doi:10.1016/S0022-2836(02)01223-8 J. Mol. Biol. (2003) 325, 377–387

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Analysing Six Types of Protein–Protein Interfaces

Yanay Ofran[1,2*] and Burkhard Rost[1,3,4*]

[1]*CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168th Street BB217, New York, NY 10032, USA*

[2]*Department of Medical Informatics, Columbia University, 622 West 168th Street, New York, NY 10032 USA*

[3]*Columbia University Center for Computational Biology and Bioinformatics (C2B2), Russ Berrie Pavilion, 1150 St. Nicholas Avenue, New York NY 10032, USA*

[4]*North East Structural Genomics Consortium (NESG) Department of Biochemistry and Molecular Biophysics Columbia University, 650 West 168th Street BB217, New York NY 10032, USA*

**Corresponding authors*

Non-covalent residue side-chain interactions occur in many different types of proteins and facilitate many biological functions. Are these differences manifested in the sequence compositions and/or the residue–residue contact preferences of the interfaces? Previous studies analysed small data sets and gave contradictory answers. Here, we introduced a new data-mining method that yielded the largest high-resolution data set of interactions analysed. We introduced an information theory-based analysis method. On the basis of sequence features, we were able to differentiate six types of protein interfaces, each corresponding to a different functional or structural association between residues. Particularly, we found significant differences in amino acid composition and residue–residue preferences between interactions of residues within the same structural domain and between different domains, between permanent and transient interfaces, and between interactions associating homo-oligomers and hetero-oligomers. The differences between the six types were so substantial that, using amino acid composition alone, we could predict statistically to which of the six types of interfaces a pool of 1000 residues belongs at 63–100% accuracy. All interfaces differed significantly from the background of all residues in SWISS-PROT, from the group of surface residues, and from internal residues that were not involved in non-trivial interactions. Overall, our results suggest that the interface type could be predicted from sequence and that interface-type specific mean-field potentials may be adequate for certain applications.

Keywords: protein–protein interaction; protein complexes; protein interface; protein folding; drug design

Introduction

Do different types of interactions use different biochemical mechanisms?

Non-covalent contacts between residue side-chains are the basis for protein folding, protein assembly, and protein–protein interaction. These contacts occur under many different conditions, and facilitate a variety of interactions and associations within and between proteins. For example, residue–residue contacts determine protein structure by a myriad of interactions between residue side-chains. Non-covalent interactions between side-chains mediate the assembly of folded chains into multi-chain proteins. In these two instances, the interactions are permanent, in the sense that they typically last for the lifetime of a protein. However, non-covalent residue–residue interactions can be transient, as in receptor–ligand interaction or in signal transduction. These interactions typically last for only short times. Given the wide range of interfaces, one may hypothesise that different types of interactions are facilitated by different biochemical mechanisms.

Previous studies

Many studies have investigated whether the characteristics of interfaces differ between e.g. internal (within the same chain) and external (between different chains) interactions.[5–11]

Abbreviations used: DIP, database of interacting proteins[1]; PDB, Protein Data Bank of experimentally determined 3D structures of proteins[2,3]; SWISS-PROT, human curated data base of annotated protein sequences[4].

E-mail addresses of the corresponding authors: yo135@columbia.edu; rost@columbia.edu
http://cubic.bioc.columbia.edu/

Although all studies analysed proteins of known structure, their results were contradictory. Three theoretical, technical and computational problems may account for these differences. (1) In order to draw veritable conclusions from the available structural data it is necessary to analyse as many proteins as possible. However, none of the studies fully exploited the wealth of data available in the Protein Data Bank (PDB);[2,3] most analyses have been limited to relatively small, hand-selected data sets. One reason for analysing small data sets was that there is no simple way to distinguish automatically between (i) interfaces between two chains that belong to one multi-chain protein and (ii) interfaces between two different proteins. (2) Due to small datasets, most studies could not distinguish between homo-multimers and hetero-multimers or between permanent interactions and transient interactions. Instead, they had to focus on comparing internal interactions (within one chain) *versus* external (between chains) interactions. (3) Most studies have described external interactions through surface patches. Such surface patches may not capture all aspects of protein interactions. For example, slightly buried residues with long side-chains may be missed, although they participate in interfaces. Furthermore, analyses of residue mutations have indicated that the contribution to the free energy of binding is not distributed evenly across the interface.[12] Some residues identified as part of a surface patch may form important contacts, while others may not form contacts at all. Therefore, the analysis of surface patches may not capture all residue–residue contacts that underlie the interaction.

Different conclusions from analysing surface patches

Comparisons of protein interfaces have yielded contradictory results. Some studies report that the amino acid composition of different types of interfaces are similar;[7,8,13] others report significant differences.[6,10] Most studies are focused on comparing internal and external interfaces. A few studies distinguished external interfaces in more detail. For instance, Jones & Thornton[5] proposed a distinction between "obligatory" interactions, i.e. interfaces between chains that are in permanent contact (e.g. multi-chain proteins), and transient interactions, i.e. interfaces between separate proteins that interact only transiently to carry out a particular biological task (e.g. signal transduction or receptor–ligand binding). Unfortunately, such a detailed distinction of external interfaces reduced the available hand-selected data sets even further. Nevertheless, two groups suggested that the composition differs between internal, transient, and obligatory interfaces.[6,10] It may be suggested to surmount the problem of non-representative data sets by assuming that all homo-oligomers constitute permanent interactions and all hetero-oligomers constitute transient interactions. If so, we could classify the whole PDB automatically into transient and permanent oligomers. However, there are many examples of permanent hetero-oligomers and transient homo-oligomers. Furthermore, even if we accept this assumption, the literature still gives conflicting answers to the question of whether residue–residue preferences differ between homo-oligomers and hetero-oligomers.

We developed a simple data-mining method to analyse and sort structural data in a way that allows analysis of interfaces in very large data sets of high-resolution structures. In particular, we sorted the data into different groups of homo-oligomers *versus* hetero-oligomers and permanent interactions *versus* transient interactions. To our knowledge, this is the largest non-redundant data set of residue–residue contacts analysed thus far. We found significant differences in the sequence features between the following six types of interfaces: (1) intra-domain: interfaces within one structural domain; (2) domain–domain: interfaces between different domains within one chain; (3) homo-obligomer: interfaces between permanently interacting identical chains; (4) homo-complex: interfaces between transiently interacting identical protein chains; (5) hetero-obligomer: interfaces between permanently interacting different protein chains; (6) hetero-complex: interfaces between different transiently interacting protein chains.

We introduced the term "obligomer" to denote interfaces between residues from two chains that are "obligatory" in the sense introduced by Jones & Thornton.[5] In contrast, we refer to complexes as interfaces between transiently interacting chains. In the literature, all interfaces between different chains (hetero) are often referred to as protein–protein interactions. Note that, while results from experiments such as yeast two-hybrid systems[14,15] are usually thought to reflect generic protein–protein interactions, these experimental means may detect interfaces between identical chains (homo).[1,2]

Results and Discussion

Accurate automatic distinction between homo-interfaces and hetero-interfaces

Most PDB records that describe the structure of more than one chain do not specify whether the different chains belong to a single protein (interacting permanently), or to several proteins (interacting transiently). This data-mining problem has often been quoted as the reason for using small data sets and/or for the particular way in which external interfaces were distinguished.[5–8,10,11,16–20] Here, we propose an extremely simple solution: profit from the biological expertise that is at the heart of the SWISS-PROT database (see Methods).[4] This simple procedure reproduced correctly and automatically the small data sets that were hand-selected for previous publications;[5,6,10] i.e. we

Table 1. Data set statistics

Type of interface	Number of contacts
Internal	
Intra-domain	3,340,485
Domain–domain	255,144
External	
Homo-obligomers	218,104
Homo-complexes	3077
Hetero-obligomers	18,886
Hetero-complexes	166,412

Contacts: residues were defined as in contact if the separation between the two closest atoms was ≤6 Å. We separated the following six types of interfaces (see Methods for details). (1) Intra-domain: contacts between residues in the same structural domains (according to the domain definition of PrISM[43]). (2) Domain–domain: contacts between residues in different structural domains in the same chain. (3) Homo-obligomers: contacts between residues on two different chains that have identical sequence and are permanent in the sense that we have no evidence for any biological interaction of the monomer. (4) Homo-complexes: contacts between residues on two different chains that have identical sequence and are transient in the sense that another chain of that sequence is observed in the cell as functional monomers. (5) Hetero-obligomers: contacts between two non-identical chains from the same protein (transient). (6) Hetero-complexes: contacts between two non-identical chains from two different proteins (permanent).

found all the complexes identified in the literature to be classified as complexes by our simple assignment method. Moreover, the resulting data sets were more than one order of magnitude larger than data sets analysed in most previous studies (Table 1). Thus, we could analyse statistically significant sets even for a very fine-grained separation of six interface types.

One generic problem of bioinformatics is the lack of suitable statistical tools. Significance tests such as χ^2 are sensitive to sample size.[21] Consequently, applying these tests to very large data sets is prone to inferential errors, especially when the number of degrees of freedom is substantially low compared to the number of data points.[21] When we applied the standard χ^2 test to our data, we found that the differences between the amino acid compositions of the six interfaces were extremely significant statistically (unpublished results). However, even when we randomly reshuffled our data sets, χ^2 indicated significant differences between such nonsense splits. Therefore, we introduced a method that used the Jensen–Shannon information to explore the self-consistency of the data (find-self procedure; see Methods). This procedure revealed that the amino acid compositions differed significantly between the six interface types: samples of 1000 residues taken at random from each type of interface identified their own type correctly in over 63–100% of the cases (Table 2). Note that this did not imply that >63% of the individual interfaces were classified correctly, rather that the pool of contacts from each type of interface was consistent to a certain extent. Note, furthermore, that the absolute values of the percentages depend on the size of the samples drawn at each iteration (i.e. 1000; see Methods).

Differences in sequence on two levels: amino acid composition and contact preferences

In general, the concept of difference in sequence has two aspects: interfaces may differ in their amino acid composition and/or their residue–residue contact preferences. For example, complexes might have fewer negatively charged residues than obligomers; however, complexes might incorporate these fewer negative charges more often into salt-bridges. We investigated these two aspects (residue composition and residue–residue contact preferences) separately.

The six interface types differed significantly in their amino acid compositions

The find-self procedure unravelled the fact that there are at least six types of interface that differ significantly in their amino acid composition

Table 2. Significant difference between amino acid composition of six interface types

	Internal		External			
	Intra-domain	Domain–domain	Homo-obligomer	Homo-complex	Hetero-obligomer	Hetero-complex
Intra-domain	75.9	19.4	0.4	–	4.3	0.2
Domain–domain	18.6	62.7	0.9	–	16.4	1.4
Homo-obligomer	0.9	1.5	78.0	–	2.2	17.4
Homo-complex	–	–	0.2	99.8	–	–
Hetero-obligomer	3.9	17.2	1.9	–	70.8	6.2
Hetero-complex	0.1	1.2	16.3	–	6.3	76.1

Numbers indicate how often (in percentage points) the amino acid compositions of 1000 residues drawn at random from one interface type was most similar to a different set of 1000 residues drawn randomly from another interface (a dash indicates a value ≤0.1). For example, in 75.9% of the cases, the composition of the 1000 residues from intra-domain contacts were more similar to the composition of another 1000 residue sample from the same data set than to 1000 residue samples from any other interface type. All values on the diagonal reflect correct identification of the respective class; off-diagonal elements reflect the confusion. For instance, 19.4% of the misclassified intra-domain contacts were misclassified as domain–domain contacts. The symmetry of the table indicates a high level of consistency in the stochastic find-self procedure. The maximal standard deviation for each cell in the table was 4.4 percentage points. Note that the percentages should not be misread to mean retrieval of individual interfaces; rather, they refer to the retrieval of contacts from pools of 1000 contacts. The absolute numbers of the percentages obviously depend on the size of the randomly sampled pool (here 1000).

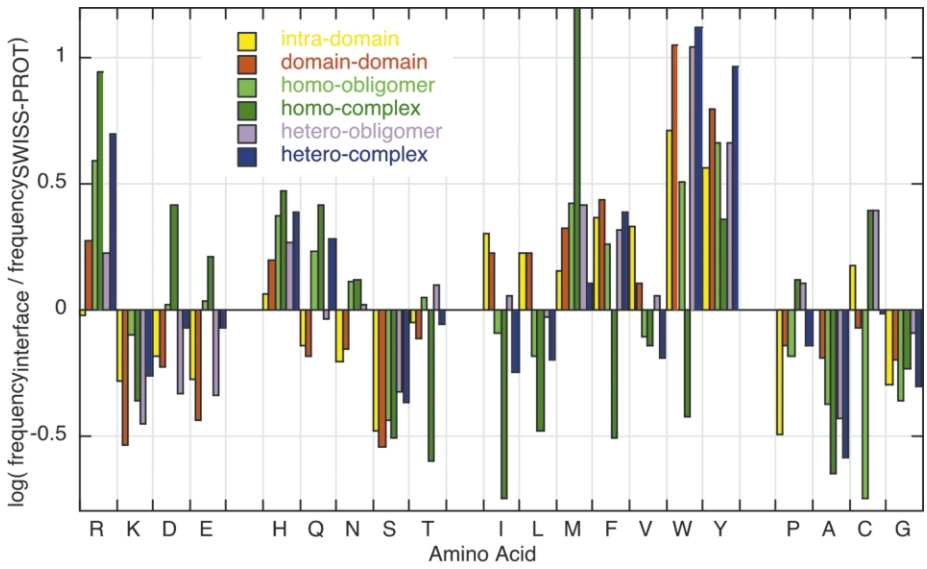

Figure 1. Amino acid composition of six interface types. The propensities of all residues found in SWISS-PROT were used as background. If the frequency of an amino acid is similar to its frequency in SWISS-PROT, the height of the bar is close to zero. Over-representation results in a positive bar, and under-representation results in a negative bar. The amino acid residues are identified by their one-letter code, sorted by biophysical features.

(Table 2: note that all values on the diagonal are substantially higher than the off-diagonal counts). In terms of the most coarse-grained separation, the sequence compositions differed most strongly between internal and external interfaces (Table 1S in Supplementary Material). The least-distinct types were (a) domain–domain interfaces that overlapped with intra-domain interfaces and (b) hetero-obligomers that resembled domain–domain interfaces remarkably often (off-diagonal elements in Table 2). The former was not surprising, since domain–domain interfaces are formed between residues on the same chain, and thus are likely to be similar to intra-domain contacts. On the other hand, we can view domain–domain interfaces as permanent interactions between independently folded units. Therefore, we may expect them to be similar to hetero-obligomers, which by definition, associate independently folded chains. Transient interfaces between identical chains (homo-complexes) constituted the seemingly most distinct interface type. The fact that our method could distinguish automatically between the interfaces of different chains of the same protein (hetero-obligomers) and different chains of different proteins (hetero-complexes) indicated strongly that the success of our data-mining approach in distinguishing complexes from obligomers was not restricted to the expert-curated data sets.

All interfaces differed in residue composition from background and surface

We used the residue compositions of all proteins in SWISS-PROT as the background to compare the compositions between the six interface types. We found that the composition of all interface types differed substantially from the composition of SWISS-PROT (Figure 1; Table 2S in Supplementary Material shows that this difference was highly significant). Nonetheless, our results showed why some studies report a strong correlation between the propensities of residues in internal and external interactions.[7,8] Most residues show similar trends in internal contacts and many other types of interfaces (Figure 1(b)). Indeed, we observed very strong correlations (equation (3)) between the amino acid residue distributions in all interface types ($r > 0.8$ in all pairwise comparisons) except for homo-complexes. Furthermore, all interface types were highly correlated to the distribution of amino acid residues in SWISS-PROT ($r > 0.8$). Nevertheless, the find-self procedure indicated substantial differences between these distributions, suggesting that the correlation coefficients were not sensitive enough for this comparison. All interfaces differed significantly from exposed residues (Table 3). Interestingly, about 1.3% of all the internal residues were found not to be in any non-trivial contact by our definition of contact. Most of these were at the ends of chains. When we added these "free" residues as a separate class, we found that they again differed significantly from all other classes.

Similarities and differences in and to the literature for composition

Some studies report substantial differences between internal and external interfaces,[22] while others report a high level of similarity.[8] Xu *et al.* conclude from the differences they found that internal contacts are facilitated by other than external mechanisms.[22] The substantial differences

Table 3. Significant difference in composition of six interfaces and surface residues

	Internal		External				
	Intra-domain	Domain–domain	Homo-obligomer	Homo-complex	Hetero-obligomer	Hetero-complex	Surface exposed
Intra-domain	75.7	19.6	0.3		4.4		
Domain–domain	16.7	64.5	1.6		16.6	0.6	
Homo obligomer	0.4	2.1	73.6		3.4	20.5	
Homo complex			0.1	99.9			
Hetero obligomer	3.2	15.2	3.4		73.1	5.1	
Hetero-complex		1.9	17.0		4.5	76.6	
Exposed							100.0

Same procedure as for Table 2; however, we included all the exposed residues in our data set as a separate category. Exposure was defined based on DSSP,[44] residues with a relative solvent accessibility[45] $\geq$16% were considered to be on the surface. Note: the maximal standard deviation for each cell was $<$5 percentage points.

we found between internal and external contacts support this view. Theoretical and experimental works attempted to identify the residues that play key roles in each type of interaction. A few groups found polar and charged residues as well as salt-bridges to be the major contributors for the formation of interactions.[11,17,22] Other studies reported that salt-bridges are not an important factor in protein–protein interaction,[9] or that interfaces favour non-polar residues.[23] The detailed separation of six types of interfaces explained these contradictions: while we identified some general trends, most of the residues showed different behaviours in different types of interfaces. In particular, we found no clear common denominator for charged residues: lysine was under-represented in all types of interfaces, while arginine was over-represented. Most large hydrophobic residues were favoured in all types of interactions (in particular, histidine, methionine, and tyrosine). In contrast, serine, alanine and glycine were under-represented. The other residues demonstrated different trends in different types of interfaces, yet, bio-physically similar residues, such as leucine and isoleucine, or aspartic acid and glutamic acid, usually showed similar trends, indicating the reliability of the data. Overall, the composition of homo-complexes was most exceptional in that it frequently differed from the trends of all other interface types. Jones & Thornton compared the propensities of residues in homo- and hetero-multimer interfaces.[5] They conclude that hydrophobic residues often are more abundant in homo-multimers than in hetero-multimers. When grouping all homo-multimeric and all hetero-multimeric interfaces, we found a similar trend. However, when we separated permanent interactions and transient interactions, this distinction disappeared. Jones *et al*. reported significant differences between the compositions of domain–domain interfaces and of the protein cores.[13] Their conclusion was based on a standard χ^2 test. To revisit this point, we checked whether our six data sets of contacts differed in composition from (a) SWISS-PROT as a whole (see Table 2S in Supplementary Material), (b) from exposed residues (Table 3) and (c) from residues that do not form any contact (see Table 3S in Supplementary Material). Our results indicated that each of these biophysical categories is characterised by unique residue compositions. In particular, we confirmed the earlier results,[5,6] that domain–domain interfaces differed from both the background and from intra-domain interfaces.

Analysing hot-spot residues

Bogan & Thorn reported hot-spots in binding energy for protein interfaces.[12] These spots are reported to be abundant in tryptophan, tyrosine and arginine, and depleted of serine, threonine, leucine and valine. Chakrabarti & Janin take an approach similar to that of Bogan & Thorn,[12] and differentiate between the core of the interface and its rim.[24] They find tryptophan and tyrosine to be over-represented in the core, and leucine and valine to be under-represented. While arginine appears abundant in the core, its propensity does not exceed the expected level on protein surfaces. Overall, our data confirmed these findings. However, when looking at the different types of interfaces, some exceptions were revealed. For example, tryptophan, which was extremely over-represented in most interface types, was under-represented in homo-complexes. Leucine, valine, and threonine showed different trends in different types of interfaces. In contrast, leucine, isoleucine and valine were remarkably similar in their preferences for all interfaces.

Residue contact preferences differed statistically

The residue–residue preferences differed remarkably between the six types of interfaces. When we repeated the find-self procedure for the set of preferences, the results were very similar to those obtained for composition (Table 4). Overall, the off-diagonal pattern was similar to that for residue composition (Table 2). The contact preferences were slightly more similar between the intra-domain and the domain–domain interfaces than

Table 4. Significant difference between contact preferences of six interface types

	Internal		External			
	Intra-domain	Domain–domain	Homo-obligomer	Homo-complex	Hetero-obligomers	Hetero-complex
Intra-domain	63.0	29.5	1.7	–	4.7	1.1
Domain–domain	21.8	61.2	0.9	–	12.6	3.5
Homo obligomer	1.5	1.6	79.6	–	1.9	15.4
Homo complex	–	–	–	100.0	–	–
Hetero obligomers	4.2	9.6	0.9	–	81.6	3.7
Hetero-complex	0.6	4.6	15.9	–	6.9	72.0

Same procedure as for Table 2; however, now the randomly chosen samples were 1000 contact preferences rather than 1000 amino acid residues. Note: the maximal standard deviation for each cell was <5 percentage points.

between domain–domain and hetero-obligomer interfaces. However, there was still a clear similarity between the latter two. For residue compositions, we noticed a trend to confuse homo-obligomers with hetero-complexes. For contact preferences, this trend was intensified: over 15% of the samples of homo-obligomer contacts were misidentified as hetero-complexes and *vice versa*. Grouping the two internal interface types into one, and the four external types into another class, we found that the residue contact preferences between these two classes differed substantially.

Homo-complexes depleted in salt-bridges and rich in contacts between identical residues

Hydrophobic–hydrophilic interactions dominated intra-domain, domain–domain and hetero-complex interfaces (Figures 2(A), 3(B) and (F); red squares indicate highly preferred interactions, blue squares indicate highly unlikely interactions). Cysteine bridges were observed more often than expected for all interface types. Similarly, salt-bridges were common with the exception of homo-complexes, for which they were observed less often than expected. Homo-complexes exhibited an extreme general preference for interactions between identical amino acid residues. Furthermore, overall the contact preferences also stood out most for homo-complexes.

Similarities and differences in and to the literature for contact preferences

The assessment of residue–residue preferences under different circumstances may be crucial for successful structure prediction, as well as for protein threading and drug design. Previous attempts to determine residue–residue preferences have yielded many scales and matrices.[7,8,17,25–27] Most of these matrices use the same set of preferences for all the different types of interactions or focus on internal interfaces. However, recently a few studies demonstrated the success of including data from external interactions to compile mean-field potentials for improving docking.[28–30] Bahar, Jernigan and colleagues present matrices of contact energies reflecting the attraction or repulsion between each residue pair.[17] Those matrices are given in RT units and hence are not fully comparable with our results, which are based on log odds. Yet, some interesting similarities and differences are noticeable. Bahar *et al*. find a high preference in the pairing between identical amino acid residues. This observation may appear rather odd, because an interaction between two identically charged residues appears to be energetically extremely unfavourable.[31,32] Our results appear to explain this oddity: homo-obligomers were the only type of interaction for which we observed strong preferences for interactions between identical amino acid residues (Figure 2(C)). This observation might be explained by the evolutionary advantage of favouring identical residue pairs in contacts between identical chains: while the conservation of non-identical contacts requires two neutral/beneficial point mutations, identical contacts need only one (Shoshana Wodak, Brussels, personal communication). For the other interface types, we observed strong self-interaction preferences exclusively for cysteine residues, which are known to stabilise interactions through forming cysteine-bridges (Figure 2). Confirming earlier findings, we found salt-bridges abundant in interfaces.[9,11,33] Based on amino acid composition, some studies hypothesise that hydrophobic interactions are more frequent in permanent interactions than in transient interactions.[5] This hypothesis is confirmed by our residue–residue preference data, both for homo-obligomers and for hetero-obligomers.

Can we predict interfaces from sequence?

Usually, interfaces have been defined structurally, i.e. according to the topography of the interacting macromolecules (surface patches). Jones & Thornton attempted to predict external interfaces from protein structures.[13,34] Ultimately, we pursue a different objective; namely, to predict protein–protein interactions directly from sequence. Hence, we had to replace the concept of external surface patches by that of sequence-consecutive interface segments (note that the data for the explicit analysis of segments are not shown). The similarities of our results to those obtained by some of the groups that analysed interface patches may or may not indicate that the two concepts are

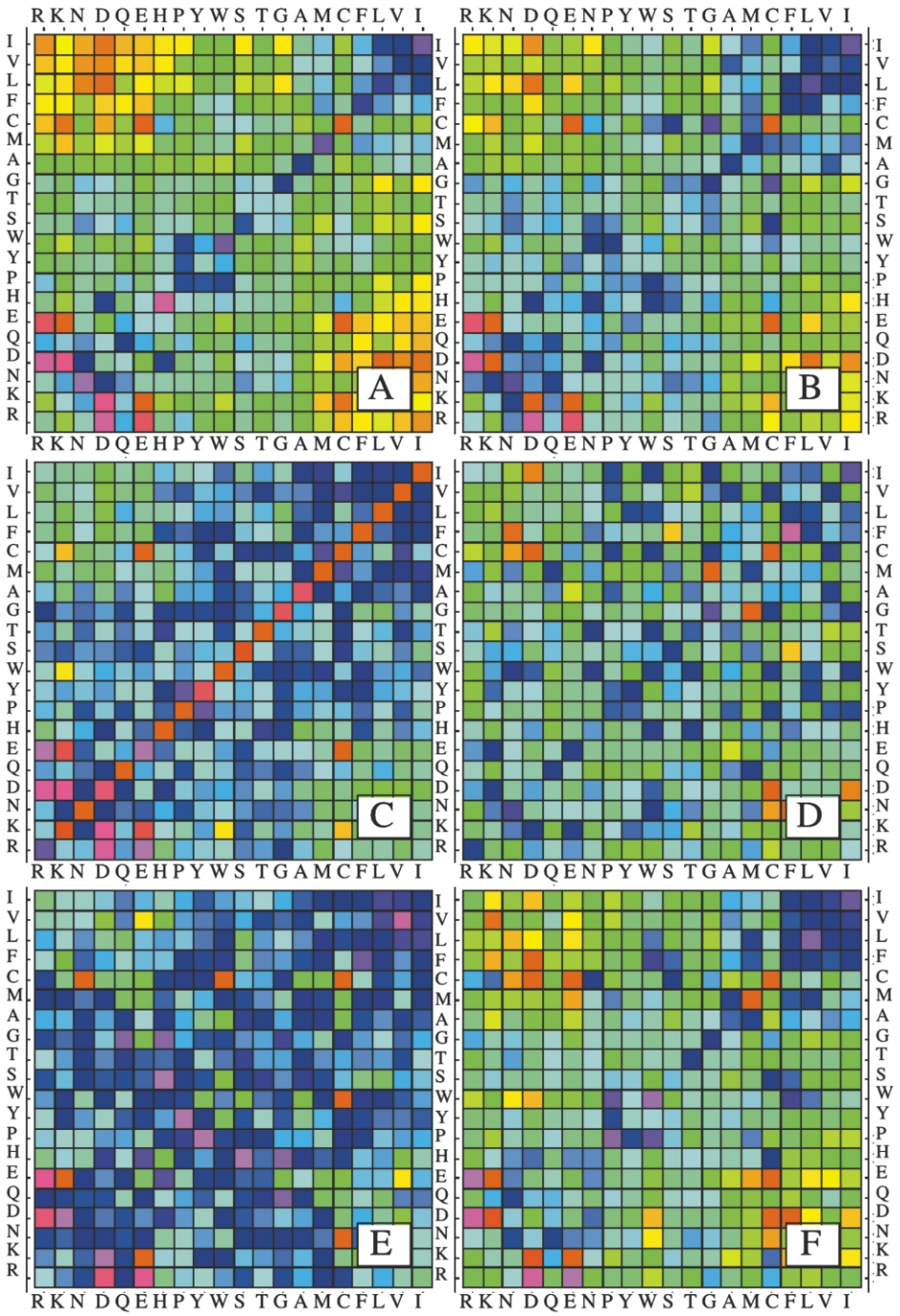

Figure 2. Residue–residue preferences. (A) Intra-domain, (B) domain–domain, (C) obligatory homo-oligomers (homo-obligomers), (D) transient homo-oligomers (homo-complexes), (E) obligatory hetero-oligomers (hetero-obligomers), and (F) transient hetero-oligomers (hetero-complexes). A red square indicates that the interaction occurs more frequently than expected; a blue square indicates that it occurs less frequently than expected. The amino acid residues are ordered according to hydrophobicity,[42] with isoleucine as the most hydrophobic and arginine as the least hydrophobic.

not that different, after all. Our direct analysis of sequence composition revealed significant differences between the types of interfaces, and between all interfaces and the background distribution. The method we used to explore statistical significance conceptually resembled a prediction method: we used the sequence-composition entropy of a pooled sample to "predict" its interface type. Obviously, the high level of success (between 62% and 100%) did not imply that we could predict individual interfaces at this level of precision. Nevertheless, our data may suggest the feasibility of such a prediction method. Even if this speculation turns out to be over-optimistic, our results still may have important impacts for methods that attempt to infer protein function from protein

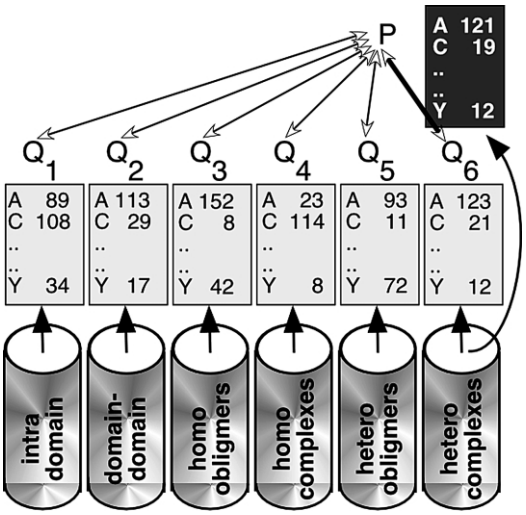

Figure 3. A sketch of the find-self procedure. We sampled 1000 contacts from one data set (*P*, here hetero-complexes) and then another 1000 from each of the six interface types (Q_1–Q_6). Then, we measured the divergence between the amino acid composition of *P* and each of the samples Q_1–Q_6. If the data set from which *P* was sampled had a unique and distinguishable composition, we expect that the sample most similar to *P* was Q_6. This process was repeated 1000 times for each type of interface (6000 total).

structure and/or attempt to predict aspects of protein structure and function.

Conclusions

Our study differed from previous analyses in four important ways. (1) We data-mined a set of interfaces from PDB that was, to our knowledge, far larger than data sets analysed before. (2) This large data set enabled us to base our analysis on a more finely grained distinction of interfaces than explored previously. In particular, we distinguished between two types of internal interactions (intra-domain, domain–domain) and between four types of external interactions (homo-obligomers, homo-complexes, hetero-obligomers, and hetero-complexes). (3) We analysed interface contacts rather than surface patches. (4) We established the statistical significance of differences through a rigorous information theory-derived procedure. These four novel components together yielded results that appeared to establish unambiguously that the six types of interfaces analysed differed in both their amino acid compositions and their residue-contact preferences. It was suggested in the past that there are many different types of interactions and that each of them is based on different biophysical mechanisms. The results of the find-self procedure may confirm this intuition. Thus, our data may be considered *a posteriori* as expected by many readers. However, it was encouraging how cleanly our algorithm distinguishing between multi-chain proteins and complexes of different proteins generated very consistent results. The success of this automatic method might eventually become the aspect of our work that will influence prediction methods most, as it allows creating large data sets of high resolution.

Methods

Generation of the data set

Today's PDB[2,3] is biased; and such bias can seriously impact statistical analyses.[35] To reduce the bias, we compiled the largest possible non-redundant subset of PDB: no pair of proteins in that set had more than 25% identity over 100 aligned residues.[36] The non-redundant set included 1812 high-resolution structures. We excluded NMR structures, theoretical models, and chains shorter than 30 residues. Of these proteins, 936 (51%) had resolutions below 2 Å, 74 proteins (4%) had resolutions above 3 Å. Our results did not change qualitatively when restricting the analysis to structures with higher resolution. We included all 1812 for the data shown in order to guarantee better statistics.

Elimination of packing complexes

When parsing PDB files, it is very hard to determine when a pair of chains is merely a packing multimer and when it is genuinely a biologically functional multimer. The problem is intensified when attempting to automatically parse hundreds of PDB files. Two approaches are suggested for coping with this problem. One is based on calculating the reduction of solvent accessibility due to oligomerisation[37] and the other is based on measuring the conservation of contacting residues.[38] We used the PQS server,[37] which applies the first method, to eliminate PDB files that appear to be packing complexes rather than biologically functional multimers.

Analysing interface contacts rather than interface patches

Typically, internal interfaces are defined in terms of contacting residues. External interfaces, however, are most often defined according to geometrically continuous patches of residues on the surface of a protein that exclude solvent by binding to another chain. The difference in definitions hampers the comparison between these two types of interactions. Furthermore, patch analysis might include some residues that are not really involved in the interactions (i.e. do not form inter-chain residue–residue contacts). They might exclude residues that play a key role in the interaction. We replaced the notion of patches by defining the interface in terms of contacting residues both for internal and for external interfaces. We defined a residue pair to be in contact if the distance between the closest of their respective atoms was ≤ 6 Å and their sequence separation was three or more residues. Note that this particular definition included contacts between β-sheets, while it ignored the contacts responsible for sharp β-turns. Note furthermore that the same residue may participate in different interfaces. The choice of the distance cut-off threshold that defines a contact is not straightforward. Previous studies used distances between 4 Å and 12 Å between two C^{α} or C^{β} atoms. However, the

variations in the sizes of side-chains might result in an under-representation of large residues in the data, as their side-chains themselves can extend several ångstrøm units. Hence, we defined contacts based on the distance between the closest pair of atoms of any two residues. This definition is more permissive than those used in other studies, thus classifying more residues to be in contact. However, it is not biased towards amino acid residues of any size. Thus, rather than biasing the data towards some residues, our permissive definition merely introduces "white noise". Using this definition, we parsed the set of 1812 PDB files to obtain all the contacting residues. Once we obtained the list of all pairs of contacting residues in these PDB files, we classified them into six types using the methods described below.

Homo-multimers *versus* hetero-multimers

Using simple sequence comparisons we differentiated between homo-multimers and hetero-multimers. Interactions between chains with more than 10% difference in sequence were defined as hetero-multimers. All other interactions were classified as homo-multimers.

Expert-driven automatic distinction between hetero-obligomers and hetero-complexes

A multimer can be permanent, i.e. all the functions of each of the chains can be carried out only in this multimeric state. Alternatively, it can be transient, i.e. one or more of the chains can be functional in different contexts and only in this particular multimer. Several studies have hypothesised that these two different types of interactions are based on different residue–residue contacts. However, it is hard to determine from the PDB file of a multimer which is the case. We introduced the following simple idea to achieve such a distinction automatically. SWISS-PROT files describe the sequence of a protein as it was studied in the laboratory. If the protein is studied in its multimeric state, then the sequence of all the chains will be submitted to SWISS-PROT in a single file. We hypothesised that experts typically add a new entry to the database if they study one of the chains by itself. That is, if there is experimental evidence identifying this chain as a separate functional protein. If this is true, we only have to map all chains in our data set to SWISS-PROT[4] and label chains by their respective SWISS-PROT identifiers. If non-identical chains from one PDB file appear in the same SWISS-PROT file, this indicates that there is no known situation in which they function separately in the cell. Hence, if we found two or more chains in the same SWISS-PROT file, we assumed that their association is obligatory (hetero-obligomer), otherwise we assumed that their association is transient (hetero-complexes). Following this logic, we divided our date set of hetero-multimers into two subsets. Contacts between chains in a PDB file that appear in the same SWISS-PROT file were classified as permanent interactions, or hetero-obligomers. Contacts between chains that appear in different SWISS-PROT files were classified as transient interactions, or hetero-complexes.

Database-driven distinction between homo-obligomers and homo-complexes

The same problem of differentiating permanent interactions from transient interactions exists also with homo-multimers. However, the cross-reference to SWISS-PROT is not applicable for homo-multimers. To distinguish between homo-multimers that are obligatory and those that are transient, we used the DIP database of interacting proteins.[1] We used DIP to detect those among our homo-multimers that, according to DIP, appear as functional monomers in the cell. As we obtained our data from PDB and PQS, we can assume that all the homo-multimers in our dataset appear in the cell as functional multimers. Therefore, those among them that are annotated by DIP to appear also as monomers should be classified as non-obligatory homo-multimers, or homo-complexes. All homomers that were not annotated as monomers in DIP were then classified as obligatory homomers, or homo-obligomers. Thus, we classified all the homo-multimer residue–residue contacts in our datasets to be either homo-complexes or homo-obligomers.

Establish statistical significance by find-self procedure based on Jensen–Shannon divergence

Most researchers are aware that standard significance tests are problematic when the data sets are small.[21] Another problem with these tests that is not commonly noted, is that the application of significance tests to very large data sets with a few degrees of freedom, like those analysed here, can lead to severe inferential mistakes (unpublished results).[21] Therefore, we introduced a simple information theory-based procedure (Figure 3) in order to answer the question of whether the amino acid compositions and contact preferences differed significantly between the six groups of interfaces. We considered two groups to differ if we could sort the interfaces correctly into their respective group using only its amino acid composition. Conceptually, the procedure resembles boot-strapping techniques.[39] Technically, it measured the Jensen–Shannon (JS) divergence[40] between random samples from each data set. For a pair of distributions p_1 and p_2, with prior probabilities π_1 and π_2, this measure is defined as:

$$\mathrm{JS}(p_1, p_2) = H(\pi_1 p_1 + \pi_2 p_1) - \pi_1 H(p_1) - \pi_2 H(p_2) \qquad (1)$$

with: π_1 and $\pi_2 \geq 0$, and $\pi_1 + \pi_2 = 1$ and $H(p) = -\Sigma_i p(x_i)\log_2(x_i)$, where π_1 and π_2 are the weights of the two probability distributions p_1 and p_2, respectively, and $H(x)$ is the Shannon entropy.[41]

The following procedure measured how often interfaces from one group were most similar in their amino acid composition to interfaces from the same or any other group.

1. Pick a random sample P with 1000 residues from one of the six sets of residue–residue contacts.
2. Pick six random samples Q_1–Q_6 with 1000 residues from each of the six sets.
3. Find the set Q_p from Q_1–Q_6 least divergent from set P by measuring the JS divergence between P and Q_1–Q_6.
4. Record the types of interactions from which P and Q_P were sampled.

We repeated this procedure 6000 times (1000 for each of the six types of interactions). If the residue composition differed significantly between the types, we expect Q_p to be, in most cases, the sample that was sampled from the same population of P. That is, we expect that in most cases, P and the sample most similar to P were sampled from contacts of the same interface type.

Measuring residue–residue preferences

After we had established that the amino acid composition differed between the six interface types, we used the six lists to compute the likelihood of forming contacts between each pair of amino acid residues. In particular, we compiled the log odds ratio of the observed frequency of the pair over its expected frequency:

$$L_x(i,j) = \log_2(P(i,j)/(P(i)P(j))) \quad (2)$$

where the subscript x represents one of the six types of interfaces (intra-domain, domain–domain, homo-obligomers, homo-complex, hetero-obligomers and hetero-complex), i and j are types of amino acid residue, $P(i,j)$ is the probability of a contact between amino acids of type i and j in interfaces of type x, and $P(i)$ and $P(j)$ are the probability of occurrence for amino acid residues i and j, respectively, in the interfaces of type x. Hence, the denominator describes the probability of a contact between i and j if the formation of contacts between i and j were random. On the basis of this equation, we generated six matrices for the likelihood for all possible contacts in each interface type.

Standard correlation

We applied the following standard correlation coefficient to compare our results to the literature:

$$r_{xy} = \frac{\sum_{i=1}^{20}(x_i - \langle x\rangle) \times (y_i - \langle y\rangle)}{\sqrt{\sum(x_i - \langle x\rangle)^2 \times \sum(y_i - \langle y\rangle)^2}} \quad (3)$$

where x and y are two data sets (e.g. internal SWISS-PROT), x_i and y_i are the propensities of amino acid i, and $\langle x\rangle$, $\langle y\rangle$ denote the mean over all 20 amino acids.

Acknowledgements

Thanks to Lukasz Salwinski (UCLA) and Ioannis Xenarios (UCLA, Lausanne) for their help in obtaining homo-complexes from DIP; thanks to Jinfeng Liu (Columbia) for computer assistance and Henry Bigelow (Columbia) for invaluable comments on the manuscript. We are grateful for the invaluable comments from two unknown referees, and from Shoshana Wodak (Brussels), and from Barry Honig (Columbia). This work was supported by grants 1-P50-GM62413-01 and RO1-GM63029-01 from the National Institute of Health. Last, not least, thanks to all those who deposit their experimental data in public databases, and to those who maintain these databases, in particular to Phil Bourne (UCSD), Amos Bairoch (Geneva), Rolf Apweiler (EBI) and their teams.

References

1. Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M. & Eisenberg, D. (2000). DIP: the database of interacting proteins. *Nucl. Acids Res.* **28**, 289–291.
2. Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Jr, Brice, M. D., Rodgers, J. R. *et al.* (1977). The Protein Data Bank. A computer-based archival file for macromolecular structures. *Eur. J. Biochem.* **80**, 319–324.
3. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al.* (2000). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235–242.
4. Bairoch, A. & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucl. Acids Res.* **28**, 45–48.
5. Jones, S. & Thornton, J. M. (1996). Principles of protein–protein interactions. *Proc. Natl Acad. Sci. USA*, **93**, 13–20.
6. Jones, S. & Thornton, J. M. (1997). Analysis of protein–protein interaction sites using surface patches. *J. Mol. Biol.* **272**, 121–132.
7. Keskin, O., Bahar, I., Badretdinov, A. Y., Ptitsyn, O. B. & Jernigan, R. L. (1998). Empirical solvent-mediated potentials hold for both intra-molecular and inter-molecular inter-residue interactions. *Protein Sci.* **7**, 2578–2586.
8. Glaser, F., Steinberg, D. M., Vakser, I. A. & Ben-Tal, N. (2001). Residue frequencies and pairing preferences at protein–protein interfaces. *Proteins: Struct. Funct. Genet.* **43**, 89–102.
9. McCoy, A. J., Chandana Epa, V. & Colman, P. M. (1997). Electrostatic complementarity at protein/protein interfaces. *J. Mol. Biol.* **268**, 570–584.
10. Lo Conte, L., Chothia, C. & Janin, J. (1999). The atomic structure of protein–protein recognition sites. *J. Mol. Biol.* **285**, 2177–2198.
11. Sheinerman, F. B., Norel, R. & Honig, B. (2000). Electrostatic aspects of protein–protein interactions. *Curr. Opin. Struct. Biol.* **10**, 153–159.
12. Bogan, A. A. & Thorn, K. S. (1998). Anatomy of hot spots in protein interfaces. *J. Mol. Biol.* **280**, 1–9.
13. Jones, S., Marin, A. & Thornton, J. M. (2000). Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. *Protein Eng.* **13**, 77–82.
14. Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. & Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.
15. Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R. *et al.* (2000). A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.
16. Jernigan, R. L. & Bahar, I. (1996). Structure-derived potentials and protein simulations. *J. Mol. Biol.* **6**, 195–209.
17. Bahar, I. & Jernigan, R. L. (1997). Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. *J. Mol. Biol.* **266**, 195–214.
18. Bahar, I. & Jernigan, R. L. (1996). Structure-derived potentials and protein simulations. *J. Mol. Biol.* **6**, 195–209.
19. Nayal, M., Hitz, B. C. & Honig, B. (1999). GRASS: a server for the graphical representation and analysis of structures. *Protein Sci.* **8**, 676–679.
20. Fetrow, J. S., Siew, N., Di Gennaro, J. A., Martinez-Yamout, M., Dyson, H. J. & Skolnick, J. (2001). Genomic-scale comparison of sequence- and structure-based methods of function prediction: does structure provide additional insight? *Protein Sci.* **10**, 1005–1014.

21. Royall, R. M. (1986). The effect of sample size on the meaning of significance tests. *The American Statistician*, **40**, 313–315.
22. Xu, D., Lin, S. L. & Nussinov, R. (1997). Protein binding *versus* protein folding: the role of hydrophilic bridges in protein associations. *J. Mol. Biol.* **265**, 68–84.
23. Zhou, H. X. & Shan, Y. (2001). Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins: Struct. Funct. Genet.* **44**, 336–343.
24. Chakrabarti, P. & Janin, J. (2002). Dissecting protein–protein recognition sites. *Proteins: Struct. Funct. Genet.* **47**, 334–343.
25. Hendlich, M., Lackner, P., Weitckus, S., Flöckner, H., Froschauer, R., Gottsbacher, K. *et al.* (1990). Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. *J. Mol. Biol.* **216**, 167–180.
26. Prlic, A., Domingues, F. S. & Sippl, M. J. (2000). Structure-derived substitution matrices for alignment of distantly related sequences. *Protein Eng.* **13**, 545–550.
27. Sippl, M. J. (1995). Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.* **5**, 229–235.
28. Moont, G., Gabb, H. A. & Sternberg, M. J. (1999). Use of pair potentials across protein interfaces in screening predicted docked complexes. *Proteins: Struct. Funct. Genet.* **35**, 364–373.
29. Aloy, P. & Russell, R. B. (2002). Interrogating protein interaction networks through structural biology. *Proc. Natl Acad. Sci. USA*, **99**, 5896–5901.
30. Aloy, P., Querol, E., Aviles, F. X. & Sternberg, M. J. (2001). Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J. Mol. Biol.* **311**, 395–408.
31. Polticelli, F., Ascenzi, P., Bolognesi, M. & Honig, B. (1999). Structural determinants of trypsin affinity and specificity for cationic inhibitors. *Protein Sci.* **8**, 2621–2629.
32. Schueler, O. & Margalit, H. (1995). Conservation of salt bridges in protein families. *J. Mol. Biol.* **248**, 125–135.
33. Honig, B. & Nicholls, A. (1995). Classical electrostatics in biology and chemistry. *Science*, **268**, 1144–1149.
34. Jones, S. & Thornton, J. M. (1997). Prediction of protein–protein interaction sites using patch analysis. *J. Mol. Biol.* **272**, 133–143.
35. Rost, B. (2002). Enzyme function less conserved than anticipated. *J. Mol. Biol.* **318**, 595–608.
36. Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Eng.* **12**, 85–94.
37. Henrick, K. & Thornton, J. M. (1998). PQS: a protein quaternary structure file server. *Trends Biochem. Sci.* **23**, 358–361.
38. Elcock, A. H. & McCammon, J. A. (2001). Identification of protein oligomerization states by analysis of interface conservation. *Proc. Natl Acad. Sci. USA*, **98**, 2990–2994.
39. Efron, B., Halloran, E. & Holmes, S. (1996). Bootstrap confidence levels for phylogenetic trees. *Proc. Natl Acad. Sci. USA*, **93**, 13429–13434.
40. Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Trans. Inform. Theory,* **37**, 145–151.
41. Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Tech. J.* **27**, 379–423; see also pp. 623–656.
42. Kyte, J. & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132.
43. Yang, A. S. & Honig, B. (2000). An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. *J. Mol. Biol.* **301**, 665–678.
44. Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
45. Rost, B. & Sander, C. (1994). Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins: Struct. Funct. Genet.* **19**, 55–72.

Edited by J. Thornton

(Received 25 July 2002; received in revised form 22 October 2002; accepted 25 October 2002)

Supplementary Material comprising three Tables is available