Journal of
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

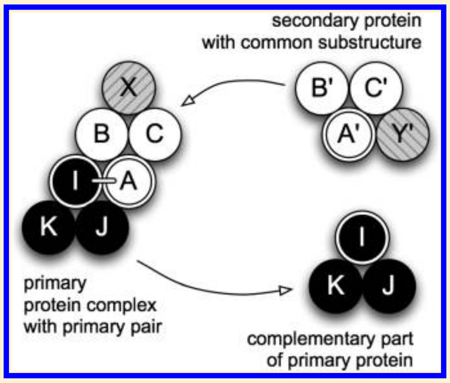
ProPairs: A Data Set for Protein−Protein Docking

Florian Krull, Gerrit Korff, Nadia Elghobashi-Meinhardt, and Ernst-Walter Knapp*

Institute of Chemistry and Biochemistry, Freie Universität Berlin, Fabeckstrasse 36a, 14195 Berlin, Germany

S *Supporting Information*

ABSTRACT: ProPairs is a data set of crystal structures of protein complexes defined as biological assemblies in the protein data bank (PDB), which are classified as legitimate protein−protein docking complexes by also identifying the corresponding unbound protein structures in the PDB. The underlying program selecting suitable protein complexes, also called ProPairs, is an automated method to extract structures of legitimate protein docking complexes and their unbound partner proteins from the PDB which fulfill specific criteria. In this way a total of 5,642 protein complexes have been identified with 11,600 different decompositions in unbound protein pairs yielding legitimate protein docking partners. After removing sequence redundancy (requiring a sequence identity of the residues in the interface of less than 40%), 2,070 different legitimate protein docking complexes remain. For 810 of these protein docking complexes, both docking partners possess corresponding unbound structures in the PDB. From the 2,070 nonredundant protein docking complexes there are 417 which possess a cofactor at the interface. From the 176 protein docking complexes of the Protein−Protein Docking Benchmark 4.0 (DB4.0) data set, 13 differ from the ProPairs data set. Twelve of them differ with respect to the composition of the unbound structures but are contained in the large redundant ProPairs data set. One protein docking complex of the DB4.0 data set is not contained in ProPairs since the biological assembly specified in the PDB is wrong (PDB id 1d6r). For one protein complex (PDB id 1bgx) the DB4.0 data set uses a fabricated unbound structure. For public use interactive online access is provided to the ProPairs data set of nonredundant protein docking complexes along with the source code of the underlying method [http://propairs.github.io].



■ INTRODUCTION

Protein−protein interaction is the basic mechanism underlying signal transduction and regulation of transcription and metabolic processes in the living cell.[1] To understand the functional role of protein−protein docking events in the living cell, knowledge of the docking geometries of protein complexes may not be necessary. Knowledge of the network of protein−protein encounters in the living cell studied by system biology is extensive,[2−5] but it is nevertheless notoriously incomplete. If the structures of the individual players - the proteins - are known, methods identifying which proteins interact by forming the corresponding complexes can be useful.[6] These methods are even more important if one is interested in targeting signal transduction and regulation with specific drugs. Here, one needs to know the drug targets in more detail.[7,8] Drugs that bind at the interface of the interacting proteins that form a complex may be particularly effective, but designing such drugs requires precise knowledge of the protein−protein interface.[9−12] Particularly critical are functional protein complexes that form only transiently and are therefore very difficult to characterize structurally. For such systems, the chances of finding crystal structures of the unbound proteins that form the prospective metastable protein complex are higher than the chances of finding the structure of the protein complex itself. Predicting protein docking geometries based on the unbound protein structures and simulating such docking events can fill this information gap and is therefore an important research area.[13−19]

The prediction of protein complex geometries from knowledge of the individual unbound protein structures requires test cases[20,21] which involve not only the protein docking complexes but also the corresponding unbound protein structures. These test cases are not only valid for empirical methods that need large databases but also for traditional methods that simulate the docking complex with physical force fields[22−24] or analogous modern methods, which employ a combination of physical force fields and specific contact information obtained experimentally.[25,26]

Alternatively, there are numerous approaches to identify protein docking partners using machine learning methods. These usually require a large body of protein docking complexes for which structure and contact region (interface) are precisely defined. Although the Protein Data Bank (PDB)[27] contains many protein complexes, it is not directly clear which ones are also protein docking complexes, i.e. whether a decomposed state of the complex also exists under physiological conditions. Biochemical-based data would be most helpful for identifying the appropriate decompositions of protein complex structures. However, such information is spread over a wide range of literature and difficult to collect.

Received: February 13, 2015
Published: June 2, 2015

DOI: 10.1021/acs.jcim.5b00082
J. Chem. Inf. Model. 2015, 55, 1495−1507

Alternatively, one can use information about the availability of (unbound) protein structures in the PDB,[27] which correspond to a decomposition of the targeted protein complex, where one can use the PDB's information regarding the biological assembly.[28] The latter method of verifying a specific decomposition of a protein complex has the advantage that one also directly obtains the structures of the unbound partner proteins. These allow to build a realistic prediction scenario based on known unbound protein structures. Similar strategies were used to generate structure data sets of protein docking complexes including the corresponding unbound structures that were hand- or semiautomatically selected[29] from the PDB[27] database. Alternatively also fully automatic selection procedures[30,31] were applied to the PDB database. The purpose of the present contribution is to introduce an up-to-date, large, and reliable set of protein docking complexes with unbound partner proteins. Although this set is generated by an automatic procedure, it nevertheless includes practically all structures of much smaller hand curated data sets. Although we applied much care, we would like to emphasize that the ProPairs data set may contain incorrect data, as is not uncommon for large and complex data sets.

■ METHODS

A. General Considerations and Definitions. To establish a benchmark set for protein−protein docking we aim to construct a set of protein docking partners. These should involve known three-dimensional structures of the isolated proteins and the corresponding protein complex such that one can assume the protein complex forms by a protein docking procedure. Note that different protein docking complexes may relate to the same protein complex identified by its PDB id, if the complex can be composed by different pairs of unbound structures involving different interfaces. For this purpose we first explore the PDB (containing 95,280 protein structures in November 2013) to collect protein complex structures that are identified by involvement of several polypeptide chains, where at least two are larger than 19 amino acids yielding 44,088 protein complex structures. We analyze these multichain protein complex structures according to their biological assemblies as given in the PDB[27] considering crystal but not NMR structures. If several biological assembly structures are listed only the top one is considered.[32]

Homodimeric protein structures may exist occasionally only under specific crystallographic conditions and not in solution. To avoid artifacts from such protein complexes, homodimers are ignored completely if the corresponding monomer consists of a single polypeptide chain. However, if the monomer of a homodimeric protein consists of several polypeptide chains, the monomer can still be a candidate for protein docking and is considered. Thus, we reduce the number of potential protein docking complexes from 44,088 to 31,669. From this set we aim to select protein complexes possessing the following features: (i) the complexes are easily separable in two independent domain structures which exist as independent proteins; (ii) the complexes form well-defined contact surfaces (interfaces); (iii) the complexes consist of different polypeptide chains (i.e., no polypeptide chain is cut).

The interface of a protein docking complex is regarded as a special case of a contact surface. A contact surface (used in the Method section parts E and F) can exist between two chains of the same docking partner, while in contrast an interface (used in this and the following paragraphs) is a contact surface between two legitimate docking units. Residues are considered to be part of a contact surface if the distance ($d_{\text{res-res}}$) of at least one pair of heavy atoms belonging to residues on different sides of the contact fulfills $d_{\text{res-res}} < 10$ Å.[33] Such contact surface residues can constitute a legitimate interface of a protein docking complex and are thus labeled interface residues. A more precise definition of a legitimate interface follows in part E.

To fulfill conditions (i)-(iii) for a pair of proteins forming a legitimate protein docking complex with known structure, we require that, at least for one of the proteins forming the complex, a corresponding unbound structure is also available in the PDB.[27] In contrast to protein complex structures, we also consider NMR data for the unbound structures. Protein complexes that contain complicated interfaces, since different polypeptide chains penetrate each other, are likely excluded if the corresponding unbound structures exist. For a multichain protein complex different modes of decomposition in two pairs of proteins are possible; one or even both of the individual proteins of the decomposition may still involve several polypeptide chains, rendering this problem more difficult than for two-chain protein complexes. An example is shown schematically in Figure 1. We define the identity of a protein docking complex with the help of the interface formed by the protein docking partners, as will be explained in more detail in the following.

B. The Role and Types of Nonproteogenic Compounds in Protein Interfaces. Protein structures may contain cofactors or involve bound DNA or RNA. If DNA or

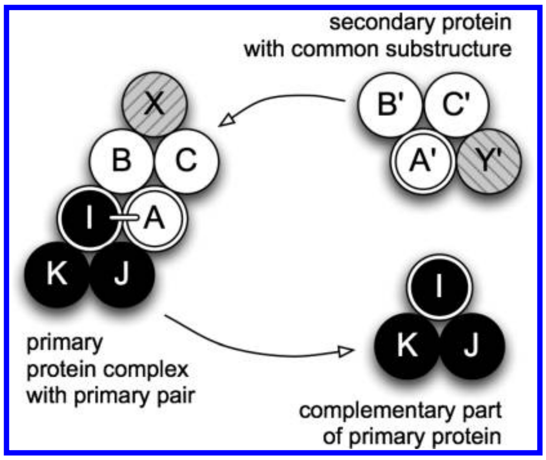


Figure 1. Exploring the potential decomposition of a primary protein. Starting from a primary pair of polypeptide chains (chains A and I), which have a contact surface in common, one can identify by sequence similarity a secondary protein in the PDB[27] which can be a candidate of an unbound protein structure. The secondary protein must possess a polypeptide chain (A′) that is sequentially similar to chain A with maximum common connected substructure (CCsub) (i.e., A′, B′, C′ are similar to A, B, C of the primary protein). The chains X and Y′ are dissimilar (sequence identity less than 40%) and are therefore not part of the CCsub. Thus, chain X is removed from the protein complex, while chain Y′ is kept, since it would correspond to a realistic prediction scenario. In the present example the resulting complete interface is formed by the two sets of chains (A, B) and (I, J). Hence, the polypeptide chains (A, B, C) constitute the first binding partner and (I, J, K) form the complementary second binding partner of the primary protein complex. For the second binding partner, an unbound protein can be found with the same procedure starting from the polypeptide chain I or J.

RNA is located at the potential interface of a docking complex, we do not consider this interface. Similarly, we ignore unbound protein structures that have DNA or RNA in the interface region. There are special databases that consider docking between proteins and DNA[34] or RNA.[35] Cofactors are smaller nonpeptidic compounds like cations, heme, GTP, or others. They can have a significant influence on protein−protein docking if they are localized at the interface and must be considered in this case. Cofactors that are not at the interface are ignored in the selection procedure for unbound structures although they are not removed from the complex structures in the ProPairs data set. Cofactors like GMP and AMP are considered to be equivalent [a list of equivalent cofactors can be found in the Supporting Information, Table S1]. Some cofactors, like H_2O, SO_4^{2-}, or Cl^- [listed in the Supporting Information, Table S2], are often found in protein crystal structures. Their presence at the interface is not generally connected with a specific biochemical function. Therefore, these cofactors are ignored even if they occur at the interface.

C. Protein Sequence Alignments. To find appropriate isolated protein structures which are part of a protein docking complex a pairwise sequence alignment is performed across all polypeptide chains belonging to proteins stored in the PDB.[27] In this procedure we ignore polypeptide chains that contain fewer than 20 amino acids, since small polypeptide chains often do not possess a definite three-dimensional structure that can undergo meaningful sequence alignment. A larger threshold would also be useful. However, in order not to miss any of the protein docking complexes (for instance the protein with PDB id 1ppe) from the Protein−Protein Docking Benchmark 4.0 (DB4.0)[29] involving a polypeptide of length 29 residues we used the chain length of 20 residues. The PDB[27] contains protein structures involving $N_{\text{chain}} = 2.2 \times 10^5$ polypeptide chains (November 2013) resulting in $(N_{\text{chain}}^2 - N_{\text{chain}})/2 = 2.4 \times 10^{10}$ sequence alignment tasks. The alignment is carried out with the Smith-Waterman algorithm[36] which finds the optimal local sequence alignment using standard scoring parameters. From the pairwise sequence alignment of polypeptide chains n_1 and n_2 involving r_1 and r_2 residues, respectively, we obtain the relative sequence identity, $S_{\text{chain}}(n_1, n_2)$, which is the number of matching residues of a sequence alignment divided by the number of residues of the shorter polypeptide chain $\min(r_1, r_2)$. The results of the sequence alignments were also used in a subsequent step to remove redundant data.

D. Polypeptide Chains Forming a Contact. To identify a contact between two polypeptide chains of a protein complex we first introduce residue pair contacts of residues belonging to different chains. Two such residues are in contact if at least two non-hydrogen atoms of these residues are separated by a distance $d_{\text{res-res}} < 5.5$ Å. Two polypeptide chains are in contact if, for each chain, larger patches of residues form such pair contacts with the other chain. More precisely the product of the number of residues - involved in such pair contacts - from one chain with the other chain must be larger than 24.[37] We search the PDB[27] for protein complexes for which at least two polypeptide chains are in contact and call them henceforth primary protein complexes. Two polypeptide chains in the primary protein complex that are in contact are called the primary pair (chains A and I, Figure 1). The interface for a specific decomposition of a primary protein complex in two parts can generally involve several polypeptide chains for each of the two resulting potential binding partners (Figure 1). To avoid the combinatorial explosion that would occur if we allowed all possible decompositions of the primary protein complex in two potential binding partners, we first identify all pairs of polypeptide chains that are in contact.

E. Finding Unbound Protein Structures That Fit in a Protein Complex. The validity of a protein complex to serve as a legitimate docking complex is verified as follows. We consider all pairs of polypeptide chains in a primary protein complex that are in contact as defined in section D. The two chains of such a primary pair potentially belong to different binding partners of a protein complex such that the contact surface of the primary pair would be part of a legitimate interface (see for instance A and I in Figure 1). For these two polypeptide chains we try to find, at least for one chain (A or I), a counterpart (for instance A′ for A in Figure 1) with more than 70% sequence identity in a secondary protein (consisting of chains A′, B′, C′, Y′ in Figure 1) of the PDB,[27] which may serve as an unbound structure after several further checks. One of these checks is to require that from the secondary protein no chain is similar to chains of the other binding partner of the primary pair, i.e. the sequence identity should be less than 40%. As a consequence, chain I is sequentially dissimilar to the chains A′, B′, C′, Y′ in Figure 1. This requirement prevents us from accidentally considering two pairs of a homodimeric multichain protein that are essentially identical.

Our goal is to establish the largest common substructure of polypeptide chains that belongs to the two proteins and that involves chains A and A′, respectively. The chains A, B, C of the primary protein complex and A′, B′, C′ of a potential unbound, secondary protein are similar, while the chains X and Y′ are dissimilar, see Figure 1. We further require that the polypeptide chains of the common substructure are mutually in contact as defined before (section D) and thus label it the common connected substructure (CCsub). This CCsub may, for both proteins (primary and secondary), involve only a fraction of all polypeptide chains (X of the primary and Y′ of the secondary protein do not belong to the CCsub in Figure 1). A CCsub is identified by checking the similarity of the two proteins with respect to sequences of all polypeptide chains and of the residues in their pairwise contact surfaces. We require that the polypeptide chains (chain identifier n_1) of the primary protein complex ($i = 1$) belonging to the CCsub have a sequence identity of

$$S_{\text{chain}}(n_1, n_2) \geq 0.7 \tag{1}$$

with a corresponding chain (chain identifier n_2) of the secondary protein ($i = 2$) (Figure 1).

F. Similarity of the Contact Surface of Pairs of Polypeptide Chains. For residues in the contact surfaces of two polypeptide chain pairs (n_1, m_1 and n_2, m_2) belonging to a primary ($i = 1$) and secondary protein ($i = 2$), we evaluate the contact surface similarity $S_{\text{2-inter}}$ by first computing the relative sequence identities $S_{\text{1-contact}}(n_1, n_2)$ and $S_{\text{1-contact}}(m_1, m_2)$ of the one-sided contact surface (1-contact means it comprises the surface of only one of the contact partners) of the corresponding chain pairs (n_1, m_1 and n_2, m_2) in the two proteins ($i = 1, 2$). $S_{\text{1-contact}}(n_1, n_2)$ is defined as the number of identical contact surface residues in chains n_1 and n_2 divided by the smaller number of contact surface residues in chains n_1 and n_2. The contact surface similarity measure of the two chain pairs (n_1, m_1 and n_2, m_2) is defined as two-sided contact surface similarity (2-contact means it comprises the surfaces of both contact partners; see also Figure 2)

$$S_{2\text{-contact}}(n_1, n_2, m_1, m_2) = S_{2\text{-sided}}[S_{1\text{-contact}}(n_1, n_2), S_{1\text{-contact}}(m_1, m_2)] \tag{2a}$$

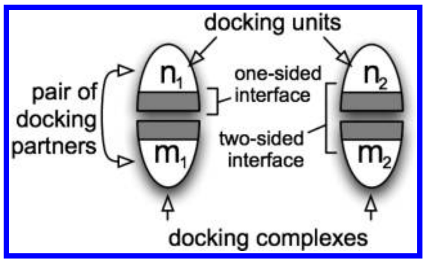


Figure 2. Schematic structure of two protein docking complexes 1 (n_1, m_1) and 2 (n_2, m_2). Each one consists of a pair of docking partners which are bound together. Individual docking partners are called docking units. Each docking unit possesses a one-sided interface that can get in contact with the suitable docking partner. The intact protein docking complex possesses a two-sided interface consisting of two one-sided interfaces that are in contact. In the following we refer to two-sided interfaces simply as interfaces, while the remainder are referred to as one-sided interfaces if applicable.

which uses the function

$$S_{2\text{-sided}}(u, v) = [2u - u^2] \times [2v - v^2] \tag{2b}$$

The behavior of eq 2b is shown in Figure S1 of the Supporting Information. In the subsequent discussion, we will use the contact surface similarity measures, eqs 2a and 2b, referring above to pairs of chains. We also consider the more general case of interfaces that are formed by several chains for each of the two one-sided interfaces that make up the complete two-sided interface.

For two chain pairs (n_1, m_1 and n_2, m_2) having a similar contact surface we require

$$0.5 \leq S_{2\text{-contact}}(n_1, n_2, m_1, m_2) \tag{3}$$

Hence, the requirement for contact surface similarity is set lower than for simple chain similarity [$S_{\text{chain}}(n_1, n_2)$, eq 1] to account for ambiguities in the definition of contact surface residues. For instance, in the symmetric case $S_{1\text{-contact}}(n_1, n_2) = S_{1\text{-contact}}(m_1, m_2)$ the similarity measure, eq 3, is relaxed, since it is sufficient to require that $S_{1\text{-contact}}(n_1, n_2) \geq 0.46$ to fulfill $0.5 \leq S_{2\text{-contact}}(n_1, n_2, m_1, m_2)$, as compared to simple chain similarity, for which we require $S_{\text{chain}}(n_1, n_2) \geq 0.7$, eq 1. However, in an asymmetric case the inequality (3) strongly penalizes dissimilar sequences. For instance with $S_{1\text{-contact}}(n_1, n_2) < 0.29$ the inequality (3) cannot be fulfilled even if the second chain pair (m_1, m_2) has maximum similarity $S_{1\text{-contact}}(m_1, m_2) = 1.0$.

G. Construction of a Complete Interface of the Primary Protein Complex and Identification of Corresponding Unbound Structures. We construct the complete interface of a protein complex decomposition from knowledge of a primary pair of polypeptide chains (for instance chains A and I in Figure 1) and knowledge of the CCsub in a secondary protein containing one (for instance A′ in Figure 1) of the two (A or I) polypeptide chains of the primary pairs (sequence identity between A and A′ is above 70%). To identify the chains belonging to the other potential binding partner, we start with the polypeptide chain (I) of the primary pair which does not belong to the CCsub (first binding partner) in the primary protein complex (Figure 1) and merge it successively with all polypeptide chains that are mutually in contact. To exclude the possibility that homodimeric primary protein complexes are also identified as unbound secondary protein by our algorithm, we require that the chains (J, K) merged with chain I have a sequence identity of less than 40% with any chain (A, B, C) in the CCsub (Figure 1). This set of mutually connected (being in contact as defined in section D) polypeptide chains (I, J, K) constitutes the complementary part of a decomposition of the primary protein, while the first part is defined by the CCsub. All chains of the primary protein that do not belong to the CCsub or to the complementary part are henceforth ignored. Accordingly, chain X of the primary protein (Figure 1) is ignored, since is has no match in the unbound structure and is therefore a spectator in the resulting decomposition of the primary protein. We call this ensemble of polypeptide chains the reduced primary protein complex (A, B, C, I, J, K in Figure 1). With this step we decomposed the primary protein complex in two parts with a corresponding specific interface. For this decomposition, we have found so far one unbound candidate protein (i.e., the CCsub of the secondary protein, A′, B′, C′, Y′) corresponding to the first part of the decomposed primary protein complex. Chain Y′ of the unbound structure is a spectator like chain X of the primary protein, not directly involved in the identified interface (Figure 1). However, when using unbound proteins in a realistic prediction scenario this information is not available. Hence spectator chains are kept in unbound structures. In case no unbound structure is found for the complementary part of the primary protein no spectator chain is identified and all chains (I, J, K, in Figure 1) are considered to constitute the binding partner. If an unbound structure for the complementary part of the primary protein is also found, spectator chains of the complementary part can be identified and are eliminated.

We still need to check whether, in the potential unbound secondary protein, the interface of the CCsub is not covered by other polypeptide chains belonging to the secondary protein. For this purpose, we first perform sequence alignments to identify the residues of the secondary protein that correspond to the potential interface residues of the primary protein complex. Based on this information we can structurally superimpose the C_α atom pairs of the corresponding interface residues from the primary and secondary protein using the Kabsch algorithm.[38] We thus obtain a one-to-one assignment of equivalent residues belonging to the interface patch of the primary protein complex and the unbound secondary protein. However, we discard unbound secondary proteins that have less than 55% interface residues in common with the primary protein.

A secondary protein is considered to be a legitimate unbound structure if all polypeptide chains that belong to the CCsub (A′, B′, C′ in Figure 1) or are connected with it (for instance Y′ in Figure 1) do not clash with the complementary part of the primary protein complex (i.e., chain A′, B′, C′, Y′ should not collide with any of the chains I, J, K in Figure 1). A polypeptide chain is considered to collide with another protein if more than 25% of its C_α atoms are closer than 5 Å to any C_α atoms of the other protein. This relative tolerant value of 25% is chosen so that unbound structures exhibiting a larger conformational difference to the bound counterpart are not automatically excluded. The corresponding protein docking complexes constitute the more difficult docking problems. Repeating this procedure for all primary polypeptide chain pairs (as for instance described for A-I in Figure 1) with a common interface may yield the same decomposition several times and may also yield unbound secondary proteins for the complementary part of the primary protein complex. With this procedure a primary protein complex consisting of more than two polypeptide chains may be decomposed in more than one way. We consider all decompositions of the primary protein complexes found with the above procedure as legitimate protein partners for protein docking, for which a corresponding legitimate unbound protein is available at least for one partner. Using these criteria, 11,600 different legitimate pairs of protein docking partners were identified from the PDB. For each of these docking partners more than one equivalent unbound structure may be available. All of them are incorporated in the large ProPairs data set of 11,600 protein complexes.

H. Similarity Measure of One-Sided Interface Surfaces of Two Docking Units. We discriminate between protein pairs that are docking partners and protein pairs that are only docking units. The former pair always constitutes a legitimate protein docking complex, whereas docking units can also belong to different protein docking complexes (see Figure 2). Each docking unit possesses at least one one-sided interface surface that may involve interface residues that belong to more than one polypeptide chain. In a protein docking complex each docking partner possesses such a one-sided interface surface (see Figure 2). To remove redundancies between protein complexes with similar interfaces, we compare one-sided interfaces of docking units pairwise. Two docking units (1 and 2) have a total of r_1 and r_2 residues belonging to N_1 and N_2 polypeptide chains in the one-sided interfaces, where the chains are denoted by ${n_1}^{(j)}$, $j = 1, . . N_1$ and ${n_2}^{(k)}$, $k = 1, . . N_2$ and are collected in vector notation $\vec{n}_i = (n_i^{(1)}, n_i^{(2)}, \ldots n_i^{(N_i)})$. With these definitions, we can express the maximum sequence similarity of the one-sided interfaces (1-inter) of two docking units 1 and 2 (for a definition see Figure 2) assuming that $N_1 \leq N_2$ according to

$$S_{1\text{-inter}}(\vec{n}_1, \vec{n}_2) = \frac{1}{\max(r_1, r_2)} \max_{\pi}\left(\sum_{i=1,N_1} \hat{s}_{1\text{-inter}}(n_1^{(i)}, n_2^{(\pi(i))}) \right) \tag{4}$$

where $\hat{s}_{1\text{-inter}}\ (n_1^{(i)}, n_2^{(\pi(i))})$ is the number of residue identities between the polypeptide chains ${n_1}^{(i)}$ and ${n_2}^{(\pi(i))}$. In eq 4, $\pi(i)$ is a specific selection of N_1 integers out of the $i = 1, . . N_2$ integers and with $\max_\pi$ all combinations are used. In the case of $N_1 = N_2$ π denotes permutations. Thus, we guarantee that all combinations of chain assignments between the two docking units (1 and 2) are considered.

RESULTS

A. Removing Redundancies of Protein Docking Complexes and Unbound Proteins. It is important to remove redundancies of protein docking complexes to avoid biasing of structures that are over-represented in the PDB.[27] Redundancies can be identified by structure or sequence similarity, or by a combination of both, or by the physicochemical composition.[39−43] In the Protein−Protein Docking Benchmark 4.0 (DB4.0)[29] redundancies were avoided by considering only protein complexes belonging to different protein families according to the criteria used by SCOP.[44,45] With the database DOCKGROUND[30,31] one can use protein families or sequence identities of the whole polypeptide chains to control redundancies. ProPairs uses the sequence identity in the interface to remove redundancies as outlined in detail in the following sections. There are two types of redundancies: (i) among the 11,600 protein docking partners found in the PDB[27] similar interfaces may exist; (ii) for one binding partner (or even both) of a protein docking pair more than one appropriate unbound structure may be found in the PDB.[27] In the following we discuss the removal of both types of redundancies in more detail.

B. Representative Protein Docking Complexes. In the present application redundancies of protein docking complexes with high interface similarity [as defined in eq 2a of the Method section part F for chain pairs] are removed using appropriate sequence information. We first remove redundancies between pairs of protein docking partners. For this purpose, we group protein docking partners with respect to their interface similarity forming clusters of interfaces using the single-linkage cluster algorithm[46] with the distance function $d_{\text{similar}} = 1 - S_{2\text{-inter}}$. The two-sided interface similarity $S_{2\text{-inter}}$ between two distinct pairs ($i = 1, 2$) of protein docking partners, each one consisting of a pair of docking units $(\vec{n}_1, \vec{m}_1)$ and $(\vec{n}_2, \vec{m}_2)$, is defined as

$$\begin{aligned} S_{2\text{-inter}}(\vec{n}_1, \vec{m}_1, \vec{n}_2, \vec{m}_2) = \max[S_{2\text{-sided}}(S_{1\text{-inter}}(\vec{n}_1, \vec{n}_2), \\ S_{1\text{-inter}}(\vec{m}_1, \vec{m}_2)), S_{2\text{-sided}}(S_{1\text{-inter}}(\vec{n}_1, \vec{m}_2), S_{1\text{-inter}}(\vec{n}_2, \vec{m}_1))] \end{aligned} \tag{5}$$

$S_{1\text{-inter}}$, defined in eq 4, describes the sequence similarity between two one-sided interfaces belonging to protein complexes 1 and 2, respectively. We use $S_{2\text{-sided}}(\text{u, v})$, defined by eq 2b, to combine the two similarities of four one-sided interfaces to obtain the similarities of two two-sided interfaces. However, in contrast to the previous usage of eq 2b we also consider here the cross combinations of the docking units $(\vec{n}_1, \vec{m}_2$ and $\vec{n}_2, \vec{m}_1)$. We use the maximum sequence identity of both combinations as our measure of similarity $S_{2-\text{inter}}$ between two protein docking complexes.

To control the number of different clusters of interfaces, we require that at least one pair of protein docking partners belonging to the same cluster should have an interface similarity obeying the inequality

$$S_{2\text{-inter}} \geq 0.4 \tag{6}$$

As a result, from the 11,600 different pairs of protein docking partners, 2,070 clusters of interfaces were generated. More details on the number of protein complexes and their interfaces are listed in Table 1. The PDB ids of the protein docking

Table 1. Number of Protein Docking Complexes and Their Corresponding Interfaces and Cofactors

complete (redundant) set	occurrence
number of legitimate interfaces	11, 600
number of legitimate interfaces with no redundancies due to symmetry	7,035
number of protein complexes with at least one legitimate interface	5,642
legitimate interfaces with cofactors in the interface region	3,605
with only one cofactor	2,190
representative (nonredundant) set	occurrence
number of interfaces	2,070
only for one binding partner an unbound structure is assigned	1,260
both binding partners possess unbound structures	810
average *i*RMSD where both unbound structures are available	1.88 Å
interfaces with cofactors	417
interfaces with cofactors where for only one binding partner an unbound structure is assigned	274

complexes belonging to the different clusters can be obtained from the Web page [http://propairs.github.io]. In Figure 3 the distribution of cluster sizes are displayed. The six largest clusters are listed in Table 2.

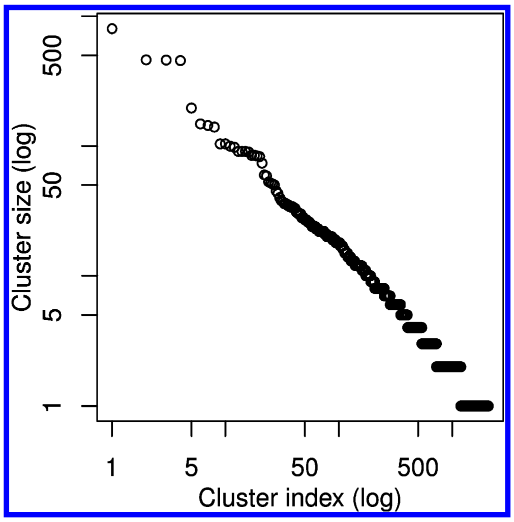


Figure 3. Size distribution of clusters containing similar interfaces of protein docking partners. 2,070 of such clusters were generated from the 11,600 pairs of docking partners. A cluster index is assigned according to its size to obtain a monotonic dependence of cluster sizes. The cluster indexes are plotted against the cluster sizes. The largest cluster (with index 1) consists of 801 members (Table 2); 882 clusters contain only one interface.

Table 2. Six Largest Clusters of Redundant Protein Docking Complexes Contained in the Large ProPairs Data Set of 11,600 Complexes

size	category of interaction partners
801	hemoglobin alpha chain/beta chain
462	ATCase catalytic chain/regulatory chain
460	MHC alpha chain/microglobulin
456	antibody heavy chain/light chain
196	insulin alpha chain/beta chain
148	trypsin/trypsin inhibitor

From each cluster of similar protein docking partners we choose one representative member that is assigned to the benchmark set of protein docking partners according to criteria which are applied in the following sequential order: (i) it has the largest number of residues in the interface of the two protein docking partners; (ii) it has the minimum number of residues belonging to gaps that are localized at the interface; (iii) it has the minimum number of polypeptide chains belonging to the interface of the two protein docking partners; (iv) it has the lowest number of polypeptide chains which are not part of the common interface; (v) it has the highest similarity to the other members of the same cluster.

C. Representative Unbound Proteins for Protein Docking Complexes. So far, the protein docking partners which enter the benchmark set may possess more than one appropriate unbound protein structure that is similar to one of the two protein docking partners. For each protein docking partner, these unbound protein structures are now ranked and selected according to criteria that are applied in the following sequential order: (i) they consist of the smallest number of polypeptide chains; (ii) they have the minimum number of residues belonging to gaps corresponding to residues in the interface of the primary protein; (iii) they have the maximum number of residues corresponding to the interface of the primary protein; (iv) they possess a minimum number of cofactors located in the interface regime that do not appear at the interface of the protein docking complex (These cofactors are subsequently removed from the unbound protein to obtain a legitimate structure.); (v) they have the lowest interface C_α atom root-mean-square deviation (*i*RMSD) relative to the corresponding part of the primary protein complex. The latter requirement is also one of the major quality criteria used to classify a prediction result in the CAPRI contest.[33] As a result, for each pair of the 2,070 protein docking partners without cofactors at the interface, at least one docking partner possesses an appropriate unbound protein structure. The selection procedure of unbound proteins corresponding to primary proteins with cofactors at the interface is described below.

D. Cofactors in the Interface Region of the Primary Protein and Similarity of Cofactors in Primary and Secondary Protein. Cofactors located at the interface can play a significant role in protein–protein docking. Therefore, we register the cofactor compositions located at the interface of the nonredundant primary proteins. We consider a cofactor to be located in the interface region of a primary protein if it is in contact with at least three interface residues, regardless of which part of the decomposed primary protein the residues belong to. As for residue pair contacts, we consider a cofactor-residue contact to exist if the distance between any pair of heavy atoms of cofactor and residue fulfills $d_{\text{cof-res}} < 5.5$ Å.

To find corresponding unbound structures that possess a cofactor similar to the one present at the interface of the primary protein, we first need to define cofactor similarity. We consider two cofactors ($c_{1,}$ c_2) to be chemically similar if they fulfill one of three criteria: (i) they have the same name assigned in the PDBbank;[27] (ii) they belong to the same group of cofactors that we have defined (see Table S1 in the Supporting Information); (iii) they have a similar composition of atom types [$\text{Sim}_{\text{cof}}(c_1, c_2) > 0.7$, eq 7]. For two cofactors c_1 and c_2 the similarity of the atomic composition is defined as

$$\text{Sim}_{\text{cof}}(c_1, c_2) = 1 - \frac{\sum_i |N_i^{type}(c_1) - N_i^{type}(c_2)|}{\sum_i [N_i^{type}(c_1) + N_i^{type}(c_2)]} \tag{7}$$

where $N_i^{type}(c_j)$ is the number of atoms of cofactor c_j that belong to type *i*.

E. Equivalence of Cofactor Positions at the Interface of Primary Protein and Unbound Protein. We consider only cofactors in the unbound protein that are in the corresponding interface patch, i.e. the distance between any pair of heavy atoms of cofactor and at least two interface residues fulfills $d_{\text{cof-res}} < 6$ Å. This cutoff distance is slightly larger than the value used for the primary protein complex (5.5 Å) since we do not want to miss cofactors in the interface patch of the unbound protein. We now have identified pairs of primary proteins and unbound proteins that have equivalent cofactors in their interface regions.

Next, for each cofactor at the interface of the primary protein we try to assign a similar cofactor in an unbound protein that is at the equivalent position. A cofactor of the decomposed primary protein is assigned to a corresponding cofactor of the unbound protein if at least 25% of the interface residues that are in contact with the cofactor ($d_{\text{cof-res}} < 10$ Å) are equivalent with respect to the sequence alignment between primary and

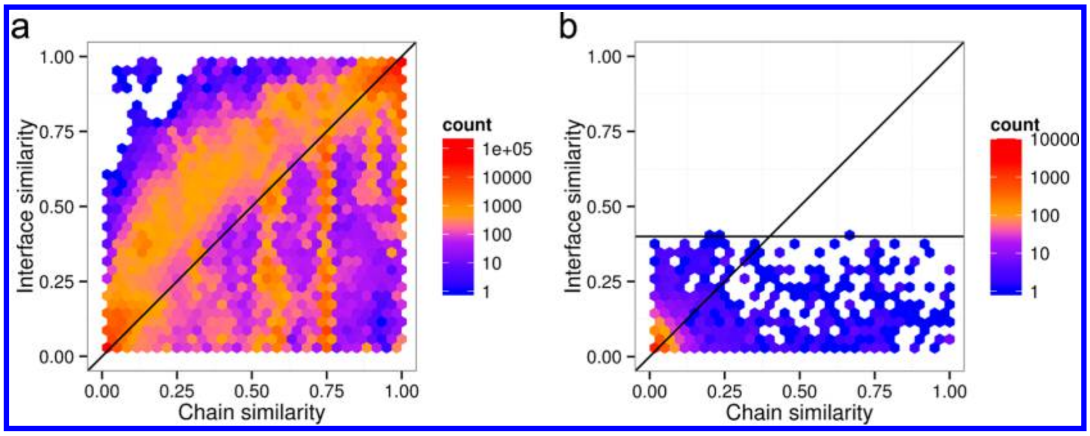


Figure 4. Correlation diagram between two-sided interface similarity [defined by eq 5, that uses only sequence identity of the residues at the interface] and chain similarity [given by the sequence identity of the whole polypeptide chains that possess interface residues]. a) All pairs of the 11,600 protein docking complexes found in the PDB[27] are considered. b) Only the 2,070 nonredundant protein docking partners are considered. Here, the interface similarity is below 40% using the inequality in eq 6 in a single-linkage cluster algorithm.[46]

unbound protein. Cofactors at the interface of a protein complex may have more than one analogue in corresponding unbound structures and vice versa. In such cases, the assignment involving more matching interface residues is favored.

F. Representative Unbound Structures for Protein Docking Complexes with Cofactors at the Interface. We search for complementary pairs of unbound proteins that together possess the complete set of cofactors present in the primary protein and whose positions correspond to the locations at the interface of the primary protein. If this search is successful and we find a single pair of unbound proteins, the pair is selected. For the case that several pairs of such unbound proteins are found, the selection criteria applied for unbound structures of docking complexes without cofactors at the interface listed above (section C of Results) are used. If the search is unsuccessful, we consider only one unbound structure involving the maximum number of cofactors corresponding to the ones that are present at the interface of the primary protein. For the case that several unbound structures with the same maximum number of cofactors are found, we apply the selection criteria described above for unbound structures of primary proteins with no cofactor at the interface.

■ DISCUSSION

The ProPairs Data Set of Protein Docking Complexes. With our proposed procedure to identify legitimate protein docking partners, we find 5,642 different protein complexes in the PDB[27] with at least one legitimate docking interface (for more details see Table 1). Since a considerable number of these protein complexes possess more than one such interface, we obtain for the large ProPairs data set 11,600 decompositions of protein complexes yielding different legitimate docking interfaces. Of the 11,600 protein docking complexes 3,605 contain cofactors at the interface. Hence, considering protein docking complexes involving cofactors is quite important. To each protein docking complex in the large ProPairs data set a potentially larger number of unbound structures may be assigned. However, the individual unbound structures are not yet combined to complementary pairs that can form the whole protein complex. This pairing takes place when the data set of protein complexes is further reduced.

The large ProPairs data set of N_{redun} = 11,600 protein docking complexes contains a considerable number of redundancies. A large part of the redundancies is simply the result of the symmetry of protein complexes that may, for instance, involve a C_n rotational axis of symmetry. Removing these initial redundancies leads to 7,035 independent protein docking complexes that may still involve other types of redundancies. To detect and remove all redundancies in the large ProPairs data set, we compute all pairwise similarities of the interface residues performing $(N_{\text{redun}}^2 - N_{\text{redun}})/2$ = 67,274,200 pairwise comparisons of sequences. Several methods can be applied to compute the similarity between protein–protein interfaces.[39−43] In our approach we compute sequence alignments between all chain pairs belonging to different protein complexes that are at the interface. In this way we obtain an interface similarity measure analogous to eq 5, namely by evaluating the fraction of aligned interface residues. We use this similarity measure in a single linkage cluster algorithm as described in the Results section B. As a result, we obtain 2,070 nonredundant protein docking complexes, constituting the small ProPairs data set.

The large ProPairs data set provides for each pair of docking units all possible unbound structures found in the PDB. In contrast, the small ProPairs data set selects one corresponding unbound structure for each docking unit. Thus, the small ProPairs data set can either contain a single unbound structure or a single complementary pair of unbound structures that can form the complete protein complex. In the small ProPairs data set, 810 protein docking complexes contain unbound structures for both binding partners and 417 of them possess at least one cofactor at the interface (Table 1). The average *i*RMSD between the unbound structures and the corresponding binding partners of the protein docking complexes is 1.88 Å for the 810 protein docking complexes that possess an unbound structure for both binding partners.

Comparing the Large and the Small ProPairs Data Set. To analyze the relation between pairwise interface similarity $S_{2\text{-inter}}$ and similarity of all polypeptide chains $S_{2\text{-chain}}$ belonging to the interface [defined analogous to eq 5], we generate a two-dimensional correlation diagram of occurrences of these two quantities for the large ProPairs data set of 11,600 redundant and the small ProPairs data set of 2,070 nonredundant interfaces of protein docking complexes (Figure 4a and b). In Figure 4a we observe that the majority of pairs of interfaces are either very dissimilar or similar with respect to both: the interface ($S_{2\text{-inter}}$, from both binding partners) and the chains participating in the interface ($S_{2\text{-chain}}$). The dissimilarity originates from unrelated complexes. The similarity corresponds to protein complexes with a high degree of redundancy, since they may differ only by some point mutations or are the same type of protein belonging to different species. A third cause for high similarity is intrinsic similarity within protein complexes (for instance due to rotational symmetry) leading to multiple occurrences of an interface. The vertical stripes appearing in Figure 4a originate mainly from pairs of protein docking complexes of hemoglobin structures of which the PDB contains many structures. Since hemoglobin involves the α- and β-chains, more than one stripe occurs. The cluster of hemoglobin of the large ProPairs data set contains 801 docking complexes.

At the bottom right corner of Figure 4a we observe an accumulation of interface pairs that possess practically no interface similarity but a high similarity in the sequences of the chains that are part of the interface. At first sight this relationship seems to be counterintuitive. If two single chain binding partners can form protein docking complexes in alternative geometries, the corresponding interface similarity of the two possible binding modes vanishes. The chain similarity, however, remains unity, leading to an accumulation of occurrences in the lower right corner of the similarity correlation diagram in Figure 4a. An example of such a protein complex is shown in Figure 5. In Figure 4b the same correlation plot between interface and chain similarity is shown after removing the interface redundancies. Now the interface similarity is below 0.4 as required by the condition in eq 6; the stripes are gone, but the accumulation of interface pairs with vanishing interface and chain sequence similarity remains.

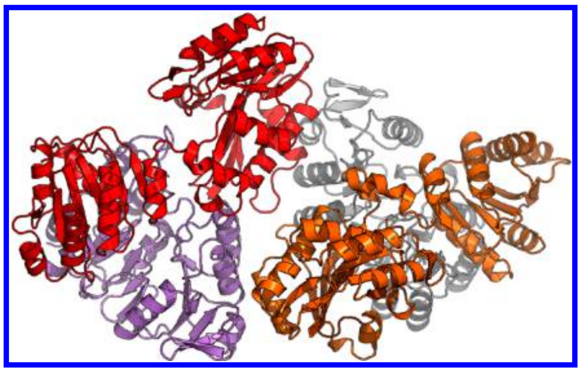


Figure 5. Protein complex (PDB id 3ufx, succinyl-CoA-synthetase) with high chain similarity ($S_{2\text{-chain}}$ = 1.0) and low interface sequence identity ($S_{2\text{-inter}}$ = 0.0). The protein complex is a heterotetramer consisting of four polypeptide chains, two α subunits (violet and gray), and two β subunits (red and orange). Each α subunit forms two inequivalent interfaces, each one with a different β subunit. The same applies to the β subunits.

Above the diagonal (black line in Figure 4a) there is also a considerable abundance of pairs of protein docking complexes. These have an interface similarity that is larger than the corresponding chain similarity. This relationship is an indication of the importance of such interfaces that possess a larger fraction of conserved residues than the whole polypeptide chains participating in the interface.[47−50] An example of a pair of protein docking complexes with very similar interfaces but relatively small similarity of the chain sequences of chains participating in the interface is shown in Figure 6. Interestingly, the structures of the two protein complexes, which are trypsines with inhibitors, are very similar. According to our similarity measure based on interface similarity, this pair of protein complexes is redundant. Since the PDB lists a large number of redundant protein complexes of trypsines with inhibitors, none of the two protein docking complexes shown in Figure 6 is in the small nonredundant ProPairs data set of 2070 entries. In contrast, both complexes are listed in the large ProPairs data set. These redundant protein docking complexes are removed with a single-linkage cluster algorithm[46] as described in the Results section B.

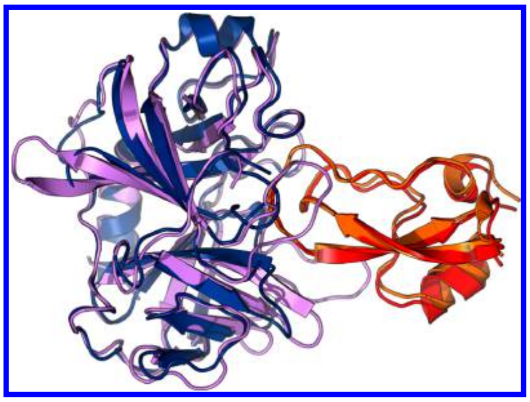


Figure 6. Protein complex (PDB id 1fy8) with low chain similarity ($S_{2\text{-chain}}$ = 0.259) and high interface similarity ($S_{2\text{-inter}}$ = 0.678). The complex of trypsin (blue) bound to trypsin inhibitor (red) is superimposed on a complex (PDB id 1eaw) consisting of a protein with trypsin-like binding pocket (violet) bound to the same trypsin inhibitor (orange). Therefore, the latter contributes considerably to the two-sided chain similarity $S_{2\text{-chain}}$, while the one-sided chain similarity between trypsin and the trypsin-like protein is only $S_{1\text{-chain}}$ = 0.139.

Comparison with Protein−Protein Docking Benchmark 4.0 (DB4.0). We compare the ProPairs data set of protein docking complexes with the Protein−Protein Docking Benchmark 4.0 (DB4.0)[29] which, together with DOCKGROUND,[30,31] is one of the most widely used data sets in the docking community.[16,20] This data set consists of 176 protein docking complexes from which 164 possess two unbound structures and 12 possess only one unbound structure. The latter 12 are protein complexes for which one binding partner is an immunoglobulin that cannot be crystallized in the unbound state. For eight protein docking complexes in the DB4.0 data set, the coordinates of one of the two unbound structures were altered by modeling [see http://zlab.umassmed.edu/benchmark/].

It is of interest to find out to which degree the DB4.0 protein complexes[29] are part of the large ProPairs data set of 11,600 legitimate protein complexes of the present study. It is also useful to know how many of the DB4.0 protein complexes can still be found in the small ProPairs data set of 2,070 representative protein complexes with nonredundant interfaces selected from the large data set. From the 176 protein docking complexes in the DB4.0 data set, we also find 163 in the redundant large ProPairs data set possessing identical unbound structures. If we also include the ProPairs protein docking complexes, for which only one unbound partner protein is available, 175 of the total 176 protein docking complexes are

Table 3. Deviations between DB4.0[29] and the Present Data Set Due to Deficiencies of the ProPairs Method[c]

no.	protein docking complex	1. unbound protein	2. unbound protein	reason for deviation
1	1d6r A:I[a]	2tgt A[b]	1k9b A[b]	Complex was not found since the first biological assembly (boiunit) of 1d6r is incorrect (see FigureS3.1).
2	1azs AB:C	1ab8 AB	1azt A	Complex was found. However, the first unbound structure was not found since chain A in 1azs is part of the interface but the corresponding chain A in 1ab8 has a very dissimilar sequence (see FigureS3.2).

[a]The first four characters denote the PDB id of the protein. The characters behind denote the polypeptide chains involved in the interface of a protein docking complex. The chains belonging to different binding partners are separated by ":". [b]The first unbound protein refers to the set of interface chains of the complex listed before ":". The second unbound protein refers to the set of interface chains of the complex listed after ":". [c]The listed protein complexes and the notations refer to the DB4.0 data set. The structures of the protein complexes are shown with the corresponding numbers in Figure S3 of the Supporting Information.

also included in the large ProPairs data set. The reason for this small discrepancy is explained below (see Table 3). In the small nonredundant ProPairs data set, 108 from the 176 protein complexes in DB4.0 are still present. The remaining 67 protein complexes that are present in the large ProPairs data set are represented by equivalent protein complexes in the small ProPairs data set. A schematic presentation of the overlap of data sets is shown in Figure 7.

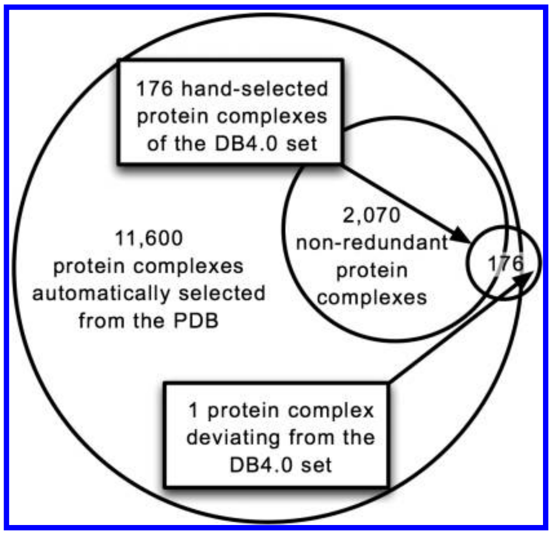

176 hand-selected protein complexes of the DB4.0 set
11,600 protein complexes automatically selected from the PDB
2,070 non-redundant protein complexes
176
1 protein complex deviating from the DB4.0 set

Figure 7. Schematic representation of the sets of interfaces of protein docking complexes identified with the ProPairs procedure in the PDB[27] and the DB4.0[29] data set. Only two docking complexes of the DB4.0 data set are not contained in the large ProPairs data set of 11,600 docking complexes. If one also considers the identity of the corresponding unbound protein structures, there are 11 additional cases in which the large ProPairs data set does not fully agree with the DB4.0 data set but offers essentially equivalent alternative data. 108 of the 176 protein complexes considered in the DB4.0 data set also appear in the small ProPairs data set of 2,070 docking complexes. However, ProPairs may be more reliable since it offers equivalent protein docking complexes obeying well-defined criteria.

Composition of Protein Docking Complexes of DB4.0 and Small ProPairs Database. In the comparison between DB4.0 and ProPairs for the small ProPairs data set, we consider only unbound proteins belonging to protein docking complexes for which both unbound proteins are available. These are a total of 2*810 = 1620 unbound proteins from the small ProPairs data set. In comparison, analogous data are shown for the DB4.0[29] data set, for which we consider the unbound structures corresponding to 175 complexes that are also contained in the large ProPairs data set of 11,600 protein docking complexes. [The complex with the wrong biological unit in the PDB (1d6r) was ignored.] In this comparison we consider the distribution of unbound proteins shown in the form of histograms as a function of the number of residues (Figure 8a), the number of C_α atoms in the interface (Figure 8b), and as a function of the interface RMSD (Figure 9).

For both data sets, the distribution of unbound proteins as a function of the number of residues of the unbound proteins (Figure 8a) decays monotonically with increasing size. The ProPairs data set possesses a few (about 30) very large unbound proteins that appear very pronounced in the logarithmic plot. Apart from this detail, the distributions of the ProPairs and DB4.0 data set are quite similar. For both data sets, the distribution of unbound proteins as a function of the number of C_α atoms at the interface (Figure 8b) shows a maximum at about 70 C_α atoms. The ProPairs data set possesses a few (about 25) unbound proteins with a larger interface area than DB4.0. Again, these appear pronounced since a logarithmic plot is used. The distributions for both the ProPairs and DB4.0 data sets are also very similar. Although a small lower limit of only 20 residues is used for unbound proteins in the ProPairs data set, no significant accumulation of very small, unbound proteins or small interfaces can be observed in the distribution.

The distribution of interface RMSDs (*i*RMSD) is one major criterion of the Capri contest to rank the predictions of protein docking complexes based on unbound protein structures.[33] Therefore, the ProPairs data set provides this parameter and users may choose it to select protein complexes. Figure 9 shows the distribution of *i*RMDSs for the 810 pairs of docking partners from the nonredundant small ProPairs data set, for which two corresponding unbound proteins are available. In comparison with the corresponding distribution from the 175 [The complex with the wrong biological unit in the PDB (1d6r) was ignored.] protein complexes of the DB4.0 benchmark set, which are also in the large redundant ProPairs data set, the number of protein docking complexes in the small ProPairs data set is considerably larger (note the logarithmic scale) for all *i*RMDS distance classes. Beyond the distance class 6 Å to 7 Å, the DB4.0 benchmark set possesses only very few complexes, while in the small ProPairs data set there are about 30 unbound proteins corresponding to different complexes with an *i*RMDS larger than 10 Å.

A more detailed comparison of the *i*RMDSs, which also includes the data from the large ProPairs data set, is shown in Figure S2 of the Supporting Information. The *i*RMDSs of the DB4.0[29] benchmark set are smaller (average 1.39 Å) than those of the small ProPairs data set (average 1.88 Å). In other words, the DB4.0 benchmark set lacks protein docking complexes whose unbound structures deviate more strongly in the interface region. These cases may not be so useful for the

DOI: 10.1021/acs.jcim.5b00082

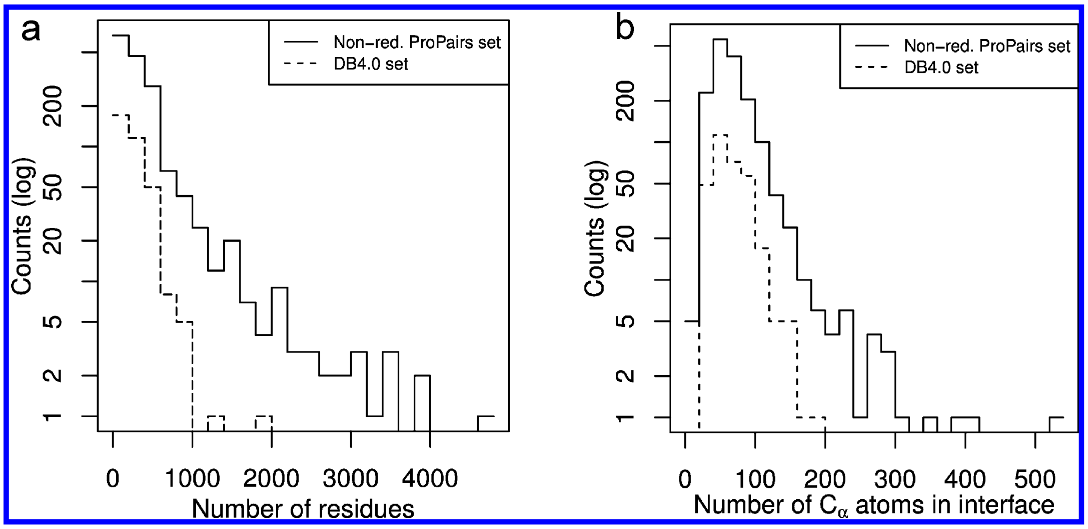


Figure 8. Comparison of the small ProPairs and DB4.0 data sets of protein docking structures using a logarithmic scale. For ProPairs (solid line) we consider only unbound proteins of protein docking complexes for which both unbound proteins are available involving a total of 1,620 different unbound proteins. In comparison analogous data are shown for the DB4.0[29] data set for which we consider the unbound structures corresponding to 175 complexes that are also contained in the large ProPairs data set of 11,600 protein docking complexes (dashed line). **a:** Distribution of unbound proteins as a function of the total number of residues in the unbound protein. **b:** Distribution of unbound proteins as a function of the number of C_α atoms at the interface.

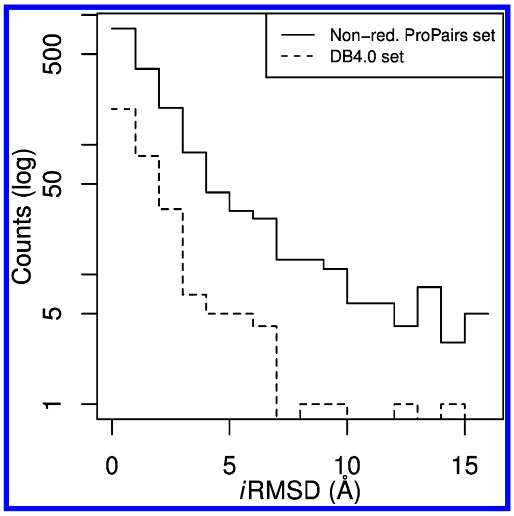


Figure 9. Histogram of the number of unbound proteins (single logarithmic scale) as a function of the interface RMSD (*i*RMSD, bin width 1 Å). 810 protein docking complexes are considered, each one possessing two unbound structures taken from the small ProPairs data set (solid line). Both unbound protein structures of a protein docking complex contribute independently to the distribution with their own *i*RMSD value. In comparison, analogous data are shown for the DB4.0[29] data set for which we consider the unbound structures corresponding to 175 complexes that are also contained in the large ProPairs data set of 11,600 protein docking complexes (dashed line).

learning of scoring or energy functions for prediction, but they can be considered as more challenging candidates for protein docking prediction.

Protein Docking Complexes of DB4.0 That Are Not Part of the Large ProPairs Data Set. There are different reasons for deviations between the DB4.0[29] data set of 176 semiautomatic selected protein docking structures and the large ProPairs data set of protein docking complexes generated with a fully automatic selection method applied to the whole set of PDB[27] structures. For example, the protein docking complex with PDB id 1d6r contained in DB4.0 was not found with the ProPairs criteria since the biological assembly listed first in the PDB is incorrect (Table 3 and Figure S3.1 in the Supporting Information). The second protein complex listed in Table 3 is contained in the DB4.0 data set with two unbound structures. In the large ProPairs data set this protein complex is listed only with the unbound structure 1azt, since the sequence of the other unbound structure (PDB id 1ab8) is too dissimilar from the corresponding bound part in the protein complex. Therefore, this unbound structure is not recognized by the ProPairs selection criteria that are based on sequence identity. The protein complex 1bgx in DB4.0 involves one fabricated unbound structure of DNA polymerase (PDB id 1cmw).[51] The same protein complex is in the small ProPairs data set;, however, the fabricated unbound structure is replaced by another structure if DNA polymerase (PDB id 1taq).

The protein docking complex (PDB id 1eer, no. 3 in Table S3) possesses two interfaces (Figure 10). Thus, this complex occurs twice in the small ProPairs data set, representing each of the two interfaces, both of which involve the complementary pairs of unbound structures that can form the whole complex. The entries in Table 4 are protein docking complexes contained in the DB4.0 data set which differ from the ProPairs

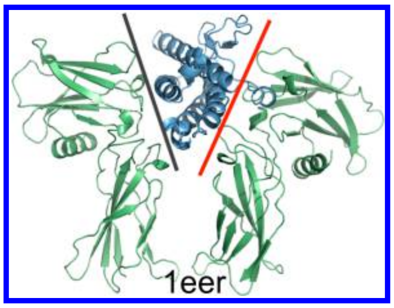


Figure 10. Protein docking complex (PDB id 1eer) exhibits two independent interfaces (black and red lines, Table S3.3), which are both contained in the small ProPairs data set.

Table 4. Deviations between the Small ProPairs Data Set and DB4.0[29] That Are Related to Deficiencies of the Latter[c]

no.	protein docking complex	1. unbound protein	2. unbound protein	reason for deviation
11	1zli A:B[a]	1kwm A[b]	2jto A[b]	First unbound structure has a covered interface region (see FigureS3.11).
12	1pxv A:C	1x9y A	1nyc A	First unbound structure has a covered interface region (see FigureS3.12).
13	1bgx HL:T	1ay1 HL	1cmw A	Second unbound structure was fabricated and withdrawn from PDB.[51]

[a] The first four characters denote PDB id of the protein. The characters behind denote the polypeptide chains involved in the interface of a protein docking complex. The chains belonging to different binding partners are separated by ":". [b] The first unbound protein refers to the set of interface chains of the complex listed before ":". The second unbound protein refers to the set of interface chains of the complex listed after ":". [c] The listed protein complexes and notations refer to the DB4.0 data set. The structures of the protein complexes are shown in Figure S3 of the Supporting Information with the corresponding numbers.

data set, since only one of the two unbound structures is assigned. For the protein complexes 4−7 in Table S3, ProPairs rejects one of the two unbound structures since a small peptide is localized in the interface region of that unbound structure. For the protein complexes involving a cofactor at the interface (complexes 8−10 in Table S3), ProPairs does not consider the pair of unbound structures since they do not contain the cofactor present at the interface.

In Table 4, three protein complexes are listed for which the DB4.0[29] possesses deficiencies. In two cases (protein complex 1zli and 1pxv) one of the unbound structures used by DB4.0 has the interface region covered by another protein domain such that the unbound structure is inappropriate. An example is shown in Figure 11. The last entry in Table 4 refers to a protein docking complex for which DB4.0 uses a fabricated and therefore obsolete protein structure for one of the unbound structures (PDB id 1cmw). The same protein docking complex is contained in the small ProPairs data set; however, the fabricated unbound structure of DNA polymerase is replaced by an alternative polymerase structure (PDB id 1taq).

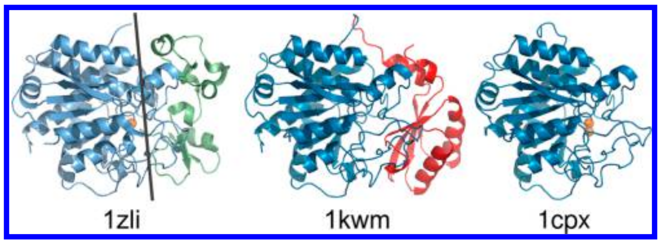


Figure 11. Protein docking complex (left panel, PDB id 1zli) (blue and green part, black line indicates interface) possesses a zinc ion (orange sphere) at the interface. In DB4.0[29] the unbound structure (middle panel, PDB id 1kwm) is assigned to the blue domain of the protein docking complex (left panel). This unbound structure involves an additional domain (red chain) that would collide with the green chain of the protein docking complex and is therefore inappropriate. The small ProPairs data set contains the same protein docking complex. However, the unfavorable unbound structure is replaced by another unbound structure (left panel, PDB id 1cpx), where the interface is freely accessible and which contains a zinc ion in the same location as in the protein docking complex.

Reliability of Data in the ProPairs Data Set. We applied much care in designing the automatic generation of the ProPairs data set. During development stages of the automatic selection procedure the resulting data sets were selectively inspected, and whenever we found problematic data the selection procedure was adjusted. Nevertheless, we cannot rule out the possibility that the data sets generated by our method still contain incorrect data. Apart from improperly handled scenarios in our procedure, the main source of errors is wrong data in the PDB. These can be wrong biological assemblies or even wrong protein structures. Comparing the content of ProPairs with DB4.0 we found one example for each of these types of errors. The DB4.0 database contains a protein structure which was fabricated. In a preliminary version ProPairs contained a protein complex based on a wrong biological assembly in the PDB, which yields colliding protein structures. This was recognized and removed with a collision check that was added to the ProPairs selection procedure. We recommend users of ProPairs to be aware of errors and encourage them to communicate possible problems. The authors of this study will review such problems for improvement of the ProPairs program. The complete source code is released under an open source license, so users in the scientific community are also able to modify the automatic selection procedure and are very welcome to share their improvements.

SUMMARY

ProPairs identifies protein docking complexes using biological assembly information with corresponding unbound structures from the PDB.[27] The fully automatic selection procedure is based on sequence identity. In contrast to other approaches in which redundancies of protein docking complexes are removed by protein families or sequence identities of the whole proteins, ProPairs removes redundancies using sequence identities at the interface only. This selection criterion may be one reason why the ProPairs data set is larger than other data sets of protein docking complexes. Another reason for the larger size of ProPairs is the continuous growth of the PDB. The present study is based on protein structures that were available in the PDB in November 2013. The existence of corresponding unbound structures in the PDB is used as the criterion for legitimate decompositions of the considered protein complexes. Simultaneously, the unbound protein structures can be used as benchmarks to develop and test algorithms that predict protein docking geometries. A protein complex may be decomposed in several ways, and for each decomposition there may be more than one unbound structure for each of the two decomposed parts of the protein complex. The large ProPairs data set contains this information and thus consists of 11,600 different decompositions of protein complexes that are found based on the existence of corresponding unbound protein structures in the PDB. This data set can be downloaded at http://propairs.github.io. Aside from a few exceptions, the large ProPairs data set comprises all protein docking complexes of the DB4.0[29] benchmark set that was obtained in a semiautomatic way and hand-curated. In contrast to DB4.0, ProPairs is not explicitly hand-curated though its selection procedures were tested for quality and performance. As a result, the large ProPairs data set includes nearly the complete DB4.0 benchmark set, demonstrating the reliability of the selection criteria used in ProPairs.

By removing redundancies in the large ProPairs data set of 11,600 docking complexes we obtain the small ProPairs data set involving 2,070 nonredundant protein docking complexes. The redundancies between the docking complexes are identified by computing sequence identities in their interface regions. For 810 protein docking complexes in the small ProPairs data set both unbound partner proteins are available. The small nonredundant ProPairs data set is thus more than ten times larger than that of DB4.0.[29] The small ProPairs data set is available as a Web application at http://propairs.github.io. Various filters exist and can be applied to obtain specific subsets. Furthermore, the overlaid structures of bound and corresponding unbound partner proteins are visualized. The ProPairs data set is generated by a computer program using well-defined rules. This program can be applied to the PDB to generate an up-to-date data set of protein docking complexes. For the approximately 90,000 protein structures available in the PDB in November 2013, the execution time required by 32 CPU cores is on the order of 1 week.

The ProPairs data set is not only useful as a benchmark set of protein docking complexes to design and test protein docking algorithms. It may also be instrumental in identifying new interactions between proteins forming transient protein complexes in metabolic and signaling networks. Furthermore, ProPairs may provide valuable information for identifying new drug targets. Finally, perhaps more significant is the possibility that ProPairs could aid in designing more effective drugs that are engineered to act at the interface of docking complexes, thus inhibiting protein function most efficiently.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

A description is given of the types of cofactors at the interface and how the various cofactors are treated. Detailed information regarding the difference between the ProPairs and DB4.0 data set is provided, and the corresponding structures are displayed. The function $S_{2\text{-sided}}(u, v)$, eq 2b, is displayed as a contour plot. The distribution of the proteins in the large ProPairs data set is given as a function of the *i*RMSD.The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00082.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: knapp@chemie.fu-berlin.de.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

F.K. would like to thank Dr. Dima Kozakov and Dr. Sandor Vajda for their hospitality in the winter 2009/2010. The authors gratefully acknowledge financial support from the Deutsche Forschungsgemeinschaft (DFG) in the framework of the research training group “Computational System Biology” (CSB), GRK 1772.

■ ABBREVIATIONS:

protein data bank, PDB; common connected substructure, CCsub; interface root-mean-square deviation, *i*RMSD

■ REFERENCES

(1) Lodish, H.; Berk, A.; Zipursky, S. L.; Matsudaira, P.; Baltimore, D.; Darnell, J. *Molecular Cell Biology*; Freeman: New York, 2000.

(2) Xenarios, I.; Salwinski, L.; Duan, X. J.; Higney, P.; Kim, S.-M.; Eisenberg, D. Dip, the Database of Interacting Proteins: A Research Tool for Studying Cellular Networks of Protein Interactions. *Nucleic Acids Res.* **2002**, *30*, 303−305.

(3) Zanzoni, A.; Montecchi-Palazzi, L.; Quondam, M.; Ausiello, G.; Helmer-Citterich, M.; Cesareni, G. Mint: A Molecular Interaction Database. *FEBS Lett.* **2002**, *513*, 135−140.

(4) Hermjakob, H.; Montecchi-Palazzi, L.; Lewington, C.; Mudali, S.; Kerrien, S.; Orchard, S.; Vingron, M.; Roechert, B.; Roepstorff, P.; Valencia, A. Intact: An Open Source Molecular Interaction Database. *Nucleic Acids Res.* **2004**, *32*, D452−D455.

(5) Cusick, M. E.; Yu, H.; Smolyar, A.; Venkatesan, K.; Carvunis, A.-R.; Simonis, N.; Rual, J.-F.; Borick, H.; Braun, P.; Dreze, M. Literature-Curated Protein Interaction Datasets. *Nat. Methods* **2008**, *6*, 39−46.

(6) Aloy, P.; Russell, R. B. Structural Systems Biology: Modelling Protein Interactions. *Nat. Rev. Mol. Cell Biol.* **2006**, *7*, 188−197.

(7) Imming, P.; Sinning, C.; Meyer, A. Drugs, Their Targets and the Nature and Number of Drug Targets. *Nat. Rev. Drug Discovery* **2006**, *5*, 821−834.

(8) Morelli, X.; Bourgeas, R.; Roche, P. Chemical and Structural Lessons from Recent Successes in Protein−Protein Interaction Inhibition (2p2i). *Curr. Opin. Chem. Biol.* **2011**, *15*, 475−481.

(9) Wells, J. A.; McClendon, C. L. Reaching for High-Hanging Fruit in Drug Discovery at Protein-Protein Interfaces. *Nature* **2007**, *450*, 1001−1009.

(10) Metz, A.; Ciglia, E.; Gohlke, H. Modulating Protein-Protein Interactions: From Structural Determinants of Binding to Druggability Prediction to Application. *Curr. Pharm. Des.* **2012**, *18*, 4630−4647.

(11) Zinzalla, G.; Thurston, D. E. Targeting Protein-Protein Interactions for Therapeutic Intervention: A Challenge for the Future. *Future Med. Chem.* **2009**, *1*, 65−93.

(12) Milroy, L.-G.; Grossmann, T. N.; Hennig, S.; Brunsveld, L.; Ottmann, C. Modulators of Protein-Protein Interactions. *Chem. Rev.* **2014**, *114*, 4695−4748.

(13) Katchalski-Katzir, E.; Shariv, I.; Eisenstein, M.; Friesem, A. A.; Aflalo, C.; Vakser, I. A. Molecular Surface Recognition: Determination of Geometric Fit between Proteins and Their Ligands by Correlation Techniques. *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89*, 2195−2199.

(14) Aloy, P.; Pichaud, M.; Russell, R. B. Protein Complexes: Structure Prediction Challenges for the 21st Century. *Curr. Opin. Struct. Biol.* **2005**, *15*, 15−22.

(15) Ritchie, D. W. Recent Progress and Future Directions in Protein-Protein Docking. *Curr. Protein Pept. Sci.* **2008**, *9*, 1−15.

(16) Janin, J. Protein−Protein Docking Tested in Blind Predictions: The Capri Experiment. *Mol. BioSyst.* **2010**, *6*, 2351−2362.

(17) Wass, M. N.; David, A.; Sternberg, M. J. Challenges for the Prediction of Macromolecular Interactions. *Curr. Opin. Struct. Biol.* **2011**, *21*, 382−390.

(18) Vajda, S.; Hall, D. R.; Kozakov, D. Sampling and Scoring: A Marriage Made in Heaven. *Proteins: Struct., Funct., Bioinf.* **2013**, *81*, 1874−1884.

(19) Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W. Computational Methods in Drug Discovery. *Pharmacol. Rev.* **2014**, *66*, 334−395.

(20) Vakser, I. A. Protein-Protein Docking: From Interaction to Interactome. *Biophys. J.* **2014**, *107*, 1785−1793.

(21) Vakser, I. A. Low-Resolution Structural Modeling of Protein Interactome. *Curr. Opin. Struct. Biol.* **2013**, *23*, 198−205.

(22) Wodak, S. J.; Janin, J. Computer Analysis of Protein-Protein Interaction. *J. Mol. Biol.* **1978**, *124*, 323−342.

(23) Northrup, S. H.; Allison, S. A.; McCammon, J. A. Brownian Dynamics Simulation of Diffusion-Influenced Bimolecular Reactions. *J. Chem. Phys.* **1984**, *80*, 1517−1524.

(24) Ullmann, G. M.; Knapp, E. W.; Kostic, N. M. Computational Simulation and Analysis of Dynamic Association between Plastocyanin

and Cytochrome F. Consequences for the Electron-Transfer Reaction. *J. Am. Chem. Soc.* **1997**, *119*, 42−52.

(25) Dominguez, C.; Boelens, R.; Bonvin, A. M. Haddock: A Protein-Protein Docking Approach Based on Biochemical or Biophysical Information. *J. Am. Chem. Soc.* **2003**, *125*, 1731−1737.

(26) Schneider, S.; Zacharias, M. Scoring Optimisation of Unbound Protein−Protein Docking Including Protein Binding Site Predictions. *J. Mol. Recognit.* **2012**, *25*, 15−23.

(27) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.

(28) Levy, E. D.; Pereira-Leal, J. B.; Chothia, C.; Teichmann, S. A. 3d Complex: A Structural Classification of Protein Complexes. *PLoS Comput. Biol.* **2006**, *2*, e155−e155.

(29) Hwang, H.; Vreven, T.; Janin, J.; Weng, Z. Protein-Protein Docking Benchmark Version 4.0. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 3111−3114.

(30) Douguet, D.; Chen, H.-C.; Tovchigrechko, A.; Vakser, I. A. Dockground Resource for Studying Protein-Protein Interfaces. *Bioinformatics* **2006**, *22*, 2612−2618.

(31) Gao, Y.; Douguet, D.; Tovchigrechko, A.; Vakser, I. A. Dockground System of Databases for Protein Recognition Studies: Unbound Structures for Docking. *Proteins: Struct., Funct., Bioinf.* **2007**, *69*, 845−851.

(32) Meyer, T.; Knapp, E. W. Database of Protein Complexes with Multivalent Binding Ability: Bival-Bind. *Proteins: Struct., Funct., Bioinf.* **2014**, *82*, 744−751.

(33) Méndez, R.; Leplae, R.; De Maria, L.; Wodak, S. J. Assessment of Blind Predictions of Protein-Protein Interactions: Current Status of Docking Methods. *Proteins: Struct., Funct., Bioinf.* **2003**, *52*, 51−67.

(34) van Dijk, M.; Bonvin, A. M. A Protein−DNA Docking Benchmark. *Nucleic Acids Res.* **2008**, *36*, e88−e88.

(35) Lewis, B. A.; Walia, R. R.; Terribilini, M.; Ferguson, J.; Zheng, C.; Honavar, V.; Dobbs, D. Pridb: A Protein−Rna Interface Database. *Nucleic Acids Res.* **2011**, *39*, D277−D282.

(36) Smith, T. F.; Waterman, M. S. Identification of Common Molecular Subsequences. *J. Mol. Biol.* **1981**, *147*, 195−197.

(37) Bickerton, G. R.; Higueruelo, A. P.; Blundell, T. L. Comprehensive, Atomic-Level Characterization of Structurally Characterized Protein-Protein Interactions: The Piccolo Database. *BMC Bioinf.* **2011**, *12*, 313−313.

(38) Kabsch, W. A Solution for the Best Rotation to Relate Two Sets of Vectors. *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.* **1976**, *32*, 922−923.

(39) Zhu, H.; Sommer, I.; Lengauer, T.; Domingues, F. S. Alignment of Non-Covalent Interactions at Protein-Protein Interfaces. *PLoS One* **2008**, *3*, e1926.

(40) Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J. Siteengines: Recognition and Comparison of Binding Sites and Protein-Protein Interfaces. *Nucleic Acids Res.* **2005**, *33*, W337−W341.

(41) Pulim, V.; Berger, B.; Bienkowska, J. Optimal Contact Map Alignment of Protein-Protein Interfaces. *Bioinformatics* **2008**, *24*, 2324−2328.

(42) Gao, M.; Skolnick, J. Ialign: A Method for the Structural Comparison of Protein-Protein Interfaces. *Bioinformatics* **2010**, *26*, 2259−2265.

(43) Tsai, C.-J.; Lin, S. L.; Wolfson, H. J.; Nussinov, R. A Dataset of Protein-Protein Interfaces Generated with a Sequence-Order-Independent Comparison Technique. *J. Mol. Biol.* **1996**, *260*, 604−620.

(44) Murzin, A. G.; Brenner, S. E.; Hubbard, T.; Chothia, C. Scop: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures. *J. Mol. Biol.* **1995**, *247*, 536−540.

(45) Fox, N. K.; Brenner, S. E.; Chandonia, J.-M. Scope: Structural Classification of Proteins - Extended, Integrating Scop and Astral Data and Classification of New Structures. *Nucleic Acids Res.* **2014**, *42*, D304−D309.

(46) Jain, A. K.; Dubes, R. C. *Algorithms for Clustering Data*; Prentice-Hall, Inc.: Englewood Cliffs, NJ, 1988.

(47) Grishin, N. V.; Phillips, M. A. The Subunit Interfaces of Oligomeric Enzymes Are Conserved to a Similar Extent to the Overall Protein Sequences. *Protein Sci.* **1994**, *3*, 2455−2458.

(48) Valdar, W. S.; Thornton, J. M. Protein-Protein Interfaces: Analysis of Amino Acid Conservation in Homodimers. *Proteins: Struct., Funct., Bioinf.* **2001**, *42*, 108−124.

(49) Ofran, Y.; Rost, B. Analysing Six Types of Protein-Protein Interfaces. *J. Mol. Biol.* **2003**, *325*, 377−387.

(50) Caffrey, D. R.; Somaroo, S.; Hughes, J. D.; Mintseris, J.; Huang, E. S. Are Protein-Protein Interfaces More Conserved in Sequence Than the Rest of the Protein Surface? *Protein Sci.* **2004**, *13*, 190−202.

(51) Berman, H. M.; Kleywegt, G. J.; Nakamura, H.; Markley, J. L.; Burley, S. K. Safeguarding the Integrity of Protein Archive. *Nature* **2010**, *463*, 425−425.