Are protein-protein interfaces special regions on a protein's surface?

Sam Tonddast-Navaei and Jeffrey Skolnick^{a)}

*Center for the Study of Systems Biology, School of Biology, Georgia Institute of Technology,
250 14th Street N.W., Atlanta, Georgia 30318, USA*

(Received 27 August 2015; accepted 25 November 2015; published online 15 December 2015)

Protein-protein interactions (PPIs) are involved in many cellular processes. Experimentally obtained protein quaternary structures provide the location of protein-protein interfaces, the surface region of a given protein that interacts with another. These regions are termed half-interfaces (HIs). Canonical HIs cover roughly one third of a protein's surface and were found to have more hydrophobic residues than the non-interface surface region. In addition, the classical view of protein HIs was that there are a few (if not one) HIs per protein that are structurally and chemically unique. However, on average, a given protein interacts with at least a dozen others. This raises the question of whether they use the same or other HIs. By copying HIs from monomers with the same folds in solved quaternary structures, we introduce the concept of geometric HIs (HIs whose geometry has a significant match to other known interfaces) and show that on average they cover three quarters of a protein's surface. We then demonstrate that in some cases, these geometric HI could result in real physical interactions (which may or may not be biologically relevant). The composition of the new HIs is on average more charged compared to most known ones, suggesting that the current protein interface database is biased towards more hydrophobic, possibly more obligate, complexes. Finally, our results provide evidence for interface fuzziness and PPI promiscuity. Thus, the classical view of unique, well defined HIs needs to be revisited as HIs are another example of coarse-graining that is used by nature. © 2015 AIP Publishing LLC. [<http://dx.doi.org/10.1063/1.4937428>]

NOMENCLATURE

EF	=	Enrichment Factor
GI	=	Gene ID
HI	=	half-interface
HIPPIE	=	human integrated protein-protein interaction reference
PPIs	=	protein-protein interactions
rASA	=	relative Accessible Solvent Area
PDB	=	protein data bank
RMSD	=	root-mean-square deviation

I. INTRODUCTION

Comprehensive understanding of many complex cellular processes, such as signaling, transcription, inhibition, translation, and regulation, is not possible without considering ligand-macromolecule and macromolecular interactions, among which are protein-protein interactions (PPIs).¹ The existence of metabolic pathways and their precise regulation is governed by these interactions. PPIs provide an effective means of regulation and minimization of genome size while retaining the advantages conferred by modular oligomerization. PPIs occur when at least two individual proteins make physical contacts via their surface amino acids, also referred to as half-interfaces (HIs), and form a protein-protein interface due to hydrophobic or complementary

electrostatic interactions. Based on their binding affinity, these PPIs are categorized as being transient or obligate.^{2,3} Irrespective of the nature of the interactions and their strength, mapping the PPI network (the interactome) is not feasible without the knowledge of which proteins interact, with a deeper understanding provided when their interacting interface is known. A recent study shows that a significant number of small molecule (ligand) binding sites are located near these interfaces,⁴ making them important in terms of the small molecule regulation provided by macromolecular association.

In the last decade, there has been an exponential growth in the number of protein structures⁵ solved using X-ray crystallography and NMR methods to the point that it has been declared that fold space for the monomeric state of proteins is essentially complete.^{6–8} Quaternary structure space, on the other hand, is far from complete when we compare the non-redundant number of available multimeric complexes ($\sim 10^4$) to the theoretical upper bound of possible complexes ($\sim 10^7$).⁹ This incompleteness is reflected in the status of currently known interactome maps. Indeed, a number of studies^{10–15} suggest that around 80% of actual pairwise protein interactions are yet to be determined. Despite this significant limitation, the existing library of multimeric protein structures has been the major source of our concepts of the features of HIs.

Protein HIs were considered to be a “special” subset of the protein surface that covers roughly 25%–30% of the whole surface on average. Given this presumption, surfaces were classified into interacting (HI) and non-interacting (non-HI) regions.¹⁶ Therefore, many efforts have been made to

^{a)} Author to whom correspondence should be addressed. Electronic mail: skolnick@gatech.edu. Tel.: (404) 407-8975. Fax: (404) 385-7478.

FIG. 1. Schematic representation of copying HIs and validation process. Left panel shows the known interaction and HI for query protein Q shown in black. The middle panel shows the process of copying geometric HIs. HIs copied from T_1 and T_4 are “new” with respect to the known HI of Q. Those copied from T_3 and T_2 are “known” and “overlapped” HIs, respectively. The right panel shows the process of validating the geometric HIs by copying the corresponding partner and matching to the HIPPIE dataset.

predict the HI for a given protein, when either the protein’s sequence or structure is provided as input.^{17–20} This is a reasonable goal only if the HI comprises a relatively small portion of the entire protein’s surface. The majority of HI prediction approaches use machine-learning algorithms and utilize the existing, incomplete library of multimeric structures for training.^{17,21–24} In these studies, many different features such as sequence conservation, solvent accessible surface area, residue propensity, and geometric properties were employed. Though there were some signals in each of the mentioned features, none was strong enough to make HI residues dramatically stand out from the non-interacting ones. Possibly, this reflects the fact that the descriptors were inadequate, the training set was too small or alternatively, HIs are perhaps less well defined than originally conjectured. If so, ideas from coarse-graining as applied to other protein problems including protein structure prediction and ligand virtual screening might be appropriate.^{25–29} These methods identify key descriptors of a system which capture the essential physics. Often due to inaccuracies in the force field, such lower resolution descriptors perform better than detailed atomic approaches.^{25–29} We adopt this same philosophy in what follows.

Rather than attempting to identify the HI on a protein without knowledge of its partner(s), other studies focused on predicting the interaction pose and thereby the quaternary structure for a given set of proteins.^{24,30–35} Roughly, there are two types of approaches: docking algorithms^{30,31,35} whose goal is to predict the quaternary structure *a priori* and are template free, i.e., they do not require that any previous example of the quaternary structure be known. Others are template based^{24,32–34} and predict the interaction pose by copying from a related quaternary structure template. While docking methods suffer from the lack of a good scoring function to pick the right decoy,³⁶ the latter’s performance is limited by the small number of available quaternary structures. Nevertheless, all methods implicitly assume that there is just a handful of interfaces per protein with relatively well-defined boundaries. In addition, the performance of these methods was relatively poor compared to computational methods for predicting protein structures^{25,37,38} or determining the catalytic sites for an enzyme.^{39–41} The low success rates¹⁶

of predicting PPIs suggest that there might be flaws in our understanding of the nature of PPIs. Thus, it is appropriate to revisit the fundamental concepts and assumptions about protein interfaces.

Using structural analysis, this work demonstrates that potential HIs cover roughly three quarters of a protein’s surface. In addition, the ratio of the number of interacting amino acids in HIs to non-interacting ones (also referred to as the interface to surface ratio in this paper) grows as the number of known interaction partners increases. This correlation suggests that as we move towards a more complete interactome and quaternary structure library, we will find many novel interaction sites on a given protein’s surface. Thus, HIs are not “special” as had been implicitly assumed in prior work. If this conclusion holds, then the classical view of protein HIs needs substantial revision. Our conclusion is based on a very simple idea: For a given protein, we merely copy all the HIs (Figure 1) from structures having a similar fold independent of sequence identity. Since the only condition is geometric similarity between the query and template proteins, we henceforth refer to them as “geometric half-interfaces.” Next, we used available experimental data to validate whether the predicted geometric HIs are also relevant physical interfaces with potential biological relevance. This subset of validated HIs is further discussed and categorized into three groups: novel, known, and overlapped, viz., having both novel and known interfacial residues. Finally, validation examples of proteins are presented in which geometric HIs are used to suggest the interaction pose of two proteins that are *a priori* predicted to interact (without use of any of their known binding information).

II. METHODS

A. DL6k set

All dimer complexes in the Protein Data Bank (PDB)⁵ obtained as of March 2014 via X-ray diffraction were selected. To avoid complexes with peptides, only those where both participating chains are greater than 40 amino acids in length are included. Next, the interaction energy between chains was calculated using a residue based statistical potential⁴² and a

cutoff of -12 kcal/mol was used to filter the structures. Finally, to remove redundancy, only complexes having at least one chain with a sequence identity of less than 35% to the other members is kept. This gives 5938 dimer complexes made of 6154 unique monomers. The list of complexes is available at <http://cssb2.biology.gatech.edu/jcp-ppi/DL6k.txt>.

B. BU21k set

The BU-set ($\sim 21k$) is built by clustering⁴³ PDB entries with at least one multimeric biological assembly unit at the 90% sequence identity level. Next, the HIs of all the members in each cluster were assigned to the cluster representative and considered to be the “known” HI for that cluster. The BU21k set is used as templates for copying HIs. The list of complexes is available at <http://cssb2.biology.gatech.edu/jcp-ppi/BU21K.txt>.

C. HI residue definition

Surface residues are defined as those with relative Accessible Solvent Area (rASA) greater or equal to 20% in their monomeric state. The rASA was calculated using the NACCESS software program. A surface residue is considered to be part of a HI if there is a minimum of $5 \text{ \AA}^2$ loss in its absolute ASA after multimer formation.

D. Copying HI from homologs/analogs

All proteins are described at the coarse-grained $C\alpha$ and $C\beta$ level. This enables us to find pairs of HIs whose backbone geometries are very similar, yet whose residue identity at structurally equivalent backbone positions can be very different. For a given query protein, HIs are copied (Figure 1) from their template proteins if the following two conditions are met: (a) The query and templates share globally similar structures. TM-align⁴⁴ is used to align the structure of the query protein against the BU21k set. To ensure that the query and template have the same fold, a TM-score cutoff of

greater than or equal to 0.5 (p-value of 5.5×10^{-7})⁴⁵ is used and the query residues aligned to the template's HI are marked for possible copying. (b) Local similarity of the HI between the query and the template. The previously marked residues as belonging to a HI are aligned against the template's HI through a non-sequential alignment algorithm using APoc.⁴⁶ Next, if the alignment has a PS-score of 0.4 (p-value of 2×10^{-4}) (Figure S1 of the supplementary material⁴⁷) or above and the number of aligned residues is at least 10 amino acids, then the corresponding residues are marked as HI residues for the query protein. The cutoff values for both TM-score and PS-score were chosen to be relatively high to reflect our conservative approach for copying HIs. This will yield structurally similar HIs at a coarse-grained level, which is typically in the range of $C\alpha$ RMSDs (root-mean-square deviations) of 2–4 Å.

We note that APoc was originally developed and tuned to measure the geometric similarity between protein small molecule ligand binding pockets at a coarse-grained $C\alpha$ and $C\beta$ level. Calculating the mean of random PS-score (Figure S2 of the supplementary material⁴⁷) for the case of HIs resulted in the value of 0.289, which is less than the value for pockets (0.308) as reported by Gao and Skolnick.⁴⁶ Therefore, for the same PS-score, the p-value for a HI is at least as significant as that reported for small molecule ligand binding pockets.

E. Evaluation of the half interface predictions

To assess the quality of half-interfaces, we report the average precision and recall

$$\langle Precision \rangle = \frac{\sum_i^{N_{total}} Precision_i}{N_{total}}; \quad (1)$$

$$\langle Recall \rangle = \frac{\sum_i^{N_{total}} Recall_i}{N_{total}}.$$

N_{total} is the total number of cases with at least one predicted interaction and one known experimentally verified interaction. Also, the precision and recall for individual cases are calculated by

$$Precision_i = \frac{\text{Number of known true positive predictions for } i}{\text{Total number of predictions for } i}, \quad (2)$$

$$Recall_i = \frac{\text{Number of known true positive predictions for } i}{\text{Total number of known interactions for } i}. \quad (3)$$

Finally, the Enrichment Factor (EF) is defined as

$$EF = \frac{\text{Number of true positive predictions/Total number of predictions}}{\text{Number of known interactions/Number of possible pair interactions}}. \quad (4)$$

F. Calculation of predicted HIs coverage with respect to a reference HI

The coverage of a given HI with respect to a reference (template HI or “known” HI) is calculated by the

following formula:

$$Coverage = \frac{N_{cp}}{N_{tot}} \quad (5)$$

in which N_{cp} corresponds to the number of copied residues from the template's HI (that satisfies the conditions in Section II D) and N_{tot}^{temp} is the total number of residues in template HI.

G. Sequence conservation score

The Jensen-Shannon divergence (JSD)⁴⁸ is used to measure the evolutionary sequence conservation by calculating the divergence between the amino acid distribution p_{ia} and the background distribution (frequency of amino acid occurrence) for a given amino acid type (f_a),

$$JSD_i = SE\left(\frac{p_{ia} + f_a}{2}\right) - \frac{1}{2}SE(p_{ia}) - \frac{1}{2}SE(f_a), \quad (6)$$

$$SE(p_{ia}) = -\sum_{a=1}^{20} p_{ia} \log p_{ia}. \quad (7)$$

Our calculation is inspired by the code provided in the work of Capra and Singh³⁹ using the same background distribution (f_a) and gap penalty. The final raw score is converted to Z-score using the mean (μ) and standard deviation (σ) for each case,

$$Z_i = \frac{JSD_i - \mu(JSD)}{\sigma(JSD)}. \quad (8)$$

Multiple sequence alignments for sequence conservation calculations were calculated using PSI-BLAST⁴⁹ with an E-value inclusion of 10^{-4} and considering the maximum number of homologs to be 1000.

H. Sequence similarity calculation

The sequence similarity score between the HI of query “ i ” and its template is calculated using the following equation:

$$sim_i = \frac{\sum_j^{N_{intf}} s_j}{N_{intf}} \quad (9)$$

in which N_{intf} is the total number of amino acids in the HI of the query and s_j is the similarity score between the amino acid “ j ” in the query and the corresponding amino acid in the

HI of the template “ j' .” The value of s_j equals one or zero if the BLOSUM62⁵⁰ element for jj' is greater than zero or less than or equal to zero, respectively.

III. RESULTS AND ANALYSIS

A. Interface to surface ratio for geometric HIs

For two proteins to form a protein-protein complex, their interacting HIs must be both geometrically and chemically compatible so that their binding free energy is favorable. In this section, we ignore the second condition (favorable binding free energy) and only focus on the geometric features of the HI. We examine the fraction of a protein's surface that is geometrically consistent with protein-protein interactions and compare this to the currently known ratio of interface to surface residues. Though different studies¹⁶ use different definitions for surface and interface residues, all report the same range of values (between 25% and 35% on average) for the percent of known interface to surface amino acids.

To address this question, for a given protein, we copied all HIs from the query protein's homologs and analogs that have the same fold irrespective of the sequence identity, as described in Sec. II. The idea of copying HIs was introduced more than a decade ago,^{51,52} however, we now add the condition of local HI structural similarity as an additional filter. Our analysis has been done on the DL6k set. The average ratio of interface to surface amino acid residues for a given sequence identity cutoff is shown in Figure 2(a). When copying HIs from high sequence identity templates, the interface to surface ratio is almost one third, agreeing well with ratios reported in previous studies.¹⁶ As the sequence identity cutoff of the templates for copying HIs is relaxed by including distant homologous and analogous proteins, the ratio of interface to surface residues starts to increase. The growth is at first slow (around 10%) until 30% sequence identity. However, below the 30% sequence identity threshold, the number of available templates increases rapidly, and the interface to surface ratio also experiences a sharp growth and reaches a surprising value of ~75%. “Known HIs” are defined as those with a minimum

FIG. 2. (a) Interface to surface ratio with respect to the global sequence identity of the query to the template from which the HIs were copied. N_{surf} is defined as the total number of amino acids that are exposed on the surface in the monomeric state of the protein. (b) Distribution of interface to surface ratios for known (green) and geometric HIs (red).

FIG. 3. Number of available templates for known (a) and geometric (b) HIs with respect to the interface to surface ratio. Data are shown in black dots, with the average is shown in red. For known interfaces, the number of templates is the number of distinct PDB files that have a sequence identity greater than 90% to the query. For geometric templates, the number of templates is defined as the number of cases with the same fold (TM-score >0.5) and sequence identity of less than 90%.

sequence identity of 90% to the template from which the interface was copied, while there is no sequence cutoff for copying HIs in the case of geometrically compatible ones (geometric HIs). Figure 2(b) shows the distributions of the interface to surface ratio per protein for known (green) and geometric (red) HIs. It clearly shows the dramatic difference between the known HIs and geometrically compatible ones. In the case of known HIs, for half of the population, ~30% of the protein's surface participates in PPIs. In contrast, for geometric HIs, half of the proteins have at least 70% of their surface involved in PPIs.

For geometric HIs, we need to point out that the ratios of interface to surface residues are likely the lower limit due to the following two reasons: First, we only copy similar HIs from proteins with the same monomeric fold. Copying HIs from different folds could only increase the fraction of surface residues that are in a geometric HI. Second, the ratio of interface to surface residues is correlated with the number of templates available for a given protein for both known, see Figure 3(a), and geometric HIs, see Figure 3(b). Hence, cases with lower or unchanged interface to surface ratio are simply due to the current limitations in the protein quaternary structure database. Our results demonstrate that more than three quarters of a protein's surface satisfies the geometric condition for participating in a PPI.

B. Comparison to Human Integrated Protein-Protein Interaction rEference (HIPPIE) database

While we just showed that, on average, 75% of a protein's surface could potentially be used for PPIs, the question is whether these new HIs are actually used by nature. In other words, is there a known protein sequence that interacts with the new proposed HI? The simplest way to address this is to check for interactions with the interacting partners of the same protein templates from which the HI was copied. To do so, we focused on the human protein subset of the (980 protein) DL6k data set with unique gene identities (GIs). Also, we only considered HIs copied from human

proteins in the template database BU21k. Next, we checked if there is an experimentally determined interaction between the query protein and the partner of the template protein from which the HI was copied (results can be found at <http://cssb2.biology.gatech.edu/jcp-ppi/>).

The HIPPIE^{53,54} provided the source of experimental interactions. The HIPPIE data set is comprised of ~200 000 interactions made by ~15 000 human proteins, providing 13 interactions per protein on average. The overlap between the HIPPIE set and BU21k consists of 2835 proteins and includes ~33 000 interactions. Out of 980 human proteins in the DL6k set, 736 (75%) were able to find at least one template in HIPPIE with a global sequence identity of less than 90% to copy an interface. 360 of 736 (49%) proteins had at least one of their predicted interactions verified by HIPPIE. Without enforcing additional conditions for copying interfaces, there are 962 total hits for the 736 query proteins with at least one good template. This corresponds to an average precision and recall of 13.47% and 6.70%, respectively. The probability of finding these interactions by chance is 4×10^{-3} , giving to an EF of 31.4. Furthermore, the reported precision is really the “observed precision” rather than the true precision.⁵⁵ Since only N_{obs} of N_{true} protein pair interactions are known (due to the incompleteness of the interactome), the true precision defined as the ratio of “observed precision”/($N_{\text{obs}}/N_{\text{true}}$) could be larger. Considering studies,^{10–15} which suggest that roughly 20% of the interactome is known, the true precision for this study could be higher by almost a factor of five. The correlation between the “observed precision” and N_{obs} can be seen in Figure S3 of the supplementary material⁴⁷ and shows a smooth increase in the average observed precision as the number of known interacting partners increases.

Although predicting protein-protein interactions and their associated poses is not the purpose of this study, as shown in Table I, our results show that introducing additional conditions (both geometric and chemical) for filtering the predictions can increase the precision. For filtering by geometric conditions, we have used the “coverage” of the copied HIs with respect to their reference template as defined in Sec. II F. We

TABLE I. Observed precision and recall of the predicted PPIs as inferred by similar geometric HIs as assessed by the HIPPIE database.

Cov ^b	Sim ^a															
	0.0				0.2				0.4				0.6			
	$\bar{P}^c$	$\bar{R}^d$	N _H ^e	N _C ^f	$\bar{P}$	$\bar{R}$	N _H	N _C	$\bar{P}$	$\bar{R}$	N _H	N _C	$\bar{P}$	$\bar{R}$	N _H	N _C
0.0	0.13	0.07	962	736	0.15	0.06	848	707	0.22	0.05	530	576	0.34	0.04	240	307
0.2	0.14	0.07	960	734	0.15	0.06	847	705	0.22	0.05	530	574	0.34	0.04	240	307
0.4	0.14	0.07	933	713	0.16	0.06	830	682	0.23	0.05	523	558	0.34	0.04	239	303
0.6	0.16	0.06	920	667	0.17	0.06	739	641	0.24	0.05	486	532	0.35	0.04	226	287
0.8	0.20	0.05	553	571	0.21	0.05	521	550	0.28	0.05	382	450	0.37	0.04	190	249

^aSim (Sequence similarity): Defined with respect to the template's HI (see Sec. II) used as a measure for chemical compatibility.^bCov (Coverage): Defined with respect to the template's HI (see Sec. II) used as a measure for geometric compatibility.^c $\bar{P}$: Average precision over the number of cases.^d $\bar{R}$: Average recall over the number of cases.^eN_H: Number of hits defined as the number of true positive HI (there could be more than one per query).^fN_C: Number of cases defined as the number of queries with non-zero predictions and HIPPIE entries.

also introduced the chemical similarity (Sec. II H) score between the sequence of the HI and its template. Using the relationship between these two factors (Figures S4 and S5 of the supplementary material⁴⁷), one can increase the precision by a factor of 2.6. Therefore, as the geometric and chemical similarities between two HIs increase, the chance of them being involved in an interaction with the same protein partner increases as well.

C. Similar geometric HIs sharing the same partners

In Sec. III B, we showed that it is likely for two proteins to interchange protein partners given that their HIs are geometrically similar (coverage greater than 0.6 as shown in Figure S4 of the supplementary material⁴⁷). One can ask the following: what is the level of chemical similarity in such a case? In particular, what is the sequence identity at the HI with respect to the global sequence identity of its template? Interestingly, our results for true positive cases show (see Figure S6 of the supplementary material⁴⁷) that the sequence identities at the global and interface level are low and on average comparable (0.32 ± 0.20 and 0.30 ± 0.24 ,

respectively), having a Pearson correlation coefficient of 0.84 and a P-value of 2.2×10^{-16} . Compared to the sequence identity in the HI, global sequence identity is higher for 58% of the cases. Note that 45% of the HIs have a sequence identity less than or equal to 0.2; this ratio increases to 75% for HIs less than 0.4 sequence identity. However, when we measure the sequence similarity for HIs (Figure 4(a)), 12% have a sequence similarity below or equal to 0.2% and 45% have a sequence similarity less than or equal to 0.4 for true positive measurements. This is an indication of the plasticity in PPIs and shows that having met the geometric condition, there is more than one way of obtaining a PPI; thus, rather than their interactions being unique, HI behave in a coarse-grained, degenerate fashion.

D. Known, new, and overlapped HIs

So far, we showed that proteins with similar geometric HIs can exchange partners and form new interactions despite the low sequence identity in their HIs. However, we have not discussed whether the HI used in the interaction is a "known" HI or a "new" one. Known HIs are those for which

FIG. 4. (a) Cumulative distribution of sequence similarity of the query's HI to template for TP (green) and FP (red) with respect to the HIPPIE dataset. (b) Probability density of sequence similarity scores of TP cases for new (<20% overlap with known HIs in black) and known (>80% overlap with known HIs in red) HIs.

FIG. 5. (a) Distribution of the overlap ratio of geometric HIs with known HIs. (b) Distribution of the overlapped patch sizes (amino acids).

there is a solved quaternary structure of the query protein (or are homologs with greater than 90% sequence identity) that uses that particular HI, and new HIs for a query protein are those not observed in any known multimeric structures of the protein. To address this issue, we calculated the residue overlap of the geometric HIs with the known ones. Overlap is defined as the number of same residues in a geometric, known HI divided by the size of the geometric HI. The overlap distributions (Figure 5(a)) show that there is no clear distinction, and all states are more or less populated. If we use the strict overlap values of zero for “new” and one for “known,” then 11% of cases fall into the “new” HI category and 16% into “known” ones. However, the majority of cases (73%) have partial overlap with known HIs. Thus, we define a new class and refer to them as “overlapped HIs.” For better illustration of these classes of HIs, we provide examples as shown in Figure 1. The HI copied from T_3 is a “known” HI for Q and the ones copied from T_1 and T_4 are “new,” while the HI from T_2 represents an “overlapped” HI.

Since a subset of overlapped HIs participate in at least two distinct interactions, it would be interesting to study the overlapped region in more detail. Figure 5(b) shows the size distribution of these overlapping patches, excluding the cases with no overlap. The size of these patches on average is 6 ± 7 (median of 10 residues) residues compared to the total size of the interface that is 18 ± 10 residues. We also measured the sequence conservation score (see Sec. II) for these patches for each query case and compared it to the residues in the known HIs for that query protein. Our result (Figure 6) shows that these patches are more conserved in 70% of cases when compared to the known HI fraction of the surface.

Interestingly, the overlapped HIs are not only observed when copying HIs from the distant homologs/analogs with low sequence identity but also for close homologs with high sequence identity. As shown in Figure S7 of the supplementary material,⁴⁷ new and overlapped HIs are observed to be independent of the sequence identity of the template from which they were copied. In addition, the same HI can be involved in forming both hetero and homo complexes for 40% of the cases. In other words, all or part of the same group of residues that form a homo-dimer complex can also participate

in forming a hetero-dimer complex (a complex is considered hetero if the proteins involved in the interaction have different GIs).

E. Current PDB complexes underrepresent charged interfaces

We have categorized validated geometric HIs into known, new, and overlapped HIs. It will be helpful to know whether there are any characteristic differences between these HIs, regarding HI sizes (Figure S8 of the supplementary material⁴⁷). The distribution of the HI sizes are similar between cases with low (less than 20% shown in black) and high (greater than 80% shown in red) overlap, with median of 14 and 17 amino acids, respectively. However,

FIG. 6. Conservation Z-score for the overlapped region of HI (y-axis) versus all of the HI (x-axis) for a given query. The probability density for each axis is plotted on each side.

TABLE II. Fraction of charged (ARG, LYS, GLU, ASP), hydrophobic acids (PHE, ILE, LEU, TRP, VAL, MET), and amino acids in different HIs. “New HIs” have more charged amino acids and are copied from template HIs with the same amino acid composition.

Type	Q ₀ ^a (new) (%)	T ₀ ^b (%)	Q ₁ ^c (known) (%)	T ₁ ^d (%)
Hydrophobic	18	21	26	26
Charged	38	35	29	30

^aQ₀: Predicted HIs for the query protein with no overlap with the known interfaces.

^bT₀: The HIs of the templates for Q₀.

^cQ₁: Predicted HIs for the query protein with complete overlap with known interfaces.

^dT₁: The HIs of the templates for Q₁.

the amino acid composition for the HIs changes significantly from more hydrophobic to more charged as the overlap with known HIs decreases. Focusing on the extreme cases of “new” (overlap = 0) and “known” (overlap = 1) HIs, the fraction of hydrophobic amino acids (PHE, ILE, LEU, TRP, VAL, MET) are 18% and 26%, respectively, whereas the values for charged amino acids (ARG, LYS, GLU, ASP) are 38% and 29%.

It gets more interesting when we trace back the compositions to the original templates from which they were copied. Table II shows that the templates of the new HIs have the same composition of charged and hydrophobic amino acids as the templates with known HIs. Moreover, Figure 4(b) shows that the same level of chemical similarity between the sequences is observed between cases with high (>80%, and shown in red) and low (<20%, and shown in black) overlap values. The results indicate that the most current known HIs in the PDB tend to favor hydrophobic more than charged HIs. This could be due to the fact that crystallizing complexes with charged HIs is more challenging due to the transient nature of the interaction and relatively smaller contact area.¹ While it is certainly possible that more charged HIs are less likely to form (implicit in earlier ideas regarding interfaces), the fact that known template HIs that generate new query HIs are more charged as well suggest that this is not the case.

F. Examples

In this section, we provide examples that illustrate the concepts of “new” and “overlapped” HIs. We superpose the structure of the query protein (color coded in pink) onto its template protein (color coded in blue). In all examples, the partner of the template protein is shown in green. While the blue-green complexes are the experimentally solved structure, the pink-green complexes are the predicted ones that are known to interact from the HIPPIE dataset. The first example (Figures 7(a) and 7(b)) shows an overlapped HI. Here, the full-length structure of human RPA14 (pink) that has the same fold (TM-score = 0.7) as RecQ2 (blue). While the global and the HI sequence identities are 0.16 and 0.0, respectively, in the HIPPIE dataset, RPA14 also interacts with RecQ1 (green), the partner of RecQ2. Here, the copied HI for RPA14 has 80% overlap with its already known crystal structure HIs, which classifies this HI as “overlapped.” Also, this example shows how the same set of amino acids (the overlapped region) could be involved in interacting with different partners.

FIG. 7. (a) The complex of RecQ2 (blue) and RecQ1 (green) (PDB_ID: 3MXN). (b) Structure of RPA14 in pink (PDB_ID: 2PI2) superposed on its template RecQ2 (blue). (c) The complex of ATG12 (blue) and ATG5 (green) (PDB_ID of 4GDK). (d) Structure of RPA14 in pink (PDB_ID: 2PI2) superposed on its template ATG12 (blue). (e) The complex of insulin receptor (blue) and tyrosine-protein phosphatase (green) (PDB_ID: 2B4S). (f) Structure of IGF1R in pink (PDB_ID: 1P4O) superposed on its template in blue. The pictures were rendered using Tachyon and VMD was used for visualization.^{65,66}

In a case of a novel HI (Figures 7(c) and 7(d)), the query protein (pink) is a human transport protein LC3C_8-125 (also known as RPA14) that has the same fold (TM-score = 0.8) as its template (blue) protein ATG12. The query and template proteins have global and HI sequence identities of 0.17 and 0.14, respectively. This is a novel HI since it has no overlap with any known HIs of the query. Despite the low HI sequence identity between RPA14 and ATG12, both have similar amino acid composition (29% charged amino acids and only 21% hydrophobic residues). This could indicate that when the geometric condition for a HI is met, there is more than one way of obtaining a favorable protein-protein interaction free energy.

To show that novel HIs are not limited to copying from distant analogs with low sequence identity, we present a third example in Figures 7(e) and 7(f). In this case, the sequence identity between the query and the template is high and yet the HI is novel. The query is an insulin receptor protein, IGF1R (pink), that shares a geometric HI with its template that is another type of insulin receptor protein, INSR (blue)

(TM-score = 0.87), with 79% global sequence identity and 50% sequence identity at the HI. This HI is 100% novel for IGF1R when compared to the existing known HI available in PDB, adding an extra 10% to the ratio of interface to surface amino acids.

IV. DISCUSSION AND CONCLUSIONS

Knowing the HIs involved in protein-protein interactions can greatly assist in determining the interaction pose of two proteins. According to currently available solved structures of multimeric complexes, on average, approximately one third of a protein's surface is known to form an interface in quaternary complexes. Thus, protein-protein interfaces were viewed as being special, and many efforts were made to predict these surface regions.^{16,56,57} In this study, we showed that the ratio of the interface to surface residues for currently known HIs is correlated (Pearson coefficient of 0.31 with P-value of 10^{-16}) with the number of experimentally solved structures for a given protein (Figure 3(a)). Next, we calculated the interface to surface ratio including "geometric HIs" by copying HIs from the same fold templates irrespective of their sequence identity. Surprisingly, more than three quarters of the surface of a protein can participate in PPIs. In this case, the ratio of interface to surface was correlated (Pearson coefficient of 0.35 with P-value of 10^{-16}) with the number of available templates with the same fold (Figure 3(b)). This ratio is a lower bound as we are only copying HIs from the same global fold templates with high local similarity in their HI. Copying the same geometric HIs from different folds could only increase this number.

For those cases in which both query and template were known to interact with the same partner (experimental evidence is provided by the HIPPIE dataset), we inferred that the same geometric HI has been used. Although the sequence identity in the HI with respect to its template is usually low, using a sequence similarity score, we showed that the same chemical environment is conserved when replacing the template protein with the query. The main assumption behind this analysis is as follows: if two proteins have highly similar folds and are known to interact with the same protein partner, it is highly likely that they use the same HI for this interaction. Further analysis shows that copied HIs fall into three categories: known, overlapped, and new HIs. Comparison of known with novel HIs shows that they have a similar size distribution but differ in their amino acid composition. These novel HIs have more charged amino acids compared to the known ones. Finally, in the case of overlapped HIs, we have shown that the amino acids in the overlapped regions are on average more conserved than the rest of the protein-protein binding sites.

We believe that the current low (~30%) interface to surface residue ratio is due to the relative paucity of experimentally solved multimeric structures. Even with the current library of quaternary structures, the ratio of interface to surface residues is high for some cases, as shown in the right tail of the green curve in Figure 2(b). These cases usually involve higher oligomeric states or have multiple quaternary structures deposited in the PDB. This suggests that as the size

of the library of multimeric structure increases, the number of HI residues in a protein grows as well. In a study by Bohnuud *et al.*,⁵⁸ it was shown that alternative binding modes exist for a given pair of proteins and that conformational selection happens through the formation of ligand binding sites. Another study by Gao and Skolnick⁵⁹ shows that a significant portion of disease-associated mutations in the human proteome are located on the portion of a protein's surface which are not covered by the currently known HIs. Therefore, both studies provide indirect evidence which suggest alternative protein HIs.

We propose that the poor performance of the current algorithms for predicting HIs reflects an underestimation of the actual interface to surface ratio due to the following: first, unlike catalytic sites that require specific interactions by specific amino acids to enable catalysis, PPIs tend to be more flexible in choosing which set of amino acids to use for a particular interaction. Second, much of the work on predicting HI's is based on the assumption that HI is predominantly hydrophobic and charged surface regions are unlikely to participate in protein-protein interactions. Our results show that many novel interfaces are charged, and *most importantly* such charged HIs are also seen in the known HIs from which they were copied. Thus, they are not an artifact of the conservation of HI geometric similarity. On the other hand, if we accept that the lower bound ratio of interface to surface residues is about three quarters, it is not reasonable to ask what were the interface residues for protein Q, as most surface residues likely participate in a HI. Rather, the question should be rephrased as to what set of amino acids were used in the interface for protein Q given that it interacts with protein P.

In conclusion, our results show that the majority of a protein's surface is geometrically capable of forming PPIs. We were able to partially validate some of the new HIs by cross-referencing to known HIs. For both known and geometric HIs, the fraction of interfacial to surface amino acids is correlated with the number of solved structures and available templates, respectively. In a given protein, HIs can overlap. These overlap regions (6-10 amino acids in size) tend to be more conserved than the rest of the HIs. Perhaps this reflects the fact that these regions have lower interaction free energies with their partners and thus tend to be more conserved. This will allow for a protein to interact with multiple partners while preserving the most favorable interaction region. This might be the origin of PPI hotspots that have been observed previously.⁶⁰⁻⁶³ However, we now conjecture that such hot spots are comprised of multiple as opposed to one or two residues. Finally, we believe that the classical concept of protein-protein binding sites, with just a few HIs per protein with clear borders, needs to be revisited. HIs likely involve a significant fraction of the protein's surface with a given sub-region, participating in multiple HIs. Thus, HIs are not unique but are just another generic feature of proteins. As shown previously, HIs occur in protein models that experience no selection for PPI.⁹ They are just a consequence of the packing of secondary structural elements that give rise to the library of small molecule ligand binding sites as well.⁶⁴ Thus, once again, HIs are also seen to be nothing special, rather they

are an intrinsic coarse-grained feature of the majority of a protein's surface that nature has taken advantage of.

ACKNOWLEDGMENTS

This research was supported in part by Grant No. GM-37408 of the Division of General Medical Sciences of the NIH.

- ¹I. M. Nooren and J. M. Thornton, *EMBO J.* **22**, 3486 (2003).
- ²I. M. Nooren and J. M. Thornton, *J. Mol. Biol.* **325**, 991 (2003).
- ³O. Keskin, A. Gursoy, B. Ma, and R. Nussinov, *Chem. Rev.* **108**, 1225 (2008).
- ⁴M. Gao and J. Skolnick, *Proc. Natl. Acad. Sci. U. S. A.* **109**, 3784 (2012).
- ⁵H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, *Nucleic Acids Res.* **28**, 235 (2000).
- ⁶D. Kihara and J. Skolnick, *J. Mol. Biol.* **334**, 793 (2003).
- ⁷Y. Zhang, I. A. Hubner, A. K. Arakaki, E. Shakhnovich, and J. Skolnick, *Proc. Natl. Acad. Sci. U. S. A.* **103**, 2605 (2006).
- ⁸J. Skolnick, A. K. Arakaki, S. Y. Lee, and M. Brylinski, *Proc. Natl. Acad. Sci. U. S. A.* **106**, 15690 (2009).
- ⁹M. Gao and J. Skolnick, *Proc. Natl. Acad. Sci. U. S. A.* **107**, 22517 (2010).
- ¹⁰R. Mosca, T. Pons, A. Ceol, A. Valencia, and P. Aloy, *Curr. Opin. Struct. Biol.* **23**, 929 (2013).
- ¹¹T. Rolland *et al.*, *Cell* **159**, 1212 (2014).
- ¹²G. T. Hart, A. K. Ramani, and E. M. Marcotte, *Genome Biol.* **7**, 120 (2006).
- ¹³K. Venkatesan *et al.*, *Nat. Methods* **6**, 83 (2009).
- ¹⁴M. P. Stumpf, T. Thorne, E. de Silva, R. Stewart, H. J. An, M. Lappe, and C. Wiuf, *Proc. Natl. Acad. Sci. U. S. A.* **105**, 6959 (2008).
- ¹⁵M. N. Wass, A. David, and M. J. Sternberg, *Curr. Opin. Struct. Biol.* **21**, 382 (2011).
- ¹⁶T. Aumentado-Armstrong, B. Istrate, and R. A. Murgita, *Algorithms Mol. Biol.* **10**, 7 (2015).
- ¹⁷Q. C. Zhang, D. Petrey, R. Norel, and B. H. Honig, *Proc. Natl. Acad. Sci. U. S. A.* **107**, 10896 (2010).
- ¹⁸M. Guharoy and P. Chakrabarti, *BMC Bioinf.* **11**, 286 (2010).
- ¹⁹J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li, and H. Jiang, *Proc. Natl. Acad. Sci. U. S. A.* **104**, 4337 (2007).
- ²⁰A. Baspinar, E. Cukuroglu, R. Nussinov, O. Keskin, and A. Gursoy, *Nucleic Acids Res.* **42**, W285 (2014).
- ²¹M. Sikic, S. Tomic, and K. Vlahovicek, *PLoS Comput. Biol.* **5**, e1000278 (2009).
- ²²H. Zellner, M. Staudigel, T. Trenner, M. Bittkowski, V. Wolowski, C. Icking, and R. Merkl, *Proteins* **80**, 154 (2012).
- ²³B. Q. Li, K. Y. Feng, L. Chen, T. Huang, and Y. D. Cai, *PLoS One* **7**, e43927 (2012).
- ²⁴S. Maheshwari and M. Brylinski, *Brief Bioinform.* **16**, 1025 (2015).
- ²⁵Y. Zhang and J. Skolnick, *Proc. Natl. Acad. Sci. U. S. A.* **101**, 7594 (2004).
- ²⁶J. Skolnick and M. Brylinski, *Briefings Bioinform.* **10**, 378 (2009).
- ²⁷J. K. Park, R. Jernigan, and Z. Wu, *Bull. Math. Biol.* **75**, 124 (2013).
- ²⁸L. Yang, G. Song, and R. L. Jernigan, *Biophys. J.* **93**, 920 (2007).
- ²⁹R. I. Dima and H. Joshi, *Proc. Natl. Acad. Sci. U. S. A.* **105**, 15743 (2008).
- ³⁰B. G. Pierce, K. Wiehe, H. Hwang, B. H. Kim, T. Vreven, and Z. Weng, *Bioinformatics* **30**, 1771 (2014).
- ³¹C. Dominguez, R. Boelens, and A. M. Bonvin, *J. Am. Chem. Soc.* **125**, 1731 (2003).
- ³²S. Mukherjee and Y. Zhang, *Structure* **19**, 955 (2011).
- ³³H. Chen and J. Skolnick, *Biophys. J.* **94**, 918 (2008).
- ³⁴N. Tuncbag, A. Gursoy, R. Nussinov, and O. Keskin, *Nat. Protoc.* **6**, 1341 (2011).
- ³⁵S. Viswanath, D. V. Ravikant, and R. Elber, *Methods Mol. Biol.* **1137**, 199 (2014).
- ³⁶J. Janin, K. Henrick, J. Moult, L. T. Eyck, M. J. Sternberg, S. Vajda, I. Vakser, and S. J. Wodak, *Proteins: Struct., Funct., Genet.* **52**, 2 (2003).
- ³⁷C. A. Rohl, C. E. Strauss, K. M. Misura, and D. Baker, *Methods Enzymol.* **383**, 66 (2004).
- ³⁸B. K. Vallat, J. Pillardy, P. Majek, J. Meller, T. Blom, B. Cao, and R. Elber, *Proteins* **76**, 930 (2009).
- ³⁹J. A. Capra and M. Singh, *Bioinformatics* **23**, 1875 (2007).
- ⁴⁰M. Brylinski and J. Skolnick, *Proc. Natl. Acad. Sci. U. S. A.* **105**, 129 (2008).
- ⁴¹A. Roy, J. Yang, and Y. Zhang, *Nucleic Acids Res.* **40**, W471 (2012).
- ⁴²H. Lu, L. Lu, and J. Skolnick, *Biophys. J.* **84**, 1895 (2003).
- ⁴³W. Li and A. Godzik, *Bioinformatics* **22**, 1658 (2006).
- ⁴⁴Y. Zhang and J. Skolnick, *Nucleic Acids Res.* **33**, 2302 (2005).
- ⁴⁵J. Xu and Y. Zhang, *Bioinformatics* **26**, 889 (2010).
- ⁴⁶M. Gao and J. Skolnick, *Bioinformatics* **29**, 597 (2013).
- ⁴⁷See supplementary material at <http://dx.doi.org/10.1063/1.4937428> for Figures S1–S8.
- ⁴⁸J. H. Lin, *IEEE Trans. Inf. Theory* **37**, 145 (1991).
- ⁴⁹S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, *Nucleic Acids Res.* **25**, 3389 (1997).
- ⁵⁰S. Henikoff and J. G. Henikoff, *Proc. Natl. Acad. Sci. U. S. A.* **89**, 10915 (1992).
- ⁵¹L. Lu, H. Lu, and J. Skolnick, *Proteins* **49**, 350 (2002).
- ⁵²P. Aloy and R. B. Russell, *Proc. Natl. Acad. Sci. U. S. A.* **99**, 5896 (2002).
- ⁵³M. H. Schaefer, J. F. Fontaine, A. Vinayagam, P. Porras, E. E. Wanker, and M. A. Andrade-Navarro, *PLoS One* **7**, e31826 (2012).
- ⁵⁴Y. C. Hwang, C. F. Lin, O. Valladares, J. Malamon, P. P. Kuksa, Q. Zheng, B. D. Gregory, and L. S. Wang, *Bioinformatics* **31**, 1290 (2015).
- ⁵⁵H. Zhou, M. Gao, and J. Skolnick, *Sci. Rep.* **5**, 11090 (2015).
- ⁵⁶S. Jones and J. M. Thornton, *J. Mol. Biol.* **272**, 133 (1997).
- ⁵⁷S. Jones and J. M. Thornton, *J. Mol. Biol.* **272**, 121 (1997).
- ⁵⁸T. Bohnuud, D. Kozakov, and S. Vajda, *PLoS Comput. Biol.* **10**, e1003872 (2014).
- ⁵⁹M. Gao, H. Zhou, and J. Skolnick, *Structure* **23**, 1362 (2015).
- ⁶⁰K. S. Thorn and A. A. Bogan, *Bioinformatics* **17**, 284 (2001).
- ⁶¹A. A. Bogan and K. S. Thorn, *J. Mol. Biol.* **280**, 1 (1998).
- ⁶²O. Keskin, B. Ma, and R. Nussinov, *J. Mol. Biol.* **345**, 1281 (2005).
- ⁶³T. Kortemme and D. Baker, *Proc. Natl. Acad. Sci. U. S. A.* **99**, 14116 (2002).
- ⁶⁴J. Skolnick and M. Gao, *Proc. Natl. Acad. Sci. U. S. A.* **110**, 9344 (2013).
- ⁶⁵J. Stone, "An efficient library for parallel ray tracing and animation," thesis, University of Missouri, 1998.
- ⁶⁶W. Humphrey, A. Dalke, and K. Schulten, *J. Mol. Graphics* **14**, 33 (1996).