PROTEINS: Structure, Function, and Bioinformatics 65:593–606 (2006)

Specificity of Molecular Interactions in Transient Protein–Protein Interaction Interfaces

Kyu-il Cho,[1] KiYoung Lee,[2] Kwang H. Lee,[1,2] Dongsup Kim,[3]* and Doheon Lee[1]*

[1]*Bio-Information System Laboratory, Department of BioSystems, KAIST, Guseong-dong, Yuseong-gu, 305-701, Daejeon, Korea*
[2]*Bio-Information System Laboratory, Department of Electrical Engineering and Computer Science, KAIST, Guseong-dong, Yuseong-gu, 305-701 Daejeon, Korea*
[3]*Protein BioInformatics Laboratory, Department of BioSystems, KAIST Guseong-dong, Yuseong-gu, 305-701, Daejeon, Korea*

ABSTRACT **In this study, we investigate what types of interactions are specific to their biological function, and what types of interactions are persistent regardless of their functional category in transient protein–protein heterocomplexes. This is the first approach to analyze protein–protein interfaces systematically at the molecular interaction level in the context of protein functions. We perform systematic analysis at the molecular interaction level using classification and feature subset selection technique prevalent in the field of pattern recognition. To represent the physicochemical properties of protein–protein interfaces, we design 18 molecular interaction types using canonical and noncanonical interactions. Then, we construct input vector using the frequency of each interaction type in protein–protein interface. We analyze the 131 interfaces of transient protein–protein heterocomplexes in PDB: 33 protease-inhibitors, 52 antibody-antigens, 46 signaling proteins including 4 cyclin dependent kinase and 26 G-protein. Using kNN classification and feature subset selection technique, we show that there are specific interaction types based on their functional category, and such interaction types are conserved through the common binding mechanism, rather than through the sequence or structure conservation. The extracted interaction types are C^{α}—H···O=C interaction, cation···anion interaction, amine···amine interaction, and amine···cation interaction. With these four interaction types, we achieve the classification success rate up to 83.2% with leave-one-out cross-validation at $k = 15$. Of these four interaction types, C^{α}—H···O=C shows binding specificity for protease-inhibitor complexes, while cation–anion interaction is predominant in signaling complexes. The amine ··· amine and amine···cation interaction give a minor contribution to the classification accuracy. When combined with these two interactions, they increase the accuracy by 3.8%. In the case of antibody–antigen complexes, the sign is somewhat ambiguous. From the evolutionary perspective, while protease-inhibitors and sig-naling proteins have optimized their interfaces to suit their biological functions, antibody–antigen interactions are the happenstance, implying that antibody–antigen complexes do not show distinctive interaction types. Persistent interaction types such as π···π, amide-carbonyl, and hydroxyl-carbonyl interaction, are also investigated. Analyzing the structural orientations of the π···π stacking interactions, we find that herringbone shape is a major configuration in transient protein–protein interfaces. This result is different from that of protein core, where parallel-displaced configurations are the major configuration. We also analyze overall trend of amide-carbonyl and hydroxyl-carbonyl interactions. It is noticeable that nearly 82% of the interfaces have at least one hydroxyl-carbonyl interactions. Proteins 2006;65: 593–606.** © 2006 Wiley-Liss, Inc.

Key words: protein–protein interfaces; conservation; feature subset selection; k-nearest neighbor; classification; specific binding interaction; persistent interaction

INTRODUCTION

Proteins play a key role in controlling all the cellular processes. When controlling biological processes, proteins can specifically associate or dissociate depending on their local environments and physiological conditions. A number of studies have been carried out to investigate the fundamental principles of protein–protein interactions.

Many research groups have looked into the representative parameters of protein–protein interfaces such as interface size, shape, complementarity, residue propensity, hydrophobicity, segmentation, secondary structures, packing density. Using these representative parameters, they have made a comparative study to identify the fea-

The Supplementary Material referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/

Grant sponsor: National Research Laboratory; Grant number: 2005-01450.

*Correspondence to: Dongsup Kim, Protein BioInformatics Laboratory, Department of BioSystems, KAIST Guseong-dong, Yuseong-gu, 305-701, Daejeon, Korea. E-mail: kds@kaist.ac.kr or Doheon Lee, Bio-Information System Laboratory, Department of BioSystems, KAIST, Guseong-dong, Yuseong-gu, 305-701, Daejeon, Korea. E-mail: dhlee@biosoft.kaist.ac.kr

Received 12 December 2005; Revised 13 March 2006; Accepted 20 April 2006

Published online 31 August 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21056

tures that could potentially distinguish the different types of protein–protein interfaces.[1–9]

For example, in one study,[1] researchers have investigated the interfaces using parameters such as accessible surface area, chemical characters, packing density, and free energy of association, and have shown that hydrophobicity is the major factor for the stability of protein–protein interactions, while complementarity of hydrogen bonds and van der Waals' interactions are responsible for the selectivity. More extensive study has been carried out for the 15 protease-inhibitor complexes and 4 antibody–antigen complexes.[3] According to this study, the sizes of protein–protein interfaces are similar, and the interfaces in oligomeric proteins are usually more hydrophobic than those involved in functional recognition. In protein–protein complexes deposited in protein three-dimensional structural database PDB,[10] there are two different types of protein complexes; homocomplexes and heterocomplexes. Homocomplexes are usually considered as permanent and optimized, whereas heterocomplexes can also have such properties, or they can be transient. Several parameters, such as size, shape, complementarity, residue propensities, hydrophobicity, segmentation, secondary structure, and conformational changes, have been used to examine the characteristics between the two types of protein interfaces.[7] As the number of known three-dimensional structures has been increased, the more detail characterization among the transient protein–protein complexes has been carried out.[11,12] On average, the interfaces have chemical properties that are close to the average protein surface, and the same packing density as the protein interior, although the properties of individual complexes widely vary.[11] In addition, it has been shown that it is difficult to discriminate between different types of protein–protein complexes based on the physicochemical and geometrical interface properties such as the contact area, planarity, polarity, shape complementarity, and pair potential.[12] Moreover, although all these studies analyzed the proteins of known structure from PDB, their results were contradictory in some cases. For example, some reports argue that amino acid compositions of the interfaces from the different types of protein complexes are similar,[13–15] but others report significant differences.[8,11]

Other research groups took notice of the role of an individual residue in protein–protein interactions. Alanine scanning mutagenesis study[16] has shown that despite the large size of binding interfaces, individual single side chains can energetically contribute a large fraction of the binding free energy.[17,18] A computational alanine scanning was also performed to probe protein–protein interactions by calculating binding free energies.[19,20] More systematic analysis has been carried out using a alanine mutation database. It showed that, at the level of individual side chains, there is little correlation between buried hydrophobic surface area and free energy of binding, contrary to the results for whole surfaces. The free energy of binding is not evenly distributed across interfaces. Instead, there are hot spots of binding energy composed of a small subset of residues. Tryptophan, tyrosine and arginine are enriched in these hot spots and are surrounded by a shell of energetically less important residues that most likely serve to occlude bulk solvent from the hot spots.[14]

There were some efforts to correlate residues conservation in spatially similar environments with binging hot spots.[21–23] The structure of protein–protein interfaces has been analyzed using geometrical hashing technique.[24–26] The analysis of interface families showed a preference for conservation of polar residues at their interfaces. In addition, structurally conserved interface residues are strongly correlated with the experimentally identified hot spots.[21] The number of structurally conserved residues, particularly of high ranking energy hot spots, increase with the interface size, which implies that hot spots are effectively distributed within the interface rather than compactly clustered in them. Furthermore, the phenomenon that similar residue hot spots occur across different protein families may suggest that affinity and specificity are not necessarily coupled.[22] Simple analysis of conservation patterns of protein–protein interfaces with respect to the protein surface have shown that the interfaces have been relatively more conserved than the protein surface during evolution.[27,28]

Instead of looking into the protein structures at the level of amino acid residues or more coarse-grained level of descriptions, several studies were concentrated on the role of specific molecular interaction types such as hydrogen bonds, salt bridges, and hydrophobic interactions. Comprehensive studies with such type of approach have been carried out in a number of studies on protein stability, while a few studies on protein interfaces with limited scope have been reported. Although hydrogen bonds, salt bridges, and hydrophobic interactions are considered to be the major determinants of protein stability, in recent years noncanonical interactions, such as interactions involving π-system and C^{α}—H···O hydrogen bonds, have been shown to be of much greater importance than previously thought.[29–36] A comprehensive structural analysis of X[]H···π hydrogen bonding(X=N, O, S) in three-dimensional protein structure has been performed. It has shown that the most efficient π-acceptor is the side chain of Trp. Numerous examples are found where peptide X–H···π interactions play important roles in stabilization of helix termini, strand edges, β-bulges, and regular turns.[29] An investigation of the π—π stacking interactions in protein core has shown that the relative orientation is an off-centered parallel orientation.[30] A detail analysis of the extent and nature of cation—π interactions has been performed using energy-based criterion. It is demonstrated that when a cationic side chain (Lys or Arg) is near an aromatic side chain (Phe, Tyr, or Trp), the geometry is biased toward one that would experience a favorable cation—π interaction.[31] The geometrical analysis of cation—π and amide—π interactions has also been performed.[37–39] Recently, some research groups tried to investigate the role of noncanonical interactions in protein–protein interfaces. A statistical potential has

been developed to quantitatively describe the C^{α}—H···O hydrogen bonding interaction in the protein–protein interfaces. It suggests that the weak C^{α}—H···O hydrogen bond makes an important contribution to the association and stability of protein complexes.[33] The cation—π interactions involving arginine in protein–protein interface has been investigated.[36]

The values of the representative parameters mentioned earlier are largely overlapped, though on average, they show a tendency depending on types of protein complexes.[7,11,12] Consequently, they have limited power for distinguishing the interfaces of different types of protein complexes. Hot spot analysis can be used to distinguish interface from the remainder of surface, but it can have difficulties to distinguish the interfaces of different types of protein complexes. This is because similar residue hot spots occur across different protein family.[22]

Most of the studies about specific molecular interaction types have been focused on the core of the protein more than the protein–protein interface, primarily because the absolute size of data set of protein–protein interfaces is relatively small. Recently, the number of three dimensional crystal structures has been sharply increased with the development of structural genomics, but few attempts have been made to investigate the meaning of a specific interaction type in the context of protein functions. There was a paper that presents classifications between crystal contacts and biological contacts on the basis of feature vectors.[40] But, there are no attempts to correlate interaction types with function of proteins. The systematic analysis of protein–protein interfaces at the interaction level can give us more information about specificity and selectivity of protein–protein interactions.

In this study, we perform systematic analysis of the interfaces of different types of transient protein heterocomplexes at the molecular interaction level to find out whether there are some distinct interaction types specific to the functional category.

We think that the local environments or physiological conditions of a protein as well as the properties of a protein itself are important to the protein's specific function, and proteins in the same functional category perform their functions under the similar environments. Transient protein–protein interfaces associate or dissociate at least once during their cellular processes, and so they have more chance to reflect their local environments or physiological conditions into their interfaces than permanent ones. As a result, the differences of molecular interaction types in transient protein–protein interfaces can be more apparent than those of molecular interaction types in the permament ones. In addition, although in some cases, homocomplexes are transient, in many cases, transient complexes are restricted to heterocomplexes only.[40] For these reasons, we choose transient heterocomplexes as our dataset.

Transient heterocomplexes can be subdivided into subfunctional categories, but as usual, the distinction of such complexes is not clear-cut in biology. Furthermore, subcategories have limited size of data to be analyzed. So, in this study, we utilize a traditional but rough functional category such as protease-inhibitors, antibody–antigens, and signaling complexes.

To represent the physicochemical properties of protein–protein interfaces, we design 18 plausible interaction types using canonical and noncanonical interactions. Then, we construct input vector, so called feature vector, using the frequency of each interaction type. We apply these input vectors to the analysis of the 131 transient protein–protein interfaces in PDB: 33 protease-inhibitor, 52 antibody–antigen, 46 signal transduction including 4 cyclin dependent kinase, and 26 G-protein. By using k-nearest neighbor (kNN) classifier and feature selection method, we show that there are function specific interaction types in transient protein complexes, and we also examine some persistent interations. In addition, we discuss the biological role and evolutionary perspectives of specific binding interactions.

MATERIALS AND METHODS

Generation of the Data Set

The initial data set of three functional classes of transient protein–protein heterocomplexes is obtained from the Protein Data Bank[10] query system. To search for protease-inhibitor complexes, keywords such as "protease inhibitor" and "protease AND inhibitor" are used. For antibody–antigen complexes, the keywords "antibody antigen" and "antibody AND antigen" are used. For signalling complexes, keywords such as "cyclin dependent kinase", "G-protein", and "signal transduction" are used. During query process, we use 50% sequence identity as a cutoff value, and experiment techniques are confined only to X-ray diffraction. To this initial data set, we add transient heterocomplexes data used in the previous studies[11,12] according to their functional category. Of these collected data, we retain the data that have better resolution than 3.5 Å. When more than one complex is present in the asymmetric unit, only one copy is retained. Protein heterodimers are selected if the interfaces have more than 25 interacting residue pairs. The proteins are considered as nonhomologous on the basis that they have a sequence identity of $<30\%$ and SSAP[41] score $\leq 80\%$. The protein heterodimers are selected such that one or both components of each complex are nonhomologous to the components in the other complexes, and so within the data set, the interfaces in each functional class are nonhomologous. We also manually confirm this using SCOP database,[42] and present as a table, with the SCOP fold classification of each complexes, in the supplementary materials (Table S1–Table S6). The sequence identity and SSAP score can be obtained using CATH[43] query system. In the case of antibody–antigen complexes, although homologous pairs are included (e.g., antibody-lysozyme complexes), the sites of recognition on the lysozyme are different. Theoretical positions of hydrogen atoms are then added with program REDUCE.[44] Finally, we obtain 131 nonredundant protein–protein

TABLE I. PDB Codes and Chain ID of Three Functional Classes of Transient Protein-Protein Complexes

Protease-inhibitors			Antibody–antigens				Signaling proteins		
1acb:EI	1hia:YJ	4cpa:–I	1ao7:AC	1jhl:LA	1qle:BH	3hfl:LY	1a02:NF	1g3n:AC	1n4m:AC
1avg:HI	1jmo:HA		1cz8:VY	1jhl:HA	1r3j:AC	3hfl:HY	1a0o:AB	1g6g:AE	1ol5:AB
1avw:AB	1mct:AI		1dee:DG	1jps:LT	1r3j:BC	3hfm:LY	1a2k:AD	1gg2:BG	1omw:AB
1bai:AC	1mkw:HK		1dvf:BD	1jps:HT	1sbb:AB	3hfm:HY	1agr:AE	1got:AB	1rrp:AB
1cbw:GI	1oyv:AI		1fdl:LY	1lk3:AL	1sy6:LA		1am4:AD	1got:BG	1tx4:AB
1cbw:HI	1pxv:BD		1fdl:HY	1lk3:AH	1sy6:HA		1azs:AC	1gzs:AB	1ukv:GY
1cho:EI	1r0r:EI		1fe8:AH	1mq8:AB	1tqb:AB		1bkd:RS	1h4l:AD	1wa5:AB
1cse:EI	1slu:AB		1fe8:AL	1mvf:AD	1tqb:AC		1blx:AB	1hel:AC	1wa5:AC
1df9:AC	1stf:EI		1fjl:AF	1nca:NL	1txv:AL		1cly:AB	1i2m:CD	1wmh:AB
1dpj:AB	1tmq:AB		1fjl:BF	1nca:NH	1txv:AL		1cxz:AB	1ibr:AB	1wq1:RG
1dtd:AB	1toc:BR		1fsk:AB	1nsn:LS	1v7m:LV		1doa:AB	1ikn:AC	1ycs:AB
1eai:AC	1tx6:BI		1fsk:AC	1nsn:HS	1v7m:HV		1e96:AB	1ikn:AD	1zbd:AB
1f34:AB	1v5i:AB		1gh6:AB	1obl:AC	1w72:DI		1efn:AB	1kjy:AB	2trc:BP
1fle:EI	2sic:EI		1hez:AE	1obl:BC	1w72:DM		1efu:AB	1kps:CD	3fap:AB
1gl1:AI	3sgb:EI		1i9r:AH	1ots:BE	1wej:HF		1f5q:AB	1kz7:AB	
1hia:XJ	3tpi:ZI		1i9r:AL	1pqz:AB	2jel:LP		1foe:AB	1m2o:AB	

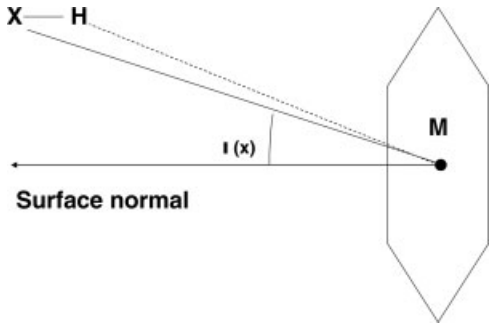

Fig. 1. Geometric parameters in X—H···π hydrogen bonds. *M* is the ring center, ω(x) is the angle between the X···M line and the surface normal.

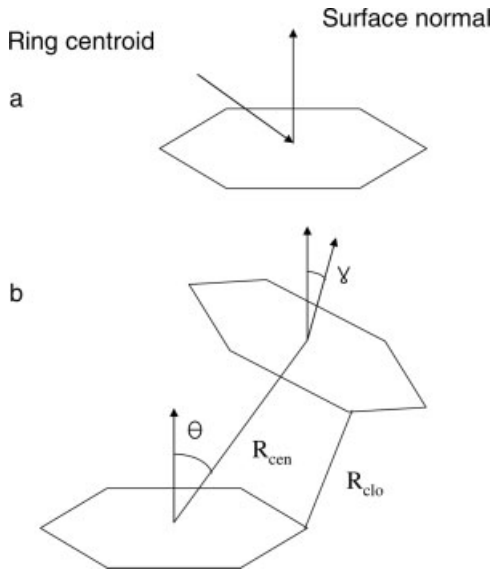

Fig. 2. Geometric parameters in π···π system. (a) Structural parameters to represent axially symmetric system. (b) Spherical polar coordinates for the relative pair orientation of π–π system.

interfaces shown in Table I, which consist of 33 protease-inhibitor, 52 antibody–antigen, and 46 signal transduction.

Definition of Interaction Type

We define 18 molecular interaction types composed of canonical hydrogen bondings, noncanonical hydrogen bondings, ion–ion interactions, and π-ring system related interactions. The selection criteria for the 18 interaction types are referred to the previous work.[29–31,33,34] Canonical hydrogen bonds involve X—H···O=C(X=N, O, CONH—) and X—H···Y (X or Y=N, O, S, CONH—) hydrogen bonds. In X—H···O=C, the selection criteria are H···O distance (≤4.5 Å) and X—H···O angle (>90°). In the case of X—H···Y, H···Y distance (≤4.5 Å) and X—H···Y angle (>120°), are adopted. They are known as an energetically stable criteria in canonical hydrogen bonding. H—N···Cation (cation = LYS, ARG), and N—H···Anion (anion = ASP, GLU) are also involved in the canonical hydrogen bonds. The selection criteria are same as those of X—H···Y type hydrogen bond. The effect of distance and angle criteria for the conventional hydrogen bonds is shown in the supplementary materials (Figure S2 and Table S7).

Noncanonical hydrogen bonds considered in this work are C^{α}—H···O=C, C^{α}—H···π, and X—H···π(X=N, O, CONH—) hydrogen bonds. For C^{α}—H···O=C hydrogen bond, we choose the distance cutoff as ≤4.5 Å and the angular cutoff as >90°, which have been proved to be an energetically stable criteria in many experiments.[33,34,45] For X—H···π(X=C^{α}, N, O, —CO—NH), the distance and angular cutoff criteria are shown in Figure 1. As a distance cutoff, the limit X—M<4.3 Å was selected and a cutoff angle ω(X) < 25° was chosen based on the previous work.[29,37,38]

We define canonical ion–ion interactions as the interactions between cationic (LYS, ARG) and anionic resi-

TABLE II. 18 Plausible Interaction Types

		Main chain					Side chain			
Interaction type		N—H	O=C	C^{α}—H	π	—OH	$-CONH_2$	—SH	⊕–charge	⊖–charge
Main chain	N—H	x[a]	x	–	x	x	x	x	x	–
	O=C		–	x	–	x	x	–	–	–
	C^{α}—H			–	x	–	–	–	–	–
	π				x	x	x	–	x	–
	—OH					–	x	–	–	–
Side chain	$-CONH_2$						–	–	–	–
	—SH							–	–	–
	⊕–charge								–	x
	⊖–charge									–

[a] "x" means interaction type involved in the analysis.

dues (ASP, GLU) with a distance less than 6.0 Å.[31] When calculating the distance, we consider the $C^{\varepsilon}/N^{\zeta}$ of LYS, $C^{\delta}/N^{\varepsilon}/C^{\zeta}/N^{\gamma}$ of ARG, C^{γ}/O^{δ} of ASP, and $C^{\delta}/O^{\varepsilon}$ of GLU. π-ring system related interactions include π···π and π···Cation interaction. A large number of experimental and theoretical studies have demonstrated that π···π interactions play an important role in molecular recognition.[30,46,47] There was a study about π···π stacking interaction for the protein core,[30] but no study on π···π stacking interaction in protein–protein interfaces has been reported. π···ring system can be represented in terms of ring centroid, surface normal vector. The relative structural orientation between π-ring system can be described by the centroid–centroid separation, R_{cen}, closest contact distances, R_{clo}, a center to normal angle, θ, and a normal to normal angle, γ. This is depicted in Figure 2. In this study, we use $R_{cen} < 7.5$ Å or $R_{clo} < 4.5$ Å, which is described in the previous work[30] and confirmed by our analysis with 3799 nonredundant interface data set from Nussinov's group.[26]

In π···Cation, we calculate the distance between the center of the π-ring and $C^{\varepsilon}/N^{\zeta}$ of LYS or $C^{\delta}/N^{\varepsilon}/C^{\zeta}/N^{\gamma}$ of ARG. The cutoff range we used is <6.[31] We summarized these interaction types in Table II, and each interaction type is represented with index in Table III. Hereafter we use the index to represent each interaction type.

TABLE III. Index of each Interaction Type

Index	Interaction type
0	π···π
1	π···cation
2	π···amide
3	π···hydroxyl
4	π···amine
5	π···H—C^{α}
6	amine···amine
7	amine···carbonyl
8	amine···hydroxyl
9	amine···cation
10	amine···anion
11	amide···carbonyl
12	amide···amine
13	amide···hydroxyl
14	carbonyl···hydroxyl
15	C^{α}—H···O=C
16	cation···anion
17	amine···thiol

Classification

Once we represent a protein–protein interaction interface as a vector representation using 18 plausible interaction types, we can take advantage of many useful methods, which is prevalent in the field of multivariate analysis and pattern recognition. k-Nearest neighbor[48] is a nonparametric classification technique that has been shown to be effective in statistical pattern recognition applications. It can achieve a high classification accuracy in problems that have unknown and non-normal distributions. The principle is very simple. Given a data set ($\mathbf{x}_i$, $\mathbf{y}_i$), it estimates values of $\mathbf{y}$ for $\mathbf{x}$ other than those in the sample using the distance metric and majority vote. Usually, distance metric is Euclidean (Eq. 1), though any other L_p-norm, such as Mahalanobis distance, can be used.

$$\|\mathbf{x} - \mathbf{x}_i\| = \sqrt{\sum_j^n (x_j - x_{ij})^2} \quad (1)$$

We construct input vector, so called, feature vector, using the frequency of each interaction type in the protein–protein interfaces. To select the appropriate distance metric, and to view the distribution of data prior to classification, principal component analysis (PCA) is carried out using singular value decomposition (SVD) technique. The results indicates that if we use 18 plausible interaction type, we can discriminate among three functional categories and it is possible to reduce dimension of input vector sharply. We apply kNN classification with leave-one-out cross-validation.

Feature Subset Selection

To extract function-specific interaction types, we use feature subset selection techniques, which are roughly

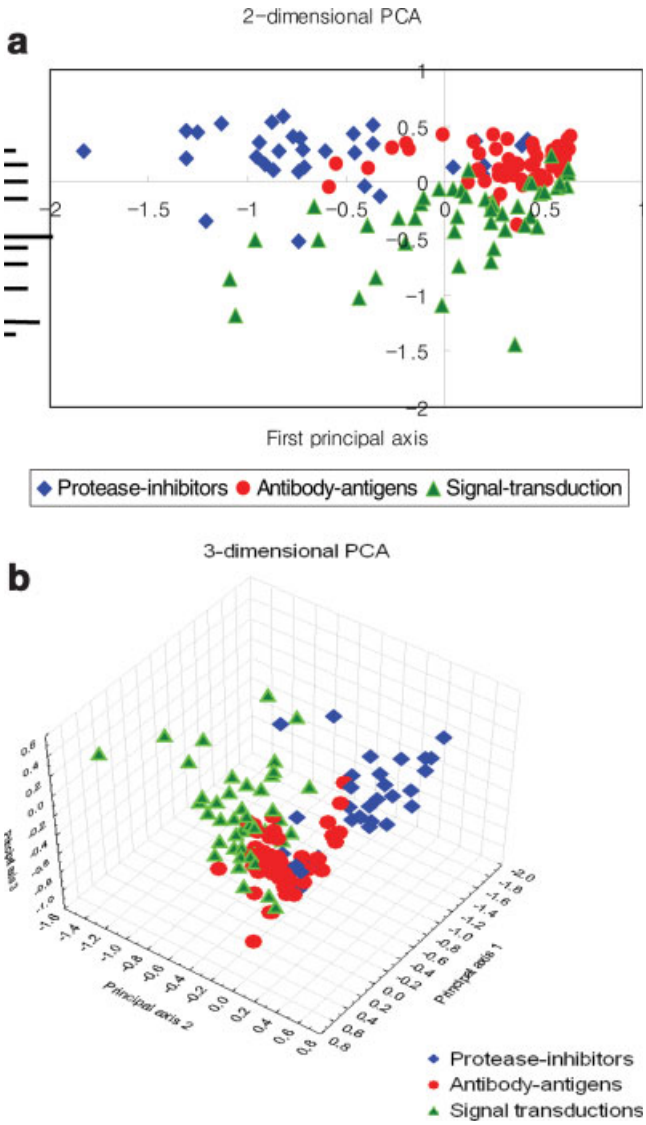

Fig. 3. 2-dimensional and 3-dimensional principal component analysis (PCA). Protease-inhibitor complexes are represented by blue labels, antibody–antigen complexes by red labels, and signaling complexes by green labels. In 2-dimensional PCA, antibody–antigens seem to be mixed up with signaling compelxes, but in 3-dimensional PCA, there is a clear separation among three functional classes of transient protein–protein interfaces. The shape of data distribution shows that it is possible to distinguish three functionally different classes of transient protein–protein interfaces using 18 plausible interaction types.

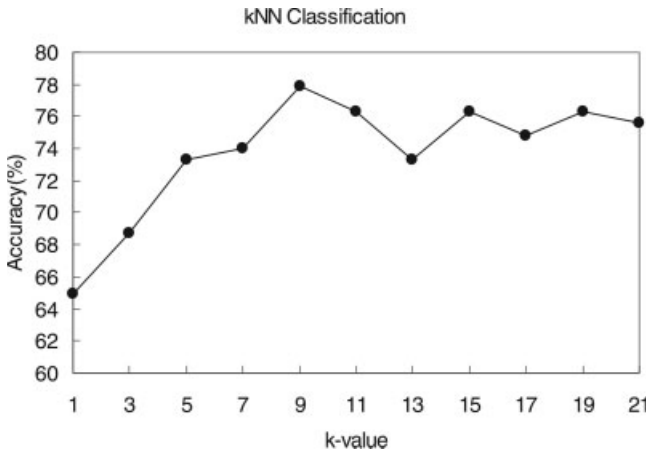

Fig. 4. Discriminating power of 18 plausible interaction types. The accuracy distribution according to *k*-value shows the usefulness of 18 plausible interaction types as a discriminator for three functionally different classes of transient protein–protein interfaces. kNN classification is performed with leave-one-out cross-validation, and 78% of accuracy is obtained at $k = 9$.

categorized into three approaches: wrapper, filter, and embedded methods.[49,50] Wrapper methods utilize learning machine as a black box to select subsets of variables based on their predictive power. Filter methods are independent of the classifier and select features based on properties that good feature sets are presumed to have, such as class separability or high correlation with the target. Embedded methods perform variable selection in the process of training and are usually specific to given learning machines. In our study, we adopt sequential forward selection, which is classic greedy wrapper, with kNN classifier as a predictor. In the first iteration, sequential forward selection tests all feature subsets with only one feature. The feature subset with the highest accuracy is chosen as the basis for the next iteration. In each iteration, the algorithm tentatively adds each feature, which is not previously selected, to the basis and retains the feature subset that results in the highest prediction accuracy. We carry out this experiment with the help of free software package LNKnet from MIT Lincoln Laboratory, which contains many useful classification and feature se- lection algorithm. With the help of feature subset selection technique, we can identify distinct interaction types specific to their functional categories.

RESULTS

Specific Binding Interaction Types

Before classification, we perform principal component analysis (PCA) using singular value decomposition (SVD) technique to test the usefulness of 18 plausible interaction types and to check the shape of data distribution. As we can see from Figure 3, PCA clearly shows that 18 plausible interaction types can have a good discriminating power for three different types of protein–protein interfaces. We perform kNN classification with leave-one-out cross-validation, and we can get 78% of accuracy at $k = 9$. In Figure 4, we report the classification accuracy according to the change of *k*-value. This clearly proves the usefulness of a set of interaction types as a discriminator for transient protein–protein interfaces.

Though PCA and kNN classification are quite effective to show the usefulness of 18 interaction types to discriminate three functional classes of transient protein–protein interfaces, they are not effective to investigate whether there are distinct interaction types specific to their functional category in original input space. This is because newly generated variables in PCA are the results of linear combination of original variables. In addition, if the num-

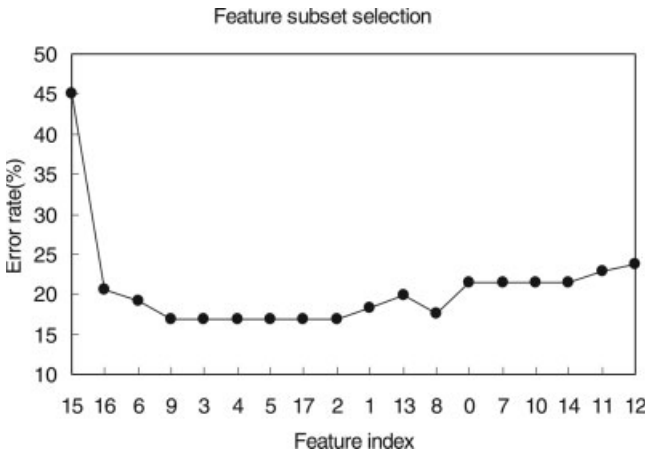

Fig. 5. Feature subset selection. The cross validation error rate achieved as each feature is added. Each feature is tested to find the one which is most effective in classifying the data by itself. The remaining features are then paired with the first and the best is selected as the second feature. Features are added in this way until there are no more left. The feature sets are tested using kNN classifier using leave-one-out cross validation. The most favorable feature set's indexes are 15, 16, 6, and 9, corresponding to C^{α}—H···O=C, cation–anion, NH···NH, and Amine···cation interaction.

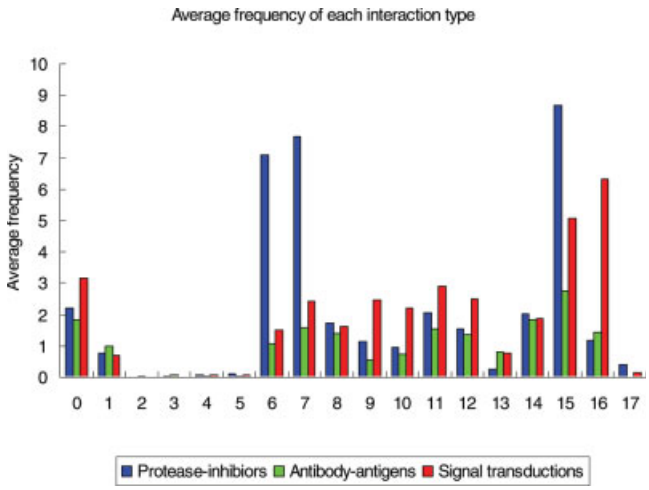

Fig. 6. Average frequency of each interaction type. The C^{α}—H···O=C (index = 15) interaction shows the highest frequency in protease-inhibitor complexes, and the differences of average frequency between the three functional classes are significant. The NH···NH (index = 6) and NH···O=C (index = 7) interaction show similar pattern, but the difference of average frequency between antibody–antigen complexes and signaling complexes is relatively small. The cation–anion interactions (index = 16) show the highest frequency in signaling complexes and significant difference over other functional classes. Other interactions show similar pattern of frequency regardless of the functional classes.

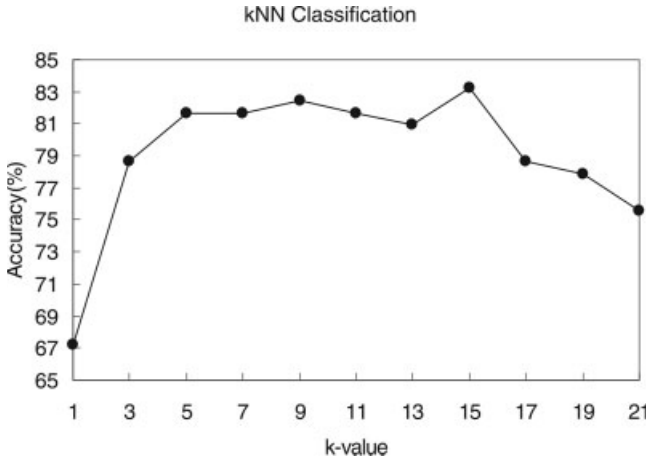

Fig. 7. Accuracy of kNN classification using only four specific binding interactions. We perform kNN classification with leave-one-out cross validation, and achieved 83.2% classification accuracy at $k = 15$.

ber of samples are relatively small with respect to the number of features, overfitting problem may happen. To identify relatively distinct interaction types from 18 interaction types and to reduce the dimensionality of the input data, we adopt feature subset selection technique, which is widely used in the field of pattern recognition. There are a lot of method to use, but in this study, we use kNN classifier as learning machine, and we use forward direction as a search direction with leave-one-out cross-validation. The cross-validation error rate is calculated as each feature is added. Each feature is tested to find the one which is most effective in classifying the data by itself. The remaining features are then paired with the first and the best is selected as the second feature. Features are added in this way until there are no more left. The result is shown in Figure 5. When we use the first four interactions(index = 15, 16, 6, and 9), the experiment shows the lowest error rate.

It means that, of 18 plausible interaction types, we achieve the best performance when using four in- teraction types, which involve C^{α}—H···O=C interaction (index = 15), ion–ion interaction (index = 16), NH···NH (index = 6), and amine···cation (index = 9) interaction. By doing so, we can reduce input dimension from eighteen to only four, and show that these four interaction types are the major factors to discriminate three different functional classes of protein–protein interfaces. The ion–ion interactions are specific to signal transduction complexes, and C^{α}—H···O=C interactions are relatively more specific to protease-inhibitor complexes. The NH··· NH and amine···cation interaction give a minor contribution to the classification accuracy. When combined with these two interactions, they increase the accuracy by 3.8%. In the case of amine···cation interaction, the difference of the frequency is rather small. However, when it is combined with the other interactions, it shows discriminating power, especially between signaling complexes and antibody–antigens. It shows that a variable that is largely useless by itself can provide a significant performance improvement when taken with others.[50] These results are consistent with the average frequency of each interaction type in Figure 6. Using only these specific binding interaction types, we perform kNN classification with leave-one-out cross-validation, and we obtain 83.2% classification accuracy at $k = 15$. In Figure 7, we report the classification accuracy according to the

TABLE IV. The Error Rate of each Functional Class at $K = 15$

Class	Patterns	No. errors	% Errors	Std Dev	RMS error	Label
Protease-inhibitors	33	8	24.24	7.5	0.372	0
Antibody–antigens	52	5	9.62	4.1	0.289	1
Signal transduction	46	9	19.57	5.8	0.344	2
Total	131	22	16.79	3.3	0.331	

TABLE V. The Classification Confusion Matrix of Prediction Results at $K = 15$

	Computed class			
Desired class	0[a]	1[b]	2[c]	*Total*
0[a]	25	7	1	33
1[b]	4	47	1	52
2[c]	2	7	37	46
Total	31	61	39	131

[a]Protease-inhibitors.
[b]Antibody–antigens.
[c]Signaling proteins. In leave-one-out cross-validation, 7 protease-inhibitors are misclassified as antibody-antigens, and 1 protease-inhibitor as signaling protein. As the same way, 4 antibody-antigens are misclassified as protease-inhibitors, 1 antibody-antigens as signaling proteins, 2 signaling complexes as protease-inhibitors, and 7 signaling proteins as antibody-antigens.

change of k-value. This classification accuracy shows a clear improvement over an accuracy achieved by considering amino acid pair frequency. When the frequency of 210 interacting residue pairs are used as a feature vector, the classification accuracy is 73.2% at $k = 1$, and as the k-value increases, the classification accuracy goes down. The results are shown in the supplementary materials (Figure S1).

At $k = 15$, we examine the error rate according to each functional class. The error rate of protease-inhibitors is 24.24%, and that of antibody–antigens is 9.62%. The signaling proteins have 19.57% error rate. The average error rate is 16.79%. The results are shown in Table IV. In leave-one-out cross-validation, we misclassify 7 protease-inhibitors as antibody–antigens and 1 protease-inhibitor as signaling protein. As the same way, we misclassify 4 antibody–antigens as protease-inhibitors, 1 antibody–antigens as signaling proteins, red signaling complexes as protease-inhibitors, and 7 signaling proteins as antibody–antigens. The classification confusion matrix of prediction results is reported in Table V.

The NH···NH interaction and NH···O=C interaction are highly correlated with the C^{α}—H···O=C interaction, so it does not contribute largely to the performance of classification. But they are also specific to protease-inhibitor complexes.

Persistent Interaction Types

There are some interaction types that do not show discriminating power, but show persistent occurrences in all three types of transient protein–protein interfaces. Here, "Persistent interaction" is defined as the interaction satisfying the following two equations, Eq. 2 and Eq. 3.

$$A = \frac{\sum_{i=1}^{F_c} n(f_{ci}, I)}{F_c} \geq E_a, \quad \forall c = 0, 1, 2 \tag{2}$$

$$R = \frac{\sum_{c=0}^{2} \sum_{i=1}^{F_c} m(f_{ci}, I)}{\sum_{c=0}^{2} F_c} \times 100 \geq E_r, \text{with } m(f_{ci}, I) = \begin{cases} 1 & \text{if } n(f_{ci}, I) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

In Eq. 2, "A" is defined as the average number of frequencies of a specific interaction type, I, over a specific class, c. The f_{ci} is an ith interface in class c, and I is one of the interactions involved in 18 interaction types. The F_c is the total number of interfaces in class c. The $n(f_{ci}, I)$ represents the number of interaction, I, on the interface, i, in class c. If there are 73 π···π interactions in 33 protease-inhibitors, then "A" value is 2.2, calculated by 73/33. "E_a" value is the expected occurrences of a specific interaction type, I, on a specific interface. The data set contain 131 interfaces and 18 features, and the observed total number of interactions are 3743 over the 131 interfaces. So, E_a value is 1.6, calculated by 3743/(131*18). In Eq. 3, the "R" is defined by the average ratio of the interfaces having nonzero occurrences of a specific interaction type, I, over the 131 interfaces. If there are 84 interfaces having nonzero π···π interactions in the 131 interfaces, the the "R" value is 64%, calculated by (84/131)*100. "E_r" value is defined as the expected average ratio of the interfaces having nonzero occurrences of a specific interaction type, I, over the 131 interfaces. The observed total number of interfaces having nonzero occurrences are 1202 over the 18 features, so the E_r value is 50, calculated by {1202/(131*18)}*100. The interactions satisfying the above two criteria ($A \geq 1.6 \cap R \geq 50$) are π···π, amide-carbonyl, and hydroxyl-carbonyl interaction. π···π interaction has 2.4 average number of frequencies. Even though there is a large variation depending on each interface, over 64% of the interfaces have at least more than one π···π interaction. Furthermore, in signal transduction complexes, over 80% of interfaces have nonzero π···π interactions. This result is shown in Table VI.

The relative orientation of one π-ring with respect to another is analyzed by a center-normal angle, θ, a normal-normal angle, γ. In this analysis, we only consider dimers, and do not consider trimers. The population dis-

TABLE VI. Overall Trend of Persistent Interactions

	π···π			Amide-carbonyl			Hydroxyl–carbonyl		
Category	avg[a]	n^b/t^c	ratio (%)[d]	avg[a]	n^b/t^c	ratio (%)[d]	avg[a]	n^b/t^c	ratio (%)[d]
Protease-inhibitor	2.2	19/33	57.6	2.1	24/33	72.7	2.0	31/33	93.9
Antibody–antigen	1.8	28/52	53.8	1.6	30/52	57.6	1.8	38/52	73.1
Signal transduction	3.2	37/46	80.4	2.9	35/46	76.1	1.9	38/46	82.6
Total	2.4	84/131	64.1[e]	2.2	89/131	68.8[e]	1.9	107/131	81.7[e]

[a]Average number of frequencies (A) calculated from Eq. 2.
[b]The number of interfaces having non zero occurrences.
[c]Total number of interfaces in each class.
[d]The percent ratio of n/t.
[e]R, calculated from Eq. 3.

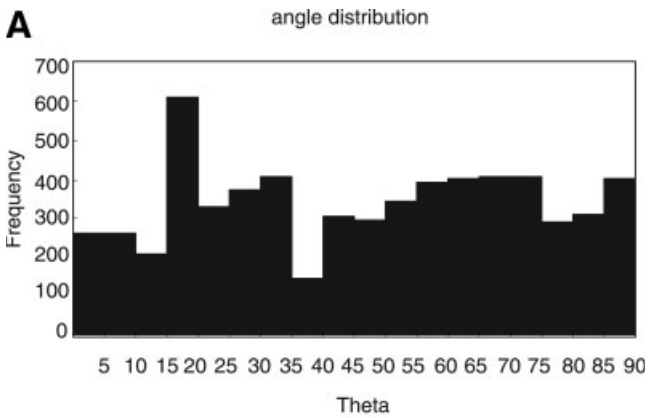

Fig. 8. θ and γ angle distribution of 314 π−π interactions in three functional classes of transient protein–protein interfaces. The θ distributions have maximum peak around 15–20° and the γ distributions have maximum peaks around 5–10°. These results show that in three functional classes of transient protein–protein interfaces, herringbone shape (Fig. 9) is a major configuration.

tributions of θ and γ, are analyzed and corrected for spherical polar and Euler angle probability bias.[30] In Figure 8, though there can be a variety of combinations of θ and γ, we can observe that θ distribution have peaks around 15–20° and the γ distributions have peaks around 5–10°. This shows that in transient protein–protein interfaces, herringbone shape is a major configuration. This result is different from the configuration in the protein core, where parallel-displaced configurations are major.

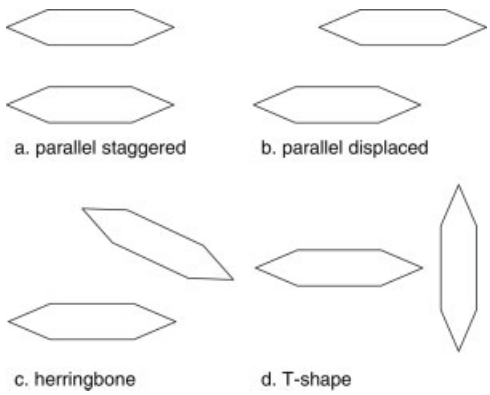

Fig. 9. Schematic diagram of the four different orientations of π···π stacking interactions.

The four different types of π···π stacking interactions are shown in Figure 9.[47] The result of amide–carbonyl and hydroxyl–carbonyl group interactions are also shown in Table VI. It is noticeable that nearly 82% of the interfaces has at least more than one hydroxyl–carbonyl interactions.

DISCUSSION

Protein–protein interfaces contain a variety of interactions, including hydrogen bonding, salt bridges, water bridging interactions and hydrophobic interactions.[51,52] It is well known that a delicate balance of various weak and strong noncovalent interactions contributes to the stability and selectivity of protein–protein complexes. protein–protein interactions are the results of dynamic processes between proteins or between proteins and their local environments. Local environments or physiological conditions of a protein as well as the properties of a protein itself, are important to the protein's specific function, and proteins in the same functional category perform their functions under the similar environments. Transient protein–protein interfaces associate or dissociate at least once during their cellular processes, so they have more chance to reflect their local environments or physiological conditions into their interfaces than permanent ones. As a result, the differences of molecular

TABLE VII. Amino Acid Residue Pairs Showing Conservation of C^{α}—H···O=C Interactions in Protease-Inhibitor Interfaces

1acb:EI	PHE(41)[a]	MET(192)	GLY(193)	SER(195)	SER(214)	TRP(215)		SER(217)	SER(218)
	ASP(46)	LEU(45)	LEU(45)	LEU(45)	THR(44)	VAL(43)		SER(41)	SER(41)
	LEU(47)	ASP(46)	ASP(46)			THR(44)			
1avw:AB	PHE(41)	GLN(192)	GLY(193)	SER(195)		TRP(215)	GLY(216)	SER(217)	SER(218)
	ILE(564)	ARG(563)	ARG(563)	ARG(563)		PRO(561)	PRO(561)		
		ILE(564)	ARG(563)						
		ARG(565)							
1cbw:HI		MET(192)	GLY(193)	SER(195)	SER(214)	TRP(215)	GLY(216)	SER(217)	SER(218)
		LYS(15)	LYS(15)	LYS(15)	CYC(14)	PRO(13)	PRO(13)	PRO(13)	GLY(12)
		ALA(16)	ALA(16)						PRO(13)
1cho:EI	PHE(41)	GLN(192)	GLY(193)	SER(195)		TRP(215)		SER(217)	
	GLU(19)	GLU(19)	LEU(18)	LEU(18)		CYS(16)		PRO(14)	
	TYR(14)		GLU(19)					ALA(15)	
1eai:AC	THR(41)	GLN(192)	GLY(193)	SER(195)	SER(214)	PHE(215)	VAL(216)	SER(217)	LEU(218)
	MET(32)	LEU(31)	LEU(31)	LEU(31)	PRO(30)	CYS(29)	PRO(28)	THR(27)	GLY(43)
1fle:EI	THR(41)		GLY(193)	SER(195)	SER(214)	PHE(215)	VAL(216)	SER(217)	
	MET(25)		ALA(24)	ALA(24)	CYS(23)	ARG(22)	ILE(21)	LEU(20)	
						CYS(23)			
1gl1:AI	PHE(41)	MET(192)	GLY(193)	SER(195)	SER(214)	TRP(215)	GLY(216)	SER(217)	SER(218)
	LYS(31)	LEU(30)	LEU(30)	LEU(30)	THR(29)	CYS(28)	ALA(27)	ALA(26)	ALA(26)
	ALA(32)	LYS(32)	LYS(31)						
1jmo:HA	LEU(41)		GLY(193)	ALA(195)	SER(214)	TRP(215)	GLY(216)	GLU(217)	
	SER(445)		LEU(444)	LEU(444)	PRO(443)	MET(442)	PHE(441)	GLY(440)	
			SER(445)						
1mct:AI	PHE(41)	GLN(192)	GLY(193)	SER(195)	SER(214)	TRP(215)	GLY(216)	TYR(217)	
	ILE(6)	PRO(4)	ARG(5)	ARG(5)	PRO(4)	CYS(3)	ILE(2)	ARG(1)	
	TRP(7)	ARG(5)							
		ILE(6)							
1slu:BA	PHE(41)	GLN(192)	GLY(193)	SER(195)		PHE(215)	GLY(216)	TYR(217)	
	MET(85)	MET(84)	MET(84)	MET(84)		SER(82)	VAL(81)	PRO(80)	
	HIS(86)	MET(85)	MET(85)						
1tx6:BI	PHE(41)	GLN(192)	GLY(193)	SER(195)	SER(214)	TRP(215)	GLY(216)	TYR(217)	
	SER(77)	THR(75)	ARG(76)	ARG(76)	THR(75)	CYS(74)	ILE(73)	ALA(72)	
	ASN(78)	SER(77)	SER(77)			THR(75)			
3sgb:EI	ARG(41)		GLY(193)	SER(195)	SER(214)	GLY(215)	GLY(216)	SER(217)	
	GLU(19)		LEU(18)	LEU(18)	THR(17)	CYS(16)	ALA(15)	PRO(14)	
						THR(17)			
3tpi:ZI	PHE(41)	GLN(192)	GLY(193)	SER(195)	SER(214)	TRP(215)	GLY(216)		
	ALA(16)	CYS(14)	LYS(15)	LYS(15)	CYS(14)	PRO(13)	PRO(13)		
	ARG(17)	LYS(15)	ALA(16)						
		ALA(16)							

[a]The number in the parenthesis represents the sequence number. The table summarizes the interacting residue pairs maintaining the conservation of C^{α}—H···O=C interactions in the interfaces composed of chymotrypsin family and its interacting inhibitor. For example, PHE(41) of α−chymotrypsin interacts with ASP(46) and LEU(47) of eglin C in 1acb, and GLU(19) and TYR(20) of PMP-C in 1gl1. There are changes in interacting residue pairs, but C^{α}—H···O=C interactions remain unchanged.

interaction types in transient protein–protein interfaces can be more apparent than those of molecular interaction types in the permanent ones. Therefore, it is reasonable that the transient proteins in the same functional category recognize their interacting partners by certain types of molecular interactions that are specific to their own protein family members and their local environments. As a result, proteins can show specific binding interactions according to their functional classes of protein–protein interfaces. If the binding region of one component of transient protein–protein interfaces remains unchanged irrespective of their partners, not only specific binding interaction but also interaction topology itself can be conserved through common binding mechanism, even though there is a large variation of interaction pairs at the residue level. In this study, we do not deal with all the possible interactions, but we just use 18 plausible interaction types, containing canonical hydrogen bondings, noncanonical hydrogen bondings, ion–ion interactions and π-ring system related interactions. Nevertheless, they show good discriminating power for the three functional classes of transient protein–protein interfaces. Furthermore, only four distinct interaction types are enough to distinguish the three functional classes of transient protein–protein interfaces.

In protease-inhibitor interfaces, backbone–backbone interactions (e.g. NH···NH, NH···O=C, and C^{α}– H···O=C interaction) are predominant, and the binding specificity is controlled by C^{α}—H···O=C interactions. The importance of C^{α}—H···O=C interactions for the stability and specificity of protein–protein interactions has been already shown in several studies.[32,33] When we examine protease-inhibitor interfaces in more detail, we find that the C^{α}—H···O=C interaction topology is conserved, while there is a large variation of interacting pair at the residue level. For example, 1acb and 1gl1 in PDB are the 3-dimensional structures of serine protease and its inhibitor. Protease components are α-chymotrypsin and inhibitor components are eglin C in 1acb and PMP-C in 1gl1, which are nonhomologous. At the residue level, the inhibitor's residues interacting with α-chymotrypsin in the interfaces, are quite different from each other. For instance, PHE^{41} of α-chymotrypsin interacts with ASP^{46} and LEU^{48} of eglin C in 1acb, and GLU^{19} and TYR^{20} of PMP-C in 1gl1. However, C^{α}— H···O=C interactions remain unchanged. In Table VII, we summarize interacting residue pairs keeping up C^{α}—H···O=C interactions in each interface. We also depict the C^{α}—H···O=C interaction topology in Figure 10. The conservation of C^{α}—H···O=C interaction topology is also observed in other protease-inhibitor interfaces.

In signalling proteins, which are mainly composed of G-proteins and cyclin dependent kinases, the binding specificity is controlled by cation–anion interactions. In Table VIII, we list interacting residue pairs keeping up cation–anion interactions in the interfaces composed of small GTPase RAN and its partner. To sustain cation–anion interactions, cation changes are limited to the LYS and ARG, and anion changes are confined to the GLU and ASP. Therefore, it is difficult to find examples to describe the conservation of cation–anion interaction topology. However, like protease-inhibitors, it also shows the conservation of cation–anion in- teraction topology. For example, 1ibr and 1i2m involve small GTPase protein RAN, and RAN's interaction partner is importin β-subunit, nuclear transport receptor, in 1ibr and regulator of chromosome condensation(RCC1) in 1i2m. The global structures of the interfaces are quite different from each other, and binding region is partly overlapped with each other. Nevertheless, cation–anion interaction topology is conserved on the overlapping region, even though there are some changes in the residue pairs. ASP(77) interacts with LYS(62) in 1ibr, but ARG(320) in 1i2m. ASP(107) interacts with two LYS(62), LYS(68) in 1ibr, but two ARG(320), ARG(325) in 1i2m. Also, they maintain ARG(106)-ASP(160), ARG(140)-ASP(288) in 1ibr, and ARG(106)-ASP(384), ARG(140)-ASP[44] It shows that even though there are changes at the residue level, proteins maintain their specific cation–anion interactions. We depict cation–anion interaction in Figure 11. In the case of antibody–antigens, the sign is somewhat ambiguous. From the evolutionary perspective, while protease-inhibitors and signaling proteins have optimized their interfaces to suit their biological function,

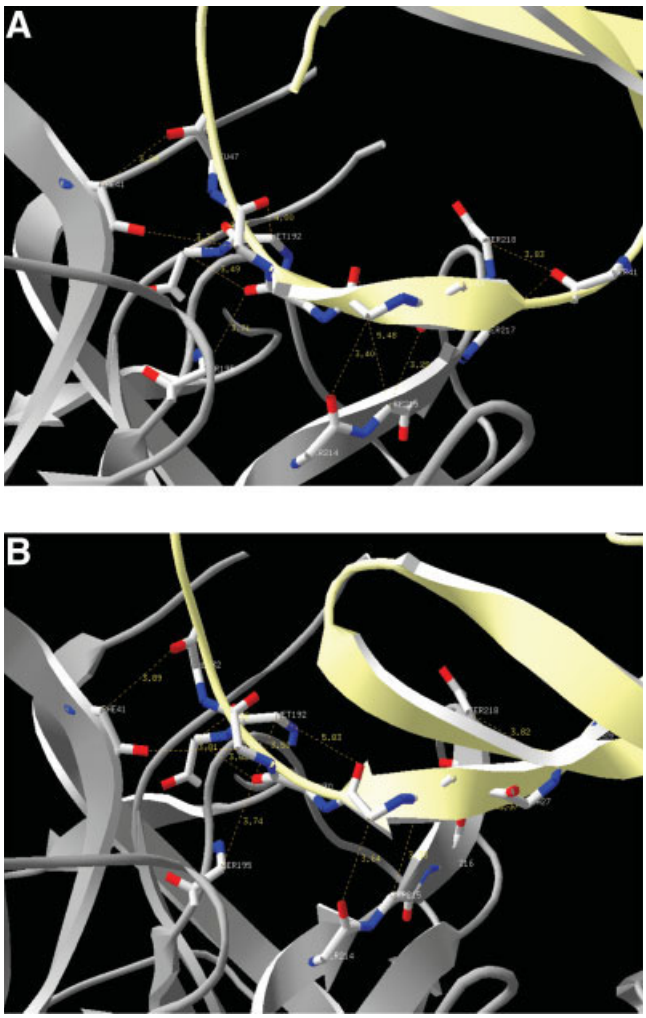

Fig. 10. Conservation of C^{α}—H···O=C interaction topology across the interface of the protease-inhibitor complexes. (a) Schematic diagram of C^{α}—H···O=C interactions of 1acb. (b) Schematic diagram of C^{α}—H···O=C interactions of 1gl1. The dashed lines represent C^{α}—H···O=C interactions. Protease components are α-chymotrypsin and inhibitor components are eglin C in 1acb and PMP-C in 1gl1, which are nonhomologous. At the residue level, the inhibitor's residues interacting with α-chymotrypsin are quite different from each other (Table VII), but C^{α}—H···O=C interactions remain unchanged. The images are created by program Swiss-PdbViewer.[53]

antibody–antigen interactions are the happenstance, implying antibody–antigen complexes do not show distintive interaction types. In limited scope, a previous work[54] has shown that the selectivity of the binding of the protein family is achieved by conserved hydrogen bonds.

It should be emphasized that our study is the first systematic approach to analyze protein–protein interfaces at the molecular interaction level in the context of protein functions. Our study clearly shows that there are specific interaction types based on their functional classes of protein–protein interfaces, and several specific interactions are conserved according to the functional classes of protein–protein interfaces through the com-

TABLE VIII. Amino Acid Residue Pairs Showing Conservation of Cation-Anion Interactions in RAN Family

1i2m:CD	ASP(77)[a]	ARG(95)	ARG(106)	ASP(107)	ARG(110)	LYS(134)	ARG(140)
	ARG(320)	ASP(95)	ASP(384)	ARG(320)	GLU(322)	ASP(95)	ASP(44)
		GLU(109)		ARG(325)			GLU(56)
1ibr:AB	ASP(77)		ARG(106)	ASP(107)	ARG(110)		ARG(140)
	LYS(62)		ASP(160)	LYS(62)	ASP(160)		GLU(281)
				LYS(68)			ASP(288)
1wa5:AB		ARG(95)				LYS(134)	
		GLU(506)				GLU(484)	
1wa5:AC	ASP(77)		ARG(106)	ASP(107)	ARG(110)	LYS(134)	
	LYS(62)		GLU(107)	LYS(62)	GLU(107)	GLU(370)	
	LYS(66)			LYS(66)			

[a]The number in the parenthesis represents the sequence number. This table lists the interacting residue pairs maintaining the conservation of cation-anion interactions in the interfaces composed of small GTPase RAN and its partner. 1ibr and 1i2m involve small GTPase protein RAN, and RAN's interacting partner is importin beta-subunit, nuclear transport receptor, in 1ibr and regulator of chromosome condensation(RCC1) in 1i2m. The global structures of the interfaces are quite different from each other, and binding region is partly overlapped with each other. However, cation-anion interaction is conserved on the overlapping region, even though there are some changes in the residue pairs. For example, ASP(77) interacts with LYS(62) in 1ibr, but ARG(320) in 1i2m. ASP(107) interacts with two LYS(62), LYS(68) in 1ibr, but two ARG(320), ARG(325) in 1i2m. Though there are changes interaction residue pairs, they try to sustain cation-anion interactions.

mon binding mechanism, rather than through the sequence or structure conservation.

CONCLUSIONS

Most of the studies about protein–protein interfaces have been carried out at the level of amino acid residues or more coarse-grained level of descriptions. In this study, we look at the interfaces at the molecular interaction level. Comprehensive studies of such type of approach have been carried out in a number of studies on protein stability, while a few studies on protein interfaces with limited scope have been reported. In addition, most of them were concentrated on the role of an individual interaction type. Furthermore, few attempts have been made to investigate the meaning of specific interaction types in the context of protein functions.

This is the first systematic approach to analyze protein–protein interfaces at the molecular interaction level in the context of protein functions. Our study clearly shows that proteins show specific binding interactions according to their functional classes of protein–protein interfaces, and specific interactions are conserved according to their functional category through the common binding mechanism, rather than through the sequence or structure conservation.

We take advantage of classification and feature subset selection technique, which are prevalent in pattern recognition and machine learning field. Three functional classes of transient protein–protein complexes can be distinguished by only four interaction types, which involve C^{α}—H···O=C interaction, ion–ion interaction, NH···NH, and amine···cation interaction. Of these four interaction types, C^{α}—H···O=C are predominant in protease-inhibitor interfaces, and cation–anion interactions appear more frequently in signaling complexes. When we examine the interfaces in more detail, these two types of interactions are conserved, while there is a large variation of interacting pair at the residue level. The NH···NH and amine···cation interaction give a minor contribution to the classification accuracy. When combined with the above two interactions, they show only 3.8% higher accuracy. In the case of amine···cation interaction, the difference of the frequency is rather small. However, when it is combined with the other interactions, it shows discriminating power, especially between signaling complexes and antibody–antigens.

We also examine persistent interaction types such as π···π interaction, amide–carbonyl and hydroxyl–carbonyl group interactions. It is noticeable that nearly 82% of the interfaces has at least more than one hydroxyl–carbonyl interactions. In π···π interaction, herringbone shape is a major configuration. This result is different from that of the protein core, where parallel-displaced configurations are major. There was a study about π···π stacking interaction for the protein core,[30] but no study on π···π stacking interaction in protein–protein interfaces has been reported.

An important implication of our work is that the analysis of protein–protein interfaces at the molecular interaction level in the context of protein functions can give us another point of view about protein–protein interactions. Proteins may selectively recognize their partners with specific binding interactions under the appropriate local environments, and sustain their stability with the help of persistent interactions. Furthermore, if the binding region of one component of transient protein–protein interfaces remains unchanged irrespective of their partners, not only interaction itself but also interaction topology can be conserved through common binding mechanism, even though there is a large variation of interacting pairs at the residue level. The findings of this study may help to design artificial drug candidates, which can block or activate biologically meaningful pathways. Moreover, our approach can be extended to investigate the effects of local environments to the protein–protein interactions. We anticipate that we can directly apply our approach to the fine-tuning of protein–protein docking problem, and to distinguish crystal contacts from biological contacts in near future.

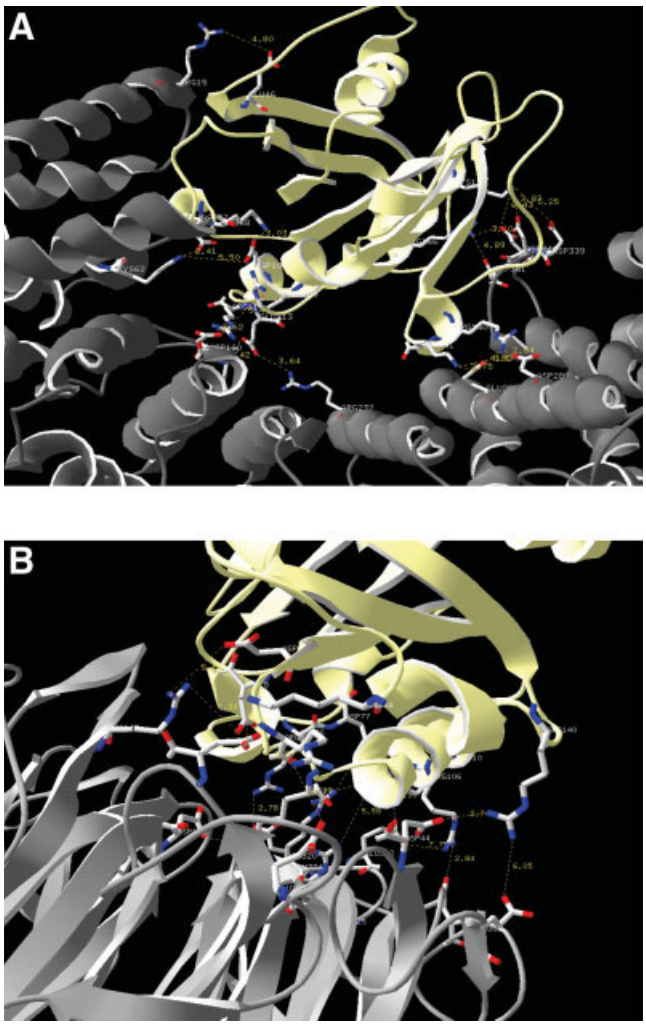

Fig. 11. Conservation of cation–anion interaction topology in signaling complexes. (a) Schematic diagram of cation–anion interaction of 1ibr. (b) Schematic diagram of cation–anion interaction of 1i2m. The dashed lines represent cation–anion interactions. 1ibr and 1i2m involve small GTPase protein RAN, and RAN's interaction partner is importin β-subunit, nuclear transport receptor, in 1ibr and regulator of chromosome condensation(RCC1) in 1i2m. The global structures of the interfaces are quite different from each other, but cation–anion interaction topology is conserved, even though there are some changes in the residue pairs (Table VIII). The images are created by program Swiss-PdbViewer[53].

ACKNOWLEDGMENTS

This work was supported by the Korean Systems Biology Research Grant (2005-00343) from the Ministry of Science and Technology, National Research Laboratory Grant (2005-01450) from the Ministry of Science and Technology. We would also like to thank CHUNG Moon Soul Center for BioInformation and BioElectronics and the IBM SUR program for providing research and computing facilities.

REFERENCES

1. Chothia C, Janin J. Principles of protein-protein recognition. Nature 1975;256:705–708.
2. Argos P. An investigation of protein subunit and domain interfaces. Protein Eng 1988;2:101–113.
3. Janin J, Chothia C. The structure of protein-protein recognition sites. J Biol Chem 1990;265:16027–16030.
4. Korn AP, Burnett RM. Distribution and complementarity of hydropathy in multisubunit proteins. Proteins 1991;9:37–55.
5. Lawrence MC, Colman PM. Shape complementarity at protein/protein interfaces. J Mol Biol 1993;234:946–950.
6. Jones S, Thornton JM. Protein-protein interactions: a review of protein dimer structures. Prog Biophys Mol Biol 1995;63:31–65.
7. Jones S, Thornton JM. Principles of protein-protein interactions. Proc Natl Acad Sci USA 1996;93:13–20.
8. Jones S, Thornton JM. Analysis of protein-protein interaction sites using surface patches. J Mol Biol 1997;272:121–132.
9. McCoy AJ, Epa VA, Colman PM. Electrostatic complementarity at protein/protein interfaces. J Mol Biol 1997;268:570–584.
10. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res 2000;28:235–242.
11. LoConte L, Chothia C, Janin J. The atomic structure of protein-protein recognition sites. J Mol Biol 1999;285:2177–2198.
12. IMA Nooren, Thornton JM. Structural characterization and functional significance of transient protein-protein interactions. J Mol Biol 2003;325:991–1018.
13. Keskin O, Bahar I, Badretdinov AY, Ptitsyn OB, Jernigan RL. Empirical solvent-mediated potentials hold for both intra-molecular and inter-molecular inter-residue interactions. Protein Sci 1998;7:2578–2586.
14. Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces: characterization and comparison with oligomeric protein interfaces. J Mol Biol 1998;280:1–9.
15. Glaser F, Steinberg DM, Vakser IA, Ben-Tal N. Residue frequencies and pairing preferences at protein- protein interfaces. Proteins 2001;43:89–102.
16. Well JA. Systematic mutational analyses of protein-protein interfaces. Meth Enzymol 1991;202:390–411.
17. Clackson T, Wells JA. A hot spot of binding energy in a hormone-receptor interface. Science 1995;267:383–386.
18. Clackson T, Ultsch MH, Wells JA, de Vos AM. Structural and functional analysis of the 1:1 growth hormone: receptor complex reveals the molecular basis for receptor affinity. J Mol Biol 1998; 277:1111–1128.
19. Massova I, Kollman PA. Computational alanine scanning to probe protein-protein interactions: a novel approach to evaluate binding free engergies. J Am Chem Soc 1999;121:8133–8143.
20. Kortemme T, Baker D. Protein-protein interfaces: analysis of amino acid conservation in homodimer. Proc Natl Acad Sci USA 2002;99:14116–14121.
21. Hu Z, Ma B, Wolfson H. Conservation of polar residues as hot spots at protein interfaces. Proteins 2000;39:331–342.
22. Ma B, Elkayam T, Wolfson H, Nussinov R. Protein-protein interactions: structually conserved residues distinguish between binding sites and exposed protein surfaces. Proc Natl Acad Sci USA 2003;100:5772–5777.
23. Halperin I, Wolfson H, Nussinov R. Protein-protein interactions: coupling of structurally conserved residues and of hot spots across interfaces. Implicaions for docking. Structure 2004;12: 1027–1038.
24. Nussinov R, Wolfson H. Efficient detection of three-dimensional structural motifs in biological macro- molecules by computer vision techniques. Proc Natl Acad Sci USA 1991;88:10495–10499.
25. Tsai CJ, Lin SL, Wolfson H, Nussinov R. A dataset of protein-protein interfaces generated with a sequence order-independent comparison technique. J Mol Biol 1996;260:604–620.
26. Keskin O, Tsai CJ, Wolfson H, Nussinov R. A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications. Protein Sci 2004;13:1043–1055.
27. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 1996;257:342–358.
28. Valdar William SJ, Thornton JM. Protein-protein interfaces: analysis of amino acid conservation in homodimer. Proteins 2001; 42:108–124.

29. Steiner T, Koellner G. Hydrogen bonds with π-acceptors in proteins: frequencies and role in stabilizing local 3D structures. J Mol Biol 2001;305:535–557.
30. McGaughey GB, Gagne M, Rappe AK. πstacking interactions: alive and well in proteins. J Biol Chem 1998;273:15458–15463.
31. Gallivan JP, Dougherty DA. Cation-π interactions in structural biology. Proc Natl Acad Sci USA 1999;96:9459–9464.
32. Senes A, Ubarretxena-Belandia I, Engelman DM. The C^{α}−H···O hydrogen bond: a determinant of stability and specificity in transmembrane helix interactions. Proc Natl Acad Sci USA 2001; 98:9056–9061.
33. Jiang L, Lai L. CH–O hydrogen bonds at protein-protein interfaces. J Biol Chem 2002;227:37732–37740.
34. Baker E, Hubbard R. Hydrogen bonding in globular proteins. Prog Biophys Mol Biol 1984;44:97–179.
35. Babu M M. NCI: a server to identify non-canonical interactions in protein structures. NAR 2003;31:3345–3348.
36. Crowley PB, Golovin A. Cation-π interactions in protein-protein interfaces. Proteins 2005;59:231–239.
37. Burley SK, Petsko GA. Amino-aromatic interactions in proteins. FEBS Lett 1986;203:139–143.
38. Flocco MM, Mowbray SL. Planar stacking interactions of arginine and aromatic side-chains in proteins. J Mol Biol 1994;235: 709–717.
39. Toth G, Watts CR, Murphy RF, Lovas S. Significance of aromatic-backbone amide interactions in protein structure. Proteins 2001;43:373–381.
40. Julian M, Zhiping W. Atomic contact vector in protein-protein recognition. Proteins 2003;53:629–639.
41. Talyor WR, Orengo CA. Protein structure alignment. J Mol Biol 1989;208:1–22.
42. Antonina A, Dave H, Steven EB, Hubbard TJP, Cyrus C, Alexey GM. SCOP database in 2004: refinements integrate structure and sequence family data. Nuclecic Acids Res 2004;32:D226–D229.
43. Pearl FM, Todd A, Sillitoe I, Dibley M, Redfern O, Lewis T, Bennett C, Marsden R, Grant A, Lee D, Akpor A, Maibaum M, Harrison A, Dallman T, Reeves G, Diboun I, Addous S, Lise S, Johnston C, Sillero A, Thornton J, Orengo C. The CATH domain structure database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. Nucleic Acids Res 2005;33:D247–D239.
44. Word JM, Lovell SC, Richardson JS, Richardson DC. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. J Mol Biol 1999;285:1733–1747.
45. Scheiner S, Kar T, Gu Y. Strength of the C^{α}H···O hydrogen bond of amino acid residues. J Biol Chem 2001;276:9832–9837.
46. Sinnokrot MO, Sherrill CD. Substituent effects in π−π interactions: sandwich and T-shaped configurations. J Am Chem Soc 2004;126:7690–7697.
47. Duda R O, Hart P E, David S. Pattern classification. New York: Wiley; 2001.174pp.
48. Kohavi R, John GH. Wrappers for feature subset selection. Artif Intell 1997;97:273–324.
49. Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003;3:1157–1182.
50. Sun S, Bernstein ER. Aromatic van der Waals clusters. J Phys Chem 1996;100:13348–13366.
51. Sheinerman FB, Norel R, Honig B. Electrostatic aspects of protein-protein interactions. Curr Opin Struct Biol 2000;10:153–159.
52. Xu D, Tsai CJ, Nussinov R. Hydrogen bonds and salt bridges across protein-protein interfaces. Protein Eng 1997;10:999–1012.
53. Guex N, Peitsch MC. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 1997;18:2714–2723.
54. Xu RM, Carmel G, Kuret J, Cheng X. Structural basis for selectivity of the isoquinoline sulfonamide family of protein kinase inhibitors. Proc Natl Acad Sci USA 1996;93:6308–6313.