3086 J. Med. Chem. 2010, 53, 3086–3094
DOI: 10.1021/jm901696w

Journal of Medicinal Chemistry Article

Chemical Fragments that Hydrogen Bond to Asp, Glu, Arg, and His Side Chains in Protein Binding Sites

A.W. Edith Chan,[†] Roman A. Laskowski,[‡] and David L. Selwood*,[†]

[†]Biological and Medicinal Chemistry Group, Wolfson Institute for Biomedical Research, University College London, Gower Street, London WC1E 6BT, U.K., and [‡]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, U.K.

Received November 16, 2009

We present an analysis of the chemical fragments from lead-like ligands in the Protein Data Bank (PDB) that form hydrogen bonds to the side chains of Asp, Glu, Arg, and His, which are the most common residues found in ligand binding sites. A fragment is defined as the largest ring assembly containing the atoms involved in hydrogen bonding. In total, 462 fragments were found in 2038 ligands from over 8000 protein−ligand structures in the PDB. The results show which fragments have a higher propensity for interaction with specific side chains. Some fragments interact with Asp but not with Glu, and vice versa, despite these side chains sharing the same chemical moiety. Arg side chains form hydrogen bonds almost exclusively with O-mediated ligands, and the fragments are the most diverse. Hydrogen bond distances from the imidazole of His showed a wider range than the other three amino acids.

Introduction

During the past 10 years, fragment-based drug discovery[1−4] has become an established and successful paradigm. It is commonly used to discover new chemical entities against a protein drug target. Small chemical structures or fragments (usually of 150−250 molecular weight and weak affinity) are screened to probe the protein's binding site to identify larger and more potent binding molecules. Although most platforms are laboratory-based, in silico techniques are emerging.[5] Most computational methods employ well-established virtual screening techniques such as docking or pharmacophore generation together with a library of low molecular weight "fragments" to identify ligands that have a high probability of binding.[6] Free-energy calculation by systematic sampling followed by de novo assembly of fragments has also been developed.[7] Small fragment databases are usually generated by analyzing drug-related databases, such as the World Drug Index.[8] Analyses of data from molecular databases have been used to characterize molecular frameworks,[9] property,[10] diversity, and privileged scaffolds for different drug-related molecular databases. These databases provide useful information on the variety of fragments and their drug-like or lead-like properties, without the linking information on target or protein−ligand interactions.

Protein−ligand interaction information and ligand chemical structures can be retrieved from many established web-based databases such as Relibase,[11] PDBsum,[12,13] and PDBe.[14,15] Searches can be by protein name, protein sequence, molecule name, formula, simplified molecular input line entry specification (SMILES[a]) string, or a sketched molecular fragment. A comprehensive review of these databases has recently been published.[16] However, given a new target protein, it is still hard to determine what kind of scaffolds or fragments chemists should try. Identification of the chemical fragments that preferentially interact with particular protein side chains in a binding site could provide a head start for library design work.

Our knowledge of the three-dimensional (3D) structures of protein targets of course plays a major role in designing and optimizing compounds that bind to specific targets. Currently, the number of macromolecular structures publicly available from the Protein Data Bank (PDB)[17] is close to 60K entries, of which about 37K entries have bound ligands. The 3D structures of protein−ligand complexes provide a wealth of information for understanding how proteins interact with different chemical fragments, and functional groups.[18] To date, analyses of specific ligand/side chain interactions in the structures of the PDB have tended to focus on single ligand atoms rather than chemical fragments. Thus the analyses that underpin SuperStar[19] or MED-SuMo,[20] the former analyses derive from the 3D distributions of specific ligand atoms types about different protein side chains while the latter applies a heuristic based on 3D representation of macromolecular surfaces such as H-bond acceptor and hydrophobic. These two methods are similar to those with pharmacophore generation methods (chemical physical properties are merged−e.g. H bond from Asp or Glu), while our new method lists the fragments from the interaction with each amino acid. The results, for example, the difference in fragment interaction with Asp and Glu will not be discovered if the interaction is combined. Relibase, too, provides superposition of a particular ligand/fragment with protein side chains/functional groups, revealing their interaction modes of a related protein family. For a medicinal chemist,

*To whom correspondence should be addressed. Phone: +44 207 679 6716. Fax: +44 207 209 0470. E-mail: d.selwood@ucl.ac.uk.

[a] Abbreviations: 3D, three-dimensional; CCP, cytochrome C peroxidase; HA, hydrogen bond acceptor; HD, hydrogen bond donor; logP, logarithm base 10 of the partition coefficient P; MW, molecular weight; PDB, Protein Data Bank; PISCES, protein sequence culling server; RO5, Lipinski's Rule of Five; SMILES, simplified molecular input line entry specification.

 Published on Web 03/15/2010

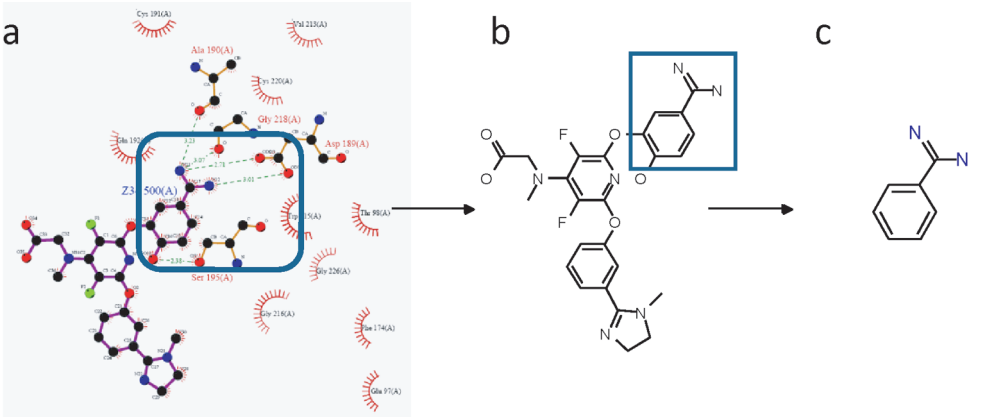


Figure 1. Definition of fragment motif: (a) LIGPLOT of PDB code: 1fjs. The hydrogen bonding interaction between the ligand Z34 and the Asp side chain is highlighted in a rounded box; (b) structure of ligand *N*-[2-[5-[amino(imino)methyl]-2-hydroxyphenoxy]-3,5-difluoro-6-[3-(4,5-dihydro-1-methyl-1*H*-imidazol-2-yl)phenoxy]pyridin-4-yl]-*N*-methylglycine (ZK-807834, ligand code: Z34). Extracted fragment in square box; and (c) extracted fragment with hydrogen bonding atoms in blue.

however, chemical fragments (and chemical functional groups) are of more interest and value than single atoms.

In this study, we identify chemical partners or fragments that commonly bind to side chains of four specific amino acids via hydrogen bonding interactions. There are, of course, many ways of defining fragments or scaffolds depending on the aim of the study.[9,21] In this paper, our aim is to bring out the most important proximate chemical features that form hydrogen bonds with the protein side chains. Therefore, a ring assembly seems to be a good choice. Simple functional groups linked to alkyl chains may be important for binding but these will be considered in a separate study. Therefore, we defined a fragment as the largest ring assembly containing the atoms involved in hydrogen bond(s) to one of the side chains. Among the 20 amino acids, at least 12 (especially charged amino acids: Asp, Glu, Arg, and Lys) can form hydrogen bonds with their side chains. However we chose Asp, Glu, Arg, and His for their reported importance in protein binding sites. For example, Villar and Kauvar found that in 50 diverse protein binding sites, His, Arg, and Asp were more frequently in contact with the ligand than other side chains.[22] In another study, Bartlett et al. have found that His, Glu, Asp, and Arg accounted for 55% of all catalytic residues in enzyme active sites.[23] In this study, we examined only hydrogen bonding interactions with the ligand. The process followed in the analysis is illustrated in Figure 1. The interaction information for a given amino acid side chain and its partner ligand is extracted from PDBsum, the interacting fragment of the ligand is identified and the fragments collated, only relatively high resolution < 2.1 Å structures were used. This process extracts the preferred fragments for a given amino acid side chain and should allow chemists to bias the compositions of a library to enable a higher hit rate. We find that fragments interacting with Asp and Glu are surprisingly conservative especially for those forming two hydrogen bonds. Fragments interacting with Arg and His are more diverse. The data can serve to inform a fragment-based approach to drug design as well as being useful for targeting binding hot spots in protein−protein interactions.[24]

Results and Discussion

Fragment Definition and Extraction. For each ligand, the chemical fragment involved in hydrogen bonding to the protein side chain of interest was identified as illustrated in Figure 1. Figure 1a shows a LIGPLOT[25] diagram of the hydrogen bonding interactions between a ligand and protein. Figure 1b shows a chemical representation of the ligand, and Figure 1c shows the interacting fragment. The chemical fragments were abstracted from ligands that interact via hydrogen bonds with protein side chains. We define the interacting fragment as the largest ring assembly containing the atoms involved in the hydrogen bond(s). Other substituents on the ring assembly are removed if they are not involved in the hydrogen bond to the relevant protein side chain. The removal of other substituents that are not involved in hydrogen bonding will reduce the number of different fragments, thus making the statistics of fragments studied more relevant. It was apparent that this definition did not capture information for functional groups such as sulfonamide, urea, and hydroxamic acid linked to alkyl chains. However, the choice of our fragment definition was designed to maximize the information retrieval of drug like fragments.[26] Alkyl chains introduce greater degrees of conformational freedom and might be expected to be disfavored as drug fragments. In our analysis, we had to strike a balance between retrieval of "drug like" fragments and rejecting useful binding information. Thus the rules utilized are broader than a simple Rule of Three definition[27] such as employed for a fragment library for screening. The molecular weight range 100−800 is far outside a Rule of Three definition or even Lipinski[28] drug likeness guidelines. In addition, we collected the information on motifs such as amidines, even though they may have problems in terms of oral bioavailability. The purpose of this study is to identify the most common binding units (or fragments) without prejudice. If the information is being used to design a screening library, then factors such as the suitability of groups for an oral indication can come into play. A detailed description of the fragment extraction process is in the Materials and Methods section.

Acidic and Negatively Charged Residues−Asp and Glu. The pK_a values for Asp and Glu side chain carboxylates are 3.9 and 3.3, respectively,[29] and are negatively charged at physiological pH. The carboxylic group is capable of forming hydrogen bonds with other molecules either through one or both oxygens. In this study, we will examine these two categories separately. In general, from more than 5000 PDB complexes 159 and 143 unique fragments were found for Asp

Table 1. Fragments Showing Two Hydrogen Bonds to Asp and Glu Side Chains[d]

Generic fragment (f[a])	Highest frequency examples[b] n[c]		
Aryl amidines			
Asp (11)	72	30	12
Glu (2)	2	0	0
Guanidines			
Asp (2)	14	0	1
Glu (5)	2	1	1
1-aza-2-aminoaryls			
Asp (8)	9	7	4
Glu (6)	0	10	0
azaheteroaryl-7-amines			
Asp (2)	0	2	
Glu (1)	3	0	
Dihydropyridazines			
Asp (2)	2	1	
Glu (0)	0	0	
Cyclic diols			
Asp (20)	69	54	10
Glu (13)	65	105	5

[a] f is the frequency of the generic fragment found in the Asp and Glu data set. [b] Atoms involved in hydrogen bonding are in bold-type. Hydrogen atoms are not shown for chemical structures as these are not visible in most X-ray data. [c] n is the frequency of the unique ligand containing the fragment appearing in the PDB bound to the Asp/Glu side chains. [d] Wn represents an inserted CH_2 or a functional group.

and Glu, respectively. These fragments together with their frequency of occurrence, ligand codes, PDB codes, and PDBsum links are listed in Tables S1 and S2 in the Supporting Information. The atoms that can mediate hydrogen bonds are nitrogen (most frequent), oxygen, and sulfur. There are only four S-mediated fragments for Asp and Glu combined.

Fragments Binding Both Oxygens of Asp and Glu Side Chains. One would expect that the hydrogen bond fragments found for Asp and Glu would be similar because they both have carboxylic acid side chains. Indeed, their O-mediated fragments are similar (Table 1). However, their N-mediated fragments are quite different when both oxygens are engaged in hydrogen bonding.

In Asp, when both oxygens of the carboxylic acid moiety are involved in hydrogen bonding with ligands, the aryl amidine-mediated fragments (Table 1) occur the most frequently (11 fragments). The amidine group forms two hydrogen bonds beautifully with the carboxylic acid group in a typical linear fashion through a pair of N−H···O interactions as shown in Figure 1a. In contrast, this type of interaction is not common for Glu, as only two fragments are found with the amidine group and their frequency is low. The amidine group fragment is mostly found in binding with serine proteases, such as trypsin, thrombin, and factor Xa, where it binds with Asp in the active site. Serine proteases accounted for almost 1200 PDB entries (out of 3851) in this

Fragment	a	b	c
Ligand code	270	AIM	IK8
PDB code	1qa0	1d6w	1z6e
Protein	Trypsin	Thrombin	Factor Xa

Figure 2. Fragments that form a single hydrogen bond with the Asp side chain found in serine proteases. The amino group only interacts with one oxygen to the Asp side chain.

study. Other nonamidine containing groups, either in ring structures or alkyl chains, have been used to target these families.[30] Nonamidine containing ring fragments in low frequency were found for these proteins. Figure 2 lists one example for trypsin, thrombin, and factor Xa. However, these fragments form H-bonds with only one oxygen from the side chain of Asp. For example, the fragment aminobenzisoxazol (Figure 2c) has been used to target various serine proteases but depending on what scaffolds it was attached to showed different activities in different serine proteases.[31] The low frequency of these fragments in our data set is probably due to the pragmatism of the experimentalists; inhibitors known to bind to a given protein are likely to be tried for related family members.

Only 2 and 5 guanidine containing fragments are found for Asp and Glu, respectively, with the most common ones shown in Table 1. The phenylguanidine fragment extracted from interaction with Asp side chains came from 14 unique ligands, 9 (corresponding to 9 PDB entries) of them bound to the carboxypetidase B family while the rest to serine proteases (4 unique ligands, 6 PDB entries). For Glu, the five guanidine containing fragments came from nitric oxide synthase (ligand ARR in 1vaf and 1vag), thrombin (4CP in 2bvr), and neuraminidase (e.g BCZ in 1l7f, G20 in 2qwf, and GNA in 2qw3). In the case of thrombin, Glu192 is not at the center of the active site but rather near the P5′ and P6′ pockets.[32]

The most frequent N-mediated fragments found for Glu are mainly guanine-like and adenine-like bases, which allow two hydrogen bonds to form to the carboxylate through two adjacent nitrogens (Table 1). Only one each fragment of hydroxamic acid and urea are found for Glu (see Supporting Information Table S2) but none for Asp. Although relatively rare, the 1,2-dihydropyrazines (found only in the Asp data set) are an attractive double H-bonding group to carboxylates.[33]

The cyclic diols are extremely common double H-bond forming fragments. The three most common examples shown in Table 1 do not display the full diversity of this group (see Supporting Information Tables S1 and S2). Natural sugars (see the ribose like fragment in Table 1, bottom) are predominant, but other cyclic diols are very common. Although the diol is the interacting group here, the ease (in chemical synthesis terms) with which this can be accommodated within larger scaffolds goes some way to explaining the high frequency of this fragment.

Fragments Binding a Single Oxygen of Asp or Glu Side Chains. The N-mediated fragments of both Asp and Glu that bond via one oxygen of the carboxylic acid to the nitrogen show, as might be expected, a greater diversity than the double H-bond fragments. The highest frequency examples are shown in Table 2. Fragments can be from heterocyclic rings or amines, sulfonamides, or amide groups attached to ring systems. Some proteins bind to a diversity of fragments. For example, haem enzyme cytochrome C peroxidase (CCP)

Table 2. Fragments Showing One Hydrogen Bond to Asp and Glu Side Chains[d]

Generic fragment (f[a])	Highest frequency examples[b] n[c]		
aminoaryls			
Asp (29)	36	9	10
Glu (12)	9	8	5
Cyclic amines			
Asp (8)	8	5	4
Glu (10)	10	0	7
Azaheterocycles			
Asp (12)	7	6	5
Glu (13)	4	0	3
Sulfonamides			
Asp (3)	8	5	1
Glu (0)	0	0	0
Cyclic alcohols			
Asp (20)	23	4	2
Glu (26)	76	5	0
Phenols			
Asp (11)	54	5	1
Glu (16)	77	0	1
Carboxylic acids			
Asp (10)	3	5	1
Glu (7)	0	4	1

[a] f is the frequency of the generic fragment found in the Asp/Glu data set. [b] Atoms involved in hydrogen bonding are in bold-type. Hydrogen atoms are not shown for chemical structures as these are not visible in most X-ray data. [c] n is the frequency of the unique ligand containing the fragment appearing in the PDB bound to the Asp/Glu side chains. [d] Wn represents an inserted CH_2 or a functional group.

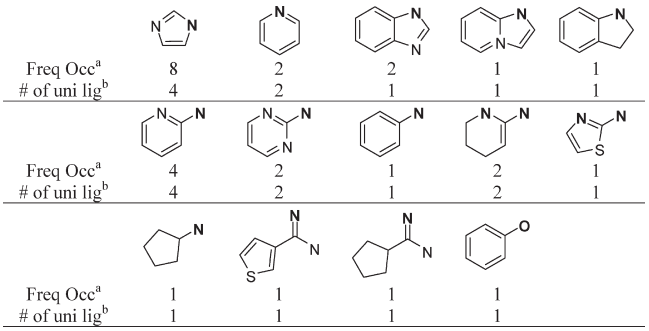


Figure 3. Fragments hydrogen bonding with Asp side chain found in cytochrome c peroxidases. (a) Frequency of occurrence (number of PDB retrieved) is listed for each fragment. (b) Number of unique ligands is listed for that fragment.

is a membrane-bound hemoprotein that is essential for electron transport. For this single protein, 14 fragments are found as shown in Figure 3.

For O-mediated interactions, the interacting fragments involving either single or both oxygens are similar. Both Asp and Glu interact mainly with hydroxyl oxygen (50 fragments). Most of them are sugar-like structures. The rest of the fragments are carboxylic acids (9 fragments). In cases like these, either the carboxylic acid from the Asp/Glu side chains or the ligand must be protonated. For example, the carbonyl group of one carboxylic acid is hydrogen bonded to the hydroxyl group of the other (see for example 1oxr in PDBsum). Unfortunately, it is impossible to distinguish which is which from the X-ray PDB structure. In the case of S-mediated fragments, two each are found for Asp and Glu.

Table 3. Fragments Showing Hydrogen Bonds to Arg Side Chain[d]

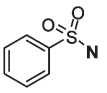
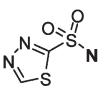
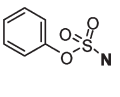
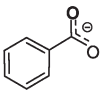
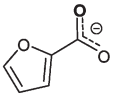
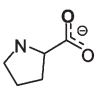
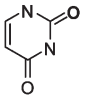
Generic fragment (f[a])	Highest frequency examples[b] n[c]		
Aza heterocycles			
(11)	9	3	2
Sulfonyls			
(9)	3	2	1
Cyclic alcohols			
(15)	11	7	7
Phenols			
(14)	20	5	3
Carboxylic acids			
(52)	22	11	8
Carbonyls			
(27)	4	3	1

[a] f is the frequency of the generic fragment found in the Arg data set. [b] Atoms involved in hydrogen bonding are in bold-type. Hydrogen atoms are not shown for chemical structures as these are not visible in most X-ray data. [c] n is the frequency of the unique ligand containing the fragment appearing in the PDB bound to the Arg side chain. [d] Wn represents an inserted CH_2 or a functional group.

Basic and Positively Charged Arg. Arg is positively charged at pH values below its pK_a, which is 12. Therefore, Arg is entirely a hydrogen bond donor in proteins.[34] Arginine is well designed to bind the phosphate anion. It is often found in the active centers of proteins that bind phosphorylated substrates[35−37] as well as in DNA binding to bases, such as guanine, thymine, and adenine.[38] In fact, the most frequent functional groups that interact with Arg are phosphate and phosphonate groups and they are exclusively alkyl linked so are not included in the present analysis. The whole fragment set found for Arg, together with their frequency of occurrence, ligand codes, PDB codes, and PDBsum links are listed in Supporting Information Table S3.

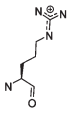
A total of 130 unique fragments were found for Arg. Arg can form hydrogen bonds through O, N, F, Cl, and S-mediated ligands. Table 3 shows the generic structures and examples of highest frequency fragments. Almost 90% of them are formed through O-mediated fragments (see Table 5). This suggests that it will be most efficient to target Arg with O-mediated fragments/ligands. Among the O-mediated fragments, the most popular ones are carboxylic acids (52 cases), hydroxyl (29 cases), and carbonyl (27 cases) groups. Only 12 N-mediated fragments are found. Other most common functional groups linked fragments found are nitro, sulfonyl, and amide.

Table 4. Fragments Showing Hydrogen Bonds to His Side Chain[d]

Generic fragment (f[a])	Highest frequency examples[b] n[c]		
Aza heterocycle			
 (8)	 14	 7	 4
Sulfonamides			
 (9)	 5	 1	 1
Cyclic alcohols			
 (24)	 4	 3	 3
Phenols			
 (20)	 88	 13	 2
Carboxylic acids			
 (18)	 17	 8	 2
Carbonyls			
 (28)	 5	 4	 4

[a] f is the frequency of the generic fragment found in the His data set. [b] Atoms involved in hydrogen bonding are in bold-type. Hydrogen atoms are not shown for chemical structures as these are not visible in most X-ray data. [c] n is the frequency of the unique ligand containing the fragment appearing in the PDB bound to the His side chain. [d] Wn represents an inserted CH_2 or a functional group.

Table 5. Analysis and Statistics for the Four Amino Acids

	Asp	Glu	Arg	His
				
Side chain moiety and property	carboxylic acid negatively charged, acidic	carboxylic acid negatively charged, acidic	guanidinium positively charged, basic	imidazole polar, basic
PDB[a]	3851	2541	3764	2736
Non-redundant PDB[b]	2428	1043	1568	1019
Protein family[c]	186	150	214	166
Unique ligand[d]	992	893	710	883
Unique motif	159	143	130	140
Diversity ratio[e]	0.16	0.16	0.18	0.16
Hydrogen bond mediated atom				
N	106 (66%)	80 (56%)	7 (5%)	36 (26%)
O	53 (33%)	56 (39%)	124 (91%)	86 (63%)
F/Cl	0	0	5	2
S	2	2	1	5
Mixed[f]	0	6	0	6

[a] The number of PDB entries was counted before filtering of unwanted ligands. This included all PDB entries where ligands have hydrogen bonding with the amino acid side chains. [b] The number of nonredundant PDB entries was calculated using PISCES[33] as described in text. [c] The name of protein family was retrieved directly from the title entry in PDB. [d] The number of unique ligands was counted after filtering of unwanted ligands. [e] The ratio is obtained by (number of unique fragment/number of unique ligands). It measures the diversity of the fragment motifs for all the ligands of each amino acid. [f] Mixed category involves more than two heteroatoms in the same ring assembly having hydrogen bonds with the same amino acid. The two heteroatoms are usually N and O.

Basic Residue His. His is the only amino acid with a pK_a in the physiological range, with a pK_a of 6. The imidazole of His makes it a common participant in enzyme catalyzed reactions. His has a high frequency coordinating to almost any metal.[39] The most frequent ones are Zn, Cu, Fe, and Mn. The unprotonated imidazole is nucleophilic and can serve as a general base, hydrogen bond donor and acceptor, while the protonated form can serve as a general acid and hydrogen bond donor. In any case, either hydrogen can serve as a hydrogen bond donor or acceptor. However, X-ray crystallography cannot resolve hydrogen atoms in most protein crystals and so hydrogens are absent from the coordinates. Thus it is difficult to determine which state the His is in when bonding to another molecule. In this study, we did not try to distinguish if the His side chains are hydrogen bond donors or acceptors.

His, unlike Arg, can form hydrogen bonds favorably with both N and O-mediated fragments. His can form hydrogen bonds through O, N, F, Cl, and S-mediated ligands. The whole fragment set found for His, together with their frequency of occurrence, ligand codes, PDB codes, and PDBsum links, are listed in Supporting Information Table S4. Table 4 shows the generic structures and examples of highest frequency fragments. Among the 140 unique fragments found for His, there are more O-mediated (86 cases, 63%) fragments than N-mediated (36 cases, 26%) ones. Like Arg, the most frequent functional group is the phosphate group (129 cases), yet the phosphate group was not observed linked to any ring system. Most of the O-mediated fragments are hydroxyl groups (44 cases), followed by 28 cases of carbonyl and 18 cases of acids (numbers may not total 86 because some fragments have more than one functional group that can form hydrogen bonds with His). Other functional groups linked fragments such as sulfonyl, sulfinyl, sulfonamide, and nitro are observed at low frequency.

The N-mediated fragments are mainly nitrogen containing heterocyclics, except for four cases where the interacting amine (three fragments) or guanidine (one fragment) group is attached to a ring system. The most frequent fragment is imidazole (15 cases). Because His frequently coordinates to metal in proteins, the ligands that bind to these active/binding sites have a high occurrence of binding to metal too. Nine sulfonamide fragments are found in our data set that attach to a ring system, and only one hydroxamic acid containing fragment was found in this study.

Analysis of the Whole Data Set. Table 5 shows some general statistics of the data retrieved. At first glance, it seems that Asp and Arg both have far more PDB entries than those of Glu and His. Using PISCES (protein sequence culling server)[40] to remove the redundant entries (sequence identity between two protein sequences of more than 80%[41]), it emerged that there are many more entries retrieved from Asp (2428 entries) over Glu (1043), Arg (1568), and His (1019).

Another interesting observation is that there are relatively few unique ligands from Arg compared to those of Asp, Glu, and His. The diversity ratio (number of unique fragments/number of unique ligands) in Table 5 measures the diversity of the fragments for all the ligands of each amino acid. When the diversity ratio is high or approaching 1, all the fragments

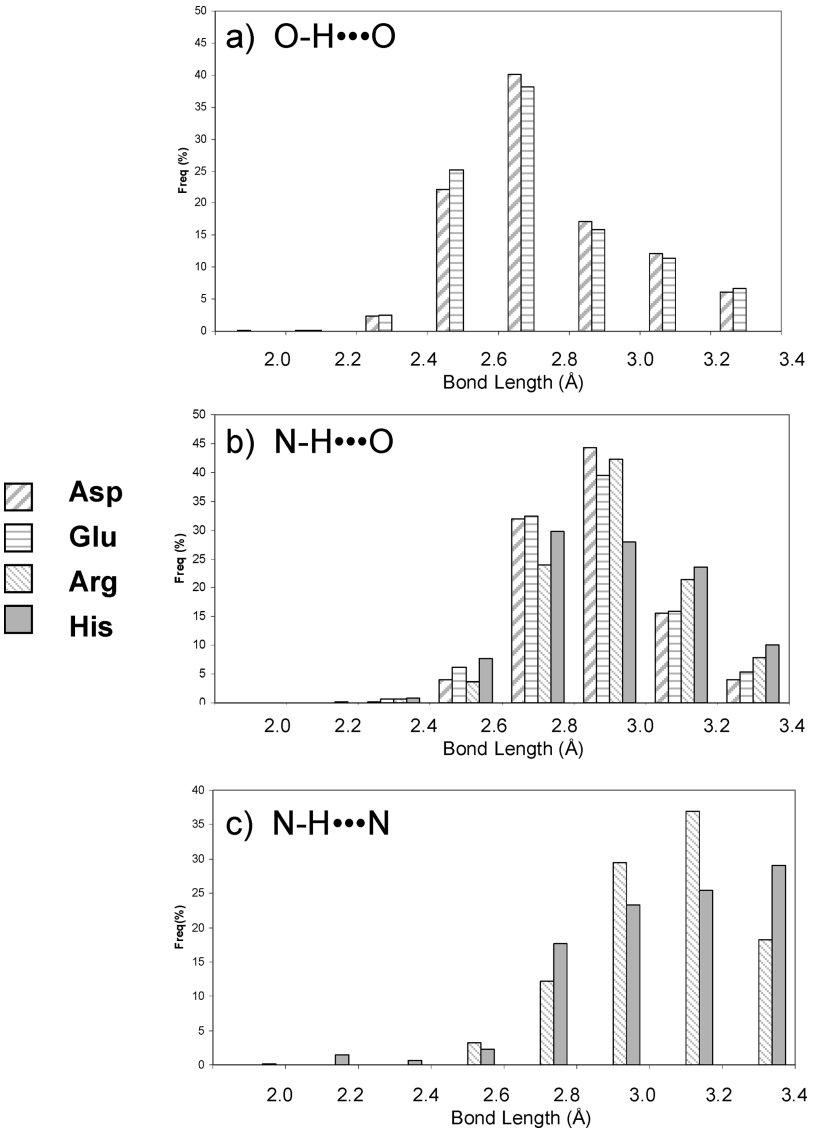


Figure 4. Hydrogen bond distances between ligand and side chain. (a) O−O type for hydrogen bonds for Asp and Glu; (b) N−O type for Arg and His, and O−N for Asp and Glu; (c) N−N-type for Arg and His where X-Y corresponds to X from side chain and Y from ligand.

from their individual ligands will be different from each other (most diverse). When the diversity ratio is small and approaching zero, there are more redundant fragments from different ligands. Asp, Glu, and His have a similar ratio, 0.16, while Arg (0.18) is higher. It suggests that there is more variety of fragments (more diverse) that interact with Arg in the PDB even though the number of unique ligands for Arg is the smallest in this study.

The four amino acids form hydrogen bonds with ligands through electronegative heteroatoms N, O, F, Cl, and S. The number of fragments via different heteroatoms is analyzed in Table 5, together with their percentage. In general, the acidic side chains of Asp and Glu have a higher tendency to form hydrogen bonds with N-mediated fragments, while basic residues, Arg and His, have a higher tendency to form hydrogen bonds with O-mediated ones. Most interesting is that Arg almost exclusively forms hydrogen bonds with O-mediated ligands (91%), suggesting it could be most effective to use O-mediated fragments to form hydrogen bonds with Arg. Fluorine and chlorine atoms have interesting bonding characters in protein−ligand interaction.[42−44] Despite the traditional view of being hydrogen bond acceptors, there is also evidence that they may be hydrogen bond donors. Among all our ring fragments, F and Cl attached to ring systems are only detected with low frequency for Arg and His side chains but not at all with those of Asp and Glu. This may suggest that it is not effective to target the side chains of the amino acids in this study with F or Cl.

In Figure 4, the hydrogen bond distances between ligand and side chain measured between two heteroatoms are reported. The measured distances were between O−O, N−O, and N−N atoms, as coordinates of H atoms are not usually reported in X-ray structures. Typical distances between two heteroatoms ranged from 2.5−3.5 Å. The range of bond lengths of different types of hydrogen bond found in this study is in line with the typical values.[45] The trend is that O−O being the shortest, while N−N the longest due to the atomic radii of these atoms.

In the N−O category, the mean hydrogen bond distances for Asp, Glu, and Arg are around 2.8−3.0 Å. However, in His, the mean value is slightly shorter, around 2.6−2.8 Å. Also its bond value distribution is more spread out. This implies that His, as a hydrogen bond acceptor or donor, has the ability to form a wider range (stronger to weaker) of

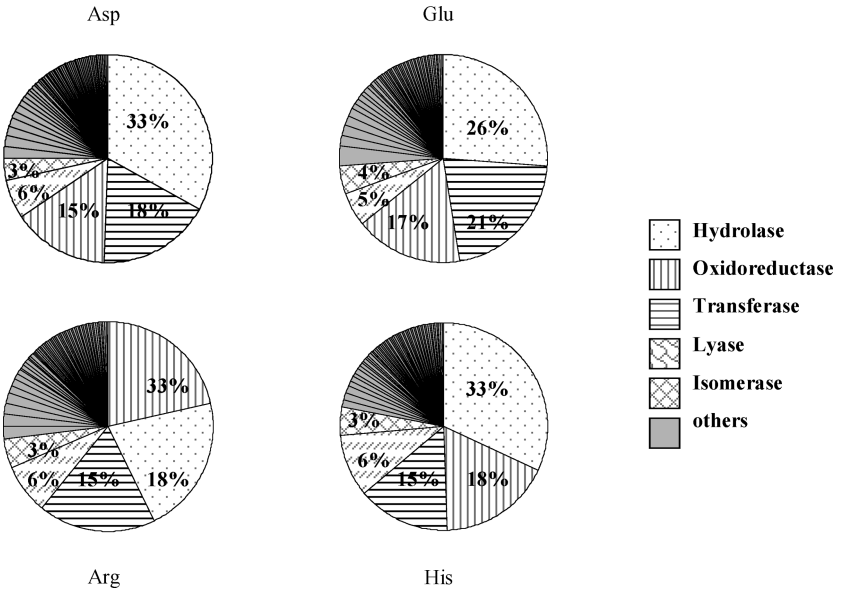


Figure 5. Family distribution based on the title information abstracted from the PDB.

hydrogen bonds with ligands, depending on the bonding environment in the ligand binding sites.

In the N−N category, the mean hydrogen bond value (3.2 to 3.4 Å) between His and other N-containing ligands tends to be longer. This suggests weaker hydrogen bonds compared to those formed by Arg. Most of the PDB entries in this category are Zn binding proteins. His coordinates to Zn as well as ligand. In this case, His is donating electrons to the metal and therefore forms a weaker hydrogen bond with ligand.

Our data set contains structures from all enzyme classes as well as from various receptor families. In Figure 5, the protein family distributions for each amino acid are shown as pie charts. In general, five protein families dominate: hydrolase, oxidoreductase, transferase, lyase, and isomerase. In Asp and Glu, their family distribution profiles are similar. However, in Arg, the top protein family is oxidoreductase. The domination of enzymatic classes over receptor families indicates that in almost all cases, hydrogen bonding interaction is a requirement in enzymatic active sites.

Use of the Data. The fragments compiled here may be of use to medicinal, computational scientists, and biological science researchers as a compendium of fragments known to interact with the four side chains analyzed here. The interactions these fragments make with the side chains can be examined in 3D or using the 2D schematics diagrams of LIGPLOT.

Another way of using this analysis might be to direct de novo fragment-based design efforts, where a typical work flow consists of docking a fragment into the binding site of the target protein, choosing the best orientation and then using this as a starting point for the attachment of substituents with the aim of targeting a new area where other interactions might be made within the binding site. For example, in the case of an Asp in the binding site, one might begin an in silico exercise by docking all the fragments (for Asp) provided from this study into the binding site. Our fragments will represent a more focused (for Asp) subset compared to the common fragments generated from known drug databases. Many docking programs and scoring functions have been reviewed for their performance in fragment-based drug design.[46,47] Next, one could grow the fragment into a single ligand using one of several algorithms, such as SPROUT,[48] FieldStere,[49] or ReCore.[50]

In addition, the statistics from the study provide powerful suggestions. For example, when targeting Arg for successful hydrogen bonding, O-mediated ligands should be used. The modification of fragments, for the creation of de novo fragments, or ease of synthesis, or accommodation of other substituents, could be easily done by using chemist's intuition and experience or by computational analysis such as bioisostere analysis. However, if 3D binding site information is available, our fragment atlas will add weight to a 3D computational approach.

Conclusions

We have gathered hydrogen bonding information between ligands and four types of protein amino acid side chain from the PDB database. The vast amount of data retrieved enables us to analyze the fragments that form hydrogen bonds with the amino acid side chain in question. Our results enable researchers to quickly generate hypotheses regarding binding fragments for a given binding site.

The choice of our fragment definition was designed to maximize the information retrieval of drug like fragments (amino acids and peptides were excluded). This analysis also identifies binding from diverse heterocyclic fragments such as observed for the Asp carboxylate group.

From the perspective of drug design, the similarity in residue utilization at binding sites for unrelated proteins observed in this study indicates that limits may exist to the possible types of interactions with other molecules. For example in the case of trypsin, a benzamidine group is the most frequently utilized fragment to bind to Asp in the active site. Consequently, some types of chemical structures should be favored for interaction with a macromolecule. Such limitations may account for the observed presence of particular substructural fragments across pharmacologically diverse classes of chemicals that elicit their action by blocking a recognition site. Alternatively, the appearance of common fragments may only reflect the limited variability of currently available chemical libraries from which drugs are derived. The results of this paper will be useful for directing new synthetic chemical efforts based on knowledge from the past.

In summary, this analysis of chemical fragments provides information that is useful for drug design, especially in the

area of fragment-based design, chemical library design, and selection of screening compounds. Research into the fragment preferences of different amino acids for other interactions, such as hydrophobic and cation−π would complement this study.

Materials and Methods

Data Set. The data set for this study was compiled in April 2009 and was taken to be all released PDB entries at that time, solved by X-ray crystallography to a resolution of 2.1 Å or better, and containing proteins in complex with small-molecule ligands. This gave a set of approximately 26000 structural models. Hydrogen bond interactions between the ligands and specific protein side chains in these complexes were extracted from the data files in PDBsum using a series of in-house Perl scripts. PDBsum is a web atlas of all PDB entries and provides various structural analyses of the models, including schematic LIGPLOT diagrams of protein−ligand interactions. The interactions are calculated using the HBPLUS program,[51] which identifies potential hydrogen bonds and nonbonded contacts. These data are stored in text files which are publicly accessible from the PDBsum[13] web site. The data were scanned to identify all ligands interacting with any of the four side chains of interest. For each ligand, its 3D coordinates, together with those of the interacting side chain were extracted from the parent PDB file and translated into MDL SD format[52] for further processing.

Filtering Rules for Ligands. The side chains examined in this study were those of amino acids Asp, Glu, Arg, and His. The analysis of this paper focused on lead-like ligands. There are many definitions of lead-like properties. Among them, the Lipinski Rule of Five (Ro5) is probably the best known and most used. The Ro5 is MW ≤ 500, number of hydrogen bond donors (HD) ≤ 5, number of hydrogen bond acceptors (HA) (or number of N and O together) ≤ 10, and logP ≤ 5. Oprea and his group have also extensively analyzed lead-like and drug-like properties. In one of their recent papers,[53] they included in their data set compounds from phases I to III. Their properties extend to a larger range of values, e.g., MW can be as large as 760 and clogP > 6. In our study, we wanted to capture more compounds that satisfied lead-likeness. In Table 6, rules 1−7 have reflected this selection.

Interacting ligands were filtered to leave those considered most relevant to medicinal chemists and were identified using the filtering rules shown in Table 6. These include selection of physical and chemical properties (Table 6, rules 1−7) to ensure lead-like properties and rules to exclude artifacts such as solvent molecules and impurities (rules 8−10). Other irrelevant macromolecules (rules 11−13) were also removed. The filtering was done using scripts written for the cheminformatics software package Pipeline Pilot. A full list of excluded ligands and het groups is provided in Supporting Information Table S5. For each ligand, the "interacting fragment" was then manually defined, as illustrated in Figure 1, and the counts of each fragments' hydrogen-bonded interactions with the four side chains of interest in this study were compiled and analyzed.

Fragment Identification and Extraction. The process of identification of fragments was manual. Initially the interaction between a ligand and the protein was visually examined using the LIGPLOT (2D) diagrams in PDBsum, in 3D using RasMol (again from the relevant PDBsum ligand page), and the text file generated by LIGPLOT detailing the H-bond interactions (bond length, interacting atoms, etc) as mentioned in the section of Data Set. Once all the fragments had been identified and classified they were drawn with ISISDraw and saved as mol, SMILES, and sd files. Retrospective computer programs were written after this effort. In general, two Pipeline Pilot scripts were written for each amino acid separating the N- and O-mediated interactions. For Asp and Glu, two extra Pipeline Pilot scripts were written to separate the 2 and 1-oxygen interactions. This script performed a series of 2D substructure searches (using the Substructure Filter from File component) from the mol files generated above and gave statistics of the frequency of fragments observed (using the Generate Frequencies component). For example (as illustrated in Figure 6), for the set of ligands that has O-mediated interaction with Arg side chain, the script will read in the SMILES strings (using the SMILES Reader component) of the ligands that has only O-mediated interaction with Arg side chain, then the program performs substructure searches of all the fragments we have reported in Supporting Information Table S3.2 and S3.4. The scripts allowed us to examine if any ligand had not been categorized by fragment type. Each ligand will only be categorized once. In addition, a series of C-programs were written. They use the mol files generated and performed a graph matching to the original ligand to retrieve the atom names lost during the conversion to mol files and allow the analysis of which fragments H-bond to which side chains and generate the tables in the Supporting Information.

Table 6. Filter[a] Criteria Used to Retrieve Relevant Ligands from the PDB

rule	criteria
1.	100 < molecular weight (Da) < 800
2.	number of atoms > 5
3.	only atoms H, C, N, O, P, S, F, Cl, Br, I
4.	number of oxygen and nitrogen atoms < 16
5.	number of hydrogen donor atoms < 8
6.	number of rotatable bonds < 16
7.	not metal ion or inorganic compound, such as AlF_3
8.	not a common solvent used in X-ray, such as GOL (glycerol), EDO (1,2-ethanediol), TRS (2-amino-2-hydroxymethyl-propane-1,3-diol)
9.	not an impurity or unknown, such as UNX, UNK, UNL, ARG, O, C, N
10.	not negative ion, such as NO_3^- (nitrate), SO_4^{2-} (sulfate)
11.	not an amino acid
12.	not a common sugar or lipid
13.	not a common cofactor, such as ATP, ADP, SAM, SAH, NAD, and FAM

[a] A complete list of filtered ligands and het groups is included in Table S5 in the Supporting Information.

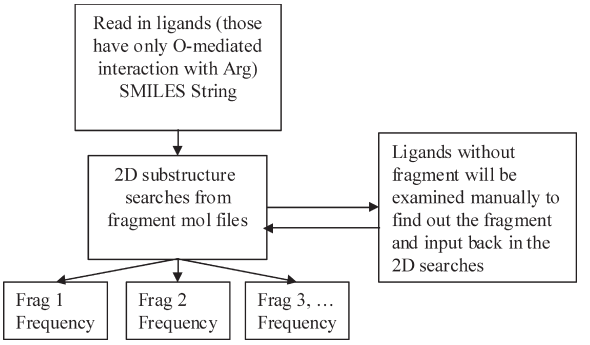


Figure 6. Flowchart to illustrate fragment 2D substructure search process in Pipeline Pilot.

Acknowledgment. We thank Paul Gane for useful discussions. This work was supported by Cancer Research UK grant no. C19746/A9867.

Note Added after ASAP Publication. This paper was published on March 15, 2010 with an incorrect version of the synopsis graphic and Figure 1. The revised version was published on March 18, 2010.

Supporting Information Available: Tables S1−S4 listing all fragments, their frequency of occurrence, related ligand codes, PDB codes, and their PDBsum links for Asp, Glu, Arg, and His, respectively. Table S5 lists the excluded ligands and het groups. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Congreve, M.; Chessari, G.; Tisi, D.; Woodhead, A. J. Recent developments in fragment-based drug discovery. *J. Med. Chem.* **2008**, *51*, 3661–3680.
(2) Hajduk, P. J.; Greer, J. A decade of fragment-based drug design: strategic advances and lessons learned. *Nat. Rev. Drug Discovery* **2007**, *6*, 211–219.
(3) Warr, W. A. Fragment-based drug discovery. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 453–451.
(4) Murray, C. W.; Rees, D. C. The rise of fragment-based drug discovery. *Nature Chem.* **2009**, *1*, 187–192.
(5) Joseph-McCarthy, D. Challenges of fragment screening. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 453–458.
(6) See examples in special issue: Fragment-based drug discovery. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 449−620.
(7) Clark, M.; Meshkat, S.; Talbot, G. T.; Carnevali, P.; Wiseman, J. S. Fragment-based computation of binding free energies by systematic sampling. *J. Chem. Inf. Model* **2009**, *49*, 1901–1913.
(8) Mauser, H.; Stahl, M. Chemical fragment spaces for de novo design. *J. Chem. Inf. Model.* **2007**, *47*, 318–324.
(9) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.
(10) Lameijer, E. W.; Kok, J. N.; Back, T.; Ijzerman, A. P. Mining a chemical database for fragment co-occurrence: discovery of "chemical cliches". *J. Chem. Inf. Model.* **2006**, *46*, 553–562.
(11) Hendlich, M.; Bergner, A.; Günther, J.; Klebe, G. Relibase—design and development of a database for comprehensive analysis of protein-ligand interactions. *J. Mol. Biol.* **2003**, *326*, 607–620.
(12) Laskowski, R. A. PDBsum: summaries and analyses of PDB structures. *Nucleic Acids Res.* **2001**, *29*, 221–222.
(13) PDBsum on the web. http://www.ebi.ac.uk/pdbsum/.
(14) Velankar, S.; Best, C.; Beuth, B.; Boutselakis, C. H.; Cobley, N.; Sousa Da Silva, A. W.; Dimitropoulos, D.; Golovin, A.; Hirshberg, M.; John, M.; Krissinel, E. B.; Newman, R.; Oldfield, T.; Pajon, A.; Penkett, C. J.; Pineda-Castillo, J.; Sahni, G.; Sen, S.; Slowley, R.; Suarez-Uruena, A.; Swaminathan, J.; van Ginkel, G.; Vranken, W. F.; Henrick, K.; Kleywegt, G. J. PDBe: Protein databank in Europe. *Nucleic Acids Res.* **2010**, *38*, D308–D317.
(15) Protein databank in Europe (PDBe) on the web. http://www.ebi.ac.uk/pdbe/.
(16) Kirchmair, J.; Markt, P.; Distinto, S.; Schuster, D.; Spitzer, G. M.; Liedl, K. R.; Langer, T.; Wolber, G. The Protein Data Bank (PDB), its related services and software tools as key components for in silico guided drug discovery. *J. Med. Chem.* **2008**, *51*, 7021–7040.
(17) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
(18) Chalk, A. J.; Worth, C. L.; Overington, J. P.; Chan, A. W. E. PDBLIG: classification of small molecular protein binding in the Protein Data Bank. *J. Med. Chem.* **2004**, *47*, 3807–3816.
(19) Verdonk., M. L.; Cole, J. C.; Taylor, R. SuperStar: A knowledge-based approach for identifying interaction sites in proteins. *J. Mol. Biol.* **1999**, *289*, 1093–1108.
(20) Moriaud, F.; Doppelt-Azeroual, O.; Martin, L.; Oguievetskaia, K.; Koch, K.; Vorotyntsev, A.; Adcock, S. A.; Delfaud, F. Computational fragment-based approach at PDB scale by protein local similarity. *J. Chem. Inf. Model.* **2009**, *49*, 280–294.
(21) Vietg, M.; Siegel, M. G.; Higgs, R. E.; Watson, I. A.; Robertson, D. H.; Savin, K. A.; Durst, G. L.; Hopskind, P. A. Characteristic physical properties and structural fragments of marketed oral drugs. *J. Med. Chem.* **2004**, *47*, 224–232.
(22) Villar, H. O.; Kauvar, L. M. Amino acid preferences at protein binding sites. *FEBS Lett.* **1994**, *349*, 125–130.
(23) Bartlett, G. J.; Porter, C. T.; Borkakoti, N.; Thornton, J. M. Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.* **2002**, *324*, 105–121.
(24) Moreira, I. S.; Fernandes, P. A.; Ramos, M. J. Hot spots—a review of the protein−protein interface determinant amino acid residues. *Proteins* **2007**, *68*, 803–812.
(25) Wallace, A. C.; Laskowski, R. A.; Thornton, J. M. LIGPLOT: a program to generate schematic diagrams of protein−ligand interactions. *Protein Eng.* **1995**, *8*, 127–134.
(26) Muegge, I. Selection criteria for drug-like compounds. *Med, Res. Rev.* **2003**, *23*, 302–321.
(27) Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A "Rule of three" for fragment-based lead discovery? *Drug Discovery Today* **2003**, *8*, 876–877.
(28) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug. Delivery Rev.* **1997**, *23*, 3–25.
(29) Onufriev, A.; Case, D. A.; Ullmann, G. M. A novel view of pH titration in biomolecules. *Biochemistry* **2001**, *40*, 3413–3419.
(30) Trujillo, J. I.; Huang, H.-C.; Neumann, W. L.; Mahoney, M. W.; Long, S.; Huang, W.; Garland, D. J.; Kusturin, C.; Abbas, Z.; South, M. S.; Reitz, D. B. Design, synthesis, and biological evaluation of pyrazinones containing novel P1 needles as inhibitors of TF/VIIa. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 4568–4574.
(31) Quan, M. L.; Lam, P. Y.; Han, Q.; Pinto, D. J.; He, M. Y.; Li, R.; Ellis, C. D.; Clark, C. G.; Teleha, C. A.; Sun, J. H.; Alexander, R. S.; Bai, S.; Luettgen, J. M.; Knabb, R. M.; Wong, P. C.; Wexler, R. R. Discovery of 1-(3′-aminobenzisoxazol-5′-yl)-3-trifluoromethyl-*N*-[2-fluoro-4- [(2′-dimethylaminomethyl)imidazol-1-yl]-phenyl]-1*H*-pyrazole-5-carboxyamide hydrochloride (razaxaban), a highly potent, selective, and orally bioavailable factor Xa inhibitor. *J. Med. Chem.* **2005**, *48*, 1729–1744.
(32) Baglin, T. P.; Carrell, R. W.; Church, F. C.; Esmon, C. T.; Huntington, J. A. Crystal structure of native and thrombin-complexed heparin cofactor II reveal a multistep allosteric mechanism. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 11079–11084.
(33) Brenk, R.; Naerum, L.; Gradler, U.; Gerber, H. D.; Garcia, G. A.; Reuter, K.; Stubbs, M. T.; Klebe, G. Virtual screening for submicromolar leads of tRNA-guanine transglycosylase based on a new unexpected binding mode detected by crystal structure analysis. *J. Med. Chem.* **2003**, *46*, 1133–1143.
(34) Cantor, C. R.; Schimmel, P. R. *Biophysical Chemistry, Part I*; W. H. Freeman: San Francisco, 1980; pp 49.
(35) Cotton, F. A.; Hazen, E. E., Jr.; Legg, M. J. Staphylococcal nuclease: Proposed mechanism of action based on structure of enzyme-thymidine 3′,5′-bisphosphate-calcium ion complex at 1.5-Å resolution. *Proc. Natl. Acad. Sci. U.S.A.* **1979**, *76*, 2551–2555.
(36) Kim, E. E.; Wyckoff, H. W. Structure of alkaline phosphatases. *Clin. Chim. Acta* **1990**, *186*, 175–187.
(37) Tamaru, S.; Hamachi, I. Recent progress of phosphate derivatives recognition utilizing artificial small molecular receptors in aqueous media. In *Recognition of Anions*; Vilar, R., Ed; Springer: Berlin, Heidelberg, 2008; pp 95−127.
(38) Luscombe, N. M.; Laskowski, R. A.; Thornton, J. M. Amino acid−base interactions: a three-dimensional analysis of protein−DNA interactions at an atomic level. *Nucleic Acids Res.* **2001**, *29*, 2860–2874.
(39) Harding, M. M. The architecture of metal coordination groups in proteins. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2004**, *60*, 849–859.
(40) Wang, G.; Dunbrack, R. L., Jr. PISCES: a protein sequence culling server. *Bioinformatics* **2003**, *19*, 1589–1591.
(41) Chothia, C.; Lesk, A. M. The relation between the divergence of sequence and structure in proteins. *EMBO J.* **1986**, *5*, 823–826.
(42) Politzer, P.; Lane, P.; Concha, M. C.; Ma, Y.; Murray, J. S. An overview of halogen bonding. *J. Mol. Model.* **2007**, *13*, 305–311.
(43) Politzer, P.; Murray, J. S.; Concha, M. C. Halogen bonding and the design of new materials: organic bromides, chlorides and perhaps even fluorides as donors. *J. Mol. Model.* **2007**, *13*, 643–650.
(44) Zhou, P.; Zou, J.; Tian, F.; Shang, Z. Fluorine bonding—how does it work in protein−ligand interactions? *J. Chem. Inf. Model.* **2009**, *49*, 2344–2355.
(45) Brown; I. D. Hydrogen bonds. In *The Chemical Bond in Inorganic Chemistry: The Bond Valence Model*; International Union of Crystallography Monographs on Crystallography, 12; Oxford University Press: Oxford, UK, 2006; pp 75−90.
(46) Hubbard, R. E.; Chen, I.; Davis, B. Informatics and modeling challenges in fragment-based drug discovery. *Curr. Opin. Drug Discovery Dev.* **2007**, *10*, 289–297.
(47) Vangrevelinghe, E.; Rudisser, S. Computational approaches for fragment optimization. *Curr. Comput.-Aided Drug Des.* **2007**, *3*, 69–83.
(48) Boda, K.; Johnson, A. P. Molecular complexity analysis of de novo designed ligands. *J. Med. Chem.* **2006**, *49*, 5869–5879.
(49) Cheeseright, T. The identification of bioisosteres as drug development candidates. *Innovations in Pharmaceutical Technology*, March, **2009**, 22−26.
(50) Maass, P.; Schulz-Gasch, T.; Stahl, M.; Rarey, M.; Recore, A Fast and versatile method for scaffold hopping based on small molecule crystal structure conformations. *J. Chem. Inf. Model.* **2007**, *27*, 390–399.
(51) McDonald, I. K.; Thornton, J. M. Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* **1994**, *238*, 777–793.
(52) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244–255.
(53) Oprea, T. I.; Allu, T. K.; Fara, D. C.; Rad, R. F.; Ostopovici, L.; Bologa, C. G. Lead-like, drug-like or pub-like: how different are they? *J. Comput.-Aided Mol. Des.* **2007**, *21*, 113–119.