HHS Public Access

Author manuscript

J Chem Inf Model. Author manuscript; available in PMC 2018 July 13.

Published in final edited form as:

J Chem Inf Model. 2018 February 26; 58(2): 532–542. doi:10.1021/acs.jcim.7b00580.

Structural Characterization and Function Prediction of Immunoglobulin-like Fold in Cell Adhesion and Cell Signaling

Jiawen Chen[1], **Bo Wang**[1], and **Yinghao Wu**[1,*]

[1]Department of Systems and Computational Biology, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, NY, 10461

Abstract

Domains that belong to immunoglobulin (Ig) fold are extremely abundant in cell surface receptors, which play significant roles in cell-cell adhesion and signaling. Although the structures of domains in Ig fold share common topology of β-barrels, functions of receptors in adhesion and signaling are regulated by the very heterogeneous binding between these domains. Additionally, only a small number of domains are directly involved in the binding between two multi-domain receptors. It is challenging and time-consuming to experimentally detect the binding partners of a given receptor, and further determine which specific domains in this receptor are responsible for binding. Therefore, current knowledge in binding mechanism of Ig-fold domains and their impacts on cell adhesion and signaling is very limited. A bioinformatics study can shed lights on this topic from a systematic point of view. However, there is so far no computational analysis on the structural and functional characteristics of entire Ig fold. We constructed non-redundant structural datasets for all domains in Ig fold, depending on their functions in cell adhesion and signaling. We found that datasets of domains in adhesion receptors show different binding preference from domains in signaling receptors. Using structural alignment, we further built a common structural template for each group of domain dataset. By mapping the protein-protein binding interface of each domain in a group onto the surface of its structural template, we found binding interfaces are highly overlapped within each specific group. These overlapped interfaces, as we called consensus binding interfaces, are distinguishable among different datasets of domains. Finally, the residue compositions on the consensus interfaces was used as indicators for multiple machine learning algorithms to predict if they can form homotypic interactions with each other. The overall performance of the cross-validation tests shows that our prediction accuracies are ranged between 0.6 and 0.8.

1. Introduction

Cells adapt to their surrounding environments by forming dynamic contacts with each other[1, 2]. The process of forming these intercellular contacts, called cell adhesion, is maintained by the molecular interactions between receptors expressed on surfaces of respective cells[3, 4]. The binding of cell surface receptors further triggers multiple intracellular signaling pathways[5] and finally leads to the phenotypic variation of cells[6]

*Corresponding authors: Yinghao Wu, Phone: (718) 678-1232, Fax: (718) 678-1018, yinghao.wu@einstein.yu.edu.

(Figure 1a). Based on these facts, it is appreciated that cell surface receptors are critical components in adhesion and signaling[7, 8]. Immunoglobulin (Ig) fold, as the largest and the most typical class of domains for cell surface recognition[9], are widely distributed in many types cell surface receptors that involve in adhesion and signaling[10]. The genes encoding domains that belong to Ig fold include both multigene and single-gene representatives[11]. They possess of common structural features of a β-sandwich framework (Figure 1e) with hypervariable loops[12]. Ig-fold cell surface receptors play essential roles in regulating diversified functions that are closely relevant to human health. For instance, the binding between Ig domains in T cell receptors (TCR) and major histocompatibility complex (MHC) triggers the T cell signaling pathway during immune response[13-16], while the differential binding between domains of specific types of cadherin superfamily are the major driving force of tissue morphogenesis during embryonic development[17]. As a result, the diverse functions of receptors in cell adhesion and signaling are conducted by the binding of their Ig-fold domains which share high structural similarity. The underlying question is: how functional diversity of Ig-fold domains is encoded in their common structural template.

There is a more straightforward way to ask above question: if it is possible to distinguish binding between different Ig domains based on their sequences which all end up with similar structures. One extreme case is to identify the Ig-fold domains that are directly involved in binding from others that have no binding targets. The reason for doing this is due to the fact that most cell surface receptors contain multiple copies of Ig-fold domains in their extracellular regions[18-20]. However, not all these domains are involved in the functional dimerization of receptors. For an example, the extracellular region of the Down syndrome cell adhesion molecule (DSCAM) consists of 10 Ig-like domains and 6 Fibronectin domains. Among these domains, only domains Ig2, Ig3 and Ig7 are directly involved in the intermolecular contact[21]. Generally, in order to elucidate the function of a cell surface receptor, it is necessary to determine which specific domains of the receptor are responsible for binding with its partner[22]. This task is traditionally accessed either by structural determination of entire complex of the receptor and its binding partner, which is only succeeded in a very few limited cases[23-25], or by carrying out site-directed mutagenesis on residues in each domain to detect whether these mutations can affect binding of the receptor[26-28]. Unfortunately, it is intractable to enumerate all possible domain combinations by brute-force screening every single residue, not to mention the fact that many of the receptor's binding targets are unknown. Therefore, current knowledge on binding of Ig-fold domains and their impacts on cell adhesion and signaling is very limited. Computational approaches that capable of offering predictive analysis to the functions of Ig-fold domains provide a complementary strategy. In practice, these methods can to a great extent reduce the complexity of experimental tests by providing a shortlist of functional domains in a receptor for verification. This can greatly facilitate our understanding to the molecular mechanism of receptors in cell adhesion and signaling. Cyrus Chothia and coworkers made pioneering analysis to the structural determinants in sequences of Ig domains[29, 30]. Many computational and bioinformatics efforts were followed on structural and functional characterization of domains in different superfamily of Ig fold[31-35]. To the best of our knowledge, a systematic evaluation to the structural similarity of the entire Ig fold, especially in the context of their functional allocation in cell adhesion and cell signaling, has not been documented.

In this article, we analyzed the structure-function relationship for protein domains in the entire Ig fold. All cell surface receptors that participate in cell-cell adhesion and signaling have been selected from the UniProt database[36]. All domains that belong to the Ig fold were curated from these proteins. These domains were classified into different groups based on their sequence and function similarity. For each specific group, a non-redundant structural dataset was constructed. Each item in the dataset contains not only the spatial coordinates, but also the information about its state in domain-domain interactions. Each group of protein domains shows different binding preference, corresponding to their cellular functions. Using structural alignment, we further built a common structural template for each group of domain dataset. By mapping the protein-protein binding interface of each domain in a group onto the surface of its common structural template, we found these binding interfaces are highly overlapped for each specific group. These overlapped interfaces, as we called consensus binding interfaces, are distinguishable among different groups of domains. We used the residue compositions on the consensus interfaces of an Ig-fold domain as indicators for multiple machine learning algorithms to predict if it can bind to different types of domain targets. Cross-validation results show that we achieved reasonably high accuracy for domains that are involved in homo-dimerization. Therefore, our study served as the first predictive method that can recognize homotypic binding between domains in different functional classes of Ig fold. The results brought valuable insights to the molecular mechanisms of protein-protein interactions in the diverse functions of cell surface receptors.

2. Models and Methods

2.1. Construct non-redundant datasets for Ig-fold domain structures

Our study mainly focuses on Ig-fold domains in membrane proteins which function as receptors in adhesion and signaling on cell surfaces. The first step is to generate the lists for all adhesion or signaling receptors by searching the UniProt database[36]. The next step is to find Ig-fold domains from the list by searching the Pfam database[37]. Ig-fold domains in most cell surface receptors are from two Pfam clans: the immunoglobulin superfamily (CL0011) and the immunoglobulin-like fold superfamily (CL0159). Structures of domains selected from these two clans were further obtained by downloading its three-dimensional coordinates from the Protein Data Bank (PDB)[38]. Different binding modes were recorded for all Ig-fold domains in the datasets. Specifically, if a domain has binding partner, three types of binding modes were denoted. The first is heterotypic binding (HETE) which indicate binding is formed between different types of protein (Figure 1b). The second is homotypic binding between different domains (HODD), which indicate the binding is formed through different domains of two proteins with the same type (Figure 1c). The third is homotypic binding between same domains (HOSD), which indicate the binding is formed between two proteins of the same type through the same domains (Figure 1d). It is worth mentioning that the binding model of a domain is not exclusive, considering it can have more than one binding partners.

There is repetitive information in the datasets derived from Pfam. The redundancy was removed in order to carry out statistically meaningful analysis on the datasets. As a result, 147 domains that belong to Ig superfamily (CL0011) were left for adhesion receptors.

Similarly, 112 domains were left for adhesion receptors that belong to Ig-like fold superfamily (CL0159). For signaling receptors, 149 domains from CL0011 were left and 47 domains that belong to CL0159 were left. The detailed information of final datasets can be downloaded online, and the procedure of dataset construction is in the supplemental documents.

2.2. Generate consensus binding interfaces of Ig-fold domains by structural alignment

A large portion of domains in the non-redundant datasets form interactions with domains in other proteins, through which these proteins can aggregate into hetero-oligomers or homo-oligomers. In order to understand the functional similarity among domains in the datasets, the first step is to structurally align domains together so that their binding interfaces can be quantitatively compared with each other. However, the structures of two domains cannot be directly superimposed together, considering the fact that each domain in the datasets has different number of amino acids. Moreover, there are small structural variations between different domains in terms of the length of each β-strand, connectivity of different strands in β-sheets and conformational diversity of each connecting loop. Therefore, we applied the algorithm TM-align to carry out the pairwise structural superposition between two domains[39].

Using TM-align as a tool for structural comparison, we selected a common structural template for each of the three datasets. Detailed procedure can be found in the supplemental documents. This structural template was used later as a platform to analyze the similarity of binding interfaces for different domain in the same dataset. After we selected the structural template for a dataset, we can project the binding interface of each domain in the dataset onto the corresponding residues of the template. Before the projection, residues that form intermolecular contacts with domains of other proteins in a complex were annotated as binding residues for all domains in three datasets. If the distance between any atom in the sidechain of a given residue and any atoms in other proteins of the complex is below the cutoff (5.5 Å), this residue will be marked as binding residue and become part of the binding interface of the corresponding domain. Otherwise, if no atom in the sidechain of a given residue forms contact with atoms in other proteins, this residue will be excluded from the binding interface of the domain. When all binding residues of a domain were identified, the entire binding interface was mapped to the structural template of the corresponding dataset. Specifically, TM-align was used to implement the superposition between the domain and the template. The algorithm generates maximal number of residue pairs between the domain and the template by dynamic programming to optimize the spatial superposition. As a result, if a residue in the domain belongs to part of the binding interface, its annotation of binding will be transformed to its paired reside in the template. The entire binding interface was projected to the template after transformations were carried out for all residue pairs.

In more detail, three types of binding interfaces are specifically differentiated, same as annotation used in *2.1.* Different types of interfaces can coexist in the same domain, considering that each domain can have more than one binding partners. For each domain, all types of binding interfaces were separately projected to the template. After projecting any of the three binding interfaces in all domains of a dataset to their corresponding structural

template, the consensus interface for all three different binding modes can be further constructed by calculating how many times each residue in the template is counted as part of these three binding interfaces. In detail, three variables were assigned to each residue in the template, corresponding to the frequencies of the residue that were involved in the three different interfaces. By superimposing a domain to the template, the values of variables for all residues were updated. The values were added by one to a variable of all residues that were marked as the corresponding type of binding interface in the aligned domain. Same was carried out for other two variables. After alignment was performed for all domains in the dataset, three profiles were generated along the residue index of the template, indicating how likely each residue was involved in the potential intermolecular interactions of three different types. Consequently, a specific type of consensus binding interface in a structural template is made up of all residues which corresponding variable has higher value than an empirically determined cutoff. In summary, we generated a structural template for each of the three non-redundant domain datasets and three types of consensus binding interfaces for each of the three structural templates.

2.3. Predict domain binding state by machine learning

Given the structural template and consensus binding interfaces for each of the three datasets, machine learning was applied to predict the binding state of an Ig-fold domain. There are three binding states for each domain. Each binding state is a binary signal indicating whether this domain forms interactions with domains of other proteins through the corresponding binding mode. While the binding state of a domain was the output of machine learning, the compositional vector of residues in the consensus binding interface was constructed as inputs for machine learning. In order to build the compositional vector of a query domain, the structure of the domain was aligned to the template of which dataset the query belongs to. This was done by TM-align. A list of residue pairs between the query and the template was attained after the structural alignment. The residues in the query were separately selected if their paired partners in the template were in one of the three consensus binding interfaces. For each interface of the query domain, the composition of selected residues constitutes the twenty-dimensional vector. Each dimension of the vector stands for the probability of finding a specific type of residue from the selected ones. Consequently, three vectors were derived for the query domain, corresponding to the three consensus binding interfaces. These vectors were the input indicators of machine learning to identify if the query domain interacts with other proteins through the corresponding binding mode.

After formatting the inputs and outputs, different algorithms of machine learning were tested to compare the prediction results, including back-propagation neural network (BPNN), support vector machine (SVM), and random forest (RF). Details of these algorithms are described in supplemental documents. In order to calibrate the behaviors of machine learning, cross-validation was separately applied to all three datasets. The leave-one-out strategy was used to avoid the potential over-fitting. In detail, three different processes of cross-validation were carried out for each dataset, corresponding to the test of three binding states. Each process consists of multiple runs of training, which is determined by the number of domains in each dataset. During each run of the leave-one-out training, one domain was selected from the dataset as the test, while the remaining domains were considered as the

training set. Domains in the training set were assigned into two groups based on the classification of the specific binding state. Both inputs and outputs of training set were fed into the three machine learning algorithms. After training, the residue compositional vector of selected test domain was used as input for prediction. The predicted outcome was compared with the real binding state. After above leave-one-out training was performed for all domains, the overall performance of individual machine learning algorithm to a specific dataset and binding state can be attained.

A weighted voting strategy was further proposed to make an integrative decision from machine learning outputs. Specifically, the integration of machine learning outputs for a given domain D were obtained by calculating the value $\sum_i w_i \delta_i(D)$, in which the summation i is carried out through all the three algorithms. The delta function $\delta_i(D)$is the binary signal of the corresponding machine learning output to the domain, which equals 1 if positive results are predicted, and −1 if negative results are predicted. The parameter w_i gives the weight of each machine learning algorithm in the voting, indicating their relative contributions to the final prediction. The range for each of the three weights is from 0 to 1. The positive or negative output of binding state from the voting, corresponding to with or without binding target under specific binding mode, depends on whether the calculated summation is larger or smaller than 0. In order to search for the best performance that the cross-validation can achieve, the weight space was discretized into small intervals (0.01) and the combinations of weights were then enumerated. The weights which optimized the cross-validation results are suggested to be used in the real test. Finally, a prediction program is available for download at: https://github.com/wujah/IgBDPredictor/. Detailed description of the package can be found in supplemental documents.

3. Results

3.1. Binding modes of Ig domains in adhesion and signaling receptors

Non-redundant datasets for different super-families of Ig-fold domains in adhesion and signaling receptors have been constructed from the integration of UniProt, Pfam and PDB databases. Each domain in these datasets might form different patterns of interactions with domains in other proteins. Three specific binding modes were designated to distinguish patterns in inter-molecular interactions. As defined in the ***Methods and Materials*** section ***2.1***, these modes are represented as HETE, HODD and HOSD. By calculating the atomic distances of inter-molecular residue pairs, all possibilities of these binding modes for a given domain in its complex were attained. The information of binding mode was collected for all domains in the three datasets. In order to compare domains in different super-families, or domains in receptors with different functional annotations, we carried out an overall statistical analysis on the likelihood of occurrence for each binding mode in each of the three datasets. The likelihood of occurrence for a given binding mode and dataset was simply derived by calculating the ratio of the number that this mode was observed through all domains in the dataset versus the total number of domains in the dataset. Consequently, the likelihoods of all three binding modes are plotted as histogram in Figure 2 for all three datasets.

In general, Figure 2 shows that a large portion of domains in all three datasets forms contacts with other proteins. For instance, over 60% domains of Ig superfamily in adhesion receptors (left columns of Figure 2) or signaling receptors (middle columns of Figure 2) interact with other proteins, while inter-molecular contacts were found in over 50% domains of Ig-like fold superfamily in adhesion receptors (right columns of Figure 2). This indicates that inter-molecular interactions play a significant role in functions of proteins that contain Ig-fold domains. Among the Ig superfamily domains of adhesion receptors, 35% of them form inter-molecular contacts through the HOSD binding mode, 23% through the HODD mode and only 9% through the HETE mode. Similarly, among the Ig-like fold superfamily domains of adhesion receptors, 26% of them form inter-molecular contacts through the HOSD binding mode, 28% through the HODD mode and only 2% through the HETE mode. Therefore, homotypic interactions are much more commonly observe in domains of adhesion receptors, no matter if they belong to Ig superfamily or Ig-like fold superfamily. The homogeneous binding between proteins of the same family is a common feature of cell adhesion molecules (CAM). These homotypic interactions are the basis of many physiological processes, such as embryonic development and tissue morphogenesis. More specifically, it is interesting to find that Ig superfamily domains prefer binding through the HOSD mode. On the contrary, Ig-like fold superfamily domains prefer binding through the HODD mode. This observation is consistent with a number of examples in which binding between different domains that belong to Ig-like fold is formed in the crystal structure of different adhesion receptors systems, including proto-cadherin[40] and receptor protein tyrosine phosphatase (RPTP)[41].

In contrast to the adhesion receptors, domains in signaling receptors show different modes of binding. In specific, a much higher portion of domains in signaling receptors are involved in heterotypic interactions than domains in adhesion receptors. As shown in Figure 2a, comparing with 9% in Ig superfamily and 2% in Ig-like fold superfamily of adhesion domains, 26% signaling domains form inter-molecular contacts through the HETE binding mode. The preference of signaling domains in heterotypic binding is resulted from their functions in cells. Different from homotypic interactions between domains of adhesion receptors which connect cells of the same type, the heterotypic interactions are formed between domains in signaling receptors and their extracellular ligands. This asymmetric mode of binding initiates the process of cell signal transduction. Although a large portion of signaling domains are involved in heterotypic interactions, it is interesting to find that there is still 32% of them form inter-molecular contacts through the HOSD binding mode (middle column of Figure 2c). This binding is formed between signaling receptors on surface of the same cell. There are examples existing in our dataset. In some cases, homo-dimerization is an initial step to their activation and ligand binding. For instance, the back-to-back binding between the membrane proximal Ig domains of immune-type receptor glycoprotein VI[42] (PDB id 2GI7) provide a structural basis of this receptor in signaling responses to ligand collagen. In contrast, many other receptors containing Ig superfamily domains perform their functions through homo-dimerization after they are activated by ligand-binding. One classic example is the receptor of human growth hormone[43] (PDB id 3HHR), in which a growth hormone ligand simultaneously bind with two receptors. In another example, dimer is

directly formed between the extracellular ligand binding regions of fibroblast growth factor receptors[44] (PDB id 1FQ9).

In summary, we carried out statistical studies on large-scale datasets of Ig-fold domains. The binding preference of these domains in adhesion receptors shows distinctive patterns from domains in signaling receptors. The differences of binding preference are originated from the functions of these domains in cell adhesion and signaling. Therefore, we show that functional characteristics of membrane receptors can be reflected from the structural basis of domain interactions, which will be further justified in the following parts.

3.2. Structural characteristic and function diversity of Ig domains

The Ig fold usually consists of 7 to 10 β-strands. The index of these strands is designated from letter *A* to letter *G*, as shown in Figure 1e. Depending on the arrangement of these strands, domains that belong to Ig fold can be classified into different groups. For instance, there are four major sets in Ig superfamily: the V-set, I-set, C1-set and C2-set, in which the I-set is a truncated V-set without the *C'* and *C"* strands (Figure 1e). In each Ig fold domain, two β-sheets form a β-sandwich framework. One sheet that contains strands *C*, *F* and *G* is called *CFG* face, while the other that contains strands *B*, *E* and *D* is called *BED* face (Figure 1e).

In order to find the common structural template, the method TM-align was applied to all pairs of domains in the datasets. The average TM-score was calculated for each domain after it was aligned to all other domains in the dataset. The domain which has the highest average score was selected as the template of the dataset. This procedure was carried out for all three datasets. Consequently, the N-terminal domain of human nectin-3 becomes the template of the dataset for adhesion domains of Ig superfamily. The domain is ranged from residue 61 to residue 167 of the protein[45] (PDB id 4FOM). The average TM-score of the template is 0.727, while a median value of this average score throughout the dataset is 0.641. According to the definition of TM-score, two proteins are usually considered to have global structural similarity if they have a score higher than 0.5[46]. Therefore, the average score of our selected template indicates that it shares significant portion of structure with all other domains in the dataset. We further plot a profile in Figure 3a, the number of which describes how many alignments were found for each residue in the template. The aligned number *n* for residue *i* in the profile means this residue was paired to residues in *n* domains during alignment. Figure 3a shows that most parts of the template, especially the two β-sheets, can be aligned to almost all domains in the dataset. The only regions with high structural variations are located close to the *C'* and *C"* strands that are highlighted with the color code of blue in Figure 3b. These regions are well known to be highly variable between different families of Ig fold[47]. Therefore, our results suggest that it is robust to use this domain as the structural representation of the dataset. For other two datasets, similar results were attained. The N-terminal domain of programmed cell death protein 1 becomes the template of the dataset for signaling domains of Ig superfamily. The domain is ranged from residue 35 to residue 145 of the protein (PDB id 5IUS). The average TM-score of the template is 0.67, while a median value of this average score throughout the dataset is 0.61. Finally, the second ectodomain of human N-cadherin becomes the template of the dataset for adhesion domains of Ig-like fold

superfamily. The domain is ranged from residue 113 to residue 214 of the protein (PDB id 3Q2W). The average TM-score of the template is 0.68, while a median value of this average score throughout the dataset is 0.61. We also calculated the average TM-score for domains of one dataset with the template from another different dataset. The results of these cross-dataset TM-scores are listed in Table S1 of supplemental documents. As shown from the table, the TM-scores calculated by templates within the same datasets are always higher than those using templates of other datasets.

As shown in Figure 3a, residues in different domains were aligned to the same position of the template. In order to check the composition of these aligned residues in different positions of the template, we calculated the participation ratio (PR) of each position to quantify its conservation of residue type. The PR[48, 49] is defined in supplemental documents. The value of PR is ranged between 0 and 1. A higher value indicates that residue types are more conservative at the corresponding position of the template. On the other hand, a lower value means that aligned residue at the position are highly variated. As a result, the profile of participation ratio is plotted in Figure 3c across all positions of the template. This plot is for adhesion domains of Ig superfamily. The sidechains of residues with the highest PR values are highlighted in the atomic representation in Figure 3d. The figure shows that some residues are highly conserved among all alignments. These residues can be grouped into three classes. A pair of salt bridge is highlighted in orange, while a di-sulfide bond is highlighted in green. Additionally, a hydrophobic core that consists of several less conserved non-polar residues is highlighted in red of Figure 3d. These sequence signatures from our structural comparison are consistent with previous analysis using sequence and structural alignments[10, 50]. We suggest that these conserved residues are the most important building blocks that stabilize the overall structural features of domains in Ig superfamily. Our results therefore reveal the physical chemical basis to the topological feature of this domain superfamily.

After the construction of structural template for each dataset, the frequency of being part of the binding interface was further calculated for each residue in the template by transferring the binding residues from each domain. In Figure 4 we projected these frequencies onto structures of the templates. In the figure, the frequency profiles are represented by transparent surfaces with color code, and the backbones of templates are in grey. The residues of high frequency are shown by blue regions of the surfaces, corresponding to the potential binding interfaces. The low frequency residues are shown in red, corresponding to the regions that are less likely involved in binding interfaces. The distributions of frequency for different types of binding interfaces in different datasets are plotted. The structures in the figure are positioned along the same orientation based on the connectivity of their β-strands. Figure 4 shows that binding interfaces are located at specific regions of template surfaces, while different types of interfaces in different datasets are highly distinctive.

In detail, the heterotypic binding interfaces (HETE) in Ig superfamily (CL0011) domain template of adhesion receptors are mainly concentrated on the surface of its *CFG* face (the left β-sheet in Figure 4a). In comparison, the homotypic binding interfaces (HOSD) in the same domain template are more extensively distributed on the surface of both *CFG* face and *BED* face (Figure 4b). Different from adhesion domains, the HETE binding interfaces in Ig

superfamily domains of signaling receptors are mainly concentrated on the surface of its *BED* face (the right β-sheet in Figure 4c). It is worth of mentioning that molecular recognitions of antibodies[51] and TCR[52] are the most important functions of Ig superfamily in signaling. The so-called complementarity determining region (CDR) loops[53] in the V-set domains of these proteins are responsible for recognizing their binding targets, which normally don't belong to Ig fold, such as MHC for TCR[54] and HIV-1 gp120 envelope glycoprotein for Antibody VRC-PG04[55]. The regions of these CDR loops, however, were not strongly highlighted here in Figure 4c. This is partly because most antibodies are not membrane proteins and not included in the datasets. The binding propensity of CDR loops can be reflected if we extend our current study by including all Ig domains that are not in membrane receptors. Finally, the HOSD interfaces in adhesion domain belonging to Ig-like fold superfamily (CL0159) are shown in Figure 4d. Similar to Ig superfamily, the HOSD interfaces of Ig-like fold superfamily are also extensively distributed. However, instead of being on the surfaces of β-sheets in Figure 4b, the binding interfaces in Figure 4d are mainly distributed on the side edges of β-sheets. Overall, these results suggest that homotypic interactions between Ig domains are formed through more extensive binding interfaces than heterotypic interactions. The heterotypic binding interfaces of Ig domains are more specifically evolved. Moreover, distinctive regions of Ig domains are used as binding interfaces when they are functioned in adhesion and signaling. Therefore, the functional diversity of Ig fold domains is reflected by the distinguish patterns of their binding interfaces, although structural features are highly conserved across different families of these domains.

3.3. The prediction accuracy of homotypic interactions between Ig domains

After calculating the frequency of being at binding interfaces for each residue in a template, the consensus interface of a specific binding mode was constructed as the region in which the frequency of all residues is higher than a predetermined cutoff value. We used the top 20% highest-frequency residues as the relative cutoff value. The residue composition in the consensus binding interface was selected as the input features to train different machine learning algorithms. We tested if these machine learning algorithms can recognize the signal of domain-domain interactions through the HOSD mode. Leave-one-out cross-validations were carried out for all three datasets, as described in the ***Methods***. In order to calibrate the performance of the cross-validation, the sensitivity, specificity, precision and overall accuracy were calculated from the testing results, as defined in supplemental documents.

The overall performance of our testing results is listed in Table 1. The table shows that the accuracies are ranged between 0.6 and 0.8 for all three datasets. The accuracies were accompanied by the qualified values of specificities and precisions, while the sensitivities are relatively low but still on the reasonable levels. The cross-validation results therefore suggest that our machine learning algorithms are able to recognize the HOSD binding mode for Ig domains in different functional groups and super-families. The high accuracy and specificity indicates the reliability of our machine learning method, while the reason of the low sensitivity will be discussed in the next paragraph. Moreover, a weighted voting strategy was proposed to integrate the test results from all three machine learning algorithms. We found that under an optimal combination, this voting mechanism can improve the

sensitivities, while maintaining the overall specificities and accuracies. The values of derived weights for SVM, RF and BPNN are 0.12, 0.8 and 0.9, respectively. Especially, for domains of Ig superfamily in adhesion receptors, we attained the final sensitivity 0.51, with the specificity 0.77, precision 0.57 and accuracy 0.69.

It is well-known that protein functions are more conserved in structure space than sequence space. In order to evaluate how much more information can be gained from structure alignment on the basis of sequence similarity, comparable predictions were performed by changing the cutoff value of highest-frequency interface residues as the inputs of machine learning algorithm. As a result, if we included more residues with lower-frequency as part of the consensus binding interfaces, the testing results became worse. Specifically, if we use the top 100% highest-frequency residues, all sequences of a domain will be included as inputs and there will be no information on binding interface as structural guidance for prediction. In this case, the prediction becomes a purely sequence-based method with the same machine learning algorithms and the same format of inputs and outputs. Using this sequence-based method, we attained the final results from weighted voting with the sensitivity 0.37, specificity 0.78, precision 0.47 and accuracy 0.63 for the dataset of adhesion receptors of Ig superfamily. Relative to the prediction against the same dataset with the sensitivity 0.51, specificity 0.77, precision 0.57, accuracy 0.69 in the original test which only contains top 20% residues in the binding interface, the purely sequence-based method resulted in much lower sensitivity, precision and accuracy. This indicates that strong sequence signals of homotypic binding between Ig domains are located on their binding surfaces. Binding is very sensitive to the sequence variations at these regions.

Because three machine learning methods use the same format of inputs and give the same format of outputs, the pairwise correlation coefficients between outputs from any of the two methods were calculated to investigate how they scored differently. Considering that the outputs from the machine learning methods are binary variables, whether or not a domain has binding partners, we used φ coefficient to quantify the correlation between different methods. The φ coefficient is defined as $\varphi = \left(n_{11}n_{00} - n_{10}n_{01}\right)/\sqrt{n_{1\bullet}n_{0\bullet}n_{\bullet 1}n_{\bullet 0}}$. In this equation, n_{11} and n_{00} indicate the numbers that both methods give positive or negative outputs, while n_{10} and n_{01} indicate the numbers that one method gives positive outputs and the other gives negative outputs. In the denominator, $n_{1\bullet}$ and $n_{\bullet 1}$ indicate the number of positive outputs from each of the two methods, and $n_{0\bullet}$ and $n_{\bullet 0}$ indicate the number of respective negative outputs. The value of coefficient is ranged from −1 to +1, whereas +1 stands for the strongest positive correlation, 0 means no correlation, and −1 indicates the strongest negative correlation. Consequently, the φ coefficient between support vector machine and random forest is 0.41. The φ coefficient between support vector machine and BP neural network is 0.54. The φ coefficient between random forest and BP neural network is 0.51. Therefore, out results show that outputs from different machine learning methods are all positively correlated, although the correlations are not very strong. This explains why the consensus voting of these positively, but weakly correlated methods can further improve the final predictions.

Comparing the results of the HOSD mode, our tests on the other two modes led into much lower accuracies. This is due to the fact that binding through the HOSD mode is formed by two domains of the same types. They share the same consensus binding interface. Therefore, information is sufficient to predict if they can interact with each other. On the contrary, binding through the HETE or HODD modes is formed by two domains of different types. The information from both interfaces should be involved as inputs of machine learning. In order to predict if two specific domains interact, machine learning inputs that contain residue compositional vectors from binding interfaces of both domains will be taken into accounts of future improvement.

Some specific cases of our test results are shown in Figure 5. The domains in the dataset for cross-validation are colored in red, while their predicted binding partners are colored in green. The rest parts of complexes are colored in grey. Figure 5a and Figure 5b are two examples of Ig domains that form homotypic interactions and were correctly recognized by our methods. A pair of N-terminal domains from Nectin-2[56] is shown in Figure 5a (PDB id 4HZA). Nectin-2 proteins at cell interfaces were found to homo-dimerize through these domains. Similarly, the crystal structure of a SYG-1 homodimer is shown in Figure 5b (PDB id 4OF3). SYG-1 is a type of multipurpose cell adhesion molecule participating in diverse physiological functions such as synapse formation[57]. Dimers are formed through the N-terminal domains of these proteins. In both cases of Nectin-2 and SYG-1, their binding modes were successfully predicted by only using the information of residue composition at their binding interfaces. Comparing with these two examples, Figure 5c and Figure 5d plot the protein complexes in which their binding modes were not correctly recognized by our methods. The crystal structure of neural cell adhesion molecules NCAM2 is shown in Figure 5c (PDB id 2WIM). In the cross-validation, they were predicted as forming homodimers through the HOSD binding mode of their N-terminal domains (red and green domains in Figure 5c). In reality, although NCAM2 proteins form homodimers, binding is taken place through the HODD mode between the N-terminal domain of one protein and the second domain of the other protein, as shown in Figure 5c[58]. Therefore, our method successfully recognized the binding state of the domain through its interface, but missed the correct binding mode. Similar cases exist in datasets as one important source of false positive during the cross-validation. Finally, the Protein Tyrosine Phosphatase δ is shown in Figure 5d (PDB id 2YD7). The prediction from our method showed that the N-terminal domain of this domain is a monomer. However, a homodimer that is formed through the HOSD mode between N-terminal domains is found in the crystal structure. In reality, it is found that this dimer is not functional under the physiological condition[59]. The dimer shown in the crystal structure is more likely an artificial complex through crystal packing. Therefore, the output of our methods on this case should be regarded as a correct recognition. We noticed this is one important source of false negative which resulted in the low sensitivity of our cross-validation results. Cautious removal of those artificial complexes will thus largely improve the sensitivity of our test. Future improvement should include methods that can identify biological binding interfaces from the crystal packing.

4. Concluding Discussions

Domains that belong to Ig fold are widely distributed in a large variety of cell surface receptors. Functions of these receptors in adhesion and signaling are fulfilled through the specific patterns of binding between their domains in the extracellular regions. Moreover, although most receptors contain multiple copies of Ig domains, only a small number of them are directly involved in binding. This is, however, an important feature of cell surface receptors. Receptors need to present their binding sites away from plasma membrane surfaces in order to reach out for their ligand. Therefore, due to the large inter-cellular distance that is commonly observed during cell adhesion, most extracellular domains of receptors serve as building blocks to support a small number of ligand-binding domains, which are normally located at the N-terminus of receptors. It is worth of mentioning that the I-set of Ig domains or the domains from FN fold often play roles as building blocks, so they naturally do not engage in recognition. In contrast, the V-set of Ig domains has high tendency to appear at the N-terminus of receptors and engage in ligand recognition.

This heterogeneity in binding of different structurally similar domains leads to the functional diversity of receptors, and increases the difficulties in decoding the molecular mechanisms of these proteins in cells. In order to bridge the structural and functional characteristics of domains in Ig fold, we constructed non-redundant structural datasets for Ig-fold domains specifically functioned in cell adhesion and signaling. We found that datasets of domains in adhesion receptors show different binding preference from domains in signaling receptors. The preference is resulted from the cellular functions of these receptors. A common structural template was further been constructed for each group of domain dataset. Comparing the template with each domain in the dataset, we found that some regions in the template are highly variated in structures, while some other residues are highly conserved in sequences across all domains in the dataset. These findings bring insights to the design of new protein sequences of this specific fold. After the construction of structural template for each dataset, the protein-protein binding interfaces of each domain in the dataset were projected onto the surface of the template. We found that distinctive regions of Ig domains are used as binding interfaces when they are functioned in adhesion and signaling, while the heterotypic binding interfaces are more specifically evolved, comparing with the more extensively distributions of homotypic binding interfaces. Finally, the residue compositions on the consensus interfaces of Ig-fold domains was used as indicators for multiple machine learning algorithms to predict if they can form homotypic interactions with each other.

The accuracies of our predictions are ranged between 0.6 and 0.8 for all datasets. The accuracies are high relative to the predictions based on random guesses, which generally lead to the accuracy of 0.5. However, it is worth of noting that the high accuracies in the study are accompanied with the relatively low sensitivities, which is due to the fact that a number of artificial complexes exist in the current datasets through crystal packing, as discussed in the results. Especially, for adhesion domains belonging to Ig-like fold superfamily (CL0159), the SVM algorithm gave a fairly good accuracy (0.71), although the method can only identify a marginal number of domains which actually exhibit a HOSD binding mode (sensitivity equals 0.01). Furthermore, if none of the domains is predicted to have a HOSD binding mode, high accuracy will be attained for all three datasets (0.65 for

the dataset of adhesion domains from CL0011, 0.68 for the dataset of signaling domains from CL0011, and 0.74 for the dataset of adhesion domains from CL0159). The high accuracies of such extreme prediction are biased by the observation that the majority of domains in the datasets don't have binding partners with HOSD mode. Moreover, the accuracies are compensated by absolutely low (0.0) sensitivity and precision. An ideal prediction should be a good strategy to balance the scores of sensitivity and specificity. In that sense, the positive likelihood ratio (LR+), which is defined as the ratio between sensitivity and 1-specificity, can be used to quantify the performance of a prediction. The LR + with a value higher than 1 indicates the prediction is better than random guess. Consequently, there will be no predictive significance by assuming no domain in a dataset has HOSD binding mode, although this assumption will result in high accuracy. In contrast, Table 1 shows that, except the SVM prediction on the dataset of adhesion domains from CL0159, the LR+ of all others are higher than 1, indicating that our prediction results are meaningful. In general, our study provides comprehensive evaluations to the structural function relationship of domains in the entire Ig fold. The machine-learning-based prediction could be a useful tool to recognize homotypic binding between Ig domains in specific functional classes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported in part by the National Institutes of Health (Grant No. R01GM120238) and a start-up grant from Albert Einstein College of Medicine. Computational support was provided by Albert Einstein College of Medicine High Performance Computing Center.

References

1. Li S. Mechanisms of Cellular Signal Transduction. Int J Biol Sci. 2005; 1:152. [PubMed: 16432552]
2. Gumbiner BM. Cell Adhesion: The Molecular Basis of Tissue Architecture and Morphogenesis. Cell. 1996; 84:345–357. [PubMed: 8608588]
3. Packard B. Receptor Phosphorylation and Signal Transduction across Plasma-Membranes. Trends in Biochemical Sciences. 1985; 10:138–138.
4. Ullrich A, Schlessinger J. Signal Transduction by Receptors with Tyrosine Kinase-Activity. Cell. 1990; 61:203–212. [PubMed: 2158859]
5. Krauss RS. Regulation of Promyogenic Signal Transduction by Cell-Cell Contact and Adhesion. Experimental Cell Research. 2010; 316:3042–3049. [PubMed: 20471976]
6. Lalli E, Sassonecorsi P. Signal-Transduction and Gene-Regulation - the Nuclear Response to Camp. Journal of Biological Chemistry. 1994; 269:17359–17362. [PubMed: 8021233]
7. Kim H, Cruz M, Bourdeau A, Dumont DJ. Cell-Cell Interactions Influence Vascular Reprogramming by Prox1 During Embryonic Development. Plos One. 2013; 8
8. Burdick MM, McCarty OJ, Jadhav S, Konstantopoulos K. Cell-Cell Interactions in Inflammation and Cancer Metastasis. Ieee Engineering in Medicine and Biology Magazine. 2001; 20:86–91. [PubMed: 11446216]
9. Williams AF, Barclay AN. The Immunoglobulin Superfamily–Domains for Cell Surface Recognition. Annu Rev Immunol. 1988; 6:381–405. [PubMed: 3289571]
10. Halaby DM, Poupon A, Mornon J. The Immunoglobulin Fold Family: Sequence Analysis and 3d Structure Comparisons. Protein Eng. 1999; 12:563–571. [PubMed: 10436082]

11. Hunkapiller T, Hood L. Diversity of the Immunoglobulin Gene Superfamily. Adv Immunol. 1989; 44:1–63. [PubMed: 2646860]
12. Bork P, Holm L, Sander C. The Immunoglobulin Fold. Structural Classification, Sequence Patterns and Common Core. J Mol Biol. 1994; 242:309–320. [PubMed: 7932691]
13. Yokosuka T, , Saito T. Immunological Synapse Vol. 340. Springer-Verlag Berlin; Berlin: 2010 The Immunological Synapse, Tcr Microclusters, and T Cell Activation; 81107
14. Hashimoto-Tane A, Yokosuka T, Ishihara C, Sakuma M, Kobayashi W, Saito T. T-Cell Receptor Microclusters Critical for T-Cell Activation Are Formed Independently of Lipid Raft Clustering. Molecular and Cellular Biology. 2010; 30:3421–3429. [PubMed: 20498282]
15. Huang J, Zarnitsyna VI, Liu BY, Edwards LJ, Jiang N, Evavold BD, Zhu C. The Kinetics of Two-Dimensional Tcr and Pmhc Interactions Determine T-Cell Responsiveness. Nature. 2010; 464:932–U156. [PubMed: 20357766]
16. Fooksman DR, Vardhana S, Vasiliver-Shamis G, Liese J, Blair DA, Waite J, Sacristan C, Victora GD, Zanin-Zhorov A, Dustin ML. Functional Anatomy of T Cell Activation and Synapse Formation. Annu Rev Immunol. 2010; 28:79–105. [PubMed: 19968559]
17. Gumbiner BM. Regulation of Cadherin-Mediated Adhesion in Morphogenesis. Nat Rev Mol Cell Biol. 2005; 6:622–634. [PubMed: 16025097]
18. Johnson CP, Fujimoto I, Perrin-Tricaud C, Rutishauser U, Leckband D. Mechanism of Homophilic Adhesion by the Neural Cell Adhesion Molecule: Use of Multiple Domains and Flexibility. Proceedings of the National Academy of Sciences of the United States of America. 2004; 101:6963–6968. [PubMed: 15118102]
19. Kocherlakota KS, Wu JM, McDermott J, Abmayr SM. Analysis of the Cell Adhesion Molecule Sticks-and-Stones Reveals Multiple Redundant Functional Domains, Protein-Interaction Motifs and Phosphorylated Tyrosines That Direct Myoblast Fusion in Drosophila Melanogaster. Genetics. 2008; 178:1371–1383. [PubMed: 18245830]
20. Ranheim TS, Edelman GM, Cunningham BA. Hemophilic Adhesion Mediated by the Neural Cell Adhesion Molecule Involves Multiple Immunoglobulin Domains. Proceedings of the National Academy of Sciences of the United States of America. 1996; 93:4071–4075. [PubMed: 8633018]
21. Zhu K, Xu YL, Liu JH, Xu Q, Ye HH. Down Syndrome Cell Adhesion Molecule and Its Functions in Neural Development. Neurosci Bull. 2011; 27:45–52. [PubMed: 21270903]
22. Katsamba P, Carroll K, Ahlsen G, Bahna F, Vendome J, Posy S, Rajebhosale M, Price S, Jessell TM, Ben-Shaul A, Shapiro L, Honig BH. Linking Molecular Affinity and Cellular Specificity in Cadherin-Mediated Adhesion. Proc Natl Acad Sci U S A. 2009; 106:11594–11599. [PubMed: 19553217]
23. Pronker MF, Lemstra S, Snijder J, Heck AJ, Thies-Weesie DM, Pasterkamp RJ, Janssen BJ. Structural Basis of Myelin-Associated Glycoprotein Adhesion and Signalling. Nature communications. 2016; 7:13584.
24. Seiradake E, Harlos K, Sutton G, Aricescu AR, Jones EY. An Extracellular Steric Seeding Mechanism for Eph-Ephrin Signaling Platform Assembly. Nature Structural & Molecular Biology. 2010; 17:398–U327.
25. Goodman KM, Yamagata M, Jin X, Mannepalli S, Katsamba PS, Ahlsen G, Sergeeva AP, Honig B, Sanes JR, Shapiro L. Molecular Basis of Sidekick-Mediated Cell-Cell Adhesion and Specificity. Elife. 2016; 5
26. Goodman KM, Rubinstein R, Thu CA, Bahna F, Mannepalli S, Ahlsen G, Rittenhouse C, Maniatis T, Honig B, Shapiro L. Structural Basis of Diverse Homophilic Recognition by Clustered Alpha- and Beta-Protocadherins. Neuron. 2016; 90:709–723. [PubMed: 27161523]
27. Goodman KM, Rubinstein R, Thu CA, Mannepalli S, Bahna F, Ahlsen G, Rittenhouse C, Maniatis T, Honig B, Shapiro L. Gamma-Protocadherin Structural Diversity and Functional Implications. Elife. 2016; 5
28. Harrison OJ, Brasch J, Lasso G, Katsamba PS, Ahlsen G, Honig B, Shapiro L. Structural Basis of Adhesive Binding by Desmocollins and Desmogleins. Proceedings of the National Academy of Sciences of the United States of America. 2016; 113:7160–7165. [PubMed: 27298358]
29. Chothia C, Gelfand I, Kister A. Structural Determinants in the Sequences of Immunoglobulin Variable Domain. J Mol Biol. 1998; 278:457–479. [PubMed: 9571064]

30. Harpaz Y, Chothia C. Many of the Immunoglobulin Superfamily Domains in Cell Adhesion Molecules and Surface Receptors Belong to a New Structural Set Which Is Close to That Containing Variable Domains. J Mol Biol. 1994; 238:528–539. [PubMed: 8176743]
31. Rubinstein R, Ramagopal UA, Nathenson SG, Almo SC, Fiser A. Functional Classification of Immune Regulatory Proteins. Structure (London, England: 1993). 2013; 21:766–776.
32. Yap EH, Fiser A. Protlid, a Residue-Based Pharmacophore Approach to Identify Cognate Protein Ligands in the Immunoglobulin Superfamily. Structure (London, England: 1993). 2016; 24:2217–2226.
33. Yap EH, Rosche T, Almo S, Fiser A. Functional Clustering of Immunoglobulin Superfamily Proteins with Protein-Protein Interaction Information Calibrated Hidden Markov Model Sequence Profiles. Journal of molecular biology. 2014; 426:945–961. [PubMed: 24246499]
34. Chailyan A, Tramontano A, Marcatili P. A Database of Immunoglobulins with Integrated Tools: Digit. Nucleic acids research. 2012; 40:D1230–1234. [PubMed: 22080506]
35. Ye J, Ma N, Madden TL, Ostell JM. Igblast: An Immunoglobulin Variable Domain Sequence Analysis Tool. Nucleic acids research. 2013; 41:W34–40. [PubMed: 23671333]
36. Uniprot: The Universal Protein Knowledgebase. Nucleic acids research. 2017; 45:D158–d169. [PubMed: 27899622]
37. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A. The Pfam Protein Families Database: Towards a More Sustainable Future. Nucleic acids research. 2016; 44:D279–285. [PubMed: 26673716]
38. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic acids research. 2000; 28:235–242. [PubMed: 10592235]
39. Zhang Y, Skolnick J. Tm-Align: A Protein Structure Alignment Algorithm Based on the Tm-Score. Nucleic acids research. 2005; 33:2302–2309. [PubMed: 15849316]
40. Rubinstein R, Thu CA, Goodman KM, Wolcott HN, Bahna F, Mannepalli S, Ahlsen G, Chevee M, Halim A, Clausen H, Maniatis T, Shapiro L, Honig B. Molecular Logic of Neuronal Self-Recognition through Protocadherin Domain Interactions. Cell. 2015; 163:629–642. [PubMed: 26478182]
41. Aricescu AR, Siebold C, Choudhuri K, Chang VT, Lu W, Davis SJ, van der Merwe PA, Jones EY. Structure of a Tyrosine Phosphatase Adhesive Interaction Reveals a Spacer-Clamp Mechanism. Science. 2007; 317:1217–1220. [PubMed: 17761881]
42. Horii K, Kahn ML, Herr AB. Structural Basis for Platelet Collagen Responses by the Immune-Type Receptor Glycoprotein Vi. Blood. 2006; 108:936–942. [PubMed: 16861347]
43. de Vos AM, Ultsch M, Kossiakoff AA. Human Growth Hormone and Extracellular Domain of Its Receptor: Crystal Structure of the Complex. Science. 1992; 255:306–312. [PubMed: 1549776]
44. Schlessinger J, Plotnikov AN, Ibrahimi OA, Eliseenkova AV, Yeh BK, Yayon A, Linhardt RJ, Mohammadi M. Crystal Structure of a Ternary Fgf-Fgfr-Heparin Complex Reveals a Dual Role for Heparin in Fgfr Binding and Dimerization. Molecular cell. 2000; 6:743–750. [PubMed: 11030354]
45. Harrison OJ, Vendome J, Brasch J, Jin X, Hong S, Katsamba PS, Ahlsen G, Troyanovsky RB, Troyanovsky SM, Honig B, Shapiro L. Nectin Ectodomain Structures Reveal a Canonical Adhesive Interface. Nature structural & molecular biology. 2012; 19:906–915.
46. Xu J, Zhang Y. How Significant Is a Protein Structure Similarity with Tm-Score = 0.5? Bioinformatics (Oxford, England). 2010; 26:889–895.
47. Wang J, Springer TA. Structural Specializations of Immunoglobulin Superfamily Members for Adhesion to Integrins and Viruses. Immunol Rev. 1998; 163:197–215. [PubMed: 9700512]
48. Wu YH, Yuan XZ, Gao X, Fang HP, Zi J. Universal Behavior of Localization of Residue Fluctuations in Globular Proteins. Physical Review E. 2003; 67:4.
49. Kramer B, Mackinnon A. Localization - Theory and Experiment. Reports on Progress in Physics. 1993; 56:1469–1564.
50. Wang JH. The Sequence Signature of an Ig-Fold. Protein Cell. 2013; 4:569–572. [PubMed: 23842991]
51. Wilson IA, Stanfield RL. Antibody-Antigen Interactions: New Structures and New Conformational Changes. Curr Opin Struct Biol. 1994; 4:857–867. [PubMed: 7536111]

52. Garcia KC, Degano M, Pease LR, Huang M, Peterson PA, Teyton L, Wilson IA. Structural Basis of Plasticity in T Cell Receptor Recognition of a Self Peptide-Mhc Antigen. Science. 1998; 279:1166–1172. [PubMed: 9469799]

53. Finlay WJ, Almagro JC. Natural and Man-Made V-Gene Repertoires for Antibody Discovery. Front Immunol. 2012; 3:342. [PubMed: 23162556]

54. Garcia KC, Adams EJ. How the T Cell Receptor Sees Antigen–a Structural View. Cell. 2005; 122:333–336. [PubMed: 16096054]

55. Joyce MG, Kanekiyo M, Xu L, Biertumpfel C, Boyington JC, Moquin S, Shi W, Wu X, Yang Y, Yang ZY, Zhang B, Zheng A, Zhou T, Zhu J, Mascola JR, Kwong PD, Nabel GJ. Outer Domain of Hiv-1 Gp120: Antigenic Optimization, Structural Malleability, and Crystal Structure with Antibody Vrc-Pg04. J Virol. 2013; 87:2294–2306. [PubMed: 23236069]

56. Samanta D, Ramagopal UA, Rubinstein R, Vigdorovich V, Nathenson SG, Almo SC. Structure of Nectin-2 Reveals Determinants of Homophilic and Heterophilic Interactions That Control Cell-Cell Adhesion. Proceedings of the National Academy of Sciences of the United States of America. 2012; 109:14836–14840. [PubMed: 22927415]

57. Ozkan E, Chia PH, Wang RR, Goriatcheva N, Borek D, Otwinowski Z, Walz T, Shen K, Garcia KC. Extracellular Architecture of the Syg-1/Syg-2 Adhesion Complex Instructs Synaptogenesis. Cell. 2014; 156:482–494. [PubMed: 24485456]

58. Kulahin N, Kristensen O, Rasmussen KK, Olsen L, Rydberg P, Vestergaard B, Kastrup JS, Berezin V, Bock E, Walmod PS, Gajhede M. Structural Model and Trans-Interaction of the Entire Ectodomain of the Olfactory Cell Adhesion Molecule. Structure (London, England: 1993). 2011; 19:203–211.

59. Coles CH, Shen Y, Tenney AP, Siebold C, Sutton GC, Lu W, Gallagher JT, Jones EY, Flanagan JG, Aricescu AR. Proteoglycan-Specific Molecular Switch for Rptpsigma Clustering and Neuronal Extension. Science (New York, NY). 2011; 332:484–488.

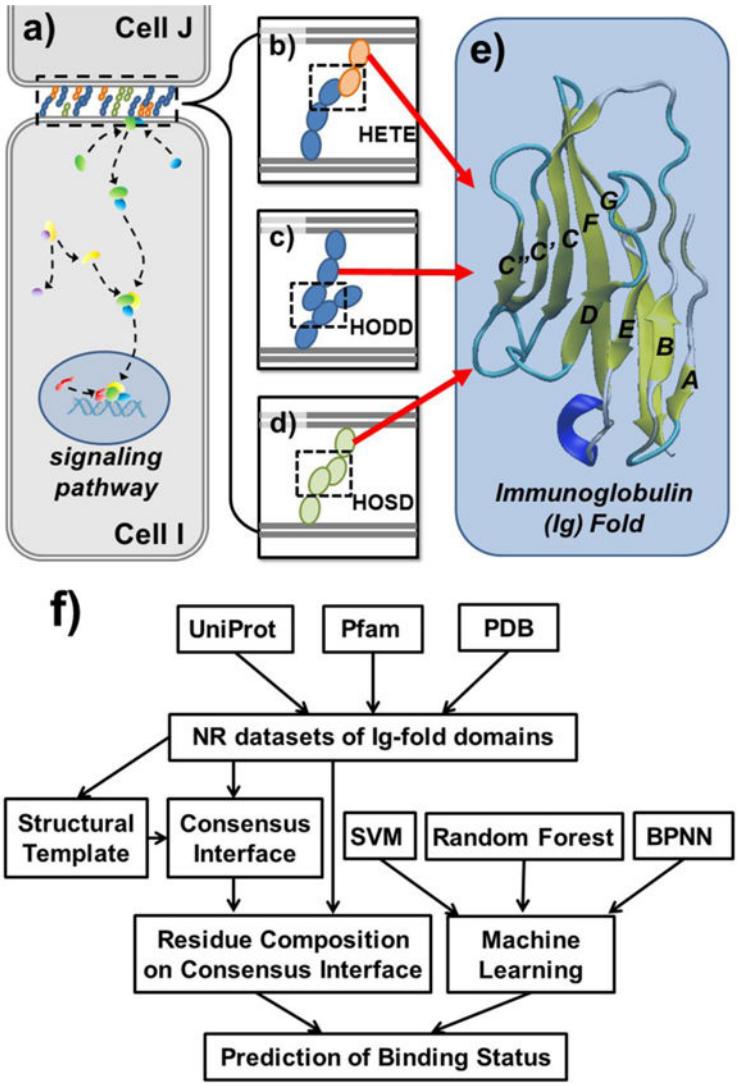

Figure 1.

(a) Cell surface receptors not only are the essential building block of intercellular adhesion, but also initiate the intracellular signaling pathway. The intermolecular interactions between receptors are conducted through different modes, including **(b)** the heterotypic binding between domains from different receptors (HETE), **(c)** the binding between different domains of two homotypic receptors (HODD), and **(d)** the binding between the same domains of two homotypic receptors (HOSD). Domains that belong to immunoglobulin fold **(e)**, which share structural features of a β-sandwich framework with hypervariable loops, is the largest group of domain families in these receptors. In this study, a computational framework **(f)** was constructed to characterize the structural similarity and functional diversity of domains in the entire Ig fold.

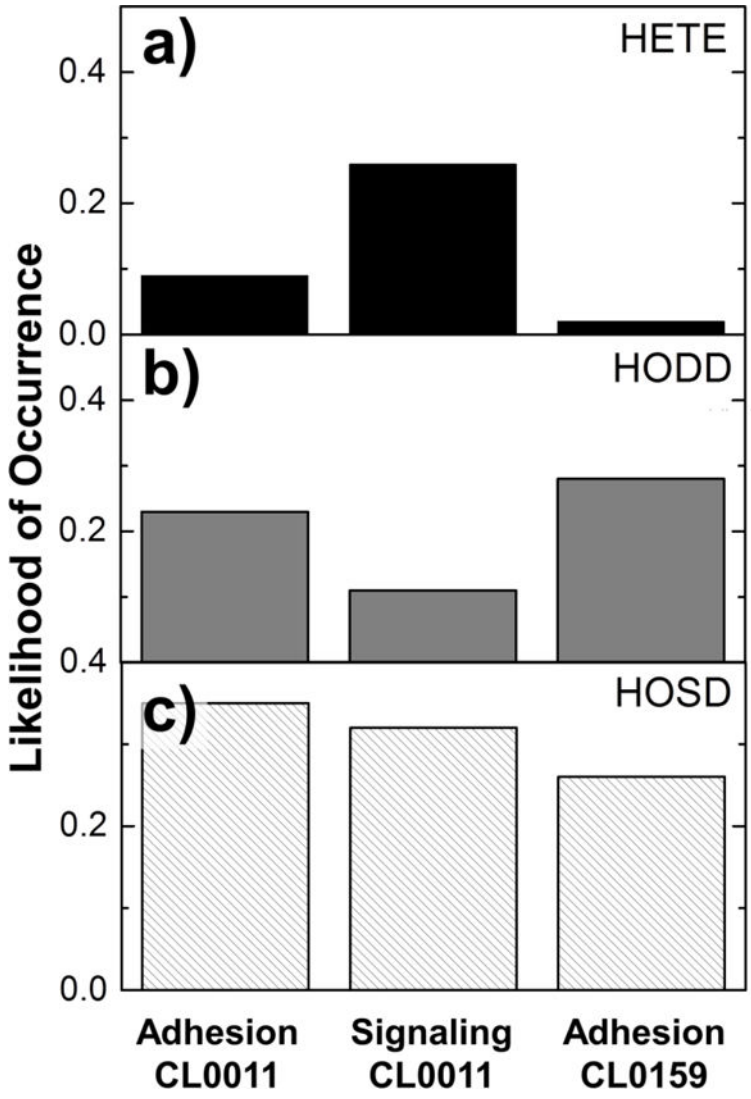

Figure 2.

An overall statistical analysis was carried out on the likelihood of occurrence for all three types of binding modes in three different datasets. The likelihood of occurrence was defined in ***Result 3.1***. The HETE mode is plotted in the black histogram **(a)**. The HETE mode is plotted in the grey histogram **(b)**. The HETE mode is plotted in the striped histogram **(c)**. The index of datasets is shown in the bottom of the figure. Different preferences of binding modes were observed for different datasets.

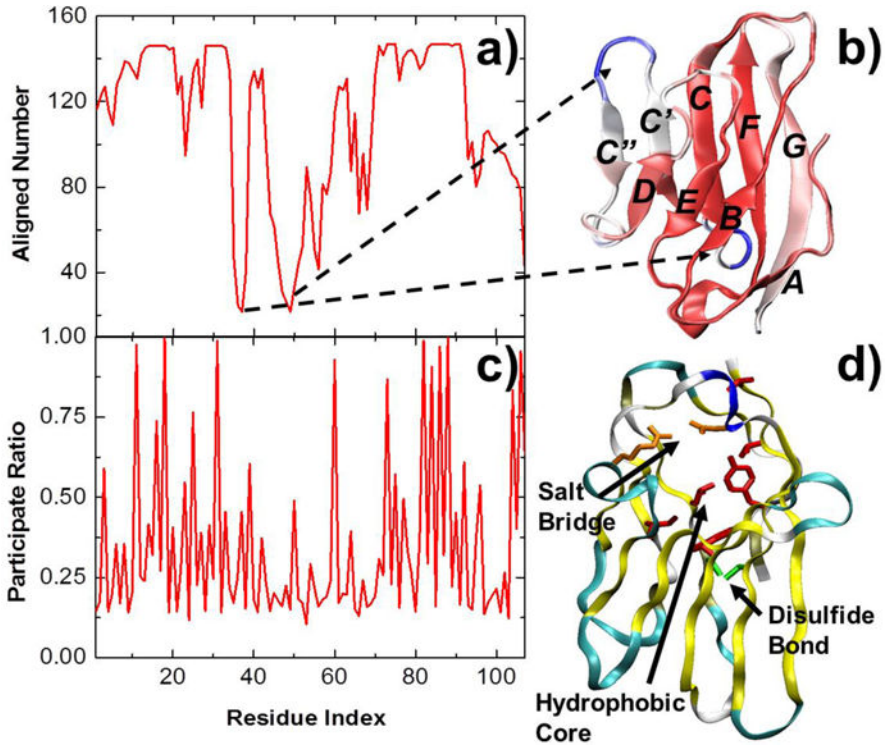

Figure 3.

A profile which describes how many alignments were found for each residue in the template is plotted in **(a)** for the dataset of Ig superfamily domains in adhesion receptors. The profile is projected to the structure of the template in **(b)** with the color index, in which blue indicates the regions with high structural variations. A profile of participation ratio (PR) which quantifies the conservation of residue type at each position is plotted in **(c)** for the same template. The sidechains of residues with the highest PR values are highlighted with the atomic representation in **(d)**.

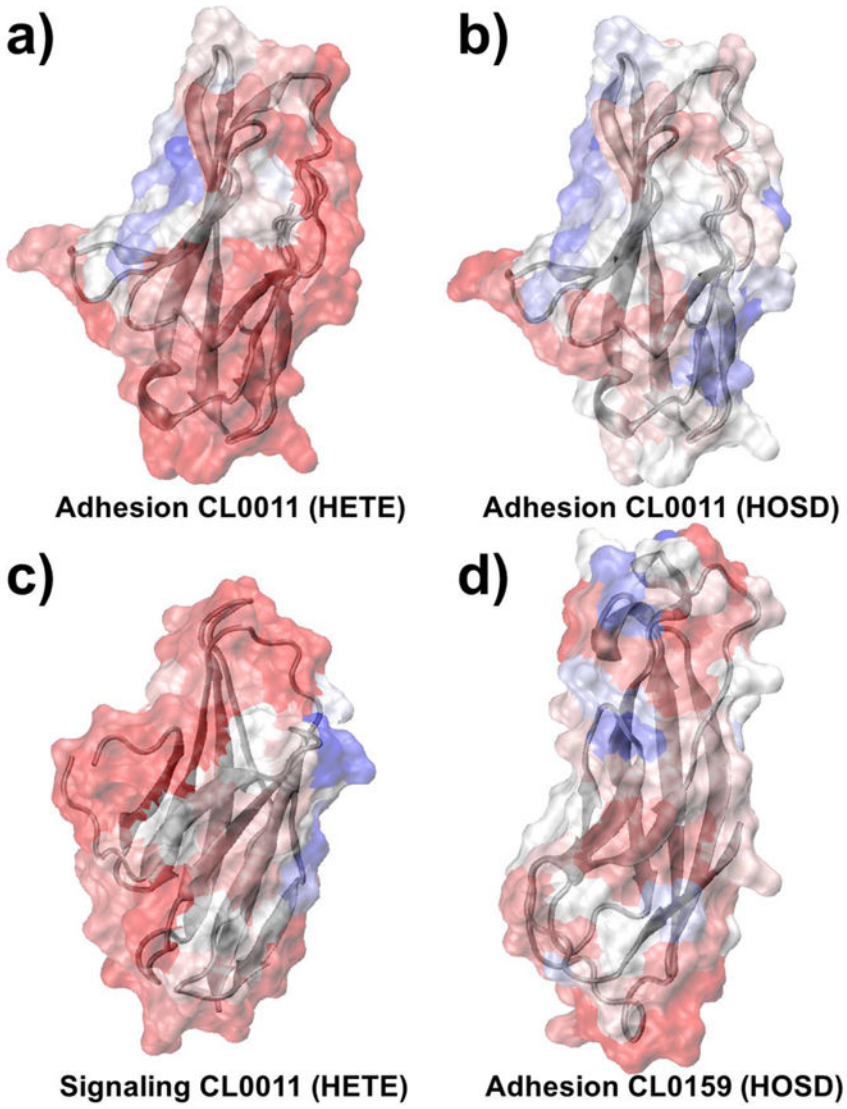

Figure 4.

The frequency of each residue as part of binding interfaces was projected onto the structure of each template. The regions of high frequencies are shown in blue, indicating the potential binding interfaces. The HETE and HOSD binding interfaces of Ig superfamily domains in adhesion receptors are plotted in **(a)** and **(b)**, respectively. The HETE binding interfaces of Ig superfamily domains in signaling receptors are plotted in **(c)**, while the HOSD binding interfaces of Ig-like fold superfamily domains in adhesion receptors are plotted in **(d)**.

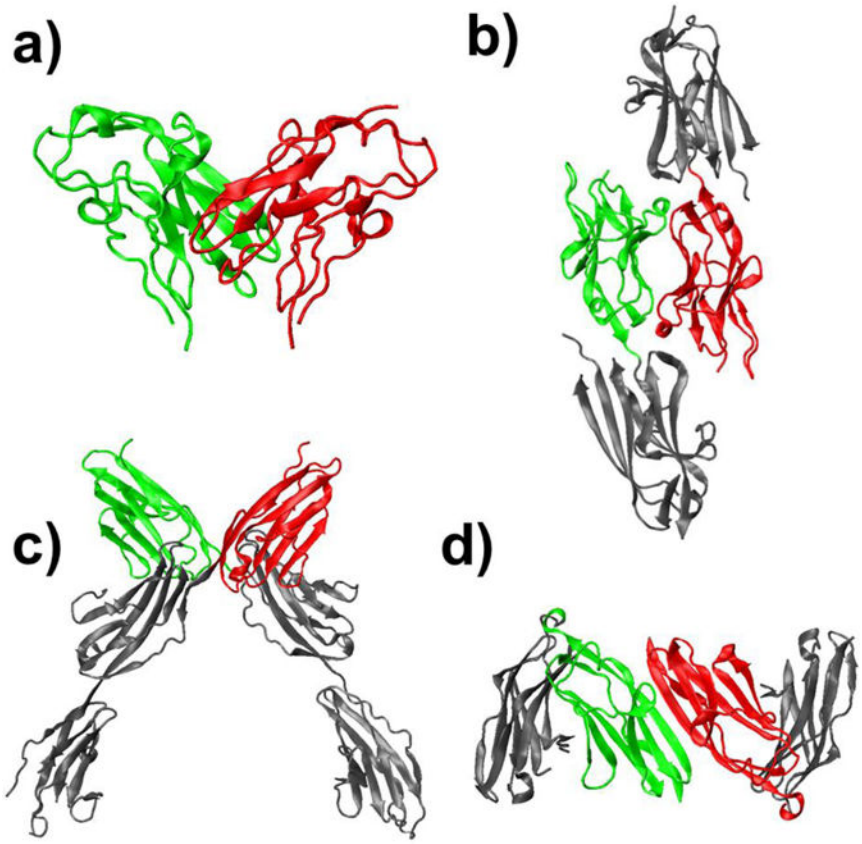

Figure 5.

Some specific cases from the datasets are shown, in which homodimers of Nectin-2 **(a)** and SYG-1 **(b)** are two examples that were correctly recognized by our machine learning methods. The domains in the dataset and their binding partners are colored in red and green, while the rest parts of complexes are colored in grey. In contrast, NCAM2 **(c)** and Protein Tyrosine Phosphatase δ **(d)** are two examples that were not correctly recognized during the cross-validation.

Table 1

The cross-validations were carried out for all three datasets to test if machine learning algorithms can recognize the signal of domain-domain interactions through the HOSD binding mode. The overall performance shows that the accuracies of different algorithms are ranged between 0.6 and 0.8. We further found the weighted voting strategy can improve the sensitivities, while maintaining the overall specificities and accuracies. The values of optimized weights for SVM, RF and BPNN are 0.12, 0.8 and 0.9, respectively.

Algorithm	Sensitivity	Specificity	Precision	Accuracy	LR+
Adhesion (CL0011)					
SVM	0.28	0.87	0.52	0.69	2.15
Random Forest	0.28	0.93	0.67	0.7	4.0
BPNN	0.45	0.79	0.54	0.67	2.14
Weighted Voting	0.51	0.77	0.57	0.69	2.2
Signaling (CL0011)					
SVM	0.21	0.86	0.42	0.66	1.5
Random Forest	0.26	0.91	0.57	0.7	2.89
BPNN	0.28	0.78	0.36	0.63	1.27
Weighted Voting	0.36	0.75	0.4	0.62	1.44
Adhesion (CL0159)					
SVM	0.01	0.93	0.25	0.71	0.14
Random Forest	0.17	0.93	0.46	0.73	2.43
BPNN	0.28	0.91	0.5	0.74	3.11
Weighted Voting	0.31	0.88	0.47	0.73	2.59

The optimized weights in the voting are 0.12, 0.8 and 0.9 for SVM, Random Forest and BPNN.