crystal Structures of *Escherichia coli* Dihydrofolate Reductase. The $NADP^+$ Holoenzime and the Folate$NADP^+$ Ternary Complex. Substrate Binding and a Model for the Transition State. *Biochemistry* **1990**, *29*, 3263–3277.

(9) Matthews, D. A.; Bolin, J. T.; Burridge, J. M.; Filman, J.; Volz, K. W.; Kaufman, B. T.; Beddell, C. R.; Champness, J. N.; Stammers, D. K.; Kraut, J. Refined Crystal Structures of *Escherichia coli* and Chicken Liver Dihydrofolate Reductase Containing Bound Trimethoprim. *J. Biol. Chem.* **1985**, *260*, 381–391.
(10) Matthews, D. A.; Bolin, J. T.; Burridge, J. M.; Filman, J.; Volz, K. W.; Kraut, J. Dihydrofolate Reductase. The Stereochemistry of Inhibitor Selectivity. *J. Biol. Chem.* **1985**, *260*, 392–399.
(11) Stammers, D. K.; Champness, J. N.; Beddell, C. R.; Dann, J. G.; Eliopoulos, E.; Geddes, A. J.; Ogg, D.; North, A. C. T. The structure of mouse L1210 dihydrofolate reductase-drug complexes and the construction of a model of human enzyme. *FEBS Lett.* **1987**, *218*, 178–184.
(12) Matthews, D. A.; Smith, S. L.; Baccanari, D. P.; Burchall, J. J.; Oatley, S. J.; Kraut, J. In *Chemistry and Biology of Pteridines;* Cooper, B. A.; Whitehead, V. M., Eds.; Walter de Gruyter & Co.: Berlin, 1986; pp 789–797.
(13) Bolin, J. T.; Filman, D. J.; Matthews, D. A.; Hamlin, R. C.; Kraut, J. Crystal Structures of *Escherichia coli* and *Lactobacillus casei* Dihydrofolate Reductase Refined a 1.7 Å Resolution. *J. Biol. Chem.* **1982**, *257*, 13650–13662.
(14) Roberts, G. C. K. Understanding the specificity of the dihydrofolate reductase binding site. *NATO ASI Ser., Ser. A.* **1989**, *183*, 209–220.
(15) Cruciani, G.; Goodford, P. J. A search for specificity in DNA-drug interactions. *J. Mol. Graphics* **1994**, *12*, 116–129.
(16) SYBYL 6.1 Molecular Modeling Software, TRIPOS, Inc., 1699 S. Hanley Rd., St. Louis, MO, 1994.
(17) Goodford, P. J. A Computational Procedure for Determining Energetically Favourable Binding Sites on Biologically Important Macromolecules. *J. Med. Chem.* **1985**, *28*, 849–857.
(18) Boobbyer, D. N. A.; Goodford, P. J.; McWhinnie, P. M.; Wade, R. C. New Hydrogen-Bond Potentials for Use in Determining Energetically Favourable Binding Sites on Molecules of Known Structure. *J. Med. Chem.* **1989**, *32*, 1083–1094.
(19) Wade, R.; Clerk, K. J.; Goodford, P. J. Further development of hydrogen bond function for use in determining energetically favourable binding sites on molecules of know structure. Ligand probe groups with the ability to form two hydrogen bonds. *J. Med. Chem.* **1993**, *36*, 140–147.
(20) GRID v.12, Molecular Discovery Ltd., West Way House, Elms Parade, Oxford, 1995.
(21) Joliffe, J. *Principal Component Analysis*; Springer: Berlin, 1986.
(22) Wold, S.; Esbensen, K.; Geladi, P. Principal Component Analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52.
(23) Baroni, M.; Costantino, G.; Cruciani, G.; Riganelli, D.; Valigi, R.; Clementi, S. Generating Optimal Linear PLS Estimations (GOLPE): An Advanced Chemometric Tool for Handling 3D-QSAR Problems. *Quant. Struct.-Act. Relat.* **1993**, *12*, 9–20.
(24) Kraulis, P. J. "MOLSCRIPT": a program to produce both detailled and schematic plots of protein structures. *J. Appl. Crystallogr.* **1991**, *24*, 946–950.

JM9504013