Published in final edited form as:

Proteins. 2020 January ; 88(1): 135–142. doi:10.1002/prot.25778.

Discovery of Receptor-Ligand Interfaces in the Immunoglobulin Superfamily

Nelson Gil, Eduardo J. Fajardo, Andras Fiser*

Department of Systems and Computational Biology, Department of Biochemistry, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, NY 10461, USA

Abstract

Cell-surface-anchored immunoglobulin superfamily (IgSF) proteins are widespread throughout the human proteome, forming crucial components of diverse biological processes including immunity, cell-cell adhesion, and carcinogenesis. IgSF proteins generally function through protein-protein interactions carried out between extracellular, membrane-bound proteins on adjacent cells, known as *trans*-binding interfaces. These protein-protein interactions constitute a class of pharmaceutical targets important in the treatment of autoimmune diseases, chronic infections, and cancer. A molecular-level understanding of IgSF protein-protein interactions would greatly benefit further drug development. A critical step towards this goal is the reliable identification of IgSF *trans*-binding interfaces.

We propose a novel combination of structure and sequence information to identify *trans*-binding interfaces in IgSF proteins. We developed a structure-based binding interface prediction approach that can identify broad regions of the protein surface that encompass the binding interfaces and suggests that IgSF proteins possess binding supersites. These interfaces could theoretically be pinpointed using sequence-based conservation analysis, with performance approaching the theoretical upper limit of binding interface prediction accuracy, but achieving this in practice is limited by the current ability to identify an appropriate multiple sequence alignment for conservation analysis. However, an important contribution of combining the two orthogonal methods is that agreement between these approaches can estimate the reliability of the predictions. This approach was benchmarked on the set of 22 IgSF proteins with experimentally solved structures in complex with their ligands. Additionally, we provide structure-based predictions and reliability scores for the 62 IgSF proteins with known structure but yet uncharacterized binding interfaces.

Keywords

Immunoglobulin Superfamily; Structure-based Mapping of Binding Site; Binding Site Conservation

*Corresponding author andras.fiser@einstein.yu.edu, Phone: 1-718-678-1068, Fax: 1-718-678-1019.

Conflict of Interest

The Authors declare no conflict of interests.

Introduction

The immunoglobulin superfamily (IgSF), one of the largest known domain classifications in the human proteome (Lander et al., 2001), includes proteins which mediate diverse biological processes, including immunity, cell-cell adhesion, and development (Aricescu & Jones, 2007; Barclay, 2003; Williams & Barclay, 1988; Zinn & Ozkan, 2017). Of special biomedical interest is the subset of IgSF proteins that are extracellular, membrane-bound, and carry out their functions through protein-protein interactions at defined, cell-cell (*trans*) binding interfaces (extracellular IgSFs) (Barclay, 2003). Prominent examples of *trans* interactions between extracellular IgSFs include the CTLA-4:B7-1/B7-2 and PD-1:PD-L1/PD-L2 systems, which downregulate the immune response following pathogen-induced activation in normal physiology in order to protect against autoimmune disease (Dai, Jia, Zhang, Fang, & Huang, 2014; Karandikar, Vanderlugt, Walunas, Miller, & Bluestone, 1996). Yet, these interactions have also been linked to cancer progression (Saresella, Rainone, Al-Daghri, Clerici, & Trabattini, 2012; Schwartz, Zhang, Nathenson, & Almo, 2002; Tanvetyanon, Gray, & Antonia, 2017). Protein interaction systems such as these constitute important pharmaceutical targets, as evidenced by the recent development of monoclonal antibodies for cancer treatment, such as ipilimumab (Lipson & Drake, 2011) and nivolumab (Brahmer, Hammers, & Lipson, 2015), which respectively target CTLA-4 and PD-1. An additional important drug class includes soluble receptor versions of IgSF proteins (e.g., abatacept (Vincenti & Luggen, 2007), a soluble CTLA-4) and their affinity-enhanced mutants (e.g. belatacept (Vincenti & Luggen, 2007)).

The rational development of therapeutics for modulating IgSF functions would benefit immensely from molecular-level insight into the relevant protein-protein interactions (Sliwoski, Kothiwale, Meiler, & Lowe, 2014). An important initial step for this task is the identification of protein-binding interfaces on individual IgSF proteins. However, progress in pharmaceutical development is limited by the paucity of available structural information. Out of the nearly 500 known cell-surface and secreted (extracellular) IgSF proteins in humans, only approximately 80 have had their crystallographic structures determined and just 22 of these have had their structures determined in complex with a molecular partner (Yap & Fiser, 2016). This class of proteins is difficult to access experimentally, as these are secreted or membrane-bound, disulfide-bond-containing structures that typically require a eukaryotic expression system and refolding studies from inclusion bodies. In particular, in order to determine a receptor-ligand complex structure, one would need to know the cognate binding partners in advance. This information is typically unknown and since there are approximately 100,000 possible combinations of interacting IgSF protein pairs, a brute-force experimental exploration would be impractical. Computational approaches to binding site determination are therefore an important step towards gaining ultimate insight into binding specificities and accelerating drug development.

Computational protein binding interface identification approaches broadly utilize some combination of sequence and structure features (Esmailbeiki, Krawczyk, Knapp, Nebel, & Deane, 2016). Sequence-based approaches most often rely on analyzing the conservation patterns of aligned sequence homolog positions to infer functionally important residues in a query protein, typically in a machine learning setting (e.g. (Ofra & Rost, 2007)). On the

other hand, the most successful structure-based approaches are template-based: they attempt to transfer known binding interfaces of structurally related proteins onto a query (e.g. (Q. C. Zhang et al., 2011)). Template-based structural approaches are more accurate at protein binding interface identification than sequence-based methods (Esmailbeiki et al., 2016), but this comes at the cost of requiring a clearly related functionally annotated structural template for a given query. Approaches that attempt to directly combine structure and sequence information (e.g. (Lichtarge, Bourne, & Cohen, 1996)) are comparatively rare (Esmailbeiki et al., 2016). On the other hand, many machine learning approaches have been developed that combine sequence and structural features to arrive at binding interface predictions (e.g. (Zellner et al., 2012)). Recent benchmarks suggest that the field of feature-based binding interface prediction appears to have saturated, as the addition of new properties results in little improvement in performance (Esmailbeiki et al., 2016; Gil & Fiser, 2018a), and argue that future improvements may be expected from customized predictors that focus on specific classes of proteins (Esmailbeiki et al., 2016).

In the current work, we combine structural and sequence information to predict the binding interfaces of IgSF proteins (Suppl. Fig. 1), taking advantage of their structural homogeneity in the context of significant sequence divergence (Barclay, 2003; Chattopadhyay et al., 2009). We present a novel structure-based interface mapping and filtering approach that we used in combination with our recently-developed multiple sequence alignment (MSA) selection pipeline, Selection of Alignment by Maximal Mutual Information (SAMMI) (Gil & Fiser, 2018b). We tested the algorithms on the set of 22 IgSF proteins with currently-known *trans*-binding interfaces (Table 1), and show that sequence-based predictions can be used to improve upon structure predictions towards the theoretical limit of binding site prediction (Gil & Fiser, 2018a). The current limiting factor in achieving this is identifying the optimal set of homologs to include in multiple sequence alignments used in conservation analysis. We also show that sequence information can provide a reliability measure for the structure-based approach – more reliable structure-based predictions will have pronounced sequence conservation signals. Finally, we provide *trans*-binding interface predictions along with reliability scores for 62 IgSF proteins of known structure for which the *trans*-binding interface is unknown.

Materials and Methods

IgSF protein dataset.

Based on previous studies (Rubinstein, Ramagopal, Nathenson, Almo, & Fiser, 2013; Yap & Fiser, 2016; Yap, Rosche, Almo, & Fiser, 2014), we compiled a set of 22 IgSF proteins forming part of known *trans*-binding heterophilic and homophilic complexes in the Protein Data Bank (Table 1) on which to benchmark the structure-based and sequence-based binding site prediction algorithms. The residues comprising the binding interfaces of the 22 IgSF proteins were obtained by applying the Contacts of Structural Units (CSU) (Sobolev, Sorokine, Prilusky, Abola, & Edelman, 1999) program with a cutoff distance of 5 Å. An additional set of 62 IgSF proteins with unknown binding interfaces (Suppl. Table 2) was obtained from the PDB as described in (Yap & Fiser, 2016). Only the N-terminal

immunoglobulin domains were considered in all analyses, as these are the ones generally involved in *trans*-binding interactions (Barclay, 2003).

Structure-based binding interface prediction.

The algorithm consists of two main steps (Suppl. Fig. 2):

1. For a given query protein, members of a set of structurally similar proteins (Q. C. Zhang et al., 2011) with known binding interfaces were structurally aligned in a pairwise manner using the “align3d” function from MODELLER 9.14 (Webb & Sali, 2016), with default parameters. The binding interface residues of each of the structural neighbors were individually mapped onto the query protein, and the union of all the mapped residues was taken as a “raw” interface prediction.
2. The raw prediction was then filtered with the aim of morphing it into a continuous patch (Suppl. Fig. 3). Every query protein atom with absolute solvent-accessible surface area (SASA) greater than $5 \text{ \AA}^2$, calculated by NACCESS (Hubbard & Thornton, 1993) using default probe size, was evaluated for the surrounding presence of atoms belonging to “raw prediction” residues that were mapped by step 1. A sphere of radius d was centered at the atom being evaluated, and the number of atomic contacts within the sphere belonging to interface-predicted (c_I) and non-interface-predicted atoms (c_N) were counted. If the total number of contacts $c_I + c_N$ exceeded a minimum number c_{min} , a ratio of interface-predicted to total contacts was calculated as $r = c_I / (c_I + c_N)$. If r exceeded a threshold value r_{min} and the atom being evaluated was not already part of the raw prediction, then it was re-classified as interface-predicted. Conversely, an atom that was part of the raw prediction was removed if r did not exceed r_{min} . The filtering process was iterated over the entire protein i times. The final interface residue prediction for the query consisted of residues with at least one atom classified as interface-predicted. The four algorithm parameters introduced by this step (d, r_{min}, c_{min}, i) were optimized (see “Benchmarking of algorithms”).

Sequence-based binding interface prediction.

Sequence-based interface predictions were performed as follows, adapted from previous studies (Gil & Fiser, 2018a, 2018b):

1. A given query protein’s amino acid sequence was searched through the NCBI “nr” database (NCBI_Resource_Coordinators, 2017) using jackhmmer 3.1 (Johnson, Eddy, & Portugaly, 2010) with a domain-based e-value cutoff of 10^{-20} and otherwise default parameters, generating a sequence profile typically including several thousand hits.
2. The jackhmmer profile was then subset into 264 alternative multiple sequence alignments (MSAs) by combinatorially applying three sequence identity filters: the minimum (set at 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, and 60%) and maximum (set at 50%, 70%, 90%, 99%) sequence identity between query and hits, and the maximum sequence identity (clustering level) among hits (set at

40%, 50%, 60%, 70%, 80%, 90%, 95%, and 99%). The total number of combinations of all parameters are 288 but the minimum and maximum sequence identities of hits to the query have an overlap in the middle range, which reduces the possible number of combinations to 264. The length of each hit also needed to cover at least 70% of the query.

3. Out of the 264 sampled MSAs, the one that scored highest using the mutual-information-based score of SAMMI (Gil & Fiser, 2018b) is hypothesized to best encode functional signals that uniquely specify a protein family. Each solvent-accessible (NACCESS-computed SASA > 5 Å²) query residue position in the SAMMI-selected MSA was scored using an algorithm based on the Jensen-Shannon Divergence (JSD), with default parameters (Capra & Singh, 2007), and the top N conserved residues were used, where N is the number of residues obtained in the structure-based prediction.

Benchmarking of standalone structure-based and sequence-based algorithms.

The F-Score (Witten, Frank, & Hall, 2011) was used to evaluate binding interface predictions. The F-Score is defined as $2 * \text{precision} * \text{recall} / (\text{precision} + \text{recall})$, where precision is the ratio of true positives to the sum of true and false positives and recall is the ratio of true positives to sum of true positives and false negatives.

The structure-based binding interface prediction algorithm was benchmarked on the set of 22 IgSF proteins with known *trans*-binding interfaces using a leave-one-out cross-validation approach. Given one of the IgSF structures as the query, the other 21 had their binding interfaces mapped onto the query and subsequently filtered to generate a continuous patch as the interface prediction. For each individual protein, the four parameters of the filtering step were optimized by brute-force exploration (d : 4 Å to 8 Å at 0.5 Å intervals, r_{min} : 0 to 1 at 0.1 intervals, c_{min} : integers from 4 to 10, i : integers from 1 to 10) to maximize the F-Score of the interface prediction. An alternative optimization of the four parameters was also performed in which F-Scores were maximized under the constraint of missing no interface residues in the prediction (i.e. having zero false negatives). Given a query protein, these zero-false-negative optimized parameters were averaged over the remaining 21 IgSF proteins in the dataset and were termed the Leave-One-Out-Averaged (LOOA) parameters for the query. Each query's LOOA parameters are expected to reflect the algorithm performance that could be expected in practice, as they remove the bias introduced by optimizing any individual query's F-Score.

The sequence-based algorithm was first benchmarked on the 22 IgSF proteins in a parallel fashion to the structure-based algorithm. The conservation-based interface prediction of the SAMMI-selected MSA from the 264 sampled for each protein was compared to the MSA whose conservation-based interface prediction yielded the highest F-Score (the "Max Sampled MSA").

Benchmarking of the combined sequence-structure algorithm.

The SAMMI-selected and Max Sampled MSAs were used to examine the sequence conservation of the residues included in the structure-based predictions. The top 15 residues

by JSD conservation score of the structure-based prediction were benchmarked as binding interface predictions.

Reliability scores for the structure-based predictions were calculated based on their agreement with sequence-based predictions originating from conservation analysis of (1) the SAMMI-selected MSA, (2) the Max Sampled MSA, and (3) the MSA that agreed the most with the structure-based prediction (the “Max Agreement MSA”). Reliability score is the ratio of predicted interface residues between the sequence and structure based approach, and as such ranges between 0 (complete disagreement) and 1 (full agreement). The relationship between reliability scores and structure-based prediction accuracy was benchmarked on the 22 IgSF proteins with known binding interfaces.

Prediction of unknown *trans*-binding interfaces.

For each of the 62 IgSF proteins with unknown *trans*-binding interfaces, structure-based interface prediction was applied by structurally aligning the 22 IgSF proteins with known *trans*-binding interfaces onto them. The known binding interfaces were mapped and subsequently filtered using the average of the LOOA parameters ($d = 7.25$, $r_{min} = 0.79$, $c_{min} = 8$, $i = 3$). Sequence-based interface prediction was applied in parallel, taking the top N conserved residues in the SAMMI-selected and Max Agreement MSAs, where N was the number of residues predicted by the structure-based approach. The fractional agreement between the residues predicted by the structure and the SAMMI / Max Agreement sequence approaches was defined as a reliability score for the structure-based prediction.

Results & Discussion

Structure-based identification of IgSF binding interfaces

First, a structure-based approach was benchmarked to identify *trans*-binding interfaces on the dataset of 22 IgSF proteins with crystallized complex structures (Table 1). The conceptual framework of this approach has been used successfully in previous studies (Hwang, Petrey, & Honig, 2016; Q. C. Zhang et al., 2011). The fundamental idea of this approach is to structurally align homologous proteins with known binding interfaces onto a query protein with an unknown binding interface (Suppl. Fig. 2), with the hypothesis that their shared evolutionary history preserved the location of the binding site. For this study, given one of the 22 IgSF proteins as the query, the *trans*-binding interfaces of the other 21 IgSF proteins in the dataset were mapped onto it. The set of all mapped query protein residues comprises a “raw” prediction that is then filtered to create a continuous patch (Suppl. Fig. 3). The results of the filtering approach are specified by a set of four parameters that, for the purposes of benchmarking, were optimized for each protein individually to maximize the binding interface prediction accuracy (measured by the F-Score (Witten et al., 2011)), both directly and under the constraint of missing no true interface residues in the predictions. To estimate the algorithm performance that could be expected in practice for a given query protein, the parameters resulting in the optimal zero-false-negative predictions for the remaining 21 IgSF proteins were averaged; these averages were termed the Leave-One-Out-Averaged (LOOA) parameters and applied to structure-based binding interface prediction of the query. The zero-false-negative parameters (Suppl. Table 2) were used as the

basis for the LOOA parameters to provide a broader set of starting residues for input to subsequent sequence-based analysis. The filtering step greatly reduces the residues matched by the raw structure interface mapping: while the mean number of raw-mapped residues was 65.4, filtering using the LOOA parameters reduces this number by 60%, to a mean of 25.8.

The structure-based algorithm is generally adept at predicting binding interfaces in the 22 IgSF dataset, with a median F-Score of 0.57 when using LOOA parameters (Fig. 1). This is in line with performance reported by recent structural-template-based protein-binding interface (Hwang et al., 2016) and small-molecule-binding site (C. Zhang, Freddolino, & Zhang, 2017) prediction methods. The number of predicted residues had a median of 25; given the median dataset interface size of 14 residues, this indicates a systematic tendency for the inclusion of false positive residues in the predictions. For comparison, when algorithm parameters were optimized under the zero-false-negative constraint, the median F-Score was 0.63; as expected, enforcing this constraint led to an increased median prediction size of 33 residues. On the other hand, when the structure-based algorithm parameters were optimized for each protein individually, the average F-Score was 0.71, and the median number of predicted residues decreased to 18. The median F-Score of 0.71 approaches ~0.80, the theoretical upper limit of binding site prediction F-Scores that generally reflects the agreement between binding site definitions among different databases (Gil & Fiser, 2018a). The structure-based algorithm's high performance on this dataset suggests that IgSF proteins possess binding interfaces conserved in geometric location despite relatively distant sequence homology, consistent with the presence of binding supersites (Russell, Sasieni, & Sternberg, 1998).

Combining structural interface mapping with sequence-based conservation analysis

Next, we benchmarked a sequence-based approach to identify protein-binding interfaces adapted from our previously published SAMMI method (Gil & Fiser, 2018b). In brief, for each query we performed a jackhmmer (Johnson et al., 2010) search on the “nr” database (NCBI_Resource_Coordinators, 2017), from which we subsampled 264 alternative multiple sequence alignments (MSAs) consisting of different subsets of search hits. SAMMI uses a mutual-information-based score to select an MSA (the “SAMMI MSA”) for subsequent conservation analysis. In conservation analysis, each query sequence position is scored, and the top N scoring positions are taken as the binding interface prediction, where N is the number of residues in the query's structure-based prediction – this was done to enable direct comparisons between approaches. In addition, we compared the performance of conservation analysis on SAMMI MSAs to the MSAs whose conservation analysis yielded the highest performance out of any of the 264 sampled (the “Max Sampled MSA”).

Sequence-based identification of binding sites was generally less successful than structure-based predictions, as expected (Suppl. Fig. 4), with median F-Scores of 0.28 and 0.45 for conservation analysis of SAMMI and Max Sampled MSAs, respectively. This result is consistent with previous benchmarking of SAMMI on a general set of protein-protein interaction interfaces (Gil & Fiser, 2018b) as well as the knowledge that protein-binding sites are generally less conserved than small-molecule-binding sites (Capra & Singh, 2007).

Particularly, the relatively low F-Scores of SAMMI MSAs could be due to competing conservation signals among IgSF binding sites other than the *trans*-binding sites.

We hypothesized that sequence conservation could pinpoint the binding residues within the broad area of the IgSF proteins identified by the structure-based method. SAMMI and Max Sampled MSAs were used to perform localized conservation analyses of the binding interface prediction patches identified by the structure-based method using the LOOA and zero-false-negative optimized parameters (Fig. 2). These predictions consisted of the 15 most conserved residues in the structure-mapped patch. The conservation analysis of SAMMI MSAs still does not appear to benefit the structure-based prediction under either set of parameters. The combination approach shows median F-Scores of 0.55 and 0.61 for the LOOA and zero-false-negative parameters, respectively, performances statistically indistinguishable from the respective median F-Scores of 0.57 ($p = 0.66$) and 0.64 ($p = 0.90$) of the structure-based method alone. On the other hand, using the Max Sampled MSAs shows a modest improvement over the structure-only predictions, with median F-Scores of 0.67 ($p = 0.03$) and 0.72 ($p = 0.03$) for the LOOA and zero-false-negative predictions, respectively. These results indicate that structure-based methods can still improve towards the theoretical upper performance limit if sequence information is utilized, and that the current limiting factor in achieving this is finding an algorithm that can consistently identify the Max Sampled MSAs.

Sequence-based predictions as a reliability measurement for structure-based predictions

Our results suggest that sequence-based conservation analysis cannot, in general, improve structure-based predictions of IgSF *trans*-binding interfaces. Similarity based prediction algorithms, such as the structure-based algorithm used here, have been proven to always be more accurate in predicting binding sites, but with the caveat that their applicability is limited to those families where annotated structural information is available. In addition, structure-based approaches can fail spectacularly if the binding site is different than what is observed in a remotely related protein. Sequence based functional site predictions do not have the same limitation, but they are overall less accurate. We hypothesized that sequence conservation would serve as an orthogonal source of information that could provide a reliability measure for the structure-based predictions. We thus examined the correlation between the performance of the structure-based predictions and the degree to which these agreed with sequence-based predictions stemming from the conservation analysis of three different classes of MSAs (Fig. 3). The SAMMI MSAs show a correlation coefficient of 0.598 (Fig. 3A), indicating that the sequence-based and structure-based predictions indeed tend to agree when the structure-based prediction is accurate. This tendency is stronger when using Max Sampled MSAs, which show a correlation coefficient of 0.70 (Fig. 3B).

We also examined the correlation between the structure-based algorithm performance and the conservation analysis prediction of the “Max Agreement MSA” – the one out of the 264 sampled MSAs for which maximum sequence-structure prediction agreement was observed (Fig. 3C). Interestingly, the correlation coefficient for Max Agreement MSAs is 0.693, which closely matches the correlation observed for the Max Sampled MSAs. This result indicates that a structure-based prediction is reliable if it is corroborated by a conservation

signal present in the alignment of some subset of available sequence space. In other words, a reliable structure-mapped patch is one onto which sequence-based conservation could be fitted. This is suggestive of a novel general sequence-structure combination algorithm to identify binding interfaces: structural template-based binding interface predictions could be assigned confidence values based on their agreement with sequence-based conservation analysis. An extension of the present work would be to apply this algorithm to a carefully curated dataset representing other sets of proteins resembling the IgSF in that they are structurally homogeneous yet sequentially divergent.

Benchmarking with alternative interface prediction approaches

A guiding principle in our approach is that superfamily-specific structural features can improve the accuracy of interface prediction. To further elaborate this point, we employed on our dataset four recently published interface prediction methods. These are not superfamily specific approaches, but provide general purpose tools for predicting protein binding interfaces. All these approaches use a fairly large number of different features in a machine learning setting to provide prediction. SPRINT-str is employing Random Forest classification of interface residues using information on accessible surface area, surface exposure, secondary structure and flexibility (Taherzadeh, Zhou, Liew, & Yang, 2018). ISPRED4 combines evolutionary information, residue conservation, interface propensity, physico-chemical properties, residue co-evolutionary signals, geometrical descriptors such as portusion, secondary structure, residue depth and B-factors (Savojardo, Fariselli, Martelli, & Casadio, 2017). INTPRED approach segments surfaces into overlapping patches and calculates the likelihood of being an interface using features of hydrophobicity, propensity, conservation, disulphide bonds, hydrogen bonds, various features of secondary structures, and planarity in a Random Forest training (Northey, Baresic, & Martin, 2017). SPPIDER is a machine learning based classifier with features including secondary structure, hydrophobicity, conservation of charge, size and residue, hydrophobicity. The performance of these four approaches on the IgSF dataset range between average F-score of 14.84% to 32.24% (Fig. 4.), significantly below what has been achieved with our approach at 57% (Fig. 2.).

Binding interface predictions for all IgSFs of known structure

Given the results of our benchmarking study, we utilized the structure-based algorithm to provide predictions of *trans*-binding interfaces for 62 IgSF proteins with known structures, but unknown binding interfaces (Suppl. Table 2). For each of these 62 proteins, we structurally mapped the binding interfaces of the 22 IgSF dataset onto them and subsequently filtered the raw mappings applying the average of the LOOA parameters. In addition, we calculated our newly established reliability scores for the predictions based on the SAMMI and Max Agreement MSAs. The average agreement with SAMMI MSAs is low, at 0.085, which suggests that the SAMMI MSAs are being influenced by the conservation signals of binding sites outside the structure-mapped patches. On the other hand, the reliability scores based on the Max Agreement MSAs average 0.284, which based on the correlation in Fig. 3C translates to an average predicted structure-based F-Score of approximately 0.40. More specifically, out of the set of 62 proteins, 26 had reliability scores of 0.30 or above, indicating that nearly half of the dataset can be expected to have accurate

binding site predictions. These binding interface predictions thus provide an important starting point for further computational and experimental study of these proteins and rational drug development.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by National Institutes of Health (NIH) grant R01 GM118709, and the Extreme Science and Engineering Discovery Environment (XSEDE) project (NSF grant ACI-1053575). NG was supported by the National Research Service Award (NRSA) individual fellowship F31GM116570 and the Medical Scientist Training Program (MSTP) grant T32GM007288.

References

- Aricescu AR, & Jones EY (2007). Immunoglobulin superfamily cell adhesion molecules: zippers and signals. *Curr Opin Cell Biol*, 19(5), 543–550. doi:10.1016/j.ceb.2007.09.010 [PubMed: 17935964]
- Barclay AN (2003). Membrane proteins with immunoglobulin-like domains--a master superfamily of interaction molecules. *Semin Immunol*, 15(4), 215–223. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/14690046> [PubMed: 14690046]
- Brahmer JR, Hammers H, & Lipson EJ (2015). Nivolumab: targeting PD-1 to bolster antitumor immunity. *Future Oncol*, 11(9), 1307–1326. doi:10.2217/fon.15.52 [PubMed: 25798726]
- Capra JA, & Singh M (2007). Predicting functionally important residues from sequence conservation. *Bioinformatics*, 23(15), 1875–1882. doi:10.1093/bioinformatics/btm270 [PubMed: 17519246]
- Chattopadhyay K, Lazar-Molnar E, Yan Q, Rubinstein R, Zhan C, Vigdorovich V, ... Almo SC (2009). Sequence, structure, function, immunity: structural genomics of costimulation. *Immunol Rev*, 229(1), 356–386. doi:10.1111/j.1600-065X.2009.00778.x [PubMed: 19426233]
- Dai S, Jia R, Zhang X, Fang Q, & Huang L (2014). The PD-1/PD-Ls pathway and autoimmune diseases. *Cell Immunol*, 290(1), 72–79. doi:10.1016/j.cellimm.2014.05.006 [PubMed: 24908630]
- Esmailbeiki R, Krawczyk K, Knapp B, Nebel JC, & Deane CM (2016). Progress and challenges in predicting protein interfaces. *Brief Bioinform*, 17(1), 117–131. doi:10.1093/bib/bbv027 [PubMed: 25971595]
- Gil N, & Fiser A (2018a). The choice of sequence homologs included in multiple sequence alignments has a dramatic impact on evolutionary conservation analysis. *Bioinformatics* doi:10.1093/bioinformatics/bty523
- Gil N, & Fiser A (2018b). Identifying functionally informative evolutionary sequence profiles. *Bioinformatics*, 34(8), 1278–1286. doi:10.1093/bioinformatics/btx779 [PubMed: 29211823]
- Hubbard S, & Thornton J (1993). NACCESS
- Hwang H, Petrey D, & Honig B (2016). A hybrid method for protein-protein interface prediction. *Protein Sci*, 25(1), 159–165. doi:10.1002/pro.2744 [PubMed: 26178156]
- Johnson LS, Eddy SR, & Portugaly E (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*, 11, 431. doi:10.1186/1471-2105-11-431 [PubMed: 20718988]
- Karandikar NJ, Vanderlugt CL, Walunas TL, Miller SD, & Bluestone JA (1996). CTLA-4: a negative regulator of autoimmune disease. *J Exp Med*, 184(2), 783–788. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/8760834> [PubMed: 8760834]
- Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, ... International Human Genome Sequencing, C. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822), 860–921. doi:10.1038/35057062 [PubMed: 11237011]

- Lichtarge O, Bourne HR, & Cohen FE (1996). An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol*, 257(2), 342–358. doi:10.1006/jmbi.1996.0167 [PubMed: 8609628]
- Lipson EJ, & Drake CG (2011). Ipilimumab: an anti-CTLA-4 antibody for metastatic melanoma. *Clin Cancer Res*, 17(22), 6958–6962. doi:10.1158/1078-0432.CCR-11-1595 [PubMed: 21900389]
- NCBI_Resource_Coordinators. (2017). Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, 45(D1), D12–D17. doi:10.1093/nar/gkw1071 [PubMed: 27899561]
- Northey T, Baresic A, & Martin ACR (2017). IntPred: a structure-based predictor of protein-protein interaction sites. *Bioinformatics* doi:10.1093/bioinformatics/btx585
- Ofran Y, & Rost B (2007). ISIS: interaction sites identified from sequence. *Bioinformatics*, 23(2), e13–16. doi:10.1093/bioinformatics/btl303 [PubMed: 17237081]
- Rubinstein R, Ramagopal UA, Nathenson SG, Almo SC, & Fiser A (2013). Functional Classification of Immune Regulatory Proteins. *Structure*, 21(5), 766–776. doi:10.1016/j.str.2013.02.022 [PubMed: 23583034]
- Russell RB, Sasieni PD, & Sternberg MJ (1998). Supersites within superfolds. Binding site similarity in the absence of homology. *J Mol Biol*, 282(4), 903–918. doi:10.1006/jmbi.1998.2043 [PubMed: 9743635]
- Saresella M, Rainone V, Al-Daghri NM, Clerici M, & Trabattini D (2012). The PD-1/PD-L1 pathway in human pathology. *Curr Mol Med*, 12(3), 259–267. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/22300137> [PubMed: 22300137]
- Savojardo C, Fariselli P, Martelli PL, & Casadio R (2017). ISPRED4: interaction sites PREDiction in protein structures with a refining grammar model. *Bioinformatics*, 33(11), 1656–1663. doi:10.1093/bioinformatics/btx044 [PubMed: 28130235]
- Schwartz JC, Zhang X, Nathenson SG, & Almo SC (2002). Structural mechanisms of costimulation. *Nat Immunol*, 3(5), 427–434. doi:10.1038/ni0502-427 [pii] [PubMed: 11976720]
- Sliwoski G, Kothiwale S, Meiler J, & Lowe EW Jr. (2014). Computational methods in drug discovery. *Pharmacol Rev*, 66(1), 334–395. doi:10.1124/pr.112.007336 [PubMed: 24381236]
- Sobolev V, Sorokine A, Prilusky J, Abola EE, & Edelman M (1999). Automated analysis of interatomic contacts in proteins. *Bioinformatics*, 15(4), 327–332. doi:10.1093/bioinformatics/btc042 [pii] [PubMed: 10320401]
- Taherzadeh G, Zhou Y, Liew AW, & Yang Y (2018). Structure-based prediction of protein-peptide binding regions using Random Forest. *Bioinformatics*, 34(3), 477–484. doi:10.1093/bioinformatics/btx614 [PubMed: 29028926]
- Tanvetyanon T, Gray JE, & Antonia SJ (2017). PD-1 checkpoint blockade alone or combined PD-1 and CTLA-4 blockade as immunotherapy for lung cancer? *Expert Opin Biol Ther*, 17(3), 305–312. doi:10.1080/14712598.2017.1280454 [PubMed: 28064556]
- Vincenti F, & Luggen M (2007). T cell costimulation: a rational target in the therapeutic armamentarium for autoimmune diseases and transplantation. *Annu Rev Med*, 58, 347–358. doi:10.1146/annurev.med.58.080205.154004 [PubMed: 17020493]
- Webb B, & Sali A (2016). Comparative Protein Structure Modeling Using MODELLER. *Curr Protoc Protein Sci*, 86, 29.1–29.37. doi:10.1002/cpps.20
- Williams AF, & Barclay AN (1988). The immunoglobulin superfamily--domains for cell surface recognition. *Annu Rev Immunol*, 6, 381–405. doi:10.1146/annurev.iy.06.040188.002121 [PubMed: 3289571]
- Witten IH, Frank E, & Hall MA (2011). *Data Mining: Practical Machine Learning Tools and Techniques* (3rd ed.) Burlington, MA: Morgan Kaufmann.
- Yap EH, & Fiser A (2016). ProtLID, a Residue-Based Pharmacophore Approach to Identify Cognate Protein Ligands in the Immunoglobulin Superfamily. *Structure*, 24(12), 2217–2226. doi:10.1016/j.str.2016.10.012 [PubMed: 27889206]
- Yap EH, Rosche T, Almo S, & Fiser A (2014). Functional clustering of immunoglobulin superfamily proteins with protein-protein interaction information calibrated hidden Markov model sequence profiles. *J Mol Biol*, 426(4), 945–961. doi:10.1016/j.jmb.2013.11.009 [PubMed: 24246499]

- Zellner H, Staudigel M, Trenner T, Bittkowski M, Wolowski V, Icking C, & Merkl R (2012). PresCont: predicting protein-protein interfaces utilizing four residue properties. *Proteins*, 80(1), 154–168. doi:10.1002/prot.23172 [PubMed: 22038731]
- Zhang C, Freddolino PL, & Zhang Y (2017). COFACTOR: improved protein function prediction by combining structure, sequence and protein-protein interaction information. *Nucleic Acids Res*, 45(W1), W291–W299. doi:10.1093/nar/gkx366 [PubMed: 28472402]
- Zhang QC, Deng L, Fisher M, Guan J, Honig B, & Petrey D (2011). PredUs: a web server for predicting protein interfaces using structural neighbors. *Nucleic Acids Res*, 39(Web Server issue), W283–287. doi:10.1093/nar/gkr311 [PubMed: 21609948]
- Zinn K, & Ozkan E (2017). Neural immunoglobulin superfamily interaction networks. *Curr Opin Neurobiol*, 45, 99–105. doi:10.1016/j.conb.2017.05.010 [PubMed: 28558267]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1.

Performance of structure-based prediction F-Scores for 22 IgSF proteins under three sets of parameters. The numbers above the bars indicate the numbers of residues included in predictions. The Leave-One-Out-Averaged (LOOA) parameter predictions show a median F-Score of 0.57 (red), with most errors being due to false positives, as seen by a median interface size of 25. Filtering parameters that maximized F-Scores for each protein show a median F-Score of 0.71 (blue) and a median number of predicted residues of 18, close to the median true interface size of 14 residues for this dataset. Optimizing for maximum F-Score under the constraint of zero false negatives (0-FN) shows a median F-Score of 0.64 (yellow), but increases the median number of predicted residues to 33.

Figure 2.

Distributions of F-Scores over the 22 IgSF dataset for six prediction categories considering either structure alone or both structure and sequence information. Using the Leave-One-Out-Averaged (LOOA) parameters, the performance of the structure-based prediction alone (STR-LOOA; median F-Score 0.57); structure-based prediction combined with conservation analysis of the SAMMI MSA (STR-LOOA-SAMMI; median F-Score 0.55, $p = 0.659$); combining the structure-based prediction with conservation analysis of the Max Sampled MSA (STR-LOOA-MaxSampled; median F-Score 0.67, $p = 0.033$); using the zero-false-negative parameters (STR-0FN; median F-Score 0.64); combining the structure-based prediction with conservation analysis based on the SAMMI MSA (STR-0FN-SAMMI; median F-Score 0.61, $p = 0.900$); utilizing the Max Sampled MSA with structure-based prediction (STR-0FN-MaxSampled; median F-Score 0.72, $p = 0.034$). All p -values were computed with the one-tailed Wilcoxon rank sum test, hypothesizing that the sequence-

structure combination prediction would be superior to the structure-based prediction alone. For each prediction category, black bars represent the median, the boxes represent the interquartile range, and the whisker tips represent the 10th and 90th percentiles. The two circles in the STR-OFN-MaxSampled category are two data points below the 10th percentile for that category.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 3.

Correlation between structure-based F-Score using Leave-One-Out-Averaged parameters (STR-LOOA) and the fraction of agreeing residues between the STR-LOOA and predictions from sequence-based conservation analysis based on SAMMI (A), Max Sampled (B), and Max Agreement (C) MSAs. The Pearson correlation coefficients are 0.598 in (A), 0.700 in (B), and 0.693 in (C).

Figure 4. Performance of four protein interface prediction method, SPRINT-Str, ISPRED, IntPred and SPPIDER on the dataset of 22 IgSF proteins. Distributions of F-Scores are shown in boxplots. For each prediction method, black bars represent the median, the boxes represent the interquartile range, and the whisker tips represent the 10th and 90th percentiles.

PDB IDs, number of binding residues, and names for the 22 IgSF proteins and their binding partners on which the structure-based and sequence-based binding interface prediction algorithms were evaluated. Names are UniProt identifiers with the species suffix removed. All proteins are human orthologs except for CD48 (2PTT.A), CD244 (2PTT.B), and PDCD1 (3BIK.B) which are murine.

Table 1.

PDB ID	Number of Binding Residues	UniProt Name	Binding Partner UniProt Name
1F5W.A	13	CXAR	CXAR
1I85.B	14	CD86	CTLA4
1I85.D	9	CTLA4	CD86
1I8L.A	13	CD80	CTLA4
1QA9.A	14	CD2	LFA3
1QA9.B	16	LFA3	CD2
1Z9M.A	14	CADM3	CADM3
2IF7.A	14	SLAF6	SLAF6
2JJS.A	21	SHPS1	CD47
2JJS.C	15	CD47	SHPS1
2PKD.A	16	SLAF5	SLAF5
2PTT.A	14	CD48	CD244
2PTT.B	18	CD244	CD48
3BIK.A	14	PD1L1	PDCD1
3BIK.B	16	PDCD1	PD1L1
3PV6.A	10	NR3L1	NCTR3
3PV6.B	12	NCTR3	NR3L1
3RBG.A	14	CRTAM	CRTAM
3UDW.A	21	TIGIT	PVR
3UDW.C	17	PVR	TIGIT
4DFH.A	20	NECT2	NECT2
4FRW.B	16	NECT4	NECT4