Protein J (2008) 27:59–70
DOI 10.1007/s10930-007-9108-x

Characterization of Protein–Protein Interfaces

Changhui Yan · Feihong Wu · Robert L. Jernigan · Drena Dobbs · Vasant Honavar

Published online: 13 September 2007

Abstract We analyze the characteristics of protein–protein interfaces using the largest datasets available from the Protein Data Bank (PDB). We start with a comparison of interfaces with protein cores and non-interface surfaces. The results show that interfaces differ from protein cores and non-interface surfaces in residue composition, sequence entropy, and secondary structure. Since interfaces, protein cores, and non-interface surfaces have different solvent accessibilities, it is important to investigate whether the observed differences are due to the differences in solvent accessibility or differences in functionality. We separate out the effect of solvent accessibility by comparing interfaces with a set of residues having the same solvent accessibility as the interfaces. This strategy reveals residue distribution propensities that are not observable by comparing interfaces with protein cores and non-interface surfaces. Our conclusions are that there are larger numbers of hydrophobic residues, particularly aromatic residues, in interfaces, and the interactions apparently favored in interfaces include the opposite charge pairs and hydrophobic pairs. Surprisingly, Pro-Trp pairs are over represented in interfaces, presumably because of favorable geometries. The analysis is repeated using three datasets having different constraints on sequence similarity and structure quality. Consistent results are obtained across these datasets. We have also investigated separately the characteristics of heteromeric interfaces and homomeric interfaces.

Keywords Heteromeric interfaces · Homomeric interfaces · Residue composition · Interface propensities · Contact preferences

Abbreviations

PDB	Protein Data Bank
PQS	Protein Quaternary Structure
ΔASA	Changes in solvent accessible surface area
rASA	Relative solvent accessibility
RIP	Raw interface propensity
NIP	Normalized interface propensity

1 Introduction

Protein–protein interactions play crucial roles in many biological functions. Elucidating the mechanisms of the

C. Yan (✉)
Department of Computer Science, Utah State University, 4205 Old Main Hill, Logan, UT 84341, USA
e-mail: cyan@cc.usu.edu

F. Wu · R. L. Jernigan · D. Dobbs · V. Honavar
Bioinformatics and Computational Biology Graduate Program, Iowa State University, Ames, IA 50010, USA

F. Wu · R. L. Jernigan · D. Dobbs · V. Honavar
Center for Computational Intelligence, Learning, and Discovery, Iowa State University, Ames, IA 50010, USA

F. Wu · V. Honavar
Artificial Intelligence Research Laboratory, Department of Computer Science, Iowa State University, Ames, IA 50010, USA

R. L. Jernigan
Department of Biochemistry, Biophysics, and Molecular Biology, Iowa State University, Ames, IA 50010, USA

R. L. Jernigan · D. Dobbs · V. Honavar
Laurence H Baker Center for Bioinformatics and Biological Statistics, Iowa State University, Ames, IA 50010, USA

D. Dobbs
Department of Genetics, Development and Cell Biology, Iowa State University, Ames, IA 50010, USA

interactions presents a challenge in molecular biology. One general approach to study the interaction between two proteins is to obtain a crystal structure of the protein–protein complex and then investigate the atomic properties of the protein–protein interface. Many studies have analyzed the characteristics of protein–protein interfaces in an effort to search for the factors that contribute to the affinity and specificity of protein–protein interactions [1–5]. These analyses show that the two surfaces of a protein–protein interface usually show high degrees of geometric and chemical complementarities. Electrostatic forces are also believed to play an important role in protein–protein interactions [6–8]. Several studies have shown that interfaces are biased in residue composition and inter-residue contacts [9, 10]. Miyazawa and Jernigan [11] developed a method to extract inter-residue potentials from frequencies of contacts between different residues in proteins. Later, Keskin et al. [12] showed that the potentials of mean force for inter-residue interactions hold for both intra-molecular and inter-molecular interactions. The important role of hydrophobic forces in protein–protein interactions has been confirmed by several researchers [13, 14]. However, a recent study [15] argues that it is the hydrophilic rather than the hydrophobic effect that makes the major contribution to protein–protein association. Another well-characterized property of interfaces is the existence of ''hot-spot'' residues, which are residues that make the largest contributions to complex formation [16].

Some studies divided the protein–protein interfaces into several subtypes and analyzed the characteristics of each subtype. Jones and Thornton [17] proposed a distinction between obligatory interactions and transient interactions. Using machine-learning methods, Block et al. [18] were able to extract physicochemical properties that are predictive of obligatory and transient interactions. Ofran and Rost [10] divided protein–protein interfaces into six types: intra-domain, domain–domain, homo-obligomer, hetero-obligomer, homo-complex, and hetero-complex. Chakrabarti and Janin [19] dissected the interfaces into a core and a rim based on solvent accessibility. Cho et al. [20] show that different functional types of protein–protein interactions have different molecular interactions specific to them.

We extracted all protein–protein interfaces from the Protein Data Bank (PDB) [21] and obtained three datasets that are much larger than any other dataset used in previous studies. Each protein was divided into three disjoint groups: interface, protein core, and non-interface surface. Comparisons show that the three groups are significantly different in residue composition, sequence entropy, and secondary structure. Since interfaces, protein cores, and non-interface surfaces have different solvent accessibilities, it is not known whether these differences are due to the differences in solvent accessibility or differences in functionality. To exclude the effect of solvent accessibility, we compared the interfaces with a set of residues that was randomly chosen from the overall residues and had the same solvent accessibility as the interfaces. The results show a clear trend that hydrophobic residues and aromatic residues are more frequent in the interfaces and hydrophilic residues are less common. Note that this trend cannot been found by comparing interfaces with protein cores and non-interface surfaces. We repeat the analysis using the three datasets and consistent results were obtained. We divided the interfaces into heteromeric interfaces and homomeric interfaces based on the similarities of the interacting chains. Comparisons show significant differences between the two types of interfaces in residue composition, sequence entropy, secondary structure, size, and contact preferences.

2 Materials and Methods

2.1 Selecting Structures for Dataset100, Dataset30, and Dataset30_3

All protein complexes in the PDB with at least two protein chains having at least 50 amino acids in each chain were obtained. We tried different thresholds of minimum length ranging from 20 to 100 amino acids. No obvious differences in interface characteristics were observed. In order to eliminate crystal packing, PDB complexes were split into individual quaternary structures based on the Protein Quaternary Structure (PQS) database [22]. In the construction of PQS database, a procedure was used to discriminate crystal packing and biological interfaces based on buried area, number of buried residue, a delta-solvation energy of folding, number of salt bridges at the interface and the presence of disulphide bridge. Then within each quaternary structure, a pair of protein chains is considered as interacting if the buried area on one chain is at least 200 $Å^2$. The same threshold of buried area was used in the SPIN-PP database (http://honiglab.cpmc.columbia.edu/SPIN/intro.html). A minimum buried area of 400 $Å^2$ on one chain has been used in some studies to define biological interfaces (reviewed in [3]). In this study, we also tried a minimum buried area of 400 $Å^2$. The only difference observed is that in the distribution of interface sizes, fewer interfaces have small sizes, since some small interfaces have been removed. No obvious differences in other properties were observed. The buried area was computed using NACCESS [23, 24]. A dataset of interfaces was thus obtained from the set of quaternary structures. Then, sequence similarity information was obtained from the sequence clusters provided by the PDB (ftp://ftp.rcsb.org/pub/pdb/derived_data/NR/). The similarity between two interfaces is defined as the highest sequence

similarity between the protein chains of the interfaces. First, redundant data were removed so that there were no identical interfaces in the dataset. The resulting dataset consists of 6,545 interfaces. This dataset is referred to as *Dataset100*, with 100 indicating that the similarity between any two pairs is below 100%. Interfaces with high similarity were removed from *Dataset100* so that the similarity between any two interfaces is below 30%. The resulting dataset (referred to as *Dataset30*) has 2,557 pairs of interacting chains. Then, all the structures having resolution >3 Å were removed from Dataset30. The resulting dataset (referred to as *Dataset30_3*) consists of 2,310 pairs of interacting chains.

2.2 Protein Cores, Interfaces, and Non-interface Surfaces

We defined residue contacts as described in Ofran and Rost [10]: two residues are in contact if the distance between them is less than 6 Å. Interface residues of a protein are the residues that contact with residues from the interacting proteins. Protein core residues are the non-interface residues whose relative solvent accessibility (rASA) is less than 25%. Non-interface surface residues are the non-interface residues whose rASA is at least 25%. The rASA of residues was calculated using the NACCESS program [23, 24]. As all the other studies, interface residues are defined based on the known interaction surfaces on PDB complexes. Some non-interface residues obtained may act as interface residues in other yet unknown interactions. To evaluate the effect of this on the analysis results, the complete knowledge of interaction sites on proteins must be known. Unfortunately, the data we have today are far from complete.

2.3 Heteromeric Interfaces and Homomeric Interfaces

An interface is a homomeric interface if the two interacting chains have a sequence identity greater than 95% and otherwise, it is a heteromeric interface. We used *Dataset100* to compare the properties of heteromeric interfaces and homomeric interfaces. *Dataset100* contains 3,990 homomeric interfaces and 2,555 heteromeric interfaces.

2.4 Interface Propensity (Raw Interface Propensity, RIP)

Let F_i be the number of residues of type i in the dataset, and f_i be the number of residues of type i in the interfaces, $w_i = f_i/\sum_m f_m$, and $W_i = F_i/\sum_m F_m$. The *interface propensity* of residue i is given by $\log_2 (w_i/W_i)$. A residue's propensities for protein cores and non-interface surfaces are computed with w_i replaced by the fractions of residue type i in the protein cores and non-interface surfaces, respectively.

2.5 Normalized Interface Propensity (NIP)

Residues are randomly extracted from the overall residues so that the resulting set had the same relative solvent accessibility (rASA) distribution as the interface residues. The resulting set will be referred to as *SetrASA*, with rASA denoting that the dataset has the same rASA distribution as the interfaces. Let s_i be the number of residues of type i in the *SetrASA*, and $S_i = s_i/\sum_m s_m$. The *normalized interface propensity* of residue type i is given by $\log_2 (w_i/S_i)$, where w_i is defined as above.

2.6 Contact Preferences

Let C_{ij} be the number of interface-crossing contacts formed by residues of types i and j. The *raw contact frequency* between residues of types i and j is given by $(C_{ij}/\sum_{m,n} C_{mn})$. The *contact preference* between residues of types i and j is given by $\log_2 \left(\left(C_{ij}/\sum_{m,n} C_{mn} \right) \Big/ (w_i \times w_j) \right)$, where w_i and w_j are defined as above. Note that contact preference is given by the logarithm of raw contact frequency divided by the frequencies of residue types i and j.

3 Results

3.1 Characteristics of Interfaces

Each protein is divided into three disjoint groups: protein core, interface, and non-interface surface. Interface properties including residue composition, secondary structure, sequence entropy, contact preferences, and size are analyzed using *Dataset100*.

3.1.1 Residue Composition

Figure 1A compares the residue compositions of protein cores, interfaces, and non-interface surfaces. Residues are placed in the order of increasing hydrophobicity based on the Kyte and Doolittle hydropathy index [25]. The comparisons show that among the three groups, protein cores

have the highest fractions of hydrophobic residues (e.g., Met, Cys, Phe, Ile, Leu, and Val) and non-interface surfaces have the least. This indicates that hydrophobic residues are preferred in protein cores and disfavored for non-interface surfaces. The opposite trend is observed for hydrophilic residues (e.g., Arg, Lys, Glu, and Asp). Figure 1B shows that all residue types have opposite propensities for protein cores and non-interface surfaces, and with His, Tyr, and Gly being notable exceptions, the propensities for interfaces are intermediate between those for protein cores and non-interface surfaces.

3.1.2 Sequence Entropy

Sequence entropy values for residues are extracted from the HSSP database (http://www.cmbi.kun.nl/gv/hssp/). The sequence entropy shows the conservation at each residue position. It is normalized over the range of 0–100, with the lowest sequence entropy values corresponding to the most conserved positions. Figure 2 compares the sequence entropy distributions of protein cores, interfaces, and non-interface surfaces. The comparisons show that among the three groups, protein cores have the highest fraction of residues in the low sequence entropy region (sequence entropy <40), and non-interface surfaces have the least. In the high sequence entropy region (sequence entropy ≥40), the opposite trend is observed. Let $A \gg B$ denotes A is more conserved than B. The results indicate that the trend of conservation is *protein core residues* » *interface residues* » *non-interface surface residues*. In a study based on a small set of transient protein–protein complexes, Nooren et al. [26] showed that interface residues are more conserved than surface residues. Here, consistent results are obtained for a larger dataset.

3.1.3 Secondary Structure

We consider eight classes of secondary structure as defined by the DSSP program [27]. Figure 3 compares the secondary structure composition of protein cores, interfaces, and non-interface surfaces. The comparisons show that among the three groups, non-interface surfaces have the highest fraction of residues in S (Bend) and T (Turn), the protein cores have the smallest, and interfaces are intermediate. The opposite trend is observed for the class E (Extended strand). No obvious location preferences are observed for the other classes of secondary structure.

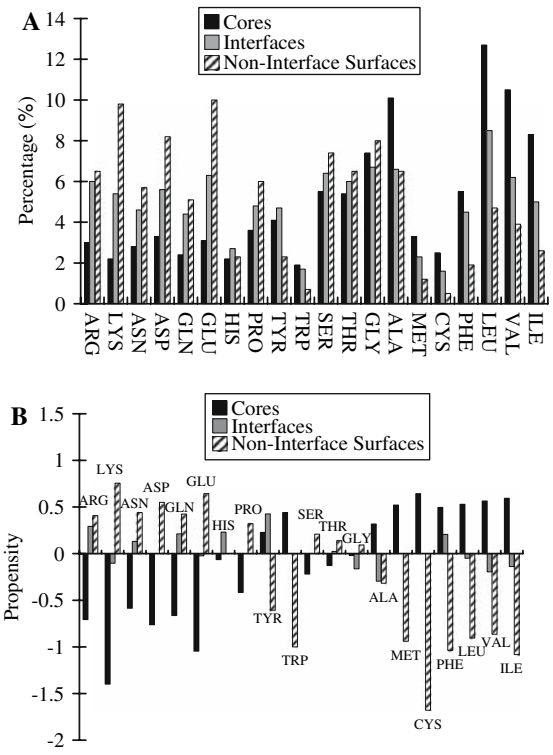

Fig. 1 Residue composition and residue propensities for different locations. (**A**) Residue compositions of protein cores, interfaces, and non-interface surfaces. (**B**) Residue propensities for protein cores, interfaces, and non-interface surfaces. Residues are ordered by their increasing hydrophobicity based on the Kyte and Doolittle hydropathy index [25]. The results are shown for *Dataset100*. The figures show that hydrophobic residues are more frequent in protein cores and less common on non-interface surfaces. The opposite trend is observed for hydrophilic residues. Residue propensities for interfaces are intermediate between those for protein cores and non-interface surfaces, with His, Tyr, and Gly being notable exceptions

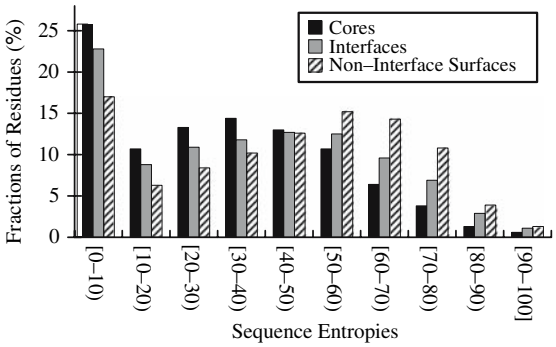

Fig. 2 Sequence entropies in protein cores, interfaces, and non-interface surfaces. Sequence entropy values for residues are extracted from the HSSP database (http://www.cmbi.kun.nl/gv/hssp/). The sequence entropy shows the conservation at each residue position in a multiple alignment. The values have been normalized over the range of 0–100, with the lowest sequence entropy values corresponding to the most conserved positions. The results are for *Dataset100*. The figure shows that among the three groups, protein cores have the highest fraction of residues with high conservation (less entropy values), non-interface surfaces have the smallest, and interfaces are intermediate

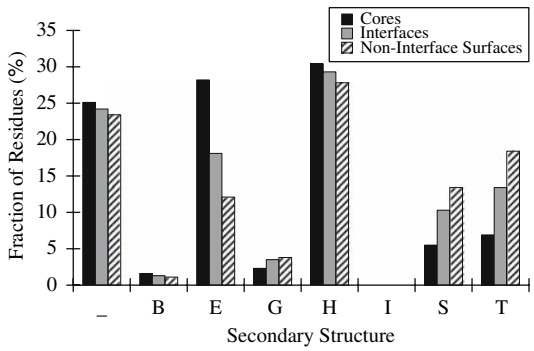

Fig. 3 Secondary structure compositions of protein cores, interfaces, and non-interface surfaces. Secondary structures of proteins are defined using the DSSP program [27]: 3_{10}-helix (G), alpha helix (H), pi helix (I), helix-turn (T), extended beta sheet (E), beta bridge (B), bend (S), and other/loop (_). Each protein is divided into interface, protein core, and non-interface surface based on solvent accessibility and whether a residue is in the interface as described in Sect. 2. The results are achieved using *Dataset100*

3.1.4 Contact Preferences

Figure 4A shows the *contact frequencies* across the interfaces given by $(C_{ij}/\sum_{m,n} C_{mn})$, where C_{ij} is the number of contacts between residues of types i and j. Figure 4B shows the *contact preferences* given by $\log_2 \left(\left(C_{ij}/\sum_{m,n} C_{mn} \right) \Big/ (w_i \times w_j) \right)$, where w_i and w_j are the frequencies of residue types i and j. In Fig. 4B, positive preferences are shown in red, negative in blue, and neutral in green. Residues are placed in order by increasing hydrophobicity. Comparison of Fig. 4A and B shows that normalizing the raw contact frequencies by the frequencies of individual residue types makes the high preferences for hydrophobic contacts, aromatic contacts and the contacts between oppositely charged residues stand out clearly (red in Fig. 4B). Figure 4B shows that the contacts between hydrophobic residues are preferred in interfaces. These highly preferred contacts correspond to the red region in the lower-right corner of Fig. 4B. The fact that Cys–Cys contacts have one of the highest preferences indicates the important role that this type of contacts has in protein–protein interactions. The contacts between residues with opposite charges (Arg–Asp, Arg–Glu, Lys–Asp, and Lys–Glu) are also preferred in interfaces. These contacts form several red entries near the upper-left corner in Fig. 4B. These results are consistent with the previous claim that disulfide bonds, salt-bridges, and hydrophobic interactions represent the main forces in protein–protein interactions [6, 9, 10, 28]. The face-to-face arrangement of two aromatic rings was reported to be favorable for interactions [9]. Here, high preferences for the contacts between different aromatic residues are observed. The interaction between a proline ring and an aromatic ring resembles the interaction between two aromatic rings [9], and this can be seen in the higher preference for the Pro–Trp (P-W) pair. Keskin et al. [12] investigated the residue contacts at protein–protein interfaces using "solvent-mediated" potentials and "residue-mediated" potentials. The abundance of the Cys–Cys contact, hydrophobic contacts, and aromatic contacts in interfaces observed in this study are consistent with the low values of the residue-mediated potentials for these contacts reported by Keskin et al. [12].

3.1.5 Interface Size

Interface size is calculated separately for each side of an interface. Figure 5 shows that interface sizes span a broad range and that the distribution has a peak in the range of 600–800 Å^2. The average interface size is 1,227 Å^2. Fourteen percent of the interfaces in the dataset have a size in the range of 600–800 Å^2. In a study based on a set of 75 hetero-complexes, Lo Conte et al. [29] found that most

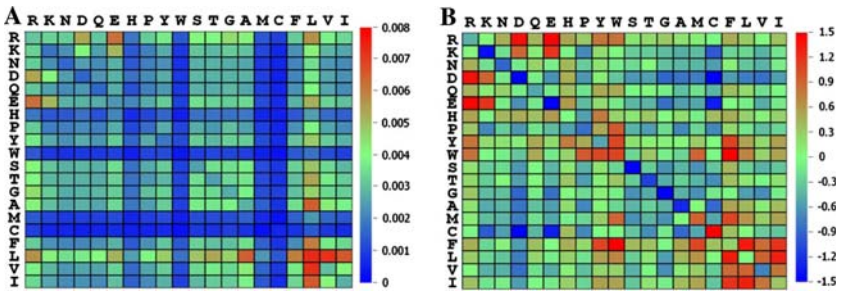

Fig. 4 Residue contact preferences for interfaces. (**A**) Raw contact frequencies given by $(C_{ij}/\sum_{m,n} C_{mn})$, where C_{ij} is the number of contacts between residue types i and j. (**B**) Contact preferences given by $\log_2 \left(\left(C_{ij}/\sum_{m,n} C_{mn} \right) \Big/ (w_i \times w_j) \right)$. The results are given for *Dataset100*. Residues are placed in order by their increasing hydrophobicity based on the Kyte and Doolittle hydropathy index [25]. Figure **B** shows that Cys–Cys contacts, the contacts between residues with opposite charges, the contacts between different aromatic residues, and those between hydrophobic residues are preferred in interfaces. These contacts are shown in red in Figure **B**. Comparison between **A** and **B** shows that normalizing raw contact frequencies by the frequencies of individual residue types makes the preferences for these contacts stand out more clearly

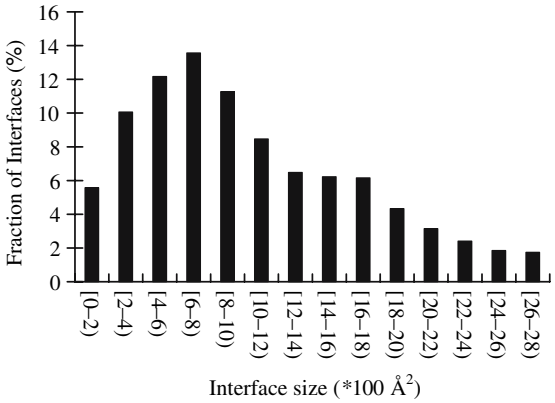

Fig. 5 Interface size distribution. Interface size is calculated separately for each side of an interface. The results are obtained for *Dastaset100*. The distribution has a peak at 600–800 Å^2. About 25% of the interfaces have a (one-sided) size in the range of 800 (±200) Å^2

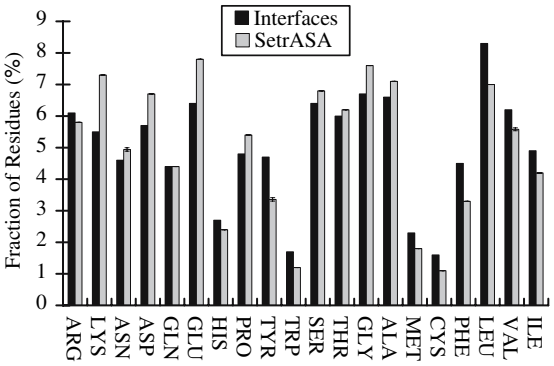

Fig. 6 Comparison of residue compositions of the *SetrASA* and at interfaces. Five *SetrASAs* are extracted from *Dataset100*. Mean values for the *SetrASAs* are displayed. The standard deviations are below 0.05 (They are shown as bars in the figure but too small to be visible). The residue types are placed in order by their increasing hydrophobicities

interfaces have a total buried area (that is, the sum buried area from both sides of the interfaces) in the range of 1,600 (±400) Å^2, which is roughly equivalent to 800 (±200) Å^2 for each side of the interface. Here, about 25% of the interfaces have a (one-sided) size in the range of 800 (±200) Å^2.

3.2 Are the Differences in Residue Composition, Conservation, and Secondary Structure Due to the Difference in Solvent Accessibility or the Difference in Functionality?

By our definition, protein core residues have a relative solvent accessibility (rASA) below 25%, non-interface surface residues have a rASA equal to or greater than 25%, and interface residues have a rASA ranging from 0% to 100%. The results from above have shown the differences in residue composition, conservation, and secondary structure among protein cores, interfaces, and non-interface surfaces. However, since these three groups have different accessibilities, it is unknown whether these differences are due to the differences in solvent accessibility or other reasons. To separate out the effect of solvent accessibility, we randomly extract residues from the overall residues so that the resulting residue set has the same rASA distribution as the interfaces. The resulting dataset will be referred to as *SetrASA*, with rASA denoting that the dataset has the same rASA distribution as interfaces. We then compare the interfaces with the *SetrASA*. Five different *SetrASAs* were independently extracted from the *Dataset100*. The size of each *SetrASA* is about 60% of that of the overall residues.

3.2.1 Residue Composition and Interface Propensity

Figure 6 compares the residue compositions of the *SetrASAs* and the interfaces. The comparisons show that the interfaces have more aromatic residues (Tyr, Trp, and Phe) and hydrophobic residues (Cys, Met, Ile, Leu, and Val) than do the *SetrASAs*. Residues with intermediate hydrophobicity (Ser, Thr, Gly, and Ala) are underrepresented in the interfaces. All charged residues, except Arg, are underrepresented in the interfaces.

We calculate the interface propensities (Normalized interface propensities, *NIP*) of residues by comparing the residue composition of the interfaces with that of the *SetrASA*, that is, *propensity* $(i) = \log_2 (w_i/S_i)$, where S_i is the fraction of residue i in the *SetrASA*s and w_i is the fraction of residue i in the interfaces. We name this propensity *normalized interface propensity (NIP)*, since the *SetrASA* can be considered as a version of the overall residues that is normalized according to the rASA distribution of the interfaces. The results are shown in Fig. 7 with residue types placed in order by their increasing hydrophobicity. Figure 7 shows that *NIP* reveals that interfaces have high preferences for hydrophobic residues and hydrophilic residues are not preferred at interfaces. On the right side (the hydrophobic end) of Fig. 7, residues have high propensities for interfaces and Cys has the highest propensity overall. On the left side (the hydrophilic end), residues (except Arg and His) have negative propensities. This indicates that the interfaces are more hydrophobic than expected based on their exposure. Figure 7 also shows aromatic residues to have high propensities for interfaces.

We compare *NIP* with the interface propensities (*raw interface propensities, RIP*) that are calculated by comparing the interfaces with all residues, which is given by

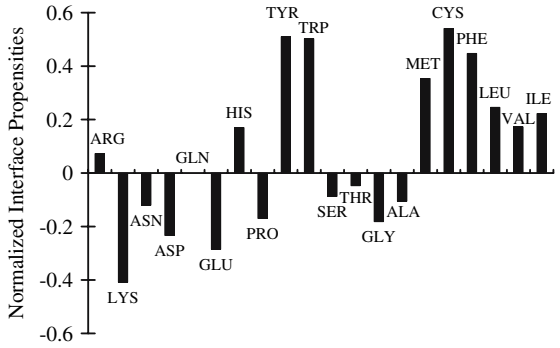

Fig. 7 Normalized interface propensities (*NIP*) of residues. The propensities are calculated by comparing interfaces with the sets (*SetrASA*) of residues that have the same relative solvent accessibility as the interfaces. Five *SetrASAs* were extracted, and mean values are displayed. The standard deviations are below 0.02 (They are shown as bars in the figure, but most of them are too small to be visible). The results show the clear trend that hydrophobic residues are preferred in interfaces and hydrophilic residues are not. Aromatic residues also have high *NIP*. The results are obtained using *Dataset100*

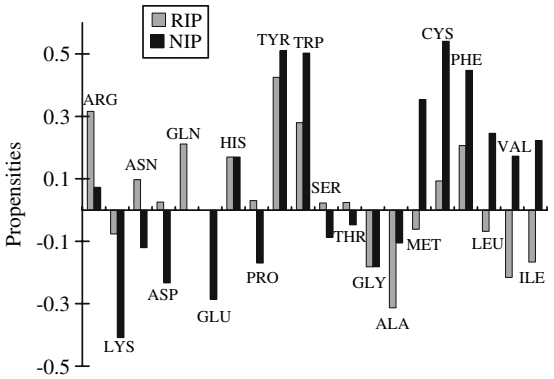

Fig. 8 Comparison of normalized interface propensities (*NIP*) and raw interface propensities (*RIP*). *NIP* are calculated by comparing interfaces with the set of residues (*SetrASA*) that has the same relative solvent accessibility as the interfaces. Five *SetrASAs* are extracted, and their mean values are displayed. The standard deviations are below 0.02 (They are shown as bars in the figure, but can barely be seen). *RIP* are calculated by comparing interfaces with the all residues. While *NIP* reveals the trend that hydrophobic residues are preferred in interfaces and hydrophilic residues are unfavorable in interfaces, this trend is not seen in the *RIP*. Many residues have opposite signs in *RIP* and *NIP*. The results were obtained for the *Dataset100*

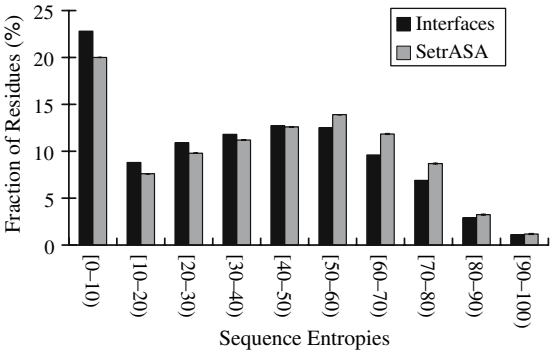

Fig. 9 Comparison of the entropies of interfaces with the *SetrASA*. Sequence entropy values for residues are extracted from the HSSP database (http://www.cmbi.kun.nl/gv/hssp/). The sequence entropy shows the conservation at each residue position in a multiple alignment. The values are normalized over the range of 0–100, with the lowest sequence entropy values corresponding to the most conserved positions. Five *SetrASAs* are extracted, and the mean values are displayed. The standard deviations are below 0.05 (They are shown as bars in the figure but too small to be visible). The results are shown for *Dataset100*

$\log_2 (w_i/W_i)$, where W_i is the fraction of residue i overall, and w_i is the fraction of residue i in the interfaces. Figure 8 shows that *NIP* reveals the trend that hydrophobic residues are preferred in interfaces and hydrophilic residues are unfavorable in interfaces, whereas this trend is not revealed by *RIP*. Many residues have opposite signs in *RIP* and *NIP*. Striking differences are seen for hydrophobic and polar residues. Ile, Val, Leu, and Met have high positive *NIP* but negative *RIP* values. Asn, Asp, Gln, and Glu have negative or neutral *NIP*, while the corresponding values of *RIP* are positive or neutral. Cys and aromatic residues (Tyr, Trp, and Phe) have high positive *NIP* but only weak positive *RIP*. The difference in the definitions of *RIP* and *NIP* is that in *NIP* interfaces are compared with a set of residues that have the same rASA distribution as the interfaces, while in *RIP* interfaces are compared with the overall residues whose solvent accessibility is different from that of the interfaces. The differences between the values of *RIP* and *NIP* indicate that solvent accessibility affects the distribution of residues. Therefore, it is crucial to account for the effect of solvent accessibility when searching for the features that can distinguish interfaces from the rest of the protein.

Previous studies have drawn contradictory conclusions on interface propensities. For example, some studies showed that Ile, Val, and Leu have high positive propensities for interfaces [17, 29, 30], while the study of Ofran and Rost [10] showed that these residues have negative or weak positive propensities for the inter-protein interfaces. Our results show that these three residues have high positive propensities when evaluated using *NIP* and negative propensities when evaluated using *RIP*. In Ofran and Rost's study, interface propensities were calculated using SWISS-PROT as background, so the results are similar to that based on *RIP* in this study, which is calculated using overall residues as background. In the studies by Jones and Thornton [31], Lo Conte et al. [29], and Bahadur et al. [32], interface propensities were calculated based on the accessible surface area of residues, and the results are similar to here based on *NIP*.

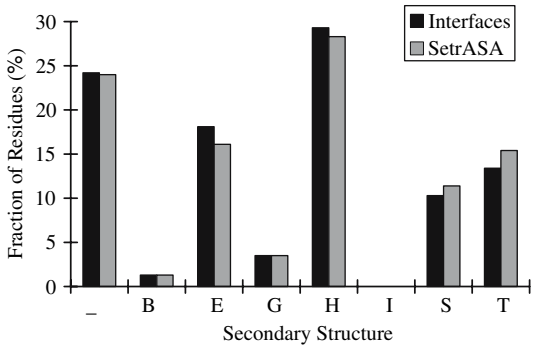

Fig. 10 Secondary structure composition of interfaces and the *SetrASA*. Five *SetrASAs* are extracted. Mean values for the *SetrASAs* are displayed. The standard deviations are less than 0.01 (They are shown as bars in the figure but too small to be visible). The results are achieved using *Dataset100*

3.2.2 Sequence Entropies

Sequence entropies of the *SetrASA* and the interfaces are compared in Fig. 9. The results show that interfaces have more residues with low sequence entropies (conserved). This indicates that interface is more conserved than *SetrASA*. The results from above (See Fig. 2) showed that protein cores are more conserved than interfaces, which in turn are more conserved than non-interface surfaces. Here, Fig. 9 shows that interfaces are more conserved than expected by their exposure.

3.2.3 Secondary Structures

A comparison of the secondary structure composition of the *SetrASAs* with that of interfaces is shown in Fig. 10. Compared with the *SetrASAs*, the interfaces have slightly more residues in E (Extended strand) and H (α helix) and fewer residues in S (Bend) and T (Turn). Despite this, there are no significant differences between the interfaces and the *SetrASAs* in terms of secondary structure composition. Although the results in a previous section (shown in Fig. 3) show some differences in secondary structure composition among protein cores, interfaces, and non-interface surfaces, here, Fig. 10 shows that interfaces do not differ much from the general situation in proteins in their secondary structure composition, after correcting for the effect of solvent accessibility. This suggests that the differences in secondary structure composition among protein cores, interfaces, and non-interface surfaces are mostly due to the differences in accessibility within the three groups rather than to different functions. Raih et al. [33] investigated the interface propensities for secondary structure types by comparing interfaces with surfaces. Their results show that _ (Loop) and S (Bend) are more frequent at interfaces. This observation may be directly attributable to the differences in the accessibilities of interfaces and surfaces.

In summary, to exclude the effect of solvent accessibility, we have compared the interfaces with residue sets (*SetrASA*) having the same relative solvent accessibility distribution as the interfaces. The results show that hydrophobic residues and aromatic residues have high propensities for interfaces; hydrophilic residues (except Arg and His) have negative propensities; and interfaces are more conserved than the remainder of the protein.

3.3 Are the Results Consistent Across Different Datasets?

So far, the results we have reported are obtained using *Dataset100*. In order to evaluate whether the results are consistent across different datasets, we analyze interface properties on three datasets with different constraints on sequence similarity and structure quality: Dataset100, Dataset30, and Dataset30_3. Figure 11 shows that the results obtained using the three datasets are consistent.

3.4 Homomeric Interfaces Compared with Heteromeric Interfaces

Some studies have shown that different types of interfaces have different characters [17, 30]. We divide *Dataset100* into heteromeric interfaces and homomeric interfaces based on the sequence identity between the interacting chains and compare the characteristics of the two types of interfaces (Fig. 12). Figure 12A shows the normalized interface propensities of residues. The results show that hydrophobic residues (Ile, Val, Leu, Phe, Cys, and Met) have high positive propensities for both homomeric interfaces and heteromeric interfaces and hydrophilic residues (Lys, Asn, Asp, Gln, and Glu) have negative propensities. This suggests that both types of interfaces are more hydrophobic than the rest of the protein. Figure 12A also shows that Cys and aromatic residues (Phe, Trp, and Tyr) have higher propensities in heteromeric interfaces than at homomeric interfaces. Hydrophobic residues (Ile, Val, Leu, and Met) have higher propensities for homomeric interfaces than for heteromeric interfaces and the opposite is observed for charged residues (except Arg). This indicates that homomeric interfaces are more hydrophobic than heteromeric interfaces. This result is consistent with the results of previous studies [17, 30]. Figure 12B shows that heteromeric interfaces have more residues with low entropies (conserved) than homomeric interfaces, suggesting that heteromeric interfaces are more conserved than homomeric interfaces. This may be related to the fact that a

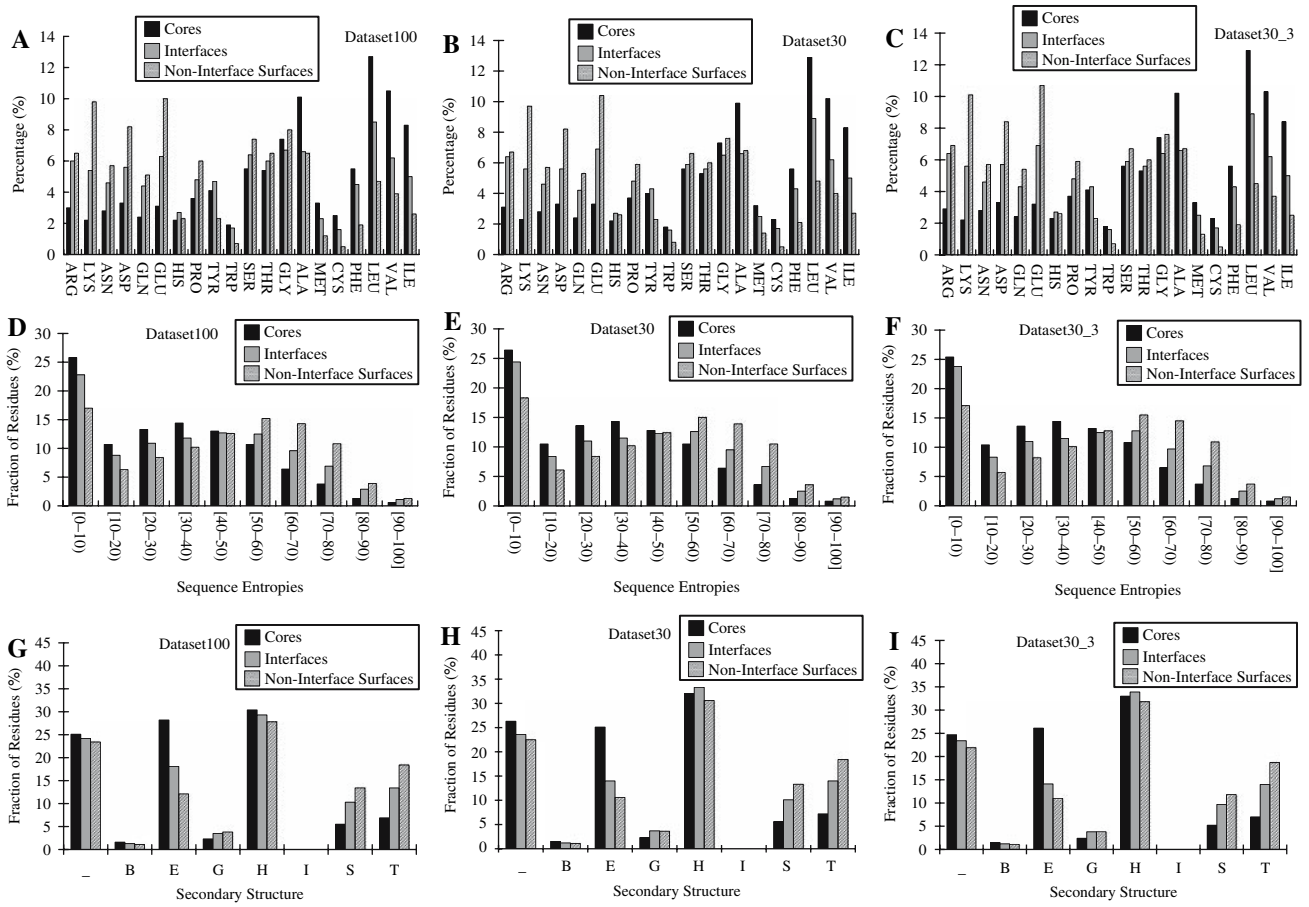

Fig. 11 The results obtained for three different datasets are consistent. (**A–C**) Residue composition. (**D–F**) Sequence entropy distribution. (**G–I**) Secondary structure composition. (**J–L**) Interface sizes. (**M–O**) Raw contact frequencies given by $(C_{ij}/\sum_{m,n} C_{mn})$, where C_{ij} is the number of contacts between residue types i and j. (**P–R**) Contact preferences given by $\log_2\left(\left(C_{ij}/\sum_{m,n} C_{mn}\right)\Big/(w_i \times w_j)\right)$, where w_i is the frequency of residue type i in the interfaces. **A**, **D**, **G**, **J**, **M**, and **P** are the results on *Dataset100*, which consists of 6,545 interfaces. **B**, **E**, **H**, **K**, **N**, and **Q** are the results on Dataset30, which consists of 2,557 interfaces. The mutual similarities among the interfaces are below 30%. **C**, **F**, **I**, **L**, **O**, and **R** are the results for Dataset30_3, which consists of 2,310 interfaces from structures having resolution better than 3.0 Å. The mutual similarities among the interfaces are below 30%

heteromeric interface involved two different proteins, and a mutation in one protein requires a complimentary mutation in the interacting protein to restore the interaction function, while a homomeric interface involves two identical chains, one mutation will affect both sides of the interface. Thus, mutations are less tolerable at heteromeric interfaces. Comparison of secondary structure composition (Fig. 12C) shows that heteromeric interfaces have more loops (_) and extended strands (E) and fewer α-helixes (H) than homomeric interfaces. Figure 12D shows the distributions of interface sizes for heteromeric interfaces and homomeric interfaces. Both types of interfaces have a peak value in the range 600–800 Å^2. However, the homomeric interfaces are larger than the heteromeric interfaces: 63% of the homomeric interfaces are larger than 800 Å^2, while only 53% of the heteromeric interfaces are larger than 800 Å^2. The average size of the homomeric interfaces is 1,311 Å^2, and the average size of the heteromeric interfaces is 1,112 Å^2. This result is consistent with the conclusion of a previous study that homomeric interfaces are larger than heteromeric interfaces [30]. Figure 12G–H show that the contacts between residues with opposite charges (Arg–Asp, Arg–Glu, Lys–Asp, and Lys–Glu) and the contacts between hydrophobic residues (the red regions at the lower-right corners of Fig. 12G–H) are preferred in both types of interfaces. Compared with homomeric interfaces, heteromeric interfaces have relatively more contacts involving Cys or aromatic residues (Phe, Tyr, and Trp). The columns and rows in Fig. 12G for these residues are more frequent (red) than the corresponding entries in Fig. 12H.

4 Discussion of Results

In this study, we compare various properties of protein cores, interfaces and non-interface surfaces, analyze

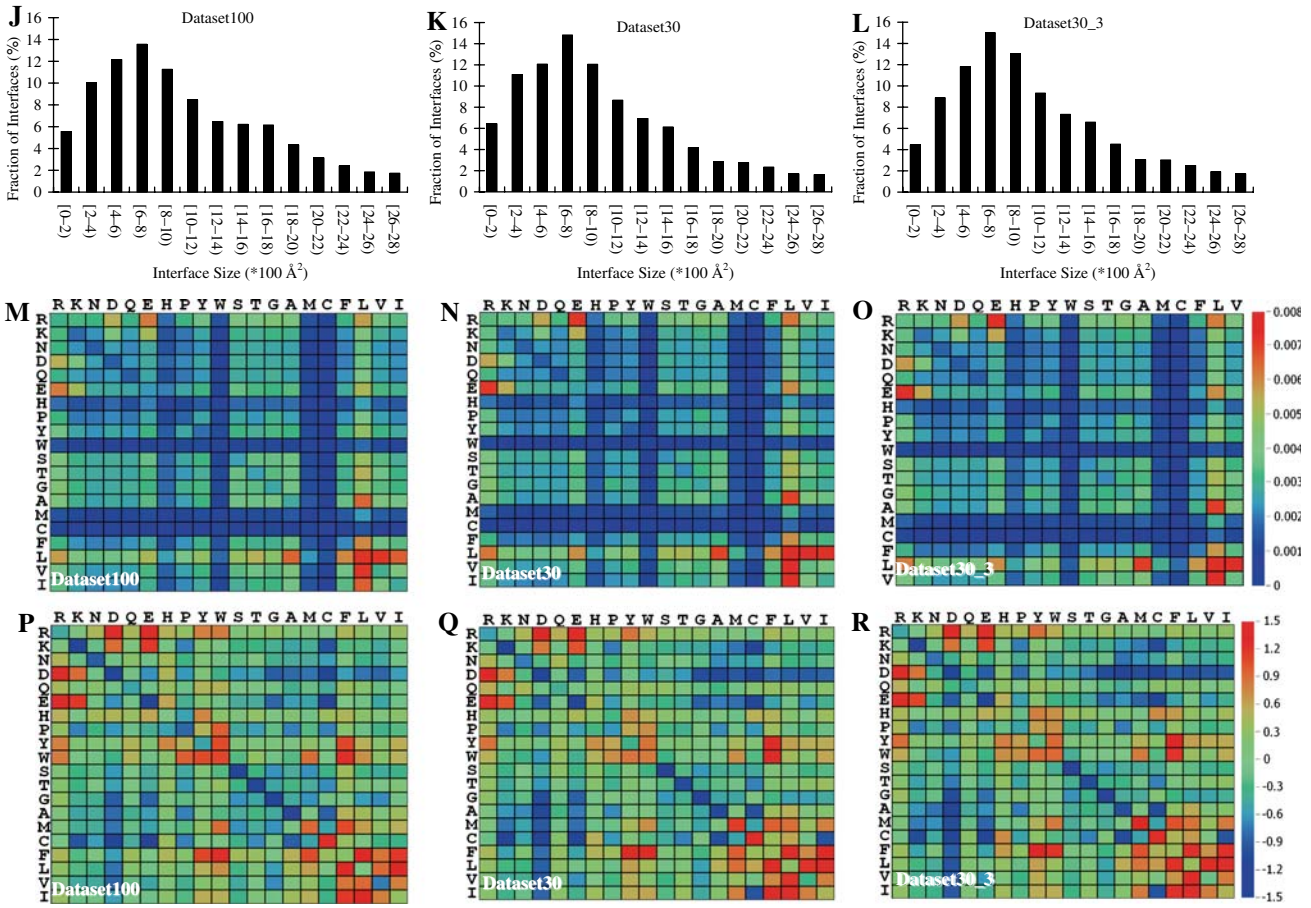

Fig. 11 continued

interface properties by separating out the effect of solvent accessibility, and investigate the differences between homomeric interfaces and heteromeric interfaces.

Compared with previous studies, the significance aspects of this study include: (1) use of large datasets of protein–protein interfaces; (2) confirming results by using three datasets with different constraints on sequence similarity and structure resolutions; and (3) separating out the effect of solvent accessibility in analyzing the characteristics of protein–protein interfaces.

We found that solvent accessibility affects the distribution of residues and it is crucial to account for the effect of solvent accessibility when searching for the features that can distinguish interfaces from the rest of the protein. Generally, hydrophilic residues are more frequent in the portions of proteins that are highly solvent accessible, and hydrophobic residues are more frequent in the buried portions. Because protein core residues have lower solvent accessibility than interface residues, and non-interface surface residues have higher solvent accessibility than interface residues, the residue distributions among these groups are affected not only by the different functions of these groups but also by the difference in their solvent accessibilities. To evaluate whether residues have special preferences for the interfaces because of the function, one must separate out the effect of solvent accessibility. Here, we do so by comparing protein–protein interfaces with a set of residues having the same solvent accessibilities. This allows us to separate out the effect of solvent accessibility on the distributions of residues, secondary structure, and sequence entropy. The comparison shows the trend that hydrophobic residues are preferred in interfaces and hydrophilic residues are not. In contrast, this trend is not observed when we compare interfaces with the overall residues, that is, when the effect of solvent accessibility is not separated out.

The result shows clearly that the interfaces have more hydrophobic residues and fewer hydrophilic residues. Interfaces with hydrophobic residues are critical for the stabilization of protein–protein complexes. The formation of a protein–protein complex in aqueous solution was reported to be an entropy-driven process [34]. The thought was that burial of hydrophobic surface patches yields a large entropy gain, providing a driving force for the

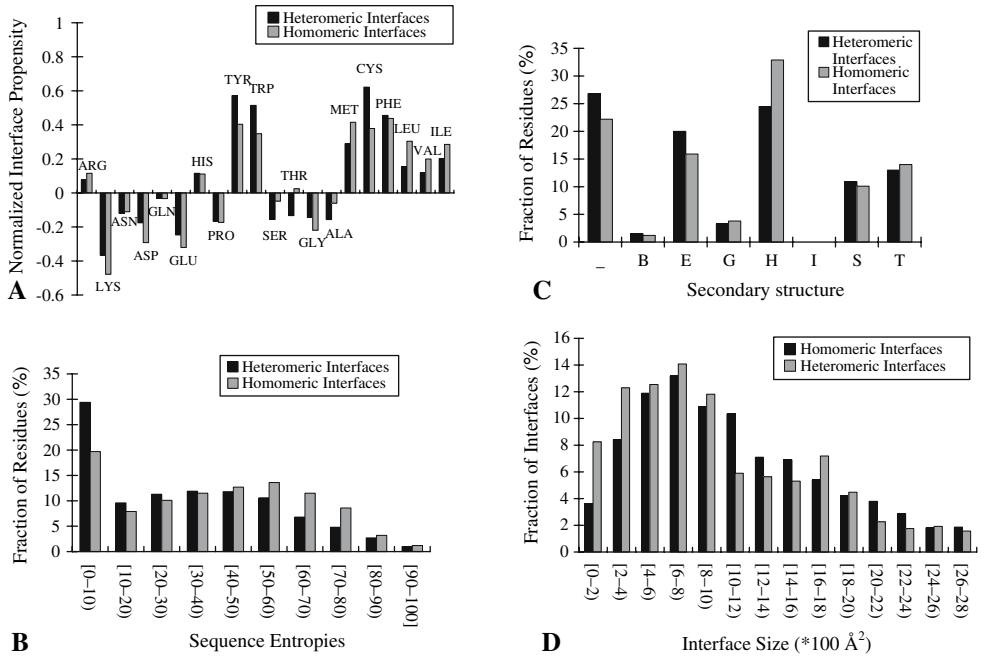

Fig. 12 Comparisons between homomeric interfaces and heteromeric interfaces. (**A**) Normalized interface propensities. (**B**) Sequence entropies. (**C**) Secondary structures. (**D**) Interface sizes. (**E–F**) Raw contact frequencies given by $(C_{ij}/\sum_{m,n} C_{mn})$, where C_{ij} is the number of contacts between residue types i and j. (**G–H**) Contact preferences given by $\log_2\left(\left(C_{ij}/\sum_{m,n} C_{mn}\right)\Big/(w_i \times w_j)\right)$. The results are obtained from *Dataset100*. Heteromeric interfaces and homomeric interfaces have been extracted from *Dataset100* based on the sequence similarities between the interacting protein chains. An interface is a homomeric interface if the two interacting chains have a sequence identity greater than 95%. Otherwise, it is considered a heteromeric interface

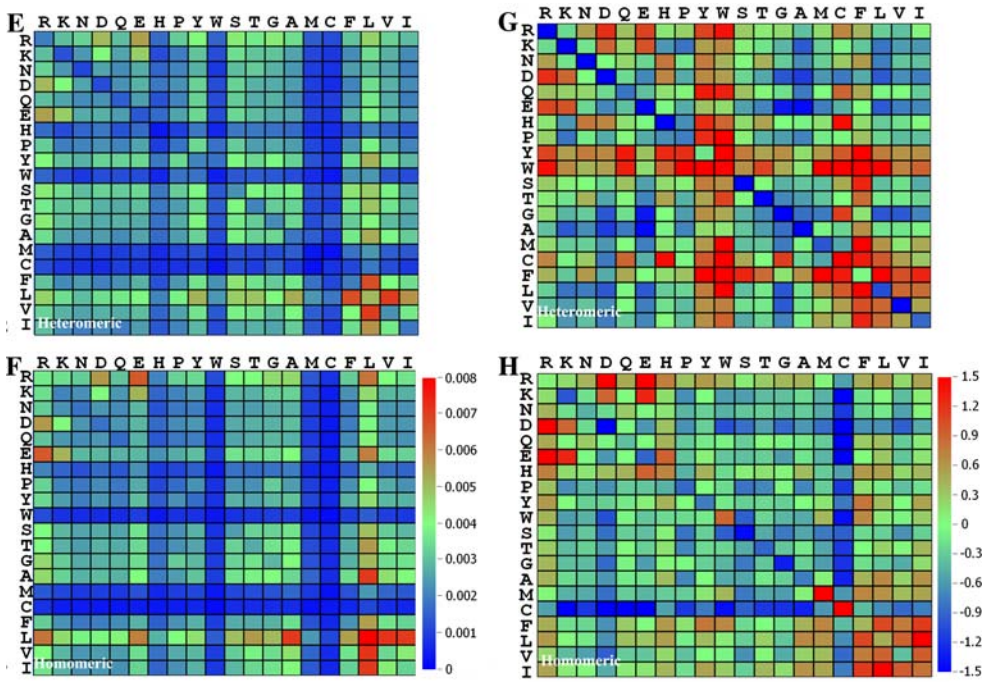

Fig. 12 continued

formation of protein complexes and thus stabilizing the resulting complexes. The results also show that the interfaces are more conserved. Conserved interfaces are crucial for the maintenance of protein–protein interactions during evolution.

We found that Cys–Cys contacts, the contacts between residues with opposite charges and the contacts between hydrophobic residues are more frequent across protein–protein interfaces. Hydrophobic interactions have been widely accepted to be the main stabilizing force for two proteins to interact. Some studies have shown that interactions between charged residues also contribute to protein–protein interactions [35, 6]. Bahar and Jernigan [35] showed that at close distances, interactions between pairs of hydrophilic residues are predominantly important; whereas hydrophobic interactions are important at longer distances. Cys–Cys pairs can contribute to the interactions by forming disulfide bonds [9]. The results we obtained confirm that disulfide bonds, salt-bridges, and hydrophobic interactions are the important forces in protein–protein interactions.

We also found that aromatic residues are more frequent at interfaces. Aromatic residues can form strong hydrophobic interactions between the bulky hydrophobic side chains. In addition to the hydrophobic interactions, the parallel arrangement of two aromatic rings makes further contributions by creating tighter packing with better geometric fit. The enhanced abundance of aromatic residues in interfaces might imply more precise geometric fits are achievable for these ring structures. Frequent interactions between aromatic residues are observed in this study.

Acknowledgements This Research was supported in part by a grant from the National Institutes of Health (GM 066387) to VH, DD, and RLJ.

References

1. Chothia C, Janin J (1975) Nature 256:705–708
2. Wodak SJ, Janin J (2002) Adv Protein Chem 61:9–73
3. Deremble C, Lavery R (2005) Curr Opin Struct Biol 15:171–175
4. Ponstingl H, Kabir T, Gorse D, Thornton JM (2005) Progr Biophys Mol Biol 89:9–35
5. Reichmann D, Rahat O, Cohen M, Neuvirth H, Schreiber G (2007) Curr Opin Struct Biol 17:67–76
6. Sheinerman FB, Norel R, Honig B (2000) Curr Opin Struct Biol 10:153–159
7. Heifetz A, Katchalski-Katzir E, Eisenstein M (2002) Protein Sci 11:571–587
8. Vizcarra CL, Mayo SL (2005) Curr Opin Chem Biol 9:622–626
9. Glaser F, Steinberg DM, Vakser IA, Ben-Tal N (2001) Proteins 43:89–102
10. Ofran Y, Rost B (2003) J Mol Biol 325:377–387
11. Miyazawa S, Jernigan RL (1985) Macromolecules 18:534–552
12. Keskin O, Bahar I, Badretinov AY, Ptitsyn OB, Jernigan RL (1998) Protein Sci 7:2578–2586
13. Young L, Jernigan RL, Covell DG (1994) Protein Sci 3:717–729
14. Berchanski A, Shapira B, Eisenstein M (2004) Proteins 56:130–142
15. Ben-Naima A (2006) J Chem Phys 125:24901
16. Keskin O, Mab B, Nussinov R (2005) J Mol Biol 345:1281–1294
17. Jones S, Thornton JM (1996) Proc Natl Acad Sci USA 93:13–20
18. Peter Block JP, Hülermeier E, Sanschagrin P, Sotriffer CA, Klebe G (2006) Proteins 65:607–622
19. Chakrabarti P, Janin J (2002) Proteins 47:334–343
20. Kyu-il Cho KL, Lee KH, Kim D, Lee D (2006) Proteins 65:593–606
21. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) Nucl Acids Res 28:235–242
22. Henrick K, Thornton JM (1998) Trends Biochem Sci 23:358–361
23. Hubbard SJ (1993) NACCESS, department of biochemistry and molecular biology. University College, London
24. Gutteridge A, Bartlett GJ, Thornton JM (2003) J Mol Biol 330:719–734
25. Kyte J, Doolittle RF (1982) J Mol Biol 157:105–132
26. Nooren IMA, Thornton JM (2003) J Mol Biol 325:991–1016
27. Kabsch W, Sander C (1983) Biopolymers 22:2577–2637
28. McCoy AJ, Chandana Epa V, Colman PM (1997) J Mol Biol 268:570–584
29. Lo Conte L, Chothia C, Janin J (1999) J Mol Biol 285:2177–2198
30. Bahadur RP, Chakrabarti P, Rodier F, Janin J (2003) Proteins 53:708–719
31. Jones S, Thornton JM (1997) J Mol Biol 272:121–132
32. Prasad Bahadur R, Chakrabarti P, Rodier F, Janin J (2004) J Mol Biol 336:943–955
33. Raih MF, Ahmad S, Zheng R, Mohamed R (2005) Biophys Chem 114:63–69
34. Creighton T (1997) Protein structures and molecular properties. WH Freeman, New York
35. Bahar I, Jernigan RL (1997) J Mol Biol 266:195–244