Atomic Contact Vectors in Protein-Protein Recognition

Julian Mintseris[1] **and Zhiping Weng**[1,2*]
[1]Bioinformatics Program, Boston University, Boston, Massachusetts
[2]Department of Biomedical Engineering, Boston University, Boston, Massachusetts

ABSTRACT **The ability to analyze and compare protein-protein interactions on the structural level is critical to our understanding of various aspects of molecular recognition and the functional interplay of components of biochemical networks. In this study, we introduce atomic contact vectors (ACVs) as an intuitive way to represent the physicochemical characteristics of a protein-protein interface as well as a way to compare interfaces to each other. We test the utility of ACVs in classification by using them to distinguish between homodimers and crystal contacts. Our results compare favorably with those reported by other authors. We then apply ACVs to mine the PDB for all known protein-protein complexes and separate transient recognition complexes from permanent oligomeric ones. Getting at the basis of this difference is important for our understanding of recognition and we achieved a success rate of 91% for distinguishing these two classes of complexes. Although accessible surface area of the interface is a major discriminating feature, we also show that there are distinct differences in the contact preferences between the two kinds of complexes. Illustrating the superiority of ACVs as a basic comparison measure over a sequence-based approach, we derive a general rule of thumb to determine whether two protein-protein interfaces are redundant. With this method, we arrive at a nonredundant set of 209 recognition complexes—the largest set reported so far. Proteins 2003;53:629–639.** © **2003 Wiley-Liss, Inc.**

Key words: protein interactions; protein recognition; protein interfaces; protein complexes; classification

INTRODUCTION

Interactions between proteins and the specificity of protein recognition lie at the base of our understanding of biological function and most cellular processes. Establishing the rules that govern the specificity and formation of protein-protein complexes necessitates an understanding of the interactions that take place on the molecular level. As the number of available complex structures grows, so does the depth of our understanding as exhibited by a number of overview studies in recent years.[1–3] These reviews analyze the general physical and chemical properties of the interfaces while trying to identify the features that could potentially distinguish the different types of protein-protein recognition sites. With the progress of structural genomics initiatives, our ability to understand the structural basis of different types of protein interaction will become critical to the understanding of functional interrelationships in the proteome.

Despite the increasing number of structures, few attempts have been made to incorporate our growing knowledge of the interface properties into a rigorous clustering or classification scheme. The most extensive *clustering* work used geometrical hashing techniques to compare protein complex structures and then aligned the interfaces with known monomer structures.[4–7] Although this large-scale approach provided fascinating insights into the nature of protein-protein interfaces, the resulting clusters are difficult to interpret. Their approach provided only a qualitative glimpse of the differences between various interfaces subtypes. The authors also pointed out the difficulty in finding an appropriate similarity measure for comparing protein interfaces.

The pioneering protein-protein *classification* work of Ponstingl et al.[8] focused on distinguishing homodimeric complexes from crystal contacts. This problem is important as crystallographers sometimes find it difficult to identify which contacts in the crystal are biologically relevant. 85-88% of the complexes could be discriminated by the change in accessible surface area (ΔASA) upon complex formation or by a statistical potential score. Valdar et al.[9] extended that work by including sequence conservation on the surface of the protomers as an additional feature and used multilayer perceptron neural networks to achieve a 92% classification accuracy based on ΔASA and sequence conservation as inputs.

A natural extension of the above work by Thornton and colleagues would be to distinguish between permanent oligomeric proteins and transient complexes. It is important to understand the differences between these two classes of complexes, especially the distinguishing properties of transient complexes since these are usually the key players in regulation of major biochemical pathways and signaling cascades. Various authors have referred to these two classes of complexes using different nomenclatures,

Grant sponsor: National Science Foundation; Grant numbers: DBI-0078194, DBI-0133834, and DBI-0116574 Grant sponsor: DOE/Krell Institute CSGF Fellowship to J.M.

*Correspondence to: Zhiping Weng, Department of Biomedical Engineering and Bioinformatics Program, Boston University, 44 Cummington Street, Boston, MA 02215. E-mail:zhiping@bu.edu

Received 15 November 2002; Accepted 23 January 2003

namely permanent and transient complexes[3] or two-state and three-state complexes.[10] Two-state complexes represent those, for which the processes of folding and binding are essentially inseparable, and three-state complexes describe two components that fold independently and then associate. Although folding-type complexes may be both homomers and heteromers, most would agree that recognition type complexes are restricted to heteromers only. Recognition type heterocomplexes can be further subdivided into smaller functional categories such as antibody-antigen, enzyme-inhibitor, enzyme-cofactor, enzyme-substrate, hormone-receptor, and signaling protein-effector. It is important to keep in mind, however, that such complex biological notions do not easily lend themselves to rigid classification schemes. It is often taken for granted that the oligomeric fold-together homo- and heterocomplexes are inherently different from recognition heterocomplexes. As usual, in biology, the distinction is not clear-cut. Sometimes, it is not easy to determine the nature of the complex even after consulting the literature, as different authors take different views or the crystallographers are not sure which form of the complex is biologically relevant.[11] Although one may often assume that homodimers are usually permanent complexes, there are known cases of receptors that dimerize only upon the initial binding of the ligand,[12] thus suggesting a recognition component.

In this study, we focus on classification of physiologically relevant complexes into two classes: folding complexes and recognition complexes. Specifically, we aim to make the distinction quantitative based on physico-chemical considerations. At first, this may seem an easy task, since the interfaces of folding complexes have been reported to be more hydrophobic than those of recognition complexes.[2,3] While on average this is a true statement, there is great overlap and we show that hydrophobicity does no hold a sufficient discriminatory power. Instead, we have taken a vector classification approach and prove that they are more suitable for resolving the differences between the two types of interfaces.

We introduce the atomic contact vectors (ACVs) derived from protein interface structures as a way to represent the physico-chemical nature of the interface. The simple vector format makes complex structures amenable to standard multivariate analysis techniques. We apply ACVs to the homodimer/crystal contact problem studied by Thornton and colleagues and obtain results at least as good as those of the previous studies. In addition, we propose a way to take into account symmetry, which turns out to be a useful tool in discriminating homodimers. We examine the ACVs as input both for classification and as a general similarity measure for clustering and illustrate their superiority over sequence-based methods. Armed with such a similarity measure, we undertake the task of extracting all possible protein-protein interfaces from the PDB.[13] Using standard clustering procedures, we sift through the available data to, in the end, produce the largest reported set of nonredundant structures of recognition-type protein-protein complexes.

THEORY AND METHODS

Atomic Contact Vectors

A protein complex interface can be described using residues or atoms from two component proteins that are within some distance cutoff or by including the distance in the formulation. Statistical knowledge-based potentials assign scores to different types of residue/atom pairs, which are assumed to be additive in order to produce a total energy score.[14–16] These approaches usually involve choosing a reference state for normalization, which is often a crucial and debatable step.

ACVs offer several advantages in analysis and classification of protein-protein interfaces. Although in many ways similar to the traditional approaches such as pair potentials and potentials of mean force, ACVs do not require the assumptions of additivity, and they are not affected by any reference state calculations. The major difference is that instead of calculating the potential for each atom type pair and then summing to obtain a final score for the interface, we leave the scores in a vector form, with a count for each atom type contact, including a zero if no such contact exists. Previously, authors avoided using such potentials at the atomic level because some of the expected pair counts would be zero.[15] However, we propose that a lack of contacts for a certain atom type pair carries as much information about the nature of the protein-protein interface as a nonzero score, if not more. Using these multidimensional vectors for classification is an intuitive way to capture the physico-chemical properties of the interface. In their raw form, the vectors implicitly carry information about the size of the interface (as well as the size of different subcomponents), although normalizing them by the total number of contacts results in vectors representing just the contact preferences and allowing us to compare the properties of interfaces of different size.

Two decisions must be made for the use of ACVs. First, should we use residue pairs or atom pairs? If the latter, how should we define atom types? A variety of atom-typing schemes exist in the literature. Zhang et al.[16] developed atomic contact energy (ACE) with applications in protein structure prediction and docking and we chose this scheme based on the considerations described in their work. Eighteen atom types are defined in ACE, for the optimal characterization of contacts in the interior of soluble proteins. The use of these atom types leads to $(C_2^{18} + 18) = 171$ types of atom pairs, thus the dimension of the interface vector. The other decision to be made is the distance cutoff. Some groups use distance-dependent contact potentials, and others define the local environment by choosing a specific distance cutoff. Zhang et al. used a fixed distance cutoff of 6 Å, supported by considering the packing density in the protein interior and comparing it with theoretical atomic volume calculations. Because it has been shown that the packing density at the interface of complexes is approximately the same as that in the interior of proteins,[2] we feel that this is a sensible cutoff value. Varying the contact radius in the 5–7 Å range did not lead to a substantial difference in our classification and clustering results (data not shown). We note that the total number of contacts (i.e., the sum over the ACVs) correlate extremely well (correlation coefficient 0.9) with the ΔASA as determined using the NACCESS program,[17] and, with the choice of a 6 Å cutoff, the slope is close to unity. We calculate ΔASA lost upon complex formation by subtracting the accessible surface area of the complex from the sum of the areas of the protomers and dividing the result by two to obtain the interface area per protomer.

Finally, we use two different versions of the contact vectors: one with simply the raw counts of the number of contacts across the interface within the cutoff radius and another with these counts normalized by the total. In the latter, each element of the vector thus represents a proportion of that contact type with respect to the total. This form of the vector makes it easier to compare interfaces that substantially differ in size but may still be similar in their chemical properties.

Classification

Representing the interface contact map as a simple vector allows us to take advantage of the wealth of methods in multivariate analysis and pattern recognition to answer biological questions. For the purposes of classification, we tried several discriminant approaches. At the base, all approaches are adaptations of the Fisher's linear discriminant (FLD).[18,19] The idea is to find the projection $\mathbf{w}$, which, upon mapping the input data $\mathbf{x}$ into a lower dimensional space, will enable us to classify the data in a simple manner. The discrimination criterion of this classic method is based on maximizing the ratio of between-class scatter (or covariance) to within-class scatter. The scatter matrix is defined as

$$\mathbf{S}_i = \sum_i (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T, \tag{1}$$

where $\mathbf{m}_i$ is the class mean. The between-class scatter, for a two-class problem, is defined as

$$\mathbf{S}_B = (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T. \tag{2}$$

The within-class scatter is traditionally defined for a two-class problem as

$$\mathbf{S}_W = p_1\mathbf{S}_1 + p_2\mathbf{S}_2. \tag{3}$$

In all cases described here, the prior probabilities p_1 and p_2 of belonging to a class are not known and can be dropped with the assumption of equal probability. We implemented the traditional linear case of FLD as well as a variation of it where we treat the classes separately and find two optimal projections, one for each class instead of a single projection for both classes as in the traditional case.[20] The optimization criteria are defined as

$$J_i(\mathbf{w}_i) = \frac{\mathbf{w}_i^T \, \mathbf{S}_B \, \mathbf{w}_i}{\mathbf{w}_i^T \, \mathbf{S}_{wi} \, \mathbf{w}}. \tag{4}$$

It can be shown that to maximize J, the vector $\mathbf{w}$ must satisfy a generalized eigenvalue problem

$$\mathbf{S}_B\mathbf{w} = \lambda \mathbf{S}_w \, \mathbf{w}. \tag{5}$$

The classification of a test vector $\mathbf{v}$ is then performed using the Mahalonobis nearest-neighbor distance:

$$Class = \min_i \{(\mathbf{w}^T \mathbf{v} - \mathbf{m}_i)\mathrm{cov}(\mathbf{x}_i)^{-1}(\mathbf{w}^T\mathbf{v} - \mathbf{m}_i)^T\} i\,. \tag{6}$$

Using the separate projections for each class essentially makes this classifier quadratic and we will refer to it as quadratic Fisher discriminant (QFD). In all cases we have seen, QFD is at least as good or better than the traditional FLD and we therefore use this classifier and do not report results of the strictly linear approach.

The second classification method we used is kernel discriminant analysis (KDA), also known as generalized discriminant analysis or kernel Fisher discriminant, as recently described.[21] There have been a number of efforts in this area in the pattern recognition community and it has been shown that KDA is closely related to the multilayer perceptron neural networks, which were used by Valdar and Thornton in the homodimer/crystal contact classification.[9] Briefly, the kernel approach allows the projection of data into a hyper-dimensional feature space where it is more easily separable. This approach was first introduced for support vector machines (SVM), but recently has also been applied to a number of classical methods,[22] including Fisher's discriminant. KDA is different from SVM in that the latter uses only the part of the data at the class boundary for the decision function, whereas KDA takes into account the whole distributions. The optimization criterion for the nonlinear case can be written similarly to Eq. 4 above, but with the data transformed by a chosen kernel function. Here, we used a radial basis function (RBF) kernel.

For both of the above methods, a problem arises due to the singularity of the S_W matrix because of the relatively small number of samples with respect to the number of features. To solve this problem, we used Principal Component Analysis (PCA) to reduce the dimensionality of the data prior to classification. We later discovered that the same approach was taken by Kriegman et al. in a face image recognition study.[23]

Dataset Retrieval

For the problem of homodimers/crystal contact classification, we used the dataset of Ponstingl et al.[8] It consists of 76 homodimers and 95 crystal contacts, which were chosen to be the largest contacts in each crystal as described by the authors. To obtain data for heterodimers, we developed a set of automated procedures for extraction of an almost exhaustive set of heteromeric interfaces from the PDB. As pointed out by Tsai et al., any such undertaking inevitably involves a number of subjective decisions, and we will describe the procedure briefly. First, we removed all records with resolution worse than 3.25 Å and all structures determined by methods other than X-ray crystallography. Then we filtered the PDB to eliminate all single chains and all chains smaller than 25 amino acids long. In cases of multiple conformations for the same residue, we kept the highest occupancy atoms. Then, we eliminated all homomeric records using the BLASTCLUST algorithm.[24] A homomeric record was defined where all chains have at

TABLE I. Summary of Classification Results

		QFD with ACV		KDA with ACV	
Classification	ΔASA (%)	Raw (%)	Normalized (%)	Raw (%)	Normalized (%)
Homodimer/Crystal Contact[a]	84.6	91.8 (92.4)	N/A	93.0 (95.3)	N/A
Recognition/Folding	75.9	80.0	69.6	91.0	74.0

[a] The values in parentheses show improved classification results after taking into account symmetry considerations. Please see text for details.

least 85% sequence identity to each other and at least 50% of the sequence was aligned. The resulting set of PDB records was used in the following nonredundant structure retrieval scheme. For each record that contains more than two chains, we constructed a matrix of all pairwise ACVs and also a matrix of distances between these ACVs.

We then defined a similarity cutoff to select a nonredundant set of interfaces for those PDB records that contain multiple instances of the same complex in the asymmetric unit. We counted all chains connected by disulfide bonds as a single unit. In addition, we used the QFD classifier described above, with the homodimer/crystal contact data[8] as the training set, to group together homomeric chains resulting in more realistic stoichiometries for complexes such as 2:1 and 2:2 PDB chains. Using the clustering methods described below, we also developed a procedure to automatically recognize antibody heavy chain/light chain complexes (one of the largest subsets) and group them together even when the disulfide bond between them does not exist or is not annotated. The procedure resulted in a dataset of 1063 heteromeric complexes, including both recognition type interfaces as well as permanent fold-together type heteromers.

Having dealt with interface redundancy at the level of a single asymmetric unit, we next cluster redundant and closely related complexes. We first clustered the complexes using sequence-based methods. Such an approach is restricted to comparison of complexes with identical stoichiometry. Following Ponstingl and Thornton, we chose a 25% sequence identity threshold and used pairwise BLAST[25] to carry out all-against-all sequence comparisons for all chains in the 1063 complexes. Although we feel that 25% is conservative, remote homology could still escape this cutoff. Using the Bron-Kerbosch algorithm,[26] we then identified groups of complexes that formed fully connected components, or cliques, meaning that all members are transitively similar to each other. To define a nonredundant set of recognition complexes, we used standard agglomerative hierarchical clustering algorithms with the Euclidean distance metric.[18]

RESULTS

Comparison of Homodimer/Crystal Contact Classification

To test the utility of ACVs in the context of classification, we applied our methods to the nonredundant dataset of 76 homodimers and 95 crystal contacts between monomers used by Ponstingl and Thornton. We applied both classification methods (QFD and KDA) using un-normalized ACVs with leave-one-out crossvalidation. The FLD method correctly classified 158/171 complexes, or 92.4%. The more sophisticated KDA method in this case showed almost no improvement, identifying just one more correct complex and resulting in 93.0% accuracy (Table I). In the framework of leave-one-out crossvalidation, we misclassified six dimers as monomers, and six monomers as dimers.

Ponstingl et al.[8] showed that an 84.6 % success rate could be achieved by discriminating on interface ΔASA because crystal contacts tend to be smaller than homodimer interfaces. With a distance dependent statistical potential score, they improved to 87.5% success rate. In a subsequent study,[9] the authors combined ΔASA and a newly introduced probabilistic measure of amino acid conservation at the interface in a multilayer perceptron neural network to obtain a 92% classification accuracy. Unfortunately, the results in the above two studies are not directly comparable. The need for a sufficient number of sequences in each of the homodimer protein families to infer a conservation score prompted the authors to reduce the number of homodimer complexes and crystal contact structures. In addition, although in the first work the 95 crystal contacts represented the largest contact for each crystal structure, Valdar and Thornton incorporated all other crystal contacts in those structures for the second work. The added crystal contacts were all smaller and therefore relatively easy to classify. Our results compare competitively with the ones reported in the above two studies. Although the difference in classification accuracy is relatively small, our structure-based method is not limited by considerations that hindered previous studies.

In the course of our analysis, we noticed that for this particular classification problem, it is useful to consider the symmetry of the complex because we expect the homodimers to be symmetric about their interface. There are some exceptions such as when one of the molecules has undergone significant conformational change, the overall symmetry may be somewhat disrupted but at the interface it is usually preserved. As a simple way to measure symmetry, independent of the specific orientation, we compared for each complex two vectors of length 18, each element representing the number of contacts a particular atom type makes with any atom of the other protomer. This approach showed that all 76 dimers in the dataset were symmetric. On the other hand, 30% of the monomer dataset is not symmetric about the interface and thus can be immediately classified as monomers. The structure of the largest crystal contact in 1KWA—a PDZ domain of a human synaptic protein, for instance, is not symmetric

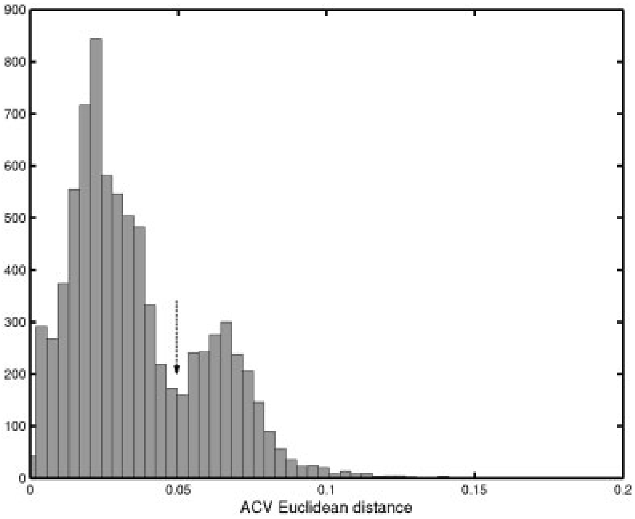

Fig. 1. Distribution of all pairwise Euclidean distances within all cliques containing more than one member. The distribution is clearly bimodal with a separation at the distance ~0.05 (arrow). The right-hand side of the distribution suggests discrepancies in the sequence-based cliques and structure-based ACVs.

and therefore the possibility of it being a homodimer can be readily ruled out. Several nonsymmetric crystal contacts can be found among the cases reported as misclassified by Ponstingl and Thornton, as well as in our misclassified complexes. In Table I we also report the classification accuracy after taking symmetry into account and it leads to a substantial improvement.

Interface Clustering and Cutoff Derivation

Having shown that the ACVs can be used for complex classification, we were interested in testing their power to identify similarity between a pair of complexes and thus their utility in constructing a nonredundant set of protein complexes. For this purpose we looked for the best independent way to define similar complexes and settled on using sequence similarity of the chains in each complex to define fully connected components or cliques as described in Methods. Using sequence similarity, we were necessarily limited to comparing only complexes with the same stoichiometry; thus two complexes were deemed similar if every chain of each protomer had a sequence identity of >25% to every corresponding chain in the protomer of the other complex in question.

Following this sequence comparison procedure, from 1063 complexes we obtained a dataset of 368 fully connected components. Approximately 80% of the dataset was covered by independent cliques, with every member of the clique transitively similar to every complex within the clique and with no member of the clique having sequence identity of =25% to any other member of the dataset. We found 158 cliques with more than one member. For each of these, we calculated all pairwise Euclidean distances within the clique using normalized ACVs. This resulted in a clearly bimodal distribution of the distances as shown in Figure 1, suggesting that some of the sequence-defined cliques are composed of several ACV clusters. If we assume that the first peak represents all closely related complexes within a clique, how do we account for the second peak? To investigate, we examined the clique that contributed the most to the second peak—this was also the largest clique in the dataset. These turned out to be major histocompatibility complexes (MHCs), complexes of an α-chain and a β-chain. For Class I MHCs, the larger α-chain is the main antigen-presenting component, and the β-chain is the nonpolymorphic β_2-microglobulin. For Class II MHCs, α- and β-chains are of equal size, and both contribute to antigen presentation. Sequences of these proteins are fairly well conserved and it was initially surprising to find that not all of them fell into a single cluster. After a PCA procedure, we projected the interface vectors of 85 members of the clique onto the first two principal components as shown in Figure 2. The data fall into three prominent clusters. Upon further investigation, we found that the cluster on the left was composed solely of Class II MHC molecules, whereas the two clusters on the right were all MHC Class I molecules, thus accounting for the most obvious data separation. Furthermore, upon investigation of the Class I clusters, we found that the data readily separated into mouse and human MHCs, and further into several subclasses, with clusters of points representing HLA A/B/C/E, and H2 K/D/L, showing a surprisingly good correlation with functional subtypes of human and mouse complexes.

Properly separating the MHC clique and checking again for all pairwise Euclidean distances of ACVs resulted in much tighter clusters than those obtained using the sequence comparison procedure; the new clusters all fit in the left peak of Figure 1 (data not shown). We found similar reasons for the other cliques that were responsible for the second peak in Figure 1. Some cliques, such as antibody-antigen complexes also showed high interface vector distance despite high sequence similarity due to complexes of antibodies targeting entirely different epitopes on the same antigen.

The above analysis, while giving strong indication of the range of the ACV distances that represent highly similar complexes, cannot be effectively used to arrive at a well-defined similarity cutoff. For this purpose we clustered normalized ACV vectors of the 1063 complexes using a standard neighbor-joining complete (farthest) linkage algorithm. The distribution of the joining distances at each linkage step of the algorithm is shown in Figure 3. There exists a clear separation between two apparent large peaks, with the left peak encompassing the same range of distances as the left peak in Figure 1. This suggests that the left peak represents complexes that are similar and the right peak accounts for clusters formed later in the procedure (with greater linking distance), and thus less similar to each other. Interestingly, two peaks in Figure 3 separate at ACV distance of ~0.05, similar to the distance separating ACVs in sequence cliques (Figure 1).

The distribution in Figure 3 can be reasonably well modeled as a mixture of two lognormal distributions and this model can be used to derive the parameters for each

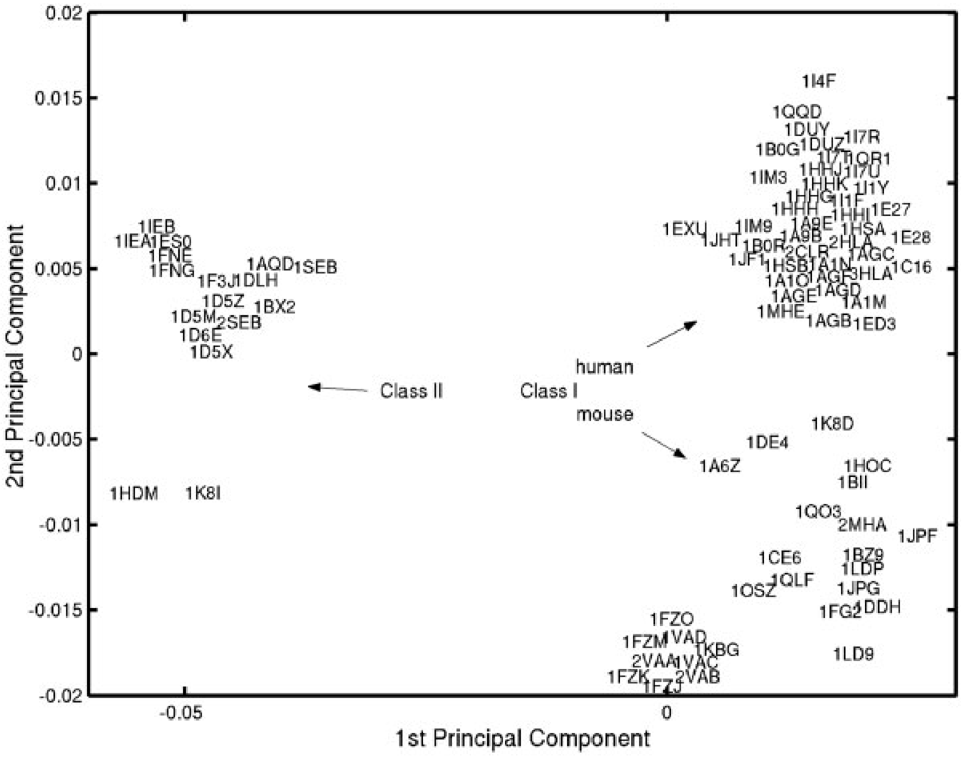

Fig. 2. The largest sequence-based clique containing 85 MHC structures (α and β chain interfaces), projected onto the first two principal components following a PCA procedure. Further analysis showed that the cluster on the left side of the plot is composed solely of the Class II MHC molecules, whereas the two on the right are all Class I MHCs, thus accounting for the most pronounced differences in the data as detected by the ACVs. The smaller separation between the two clusters on the right represents the differences in the MHC Class I interface contacts of human (top) and mouse (bottom) molecules.

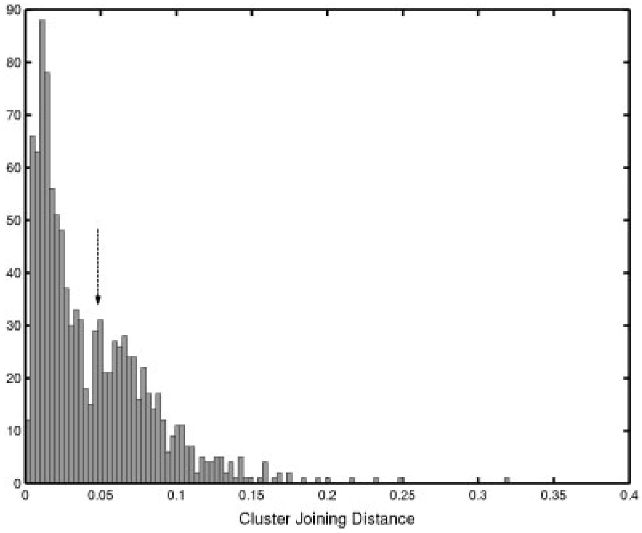

Fig. 3. Distribution of all cluster-joining distances in the process of the linkage algorithm. The left peak represents more similar vectors, joined earlier in the clustering procedure, with the right peak representing more distant clusters. Note that the cutoff distance separating the two peaks occurs again at ~0.05 (arrow), similar to that in Figure 1.

distribution. With these parameters, we can rigorously obtain the natural cutoff point, defining a certain level of similarity. Because the sizes, means and standard deviations of both hypothesized distributions are fixed, the definition of the cutoff is somewhat arbitrary within a range of values and depends only on the chosen levels of statistical significance α and β, which represent the probability that a certain ACV belongs to the "similar" distribution on the left or the "nonsimilar" distribution on the right. A plot of the relationship of cumulative probability values in the range of 0.80 to 0.99 is shown in Figure 4. It turns out that there is significant overlap between the distributions, resulting in a "gray area" where the level of similarity is uncertain. This gray area is defined by the bounds of 95% probability. Thus, given a certain ACV distance between two interfaces that is smaller than the lower bound of the "gray region," we can be 95% certain that the two interfaces are redundant. On the other hand, if the distance is larger than the upper bound of the gray region, we can be 95% certain that the two interfaces are substantially different. In our experience, a conservative cutoff of ~0.045 is a good first approximation, whereas a value of 0.05 still gives very good results with few false classifications. Using such a cutoff we can obtain clusters from the linkage analysis above, and because this interface distance value represents a natural division observed in the data, we argue that by choosing a single representative from each of these clusters we can arrive at a nonredundant set of protein-protein complexes. Note that the derivation above allows a definition of redundancy that is

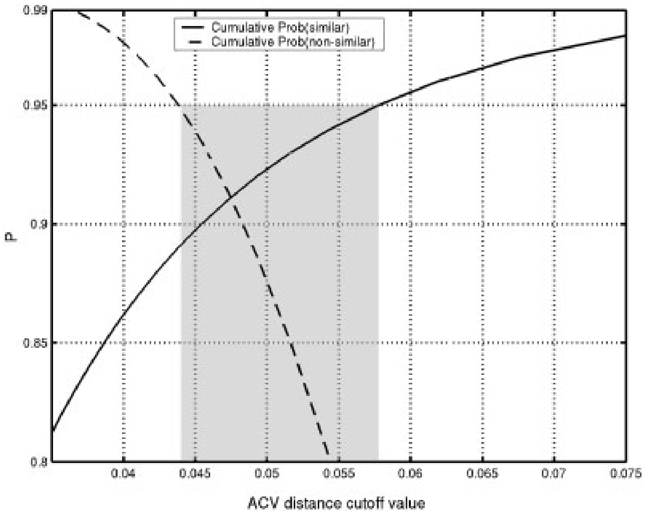

Fig. 4. Using the distribution parameters obtained from the bimodal fit of the distributions in Figure 3, the cumulative probabilities of finding an ACV distance in either distribution is presented. Solid line represents the cumulative probability of two interfaces being similar for all distances smaller than a cutoff value. Dashed line represents the cumulative probability of two interfaces being not similar for all distances greater than a cutoff value. Despite the substantial overlap, which represents the gray area of interface similarity, there is a reasonably well-defined separation, which allows a derivation of a cutoff value for complex similarity.

independent of the nature of the complexes, the size of the interacting proteins, the size of the interface, sequence similarity, or the stoichiometry of interaction. In addition, depending on the specific research needs, more strict or permissive cutoffs can be defined with the above analysis.

Comparing our results with a sequence-based approach, our ACVs with a Euclidean distance metric provide a suitable, sequence-independent method to assess similarity between protein complexes. This addresses the problem of how to assess similarity between antibody-antigen and other immune system complexes, which make up a significant proportion of solved complex structures. These complexes traditionally required special handling because of high sequence similarity of most of the antibody sequence.[27] Our approach offers a more general way to assess the similarity between two protein interfaces.

Recognition Complex/Folding Oligomer Classification

Starting with the 368 cliques generated by sequence comparison and selecting a single representative (crystal structure with the best resolution) from each clique, we manually classified all the complexes into recognition and permanent folding subclasses relying on the associated literature. In the process, we eliminated some complexes, which were deemed inappropriate because they were unnatural (i.e., synthetic, or engineered with crosslinks) or were products of cleavage. In addition, we had to eliminate 75 cliques because we could not classify them with a reasonable degree of certainty based on the available literature. This group included many nucleic acid binding complexes, chaperone complexes, and some others. In the end we were left with 198 recognition complexes and 71 folding complexes. Adding 76 (folding) homodimers from Thornton's set, we arrived at the final set of 345 complexes.

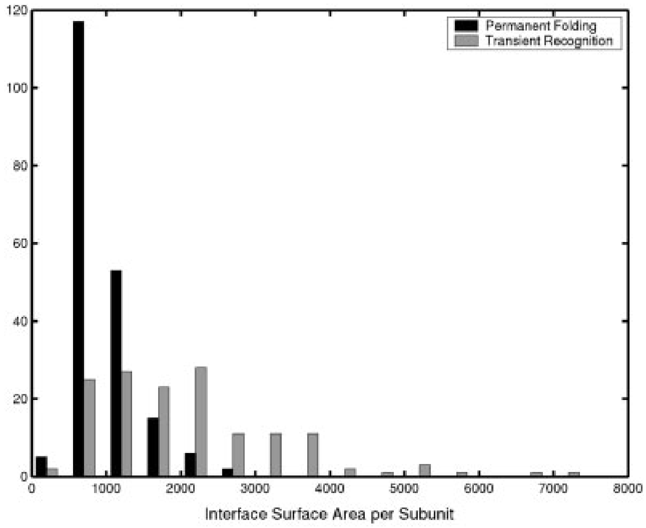

Fig. 5. Distribution of ΔASA for recognition and folding complexes.

It is often assumed that the oligomeric fold-together homo- and heterocomplexes are inherently different from recognition heterocomplexes. Oligomers are believed to have larger and more hydrophobic interfaces than recognition complexes. Thus we computed the ΔASA and hydrophobicity scores for all 345 complexes. Calculation of the latter was based on the hydrophobicity parameters of Fauchere and Pliska derived from the experimental free energies of transfer of *N*-acetyl-amino-acid amides between octanol and water.[28] Classification based on ΔASA resulted in 76% success rate, indicating that the size of the interface, although not as dominating as in the homodimer/crystal contacts case, is still a very important factor in distinguishing recognition complexes from permanent folding complexes. The hydrophobicity score proved to be a rather poor discriminating feature with only a 61.2% success rate. The distribution of interface areas and hydrophobicity scores for both classes of complexes is shown in Figures 5 and 6. From these distributions, it is apparent that recognition complexes, although smaller than folding oligomers on average, also have a smaller dynamic range. This is in agreement with previous studies,[2] which suggested that the surface area of the interface for recognition complexes is limited by the intrinsic physical requirement for each of the protomers to fold independently and exist in solution without aggregating.

We applied the QFD and KDA classification methods to the data set with leave-one-out crossvalidation and the results are presented in Table I. For this classification problem, the difference between KDA and QFD was more pronounced with 11 percentage points improvement for the kernel-based approach. KDA with un-normalized ACVs performed best, thus combining the interface size factor together with the pair contact preferences. Note that the success rate for normalized interface vectors with KDA (74.0%) is almost as good as that for ΔASA (75.9%), suggesting a significant difference in composition of pair-

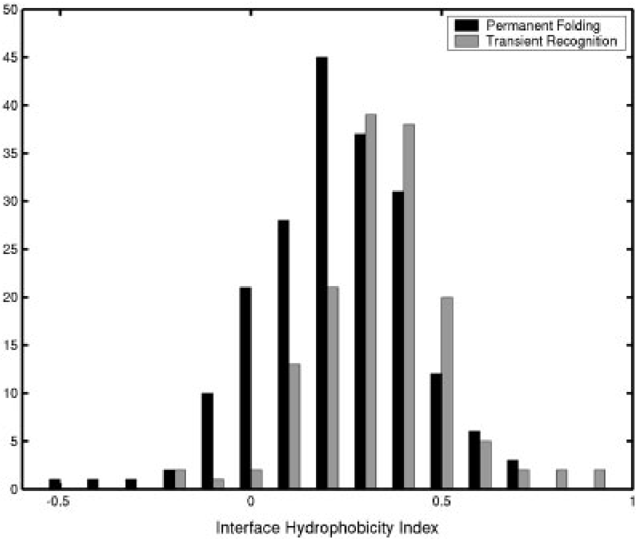

Fig. 6. Distribution of hydrophobicity scores based on Fauchere and Pliska for recognition and folding types of complexes. Although recognition complexes are certainly more hydrophilic on average, this score is clearly not a good discriminating feature.

wise interface contacts. These results are encouraging with respect to our ability to differentiate between such complex biological phenomena.

Nonredundant Dataset of Recognition Complexes

Although the dataset for classification above came from the sequence-based cliques, we used the power of ACVs to arrive at a nonredundant set of recognition complexes. Such a dataset is essential to further studies in protein recognition and docking. Large-scale studies in the past have used datasets of mixed homo- and heterodimers and both transient and recognition complexes. The manual classification of the complexes as described above served as a starting point. We used standard hierarchical clustering techniques and the similarity cutoff derived above to create a nonredundant set of recognition complexe—noncovalent transient complexes resulting from association of folded protomers. It was curated and checked manually and crosschecked with sequence-based clique clustering to ensure no redundancy. The final set (Table II) consists of 209 complexes, including 34 antibody-antigen complexes and 60 enzyme-inhibitor complexes. This set is twice as large as a recent estimate of the number of such complexes in the PDB.[29]

DISCUSSION

In classifying protein complexes, it is important to keep in mind that intricate biological phenomena that lie at the base of protein interactions do not always lend themselves to a clear-cut separation. Thornton and colleagues suggested that among the homodimers there is a fraction of complexes that exist in monomer-dimer equilibrium.[8,9] Some complexes have been crystallized where the proteins appeared as dimers or monomers, depending on the crystallization conditions[30] or as a result of a “domain swap”[31]; cases of ligand-induced dimerization are also known. Deciding whether a complex is permanent or transient is sometimes quite difficult. If proteins such as some of the transcription complexes are known to exist both as a homodimers and as heterodimers, what does that say about the nature of the complexes? If we define transient, recognition complexes as those that fold first and then associate, then how should we classify complexes of proteins with their chaperones, which assist in folding? If we use affinity measures to evaluate the nature of a complex, then how should we treat a normally obligate complex, which binds weakly after one or more key residues at the interface have been mutated? Questions of this nature account for the gray area in Figure 4. Having defined a similarity cutoff region above, we present two cases of interfaces, which fall in the gray area, indicating that a level of similarity may nevertheless exist.

The distance between ACVs for two elongation factor complexes from *Escherichia coli* (1EFU[32]) and *Thermus thermophilus* (1AIP[33]) is 0.0493. The *E. coli* complex is an Ef-Tu/Ef-Ts heterodimer, whereas the *T. thermophilus* structure is a 1:2 complex between Ef-Tu and an Ef-Ts homodimer [Fig. 7(A)]. Although the overall architecture of Ef-Ts in the two structures is different, the interface is quite similar because of a structural repeat that spans the interface of the Ef-Ts homodimer. Several key residues are conserved on the symmetric bipartite Ef-Ts as compared to the monomer, contributing to the ACV distance.

Although the above example illustrates that ACVs are capable of detecting borderline similarity for two complexes with different stoichiometry but conserved structural elements, Figure 7(B) shows that ACVs are also useful in assessing similarity between interfaces with no structural homology. The complexes are both serine protease/inhibitor complexes and are a classic example of convergent evolution of proteases. Subtilisin from *Bacillus subtilis* (1CSE[34]) and Protease B from *Streptomyces griseus* (1SGP[35]) have distinctly different folds; the former has an α-β-α structure, whereas the latter is a duplicated β-barrel. They do, however, have the same mechanism of action involving a similar catalytic triad. The inhibitor structures are also different and they represent two different protease inhibitor families. Subtilisin is in complex with Eglin C of CI-2 inhibitor family, and Protease B is inhibited by Turkey Ovomucoid Inhibitor of the Kazal family. Despite the structural differences, the ACV distance for these interfaces is 0.0498—toward the middle of the gray area in Figure 4.

Finally, although we may often tend to think that interfaces between two proteins of identical fold and similar orientation should have similar interfaces, we saw cases where this is not the case. For instance, our set in Table II contains two complexes of endonuclease domains of bacterial colicins 7 and 9 with their respective immunity proteins. Both structures and sequences of these proteins are well conserved [Fig. 7(C)], including a key Tyr residue at the interface. Despite this similarity, the ACV distance for the two interfaces is 0.074—substantially greater than the upper bound of

TABLE II. A Nonredundant Dataset of 209 Transient Recognition Complexes

1A2K A:D	1A2Y AB:C	1A4Y A:B	1ACB E:I	1ADQ A:HL	1AGR A:E
1AIP A:CD	1AK4 A:D	1ARO P:L	1ATN A:D	1AVA A:C	1AVG HL:I
1AVW A:B	1AVZ B:C	1AXI A:B	1AY7 A:B	1AZZ A:CD	1B2S A:D
1B41 A:B	1B6C A:B	1BDJ A:B	1BGX HL:T	1BJ1 HL:W	1BKD R:S
1BLX A:B	1BML A:C	1BP3 A:B	1BQQ T:M	1BUH A:B	1BVN P:T
1BZQ A:L	1C1Y A:B	1C4Z A:D	1CA0 BC:D	1CD9 A:B	1CDK A:I
1CDM A:B	1CHO E:I	1CIC AB:CD	1CLV I:A	1CMX A:B	1CN4 AB:C
1CSE I:E	1CXZ A:B	1D2Z A:B	1D5M A:C	1D6R I:A	1DE4 A:C
1DEE CD:G	1DEV A:B	1DF9 B:C	1DFJ I:E	1DHK A:B	1DKG AB:D
1DN1 A:B	1DPJ A:B	1DQJ AB:C	1DS6 A:B	1DTD A:B	1DU3 DF:A
1DX5 AM:I	1DZB X:A	1EOF AD:I	1E0O A:B	1E44 A:B	1E6J HL:P
1E96 A:B	1EAI A:C	1EAY A:C	1EBD AB:C	1EBP A:CD	1EFU A:B
1EGJ A:HL	1EJA A:B	1EMV A:B	1EO8 AB:HL	1ES7 AC:D	1EUV A:B
1EV2 A:E	1F02 I:T	1F34 A:B	1F3V A:B	1F51 AB:E	1F5Q A:B
1F60 A:B	1F7Z A:I	1F93 AB:EF	1FAK HL:T	1FAK I:HL	1FBI HL:X
1FC2 C:D	1FE8 A:HL	1FJ1 AB:F	1FLE E:I	1FLT VW:Y	1FNS HL:A
1FOE A:B	1FQ1 A:B	1FQK A:B	1FQV A:B	1FSK A:BC	1FYH A:B
1GOY I:R	1G3N A:C	1G4U R:S	1G4Y R:B	1G73 A:D	1G9M C:G
1G9M G:HL	1GCQ B:C	1GH6 A:B	1GL0 I:E	1GL4 A:B	1GOT A:B
1HCF AB:Y	1HE1 A:C	1HE8 A:B	1HEZ AB:E	1HIA AB:I	1HX1 A:B
1HYR AB:C	1I1R A:B	1I2M A:B	1I4D AB:D	1I4O A:C	1I5K A:C
1I7W A:B	1I8L A:C	1IAR A:B	1IB1 AB:E	1IBR A:B	1ICF AB:I
1IHS HL:I	1IIS AB:C	1IM3 A:D	1IM9 A:D	1IQ5 A:D	1IQD AB:C
1IRA X:Y	1ITB A:B	1J7V R:L	1JDH A:B	1JDP H:A	1JHL HL:A
1JIW P:I	1JLT A:B	1JMA A:B	1JPS HL:T	1JRH HL:I	1JTD A:B
1JTG A:B	1JTP A:L	1K4C AB:C	1K90 A:D	1K9O E:I	1KAC A:B
1KCG AB:C	1KIG HL:I	1LFD AC:B	1LPB A:B	1MLC AB:E	1NCC N:HL
1NFD EF:AB	1NRN HL:R	1NSN HL:S	1OSP HL:O	1PPF I:E	1QA9 A:B
1QAV A:B	1QBK B:C	1QFU AB:HL	1QGK A:B	1QKZ A:HL	1QMZ A:B
1QO0 A:DE	1QO3 A:CD	1RLB ABCD:F	1RRP A:B	1SBB A:B	1SGP I:E
1SLU A:B	1STF I:E	1T7P A:B	1TBR HL:R	1TMQ A:B	1TNR A:R
1TOC R:AB	1TX4 A:B	1UGH I:E	1VRK A:B	1WEJ HL:F	1WQ1 G:R
1WWW VW:Y	1XDT R:T	1YCS A:B	1ZBD A:B	2BTC I:E	2BTF P:A
2HMI B:CD	2JEL HL:P	2PCC A:B	2SIC E:I	2VIR AB:C	3BTH I:E
3HFL HL:Y	3YGS P:C	4HTC HL:I	4SGB I:E	7CEI A:B	

the gray area. Colicins and their immunity proteins are known to crossreact to some extent, and the Im7 immunity protein (7CEI[36]) can inhibit the E9 DNase (1EMV[37]), but the affinity of this interaction is 10 orders of magnitude weaker than that of Im7 with its cognate colicin (Kuhlmann et al.[37] and references therein). Despite some conserved interface residues and similar solvent accessibility of the interface (691 Å for E7/Im7 and 767 Å for E9/Im9), the general nature of interactions is rather different—mostly hydrophobic in the E9/Im9 complex and predominantly charged in E7/Im7. Hydrophobicity calculations as described above showed a hydrophobicity index of $\sim$0.255 for the former and $\sim$0.098 for the latter. Thus in this case, the ACV method distinguishes the specificity of interactions between structurally similar complexes.

CONCLUSION

In this study, we introduced ACVs and applied them to two classification problems—distinguishing homodimers from crystal contacts and obligate oligomers from transient recognition complexes. Although our methods are similar to many traditional pair potentials, we show that they perform better in classification because of our ability to use nonlinear discriminators with these multidimensional vectors. We also show that ACVs constitute a powerful tool for general comparison of protein-protein interfaces independent of the size of the interface and stoichiometry of the complex. We use them to develop an automated structure-driven procedure to retrieve a close-to-complete set of all heterocomplexes in the PDB. As a general similarity measure, we demonstrate that ACVs are capable of detecting similarities and differences where sequence-based approaches fail and derive from statistics a general rule of thumb to decide whether two interfaces are similar to each other.

Applying the above results to the dataset retrieved from the PDB, we use standard clustering techniques to arrive at a nonredundant set of transient recognition complexes. This is the largest such set ever reported and it should extremely useful for understanding the both the general principles and the diversity in protein-protein recognition. Our study of classification of permanent vs. transient recognition complexes is the first quantitative study of this kind and we hope that future

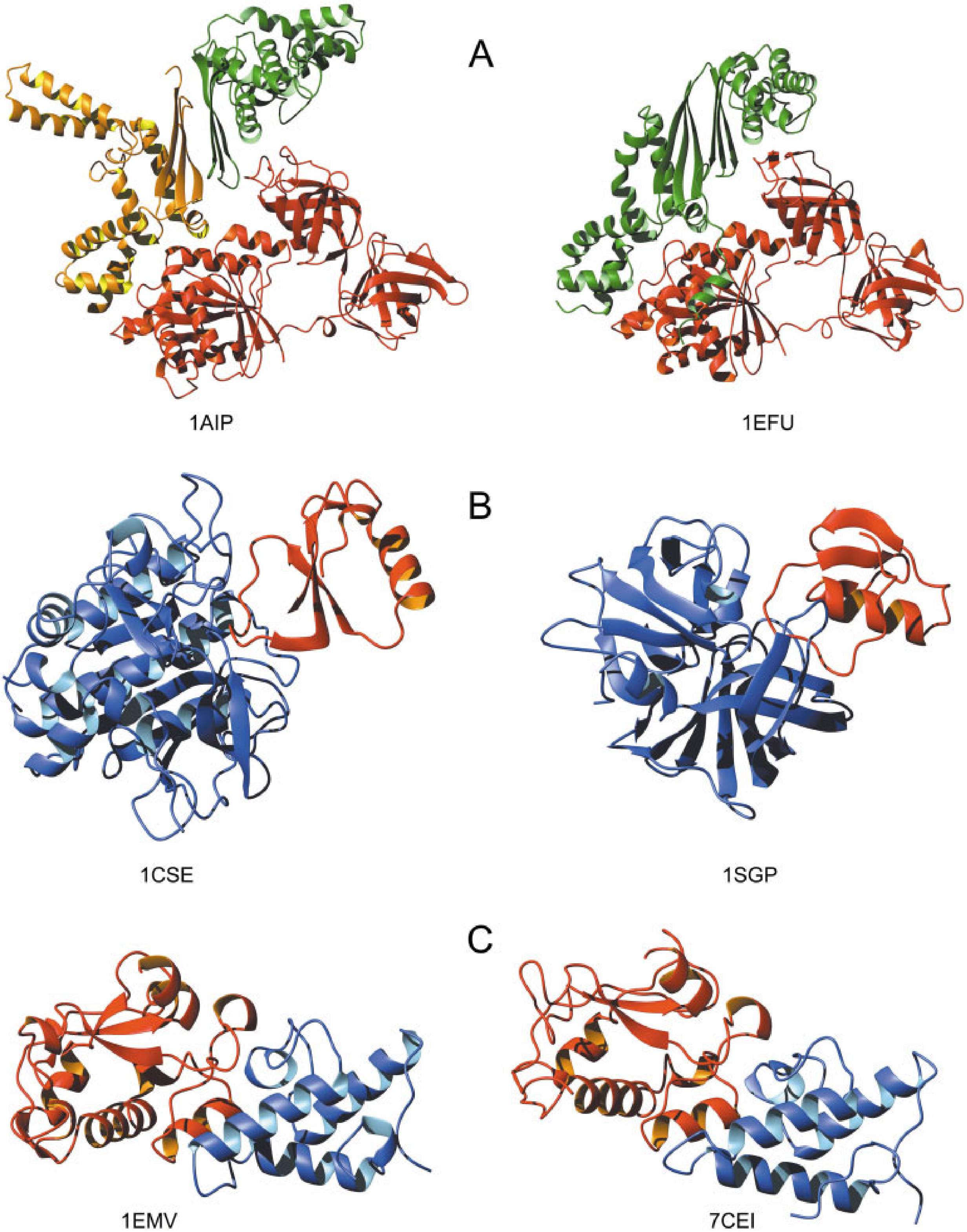

Fig. 7. Interface examples. (A) Comparison of Ef-Ts/Ef-Tu complexes from *T. thermophilus* (left) and *E. coli* (right). Ef-Tu is shown in red. Ef-Ts in 1AIP is a homodimer (shown in green and yellow), whereas in 1EFU it is a monomer (shown in green); (B) complex structures of Subtilisin-EglinC (left) and Proteinase B-Omptky3. For both complexes, the enzyme is shown in blue and the inhibitor in red; (C) DNA endonuclease domains from colicin 9 (1EMV; left) and colicin 7 (7CEI; right) with their cognate immunity proteins. Both colicin molecules are shown in red, whereas their corresponding immunity proteins are shown in blue. All images were created using MOLMOL.[38]

work in this area will help us better understand the principles that govern molecular recognition, and, in turn, principles of protein interactions that govern most biological interaction networks.

ACKNOWLEDGMENTS

We thank Hannes Ponstingl and the Thornton lab for providing the crystal contacts dataset. We appreciate the help of Gaston Baudat with KDA. We thank Joe Szustakowski for helpful discussions.

REFERENCES

1. Chakrabarti P, Janin J. Dissecting protein-protein recognition sites. Proteins 2002;47(3):334–343.
2. Lo Conte L, Chothia C, Janin J. The atomic structure of protein-protein recognition sites. J Mol Biol 1999;285(5):2177–2198.
3. Jones S, Thornton JM. Principles of protein-protein interactions. Proc Natl Acad Sci USA 1996;93(1):13–20.
4. Tsai CJ, Lin SL, Wolfson HJ, Nussinov R. A dataset of protein-protein interfaces generated with a sequence-order- independent comparison technique. J Mol Biol 1996;260(4):604–620.
5. Tsai CJ, Nussinov R. Hydrophobic folding units derived from dissimilar monomer structures and their interactions. Protein Sci 1997;6(1):24–42.
6. Tsai CJ, Nussinov R. Hydrophobic folding units at protein-protein interfaces: implications to protein folding and to protein-protein association. Protein Sci 1997;6(7):1426–1437.
7. Tsai CJ, Xu D, Nussinov R. Structural motifs at protein-protein interfaces: protein cores versus two-state and three-state model complexes. Protein Sci 1997;6(9):1793–1805.
8. Ponstingl H, Henrick K, Thornton JM. Discriminating between homodimeric and monomeric proteins in the crystalline state. Proteins 2000;41(1):47–57.
9. Valdar WS, Thornton JM. Conservation helps to identify biologically relevant crystal contacts. J Mol Biol 2001;313(2):399–416.
10. Tsai CJ, Xu D, Nussinov R. Protein folding via binding and vice versa. Fold Des 1998;3(4):R71–R80.
11. Nishida M, Nagata K, Hachimori Y, Horiuchi M, Ogura K, Mandiyan V, Schlessinger J, Inagaki F. Novel recognition mode between Vav and Grb2 SH3 domains. EMBO J 2001;20(12):2995–3007.
12. Syed RS, Reid SW, Li C, Cheetham JC, Aoki KH, Liu B, Zhan H, Osslund TD, Chirino AJ, Zhang J, et al. Efficiency of signalling through cytokine receptors depends critically on receptor orientation. Nature 1998;395(6701):511–516.
13. 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(1):235–242.
14. Miyazawa S, Jernigan RL. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. J Mol Biol 1996;256(3):623–644.
15. Moont G, Gabb HA, Sternberg MJ. Use of pair potentials across protein interfaces in screening predicted docked complexes. Proteins 1999;35(3):364–373.
16. Zhang C, Vasmatzis G, Cornette JL, DeLisi C. Determination of atomic desolvation energies from the structures of crystallized proteins. J Mol Biol 1997;267(3):707–726.
17. Hubbard SJ, Thornton JM. NACCESS. 2.1.1: Department of Biochemistry and Molecular Biology, University College London; 1993.
18. Duda RO, Hart PE, Stork DG. Pattern classification. New York: Wiley; 2001. xx, 654 p.
19. Fukunaga K. Introduction to statistical pattern recognition. Boston: Academic Press; 1990. xiii, 591 p.
20. Balakrishnama S, Ganapathiraju A, Picone J. Linear discriminant analysis for signal processing problems. IEEE Southeast Con 1999.
21. Baudat G, Anouar F. Generalized discriminant analysis using a kernel approach. Neural Computation 2000;12:2385–2404.
22. Scholkopf B, Smola AJ, Muller KR. Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation 1998;10:1299–1319.
23. Belhumeur P, Hespanha J, Kriegman D. Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. IEEE Trans PAMI 1997;19(7):711–720.
24. BLASTCLUST: NCBI standalone BLAST executables ftp://ncbi.nlm.nih.gov/blast/executables/.
25. Tatusova TA, Madden TL. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. FEMS Microbiol Lett 1999;174(2):247–250.
26. Bron C, Kerbosch J. Finding all cliques of an undirected subgraph. Commun ACM 1973;16:575–577.
27. Jiang L, Gao Y, Mao F, Liu Z, Lai L. Potential of mean force for protein-protein interaction studies. Proteins 2002;46(2):190–196.
28. Fauchere VL, Pliska V. Hydrophobic parameters pi of amino-acid side chains from the partitioning of *N*-acetyl-amino-acid amides. Eur J Med Chem 1983;18(4):369–375.
29. Halperin I, Ma B, Wolfson H, Nussinov R. Principles of docking: an overview of search algorithms and a guide to scoring functions. Proteins 2002;47(4):409–443.
30. Uhrinova S, Smith MH, Jameson GB, Uhrin D, Sawyer L, Barlow PN. Structural changes accompanying pH-induced dissociation of the beta-lactoglobulin dimer. Biochemistry 2000;39(13):3565–3574.
31. Liu Y, Eisenberg D. 3D domain swapping: as domains continue to swap. Protein Sci 2002;11(6):1285–1299.
32. Kawashima T, Berthet-Colominas C, Wulff M, Cusack S, Leberman R. The structure of the *Escherichia coli* EF-Tu.EF-Ts complex at 2.5 A resolution. Nature 1996;379(6565):511–518.
33. Wang Y, Jiang Y, Meyering-Voss M, Sprinzl M, Sigler PB. Crystal structure of the EF-Tu.EF-Ts complex from *Thermus thermophilus*. Nat Struct Biol 1997;4(8):650–656.
34. Bode W, Papamokos E, Musil D. The high-resolution X-ray crystal structure of the complex formed between subtilisin Carlsberg and eglin c, an elastase inhibitor from the leech Hirudo medicinalis. Structural analysis, subtilisin structure and interface geometry. Eur J Biochem 1987;166(3):673–692.
35. Huang K, Lu W, Anderson S, Laskowski M Jr, James MN. Water molecules participate in proteinase-inhibitor interactions: crystal structures of Leu18, Ala18, and Gly18 variants of turkey ovomucoid inhibitor third domain complexed with *Streptomyces griseus* proteinase B. Protein Sci 1995;4(10):1985–1997.
36. Ko TP, Liao CC, Ku WY, Chak KF, Yuan HS. The crystal structure of the DNase domain of colicin E7 in complex with its inhibitor Im7 protein. Structure Fold Des 1999;7(1):91–102.
37. Kuhlmann UC, Pommer AJ, Moore GR, James R, Kleanthous C. Specificity in protein-protein interactions: the structural basis for dual recognition in endonuclease colicin-immunity protein complexes. J Mol Biol 2000;301(5):1163–1178.
38. Koradi R, Billeter M, Wuthrich K. MOLMOL: a program for display and analysis of macromolecular structures. J Mol Graph 1996;14(1):51–5, 29–32.