Available online at www.sciencedirect.com
ScienceDirect
Journal of Molecular Graphics and Modelling 26 (2007) 728–739
Journal of Molecular Graphics and Modelling
www.elsevier.com/locate/JMGM

Exploring the other side of biologically relevant chemical space: Insights into carboxylic, sulfonic and phosphonic acid bioisosteric relationships

Antonio Macchiarulo, Roberto Pellicciari *

Dipartimento di Chimica e Tecnologia del Farmaco, Università di Perugia, via del Liceo 1, 06123 Perugia, Italy

Received 20 December 2006; received in revised form 27 March 2007; accepted 28 April 2007
Available online 3 May 2007

Abstract

Bioisosteric replacements have been widely and successfully applied to develop bioisosteric series of biologically active compounds in medicinal chemistry. In this work, the concept of bioisosterism is revisited using a novel approach based on charting the ''other side'' of biologically relevant chemical space. This space is composed by the ensemble of binding sites of protein structures. Explorations into the ''other side'' of biologically relevant chemical space are exploited to gain insight into the principles that rules molecular recognition and bioisosteric relationships of molecular fragments. We focused, in particular, on the construction of the ''other side'' of chemical space covered by binding sites of small molecules containing carboxylic, sulfonic, and phosphonic acidic groups. The analysis of differences in the occupation of that space by distinct types of binding sites unveils how evolution has worked in assessing principles that rule the selectivity of molecular recognition, and improves our knowledge on the molecular basis of bioisosteric relationships among carboxylic, sulfonic, and phosphonic acidic groups.

Keywords: Molecular recognition; Bioisosterism; Molecular descriptors; Chemical space; Biological space

1. Introduction

The concept of bioisosterism, firstly introduced by Friedman [1], is based on the assumption that functional groups or molecules that share similar geometrical and/or chemical properties would exert similar biological actions. It has been widely and successfully exploited to develop bioisosteric series of biologically active compounds both in medicinal chemistry and during the optimization process of lead compounds in drug discovery [2–4].

Three principal approaches, in particular, have been so far used to study the bioisosteric relationship of functional groups: (i) theoretical methods aimed at the development of descriptors of geometry and strength of non-bonded interactions of chemical groups [5]; (ii) analysis of literature data [6]; (iii) statistical analysis of frequencies and geometries of non-bonded interactions of different functional groups in crystallographic database [7–9].

In this work, we revisit the concept of bioisosterism on the assumption that binding sites with similar geometrical and/or chemical properties would recognize similar functional groups or molecules. The idea was to chart a compilation of binding sites of biological targets in order to identify which discrete population of binding sites may better recognize a given functional group or molecule and, in turn, improve the choice of the right bioisosteric replacement in a lead compound depending on the population to which its biological target belongs.

The charting of different constellations of binding sites found in cells or model systems leads to the definition of a ''binding site-based chemical space''. This can be envisaged as representing the ''other side'' of the small molecule-based biologically relevant chemical space [10,11]. Analogously to the chemical space, this space is multi-dimensional and its extent and properties will depend on the set of descriptors chosen to define the binding site itself. Previously, structural analyses of binding sites have been widely performed and employed by Thornton and co-workers in order to identify preferred interaction regions of atom probes [12], to define different types of molecular recognitions [13,14] and to predict protein functions [15,16].

* Corresponding author. Tel.: +39 075 585 5120; fax: +39 075 585 5124.
E-mail address: rp@unipg.it (R. Pellicciari).

1093-3263/$ – see front matter
doi:10.1016/j.jmgm.2007.04.010

The first endeavor in our novel approach, has been the charting of the ''other side'' of biological space encompassing small molecules containing acidic groups such as carboxylates, sulfonates or, phosphonates. These groups have been widely studied for comparative appraisals of their bioisosteric relationships. In particular, the majority of these studies have emphasized the importance of differences in the stereochemistry of interactions as one of the most important factor in determining the bioisosterism of acidic groups. Thus, sulfonic and phosphonic acidic groups have been classified as non-planar bioisosteric surrogates of the carboxylic acid function [17]. Sulfonic and phosphonic groups, indeed, share a pyramidal geometry around the central sulfur and phosphorous atom, whereas the carboxylic group displays a planar geometry.

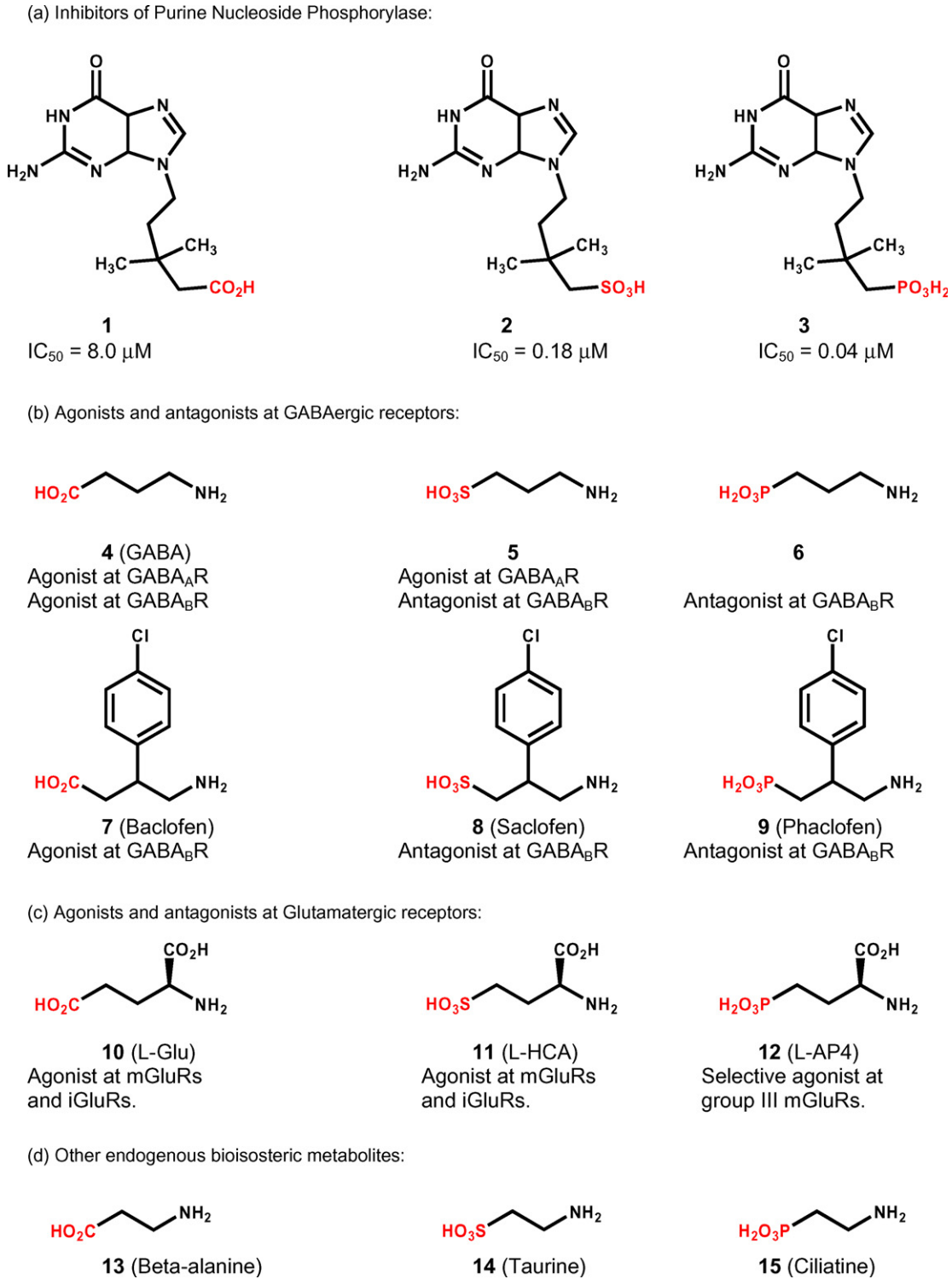

Fig. 1. (a–d) Examples of carboxylate, sulfonate and phosphonate replacement in the design and synthesis of enzyme inhibitors and receptor agonists and antagonists.

While phosphonic group is a diprotic acid ($pK_{a1}^{CH_3PO_3H_2} = 3.4$; $pK_{a1}^{CH_3PO_3H_2} = 7.8$), carboxylic and sulfonic groups are monoprotic acid, though with different pK_a values ($pK_a^{CH_3COOH} = 4.76$; $pK_a^{CH_3SO_3H} = -2.6$) [17,18]. Thus, all the three acidic groups may form both electrostatic interactions and hydrogen bonding, though differences exist among the preferred geometries of the hydrogen bond interaction. Carboxylic group, in particular, prefers hydrogen bond interactions along the lone-pairs in the *syn* position [19–21], sulfonic group adopts a nearly *eclipsed* geometry of the hydrogen bond interaction angle whereas phosphonic group prefers a *gauche* orientation of the interaction angle [9]. In addition, a crystallographic database survey of geometrical features of the hydrogen bonding interactions revealed that the sulfonic group tends to form longer interactions than carboxylic and phosphonic groups [7].

More than once the mutual replacement of these acidic moieties has been exploited to develop bioisosteric series of biologically active compounds such as enzyme inhibitors (**1**–**3**) [22], receptor agonists (**4**, **5**, **7**, **12**) and antagonists (**6**, **8**, **9**) (Fig. 1) [23–27].

For instance, carboxylic, sulfonic and phosphonic groups were used to functionalize the end group of alkyl chain attached to the nine position of guanine in order to synthesize inhibitors of purine nucleoside phosphorylase [22]. Results pinpointed that while phosphonate (**3**) and sulfonate (**2**) moieties interacted well with the binding site of the enzyme, the carboxylic group (**1**) did not yield a strong interaction. Phosponate and sulfonate were exploited as carboxylate mimics of GABAergic agonists, such as γ-amino butyric acid (**4**, GABA) and baclofen (**7**) [23–26]. Strikingly, these replacements shifted the pharmacological profiles of the resulting compounds to GABAergic antagonists (**5**, **6**, **8**, **9**). However, this was not the case in the development of glutamatergic modulators. Indeed, the phosphonate replacement of the distal carboxylic group of glutamate (**10**) led to the synthesis of L-AP4 (**12**) which turned out to be a selective subtype agonist of metabotropic glutamate receptors [27].

Interestingly, nature uses the same acidic group replacements to produce an array of active bioisosteric metabolites. For instance, L-homocysteic acid (L-HCA, **11**) is a naturally occurring bioisoster of glutamate (**10**). It has been reported that L-HCA (**11**) acts as an endogenous neurotransmitter modulating synaptic responses of the central nervous system (CNS) through the activation of glutamate receptors [28,29]. Beta-alanine (**13**), taurine (**14**) and ciliatine (**15**) are other examples of naturally occurring bioisosteric metabolites of acidic groups. In addition, sulfation and phosphorylation do not occur only in small molecule metabolites, but they are also post-translational modifications of proteins [30,31]. In contrast to phosphorylated proteins that mediate intracellular signal transduction, sulfated proteins are involved in extracellular crosstalk and cell–matrix interactions [32].

Our work consists in two parts. In the first part, we wonder if the presence of differences in some properties of binding sites recognizing carboxylic, sulfonic and phosphonic groups, furnishes novel clues in understanding the molecular recognition of acidic groups.

In the second part of the work, by constructing the "other side" of biological space of carboxylic, sulfonic, and phosphonic groups, we provide additional understanding of the bioisosteric relationships existing among the three acidic moieties. Noteworthy, the analysis of the "other side" of biological space emphasizes an underestimated property of bioisosterism: its relativity to biological target. This property was first introduced by Thornber [33], who stated that functional groups or molecules proved to be good bioisosteric replacement in one series of compounds are not necessarily useful in another.

2. Methods

2.1. Definition and collection of dataset

Specific criteria were used to define and collect three dataset of binding sites from the RCSB protein database (PDB). In particular, 198 structures in complex with small molecules containing carboxylic groups, 200 complexes with molecules containing the sulfonate moiety and 170 structures containing ligands with phosphonate groups were retrieved. Similar sequences were removed from each dataset using a cutoff of 90% of sequence identity according to the clustering protocol implemented in the PDB website (ftp://ftp.rcsb.org/pub/pdb/derived data/NR/).

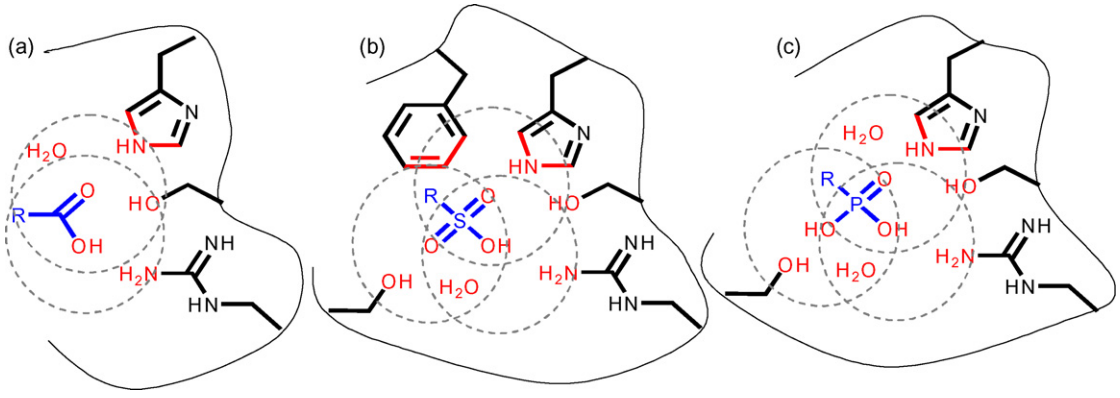

Fig. 2. (a–c) Definition of binding sites of bioisosteric groups according to the chemical environment contained in a sphere of radius 3.5 Å centered on each oxygen atom of the acidic group.

For each entry of the resulting dataset, the binding site of the bioisosteric group was defined as the chemical environment (number and type of atoms) contained in a sphere of radius 3.5 Å centered on each oxygen atom of the acidic group (Fig. 2). The value of the radius was chosen according to previously statistical analysis about the longer length of hydrogen bonding observed in crystal structures of carboxylate, sulfonate and phosphonate groups [7]. A complexity here is that many entries have multimeric chains and/or more acidic groups. In order to avoid duplicate entries, wherever more than one acidic functional group was present, all the relative binding sites were analyzed and only the unique ones were stored into the dataset. After the above operations, the final three dataset comprised respectively 176 binding sites of carboxylate groups, 180 binding sites of sulfonate moieties and 101 binding sites of phosphonate groups.

2.2. Statistical analysis

Experimental (frequency distribution of residue types, formal charge, hydrogen bonding, number of water molecules) and theoretical (hydrophobicity, flexibility, bulkiness) descriptors were calculated to compare the properties of the three dataset of binding sites. In particular, formal charge was calculated summing the charged residues and metal ions present in each binding site. We assigned a formal charge of +1 to lysine, arginine and monovalent metal ions (e.g. sodium, potassium); +2 to bivalent metal ions (e.g. magnesium, manganese, zinc, calcium); −1 to aspartate and glutamate residues. Although the protonation state of charged residues may involve histidines and varies according to the pH of the environment and the presence of cofactors, we thought that this approximation furnishes a crude estimation of the propensity of the binding site to make electrostatic interactions. Hydrogen bonding property was calculated counting the number of hydrogen bond acceptors and donors. According to the convention of atom types in pdb files, hydrogen bond donors were expressed as the sum of ''OH'' (tyrosine side chain), ''OG'' (serine and threonine side chains) and ''N'' atom types while hydrogen bond acceptors were the sum of all oxygen atom types including ''OH'' and ''OG''. Thus, the side chains of serine, threonine and tyrosine are both hydrogen bond acceptors and donors. Since histidine is endowed with a tautomeric equilibrium between the proximal and distal nitrogen of imidazole ring, both nitrogen atoms were considered as hydrogen donors to complement the hydrogen accepting property of acidic groups. Oxygen type of crystallized water was considered apart from hydrogen bonding property to reflect the specific ability of this molecule to adapt its hydrogen accepting or donating role upon the binding of acidic groups. Although different residue type classification can be adopted, in this study we classified residue type according to the polar character of the side chain. Thus, polar residues were arginine, lysine, histidine, aspargine, glutamine, serine, threonine, cysteine, tyrosine, aspartate and glutamate. The count of type and number of residues present in the binding site was accomplished analyzing the atoms falling within the defined spheres and the residues to which they belonged (Fig. 2). Thus, if any atom of a given residue falls within the spheres, the residue to which it belongs is counted for the analysis. Furthermore, we counted the presence of metal ions in the binding site. Theoretical descriptors were calculated on the basis of the profile produced by amino acid scales from literature. Briefly, amino acid scale are tables of numerical values representing different chemical and physical properties of each type of amino acid [34]. In particular, we used the Kyte–Doolittle scale to evaluate the hydropathicity of the binding site [35]; the average flexibility index of amino acid residues to determine the propensity of the site to conformational changes upon ligand binding [36]; an amino acid scale to calculate the size and bulkiness distribution of binding sites [37]. Property distributions were numerically represented using the mean, median, standard deviation and confidence limits. The type of distribution was assessed using the Pearson coefficient of asymmetry. The coefficient is calculated as the ratio of the difference between the mean and median over the standard deviation. If the coefficient is <-0.5, the distribution is negatively skewed; with a coefficient comprised in between or equal to -0.5 and 0.5, the data are approximately normally distributed; if the coefficient is >0.5, the distribution is positively skewed. Since most distributions were not normal distributions, the median was considered as more representative than the mean [38].

The above descriptors were used as variables to perform a principal component analysis (PCA). A traning set was defined from the original dataset containing 126 binding sites of carboxylate groups, 130 binding sites of sulfonate moieties and 51 binding sites of phosphonate groups. The remaining binding sites were used to define a test set (50 binding sites of carboxylate groups, 50 binding sites of sulfonate moieties and 50 binding sites of phosphonate groups) in order to cross-validate the PCA model.

In order to facilitate the interpretation of the descriptors, we checked the variables that were highly intercorrelated (coefficient of correlation >0.8) in the initial pool. In particular, the variable of flexibility turned out to be highly correlated with both bulkiness and the number of hydrogen bond donors, thus it was removed.

Cluster analysis was carried out using the relocation algorithm (k-means method) [39] on the resulting first three components of the PCA analysis. The method consists in randomly selecting three objects (binding sites) as cluster centroids. All the remaining objects are assigned to each cluster based on the closest centroid. This process represents the first step of iterative clustering cycles in which each cycle is composed by the calculation of new centroids of current clusters and assignment of objects to the cluster of the closest centroid. The iteration stops when the assignment of the objects to each cluster is stabilized at a given criteria of convergence. In this study, calculations are repeated ten times using 500 iterations in order to choose the optimal solution. The convergence criteria is set to a value of 10^{-4}. For the first iteration, the initial partition is chosen randomly (case a); dividing the whole dataset into three equal parts based on the

data order (carboxylate, sulfonate and phosphonate dataset) and taking the geometric centers of each part as cluster centroids (case b); using the geometric centers of each of the three dataset (case c). Two classification criteria of k-means clustering are used: the pooled within covariance matrix (DW) and the pooled SSPC matrix (WT). All statistical analyses were performed using the XL-STAT software. The evolutionary trace analysis of 14 sequences belonging to the class of Isomerase was carried out using the ConSurf Server [40]. Results are reported as supplementary materials (Table s2, Fig. s2) along with the multiple sequence alignment (Fig. s1) that was performed using ClustalW with the default parameters as implemented at the EBI website (http://www.ebi.ac.uk/clustalw). Table s1 of the supplementary materials contains the data about the training set, test set and the validation set which was used to test the ability of our model in assessing the correct bioisosteric replacements.

3. Results and discussion

3.1. Molecular recognition of carboxylic, sulfonic and phosphonic groups

In order to investigate the presence of specific properties within the binding sites recognizing different acidic groups, three dataset were collected from the protein databank comprising 176 binding sites of carboxylate functions, 180 binding sites of sulfonate moieties and 101 binding sites of phosphonate groups. No pair of binding sites was available where one acidic group (e.g. carboxylate) was substituted for another of this study (sulfonate and/or phosphonate) in a co-crystallized ligand.

The distribution of functions within each dataset is shown in Fig. 3 along with the averages and standard deviations of resolution factors. Enzymes are classified according to the EC number and represent the majority of each dataset, reflecting the bias of the protein database content.

The following properties were calculated for each binding site: frequency distribution of residue type and atom type, formal charge, hydrogen bonding, hydrophobicity, flexibility and bulkiness. The statistical analysis of these properties are reported in Table 1.

The inspection of Table 1 pinpoints the existence of binding site fingerprints that constitute the basis of specificity of molecular recognition of acidic groups. In particular, a selective molecular recognition of phosphonate groups is achieved by highly polar binding sites endowed with large flexibility and bulkiness. Noteworthy, these sites have the highest number of hydrogen bond donor groups compared to carboxylate and sulfonate binding sites. Similarly, we found a high positive

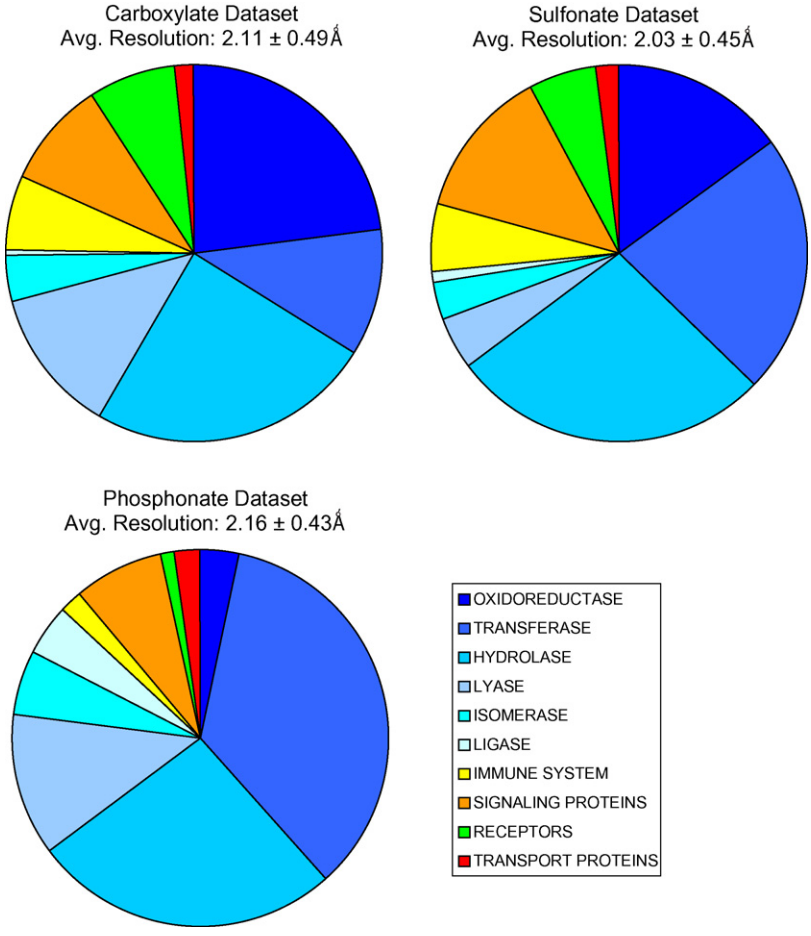

Fig. 3. Distribution of molecular functions within each dataset and relative average resolution ± standard deviation.

Table 1
Statistical analysis of molecular descriptors

Descriptors	Dataset	Mean	S.D.	Conf. 95%	Median	Pearson
Hydropathicity (Kyte–Doolittle)	Carboxylate	−4.82	5.54	0.78	−4.4	−0.24
	Sulfonate	−3.01	4.80	0.71	−3.10	0.06
	Phosphonate	−9.21	6.36	1.26	−8.80	−0.19
Bulkiness	Carboxylate	38.95	22.45	3.34	37.09	0.25
	Sulfonate	35.90	25.99	3.82	32.31	0.41
	Phosphonate	51.50	23.13	4.57	48.38	0.40
Flexibility	Carboxylate	1.27	0.74	0.11	1.31	−0.18
	Sulfonate	1.09	0.75	0.11	0.95	0.55
	Phosphonate	1.93	0.74	0.15	1.92	0.03
Polar residues	Carboxylate	2.12	1.46	0.22	**2**	0.26
	Sulfonate	1.54	1.22	0.18	**1**	1.34
	Phosphonate	3.73	1.89	0.37	**3**	1.16
Nonpolar residues	Carboxylate	0.73	0.98	0.15	0	2.22
	Sulfonate	0.97	1.13	0.17	1	−0.07
	Phosphonate	0.80	1.08	0.21	0	2.23
Metal ions	Carboxylate	0.06	0.25	0.04	0	0.67
	Sulfonate	0.01	0.11	0.02	0	0.32
	Phosphonate	0.45	0.71	0.14	0	1.87
Formal charge	Carboxylate	0.70	0.82	0.12	**1**	−1.08
	Sulfonate	0.48	0.77	0.11	**0**	1.85
	Phosphonate	1.66	1.31	0.26	**2**	−0.77
Water molecules	Carboxylate	0.78	0.94	0.14	1	−0.69
	Sulfonate	1.22	1.28	0.19	1	0.51
	Phosphonate	2.12	2.06	0.41	2	0.17
HB acceptors	Carboxylate	0.87	0.96	0.14	1	−0.41
	Sulfonate	0.54	0.78	0.11	0	2.07
	Phosphonate	1.37	1.19	0.23	1	0.92
HB donors	Carboxylate	1.99	1.27	0.19	**2**	−0.01
	Sulfonate	1.63	1.47	0.22	**1**	1.29
	Phosphonate	3.32	1.68	0.33	**3**	0.56

Experimental descriptors that promote a selective molecular recognition of each acidic group are in bold.

formal charge that underlines the strong propensity of these binding sites to make salt bridges with the double negative charge of phosphonic groups. Strikingly, the binding sites of sulfonates display a relatively poor polarity. This observation may be explained if we consider that the dataset of sulfonates suffers the bias of comprising many binding sites of sulfonate detergents that bind to hydrophobic sites. However, the lack of hydrogen bond acceptors found in sulfonate sites can be well ascribed to the negative pK_a value of sulfonic acid that affects a constant loss of the acidic proton at any protein environment and makes useless the presence of hydrogen bond acceptor functions.

The molecular recognition of carboxylic groups is achieved through narrow and mostly rigid binding sites that display a polarity in between sulfonate and phosphonate sites. In particular, the presence of two hydrogen bond donors, one acceptor and formal charge of 1, ensures an optimal anchoring of the carboxylic moiety.

Since the replacement of carboxylic, sulfonic and phosphonic groups is exploited by nature in the synthesis of metabolites and post-translational modifications of proteins, the fingerprints shown in Table 1 are the result of an evolution of recognition sites toward the achievement of selective binding interactions. In particular, the success of evolution is the exploitation of diverse acidic and hydrogen bonding properties of carboxylate, sulfonate and phosphonate groups by providing the optimal mixture and supply of complementary charge and polar groups found in binding sites. An evolutionary trace analysis of 14 sequences collected across the three dataset and belonging to the class of Isomerase, was carried out to investigate whether the amino acids responsible for acid-selectivity were conserved or exchanged during evolution. The results showed that in most of the cases the conservation scores of residues were under the statistical confidence cut-off hampering a careful analysis of their relative conservation profiles during evolution (results are reported as supplementary materials).

3.2. The "other side" of biologically relevant chemical space

In this section, we discuss how the construction and the analysis of the ''other side'' of biological space may unveil clues on molecular recognition and bioisosteric relationships of

the three acidic groups, namely carboxylic, sulfonic and phosphonic groups.

Principal component analysis (PCA) is a key tool to visualize structural diversity and similarity among different classes of objects [41–43].

In particular, we used a combination of PCA and clustering analysis to investigate if a different classification of the dataset of binding sites would be possible regardless to the acidic group of the co-crystallized ligand, but taking into account only the properties of the binding site. The idea behind this strategy was that it may exist a portion of binding sites that do not necessarily recognize with the highest affinity the acidic group present in the co-crystallized ligand. This portion of binding sites may, indeed, be endowed with properties that are better suited for the molecular recognition of another acidic group for which we do not have any co-crystallized ligand. As such, these binding sites occupy the tails of the above property distributions.

Thus, in order to evaluating the differences in the occupancy of diversity space by binding sites, we divided each of the original three dataset into a training set and test set and a PCA analysis of the whole training set was carried out. Although the 92% of the explained variance is achieved at the fifth component (Table 2), we discuss only the first three components as they cover the 73% of the initial variability of the dataset.

The inspection of the loadings (Table 3) of the original variables into the components reveals that the first component explains the effect of size and polarity of the binding site in determining the variance among the molecular recognition of carboxylate, sulfonate and phosphonate groups.

This is ascribed to the highest positive loadings shown by bulkiness, metal ions, the number of polar residues, hydrogen bond donors, hydrogen bond acceptors and formal charge on this component. On the part of binding sites, the 41% of the variance explained by the first component interprets diverse propensities to hydrogen bonding, coordination bonding and electrostatic interactions of the carboxylate, sulfonate and phosphonate recognition sites. The second component explains the effect of propensity to make van der Waals interactions by binding sites upon molecular recognition of the acidic groups. This effect is ascribed to the high positive loading of the variable accounting for the number of nonpolar residues in the binding site. It should be mentioned that this variable could also be interpreted as the role of hydrophobic interaction played in the molecular recognition of carboxylate, sulfonate and phosphonate. However, we thought that this is not the case since other variables related to the hydrophobic effect such as the hydropathicity and bulkiness of the binding site show only negligible loadings. The third component explains the 13% of the variance. The number of water molecules inside the binding site highly affects the meaning of this component. Water molecules are generally involved in molecular recognition through the formation of hydrogen bonds that bridge the ligand to the binding site. In such cases, they occupy energetically favorable and conserved positions within binding sites or can be brought by the acidic groups as

Table 2
Eigenvalues and explained variance (%)

Component	Eigenvalue	Variability (%)	Cumulative (%)
PC-1	3.683	40.927	40.927
PC-2	1.739	19.324	60.252
PC-3	1.165	12.949	73.201
PC-4	0.957	10.630	83.831
PC-5	0.709	7.878	91.710
PC-6	0.345	3.831	95.540
PC-7	0.234	2.602	98.143
PC-8	0.137	1.518	99.661
PC-9	0.031	0.339	100

Table 3
Loadings of principal component analysis (PCA)

Descriptors	PC-1	PC2	PC-3
Hydrophaticity (Kyte–Doolittle)	−0.772	0.389	0.314
Bulkiness	0.762	0.530	0.031
Polar residues	0.968	−0.035	−0.080
Nonpolar residues	−0.042	0.831	0.494
Formal charge	0.621	−0.278	0.180
Water molecules	0.053	−0.380	0.725
Metal ions	0.487	−0.422	0.499
HB acceptors	0.627	0.033	−0.090
HB donors	0.742	0.462	−0.027

Table 4
Squares of the principal components

Descriptors	PC-1	PC2	PC-3
Hydrophaticity (Kyte–Doolittle)	0.596	0.152	0.099
Bulkiness	0.581	0.281	0.001
Polar residues	**0.936**	0.001	0.006
Nonpolar residues	0.002	**0.690**	0.244
Formal charge	0.385	0.077	0.032
Water molecules	0.003	0.145	**0.526**
Metal ions	0.237	0.178	0.249
HB acceptors	0.393	0.001	0.008
HB donors	0.550	0.214	0.001

Descriptors highly associated with each component are bold typed.

Table 5
Cluster analysis of the training set

Method[a]	Dataset	Group 1 (%)	Group 2 (%)	Group 3 (%)
k-means (WT)-a	Carboxylate	18	54	28
	Sulfonate	12	63	25
	Phosphonate	49	22	29
k-means (DW)-a	Carboxylate	18	54	28
	Sulfonate	12	63	25
	Phosphonate	49	22	29
k-means (WT)-b	Carboxylate	18	54	28
	Sulfonate	11	63	26
	Phosphonate	51	20	29
k-means (WT)-c	Carboxylate	12	67	21
	Sulfonate	12	64	25
	Phosphonate	51	25	24

[a] See Section 2 for details about the method used.

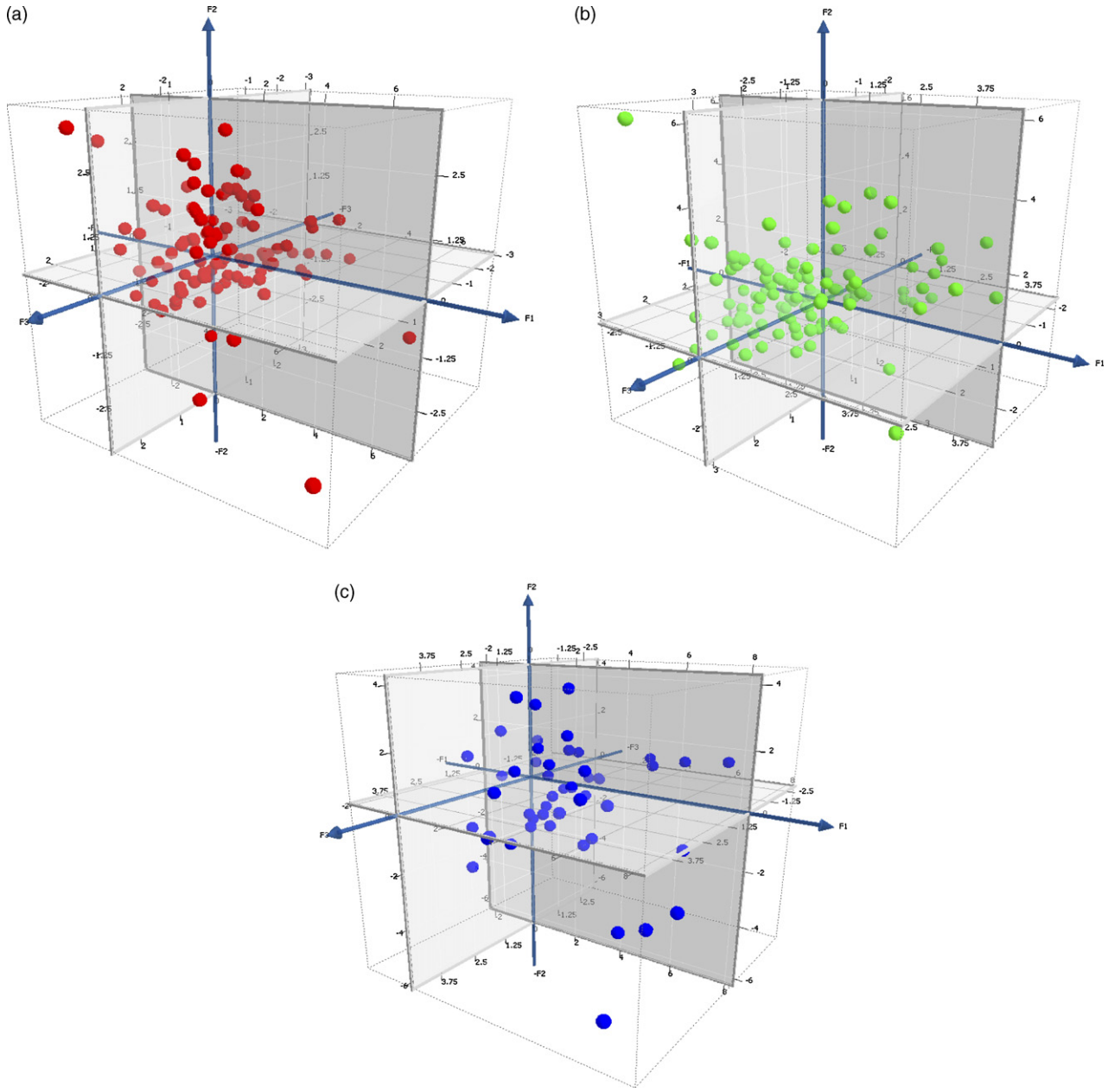

Fig. 4. Binding site constellations of the training set of acidic bioisosteric groups plotted onto the first three components (PC1 = F1, PC2 = F2 and PC3 = F3). Each dot (star) represents a binding site in the other side of biologically relevant chemical space: (a) carboxylate constellation; (b) sulfonate constellation; (c) phosphonate constellation.

part of their solvation shell. The inspection of the median and mean values in Table 1 reveals that the number of water molecules in the binding sites of phosphonate approaches the double of the corresponding values in sulfonate and carboxylate dataset. The greater occurrence of water molecules in a fraction of phosphonate dataset may be the consequence of a higher solvent accessible surface and strong polarity of these binding sites, or the consequence of a higher desolvation energy linked to a more strongly polar group, as it is the phosphonic group.

In order to confirm the above links of variables to the relative components, we analyzed the squares of the principal components (Table 4). In particular, the greater is the squared cosine, the greater is the link of the variable with the corresponding component.

The inspection of Table 5 confirms that polar interactions are viewed on the first component where the number of polar residues got the highest squared cosine; the second component encodes for the presence of hydrophobic residues whereas the third component is linked to the number of water molecules.

The next step was to plot the constellations of binding sites for each dataset of the training set on the basis of their relative score values resulting from the PCA study. Since the variance represented by the first two components is not very high (61%),

and to avoid a misinterpretation of the results, we complemented the plot of the first and second component with the third component (Fig. 4).

Sulfonate and carboxylate dataset occupy similar spaces with the latter tailing off at positive values of the first component. Conversely, phosphonate binding sites are widespread around both components and reach more positive values along the first component. In order to gain insight into what kind of acidic bioisosteric replacement would be the best on the basis of the characteristics of the binding site, the diversity space occupied by the three constellations was clustered into three regions (Fig. 5). The idea was to identify three areas of the space were it is expected to be (un-)fruitful one of the following replacements: carboxylic/sulfonic, carboxylic/phosphonic and sulfonic/phosphonic acid replacement. The inspection of population of different types of binding sites within each area unveils the specificity of molecular recognition occurring in that zone and the right bioisosteric group to be used in the design of small molecule modulators for the protein falling in that specific area.

To avoid the dependence of the resulting three areas on the clustering protocol, several protocols were comparatively used (Table 5).

In all of the clustering protocols, one class contains a net prevalence of phosphonate binding sites (49–51%) over carboxylate (12–18%) and sulfonate binding sites (11–12%). We will refer to this region as area G1. This area is localized around positive values of the first component and negative values of the second component (Fig. 5). The prevalence of binding sites that interact with phosphonate groups in area G1, pinpoints that a selective molecular recognition of this moiety occurs in these binding sites. Furthermore, we cannot rule out that the small fraction of carboxylate and sulfonate binding sites, found in this area, may better interact with phosphonic moieties (e.g. methionine aminopeptidase, which is outlier in the carboxylate constellation and is grouped in area G1).

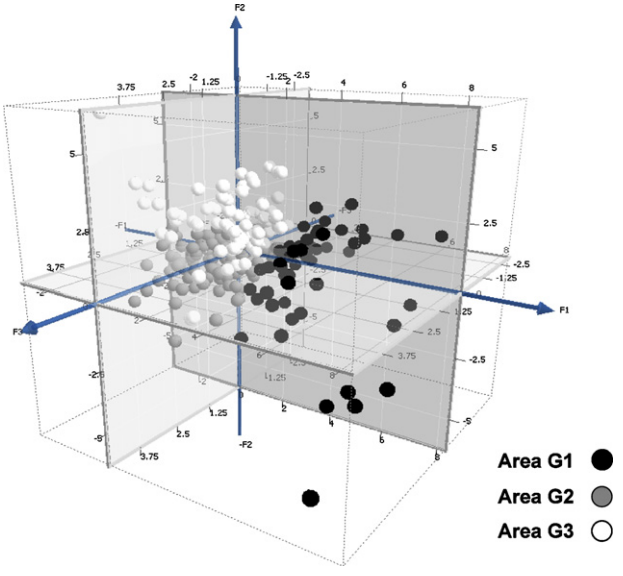

Fig. 5. Cluster analysis of the training set of binding sites within the biological space according to the method k-means (WT)-a.

Sulfonic or carboxylic replacement of phosphonic groups is likely to be unfruitful in ligands that bind to sites of area G1.

The second class (G2) is located around negative values of first and second components (Fig. 5). It is mainly composed of carboxylate (54–67%) and sulfonate (63–64%) binding sites. Phosphonate binding sites represent only a negligible portion of this class (20–25%). Thus, conversely to area G1, carboxylic and sulfonic replacements are well tolerated in small molecules that interact with sites of area G2.

Area G3 is located on top of G1 and G2, occupying the space around positive value of the second component and a short range of positive and negative values of the first component (Fig. 5). Herein, binding sites of acidic groups display almost the same abundance of population. Ligands that interact with these binding sites are suitable to tolerate carboxylic, sulfonic and phosphonic switching as good bioisosteric replacements.

It should be mentioned that limitations exist in constructing and analyzing a biological space of binding sites. These comprise both the limited number of irredundant protein complexes available in the protein data bank (for a recent review see Ref. [44]) and the lack of accurate structures and annotation of activity data for small molecules and molecular fragments. A further level of limitation is given by the fact that even when binding assay data are available, they usually refer to diverse small molecules and do not include bioisosters of a given compound.

Aware of the above limitations, we attempted to test the predictive accuracy of PCA and clustering analyses using the remaining binding sites of the test set (Table 6 and Fig. 6).

The inspection of Table 6 reveals that the same three areas (G1–G3) identified in the training set are also present in the test set. Binding sites of area G1 comprise almost exclusively phosphonate recognition sites (80–82%). Conversely, area G2 is occupied by a large fraction of carboxylate (50–52%) and sulfonate (62%) binding sites whereas area G3 does not show any marked prevalence of a specific recognition site.

As an additional validation of our study, we tested the ability of our model in assessing the correct bioisosteric replacements on the basis of the occupancy of binding sites in one of the above three regions of the space.

Table 6
Cluster analysis of the test set

Method[a]	Dataset	Group 1 (%)	Group 2 (%)	Group 3 (%)
k-means (WT)-a	Carboxylate	22	50	28
	Sulfonate	8	62	30
	Phosphonate	80	10	10
k-means (DW)-a	Carboxylate	22	50	28
	Sulfonate	8	62	30
	Phosphonate	80	10	10
k-means (WT)-b	Carboxylate	22	50	28
	Sulfonate	8	62	30
	Phosphonate	80	10	10
k-means (WT)-c	Carboxylate	32	52	16
	Sulfonate	10	62	28
	Phosphonate	82	12	6

[a] See Section 2 for details about the method used.

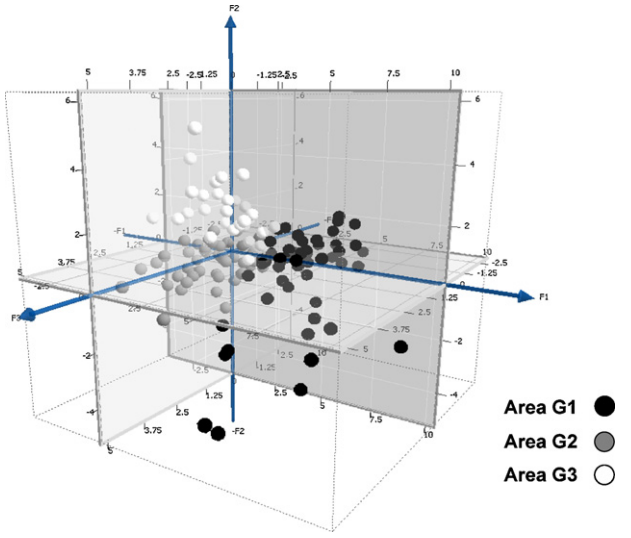

Fig. 6. Cluster analysis of the test set of binding sites within the biological space according to the method k-means (WT)-a.

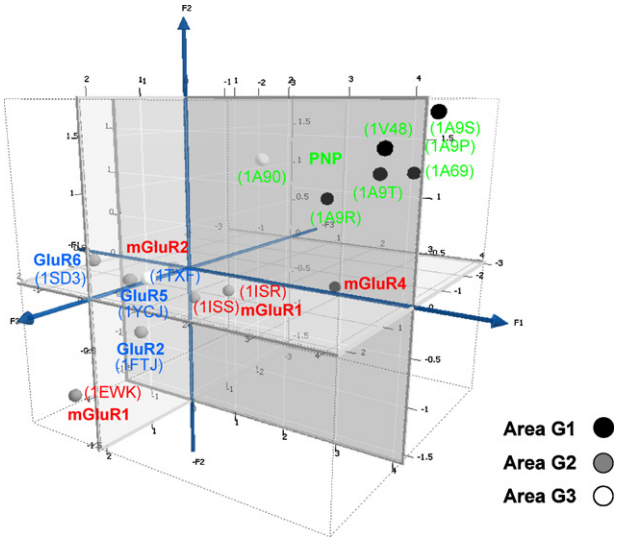

Fig. 7. Plot of the validation set into the other side of biologically relevant chemical space.

Thus, we collected an additional external test set (validation set) of drug targets for which both structures and biological data of bioisosteric ligands were available. In some case, like metabotropic glutamate receptor subtypes mGluR2 and mGluR4, we exploited the in-house availability of homology models of the binding site to complete the structural information [45]. Thus, plotting the binding site of seven crystallized purine nucleoside phosphorylases reveals the occupation of the border of area G1 (six structures) facing area G3 where only one structure (PNP, 1a90) is located (Fig. 7). In agreement with the proposed high specificity to phosphonate recognition of sites belonging to area G1, biological assays point out that phosphonic inhibitors display the highest affinity to the enzyme followed by sulfonic and carboxylic derivatives (**1**, **2**, **3**, Fig. 1) [22]. Similarly, the projection of the homology model of the binding site of mGluR4 onto the biological space reveals the occupation of area G1 whereas all remaining crystal structures of metabotropic and ionotropic glutamate receptors and the homology model of mGluR2 occupy area G2 and only one structure (GluR5, 1txf) falls in area G3 (Fig. 7). Again, phosphonic replacement of the distal carboxylic group of glutamic acid (**10**) yielded the mGluR4 subtype selective agonist, L-AP4 (**12**, Fig. 1) [27] (Fig. 7).

Finally, on the basis of the classification of the test set into the three areas G1–G3, some prospective predictions were made about which acidic group may be better recognized by a random selection of nine binding sites (Table 7).

The first six entries (2sli, 1j37, 1l7f, 1d7r, 1kkr, 1bf3) are binding sites co-crystallized with small molecules bearing a carboxylic function and that, according to the properties of the binding site, are classified in area G1.

These enzymes, though recognizing small molecules containing carboxylic groups, have strongly polar binding sites that occupy positions in the diversity space far right along the first component. For these proteins, the replacement of the carboxylic group of the ligand with a phosphonic moiety should improve the molecular recognition of the small molecule.

Conversely, the last three entries of Table 7 are binding sites (1p7p, 1kf0, 1m32) co-crystallized with small molecules bearing a phosphonic group that, according to the properties of

Table 7
Prospective predictions of a selection of binding sites of the test set

PDB code	Function	Co-crystallized compound[a]	Area of localization	Proposed replacement
2sli	Hydrolase	2-Acetylamino-7-(1,2-dihydroxy-ethyl)-3-hydroxy-6,8-dioxabicyclo[3.2.1]octane-5-**carboxylic acid**	G1	Phosphonate
1j37	Hydrolase	(2*S*)-1-[(2*R*)-2-Methyl-3-sulfanyl-propanoyl]pyrrolidine-2-**carboxylic acid** (Captopril)	G1	Phosphonate
1l7f	Hydrolase	(1*S*,2*S*,3*S*,4*R*)-3-[(1*S*)-1-Acetamido-2-ethyl-butyl]-4-(diaminomethylideneamino)-2-hydroxy-cyclopentane-1-**carboxylic acid**	G1	Phosphonate
1d7r	Lyase	*N*-Pyridoxyl-1-amino-cyclopropane**carboxylic acid**-5-monophosphate	G1	Phosphonate
1kkr	Lyase	(2*S*,3*S*)-3-Methyl**aspartic acid**	G1	Phosphonate
1bf3	Oxydoreductase	4-Hydroxyben**zoic acid**	G1	Phosphonate
1p7p	Ligase	Methionine **phosphonate**	G2	Carboxylate
1kf0	Transferase	3-**Phospho**glycerate	G2	Carboxylate
1m32	Transferase	2-Aminoethyl**phosphonate**	G2	Carboxylate

[a] In bold it is highlighted the acidic function to be replaced.

the binding site, are classified in area G2. For these proteins, the replacement of the phosphonic group of the ligand with a carboxylic or sulfonic moiety should improve the molecular recognition of the small molecule.

4. Conclusions

In a framework aimed at understanding the bioisosteric relationship among carboxylic, sulfonic and phosphonic acid groups, we envisaged a picture of biological space constituted by the ensemble of binding sites that recognize these acidic groups. The identification of three areas in space diversely populated by distinct types of binding sites, unveils how evolution has worked in assessing principles that rule the selectivity of molecular recognition. In this context, the success of evolution is based on the exploitation of different acidic and hydrogen bonding properties of carboxylate, sulfonate and phosphonate groups. Thus, depending upon the properties of the binding site, only some mutual replacement of these acidic groups will be fruitful or bioisosteric in small molecule ligand designing.

Although this is a case study restricted to binding sites recognizing acidic fragments, its extension to include other small molecule and molecular fragment binding sites will lead to map the entire ''other side'' of biological space. The analysis of such space will improve our knowledge of molecular recognition and bioisosteric relationship of small molecules and molecular fragments in the context of binding sites.

In terms of chemical biology and medicinal chemistry, the biological space of binding sites found in cells or model systems provides a framework to understand which molecule or molecular fragment is best suited to modulate orphan binding sites or replace an existing molecule or molecular fragment on the road to obtain selective chemical tools for gene products with unknown functions, and drug candidates. In addition, a map of binding sites would enforce the strategy of target family directed master-keys [46,47] by unveiling unexpected relationships among binding sites of diverse biological target families based on their molecular properties. Thus, explorations into the ''other side'' of biological space are expected to efficiently aid the identification of biologically active molecule into its counterpart, the chemical space, lowering the risk of getting lost into biologically meaningless regions.

Finally, the increasing number of emerging biological targets and the growth of structural biology projects [44] are paving the way to gain a full access into the ''other side'' of biological space.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmgm.2007.04.010.

References

[1] H.L. Friedman, Influence of isosteric replacements upon biological activity, in: Symposium on Chemical–Biological Correlation, National Academy of Sciences, Washington, DC, (1951), p. 295.
[2] X. Chen, W. Wang, The use of bioisosteric groups in lead optimisation, Annu. Rep. Med. Chem. 38 (2003) 333–344.
[3] G.A. Patani, E.J. LaVoie, Bioisosterism: a rational approach in drug design, Chem. Rev. 96 (1996) 3147–3176.
[4] P. Stefanic, M.S. Dolenc, Aspartate and glutamate mimetic structures in biologically active compounds, Curr. Med. Chem. 11 (2004) 945–968.
[5] L.B. Kier, L.H. Hall, Bioisosterism: quantitation of structure and property effects, Chem. Biodiver. 1 (1996) 138–151.
[6] Bioster, v2003-1; Accelrys Software Inc., San Diego, CA.
[7] B. Pirard, G. Baudoux, F. Durant, A database study of intermolecular NH–O hydrogen bonds for carboxylates, sulfonates and monohydrogen phosphonates, Acta Cryst. B51 (1995) 103–107.
[8] P. Watson, P. Willett, V.J. Gillet, M.L. Verdonk, Calculating the knowledge-based similarity of functional groups using crystallographic data, J. Comput. Aid. Mol. Des. 15 (2003) 835–857.
[9] Z.F. Kanyo, D.W. Christianson, Biological recognition of phosphate and sulfate, J. Biol. Chem. 266 (1991) 4264–4268.
[10] C.M. Dobson, Chemical space and biology, Nature 432 (2004) 824–828.
[11] C. Lipinski, A. Hopkins, Navigating chemical space for biology and medicine, Nature 432 (2004) 855–861.
[12] R.A. Laskowski, J.M. Thornton, C. Humblet, J. Singh, X-SITE: use of empirically derived atomic packing preferences to identify favourable interaction regions in the binding sites of proteins, J. Mol. Biol. 259 (1996) 175–201.
[13] R.A. Laskowski, N.M. Luscombe, M.B. Swindells, J.M. Thornton, Protein clefts in molecular recognition and function, Protein Sci. 5 (1996) 2438–2452.
[14] I. Nobeli, R.A. Laskowski, W.S. Valdar, J.M. Thornton, On the molecular discrimination between adenine and guanine by proteins, Nucl. Acids Res. 29 (2001) 4294–4309.
[15] R.A. Laskowski, J.D. Watson, J.M. Thornton, Protein function prediction using local 3D templates, J. Mol. Biol. 351 (2005) 614–626.
[16] F. Glaser, R.J. Morris, R.J. Najmanovich, R.A. Laskowski, J.M. Thornton, A method for localizing ligand binding pockets in protein structures, Proteins 62 (2006) 479–488.
[17] C.G. Wermuth, The Practice of Medicinal Chemistry, Academic Press, 1996.
[18] J. March, Advanced Organic Chemistry, 3rd ed., Wiley, 1985.
[19] M. Tintelnot, P. Andrews, Geometries of functional group interactions in enzyme-ligand complexes: guides for receptor modelling, J. Comput. Aid. Mol. Des. 3 (1989) 67–84.
[20] S.M. Roe, M.M. Teeter, Patterns for prediction of hydration around polar residues in proteins, J. Mol. Biol. 229 (1993) 419–427.
[21] G. Klebe, The use of composite crystal-field environments in molecular recognition and the de novo design of protein ligands, J. Mol. Biol. 237 (1994) 212–235.
[22] W.C. Guida, R.D. Elliott, H.J. Thomas, J.A. Secrist III, Y.S. Babu, et al., Structure-based design of inhibitors of purine nucleoside phosphorylase. 4. A study of phosphate mimics, J. Med. Chem. 37 (1994) 1109–1114.
[23] A. Giotti, S. Luzzi, S. Spagnesi, L. Zilletti, Homotaurine: a GABAB antagonist in guinea-pig ileum, Br. J. Pharmacol. 79 (1983) 855–862.
[24] D.I. Kerr, J. Ong, R.H. Prager, B.D. Gynther, D.R. Curtis, Phaclofen: a peripheral and central baclofen antagonist, Brain Res. 405 (1987) 150–154.
[25] D.I. Kerr, J. Ong, G.A. Johnston, R.H. Prager, GABAB-receptor-mediated actions of baclofen in rat isolated neocortical slice preparations: antagonism by phosphono-analogues of GABA, Brain Res. 480 (1989) 312–316.
[26] D.I. Kerr, J. Ong, G.A. Johnston, J. Abbenante, R.H. Prager, Antagonism at GABAB receptors by saclofen and related sulphonic analogues of baclofen and GABA, Neurosci. Lett. 107 (1989) 239–244.
[27] C. Thomsen, P. Kristensen, E. Mulvihill, B. Haldeman, P.D. Suzdak, L-2-Amino-4-phosphonobutyrate (L-AP4) is an agonist at the type IV metabotropic glutamate receptor which is negatively coupled to adenylate cyclase, Eur. J. Pharmacol. 227 (1992) 361–362.
[28] M.C. Curras, R. Dingledine, Selectivity of amino acid transmitters acting at *N*-methyl-D-aspartate and amino-3-hydroxy-5-methyl-4-isoxazolepropionate receptors, Mol. Pharmacol. 41 (1992) 520–526.

[29] Q. Shi, J.E. Savage, S.J. Hufeisen, L. Rauser, E. Grajkowska, et al., L-Homocysteine sulfinic acid and other acidic homocysteine derivatives are potent and selective metabotropic glutamate receptor agonists, J. Pharmacol. Exp. Ther. 305 (2003) 131–142.
[30] T. Hunter, Protein kinases and phosphatases: the yin and yang of protein phosphorylation and signaling, Cell 80 (1995) 225–236.
[31] K.L. Moore, The biology and enzymology of protein tyrosine *O*-sulfation, J. Biol. Chem. 278 (2003) 24243–24246.
[32] S. Hemmerich, D. Verdugo, V.L. Rath, Strategies for drug discovery by targeting sulfation pathways, Drug Discov. Today 9 (2004) 967–975.
[33] C.W. Thornber, Isosterism and molecular modification in drug design, Chem. Soc. Rev. 8 (1979) 563–580.
[34] E. Gasteiger, C. Hoogland, A. Gattiker, S. Duvaud, M.R. Wilkins, et al., Protein Identification and Analysis Tools on the ExPASy Server. The Proteomics Protocols Handbook, Humana Press, 2005, pp. 571–607.
[35] J. Kyte, R.F. Doolittle, A simple method for displaying the hydropathic character of a protein, J. Mol. Biol. 157 (1982) 105–132.
[36] R. Bhaskaran, P.K. Ponnuswammy, Positional flexibilities of amino acid residues in globular proteins, Int. J. Pept. Protein Res. 32 (1988) 242–255.
[37] J.M. Zimmerman, N. Eliezer, R. Simha, The characterization of amino acid sequences in proteins by statistical methods, J. Theor. Biol. 21 (1968) 170–201.
[38] R.F. Sokal, F.J. Rohlf, Biometry, 2nd ed., New York, 1981, pp. 43–46.
[39] B.S. Everitt, G. Dunn, Applied Multivariate Data Analysis, Oxford University Press, New York, 1992.
[40] M. Landau, I. Mayrose, Y. Rosenberg, F. Glaser, E. Martz, et al., ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures, Nucl. Acids Res. 33 (2005) W299–W302.
[41] D.S. Tan, Diversity-oriented synthesis: exploring the intersections between chemistry and biology, Nat. Chem. Biol. 1 (2005) 74–84.
[42] M. Feher, J.M. Schmidt, Property distributions: differences between drugs, natural products, and molecules from combinatorial chemistry, J. Chem. Inf. Comput. Sci. 43 (2003) 218–227.
[43] T.I. Oprea, J. Gottfries, Chemography: the art of navigating in chemical space, J. Comb. Chem. 3 (2001) 157–166.
[44] J.M. Chandonia, S.E. Brenner, The impact of structural genomics: expectations and outcomes, Science 311 (2006) 347–351.
[45] A. Macchiarulo, G. Costantino, R. Sbaglia, S. Aiello, M. Meniconi, et al., The role of electrostatic interaction in the molecular recognition of selective agonists to metabotropic glutamate receptors, Proteins 50 (2003) 609–619.
[46] P.A. Rejto, G.M. Verkhivker, Unraveling principles of lead discovery: from unfrustrated energy landscapes to novel molecular anchors, Proc. Natl. Acad. Sci. USA 93 (1996) 8945–8950.
[47] G. Muller, Medicinal chemistry of target family-directed masterkeys, Drug Discov. Today 8 (2003) 681–691.