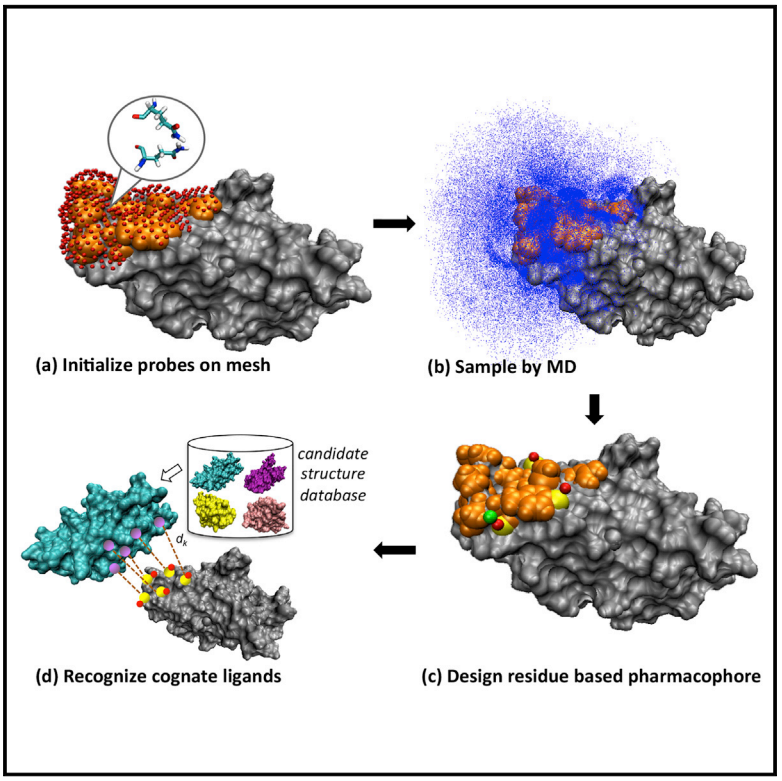
Resource

Structure

ProtLID, a Residue-Based Pharmacophore Approach to Identify Cognate Protein Ligands in the Immunoglobulin Superfamily

Graphical Abstract



Authors

Eng-Hui Yap, Andras Fiser

Correspondence

andras.fiser@einstein.yu.edu

In Brief

Yap et al. developed a method to identify cognate protein ligands for receptors from the IgSF subproteome. The method designs an optimal protein interface using residue preferences obtained from extensive molecular dynamics simulations and then identifies the best-matching ligand structures from the subproteome.

Highlights

- Structure-based method (ProtLID) is developed to find ligands in a subproteome
- ProtLID designs residue-specific pharmacophores derived from molecular dynamics
- The residue-based pharmacophores are matched to candidate ligand structures
- ProtLID is used to explore receptor-ligand interactions in the immunological synapse

Yap & Fiser, 2016, Structure *24*, 2217–2226
December 6, 2016
http://dx.doi.org/10.1016/j.str.2016.10.012

Structure
Resource

ProtLID, a Residue-Based Pharmacophore Approach to Identify Cognate Protein Ligands in the Immunoglobulin Superfamily

Eng-Hui Yap[1,2] and Andras Fiser[1,2,3,*]
[1]Department of Systems and Computational Biology
[2]Department of Biochemistry
Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, NY 10461, USA
[3]Lead Contact
*Correspondence: andras.fiser@einstein.yu.edu
http://dx.doi.org/10.1016/j.str.2016.10.012

SUMMARY

Members of the extracellular immunoglobulin superfamily (IgSF) play a key role in immune regulation through the control of the co-stimulatory pathway, and have emerged as potent drug targets in cancers, infectious diseases, and autoimmunity. Despite the difficult experimental access to this class of proteins, single structures of ectodomains of IgSF proteins are solved with regularity. However, the most biologically critical challenge for this class of proteins is the identification of their cognate ligands that communicate intercellular signals. We describe a conceptually novel method, protein-ligand interface design (ProtLID), to identify cognate ligands from a subproteome for a given target receptor protein. ProtLID designs an optimal protein interface for a given receptor by running extensive molecular dynamics simulations of single-residue probes. The type and location of residue preferences establish a residue-based pharmacophore, which is subsequently used to find potential matches among candidate ligands within a subproteome.

INTRODUCTION

The immunoglobulin superfamily (IgSF) is one of the largest domain families in the human proteome (Lander et al., 2001). Excluding antibodies, the human proteome contains 477 cell-surface and secreted IgSF proteins (extracellular IgSFs) that regulate a broad spectrum of biological processes, ranging from neural development to immune response, primarily through specific cell-to-cell (*trans*) IgSF-IgSF interactions involving their N-terminal Ig domains (Barclay, 2003). Extracellular IgSFs represent a biomedically important class of proteins implicated in cancers (Lee et al., 2012; Wai Wong et al., 2012; Xue et al., 2005), and in infectious (Bergelson et al., 1997; Mendelsohn et al., 1989; White and Littman, 1989) and autoimmune diseases (Bucciarelli et al., 2002; Lebar et al., 1986; Sharpe and Freeman, 2002). They are also proven targets and therapeutic constructs for immunosuppressive drugs (Mansh, 2011; Weber, 2007) and cancer therapies (Watanabe et al., 2005). The importance of extracellular IgSFs is well documented in immune regulation, where *trans*-cellular IgSF receptor-ligand interactions initiate the molecular cascade that culminates in the destruction of foreign pathogens and malignancies, as well as regulate the tolerance mechanisms that protect the host from harmful autoimmune responses (Lenschow et al., 1996; Salomon and Bluestone, 2001).

A systematic understanding of cognate IgSF receptor-ligand relationships, i.e., binding partners and poses, will provide critical insights into the molecular basis of these diseases and help to engineer therapeutic strategies to modulate the underlying receptor-ligand interactions, such as high-affinity mutants of cognate ligands (Larsen et al., 2005). Current knowledge of IgSF interactions is limited: our recent survey of the STRING protein-protein interaction database (Szklarczyk et al., 2011) showed that only 106 (25%) of the extracellular IgSFs have known *trans*-cellular binding partners supported by high-quality, experimental evidence (Yap et al., 2014). Of these, only 16 IgSF:IgSF complexes have had their structures solved by X-ray crystallography to provide detailed insights into the binding interfaces. These 16 IgSF:IgSF complexes are formed by 25 types of IgSF proteins, which can be classified into 9 functional families out of the 72 that we identified (Rubinstein et al., 2013). While advances in experimental automation have made possible large-scale screening of up to ~1,000 receptor-ligand interactions in anecdotal cases (Yu et al., 2009), a brute-force screening of all IgSF-IgSF pairs in the extracellular IgSF subproteome (more than a 100,000 unique combinations) remains intractable. Computational screening is necessary to shortlist potential receptor-ligand pairs for experimental verification.

Novel receptor-ligand relationships can be inferred from known relationships in homologous sequences (Rubinstein et al., 2013), orthologous receptor-ligand pairs (interology; Arslan et al., 2008), and receptor-ligand pairs sharing similar interface features (Tuncbag et al., 2011). We have previously developed sequence-based methods to assign IgSF proteins into coarse- (Rubinstein et al., 2013) and fine-grained (Yap et al., 2014) functional families. In practice, however, inference methods are limited by the paucity of known IgSF-IgSF receptor-ligand information.

In contrast, inference-free, structure-based methods can be applied to any IgSF with a known or computationally modeled

structure. These approaches face a challenging question: given an IgSF receptor structure, which protein ligand(s) could be potential binding partners among the entire extracellular IgSF subproteome? Existing structure-based methods are not suitable for predicting IgSF receptor-ligand interactions. A major class of such methods relates to small-molecule virtual drug screening, where all possible conformations of a candidate ligand are docked onto the target receptor, and an empirical scoring function is used to identify candidates with high affinity (Fradera and Mestres, 2004). Small-molecule methods cannot be directly applied to protein ligands, because the search space (ligand conformations and poses) scales exponentially with ligand size, rendering these methods computationally intractable for a protein ligand. Another class of structure-based methods is based on protein-protein docking methods, which were traditionally developed for predicting the binding poses for two proteins that are known to bind (Smith and Sternberg, 2002). These methods are extensively evaluated in the community-wide Critical Assessment of Prediction of Interaction (Janin, 2005). Attempts to re-purpose protein-protein docking methods to predict potential ligands for a given target are limited by the difficulty of predicting binding affinities accurately (de Vries et al., 2006). A recent community-wide critical assessment demonstrated that, when current state-of-the-art protein-protein docking methods were used to score and distinguish binders from non-binders, no method could consistently identify cognate binders, and the correlation between docking scores and binding affinity was at best weak (Fleishman et al., 2011). The candidate ligand library in our IgSF-specific application poses additional challenges, since all candidates from the IgSF subproteome share the same fold topology, and docking methods that rely heavily on receptor-ligand shape complementarity would be less effective in discriminating cognate from non-cognate IgSFs.

We present here a conceptually new computational approach, protein-ligand interface design (ProtLID) (Figure 1). ProtLID is inspired by two related concepts in small-molecule drug discovery: the molecular interaction field (MIF) (Goodford, 1985) and the pharmacophore (Kier, 1967). A MIF is a grid-based map reflecting the interaction energetics between a receptor and various generic moieties (e.g., water, methyl group, amine nitrogen, carboxyl oxygen, or hydroxyl). A pharmacophore is an abstract description of key atoms, groups, charged regions, and their spatial interrelations that are essential for the biological activities a drug molecule. Both the MIF and the pharmacophore describe a theoretically optimal ligand interface against which small-molecule drug candidates can be screened. To adapt these concepts to a protein ligand, we introduce the concept of residue-specific (rs)-pharmacophoric moieties, which we termed functional atoms (FAs). We focus our design effort on 26 FA types within the 20 naturally occurring amino acid types, specifically, hydrogen-bonding capable side chain oxygen/nitrogen, and hydrophobic/aromatic centers (Table S1) that contribute the most to binding interactions. The optimal FA positions on the receptor interface are determined through exhaustive sampling of small, single-residue ligand probes using molecular dynamics (MD). The resulting FA preferences constitute a unique spatial fingerprint, which we termed the rs-pharmacophore. This designed rs-pharmacophore is then used in the template-based ligand-screening step, where we combinatorially extract subtemplates from the rs-pharmacophore and match these designed, theoretical ligand subinterfaces against candidate structures in the IgSF subproteome.

ProtLID has several conceptual advantages compared with other methods. First, our sampling-derived FA preferences are correlated to free energy, as opposed to empirical scores that are based on a single static snapshot, as in the case of most MIF- and pharmacophore-based methods. This is because FA preferences are quantified by the frequencies that FAs sample in each region during all-atom MD simulations, which are reflective of the Boltzmann distributions of thermally accessible receptor surface positions. The only other known method that employs MD to generate pharmacophores was pioneered by Mackerell and coworker, but with applications for small-molecule and peptide ligands (Guvench and MacKerell, 2009). Second, in contrast to small-molecule drug discovery methods, which use generic moieties, ProtLID uses rs-moieties, thereby limiting the potential number of matches and reducing the combinatorial search space for screening each candidate protein ligand. Third, the ligand-screening step is designed to tolerate variations in crystal structures through its use of a simple energy function, subtemplate matching, and a clustering-based ranking scheme. This advantageous feature over both small-molecule and protein-protein docking methods is proven by the consistent rankings of alternative structures of the same candidate protein, a behavior specific to ProtLID.

We applied ProtLID to 11 IgSF receptors belonging to different functional families (Table 1). In each case, we evaluated how well ProtLID ranks structures of the cognate ligands among a decoy database of IgSF structures. While there is no conceptually similar approach, as the closest alternative we performed comparisons with three state-of-the-art docking algorithms, ClusPro (Comeau et al., 2004a, 2004b; Kozakov et al., 2013; Kozakov et al., 2006), ZDOCK (Pierce et al., 2011), and GRAMM (Katchalskikatzir et al., 1992).

RESULTS

Exploring Criteria for Sampling Convergence

MD simulations of single-residue ligand probes were performed to determine optimal FA positions. First, a 1 Å mesh was generated for each receptor interface, which resulted in the following numbers of mesh points (N_{mesh}): 446 (1F5W.B), 297 (1I85.B), 412 (1I85.D), 355 (1I8L.A), 462 (1I8L.C), 454 (2PTT.A), 453 (2PTT.B), 583 (3BP5.A), 610 (3BP5.B), 478 (3UDW.A), and 427 (3UDW.C). Multiple MD trajectories were launched at each mesh point for each of the 20 single-residue probe types. FA positions were collected at 1 ps intervals from the trajectories and all FA positions that fall within 5 Å of any mesh point were assigned to them. We counted n_m, the number of FAs assigned to each mesh point m. To establish how many MD runs are required for adequate sampling, we used two approaches to monitor the reproducibility of n_m for the receptor 1I85.D (Figure 2).

In the first approach, we prepared two independent datasets, each comprising ~80,000 assigned FAs from 26 FA types. From each dataset, we randomly selected a gradually increasing total number of assigned FAs (N_{FA}) and counted n_m. We repeated the

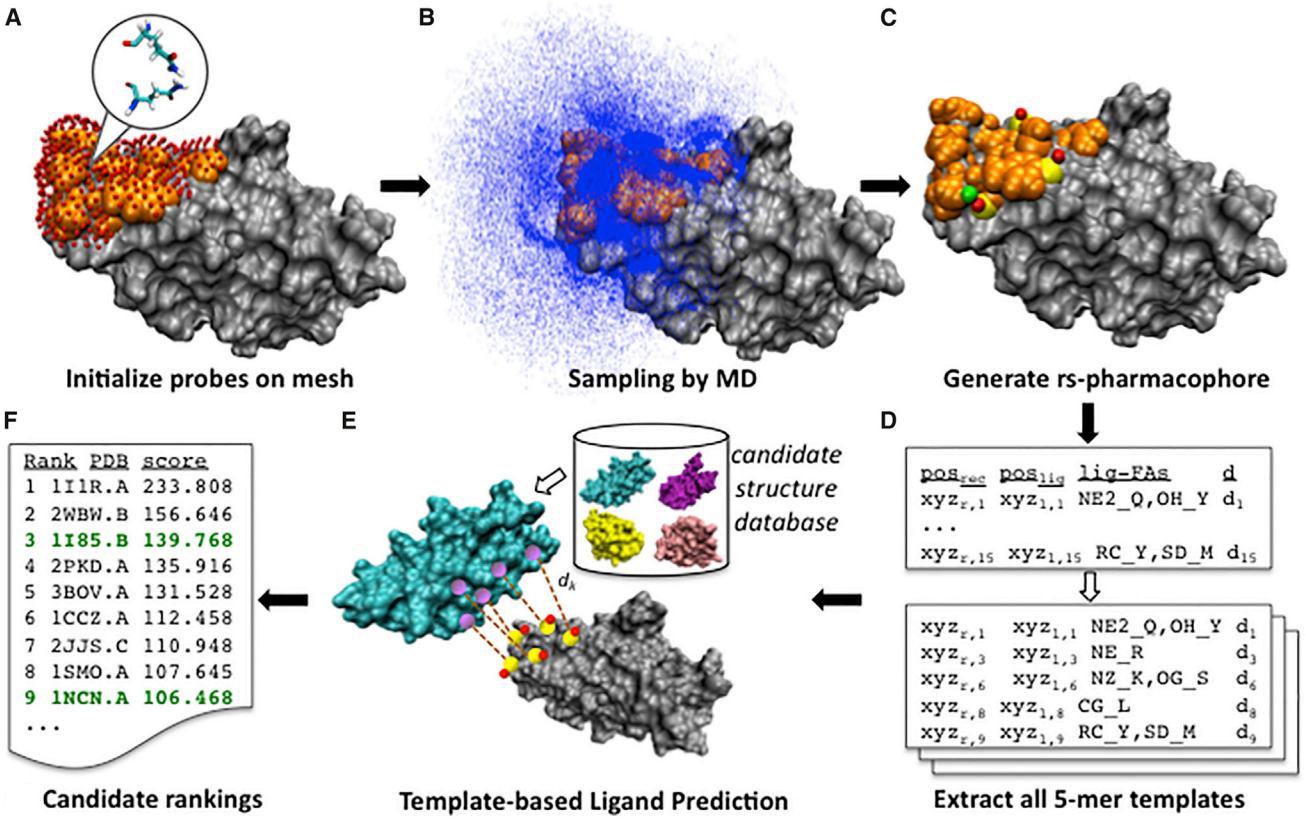


Figure 1. ProtLID Method Illustrated Using Receptor 1I85.D, Gray

(A) A 1 Å mesh (red dots) is generated over the receptor interface (orange). Single-residue probes (illustrated here using glutamine) are placed on each mesh point in different orientations.

(B) Probes are evolved using MD to collect the positions of the corresponding FA (blue for the NE2 atom of glutamine [NE2_Q]).

(C) The spatial distributions of FAs are analyzed to determine the most favored FA types at each receptor site. Three receptor sites favoring NE2_Q are shown in yellow and their corresponding predicted ligand positions in red dots. The actual Gln NE2 position in the bound cognate ligand (green dot) coincides well with the prediction.

(D) The resulting rs-pharmacophore comprises predicted ligand position, FA types, and restraint distances for each receptor site. All possible 5-mer templates are generated exhaustively from the full rs-pharmacophore.

(E) Each candidate structure in the database is matched against each template: matching 5-atom constellations on the candidate (purple) are least-square-fitted onto predicted ligand positions (red dots). Refinement is then performed to optimize steric clash and distance restraints between ligand (purple) and receptor (yellow) atoms.

(F) Qualifying poses are clustered to score and rank all candidate structures. For benchmarking, we monitored the ranks of cognate structures (green entries). See also Table S1.

random selection 20 times for each N_{FA} value and monitored the average correlation coefficient of n_m against N_{FA} (Figure 2, red line; upper axis). The plot suggests that at least ~24,000 assigned FAs are needed to have a correlation greater than 0.85. This translates to approximately seven to eight independent MD runs since each MD run yields on average ~3,000 assigned FAs (data not shown).

In a second approach, we performed 16 independent MD runs for the FA type NE_R (NE FA of Arg residue). The MD runs were divided into two sets of eight runs. Next, we randomly combined n_{runs} runs within each set and computed the correlation coefficient of n_m between the two subsets. We repeated the random combination 20 times and plotted the average correlation coefficient against n_{runs} (Figure 2, blue line; lower axis). The result suggests that about six MD runs are sufficient to reproduce a correlation $\geq$0.85.

Guided by the correlation data from both approaches, and to allow for fluctuations in the number of assignable FAs specific to different FA types, we chose to perform seven MD runs to ensure sampling convergence.

Generating rs-Pharmacophores

rs-Pharmacophores were generated from the statistical analysis of the MD snapshots. A typical rs-pharmacophore has up to 15 predicted "interactors," each comprising a receptor site (atom type and position), a corresponding predicted ligand site (allowed FA types and positions), and a receptor-to-ligand atomic distance restraint. We benchmarked the predictive precision of each pharmacophore, which is defined as the ratio of the number of true positive interactors to the total number of interactors. An interactor is considered to be a true positive if there is at least one ligand atom in the bound cognate structure that matches

Table 1. Eleven *trans*-Binding, Ig1 Interfaces Used in Ligand Prediction

Receptor Interface (Biological Source)	Receptor		Cognate Ligand		
	Protein Name	Functional Family	Protein Name	Structure(s) Available	Interface RMSD (Å)
1F5W.B (h)	CXAR	CXAR	CXAR	**1F5W.A**, 2WBW.B, 1EAJ.A, 1KAC.B, 1P6A.B, 2J12.B, 2J1K.A, 2W9L.A	0.2–2.6
1I85.B (h)	CD86	B7	CTLA4	**1I85.D**, 1I8L.C, 3BX7.C, 3OSK.A	1.2–2.2
1I85.D (h)	CTLA4	CD28/CTLA-4	CD86	**1I85.B**, 1NCN.A	1.4
3OSK.A (h)[a]	CTLA4	CD28/CTLA-4	CD86	**1I85.B**, 1NCN.A	1.4
1I8L.A (h)	CD80	B7	CTLA4	**1I8L.C**, 1I85.D, 3BX7.C, 3OSK.A	1.3–2.0
1I8L.C (h)	CTLA4	CD28/CTLA-4	CD80	**1I8L.A**, 1DR9.A	1.7
2PTT.A (m)	CD48	SLAM	CD244	**2PTT.B**, 2PTU.A	2.1
2PTT.B (m)	CD244	SLAM	CD48	**2PTT.A**, 2PTV.A	1.1
3BP5.A (m)	PDCD1	Extended CD28/CTLA-4	PD1L2	**3BP5.B**, 3BOV.A	1.3–2.4
3BP5.B (m)	PD1L2	B7	PDCD1	**3BP5.A**, 1NPU.A, 3BIK.B	1.7
3UDW.A (h)	TIGIT	Nectin-like	PVR	**3UDW.C**	NA
3UDW.C (h)	PVR	Nectin-like	TIGIT	**3UDW.A**, 3Q0H.A, 3RQ3.A, 3UCR.A	1.0–1.8

See also Table S4.

Receptor interfaces are named as PDB id.receptor chain; protein names are listed by their UniProt ID (source-suffix omitted). Bound ligand structures are bold; biological sources are denoted by h (human) or m (mouse). Interface RMSD refers to the RMSD of interface residues between bound and unbound forms of the same ligand, if available.

[a]*apo* form of the receptor.

allowed FA types and is within the stipulated restraint distance from its receptor site. The precisions of the pharmacophores are 75% (1F5W.B), 75% (1I85.B), 80% (1I85.D), 43% (1I8L.A), 75% (1I8L.C), 63% (2PTT.A), 53% (2PTT.B), 58% (3BP5.A), 70% (3BP5.B), 62% (3UDW.A), and 71% (3UDW.C).

Ligand Prediction Results

For each of the 11 IgSF receptors, we searched its designed rs-pharmacophores against a candidate database (subproteome) of IgSFs represented by their N-terminal Ig (Ig1) domain structures. All available structures of the cognate ligands (Table 1) were also included in the database. The candidates were ranked by ProtLID according to the scoring scheme described in Experimental Procedures. Table 2 reports the percentile ranks of each cognate structure, in addition to the ranking results obtained using ClusPro, ZDOCK, and GRAMM. In case of docking methods ligand candidates were docked only to the receptor interface or docked conformations were filtered to consider only those complexes that interact on the native interface (see Experimental Procedures for details).

The success of a cognate ligand prediction is measured by how often it ranks a cognate ligand better than random and, by more strict but more practical criteria, by how often it ranks a cognate ligand among the top ones (here, top 10%), so it can be shortlisted and experimentally verified (Tables 2 and S5). Of 35 cognate structures, ClusPro, ZDOCK, GRAMM, and ProtLID ranked 4, 10, 3, and 12 cognate structures in the top 10%, respectively. ProtLID's strength is more apparent when we use a more stringent test where the bound cognate ligand structures (underlined in Table 2) were omitted from the list of candidates that are not available when one searches for novel binding partners among a candidate library of apo structures. Of a total of 24 known unbound cognate ligand structures, ClusPro, ZDOCK, GRAMM, and ProtLID ranked 1, 3, 2, and 7 in the top 10%, respectively.

We next assessed the performances using the average percentile rank of a receptor's cognate structures (Figure 3A and Table S3). The motivation for this metric is that a high-quality ligand prediction method must be able to perform well with any cognate structure available in the PDB, irrespective of their structural differences at their interfaces, which can be as high as 2.6 Å (Table 1) root-mean-square deviation (RMSD). Figure 3A shows the average percentile rank of cognate ligands for the four methods, with the gray area representing a performance equal or worse than random. ProtLID, on average, ranks cognate ligands better (lower) than random in 7 out of the 11 receptors. In contrast, ZDOCK, ClusPro, and GRAMM have 2, 5, and 2 receptors with average rank better than random, respectively.

Ligand Specificity: Case Study with CTLA-4 Binding Two Different B7s

We explored how well a ligand prediction method based on a particular receptor:ligand interface can recognize a "cross-ligand," i.e., a ligand that binds to the receptor at a similar but not identical interface. All four ligand prediction methods utilize information about the receptor interface: ClusPro directs docking toward the interface through an attractive potential; ZDOCK and GRAMM poses were filtered to retain only those overlapping with the receptor interface; and ProtLID uses the receptor interface to build its rs-pharmacophore.

We used the 1I85.D and 1I8L.C receptor interfaces to investigate cross-ligand specificity. The 1I85.D interface is derived from a complex of CTLA-4 (chain D) binding to its ligand CD86 (chain B), while the 1I8L.C interface is derived from a complex of CTLA-4 (chain C) binding to its other ligand CD80 (chain A) (Table 1). Both CD86 and CD80 are IgSFs in the B7 family, sharing ~25% sequence similarity (Yap et al., 2014). While

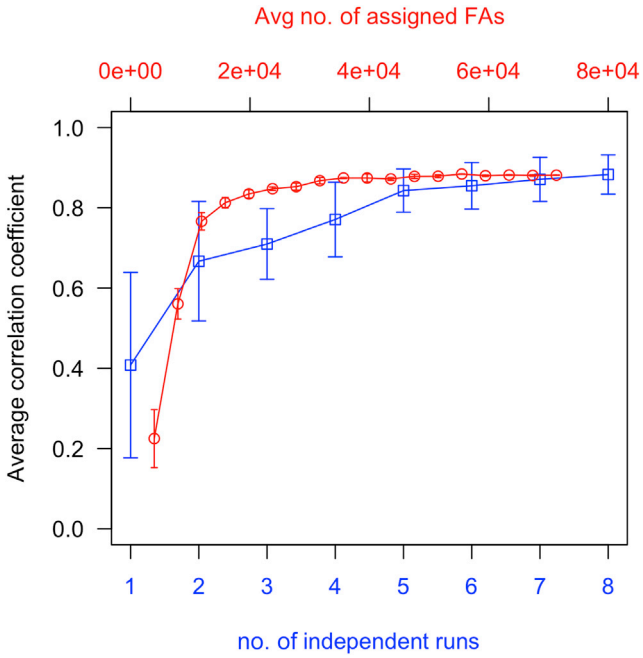


Figure 2. Reproducibility of Sampling

Average correlation coefficient between pairs of datasets combined from: (red line; upper x axis) up to 80,000 assigned FA positions from 26 FA types; or (blue line; lower x axis) up to eight runs of a single FA type (NE_R).

both bind to co-stimulatory receptors CD28 and CTLA-4, the binding affinities of CD86 are ~5- to 10-fold lower than that of CD80 (Bhatia et al., 2006). The CD80 and CD86 binding sites on CTLA-4 overlap significantly with eight common residues that are not identical (Figure 4). CD80 has two unique interface residues (Asp65 and Met97), while CD86 has one (Thr53).

We added structures of cross-ligands to the respective candidate databases for 1I85.D and 1I8L.C, and computed their percentile ranks (Table 2, entries marked with a superior a and b). For the 1I85.D (CD86-binding) interface, none of the methods identified its cross-ligand (CD80; with two available structures, 1I8L.A and 1DR9.A) within the top 10%. However, for the 1I8L.C (CD80-binding) interface, two methods successfully identified its cross-ligand, CD86: ClusPro ranked one of the two available structures, 1I85.B, in the top 5%, while ProtLID ranked both structures (1I85.B and 1NCN.A) in the top 5% and 2%, respectively.

We believe that poorer ranking of CD80 structures is primarily due to the absence of Asp65 from the interface definition of 1I85.D (Figure 4). Asp65 is critical for CTLA-4 binding to CD80, forming a salt bridge with Lys89 of CD80 (Krissinel and Henrick, 2007). The absence of Asp65 from the rs-pharmacophore means that a significant subset of specific interactions cannot be considered when searching for a cognate ligand.

Impact of Using Unbound versus Bound Receptor Structure

We explored the impact of using unbound (3OSK.A), instead of bound (1I85; chain D), receptor structure on ProtLID and the three docking methods, using CTLA-4:CD86 binding as a case study. For ProtLID, while the cognate percentile ranks based on the unbound receptor are poorer (18, 44) than that based on the bound receptor (1, 9), we see that at least one cognate ligand is ranked high, at the 18th percentile. In contrast, all three docking methods performed poorly when switching to unbound receptor structures, with cognate structures all ranking as random (ClusPro: 47, 49, ZDOCK: 62, 60, GRAMM: 43, 59).

ProtLID Ranked an Alternative Ligand for 3UDW.C in the Top 5%

In the case of 3UDW.C (PVR), ClusPro, GRAMM, and ProtLID did not recover any structure of its cognate ligand (TIGIT) within the top 10%. This is likely a consequence of our choice during method design to omit main-chain N and O from the FA definitions, due to their prevalence and lack of specificity. Analysis of the TIGIT:PVR interface shows that TIGIT uses extensively main-chain N and O to interact with PVR (Table S2).

Interestingly, ProtLID did manage to rank another IgSF candidate known to interact with PVR within the top 5%. PVR and TIGIT belong to the nectin/nectin-like IgSF family, the members of which are frequently involved in homophilic or heterophilic binding with other family members. Besides TIGIT, PVR is also reported to have binding interactions with other nectin/nectin-like family members: CD226 (Bottino et al., 2003), CD96 (Fuchs et al., 2004), nectin-1 (Harrison et al., 2012), and nectin-3 (Mueller and Wimmer, 2003). Of these, nectin-1 has a high-resolution structure available (3ALP.A) and was part of the candidate database. ProtLID ranked 3ALP.A at 4.8%. The ranks for 3ALP.A using ClusPro, ZDOCK, and GRAMM are 75.2%, 20.7%, and 87.8%, respectively.

ProtLID and ZDOCK Reliably Rank Native-like Binding Poses as Top Scorers

Two critical bottlenecks in successfully predicting cognate ligands are the adequate sampling of native-like poses and the subsequent accurate scoring of these poses. We defined a "native-like" pose using two criteria: (1) a more stringent criterion that requires the RMSD of interface residues ($RMSD_{int}$) between the bound structure and the pose be $\leq$ 5Å, and (2) a more relaxed criterion that requires 50% of interface residues on the bound structure be also found on the interface of the pose. For each criterion, we counted the number of cognate structures for a method that samples at least one native-like pose or that ranks a native-like pose as the top-scoring pose (Table 3).

According to the $RMSD_{int}$ criterion, ProtLID sampled native-like poses for all 35 cognate structures, while ClusPro, ZDOCK, and GRAMM sampled 23, 27, and 12 respectively (Table 3 i-a). When ranks were also considered, ProtLID had 14 top-scoring cognate ligand structures with native poses, more than ClusPro (5), ZDOCK (11), and GRAMM (0) (Table 3 i-b). When the second, more relaxed native-like criterion is used, it appears that all methods are able to reliably sample native poses (Table 3 ii-a). However, among these, only ZDOCK and ProtLID are able to rank native-like poses as top scorers consistently (Table 3 ii-b; 26 for ZDOCK and 22 for ProtLID). The additional hits found in the relaxed criterion (Table 3 ii-b), in comparison with the stringent criterion (Table 3 i-b), suggests that the poses sometimes predict the correct binding patch, albeit in an incorrect orientation. The strong performances of ZDOCK and ProtLID suggest that a potential side application for these methods could be to

Table 2. Cognate Ligand Prediction Results for 11 IgSF Receptors Using ClusPro, ZDOCK, GRAMM, and ProtLID

Receptor Interface	Percentile Rank of Cognate Ligand Structures			
	ClusPro	ZDOCK	GRAMM	ProtLID
1F5W.B	100, 100, 67, 52, 98, 90, 77, 83	13, **7**, 45, 34, 60, 59, 35, 87	35, 28, 99, 45, **9**, 86, 45, 56	**2**, **1**, 17, 27, 15, **9**, 13, **10**
1I85.B	45, 77, 84, 56	**9**, 29, 44, 20	79, 72, 71, 79	55, 26, 57, 46
1I85.D	18, 12 (45, 54)[a]	**7**, 62 (16, 18)[a]	42, 84 (66, 46)[a]	**1**, **9** (35, 49)[a]
1I8L.A	71, 44, 66, 46	30, 33, 54, 38	16, 12, 18, **6**	88, 74, 43, 41
1I8L.C	**5**, 64 (5, 93)[b]	**9**, 18 (41, 55)[b]	81, 20 (57, 54)[b]	**1**, **4** (5, 2)[b]
2PTT.A	14, 51	14, 36	40, 60	**2**, 20
2PTT.B	56, **9**	49, 48	85, 11	**4**, **2**
3BP5.A	**2**, 46, 28	**1**, **9**, **4**	16, 11, 34	53, 27, 49
3BP5.B	**6**, 20, 40	**5**, 69, 68	**10**, 28, 30	49, **10**, 22
3UDW.A	91	**1**	38	14
3UDW.C	89, 93, 49, 67	**1**, 73, 32, 65	29, 95, 54, 78	51, 46, 78, 13
Total no. of top 10% ranked ligands	4/35	10/35	3/35	12/35
Total no. of top 10% ranked ligands (omitting bound structure)	1/24	3/24	2/24	7/24

See also Tables S2, S3, and S5.

For each receptor interface, we reported the percentile ranks of its available cognate ligand structures, as scored using ClusPro, ZDOCK, GRAMM, and ProtLID, respectively. Percentile ranks are listed in the order of corresponding cognate ligand structures given in Table 1. Underlined, percentile rank of bound cognate structures; bold, percentile ranks less than or equal to 10.

[a]Percentile ranks of cognate structures (1I8L.A, 1DR9.A) for cross-ligand CD80.

[b]Percentile ranks of cognate structures (1I85.B, 1NCN.A) for cross-ligand CD86.

identify the likely ligand binding interface to improve docking performance (Huang and Schroeder, 2008). The breakdown of results by receptors is provided in Table S6.

DISCUSSION

We present a new computational approach, ProtLID, that predicts the cognate protein ligand for a receptor by matching a residue-specific pharmacophore that reflects FA propensities obtained from extensive MD sampling around the receptor interface. We tested our method on 11 known IgSF-IgSF receptor-ligand pairs from functionally diverse IgSF families, and compared its performance against ligand predictions using leading docking algorithms.

This conceptually new approach appears to have several practical advantages over docking methods. First, ProtLID ranked the highest fraction of unbound cognate ligand structures (29%, or 7 out of 24) within the top 10%. In contrast, the three other docking-based methods were less sensitive, ranking only 4%–13% (1–3 out of 24). Second, alternative structures of cognate ligands are ranked consistently by ProtLID, as demonstrated by the superior average cognate percentile ranks over other methods (Figure 3A). This is important because it highlights the fact that ProtLID's performance is consistent, irrespective of small structural variations that exist between similar proteins or for different structures of the same protein when solved under different experimental conditions (Sali, 2001).

These results suggest that the concept implemented in ProtLID is promising as a suitable tool for shortlisting cognate ligands from subproteomes for a given target receptor for experimental verification (Table S5). By employing a computational approach to shortlist ligand candidates one can transform an experimentally unfeasible question to a practical exploration.

The computationally intensive step in ProtLID is the MD sampling of FA propensities to design the rs-pharmacophore. However this step can be executed in a highly parallelized fashion on a computer cluster. Once an rs-pharmacophore is built, it can be used to search quickly against any large set of candidates. This makes the current method a good candidate to explore protein interactions within a variety of subproteomes or the entire set of structurally known proteins.

EXPERIMENTAL PROCEDURES

Datasets Used in This Study

List of Extracellular IgSFs

The set of extracellular human IgSF proteins was obtained as in Rubinstein et al. (2013). Briefly, human IgSF proteins were identified from the UniProt database (Apweiler et al., 2010) if they were predicted to be membrane-integral or secreted by the Phobius program (Kall et al., 2004), and their InterPro identifiers corresponded to Ig domains. Antibodies and T cell receptors were excluded, and the highly polymorphic MHC I/II proteins were represented by only one protein per gene. This resulted in an IgSF dataset with 477 proteins (Table S1 of Yap et al., 2014), of which a subset of 366 has an N-terminal Ig domain (Ig1 set).

trans-Binding IgSF-IgSF Interfaces

We compiled a list of 39 known *trans*-binding, heterophilic IgSF-IgSF pairs (Yap et al., 2014) and seven *trans*-binding, homophilic IgSF-IgSF pairs (Cao et al., 2006; Dong et al., 2006; Harrison et al., 2012; Patzke et al., 2010; Rubinstein et al., 2013; Yan et al., 2007). For each of these IgSF-IgSF pairs, we queried the UniProt database for complex structures with resolution $\leq$3.5 Å, and coverage $\geq$90% for both receptor and ligand chains. We omitted complexes that (1)involved mixed species, (2) did not interact in a *trans*-binding orientation, (3) or had interfaces that involved regions other than the Ig1 domain. Sixteen X-ray structures of *trans*-binding human or mouse IgSF-IgSF were found that contain a total of 25 unique IgSF interfaces belonging

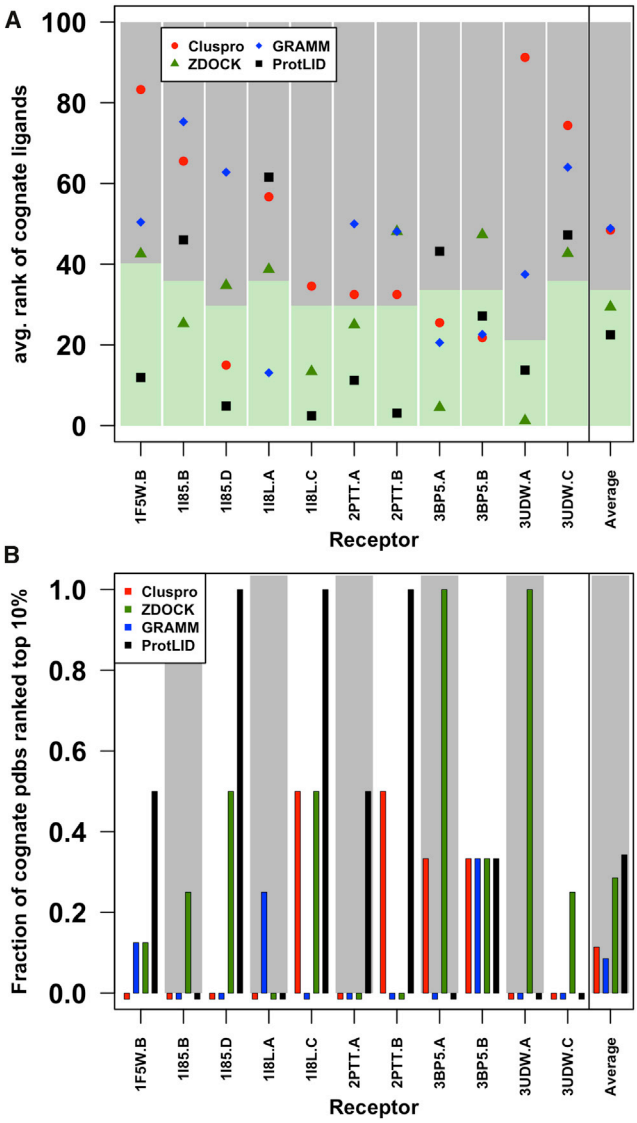


Figure 3. Performance of ProtLID versus Three Docking Approaches on 11 IgSF Receptors, x Axis

(A) Average percentile ranks of cognate structures. Insignificant (random) ranking is at the 50th percentile. Ranking equal or worse than random shown in gray.

(B) Fraction of cognate structures ranked within the top 10%. Methods with zero hits (no cognate ranked within 10%) are shown as negative bars.

to nine functional families. Eleven receptor interfaces from the largest families were selected for the current study (Table 1).

Candidate Structure Database

We queried the UniProt database for X-ray structures (resolution $\leq$3.0 Å, coverage $\geq$90%) for the 366 extracellular IgSFs in the Ig1 set. When no structure of the human protein was found, we collected available structures from mouse and rat homologs. A total of 79 Ig1 structures were found (Table S4), comprising 62 human, 15 mouse, and 2 rat structures. The performance of ProtLID for mouse and rat structures were comparable with human structures, indicating no intrinsic bias against non-human sources. For each receptor, we supplemented the candidate database with all available structures of its cognate ligand (resolution $\leq$3.5 Å, coverage $\geq$90%; ligand and receptor proteins originating from the same species). Structures with mutations were excluded. The list of cognate PDBs are provided in Table 1.

ProtLID Protocol

An overview of the method is shown in Figure 1.

Probe-Based Pharmacophore Generation

Mesh Placement over Receptor Interface. Interface atoms on the receptor were identified from solved complex structures (Figure 1A) using the CSU program (Sobolev et al., 1999) with the following additional criteria: (1) contact distance $\leq$4 Å; (2) contacting atoms are from legitimate CSU classes (Table 1 of Sobolev et al., 1999); (3) solvent-accessible surface area per NACCESS (Hubbard and Thornton, 1993) is greater than 1 Å^2; (4) carbon or sulfur atoms must come from hydrophobic residues. A 1 Å mesh was then generated on the solvent-accessible surface of the receptor using the EDTSurf software (Xu and Zhang, 2009) with a probe radius of 1 Å.

Sampling Using Molecular Dynamics. Sampling for optimal FA positions (Figure 1B) was conducted by simulating single-residue "ligand probes" on the receptor surface using the AMBER MD package (Case et al., 2005). Uncapped (N-H at N-terminal and C=O at C-terminal) ligand probes were placed on each interface mesh point. The receptor-probe system was minimized for 5,000 steps, with harmonic restraints on receptor-heavy atoms and ligand FAs gradually reduced from 5.0 kcal/mol to zero. MD simulations were performed at T = 300 K using Andersen thermal coupling and implicit solvation with no periodic boundary condition. A distance restraint of 15 Å was maintained between the ligand FA and the starting mesh point. Each trajectory was evolved for 40 ps, with energy-minimized snapshots of the probe collected every 1 ps. Seven independent MD runs were launched from each mesh point by varying the initial orientation of the probe with respect to the receptor.

rs-Pharmacophore Generation from MD Snapshots. To generate an rs-pharmacophore (Figure 1C), MD data collected for 26 FA types were analyzed in several steps. For each FA type f and MD run r, we filtered the snapshots for the presence of legitimate hydrogen bond donor-acceptor interactions or hydrophobic contacts. For each mesh point m, we computed the number of assigned FAs ($c_{f,m}$), averaged over seven MD runs.

Protrusions and indentations on the receptor surface can lead to under- and over-sampling artifacts. We addressed this problem by normalizing $c_{f,m}$ by its expected value as follows. We computed n_m, the FA-independent reference mesh count, by assigning FA positions from all snapshots to the nearest mesh point, averaging the FA counts at each mesh point over seven MD runs. We then used n_m to generate $c_{f,m,expected}$, the "expected" count at mesh point m for a FA type f that belongs to the interaction class i:

$$c_{f,m,expected} = \frac{n_m}{\sum_{m\in\{m\_i\}} n_m} \cdot \sum_{m\in\{m\_i\}} c_{f,m} \ ,$$

where $\{m\_i\}$ is the set of mesh points for which FAs of interaction type i are expected to have non-zero counts.

The ratio of actual-to-expected assigned FA counts, A/E ratio$_{f,m}$ = $c_{f,m}/c_{f,m,expected}$, represents the propensity distribution of FA f at mesh m, after normalizing for surface geometry-based sampling effects.

While the reference distribution used all 26 FA types, three FA types (O_G, N_G, and CB_C) were omitted from our final rs-pharmacophore design: O_G and N_G were not used since main-chain N and O are ubiquitous in all ligand structures and likely to produce spurious matches and extend the search time; CB_C was omitted because solvent-exposed, unpaired cysteines are infrequent in the oxidative extracellular environment (Fiser and Simon, 2002).

Of the remaining 23 N_{mesh} (f,m) combinations on the receptor interface, we retained (f,m) with above average propensities (i.e., A/E ratio $\geq$ 1) and at least one occurrence ($c_{f,m} \geq 1$). These were consolidated into "interactors" on the receptor interface, where each interactor comprises (1) a receptor site position, (2) the predicted ligand position, given by the mesh point closest to the geometric center of all assigned (f,m), (3) the permitted ligand FA types, given by the FA types found in the top 20% (f,m) when ranked by A/E ratio$_{f,m}$, and (4) a stipulated restraint distance between the receptor site and matching ligand atom. The complete set of interactors constitutes the rs-pharmacophore.

Template-Based Cognate Ligand Prediction Using rs-Pharmacophores

Each candidate structure is compared with the N-interactor rs-pharmacophore via $^{N}C_5$ 5-interactor templates generated from the full rs-pharmacophore (Figure 1D). For each template, we exhaustively searched a candidate structure for matching 5-mer atom constellations ("constellation search"), and

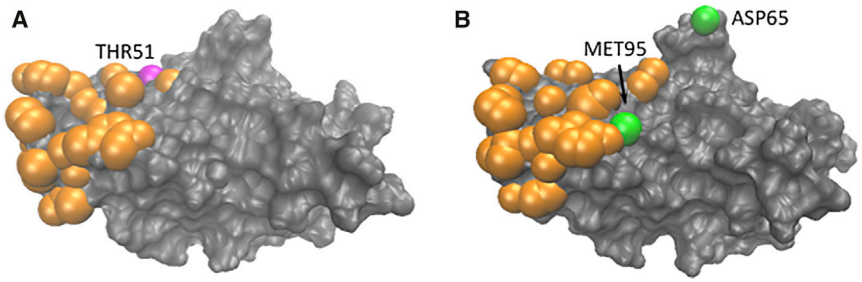


Figure 4. Overlapping Binding Interfaces of CD86 and CD80

CTLA-4 binding interfaces for (A) CD86 (PDB: 1I85) and (B) CD 80 (PDB: 1I8L). Orange, residues common to both interfaces; magenta and green, residues unique to CD86 and CD80 binding interfaces, respectively.

then refined the resulting poses of the candidate structure. Low-energy poses from all templates are clustered, and the number of poses is used as a measure of similarity to the rs-pharmacophore.

Constellation Search. The constellation search (Figure 1E) comprised the following steps: (1) For each interactor in the template we identified all candidate atoms matching its allowed FA types. (2) From these we generated exhaustively all 5-mer atom combinations, and retained those ligand matches where all of the inter-atom pairwise distances were deviating by less than 4 Å. (3) Ligands with matching constellations were least-square fitted onto their corresponding rs-pharmacophore positions, resulting in the guided docking of the full candidate structure onto the receptor.

Energy Refinement. Docked poses were refined to minimize clashes and optimize distances between the five candidate-receptor atom pairs using the following energy function:

$$E = E_{clash} + E_{restra\ int}$$

$$E_{clash} = \begin{cases} 100,000 & \text{if } p_{clash} \geq p_{clash,max} \\ 0.5\, k_{clash}\left(p_{clash} - p_{clash,\,max}\right)^2 & \text{if } p_{clash,min} \leq p_{clash} < p_{clash,max} \\ 0 & \text{if } p_{clash} < p_{clash,\,min} \end{cases}$$

$$E_{restra\ int} = \sum_{k=1}^{5} E_{restra\ int,\,k}$$

$$E_{restra\ int,\,k} = \begin{cases} 0.05\, k_{upper}\left(r_{cut,k} - r\right)^2 & \text{if } r > r_{cut,k} \\ 0 & \text{if } 3.0 \leq r < r_{cut,k} \\ 0.5\, k_{lower} r^2 & \text{if } r < 3.0 \end{cases},$$

where p_{clash} is the percentage of receptor-heavy atoms that are within 3 Å of any candidate atoms, $p_{clash,min}$ and $p_{clash,max}$ set to 5% and 20%, respectively; and $k_{clash} = 554$. The restraint energy is the sum of five harmonic distance restraints, with $k_{upper} = 10.0$, $k_{lower} = 222.22$, and $r_{cut,k} = 4.5$–4.7 Å for hydrogen bond receptor sites and 8 Å for hydrophobic receptor sites. Candidate poses with no energy violation ($E = 0$) were kept and clustered at 5 Å RMSD of interface atoms.

Table 3. Sampling of Native-like Poses

Metric	ClusPro	ZDOCK	GRAMM	ProtLID
(1) Native-like = $RMSD_{int} \leq 5$Å				
(a) No. of cognates that sampled native-like pose(s)	23/35	27/35	12/35	35/35
(b) No. of cognates whose top-scoring pose is native-like	5/35	11/35	0/35	14/35
(2) Native-like = share ≥50th percentile of native ligand interface residues				
(a) No. of cognates that sampled native-like pose(s)	35/35	35/35	35/35	35/35
(b) No. of cognates whose top-scoring pose is native-like	14/35	26/35	5/35	22/35

See also Table S6.

For each of the four methods (ClusPro, ZDOCK, GRAMM and ProtLID), the docked pose is analyzed if it is native-like using two definitions: (1) interface RMSD ($RMSD_{int}$) ≤5Å and (2) 50% overlap between a docked pose's interface residues and that of the bound ligand interface. For each definition, we report the number of cognate ligand structures that sampled at least one native-like pose and the number of cognate ligands the top-scoring (most prevalent) docked pose of which is native-like.

Scoring and Ranking Candidate Ligand Structures

For each pose, we computed the interface area, $A_{int} = A_{receptor} + A_{ligand} - A_{complex}$, using the NACCESS program, and converted this into a normalized score ($norm\_A_{int}$) between 0 and 1. The score for candidate structure *I* is then given by the sum of poses in its largest cluster, each weighted by its $norm\_A_{int}$. The percentile ranking of candidate *I* is the number of other candidates with equal or better scores, divided by the total number of candidate structures ×100 (Figure 1F).

Ligand Prediction Using Docking Algorithms

ClusPro

Each receptor was docked against each candidate structure using the ClusPro server (http://cluspro.bu.edu; Brenke et al., 2012; Comeau et al., 2004a, 2004b; Kozakov et al., 2013; Kozakov et al., 2006) with three different force fields: antigen:antibody (ab), enzyme:inhibitor (ei), and other (o). For each docking submission we also provided the list of core receptor interface residues for which attractive interactions were to be applied. For each force field category, we scored and ranked each candidate based on the number of poses in its biggest cluster. The average rank across three force fields was then used to re-rank the candidates and compute the percentile ranks, similar to ProtLID.

ZDOCK and GRAMM

The partner search procedures for ZDOCK (Pierce et al., 2011) and GRAMM were similar. For each receptor:candidate structure docking, 500 poses were generated. For GRAMM, the grid step was increased from 1.7 Å (default) to 2.0 or 2.5 Å whenever necessary to accommodate the structures within the grid. The poses were post-filtered to retain only those that docked to the receptor interface. Specifically, the receptor interface in the pose (atoms within 4 Å of any candidate atom) must share at least $N_{recIntRes,\ overlap}$ common residues with the actual receptor interface. Various values of $N_{recIntRes,\ overlap}$ were explored but there was no systematic trend in performance for ZDOCK and GRAMM. We set $N_{recIntRes,\ overlap} = 5$ to correspond with the 5-interactor template criterion used in ProtLID. The score of a candidate structure was the lowest energy among those of the remaining poses. Percentile ranks were computed as in ProtLID.

Random Ranking Benchmarks

For each receptor, we sorted structures in its candidate database randomly, and used the resulting ranks of cognate structures to compute the average cognate ranks. Up to 100,000 random trials were performed to ensure that the average and SD converged to stable values.

Native-like Pose Analysis

Poses were analyzed to determine if they werer native-like using two different criteria. For each complex structure, we first obtained the list of true ligand

interface residues, as provided by the CSU program and with the additional requirements that all contact distances be $\leq$4 Å, and that the atom pairs must be legitimate as defined by CSU. The first native-like criterion is based on $RMSD_{int}$, the RMSD of true ligand interface residues between the bound ligand structure and the ligand pose. A pose with $RMSD_{int} \leq 5$ Å is considered native-like. The second criterion is based on the percentage of shared interface residues. For each pose, we determined its interface residues, defined as residues with at least one atom within 4 Å of any receptor atom. If a pose contained $\geq$50% of the true ligand interface residues, it was considered native-like. For each partner search method, we then counted the number of cognate structures that sampled any native-like pose and for which the top-scoring pose was native-like. For ZDOCK and GRAMM, the poses analyzed always satisfied the $N_{recIntRes,\ overlap}$ requirement (see above). For ClusPro, the reported values are the averages from three force fields.

Availability

The ProtLID algorithm is available from the authors by request, free of charge for academic or non-profit use. A provisional patent application has been filed.

SUPPLEMENTAL INFORMATION

Supplemental Information includes six tables and can be found with this article online at http://dx.doi.org/10.1016/j.str.2016.10.012.

AUTHOR CONTRIBUTIONS

Conceptualization, A.F.; Methodology and Investigation, A.F. and E.Y.; Software, E.Y.; Writing, A.F. and E.Y.; Supervision and Funding Acquisition, A.F.

ACKNOWLEDGMENTS

This work was supported by NIH grant R01 GM118709, and the Extreme Science and Engineering Discovery Environment (XSEDE) project (NSF grant ACI-1053575).

Received: November 4, 2015
Revised: July 26, 2016
Accepted: October 25, 2016
Published: November 23, 2016

REFERENCES

Apweiler, R., Martin, M.J., O'Donovan, C., Magrane, M., Alam-Faruque, Y., Antunes, R., Barrell, D., Bely, B., Bingley, M., Binns, D., et al. (2010). The universal protein resource (UniProt) in 2010. Nucleic Acids Res. *38*, D142–D148.

Arslan, M., Boyce, M.C., Qi, H.J., and Ortiz, C. (2008). Constitutive modeling of the stress-stretch behavior of two-dimensional triangulated macromolecular networks containing folded domains. J. Appl. Mech-T ASME *75*, 011020.

Barclay, A.N. (2003). Membrane proteins with immunoglobulin-like domains – a master superfamily of interaction molecules. Semin. Immunol. *15*, 215–223.

Bergelson, J.M., Cunningham, J.A., Droguett, G., KurtJones, E.A., Krithivas, A., Hong, J.S., Horwitz, M.S., Crowell, R.L., and Finberg, R.W. (1997). Isolation of a common receptor for coxsackie B viruses and adenoviruses 2 and 5. Science *275*, 1320–1323.

Bhatia, S., Edidin, M., Almo, S.C., and Nathenson, S.G. (2006). B7-1 and B7-2: similar costimulatory ligands with different biochemical, oligomeric and signaling properties. Immunol. Lett. *104*, 70–75.

Bottino, C., Castriconi, R., Pende, D., Rivera, P., Nanni, M., Carnemolla, B., Cantoni, C., Grassi, J., Marcenaro, S., Reymond, N., et al. (2003). Identification of PVR (CD155) and nectin-2 (CD112) as cell surface ligands for the human DNAM-1 (CD226) activating molecule. J. Exp. Med. *198*, 557–567.

Brenke, R., Hall, D.R., Chuang, G.Y., Comeau, S.R., Bohnuud, T., Beglov, D., Schueler-Furman, O., Vajda, S., and Kozakov, D. (2012). Application of asymmetric statistical potentials to antibody-protein docking. Bioinformatics *28*, 2608–2614.

Bucciarelli, L.G., Wendt, T., Rong, L., Lalla, E., Hofmann, M.A., Goova, M.T., Taguchi, A., Yan, S.F., Yan, S.D., Stern, D.M., et al. (2002). RAGE is a multiligand receptor of the immunoglobulin superfamily: implications for homeostasis and chronic disease. Cell Mol. Life Sci. *59*, 1117–1128.

Cao, E., Ramagopal, U.A., Fedorov, A., Fedorov, E., Yan, Q.R., Lary, J.W., Cole, J.L., Nathenson, S.G., and Almo, S.C. (2006). NTB-A receptor crystal structure: insights into homophilic interactions in the signaling lymphocytic activation molecule receptor family. Immunity *25*, 559–570.

Case, D.A., Cheatham, T.E., Darden, T., Gohlke, H., Luo, R., Merz, K.M., Onufriev, A., Simmerling, C., Wang, B., and Woods, R.J. (2005). The Amber biomolecular simulation programs. J. Comput. Chem. *26*, 1668–1688.

Comeau, S.R., Gatchell, D.W., Vajda, S., and Camacho, C.J. (2004a). ClusPro: a fully automated algorithm for protein-protein docking. Nucleic Acids Res. *32*, W96–W99.

Comeau, S.R., Gatchell, D.W., Vajda, S., and Camacho, C.J. (2004b). ClusPro: an automated docking and discrimination method for the prediction of protein complexes. Bioinformatics *20*, 45–50.

de Vries, S.J., van Dijk, A.D.J., and Bonvin, A.M.J.J. (2006). WHISCY: what information does surface conservation yield? Application to data-driven docking. Proteins *63*, 479–489.

Dong, X.H., Xu, F., Gong, Y.H., Gao, J., Lin, P., Chen, T., Peng, Y., Qiang, B.Q., Yuan, J.G., Peng, X.Z., et al. (2006). Crystal structure of the v domain of human nectin-like molecule-1/Syncam3/Tsll1/Igsf4b, a neural tissue-specific immunoglobulin-like cell-cell adhesion molecule. J. Biol. Chem. *281*, 10610–10617.

Fiser, A., and Simon, I. (2002). Predicting redox state of cysteines in proteins. Method Enzymol. *353*, 10–21.

Fleishman, S.J., Whitehead, T.A., Strauch, E.M., Corn, J.E., Qin, S.B., Zhou, H.X., Mitchell, J.C., Demerdash, O.N.A., Takeda-Shitaka, M., Terashi, G., et al. (2011). Community-wide assessment of protein-interface modeling suggests improvements to design methodology. J. Mol. Biol. *414*, 289–302.

Fradera, X., and Mestres, J. (2004). Guided docking approaches to structure-based design and screening. Curr. Top Med. Chem. *4*, 687–700.

Fuchs, A., Cella, M., Giurisato, E., Shaw, A.S., and Colonna, M. (2004). Cutting edge: CD96 (Tactile) promotes NK cell-target cell adhesion by interacting with the poliovirus receptor (CD155). J. Immunol. *172*, 3994–3998.

Goodford, P.J. (1985). A computational-procedure for determining energetically favorable binding-sites on biologically important macromolecules. J. Med. Chem. *28*, 849–857.

Guvench, O., and MacKerell, A.D. (2009). Computational fragment-based binding site identification by ligand competitive saturation. PLoS Comput. Biol. *5*, e1000435.

Harrison, O.J., Vendome, J., Brasch, J., Jin, X.S., Hong, S., Katsamba, P.S., Ahlsen, G., Troyanovsky, R.B., Troyanovsky, S.M., Honig, B., et al. (2012). Nectin ectodomain structures reveal a canonical adhesive interface. Nat. Struct. Mol. Biol. *19*, 906–915.

Huang, B., and Schroeder, M. (2008). Using protein binding site prediction to improve protein docking. Gene *422*, 14–21.

Hubbard, S.J., and Thornton, J.M. (1993). 'NACCESS', Computer Program (Department of Biochemistry and Molecular Biology, University College London).

Janin, J. (2005). Assessing predictions of protein-protein interaction: the CAPRI experiment. Protein Sci. *14*, 278–283.

Kall, L., Krogh, A., and Sonnhammer, E.L.L. (2004). A combined transmembrane topology and signal peptide prediction method. J. Mol. Biol. *338*, 1027–1036.

Katchalskikatzir, E., Shariv, I., Eisenstein, M., Friesem, A.A., Aflalo, C., and Vakser, I.A. (1992). Molecular-surface recognition – determination of geometric fit between proteins and their ligands by correlation techniques. Proc. Natl. Acad. Sci. USA *89*, 2195–2199.

Kier, L.B. (1967). Molecular orbital calculation of preferred conformations of acetylcholine, muscarine, and muscarone. Mol. Pharmacol. *3*, 487–494.

Kozakov, D., Brenke, R., Comeau, S.R., and Vajda, S. (2006). PIPER: an FFT-based protein docking program with pairwise potentials. Proteins *65*, 392–406.

Kozakov, D., Beglov, D., Bohnuud, T., Mottarella, S.E., Xia, B., Hall, D.R., and Vajda, S. (2013). How good is automated protein docking? Proteins *81*, 2159–2166.

Krissinel, E., and Henrick, K. (2007). 'Protein interfaces, surfaces and assemblies' service PISA at the European Bioinformatics Institute. J. Mol. Biol. *372*, 774–797.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Larsen, C.P., Pearson, T.C., Adams, A.B., Tso, P., Shirasugi, N., Strobert, E., Anderson, D., Cowan, S., Price, K., Naemura, J., et al. (2005). Rational development of LEA29Y (belatacept), a high-affinity variant of CTLA4-Ig with potent immunosuppressive properties. Am. J. Transpl. *5*, 443–453.

Lebar, R., Lubetzki, C., Vincent, C., Lombrail, P., and Boutry, J.M. (1986). The M2 autoantigen of central-nervous-system myelin, a glycoprotein present in oligodendrocyte membrane. Clin. Exp. Immunol. *66*, 423–434.

Lee, G., Zhu, M.G., Ge, B.X., and Potzold, S. (2012). Widespread expressions of immunoglobulin superfamily proteins in cancer cells. Cancer Immunol. Immun. *61*, 89–99.

Lenschow, D.J., Walunas, T.L., and Bluestone, J.A. (1996). CD28/B7 system of T cell costimulation. Annu. Rev. Immunol. *14*, 233–258.

Mansh, M. (2011). Ipilimumab and cancer immunotherapy: a new hope for advanced stage melanoma. Yale J. Biol. Med. *84*, 381–389.

Mendelsohn, C.L., Wimmer, E., and Racaniello, V.R. (1989). Cellular receptor for poliovirus – molecular-cloning, nucleotide-sequence, and expression of a new member of the immunoglobulin superfamily. Cell *56*, 855–865.

Mueller, S., and Wimmer, E. (2003). Recruitment of nectin-3 to cell-cell junctions through trans-heterophilic interaction with CD155, a vitronectin and poliovirus receptor that localizes to alpha(v)beta(3) integrin-containing membrane microdomains. J. Biol. Chem. *278*, 31251–31260.

Patzke, C., Max, K.E.A., Behlke, J., Schreiber, J., Schmidt, H., Dorner, A.A., Kroger, S., Henning, M., Otto, A., Heinemann, U., et al. (2010). The coxsackievirus-adenovirus receptor reveals complex homophilic and heterophilic interactions on neural cells. J. Neurosci. *30*, 2897–2910.

Pierce, B.G., Hourai, Y., and Weng, Z.P. (2011). Accelerating protein docking in ZDOCK using an advanced 3D convolution library. PLoS One *6*, e24657.

Rubinstein, R., Ramagopal, U.A., Nathenson, S.G., Almo, S.C., and Fiser, A. (2013). Functional classification of immune regulatory proteins. Structure *21*, 766–776.

Sali, A. (2001). Comparative protein structure modeling of genes and genomes. Abstr. Pap. Am. Chem. S *222*, U390.

Salomon, B., and Bluestone, J.A. (2001). Complexities of CD28/B7: CTLA-4 costimulatory pathways in autoimmunity and transplantation. Annu. Rev. Immunol. *19*, 225–252.

Sharpe, A.H., and Freeman, G.J. (2002). The B7-CD28 superfamily. Nat. Rev. Immunol. *2*, 116–126.

Smith, G.R., and Sternberg, M.J.E. (2002). Prediction of protein-protein interactions by docking methods. Curr. Opin. Struct Biol. *12*, 28–35.

Sobolev, V., Sorokine, A., Prilusky, J., Abola, E.E., and Edelman, M. (1999). Automated analysis of interatomic contacts in proteins. Bioinformatics *15*, 327–332.

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P., et al. (2011). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res. *39*, D561–D568.

Tuncbag, N., Gursoy, A., Nussinov, R., and Keskin, O. (2011). Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM. Nat. Protoc. *6*, 1341–1354.

Wai Wong, C., Dye, D.E., and Coombe, D.R. (2012). The role of immunoglobulin superfamily cell adhesion molecules in cancer metastasis. Int. J. Cell Biol. *2012*, 340296.

Watanabe, T., Suda, T., Tsunoda, T., Uchida, N., Ura, K., Kato, T., Hasegawa, S., Satoh, S., Ohgi, S., Tahara, H., et al. (2005). Identification of immunoglobulin superfamily 11 (IGSF11) as a novel target for cancer immunotherapy of gastrointestinal and hepatocellular carcinomas. Cancer Sci. *96*, 498–506.

Weber, J. (2007). Review: anti-CTLA-4 antibody ipilimumab: case studies of clinical response and immune-related adverse events. Oncologist *12*, 864–872.

White, J.M., and Littman, D.R. (1989). Viral receptors of the immunoglobulin superfamily. Cell *56*, 725–728.

Xu, D., and Zhang, Y. (2009). Generating triangulated macromolecular surfaces by Euclidean distance transform. PLoS One *4*, e8140.

Xue, F., Zhang, Y., Liu, F., Jing, J., and Ma, M. (2005). Expression of IgSF in salivary adenoid cystic carcinoma and its relationship with invasion and metastasis. J. Oral Pathol. Med. *34*, 295–297.

Yan, Q.R., Malashkevich, V.N., Fedorov, A., Fedorov, E., Cao, E., Lary, J.W., Cole, J.L., Nathenson, S.G., and Almo, S.C. (2007). Structure of CD84 provides insight into SLAM family function. Proc. Natl. Acad. Sci. USA *104*, 10583–10588.

Yap, E.H., Rosche, T., Almo, S., and Fiser, A. (2014). Functional clustering of immunoglobulin superfamily proteins with protein-protein interaction information calibrated hidden Markov model sequence profiles. J. Mol. Biol. *426*, 945–961.

Yu, X., Harden, K., Gonzalez, L.C., Francesco, M., Chiang, E., Irving, B., Tom, I., Ivelja, S., Refino, C.J., Clark, H., et al. (2009). The surface protein TIGIT suppresses T cell activation by promoting the generation of mature immunoregulatory dendritic cells. Nat. Immunol. *10*, 48–57.