Chemocavity: Specific Concavity in Protein Reserved for the Binding of Biologically Functional Small Molecules

Shinji Soga,[†] Hiroki Shirai,[†] Masato Kobori,[†] and Noriaki Hirayama*[,‡]

Molecular Medicine Research Laboratories, Drug Discovery Research, Astellas Pharma Inc., 21 Miyukigaoka, Tsukuba, Ibaraki 305-8585, Japan, and Basic Medical Science and Molecular Medicine, Tokai University School of Medicine, 143 Shimokasuya, Isehara, Kanagawa 259-1143, Japan

Received April 2, 2008

The idea that there should be a specific site on a protein for a particular functional small molecule is widespread. It is, however, usually not so easy to understand what characteristics of the site determine the binding ability of the functional small molecule. We have focused on the concurrence rate of the 20 standard amino acids at such binding sites. In order to correlate the concurrence rate and the specific binding site, we have analyzed high-quality X-ray structures of complexes between proteins and small molecules. A novel index characterizing the binding site based on the concurrency rate has been introduced. Using this index we have identified that there is a specific concavity designated as a chemocavity where a specific group of small molecules, i.e., canonical molecular group, is highly inclined to be bound. This study has demonstrated that a chemocavity is reserved for a specific canonical molecular group, and the prevalent idea has been confirmed.

INTRODUCTION

Protein functions either through interactions with other biological macromolecules, mainly proteins, or small organic molecules such as amino acids, nucleic acids, sugars, and drugs. The interaction between a protein and a small molecule occurs mainly on a surface of the protein. The relevant site on the protein is usually a concavity. The accurate and efficient recognition between the protein and the small molecule is essential for the biological reaction. To realize such highly specific and efficient recognition, the characteristics of the binding site should be well formulated for the small molecule. In other words, a specific concavity must be reserved for the corresponding small molecule. Now we have many quality X-ray structures of complexes between biologically important proteins and small molecules at our disposal for detailed inspection. Not all the structures of biologically important proteins have been determined yet. However, thanks to the rapid growth of the number of protein structures in the Protein Data Bank (PDB)[1] in recent years, a sufficient number of proteins can be employed to extract a general principle regarding specific concavities closely related to biological functions.

The idea that there should be a specific site on a protein for a particular functional small molecule is widespread. Various studies have been undertaken for recognizing common functional sites in proteins. The nucleotide-binding site is one of the typical functional sites observed frequently in proteins. This site is rich in several specific amino acids and shares a common structural pattern.[2,3] As the rate at which three-dimensional structures of proteins are solved increases so rapidly, the number of protein structures without functional annotations is growing. As a result, methods for recognizing protein function from structure are more important now. Most methods currently available are reliant upon the three-dimensional arrangement of amino acids around the relevant site.[4,5]In addition, binding sites for only specific groups of small molecules have been taken into account. Shulman-Peleg *et al.*[4] focused on proteins that bind adenine, estradiol, ATP, and fatty acid. On the other hand, Zhang *et al.*[5] found the most highly connected binding sites for ten specific ligands. Although these studies have demonstrated that a specific site in a protein is recognized by a particular small organic molecule, they were confined to relatively small sets of proteins and ligands. Therefore it is highly required to expand these previous studies to generalize and confirm the prevalent idea that a specific site in a protein is reserved for a particular group of ligands. The shape and the properties of such specific sites in proteins can be determined by the amino acids flanking the cavity. Several studies have been undertaken in order to find the conserved surface structures in a set of similar proteins by translating the cavity-flanking amino acids into a set of pseudocenters.[6,7] Since these methods utilize spatially conserved physicochemical properties, they are relatively sensitive to the predefined interaction patterns and moreover computationally intensive.

Recently we have developed an index named propensity for ligand binding (PLB index) to identify the specific concavity in a protein for binding of druglike small molecules.[8,9] The PLB index is based on the characteristic appearances of the 20 standard amino acids at the concavity. Since this method is not strictly dependent on three-dimensional disposition of amino acids clustering at the concavity, it is computationally light. Despite its simplicity, it has proven very useful in identifying the specific concavity

* Corresponding author phone: +81 463 93 1121; e-mail: hirayama@is.icc.u-tokai.ac.jp.
[†] Astellas Pharma Inc.
[‡] Tokai University School of Medicine.

10.1021/ci800113c CCC: $40.75
Published on Web 07/22/2008

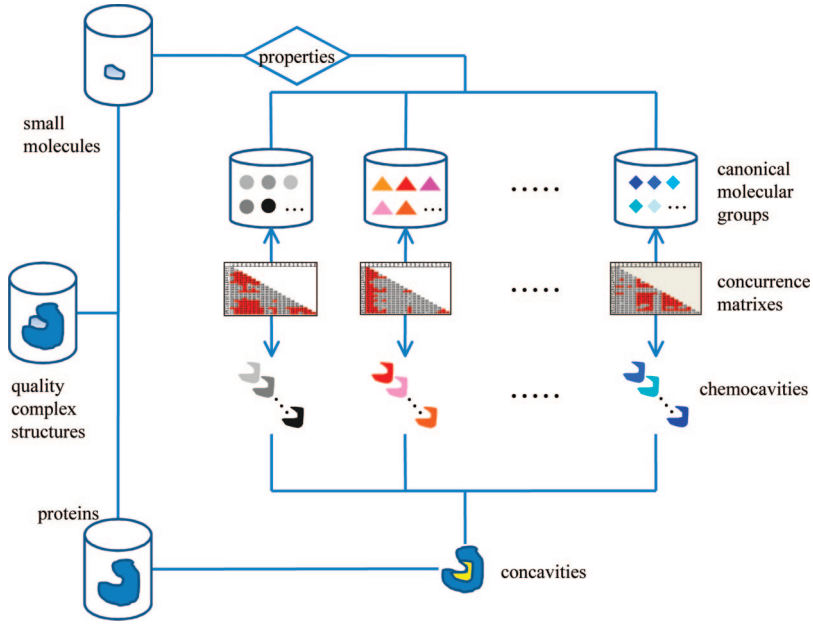

Figure 1. A schematic diagram of classification of canonical molecular groups and corresponding chemocavities. Concurrence matrixes link them.

on the surface of the protein. This concavity can be regarded as a reserved concavity for a promised small molecule. Hence, it is designated as chemocavity in this study. Chemocavity is defined as a concavity in a protein where a particular group of small molecules closely related to the function of the protein should be bound. This group of small molecules is expected to share common characteristics. In this paper, such a group is designated as a canonical molecular group. Suppose a chemocavity is an essential concavity for protein function, the corresponding canonical molecular group can bind the chemocavity so as to regulate the function of the protein. In this study, we have expanded the concept of the PLB index in order to make it sensitive enough to apply in identification of various chemocavities. In the expanded PLB index we introduced the concurrence rate of each of two of the 20 standard amino acids at the concavity. The expanded PLB index named chemocavity index has successfully identified different chemocavities corresponding to different canonical molecular groups. The present study unequivocally has shown that there is a specific chemocavity reserved in a protein structure for binding the corresponding canonical molecule which can control the function of the protein.

COMPUTATIONAL METHODS

The steps taken in this study are illustrated in Figure 1. The quality X-ray structures of complexes between proteins and small organic molecules were selected from PDB. The small organic molecules were classified according to their properties to obtain a set of canonical molecular groups. The proteins were then classified according to the canonical molecular groups that are bound to the proteins. The amino acids concurrence rates in the concavities that share the same canonical molecular group were analyzed. The amino acids concurrence rates can be expressed by concurrence matrixes. The characteristics of the concavities were expressed by the chemocavity indices based on the concurrence rates. The specific concavity for a particular canonical molecular group is designated as chemocavity. The correlations between various chemocavities were finally analyzed to check the independency of chemocavities and corresponding canonical molecular groups.

High Quality and Nonredundant X-ray Structures of Protein-Small Molecule Complexes. Reliable positions of non-hydrogen atoms in the concavities of proteins are highly required in order to accurately identify the amino acids at the concavities. A set of high quality X-ray structures of complexes between proteins and small molecules obtained from PDB on May 18, 2007 was used in this study. This quality data set was selected by use of the following criteria: the R_{free} value of less than or equal to 0.24, the resolution value of less than or equal to 2.5 Å, the occupancy factors of 1.0 for all non-hydrogen atoms of the small molecule, and the atomic displacement parameters of less than 50 Å^2 for all non-hydrogen atoms of the small molecule. The redundant proteins were removed from the data set by consulting the "Non-redundant PDB chain set (NRPDB)" resource [http://www.ncbi.nlm.nih.gov/Structure/VAST/nr-pdb.html] on May 18, 2007. A p-value of 10e-80 was used to judge sequence similarity by BLAST.[10] The homologous proteins above this threshold value were eliminated from the data set. Only the small organic molecules with molecular weight between 100 and 800 were taken into account. Finally, the 3621 complex structures including 4039 small organic molecules were selected and used as a data set for this study.

Selection of Canonical Molecular Groups. Canonical molecular groups were selected by clustering of the small organic molecules. Fingerprint and QuaSAR clustering methods implemented in the software system MOE[11] were used for this purpose. The fingerprint clustering algorithm based on the Jarvis-Patrick method[142] was applied using MACCS structural keys and Tanimoto coefficient. For QuaSAR clustering, molecular weight and SlogP[13] were used as molecular descriptors. By applying these two clustering methods, 4039 small organic molecules were brigaded to

1579 independent groups of small organic molecules. Since the groups with only a small number of member molecules are not appropriate in representing the canonical molecular groups, they are excluded from consideration in this study.

Chemocavity Index Considering Concurrence Rate of Two Standard Amino Acids at Concavity. The PLB index used in the previous studies is based on the frequency of the 20 standard amino acids at a concavity in order to judge its druggability. In this study, however, chemocavities corresponding to 48 canonical molecular groups must be distinguished. Therefore a more informative index is highly required. Gutteridge et al. have reported that combinations of different amino acid residues play important roles for catalysis at enzyme active site.[14] Therefore, based on the assumption that the concurrence rate of each of the two amino acids out of the 20 standard amino acids could augment the information, the concurrence rate of two amino acids was included in the expanded PLB index, now designated as the chemocavity index. The amino acids whose non-hydrogen atoms exist within 4.5 Å from the non-H atoms of the small molecule were used to determine the concurrence frequency of two amino acids at the concavity. Suppose the concurrence frequency of two amino acids x and y at the chemocavities corresponding to a canonical molecular group 'a' is $N_a(x,y)$, the concurrence frequency rate, $CA_a(x,y)$, is given as follows:

$$CA_a(x,y) = \frac{N_a(x,y)}{\sum_{l=1}^{20}\sum_{m=1}^{20} N_a(l,m)} \quad (1)$$

The denominator of the above equation means the total number of the concurrence frequency of every two amino acids at the concavities.

The incidence of every amino acid residue at all chemocavities in the whole data set was also calculated. The occurrence of a particular amino acid x in the chemocavities, $CA(x)$, can be defined as

$$CA(x) = \frac{N(x)}{\sum_{l=1}^{20} N(l)} \quad (2)$$

$N(x)$ denotes the number of amino acid x in the chemocavities. The denominator means the total number of all amino acids in the chemocavities. $CA(x)$ was determined using all structures in the data set.

If the occurrences of amino acids x and y are independent, $PA(x,y)$ defined in the following equation is an expected probability that the amino acid x and y appear concurrently.

$$PA(x,y) = W(x,y)CA(x)CA(y)$$
$$W(x,y) = \begin{cases} 1(x=y) \\ 2(x\neq y) \end{cases} \quad (3)$$

If the concurrence rate of the amino acids x and y at the chemocavities for the molecules belonging to canonical molecular group 'a' is more than expected, the following $RA_a(x,y)$ becomes greater than 1.

$$RA_a(x,y) = \frac{CA_a(x,y)}{PA(x,y)} \quad (4)$$

By use of a linear combination of RAs for all combinations of the two amino acids, an index expressing the tendency of a concavity to be a chemocavity 'a' is defined as follows. This chemocavity index means a propensity of a concavity i to be the chemocavity 'a'.

$$CC_{a,i} = \frac{\sum_{x=1}^{20}\sum_{y=1}^{20} N_i(x,y)RA_a(x,y)}{\sum_{x=1}^{20}\sum_{y=1}^{20} N_i(x,y)} \quad (5)$$

$N_i(x,y)$ denotes the concurrence frequency of the two amino acids x and y at a concavity i. If the chemocavity index for a concavity is high, the composition of amino acids at the concavity should be similar to that of the chemocavity. Hence, it is highly expected that the canonical molecules corresponding to this particular chemocavity should bind at this concavity.

Evaluation of Chemocavity Index. In order to assess the tendency of a concavity to be a particular chemocavity, the threshold value of $thCC_a$ was determined. The chemocavity indexes for concavities where small molecules belonging to a particular canonical molecular group were averaged to obtain the average positive CC_a index. On the other hand, chemocavity indices of concavities that do not bind the relevant canonical molecules were averaged to obtain the average negative CC_a index. The threshold value of $thCC_a$ was set to an average of these two values and used as a criterion to assess whether the concavity can be the chemocavity or not. To make the assessment simpler, 0 or 1 was assigned to each concavity as follows:

$$Z_{a,i} = \begin{cases} 1(CC_{a,i} \geq th(CC_a)) \\ 0(CC_{a,i} < th(CC_a)) \end{cases} \quad (6)$$

If the CC_a index is larger than or equal to the threshold value of CC_a, the concavity is regarded to be the chemocavity 'a'. Suppose the number of concavities in a set of concavities is N, the tendency of this set of concavities to be a particular chemocavity of 'a' is expressed as $(\sum_{i=1}^{N} Z_{a,i})/N$.

This value, designated as identification index, was used as an index to discriminate the different chemocavities belonging to different canonical molecular groups.

RESULTS AND DISCUSSION

Identification of 48 Canonical Molecular Groups. In this study, a set of the high-quality X-ray structures of complexes between proteins and small molecules was used as a source of information regarding the chemocavity and the canonical molecular group. By use of the initial set of 3621 complex structures, the 1579 independent groups of small molecules were selected based on the combination of chemical topology and physicochemical properties. For selection of molecular groups that well represent canonical molecular groups, only the groups consisting of more than nine molecules were included in the consideration. The total 48 independent groups were picked out and used in this study as canonical molecular groups. These 48 canonical groups represent functional small molecules that are highly inclined to bind their target proteins. These groups are given in Table 1. Typical biological small molecules, such as amino acids, nucleotides, and sugars, are well classified. Although buffer molecules, detergent molecules, and polyethylene glycols are not biological molecular groups, they are tightly bound at

Table 1. 48 Canonical Molecular Groups[a]

group #	molecular category	average molecular weight	average SlogP	# of compounds in the group
1	organic acid	148.07	−4.79	24
2	organic acid	189.32	−5.25	63
3	amino acid	145.18	−3.45	15
4	amino acid	208.01	−0.36	12
5	peptide	301.65	−5.59	21
6	amino sugar	207.17	−2.48	81
7	amino sugar	381.90	−5.19	11
8	amino sugar	410.40	−4.50	27
9	amino sugar	556.63	−6.44	10
10	sugar	150.13	−2.58	11
11	sugar	162.82	−2.14	12
12	sugar	169.16	−2.75	16
13	sugar	179.26	−3.18	58
14	sugar	194.18	−2.57	11
15	sugar	310.28	−4.34	10
16	sugar	341.17	−5.32	62
17	sugar	342.30	−5.40	20
18	sugar	501.19	−7.44	24
19	sugar	666.70	−9.77	17
20	sugar phosphate	225.90	−4.94	10
21	sugar phosphate	260.26	−5.48	14
22	nucleoside	267.25	−1.88	14
23	nucleoside	384.16	−3.75	42
24	nucleoside	397.39	−3.89	22
25	nucleotide	322.35	−4.71	11
26	nucleotide	343.98	−4.11	35
27	nucleotide	358.55	−6.68	35
28	nucleotide	401.49	−6.37	17
29	nucleotide	423.84	−5.77	72
30	nucleotide	441.15	−8.44	57
31	nucleotide	455.63	−3.53	55
32	nucleotide	470.83	−6.80	10
33	nucleotide	507.45	−7.29	75
34	nucleotide	518.67	−9.97	55
35	nucleotide	662.21	−6.27	99
36	nucleotide	733.99	−8.24	90
37	nucleotide	740.93	−11.27	11
38	nucleotide	759.50	−7.83	23
39	nucleotide	785.26	−5.48	89
40	thiamin	426.22	−2.73	13
41	pyridoxal phosphate	232.58	−1.16	52
42	porphyrin	563.46	1.64	198
43	higher fatty acid	265.14	4.56	10
44	buffer	195.58	−2.44	53
45	buffer	239.32	−4.69	32
46	detergent	282.55	−0.11	10
47	polyethylene glycol	150.17	−1.00	12
48	polyethylene glycol	207.56	−0.89	44

[a] The molecular group number corresponds to the chemocavity number.

specific concavities in proteins. Therefore they remain in the canonical molecular groups. It is possible that these molecules can be hints for some biological molecules. The average molecular weight and estimated octanol/water partition coefficient of SlogP of the member compounds in the 48 canonical groups are also given in Table 1. The list of all compounds included in the 48 canonical molecular groups is available in the Supporting Information. The nucleotides and sugars are the major groups in these 48 canonical groups. It is clearly due to the historical reason of protein crystallography. Many proteins with these groups of small molecules have been crystallized and their structures were determined.

Chemocavity. In the PLB index originally used for detecting the drug-binding site in the protein, only the frequency of each 20 standard amino acids was taken into account. The original PLB index was good enough for detecting the drug-binding site. In this study, however, we must characterize the concavities corresponding to the 48 canonical groups. Since the higher identification ability is required for such characterization, the expanded PLB index, termed chemocavity index, considering the concurrence rate of the two amino acids was formulated and applied in this

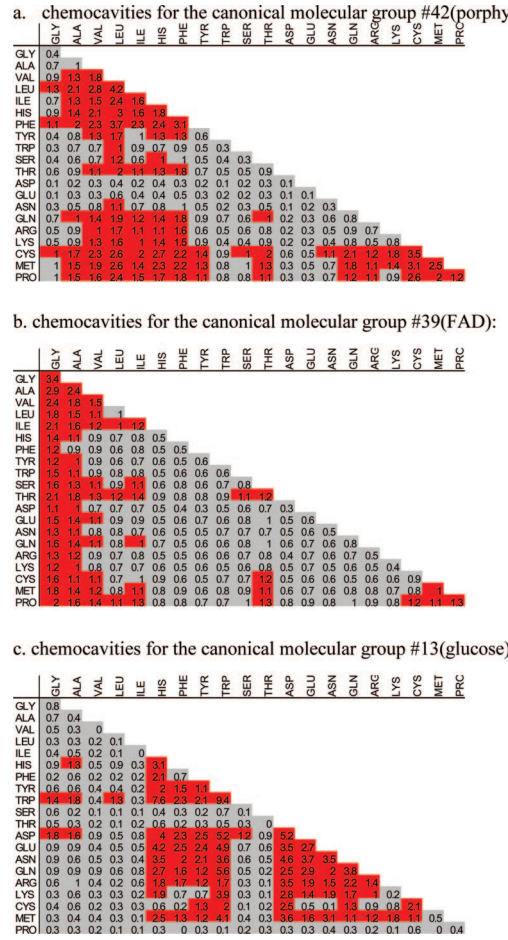

Figure 2. Concurrence matrixes of chemocavites for three canonical molecular groups including porphyrin, FAD, and glucose.

study. The concurrence rate of the two amino acids can be expressed by the concurrence matrix as shown in Figure 2. Each matrix element stands for a ratio of the concurrence frequency of a pair of two amino acids in a chemocavity to the frequency observed in other concavities on the surface of the proteins in the data set. To illustrate the characteristics of the concurrence matrixes of different chemocavities, the matrixes for three different chemocavities are compared in Figure 2. These chemocavities correspond to the canonical chemical groups of porphyrin, flavin adenine dinucleotide (FAD), and glucose. The concurrence frequencies greater than 1.0 are colored in red, and the concurrent pairs with the concurrence frequency being greater than 3.0 feature the most outstanding aspect of the chemocavity. The concurrence frequencies of Leu-Leu, Leu-Phe, Phe-Phe, Cys-Cys, and Cys-Met pairs are greater than 3.0 in the concavity of porphyrin. In the chemocavity of glucose, the concurrence frequencies of amino acids with charged and polar side chains are significantly high as expected. It is particularly interesting that the self-concurrence rate of Trp is exceptionally high. In addition, the concurrence rates of Trp with polar and charged amino acids such as Asn, Asp, Gln, Glu, Lys, and His are also markedly high. Although the self-concurrence rate of Gly is markedly high, it seems that the concurrence matrix of FAD is relatively featureless. As shown in this example, a concurrence matrix for a concavity on a protein where a specific small molecule is bound well represents the characteristics of each chemocavity.

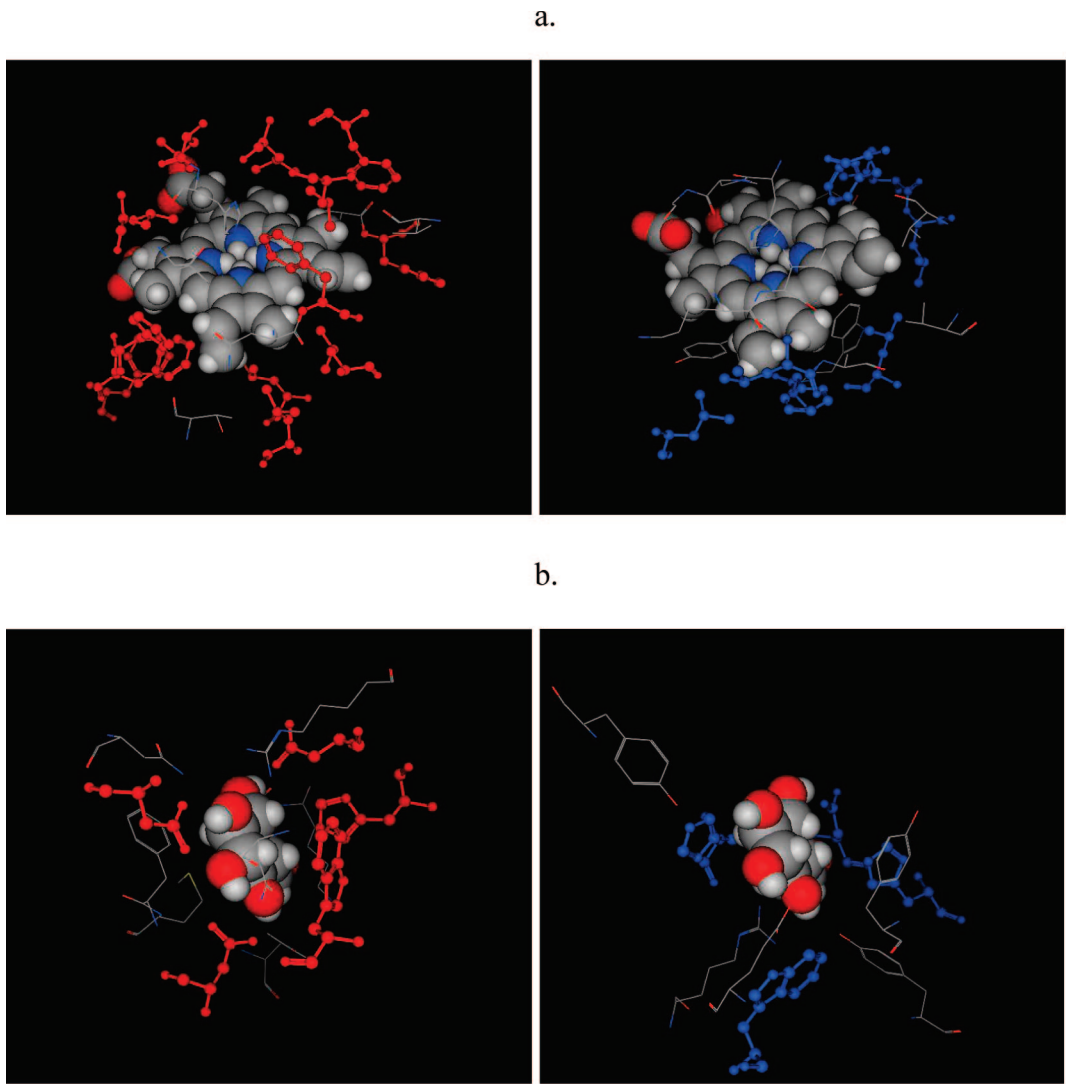

Figure 3. Concurrently existing amino acids at chemocavities of porphyrin and glucose. a. Porphyrin: The left and right structures are 1J3Y(2.57) and 1SOX(2.00), respectively. The values in the parentheses indicate the chemocavity indexes. Leu, Cys, Met, and Phe are depicted in red or blue. b. Glucose. The left and right structures are 2HPH(3.047) and 2BYO(2.759), respectively. The values in the parentheses indicate the chemocavity indexes. Trp, His, Gln, and Asp are depicted in red or blue.

It is particularly worth noting that the concurrence matrix does not directly reflect the three-dimensional structures of the concavities where small molecules are bound. Nevertheless, the concurrence matrix can characterize the chemocavity. This important feature of the concurrency of amino acids at the chemocavity is further illustrated in Figure 3. The chemocavities for porphyrin groups are rich in concurrence of Leu, Cys, Met, and Phe. In Figure 3a, these amino acids are colored in red and blue in the two different proteins. Although the three-dimensional arrangements of these amino acids around the porphyrin group are markedly different, the chemocavity indexes calculated from the concurrence rates of these amino acids clustered around the porphyrin group unequivocally suggest that these proteins share a common chemocavity for the porphyrin group. Another example of glucose is shown in Figure 3b. The chemocavity is rich in concurrence of Trp, His, Gln, and Asp. These amino acids shown in red and blue are clustered around glucose. The clustered amino acids adopt totally different three-dimensional arrangements. The chemocavity indexes, however, tell us that these concavities should belong to a chemocavity inherent to glucose.

Evaluation of Cross Interactions between Different Chemocavities. If each chemocavity is specific to a particular canonical molecular group, chemocavities corresponding to different canonical groups could be distinguished. We made an identification index based on chemocavity index as described in the method section. By use of this index, cross interactions between different chemocavities are evaluated. The results are shown in Figure 4. The chemocavity *a* is evaluated using the identification index based on the chemocavity *b* and vice versa. Therefore the (*a,b*) and (*b,a*) elements are not always symmetrical with respect to the diagonal line. The chemocavities are clustered according to the conventional classification of the corresponding canonical molecular groups. The main clusters are distinguished by different colors. The cross-element colored in red indicates the high identification index.

It is noteworthy that most of the diagonal elements are colored in red and off-diagonal elements are mostly white. Since the proteins that bind to small molecules belonging to a particular canonical molecular group are not homologous, the results clearly demonstrate that a chemocavity is highly reserved for the corresponding canonical molecular group.

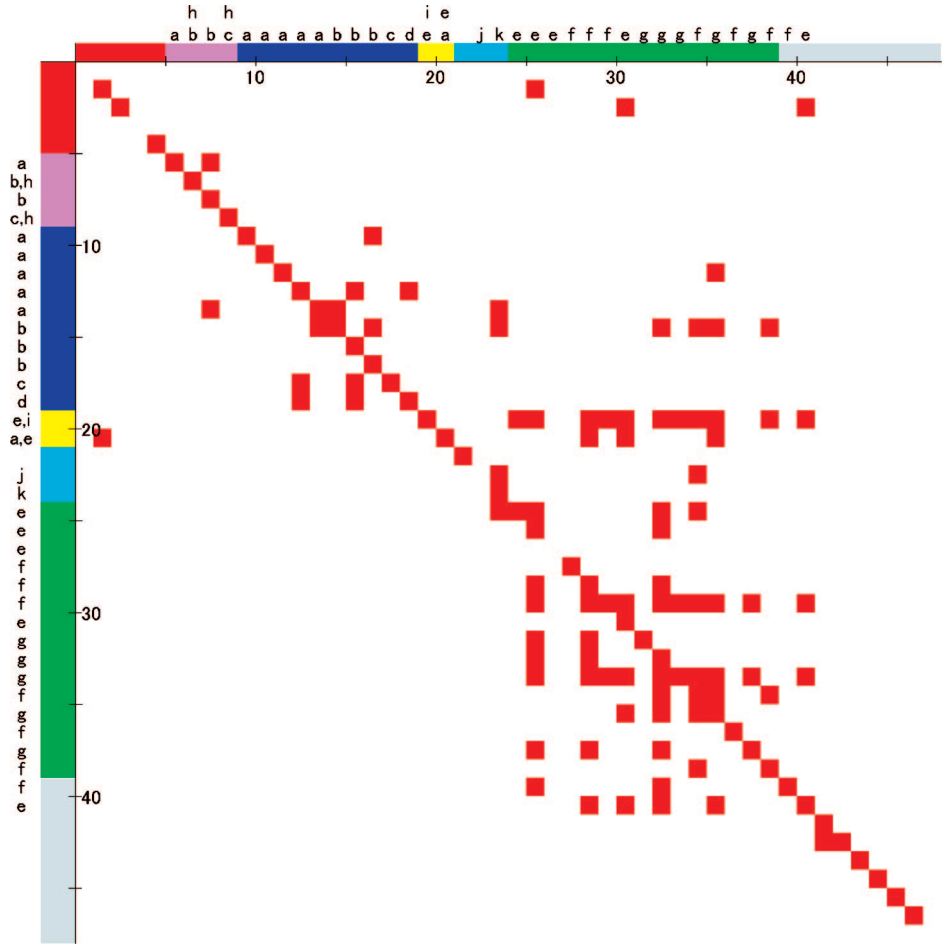

Figure 4. Chemocavity cross-interactions. Elements whose identification indexes are greater than or equal to 0.6 are indicated in red. The main clusters are distinguished by different colors. Color codes for the canonical molecular groups are as follows: orange, organic acid and amino acid (1 to 5); purple, amino sugar (6 to 9); blue, sugar (10 to 19); yellow, sugar phosphate (20 to 21); aqua, nucleoside (22 to 24); green, nucleotide (25 to 39); gray, others (40 to 48). Symbols *a* to *l* denote specific molecular groups included in each canonical molecular group: (*a*) monosaccharide, (*b*) disaccharide, (*c*) trisaccharide, (*d*) tetrasaccharide, (*e*) nucleotide monophosphate, (*f*) nucleotide diphosphate, (*g*) nucleotide triphosphate, (*h*) N-acetyl aminosugar, (*i*) open-chain form, (*j*) S-adenosyl-L-homocysteine, (*k*) S-adenosylmethionine, and (*l*) coenzyme.

In particular, the diagonal elements indicate the relevance of the chomocavity index. On the other hand, the off-diagonal elements are of interest to understand the independency of different chemocavities. Since most of the values of the off-diagonal elements are below the thresholds, major chemocavites are well discriminated to each other. Some cross-interactions are observed between different chemocavities for nucleotides. Significant cross-interactions are seen primarily in chemocavities corresponding to nucleotide triphosphates. Meanwhile, chemocavities for nucleotide monophosphates and diphosphates are well discriminated. The cross-interactions between nucleotide chemocavities and sugar phosphate chemocavities indicate that phosphate groups play important roles to gather specific amino acids into the chemocavities. On the contrary, the cross-interactions between chemocavities for nucleotides and sugars are marginal. Virtually no cross-interactions are observed between the nucleotide chemocavities and other chemocavities of non-nucleotides such as amino sugar, porphyrins, higher fatty acids, amino acids, peptides, and organic acids. Chemocavities regarding sugar groups also deserve special mention. The chemocavities corresponding to monosaccharide, disaccharide, and trisaccharide are well discriminated. It is interesting to note that the cross interactions of these chemocavities and those for amino sugar are not virtually observed.

In this study, we have demonstrated that a chemocavity is highly reserved for the corresponding canonical molecular group. By use of the identification index, we can differentiate a specific chemocavity from trivial concavities. Therefore, this method can be practically applicable to deduce a specific chemocavity in a protein whose ligand is still unknown. It may also be useful in finding a novel binding site for a particular small-molecule in a ligand-known protein. Since the concept of chemocavity is empirical-based, its accuracy and versatility solely depend on the completeness of the database of protein structures. With the increase of the number of proteins in PDB, the identification ability of this method will certainly improve.

CONCLUSIONS

We have introduced a chemocavity index based on the concurrence rate of the 20 standard amino acids at the concavity in proteins in order to identify a site in a protein inherently bound by a specific group of small organic molecules named as canonical molecular group. The chemocavity index has successfully correlated a particular

chemocavity to the corresponding canonical molecular group. But, we think it is still difficult to apply this method to any protein structures with no ligands, since accurate identification of small molecule binding sites are needed when searching against whole protein structures. That is to say, it is difficult to differentiate between small molecule interacting surface and noninteracting surface. The idea that there should be a specific site on a protein for a particular functional small molecule has been recognized widely. This study has confirmed that there should be a chemocavity in the protein to which a particular canonical molecule is highly inclined to be bound.

ACKNOWLEDGMENT

This work was partly supported by Grant-in-Aid for Scientific Research (A) (19200024) from MEXT (Ministry of Education, Culture, Sports, Science and Technology) for N.H.

Supporting Information Available: List of all compounds included in the 48 canonical molecular groups. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Berman, H. M.; Westbrook, J.; Fenz, 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.

(2) Walker, J. E.; Saraste, M.; Runswick, M. J.; Gay, N. J. Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *EMBO J.* **1982**, *1*, 945–951.

(3) Brakoulias, A.; Jackson, R. M. Towards a structural classification of phosphate binding sites in protein-nucleotide complexes: an automated all-against-all structural comparison using geometric matching. *Proteins* **2004**, *56*, 250–260.

(4) Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J. Recognition of functional sites in protein structures. *J. Mol. Biol.* **2004**, *339*, 607–633.

(5) Zhang, Z.; Grigorov, M. G. Similarity networks of protein binding sites. *Proteins* **2006**, *62*, 470–478.

(6) Davies, J. R.; Jackson, R. M.; Mardia, K. V.; Taylor, C. C. The Poisson Index: a new probabilistic model for protein ligand binding site similarity. *Bioinformatics* **2007**, *23*, 3001–3008.

(7) Konc, J.; Janezic, D. Protein-protein binding-sites prediction by protein surface structure conservation. *J. Chem. Inf. Model.* **2007**, *47*, 940–944.

(8) Soga, S.; Shirai, H.; Kobori, M.; Hirayama, N. Use of amino acid composition to predict ligand-binding sites. *J. Chem. Inf. Model.* **2007**, *47*, 400–406.

(9) Soga, S.; Shirai, H.; Kobori, M.; Hirayama, N. Identification of the druggable concavity in homology models using the PLB index. *J. Chem. Inf. Model.* **2007**, *47*, 2287–2292.

(10) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410.

(11) MOE (Molecular Operating Environment), version 2006.0801; Chemical Computing Group Inc.: Montreal, Quebec, Canada, 2006.

(12) Jarvis, R. A.; Patrick, E. A. Clustering using a similarity measure based on shared near neighbors. *IEEE Trans. Comput.* **1973**, *C22*, 1025–1034.

(13) Wildman, S. A.; Crippen, G. M. Prediction of Physiochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868–873.

(14) Gutteridge, A.; Thornton, J. M. Understanding nature's catalytic toolkit. *Trends Biochem. Sci.* **2005**, *30*, 622–629.

CI800113C