JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

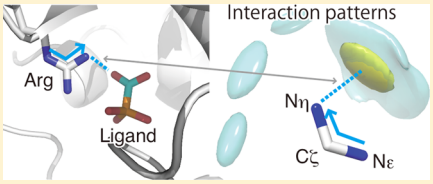
Comprehensive Classification and Diversity Assessment of Atomic Contacts in Protein−Small Ligand Interactions

Kota Kasahara,[†] Matsuyuki Shirota,[†,‡] and Kengo Kinoshita*[,†,‡,§]

[†]Department of Applied Information Sciences, Graduate School of Information Sciences, [‡]Tohoku Medical Megabank Organization, and [§]Institute of Development, Aging, and Cancer, Tohoku University, Miyagi 980-8597, Japan

Supporting Information

ABSTRACT: Elucidating the molecular mechanisms of selective ligand recognition by proteins is a long-standing problem in drug discovery. Rapid increase in the availability of three-dimensional protein structural data indicates that a data-driven approach for finding the rules that govern protein−ligand interactions is increasingly attractive. However, this approach is not straightforward because of the complexity of molecular interactions and our inadequate understanding of the diversity of molecular interactions that occur during ligand recognition. Thus, we aimed to provide a comprehensive classification of the spatial arrangements of ligand atoms based on the local coordinates of each interacting "protein fragment" consisting of three atoms with covalent bonds in each amino acid. We used a pattern recognition technique based on the Gaussian mixture model and found 13 519 patterns in the spatial arrangements of interacting ligand atoms, each of which was described as a Gaussian function of the local coordinates. Some typical well-known interaction patterns such as hydrogen bonds were ubiquitous in several hundred protein families, whereas others were only observed in a few specific protein families. After removing protein sequence redundancy from the data set, we found that 63.4% of ligand atoms interacted via one or more interaction patterns and that 25.7% of ligand atoms interacted without patterns, whereas the remainder had no direct interactions. The top 3115 major patterns included 90% of the interacting pairs of residues and ligand atoms with patterns, while the top 6229 included all of them.



INTRODUCTION

Drug discovery is an arduous process. A significant contributor to the problems encountered during drug discovery is the lack of high efficacy compounds.[1] Thus, there is a major need to design high efficacy compounds during the early stages of drug discovery. In recent years, structure-based drug design that examines the chemical space for discovering new compounds, guided by three-dimensional structural data of target proteins has been extensively studied.[2−5] However, this approach has only been partially successful because of our inadequate understanding of the molecular mechanisms of ligand recognition by proteins. In particular, one of the most intricate questions relates to protein selectivity, i.e., how each protein can recognize only few specific types of compounds from numerous candidate molecules with limited variations in the amino acid residues that form the binding pocket. Complex ligand recognition may be described based on combinations of the atomic contacts between amino acids and ligands. However, the interactions are highly complex at such an elemental level, and the diversity of such atomic interactions is not well-understood.

The Protein Data Bank (PDB)[6] is the primary resource for elucidating the diversity of atomic contacts in protein−ligand interactions, and many statistical analyses of molecular interactions have been performed using this database. This type of approach is attractive because PDB has been growing rapidly due to several major structural genomics projects in recent years.[7−10] However, the vast wealth of structural data also makes it difficult to extract information or knowledge, and hence, new methods are required to acquire insights into molecular interactions using the structural database.[11]

Previous analyses of protein−ligand interactions can be divided into a predefined classification-based approach and an unsupervised approach. In the first approach, molecular interactions are classified based on predefined geometrical and chemical patterns. For example, Panigrahi and Desiraju reported the propensities of strong and weak hydrogen bonds between proteins and ligands based on criteria related to distances between hydrogen and acceptor atoms and angles between donor, hydrogen, and acceptor atoms.[12] This approach is promising because it is easy to implement, and the results can be interpreted directly, although this approach can only be applied when the interactions are limited and predefined. However, it is not practical to list all protein−ligand interactions before analyses because many different types of interactions are used for recognizing specific molecules. Therefore, an alternative approach is necessary that does not specify the predefined classification of interactions.

The unsupervised approach can analyze a wide range of interactions without the need for predefined classifications. One of the main approaches in this category is based on the statistics of pairwise interatomic distances known as "statistical potential" or "knowledge-based potential". These potential functions have

Received: August 13, 2012

   dx.doi.org/10.1021/ci300377f | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

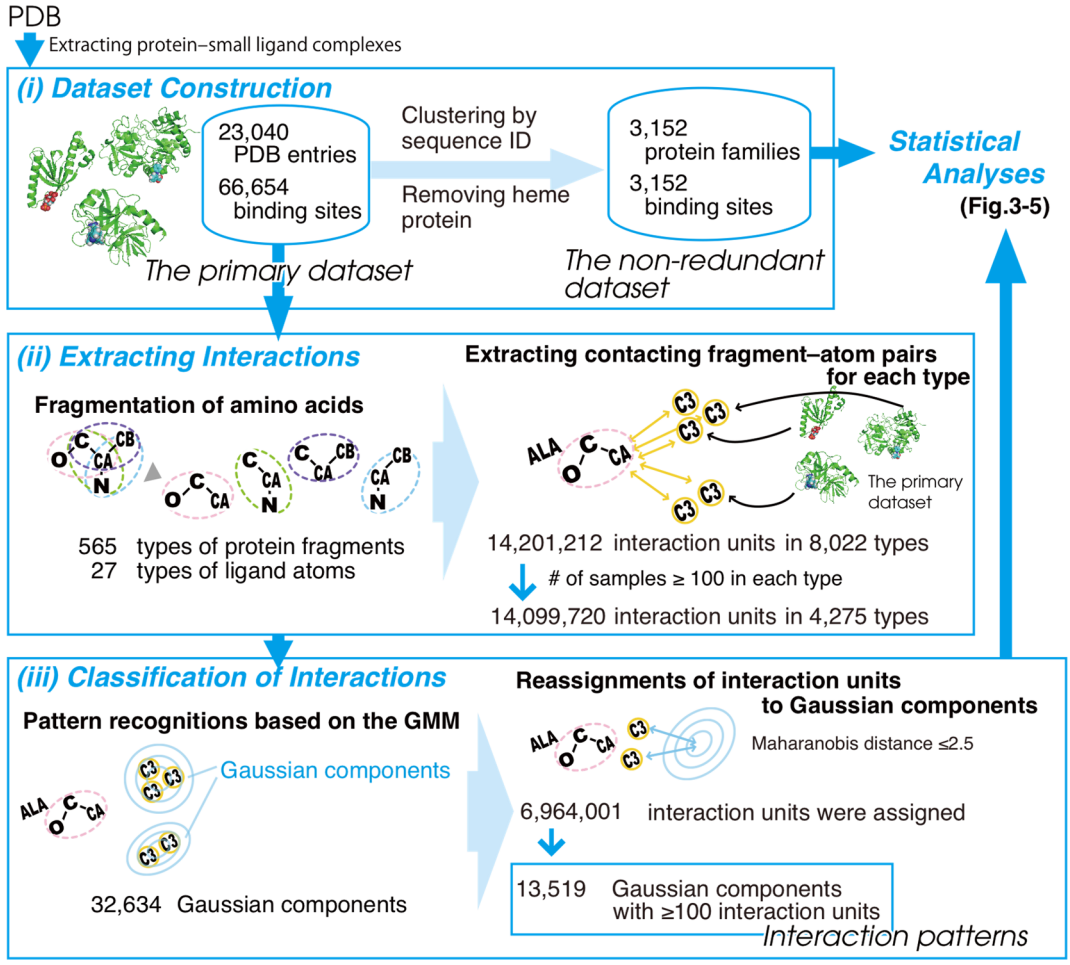


Figure 1. Method overview. The methods consist of the following steps. First, (i) two data sets were constructed. The *primary data set* consists of protein−small ligand complexes extracted from the PDB. From this data set, the *nonredundant data set* was constructed based on the clustering of protein sequence identities (in addition, complexes with heme ligands were removed). Second, (ii) information of interactions was extracted from the primary data set. In this step, amino acid residues were decomposed into fragments consisting of covalently linked three atoms. A unit of interactions was defined as a contacting pair of a protein fragment and a ligand atom. These interaction units were extracted from the data set, and they were superimposed onto the local coordination systems in each type of interaction units. Next, (iii) the interaction units were classified into patterns. Statistically overrepresented spatial arrangements of interaction units were found from superimposed distribution of interaction units, by using a pattern recognition technique based on the GMM. Each of the overrepresented interactions was defined as a Gaussian component. Then, assignments of all interaction units to each Gaussian component were performed. If a Gaussian component is assigned ≥100 interaction units, the Gaussian component is referred as an *interaction pattern*. Here, we obtained a catalogue of interaction patterns. Some statistical analyses based on this catalogue were performed; how diverse interactions are there, how diverse proteins use common interactions for the recognitions, and what kinds of interactions are preferred were discussed. In addition, applicability of the catalogue was evaluated with docking simulations.

been used for scoring protein−ligand complexes in docking tasks,[13−16] and they are useful for docking studies, although protein−ligand interactions do not always behave in an isotropic manner. Thus, a consideration of only the interatomic distances may be insufficient and an explicit consideration of the spatial arrangements of interacting pairs is necessary. To overcome this difficulty, several studies have used the spatial distributions of protein/ligand atoms around the interacting partner atoms in complexes.[17−23] These studies used statistical information to predict interacting atoms, although the knowledge extracted from the database was not summarized sufficiently well to uncover the diversity of protein−ligand interactions because the preferences of proteins and their ligand atoms were summed to generate the final prediction score.

In contrast, there have been some early attempts to obtain knowledge from the PDB directly. Imai et al. classified the interactions of some amino acid residues by using a binary vector to encode interacting ligand atoms.[24] Wang et al. reported the types of amino acid residues that preferred to interact with each ligand fragment.[25] These analytical studies have provided some useful insights into the propensities of protein−ligand interactions, but they have ignored majority of the information in the structural data. Thus, these approaches have not addressed the spatial arrangements of interacting pairs explicitly and instead interactions were discretized roughly to make interpretations.

In this study, we obtained a comprehensive classification of the spatial arrangements of interacting pairs of amino acid residues and ligand atoms using an unsupervised parametric pattern recognition technique based on the Gaussian mixture model (GMM). Using this classification, we found diverse interaction patterns; however some protein families had "unique" interaction patterns that were not observed in other families. Moreover, we performed docking calculations and

dx.doi.org/10.1021/ci300377f | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

confirmed that interactions in patterns were more preferred by native structures than by decoys. This implies that the catalogue of interaction patterns contains information about what interactions are favorable in native binding modes and it can be applicable to predict binding modes and binding sites.

MATERIALS AND METHODS

This study aimed to clarify how a variety of atomic contacts are there in the protein−small ligand complexes in the PDB. The overview of the methods is shown in Figure.1.

Data Set Construction. In this study, *primary* and *nonredundant* data sets were prepared. The primary data set contained 23 040 PDB entries, including 49 361 polypeptide chains and 66 654 protein−ligand binding sites. They were extracted from a snapshot of PDB captured on September 25, 2010, using the following criteria: (i) crystal structures with resolution ≤2.5 Å, (ii) including at least one polypeptide with ≥30 amino acid residues, (iii) including at least one small ligand with >5 heavy atoms and a molecular weight of ≥80 Da and <800 Da, and (iv) relative accessible surface area of the ligand molecule between the complex form and the isolated state <0.6.

The nonredundant data set, which was a subset of the primary data set, was constructed as follows. First, all 23 040 PDB entries in the primary data set were classified into 3219 clusters by single-linkage clustering with a threshold of 25% sequence identity using the BLASTCLUST program. These clusters are referred to as *protein families* in this study. Second, a representative complex was selected from each family to maximize the ligand diversity in the data set as follows: (i) representatives of singleton families were determined trivially; (ii) protein families of others were sorted based on the order of the number of complexes in each family; (iii) a representative complex was selected from the protein families with lower number of complex structures, which minimized the average chemical similarity measured using the Tanimoto coefficient (calculated by OpenBabel[26] software with FP2 fingerprint) between the complex and those already selected; and (iv) the process was repeated for all families. It should be noted that 67 representative complexes containing heme (HEM or HEC of the three-letter ligand ID in PDB) were removed because interactions with heme ligands are specific for heme recognition (this point is discussed in the Results and Discussion) mainly due to the amino acid residues coordinating with the iron atom. Finally, 3152 complexes were selected in the nonredundant data set.

Spatial Distributions of the Protein Fragment−Ligand Atom Interactions. The spatial distributions of interactions were obtained using the primary data set by dividing proteins into fragments and collecting interacting fragment−ligand atom pairs, where the fragments were defined as every covalently linked three heavy atoms in amino acidss. The order of the three atoms was distinguished in this process, because interacting pairs were obtained when the first atom in a fragment made contact with a ligand atom. The protein atoms and their fragments were classified into several types. Each fragment was defined based on the type of the three atoms together with the amino acid name of the first atom (e.g., Gly:Cα−N−C, Asp:Oδ−Cγ−Cβ), where the type of a protein atom was defined based on the elements and their positions in each amino acid residue (e.g., "C," "Cα," "N," and "O" as backbones and the types "Cβ" and "Cγ" as side chains; see Supporting Information Figure S1 for a list of protein atom types). The types of ligand atoms were defined based on their Tripos force field[27] (Supporting Information Table S1). Combinations of fragment and ligand atom types were used to identify the units of interaction. These definitions of fragment and atom types are the key features that determine what information to be extracted from the database. In this study, we focused on differences in interactions among amino acid types and encoded them into the type definition of protein fragments. For ligand atoms, their physicochemical property was encoded as Tripos force field type.

The interatomic contacts between proteins and ligands were detected in all the protein−ligand complexes in the primary data set using the criterion that the interatomic distance was less than the sum of van der Waals radii and offset value (1.0 Å). The relative geometry of an interacting pair was transformed into the local coordination system defined by the three atoms in the protein fragment. The spatial distributions of the interactions were determined by gathering the positions of interacting atoms in the local coordination system for each type of interaction unit. We defined the distributions as the orthogonal coordination systems, in contrast to the Rantanen's work,[22] which were based on the polar coordination systems. Interaction units containing <100 samples were discarded from the following analysis due to a lack of statistical data.

Gaussian Mixture Model-Based Pattern Recognition. The interaction patterns were analyzed based on the GMM in a manner similar to that applied by Rantannen et al.[23] who analyzed the spatial distributions of protein atoms around 30 types of predefined molecular ligand fragments using GMM. In contrast to their study, we analyzed the spatial arrangements of ligand atoms around protein fragments, which were defined collectively in an exhaustive manner, rather than using several predefined structures of fragments. In this model, we described the probability density function $p(x)$ of the event that a ligand atom at position x in the local coordinate system interacts with a protein fragment in that same frame of reference as

$$p(x) = \sum_{k=1}^{K} \pi_k \mathrm{N}(x|\mu_k, \Sigma_k)$$

where $\mathrm{N}(x|\mu_k, \Sigma_k)$ indicates a Gaussian function with the mean μ_k and the covariance matrix Σ_k, while π_k is the weight of a Gaussian function in GMM. The parameters $\mu_k\Sigma_k$ and π_k were inferred statistically based on the spatial distribution of interactions by maximizing the log-likelihood of the data points $\mathbf{X}$ in the distribution of this model using the variational Bayesian approach proposed by Attias,[28] where the variational approximation and the simultaneous probability density distribution that included all the parameters were considered to be multiples of several independent distributions (see the reference for full mathematical details). The maximum number of Gaussian components in each GMM, K, was set to 15. The number of Gaussian components was decreased during the learning processes by approaching parameter π_k to zero for unnecessary Gaussian components. This GMM parameter estimation was applied to all types of interaction units with ≥100 interactions.

Reassignment of Interactions to Patterns. Interactions in the spatial distributions were analyzed by decomposing them into a set of patterns. An interaction unit x was assigned to a Gaussian component g when the Mahalanobis distance between x and g, $D(x, g)$, was ≤2.5, as follows:

dx.doi.org/10.1021/ci300377f | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

$$D(x, g) = \sqrt{(x - \mu_g)^T \Sigma_g^{-1} (x - \mu_g)} \leq 2.5$$

Note that an interaction unit x could be assigned to more than one Gaussian component. In this study, the Gaussian components with $\pi_k \geq 0.01$ and with ≥100 interactions were considered to be interaction patterns, whereas other components were considered to be minor patterns and ignored.

Evaluation by Docking Simulations. For evaluation of the utility of interaction patterns, we performed docking simulations. The ratio of the number of interaction units in the patterns was calculated for the native and decoy structures, and their statistical significance was tested. The docking simulations were performed to the entire surface of proteins by using Autodock4.2.[29] As a data set, 3050 protein−ligand complexes were chosen from the nonredundant data set, by cutting out complexes that exhibit some difficulties during preparation process for docking simulations (for example, ligands including rare atoms that were not defined were omitted). The X-ray complex structures and docked structures with the root-mean-square deviation (RMSD) ≤ 2.5 Å are referred to as "native" structures; the other docked structures are referred to as "decoys".

RESULTS AND DISCUSSION

Construction of the Data Set and Identifying Interaction Patterns. We extracted 66 654 complexes from 23 040 PDB entries (the PDB snapshot was acquired on September 25, 2010) as described in Materials and Methods. The set of complexes was designated as the primary data set and contained 3219 protein families. The protein families were identified based on single-linkage clustering of sequence similarities with a threshold of 25% sequence identity. From the aspect of ligands, the data set contained 6842 kinds of ligands (distinguished by three-letter ligand ID). To evaluate the diversity of them, 3928 clusters were found based on single-linkage clustering of the 2D-fingerprint (FP2 fingerprint in OpenBabel[26]) with a threshold of 0.8 Tanimoto similarities. This data set contained quite diverse proteins and ligands.

All the protein−ligand complexes in the primary data set were decomposed into sets of protein fragment−ligand atom pairs with contacts. The ligand atoms were classified into 27 types based on their Tripos force field. Thus, 565 types of protein fragments were considered based on all three successive atoms in the amino acid residues, and 8022 types of fragment−atom pairs were identified as interacting unit types. After considering all possible combinations, we had 27 × 565 = 15 255 types of interaction units, hence approximately 47% of the possible pairs did not appear in the current version of PDB. This was possible because some types of ligand atoms, such as chlorine and fluorine, only appeared rarely in the data set. Of the 8022 types, 3747 types of interaction units were difficult to consider as significant "patterns" because of low observation frequencies. Most of the minor interactions arose from rare ligand atoms such as halogen, calcium, and sodium (Supporting Information Table S1, rows marked as "*"; see the columns entitled "no. interaction unit types" and "no. interaction unit types (≥100)"). We focused on the interaction unit types that included >100 samples, which yielded 18 types of ligand atoms and 413 types of protein fragments.

A pattern recognition technique was applied to each spatial distribution of the 4275 types of interaction units and 32 634 Gian components were determined. After reassigning each interaction unit to the Gaussian components and removing minor ones (<100 interaction units), we obtained 13 519 Gian components or *interaction patterns*. Details of the data set are summarized in Table 1, which includes a statistical description of the nonredundant data sets.

Table 1. Statistics for the Primary and Nonredundant Data Sets

data set	primary	nonredundant
PDB entries	23040	3152
binding sites	66654	3152
protein families (25% ID)	3219	3152
ligand variations (three-letter ID)	6842	895
interacting protein fragment types	565	462
ligand atom types	27	22
interaction unit types	8022	6129
interacting amino acid residues	832981	21564
ligand atoms	1332555	47152
interacting ligand atoms	1179357	42020
interaction units	14201211	508704
interaction patterns	13519	12619
protein fragment types in patterns	394	393
ligand atom types in patterns	17	11
interaction unit types in patterns	3003	2841
interacting residues in patterns	635597	22686
interacting ligand atoms in patterns	909551	29890
interaction units in patterns	6236420	186486

Protein Diversity in Interaction Patterns. Each interaction pattern contained a set of interaction units and each unit consisted of a protein fragment and ligand atom extracted from a complex; hence, we counted the number of families associated with each interaction pattern, which provided a good measure for evaluating the relationship between the protein diversity within the interaction patterns. As shown in Figure 2, most of the interaction patterns were commonly observed in several tens of protein families, and an interaction pattern contained protein fragments from 70.4 protein families on an average.

The most ubiquitous interaction pattern was bifurcated hydrogen bonding between a guanidinium group in an Arg side chain and a hydroxyl group, which was found in 713 protein families (Figures 2B and C). Similarly, the patterns of CH−π interactions between Trp and sp^3 carbon and that between Tyr and sp^3 carbon were found in 700 and 699 families, respectively (Figure 2D). They were present in >20% of the 3219 protein families in the data set. These ubiquitous patterns were not frequent, which is demonstrated by the long tail in Figure 2A, and only 97 patterns appeared in ≥500 protein families.

In contrast, some interaction patterns were found in only a few protein families, i.e., 105 interaction patterns were found in only a single protein family, which comprised 64 patterns in a family of heme-containing proteins, 14 in aldose reductases with NADP, 8 in dihydrofolate reductases, 7 in trypsin-inhibitor complexes, and 12 in other minor families. These patterns may be considered as family specific interaction patterns, although some family specific interaction patterns were discarded during our analyses because we rejected patterns with <100 samples. Family specific interactions were not considered to be energetically favorable if we assumed the Boltzmann distribution as the distribution of the interaction patterns. For heme binding, there were many family specific interaction patterns because heme is retained in a binding pocket by a covalent

dx.doi.org/10.1021/ci300377f | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

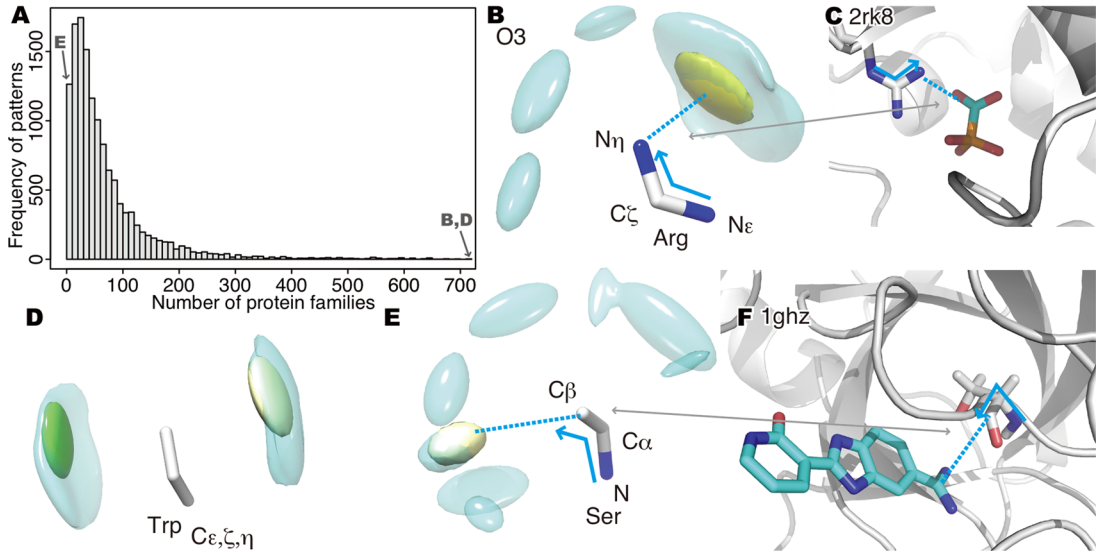


Figure 2. Ubiquitous and family specific interaction patterns. (A) Histogram of the interaction patterns for a number of protein families using each pattern. (B) Interaction patterns between an Arg:Nη−Cζ−Nε fragment and a saturated oxygen in a ligand. The contours in cyan indicate the probability density distribution of the interaction units modeled using the GMM (this is a linear combination of interaction patterns). The contour shown in yellow is the largest Gaussian component in this mixture model, i.e., the most widely used interaction pattern in diverse proteins. (C) Example of an interaction pattern shown in panel B (PDB ID: 2rk8). (D and E) Spatial distributions of interaction patterns between Trp:Cε−Cζ−Cη and sp^3 carbon and those between Ser:Cβ−Cα−N and sp^2 carbon, respectively. The interaction patterns shown in yellow and green are those described in the main text. (F) Example of the interaction units in the pattern shown in yellow in panel E (PDB ID: 1ghz).

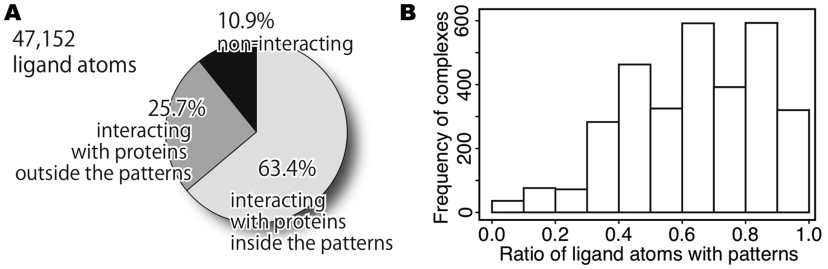


Figure 3. Coverage of interaction patterns of ligand atoms in the nonredundant data set with their assignment criteria, i.e., Mahalanobis distance ≤2.5. (A) Ratios of the three categories of ligand atoms: atoms interacting inside the patterns, those interacting outside the patterns, and those free of atomic contacts with their receptor. (B) Histogram of complexes versus the ratio of ligand atoms interacting in patterns.

bond between iron and a residue. Interactions between this residue and the four nitrogen atoms around the iron formed characteristic patterns only in heme-containing proteins. Another example of a family specific interaction was one of the patterns observed in 221 PDB entries in the trypsin family based on 62 types of inhibitors, which is shown in Figure 2E and F. This was a typical interaction pattern during the recognition of a $-C(NH_2)_2$ moiety by inhibitors via a Ser side chain in trypsin.

Ligand Recognition With/Without Patterns. After classifying the interaction patterns, we analyzed the statistics of the nonredundant data set to elucidate any general trends in the interactions, regardless of the protein families (Table 1). Initially, each ligand atom was classified into three categories (Figure 3A) to identify the contributions of interaction patterns to ligand atom recognition. The first category (63.4%) contained atoms that interacted with proteins inside the patterns, and the second category (25.7%) consisted of atoms that interacted with proteins outside the patterns, while the third category (10.9%) contained noninteracting atoms. If we assumed that higher frequencies indicated physical stability, the results suggested that most of the ligand atoms were recognized by physically favorable interactions, whereas about a quarter of the ligand atoms were not in physically suitable interactions. In addition to the atoms that made no contacts with proteins, the atoms in the latter two categories may function as linkers that construct scaffolds or that affect the physicochemical properties such as solubility and permeability, or they may play a role in tuning the binding properties.

When we focused on each complex, the ratio of ligand atoms with patterns was different for each complex, as shown in Figure 3B. Most of the complexes contained atoms with patterns but 28 of the complexes had no atoms, with typical interactions or a ratio of 0. Their ligands were either highly exposed to solvents or metal−sulfide clusters (the interaction patterns with metal atoms were not analyzed in this study).

Redundancy of Interaction Patterns. As described above, we identified 12 619 patterns, and 63.4% of the ligand atoms in the nonredundant data set interacted with these patterns (Figure 3A). In most of the complexes, more than half of the ligand atoms interacted with the patterns (Figure 3B), but it should be noted that one interaction unit could be assigned to several patterns. Indeed, each of the ligand atoms with patterns was involved in 6.58 interaction patterns on an

dx.doi.org/10.1021/ci300377f | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

average. Therefore, it was beneficial to determine the proportion of the 12 619 patterns defined that were unique, since this helped to understand the diversity of protein−ligand atomic contacts.

Furthermore, we calculated the coverage of ligand atomst with patterns and expressed it as the accumulated number of interaction patterns (Figure 4). We found that 63.4% of all

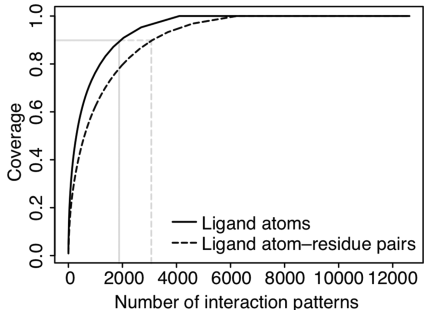


Figure 4. Coverage of interacting ligand atoms (the solid line) and residue−atom pairs (the dashed line) represented as the accumulated number of interaction patterns. The coverage = 1.0 (solid line) shows that all the ligand atoms with interaction patterns (63.9% of all ligand atoms, see Figure 3A) were involved in one of the corresponding interaction patterns (the horizontal axis). The dashed line shows the coverage of the pairs of amino acid residues and ligand atoms. The gray lines indicate the points where the coverage was 0.9.

ligand atoms had at least one interaction with one of the 12 619 patterns, but these atoms also had at least one of the frequently appearing 4104 patterns. Thus, we could not reach an overall proportion of 1.0 for the ligand atoms with 4104 patterns, as shown by the solid line in Figure 4. Furthermore, the top 1950 patterns covered 90% of the atoms with patterns. Similarly, we examined the proportion of the pairs of amino acid residues (not fragments) and ligand atoms. There were 109 976 interacting pairs and 47.0% (51 661) were assigned to at least one interaction pattern. The top 6229 patterns included all the pairs, while the top 3115 patterns included 90% of the pairs.

Our pattern recognition study found a diverse range of interaction patterns, but they were redundant; thus the number of unique patterns was limited. The 3115 patterns included most of the interacting pairs with statistically preferred spatial arrangements, which may indicate that the diversity of protein−ligand atomic contacts was around this number, or at most 6229 patterns.

Recognition Profiles of Ligand Atoms. We determined the distribution of interacting amino acids for each ligand atom type in the interaction patterns, which we refer to as *recognition profiles*. We defined a recognition profile for each type of ligand atom, which was the ratio of the number of ligand atoms interacting with 20 types of amino acids in patterns relative to the total number of ligand atoms with the considering atom type in the data set, as follows:

$$\text{profile} = \frac{N_{\text{LA}(i),\text{interact}(r)}}{N_{\text{LA}(i)}}$$

where $N_{\text{LA}(i)}$ is the total number of ligand atoms of type i in the data set, while $N_{\text{LA}(i),\text{interact}(r)}$ is the number of ligand atoms of type i interacting with residue type r. Also, r is one of 20 regular amino acids and i is one of 11 types of ligand atoms, which appeared frequently and were biologically important, i.e., saturated sp^3 carbon (C3), unsaturated sp^2 carbon (C2), aromatic carbon (CAR), saturated oxygen (O3), unsaturated oxygen (O2), amino (N3), amide (NAM), aromatic ring group (NAR), amino group with delocalized lone pair (NPL), phosphorus (P3), and sulfur (S3; the codes in parentheses were the type names defined in Tripos force field[27]). When two or more amino acids had interactions with a single ligand atom, they were counted for each amino acid; hence, the total of the ratio may exceed 1.0.

The 11 recognition profiles are shown in Figure 5. Their trends were roughly similar among the atom types with the

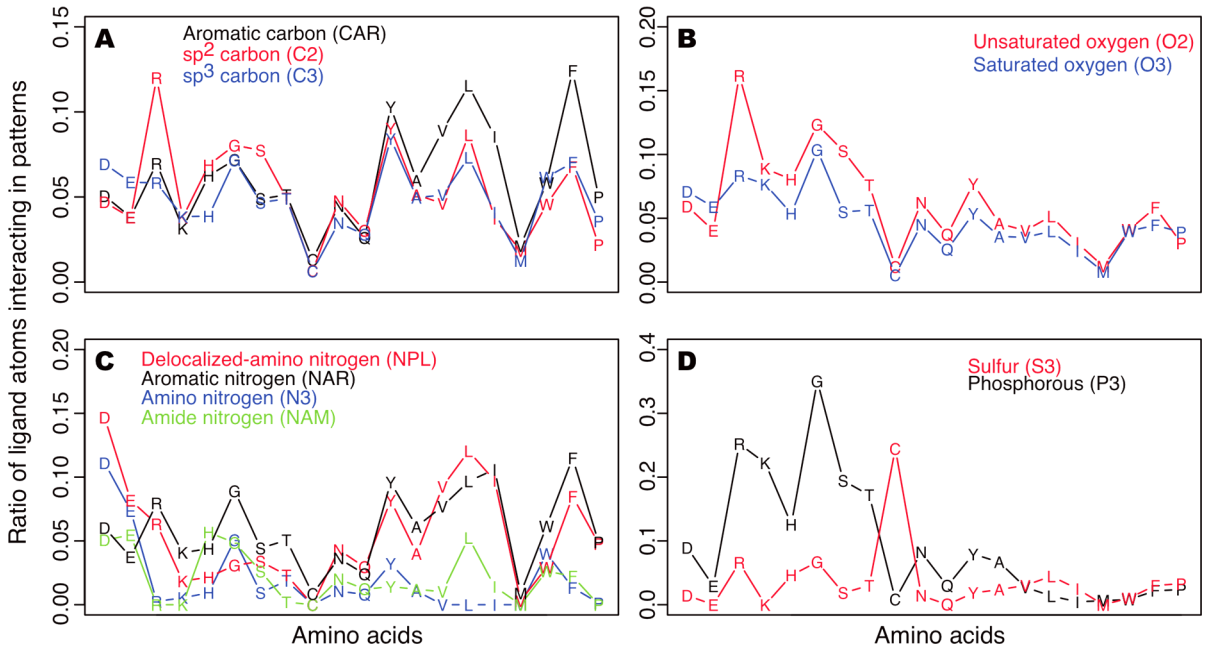


Figure 5. Ratio of ligand atoms in patterns with each type of amino acid residue for numerous ligand atoms in the same type, including noninteracting types.

dx.doi.org/10.1021/ci300377f | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

same elements. Carbon was preferred by hydrophobic groups such as Phe, Tyr, and Leu, while aromatic carbons were extraordinarily enriched during their interactions with them (Figure 5A, black line). This was due to $\pi-\pi$ interactions between ligand aromatic carbon and Phe (Figure 6B). In contrast, sp^2 carbon preferred interactions with the edge of the π-plane of Phe (Figure 6C). CH$-\pi$ interactions occurred with the aromatic carbon atoms of Leu residues (Supporting Information Figures S3K, S3L, and S3M). Among the carbon atoms, sp^2 carbon had a special preference for Arg (Figure 5A, red line), which indicated the bifurcated hydrogen bonds between a carboxyl group of a ligand and a guanidinium group (Figure 6F). Cation$-\pi$ interactions were favored during the recognition of aromatic carbon (Figure 6E).

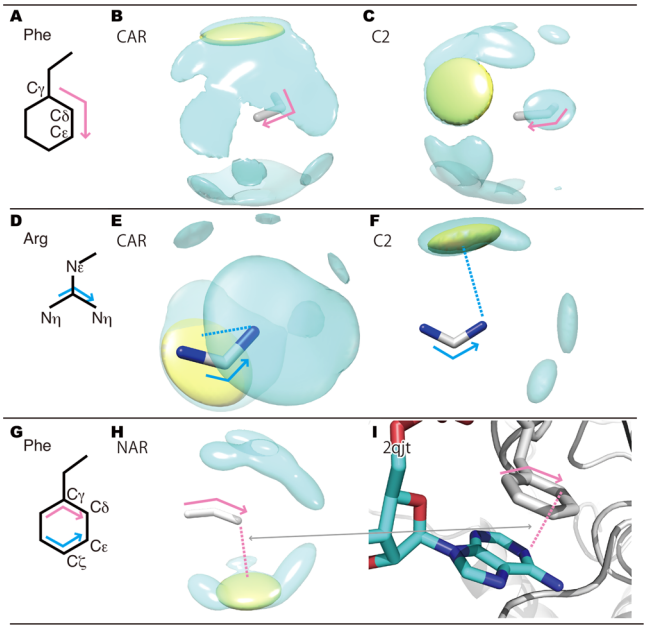


Figure 6. Examples of the interaction patterns preferred by each atom type. (A) Definitions of the protein fragments are shown in panels B and C. (B and C) The spatial distributions of GMM for interactions between the fragment Phe:Cε−Cδ−Cε and aromatic/sp^2 carbon. The contour shown in yellow is the most preferred pattern in each distribution. (D−F) Interactions between the fragment Arg:Nη−Cζ−Nη and aromatic/sp2 carbon for panels A−C. (G and H) Interactions between the fragment Phe:Cδ−Cγ−Cδ and aromatic nitrogen. (I) Example of the interacting units in the pattern shown with a yellow contour in panel H.

Polar residues were preferred by saturated oxygen, (Figure 5B, red line), while Arg residues were highly preferred by unsaturated oxygen due to the interactions between the carboxyl and guanidinium groups, as mentioned above (Figure 6F shows an example of the redundancy of interaction patterns discussed in the previous subsection). Arg also made frequent hydrogen bonds with saturated oxygen, and the interaction was the most common, as shown in Figures 2B and 2C.

During recognition of nitrogen atoms, delocalized amino nitrogen and aromatic nitrogen were preferentially recognized by hydrophobic amino acids (Figure 5C, red and black lines), whereas others (amino nitrogen and amide nitrogen) did not use interactions with hydrophobic amino acids during recognition (Figure 5C, blue and green lines). However, the latter types of nitrogen atoms were not observed frequently in the data set. These interactions with hydrophobic residues were mediated by $\pi-\pi$ and CH$-\pi$ contacts with a nitrogen-containing heterocycle in the ligands, which was found frequently in nucleotides (Figure 6H and I).

Sulfur was recognized specifically by Cys, which appeared mainly in the interaction with metal−sulfide clusters. For the recognition of phosphorus atoms, we found strong patterns by Gly to alpha phosphate in nucleotide ligands with the binding motif known as a P-loop[30] (see the Supporting Information section S8 and Figure S6A−C).

The Supporting Information provides further details on the recognition of each type of ligand atom.

Interaction Patterns were Enriched in the Native Structures over Docking Decoys. We assumed that interactions in patterns are preferred in the ligand recognitions. To evaluate this point, enrichment of interactions in patterns in native structures compared with docking decoys were tested. For the 3050 protein−ligand complexes, 570 docked structures with RMSD ≤ 2.5 Å and 12 423 structures with RMSD > 2.5 Å were obtained. The former 570 docked structures together with the 3050 crystal structures were referred to as "native structures", and the other structures were referred to as "decoys". For each structure, the ratio of number of interactions in patterns over the total number of interaction units was calculated (see Figure 7). As a result, distributions of the ratios were clearly different between the native and decoy structures (the average ratios were 0.84 and 0.49, respectively). This result indicates that the interactions in patterns were enriched in the native protein−ligand binding modes, and the validity of our assumption was confirmed. The interaction patterns defined in this work have a potential for application of a new knowledge-based scoring method to predict binding sites and binding modes.

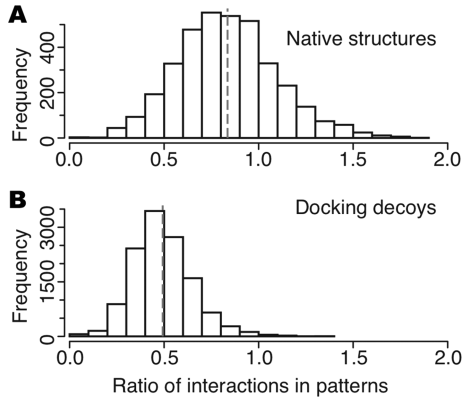


Figure 7. Ratio of the number of interaction units in patterns over the total number of interaction units. As an interaction unit can be assigned to more than one pattern, the ratio can exceed 1.0. (A) Histogram of the ratio for native structures, consisting of X-ray structures and docking structures with ≤2.5 Å in RMSD. (B) That for decoy structures, that are docking structures with >2.5 Å in RMSD. The dashed lines shows average values of the ratios, that were 0.84 and 0.49 for native and decoys, respectively.

CONCLUSION

This study used an unsupervised pattern recognition technique based on GMM to determine the spatial distributions of contacting ligand atoms around a three-atom protein fragment and the atomic contacts were classified into 13 519 patterns. After assessing the uniqueness of each interaction pattern by

dx.doi.org/10.1021/ci300377f | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

counting the number of protein families using each pattern, we found the most common interaction patterns among over 700 protein families, while many patterns were found unique to only one family. The ligands were recognized by common and family specific interactions.

We found that 63.9% of the ligand atoms in the nonredundant data set had at least one of the interaction patterns, 25.5% of the atoms interacted without patterns, while the remaining 10.6% of ligand atoms did not interact with proteins. In most of the complexes, over half of the ligand atoms were recognized by one or more patterns.

The classification of interactions was highly redundant and interacting pairs could be assigned to more than one pattern. The top 3115 main interaction patterns included 90% of the interacting residue−atom pairs with at least one pattern, while the remaining 10% were included an additional 3114 interaction patterns. Thus, the remaining 6390 patterns were considered to be redundant. After rejecting interactions that did not follow any pattern, most (90%) of the binding modes for the ligand recognition of proteins could be described using a combination of 3115 interactions patterns, while a maximum of 6229 patterns were required for all of the binding modes. The diversity of residue−atom interactions was limited.

■ ASSOCIATED CONTENT

Supporting Information

Figure S1: Definitions of atom types. Table S1 and Figure S2: Additional statistics on the primary and nonredundant data sets. Text and Figures S3−7: More detailed analysis of recognition by each element. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Tel.: +81-22-795-7179. Fax: +81-22-795-7179. E-mail: kengo@ecei.tohoku.ac.jp.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was funded by the "HD-Physiology" Grant-in-Aid for Scientific Research in Innovative Areas (22136005). The supercomputing resource was provided by the Human Genome Center (University of Tokyo). A tool for the visualization of the contours of the Gaussian functions was provided by Dr. Takeshi Kawabata.

■ REFERENCES

(1) Kola, I.; Landis, J. *Nat. Rev. Drug Discov.* **2004**, *3*, 711−716.
(2) Dailey, M. M.; Hait, C.; Holt, P. A.; Maguire, J. M.; Meier, J. B.; Miller, M. C.; Petraccone, L.; Trent, J. O. *Exp. Molec. Pathol.* **2009**, *86*, 141−150.
(3) Kalyaanamoorthy, S.; Chen, Y.-P. P. *Drug Discovery Today* **2011**, *16*, 831−839.
(4) Taboureau, O.; Baell, J. B.; Fernández-Recio, J.; Villoutreix, B. O. *Chem. Biol.* **2012**, *19*, 29−41.
(5) Schaffhausen, J. *Trends Pharmacol. Sci.* **2012**, *33*, 223.
(6) Berman, H.; Henrick, K.; Nakamura, H.; Markley, J. L. *Nucleic Acids Res.* **2007**, *35*, D301−3.
(7) Thomas, C.; Terwilliger, D. S. S. Y. *Annu. Rev. Biophys.* **2009**, *38*, 371.
(8) Dessailly, B. H.; Nair, R.; Jaroszewski, L.; Fajardo, J. E.; Kouranov, A.; Lee, D.; Fiser, A.; Godzik, A.; Rost, B.; Orengo, C. *Structure* **2009**, *17*, 869−881.
(9) Weigelt, J. *Exp. Cell Res.* **2010**, *316*, 1332−1338.
(10) Terwilliger, T. C. *J. Struct. Funct. Genomics* **2011**, *12*, 43−44.
(11) Bissantz, C.; Kuhn, B.; Stahl, M. *J. Med. Chem.* **2010**, *53*, 5061−5084.
(12) Panigrahi, S. K.; Desiraju, G. R. *Proteins* **2007**, *67*, 128−141.
(13) Zhang, C.; Liu, S.; Zhu, Q.; Zhou, Y. *J. Med. Chem.* **2005**, *48*, 2325−2335.
(14) Muegge, I. *J. Med. Chem.* **2006**, *49*, 5895−5902.
(15) Ozrin, V. D.; Subbotin, M. V.; Nikitin, S. M. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 261−270.
(16) Yang, C.-Y.; Wang, R.; Wang, S. *J. Med. Chem.* **2006**, *49*, 5903−5911.
(17) Laskowski, R. A.; Thornton, J. M.; Humblet, C.; Singh, J. *J. Mol. Biol.* **1996**, *259*, 175−201.
(18) Bruno, I. J.; Cole, J. C.; Lommerse, J. P.; Rowland, R. S.; Taylor, R.; Verdonk, M. L. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 525−537.
(19) Verdonk, M. L.; Cole, J. C.; Taylor, R. *J. Mol. Biol.* **1999**, *289*, 1093−1108.
(20) Verdonk, M. L.; Cole, J. C.; Watson, P.; Gillet, V.; Willett, P. *J. Mol. Biol.* **2001**, *307*, 841−859.
(21) Boer, D. R.; Kroon, J.; Cole, J. C.; Smith, B.; Verdonk, M. L. *J. Mol. Biol.* **2001**, *312*, 275−287.
(22) Rantanen, V. V.; Denessiouk, K. A.; Gyllenberg, M.; Koski, T.; Johnson, M. S. *J. Mol. Biol.* **2001**, *313*, 197−214.
(23) Rantanen, V.-V.; Gyllenberg, M.; Koski, T.; Johnson, M. S. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 435−461.
(24) Imai, Y. N.; Inoue, Y.; Yamamoto, Y. *J. Med. Chem.* **2007**, *50*, 1189−1196.
(25) Wang, L.; Xie, Z.; Wipf, P.; Xie, X.-Q. *J. Chem. Inf. Model.* **2011**, *51*, 807−815.
(26) O'Boyle, , N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. *J. Cheminform.* **2011**, *3*, 33.
(27) Clark, M.; Cramer, R. D.; Van Opdenbosch, N. *J. Comput. Chem.* **1989**, *10*, 982−1012.
(28) Attias, H. Inferring parameters and structure of latent variable models by variational bayes. *UAI'99 Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, Stockholm, Sweden, July 30−Aug 1, 1999; pp 21−30.
(29) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. *J. Comput. Chem.* **2009**, *30*, 2785−2791.
(30) Kinoshita, K.; Sadanami, K.; Kidera, A.; Go, N. *Protein Eng.* **1999**, *12*, 11−14.