LTHREADER: Prediction of extracellular ligand–receptor interactions in cytokines using localized threading

VINAY PULIM,¹ JADWIGA BIENKOWSKA,^{1,2} AND BONNIE BERGER^{1,3}

¹Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139 USA

²Department of Biomedical Engineering, Boston University, Boston, Massachusetts 02215, USA

³Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

(RECEIVED August 16, 2007; FINAL REVISION October 23, 2007; ACCEPTED October 23, 2007)

Abstract

Identification of extracellular ligand–receptor interactions is important for drug design and the treatment of diseases. Difficulties in detecting these interactions using high-throughput experimental techniques motivate the development of computational prediction methods. We propose a novel threading algorithm, LTHREADER, which generates accurate local sequence–structure interface alignments and integrates various statistical scores and experimental binding data to predict interactions within ligand–receptor families. LTHREADER uses a profile of secondary structure and solvent accessibility predictions with residue contact maps to guide and constrain alignments. Using a decision tree classifier and low-throughput experimental data for training, it combines information inferred from statistical interaction potentials, energy functions, correlated mutations, and conserved residue pairs to predict interactions. We apply our method to cytokines, which play a central role in the development of many diseases including cancer and inflammatory and autoimmune disorders. We tested our approach on two representative families from different structural classes (all- α and all- β proteins) of cytokines. In comparison with the state-of-the-art threader RAPTOR, LTHREADER generates on average 20% more accurate alignments of interacting residues. Furthermore, in cross-validation tests, LTHREADER correctly predicts experimentally confirmed interactions for a common binding mode within the 4-helical long-chain cytokine family with 75% sensitivity and 86% specificity with 40% gain in sensitivity compared to RAPTOR. For the TNF-like family our method achieves 70% sensitivity with 55% specificity with 70% gain in sensitivity. LTHREADER combines information from multiple complex templates when such data are available. When only one solved structure is available, a localized PSI-BLAST approach also outperforms standard threading methods with 25%–50% improvements in sensitivity.

Keywords: threading; protein interactions; statistical scores; cytokines; 4-helical bundles; TNF-like

Interaction of extracellular ligands and their receptors occupy a central role in intercellular signaling and biological processes that lead to the development and progression of many diseases. Of particular importance to human diseases are cytokines. Cytokine interactions with

their receptors are responsible for innate and adaptive immunity, hematopoiesis, and cell proliferation. Etiology of cancer and autoimmune disorders can be attributed in part to cytokine signaling through their receptors. For example, long-chain 4-helical bundle cytokines, erythropoietin, and human growth hormone are already used for the treatment of cancer and growth disorders. Many other therapies altering cytokine–receptor interactions are in clinical development (Pestka et al. 2004).

Currently, methods exist for predicting general protein–protein interactions (PPIs) that have achieved some

Reprint requests to: Jadwiga Bienkowska; MIT, CSAIL, Cambridge, MA 02139, USA; e-mail: jadwiga@csail.mit.edu; fax: (617) 258-5429; or Bonnie Berger, Department of Mathematics, MIT, Cambridge, MA 02139, USA; e-mail: bab@csail.mit.edu; fax: (617) 258-5429.

Article published online ahead of print. Article and publication date are at <http://www.proteinscience.org/cgi/doi/10.1110/ps.073178108>.

degree of success, relying mostly on data obtained from high-throughput (HTP) experiments such as yeast-two-hybrid (Y2H) screens. However, extracellular ligand–receptor interactions are much more difficult to predict than general PPIs, and methods that work well for PPIs often fail when applied to ligand–receptor binding pairs. In particular, the lack of HTP experimental data for these interactions makes it difficult to apply existing prediction methods that depend on this information (see Related work section below). The HTP experiments such as Y2H that can monitor interactions in the intracellular environment are not appropriate for investigation of extracellular proteins.

We consider the problem of predicting whether an extracellular ligand and receptor interact, given only their sequence information and several confirmed ligand–receptor PPIs among members of the same structural SCOP family (Murzin et al. 1995). Even when one or more complex structures are available within a ligand–receptor family, it is often a challenge to effectively use this information to predict interactions among other members of the family. One reason is the difficulty in identifying the interacting residues that are common among distant family members. The conformational differences that often occur at the interface of bound proteins make such identification non-obvious. In Figure 1 we compare the structural alignments for two families of cytokines. The global structural alignment methods do not generate accurate alignments at the interfaces (RMSD errors of 4.09 Å and 2.75 Å for the 4-helical and TNF-like families, respectively). The alignment of only one interacting domain (e.g., ligand or receptor) from the complex also leads to poor alignment at the interface. In comparison, when only the interaction region was considered, the alignment is much improved (RMSD errors of 1.96 Å and 1.73 Å, respectively). Clearly, the “localized alignment” based on just the interface residues is able to capture the structural variation that exists on the interface surface. By generating templates based on just the interface surface, we are able to better capture this variation.

Our approach is to thread the sequences onto the binding interface of a solved ligand–receptor complex and to evaluate the complementarity of the resulting surface. In so doing, we face four challenges: (1) identifying the residues at the binding interface that are common to a ligand–receptor family; (2) threading the query sequences onto the binding interface; (3) scoring the resulting threaded sequences in order to differentiate between binding and nonbinding partners; and (4) evaluating the significance of the predicted interaction scores.

Related work

Many computational approaches have been applied to prediction of PPIs, such as threading of structural complexes (Bowie et al. 1991; Jones et al. 1992; Lathrop and

Smith 1996; Lathrop et al. 1998; Jones 1999; Panchenko et al. 2000; Xu et al. 2003a; Bienkowska and Lathrop 2005; Zhang et al. 2005); homology modeling and statistical potentials (Aloy and Russell 2002; Lu et al. 2002; Aloy and Russell 2003; Lu et al. 2003; Aytuna et al. 2005; Davis and Sali 2005; Davis et al. 2006; Pieper et al. 2006); correlated mutations (Lichtarge et al. 1996; Pazos et al. 1997; Olmea et al. 1999; Goh et al. 2000; Tan et al. 2004); and docking methods using physical force fields (Smith and Sternberg 2002; Mendez et al. 2003; Janin 2005; Summa et al. 2005). However, the performance of all of these methods is highly dependent on the accuracy of the alignment to the structural template and, for distantly related proteins, is more prone to errors. For example, the PPI predictor InterPrets (Aloy and Russell 2003) cannot find a confident match for any of the sequences from the cytokine families that we consider. Integrative machine learning methods also have been applied to prediction of PPIs and networks (Qi et al. 2005; Singh et al. 2006). Many of these approaches rely on HTP experimental PPI data itself as a predictor, and this information is scarce for ligand–receptor pairs.

Contributions

This paper proposes a novel threading algorithm, LTHREADER, which incorporates secondary structure (SS) and relative solvent accessibility (RSA) predictions with residue contact maps to guide and constrain alignments. While existing threading algorithms (e.g., RAPTOR) (Xu et al. 2003a) are not so successful at aligning interacting residues in sequences with low homology (Xu et al. 2003b), LTHREADER achieves much higher accuracy. The improvements were achieved by introducing a concept of localized threading that, which focuses on generating accurate alignment for the putative binding interface. When multiple structural complexes are available for a ligand–receptor family, our algorithm uses alignment of contact maps to generate accurate local templates for the interaction regions. Given interaction data from gold-standard low-throughput experiments, LTHREADER predicts ligand–receptor interactions using statistical scores: statistical potential, correlated mutations, and residue conservation. We demonstrate that just with the localized threading and a single complex structure the accuracy of prediction is improved. The addition of multiple complex data further increases the accuracy.

We apply our algorithm to the cytokines, performing significantly better than existing *in silico* methods. We investigate two structurally distinct cytokine families: 4-helical bundle cytokines and the TNF-like family belonging to the all- β structural class. Cytokine interactions with receptors are particularly difficult to predict because they display a high level of structural similarity but almost no sequence similarity, preventing the effective use of

Figure 1. (A) RMSD errors for the entire structure and the interface surface for various structural alignments of the 4-helical cytokine template complexes. (Left) Alignment that minimizes the RMSD for the entire complex. In this case the RMSD over all residues is 2.58 Å but the RMSD for the interface residues is even higher, at 4.09 Å. (Middle) Alignment that minimizes the RMSD for just the receptor residues. This results in a higher overall RMSD of 3.56 Å and an RMSD for the interface residues of 2.87 Å. (Right) Alignment for the same set of complexes using just the interface residues. In this case the overall RMSD is very high (4.65 Å) while the RMSD for the interface is low (1.96 Å). (B) RMSD errors for the entire structure and the interface surface for various structural alignments of the TNF-like cytokine template complexes. (Left) Alignment that minimizes the RMSD for the entire complex. (Middle) Alignment that minimizes the RMSD for just the receptor residues. (Right) Alignment for the same set of complexes using just the interface residues. The results are similar to those for the 4-helical cytokines in that the RMSD error in the interface region is high and the global RMSD error is low for the global alignment and vice versa for the interface-based alignment.

simple homology-based methods or general threading techniques. Those methods perform very well when there is significant sequence similarity among sequences that can be determined by PSI-BLAST (Altschul et al. 1997). Furthermore, experimental interaction data for cytokines is available only from low-throughput methods, and the structures for only a few cytokine–receptor complexes have been determined. Therefore, given the variability in sequence and structure, accurate prediction of cytokine interactions is a good indicator of the success we can achieve with our algorithm. Finally, our method predicts

previously undocumented cytokine interactions which may have implications for disease. We evaluate the significance of our predictions by comparing them to those of randomized interaction surfaces.

Results

LTHREADER algorithm

We have developed a localized threading algorithm that incorporates various surface complementarity scores for

the ligand–receptor interface. Our algorithm threads two given protein sequences onto a representative template complex in order to determine and score the putative interaction surface. Our interaction prediction algorithm is divided into three stages (Fig. 2). In the first stage (Fig. 2, Stage 1) using a set of template complexes, we determined the residues that are most likely to be involved with ligand–receptor binding. We did this by generating a multiple alignment of clusters of interacting residues from each complex and determining the positions that were most conserved. We built a generalized profile for each position in the alignment of interacting residues (Przybylski and Rost 2004). In the second stage (Stage 2), the profile was used to identify the most likely location of interacting residues in the query sequences. The locations of the interacting residues in the query sequences defined the putative interaction surface. In the third stage, this surface was scored using several methods and an interaction prediction is made using a decision tree classifier which integrates these scores with experimental data (Stage 3). The significance of the classification was then evaluated by estimating the probability of predicting an interaction between the ligand–receptor pair using a randomized interaction surface.

LTHREADER was able to predict ligand–receptor interactions in two of the most challenging protein families: the hematopoietins from the SCOP family long-chain 4-helical bundle and TNF-like all- β cytokines and their corresponding receptor families. When tested on the 4-helical bundles LTHREADER was able to correctly predict interactions with 75% sensitivity and 86% specificity with a 40% gain in sensitivity compared to RAPTOR. For the TNF-like cytokines LTHREADER achieved 70% sensitivity and 55% specificity with a 70% gain in sensitivity compared to RAPTOR. These cytokine families are the most challeng-

ing test cases due to their low level of sequence similarity, and unavailability of HTP PPI data.

Alignment of interacting residues

LTHREADER employs contact maps between ligand and receptor to align interface regions in complexes of proteins belonging to distantly related families. The contact maps generated from the set of template complexes representing the 4-helical bundle and TNF-like families in our data set showed local similarity in the interface region (Fig. 3). Figure 3 illustrates that, in the interaction regions defined on ligand and receptor sequences, interacting residues have similar patterns of contacts in similar complexes. Despite the low similarity of the cytokine sequences, the similarity of the contact maps is apparent. The contact maps from multiple complexes were aligned using the algorithm described in Materials and Methods.

The alignments of interacting residues generated by LTHREADER are more accurate than those by the structural alignment program POSA and the sequence alignment program MUSCLE (Table 1). We evaluated the accuracy of aligning contacts using the measure for multiple alignment of interacting residues accuracy (*IRACC*) defined by Equation 3 in Materials and Methods. Figure 4 shows the alignments that were used to compute the *IRACC* scores. The accuracy of alignments generated by LTHREADER, POSA, and MUSCLE is shown in Table 1. In comparison to POSA, the best performing algorithm, LTHREADER improves the accuracy by 14% and 4% for 4-helical and TNF-like cytokines, respectively. LTHREADER correctly aligns interacting residues across all complexes while MUSCLE and POSA generate register errors for 2 complexes and 1 complex, respectively, from the multiple alignment. Thus, the use of structural alignments generated by LTHREADER should lead to more accurate templates of the interaction interfaces.

Next we evaluated the accuracy of the alignments of sequence pairs to the template complexes using LTHREADER, a leading threading algorithm RAPTOR, and PSI-BLAST. We used RAPTOR to thread each partner of the complex independently. This choice of approach to threading a complex structure has been guided by our previous investigations with the DBLRAP program (Singh et al. 2006). Separately threading each partner of a heterodimer complex using RAPTOR gives more accurate alignments than treating the entire complex as one structural template and threading the concatenated sequences (with a linker) of both components of the heterodimer. Straightforward threading of both sequences as one chain gives worse alignments because the larger partner contribution dominates the score optimization, leading to poor alignment of the smaller protein. For PSI-BLAST alignments, novel

Figure 2. Schematic of LTHREADER. In Stage 3, CM is the compensatory mutation score; SP, the statistical potential score; FF, the force field score; and CR, the conserved residue score.

Figure 3. (A, B) Stacked contact maps for residues from complex structures of 4-helical bundle (A) and TNF-like cytokines (B) and their receptors. The aligned two-dimensional contact maps are shown in a three-dimensional space to illustrate their similarity. Only the residue positions from the interaction cores are shown. Columns correspond to residues from receptors, and rows, ligands. The shading of the contacts corresponds to the distance between residues, defined by the shortest distance between any two heavy atoms. The lighter colors reflect shorter distance. The aligned core map shows aligned positions and the colors indicate the distance averaged among aligned pairs from all five complexes.

localized profiles were computed for both the query sequences and the template complexes within the core regions and then aligned using log-average scoring (von Öhsen and Zimmer 2001) (see Profile–profile alignments section in Materials and Methods). It is important to note that this localized PSI-BLAST method produced better alignments than the global alignments commonly performed with standard PSI-BLAST. In fact, in most cases, when one complex in the training set was PSI-BLASTed against the non-redundant (NR) sequence database at NCBI (Altschul et al. 1997), none of the remaining complexes were amongst the hits. Thus, standard PSI-BLAST cannot generate an alignment for most of the cytokine complexes in our data set. We cross-threaded the sequences from known structures (see Materials and Methods) onto the other available complexes and compared the accuracy of the threading alignment. In the case of LTHREADER, the sequence profile of interacting cores was generated based on the multiple alignment of core regions from all available template complexes. For RAPTOR and PSI-BLAST localized alignments, only a profile based on the target template complex was used. Due to the high sequence similarity and low loop length variability of the 4-helical bundle receptors, the main challenge in this case was accurately aligning the ligands. The receptors can be correctly aligned using existing sequence alignment methods such as MUSCLE or PSI-BLAST. In the case of the TNF-like cytokines, aligning the receptors is the more difficult task. Below we only report the results of these more challenging alignments.

When threading the low-similarity cytokine sequences onto the templates, we achieved better results with LTHREADER than either RAPTOR or PSI-BLAST profiles despite the fact that all methods used the same structural templates and RAPTOR used the same secondary structure and relative solvent accessibility information. Table 2 shows how successful each algorithm was at identifying the locations of interacting residues. We see that even with low sequence similarity (between 15% and 25%), LTHREADER performed well at identifying interacting residues while RAPTOR struggles. This is not surprising as RAPTOR's accuracy, like most standard threaders, decreased as the sequence similarity to the template decreased (Xu et al. 2003a). We could not compare

Table 1. Comparison of alignment accuracy IRACC for various alignment methods for the 4-helical bundle and TNF-like cytokine families^a

	4-Helical bundles	TNF-like
LTHREADER	0.85	0.70
POSA	0.72	0.66
MUSCLE	0.48	0.40
Interface structure alignment	0.70	0.60
Global structure alignment	0.48	0.35

^aAs defined in Materials and Methods, Equation 3. In both cytokine families, LTHREADER achieves the highest accuracies (0.85 and 0.70). The structure-based POSA alignments and the interface alignments perform similarly and achieve the second highest level of accuracy. The sequence-based MUSCLE alignments and the structure-based global alignment perform the worst at nearly half the accuracy of LTHREADER.

Figure 4. Comparison of interface alignments generated by LTHREADER, POSA, and MUSCLE. (A) Alignments of ligands within the 4-helical bundle cytokine family. (B) Alignments of receptors within the TNF-like cytokine family. Residues involved in an interaction are highlighted in green. Misaligned interacting residues are highlighted in red.

our threading results with the MULTIPROSPECTOR (Lu et al. 2002) threader since the program was not publicly available. The individual improvements in the accuracy of alignment by LTHREADER were substantial, ranging from 6% to 32% (Table 2). For 21 out of 24 cross-threaded complexes LTHREADER significantly improved the accuracy of the alignment at the interface. In the three cases when LTHREADER did not perform as well as RAPTOR, the accuracy is lower by 1% for the EPO–EPOR complex threaded onto the GH–GHR template and lower by 2% and 4% for threading TNFSF10–TNFRSF10B onto the TNFSF13–TNFRSF13B and TNFSF13B–TNFRSF13C templates. In the few cases where LTHREADER performed worse, the decrease in accuracy is minimal and is caused by wrongly identified core boundaries. Notably localized PSI-BLAST profiles also improve the alignments over RAPTOR by an average of 10%. It is important to note that in the case of PSI-BLAST only one complex template is required for alignment. Thus, localized threading with PSI-BLAST provides an adequate approach to address cases when only one complex template is available within a ligand–receptor family.

As further evidence of the limitations of standard threaders in handling distantly related sequences, the PPI predictor InterPrets (Aloy and Russell 2003) could not even find a confident match to a complex for any of the sequences from the cytokine families.

Prediction of ligand–receptor interactions

From the alignment of multiple complexes we have identified the core interaction regions in the sequences of both ligands and receptors. For each core region in a template complex we constructed a generalized sequence profile as described in Materials and Methods. We then aligned the query sequences to the template profiles; the query residues aligned to the interacting template residues define the putative interaction surface. This stage of LTHREADER uses the putative interaction surface to calculate the surface complementarity, scores for a pair of ligand–receptor sequences and learns using the available experimental data, what distinguishes interacting from noninteracting ligand–receptor pairs. Below we compare the performance of different scoring methods and effects of the score normalizations.

LTHREADER integrating multiple statistical scores outperforms any single-scoring method in predicting ligand–receptor interactions. First, we investigated contributions of single scores and combinations of four different surface complementarity scores: statistical potentials (SP), correlated mutations (CM), conserved residues (CR), and physical force fields (FF). Each of those scores is described in more detail in Materials and Methods. In Figure 5 we show the distributions of the calculated scores for interacting and noninteracting pairs in the principal component space for both families. From the distributions one can infer that there is some, but not an exceptional, degree of clustering within true positive (interacting) or true negative (noninteracting) pairs. Among different machine learning approaches (SVMs, decision trees, regression), we chose to use a decision tree classifier to combine our stand-alone scoring methods due to the small size of our data set. The inclusion of all scores resulted in higher prediction accuracy than the individual scoring methods, even when the latter are given the same alignments of the interaction surface. In order to measure the improvement of the integrated solution over the individual scoring methods, we compared the sensitivity and specificity of each one to that of the integrated solution for both families (Table 3). The performance was determined using leave-one-out cross validation using the data sets and structural complexes described in Materials and Methods. The initial examination of the raw scores of the interaction surface revealed that for some receptors the scores were consistently high across all putative ligands (e.g., the CM score highly depends on the variability of sequences in the multiple sequence alignment [MSA]). Normalizing scores for the interaction surface, using the method described in Materials and Methods Equation 8, greatly improved the performance of the method for both the individual and the combined scores (Table 4). In summary, using

Table 2. Comparison of threading accuracy between the RAPTOR, LTHREADER, and PSI-BLAST profile alignment algorithms

Query L–R pair (4-helical cytokines)	Template complex	% ID (% Sim)	% Acc		
			RAPTOR	LTHREADER	PSI-BLAST
GH–GHR	IL6–GP130	12(23)	33	53	17
GH–GHR	EPO–EPOR	15(15)	30	55	39
GH–GHR	CSF3–CSF3R	14(20)	31	63	57
EPO–EPOR	IL6–GP130	15(24)	29	51	24
EPO–EPOR	GH–GHR	15(25)	44	43	42
EPO–EPOR	CSF3–CSF3R	14(18)	31	58	50
IL6–GP130	GH–GHR	12(22)	28	52	58
IL6–GP130	EPO–EPOR	15(17)	40	63	29
IL6–GP130	CSF3–CSF3R	17(21)	42	66	75
CSF3–CSF3R	IL6–GP130	17(24)	51	57	50
CSF3–CSF3R	EPO–EPOR	14(19)	33	61	44
CSF3–CSF3R	GH–GHR	14(25)	32	52	53
Average		15(21)	35	56	45

Query L–R pair (TNFSFx–TNFRSFy)	Template complex (TNFSFx–TNFRSFy)	% ID (% Sim)	% Acc		
			RAPTOR	LTHREADER	PSI-BLAST
13B–17	13–13B	12(23)	41	68	52
13B–17	10–SF10B	8(16)	30	55	51
13B–17	13B–13C	16(39)	60	82	71
13B–13C	13–13B	17(29)	35	65	49
13B–13C	10–10B	9(18)	25	45	47
13B–13C	13B–17	16(39)	54	66	60
10–10B	13–13B	14(29)	56	54	45
10–10B	13B–17	8(16)	32	58	34
10–10B	13B–13C	9(18)	42	38	41
13–13B	13B–17	12(23)	56	74	62
13–13B	10–10B	14(29)	40	62	65
13–13B	13B–13C	17(29)	45	84	51
Average		13(25)	43	63	52

We have threaded 4-helical (top) and TNF-like (bottom) cytokines and their receptors (identified in column 1) onto other known template complexes (identified in column 2) and determined accuracy by the percentage of positively identified interactions out of all interacting pairs in the template complex. The identifiers for the TNF-like cytokines have been abbreviated by a generic ligand–receptor identifier TNFSFx–TNFRSFy, where x and y denote a specific family member. The sequence similarity (% Sim) is measured by the number of similar residues in the alignment generated by CLUSTAL divided by the length of the alignment. The percentage of interacting residue pairs correctly identified (% Acc) is given for RAPTOR, LTHREADER, and PSI-BLAST profile alignments.

both integrated scores and normalization leads to the best performance of the classifier. While the integrated solution had comparable specificity to the single-score-based methods, it had higher sensitivity for the 4-helical bundle and TNF-like cytokines (75% and 70%, respectively).

The significance of our interaction predictions was evaluated by estimating the probability of predicting an interaction between the ligand–receptor pair using a randomized interaction surface. If interactions between a pair were predicted with high frequency for the randomized surfaces, then an interaction predicted by LTHREADER was considered to be of low significance (see Randomized interaction surfaces section in Materials and Methods). To determine how significance affects the specificity and sensitivity of the decision tree classifier, we computed ROC curves for the two families using varying significance value

cutoffs. The results are shown in Figure 6. The 4-helical bundles perform better than the TNF-like cytokines, but in both families the true-positive rate increases significantly faster than the false-positive rate.

LTHREADER integrating multiple complex data and all scores performs superior to RAPTOR and localized PSI-BLAST. Using RAPTOR and localized PSI-BLAST alignments we predicted the putative interaction interface and calculated the corresponding surface complementarity scores. We used the same set of scores (CM, SP, FF, CR) and normalization procedure as for LTHREADER to predict ligand–receptor interactions. We employed decision trees to integrate different scores and calculated the performance using leave-one-out cross validation (Table 5). We evaluated the significance of our predictions of cytokine–receptor interactions by comparing them to

Figure 5. Plots of top three principal components of the normalized interface scores for 4-helical bundle and TNF-like cytokine pairs. Green dots correspond to interface scores for interacting pairs and red dots for noninteracting pairs. (A) In the plot for the 4-helical bundles, the interacting pairs cluster fairly well with the exception of three pairs. (B) In the plot for the TNF-like cytokines, noninteracting pairs cluster together in the middle. This clustering indicates that classifiers should be able to perform fairly well on these two families.

those of randomized interaction surfaces as described in Materials and Methods. For 4-helical bundles, the predicted LTHREADER interactions had an overall significance of 0.62 and for TNF-like cytokines, 0.81, also higher than stand-alone methods (see Table 3), RAPTOR and PSI-BLAST. In comparison, predictions obtained from RAPTOR alignments had much lower sensitivity and lower significance. This is due to the fact that very few ligand–receptor pairs (or none in the case of zero sensitivity) were predicted as interacting in these cross-validated tests by RAPTOR. Both RAPTOR and localized PSI-BLAST use one template complex structure to generate alignment. However, the localized PSI-BLAST outperforms RAPTOR and is the second-best prediction method with sensitivity higher by 25% and 60% than RAPTOR for the two cytokine families. Specificity of localized PSI-BLAST is lower than LTHREADER only for the 4-helical bundle cytokines and its sensitivity is higher than single-score predictions that start from the LTHREADER alignments. This indicates that localized PSI-BLAST with multiple scores is an adequate method for ligand–receptor interaction prediction when only one complex structure is available.

We also investigated the influence of the number of available complex structures on the accuracy of LTHREADER interaction predictions. For the 4-helical family the average sensitivity (specificity) in predicting correctly the interacting pairs using leave-one-out cross validation was 75% (86%), 50% (71%), 50% (50%), 42% (64%) for 4, 3, 2, 1 complexes used for the template construction, demonstrating the importance of inclusion of diverse complexes.

Finally, we measured the impact each scoring method had on the final prediction by recomputing all of our predictions multiple times, but removing a scoring function during each iteration. The results are shown in Table 6.

Removing FF scores from our classifier has the least effect on the overall prediction accuracy while removing any of the other scores significantly reduces specificity and/or sensitivity in at least one of the cytokine families. The lack of improvement in prediction accuracy with FF included is likely a consequence of the high level of sensitivity of the AMBER force-field function to the accuracy of the alignments and of side-chain packing following threading. This result demonstrates that FF contributions should be omitted from the combined method since they are unlikely to reflect the favorable complementarity of the ligand–receptor binding interfaces.

Novel predictions

In order to apply LTHREADER to the prediction of new ligand–receptor interactions we have trained the classifier

Table 3. Comparison of single (CM, SP, FF, CR) and combined (All) scoring methods

Algorithm Cytokine family	LTHREADER				
	Scoring function				
	CM	SP	FF	CR	All
4-Helical					
Sensitivity (%)	58	67	33	50	75
Specificity (%)	93	50	100	64	86
Significance	0.4	0.32	0.55	0.45	0.62
TNF-like					
Sensitivity (%)	10	30	30	55	70
Specificity (%)	35	35	70	30	55
Significance	0.35	0.28	0.46	0.64	0.81

Scoring methods used leave-one-out cross validation on experimentally confirmed binding and nonbinding pairs of ligands and receptors. The significance of predicting interactions is equal to 1 minus the frequency of predicting interactions between pairs using randomized interaction surfaces.

Table 4. Comparison of sensitivity and specificity for raw vs. normalized scores for stand-alone and combined methods

(%)	CM		SP		FF		CR		Combined	
	Sens	Spec	Sens	Spec	Sens	Spec	Sens	Spec	Sensitivity	Specificity
<i>4-helical</i>										
Raw	33	57	33	50	17	93	42	50	33	50
Normalized	58	93	67	50	33	100	50	64	75	86
<i>TNF-like</i>										
Raw	10	25	10	20	30	45	35	25	45	25
Normalized	10	35	30	35	30	70	55	30	70	55

using the complete set of available interaction data. FF scores were omitted for these predictions since they were shown to add no value to predictions (see above). In Figure 7 we show the resulting decision trees for 4-helical and TNF-like cytokines. Applying those classifiers to the other possible interactions in the families, LTHREADER identified several new cytokine–receptor pairs as likely binding partners. The predicted interacting partners are given in Table 8. Only predicted interactions with a significance value >0.5 are shown.

Discussion

We have shown that more accurate localized threading, and integrating several existing methods for cytokine ligand–receptor interaction prediction, can greatly improve accuracy. The strength of our method comes, partially, from leveraging a novel threading algorithm that circumvents low sequence similarity. By integrating the high-quality threading results with various kinds of statistical scores and experimental data we achieved high prediction accuracy and statistical significance.

We hope to extend LTHREADER to further enhance structure-based PPI predictors (Singh et al. 2006) by refining models of the interaction regions. We plan to apply localized threading to predict other extracellular ligand–receptor families as well as general PPIs on a genome scale of the interactome. It would seem that the success of our approach depends on the availability of structural templates and orthologous sequences. As with cytokines, other therapeutically interesting extracellular ligand–receptor families often have several complex structures available. Thus, our method helps fill a void in predicting traditionally challenging, important for human diseases, ligand–receptor interactions. In the case when multiple complex structures are not available we have shown that a localized PSI-BLAST approach can improve interaction prediction. We plan to investigate combinations of these approaches to scale to whole genomes. As an initial starting point we plan to use the SCOPPI (Winter et al. 2006) database and classification of the protein–protein interfaces observed in structural complexes.

We hope to further improve the prediction accuracy of our methods by enhancing existing and developing new scoring functions that utilize randomized surfaces to better separate signal from noise. With the current accuracy of alignments generated by LTHREADER (or localized PSI-BLAST) the important contributions to the predictions come from the statistical type scores (SP, CM, CR) while the FF contributions are clearly too noisy. It is possible that for perfectly correct alignments FFs could prove beneficial. Also, due to the high computational intensity of FF calculations, it is clear that there is no justification to apply FFs on a scale of an entire interactome. Alternatively, we will investigate smoother energy functions derived from side-chain rotamer distributions that are more tolerant to small alignment errors.

We will make the program available to the broader community.

Materials and Methods

Data sets

In the 4-helical bundle family we focused on a receptor binding site (site II) that is common to all cytokines and is the major determinant of binding. The 4-helical bundle cytokine data set

Figure 6. ROC curves for the 4-helical bundle and TNF-like families of cytokines. The curves show the change in the true-positive and false-positive rates as the different significance values are used as a threshold for training the decision tree classifier.

Table 5. Performance of LTHREADER, RAPTOR, and localized PSI-BLAST integrating all surface scoring methods

Algorithm	LTHREADER	RAPTOR	LOCAL PSI-BLAST
Cytokine family			
4-Helical			
Sensitivity (%)	75	33	58
Specificity (%)	86	36	71
Significance	0.62	0.45	0.5
TNF-Like			
Sensitivity (%)	70	0	60
Specificity (%)	55	75	55
Significance	0.81	0.55	0.68

Methods used leave-one-out cross validation on experimentally confirmed binding and nonbinding pairs of ligands and receptors. The significance of predicting interactions is equal to 1 minus the frequency of predicting interactions between pairs using randomized interaction surfaces. The results for LTHREADER are the same as in Table 3 for easier comparison.

included 12 ligands and seven receptors (list is available at the supplemental material Web site, see below). Our set of template cytokine–receptor complexes consisted of the structures from the Protein Data Bank (PDB) (Berman et al. 2002), listed in Table 7. Our gold-standard positive interaction set was obtained from the KEGG database (<http://www.genome.ad.jp/kegg>). The training set consisted of 12 positive interactions derived from low-throughput experiments and 14 putative negative interactions based on membership in different subfamilies. These data are available at the supplemental material Web site (<http://theory.csail.mit.edu/lthreader>, Supplemental Table S3).

In the TNF-like family we focused on the 90's loop binding site on the receptor common to known structural complexes (Hymowitz et al. 1999). The TNF-like cytokine data set included 13 ligands and 21 receptors (available at the supplemental material Web site). Our template complexes consisted of five PDB structures listed in Table 7. The gold-standard positive and negative interaction set was taken from the results of the flow-cytometry assays reported in Bossen et al. (2006). The training set consisted of 20 positive and 20 negative interactions determined experimentally, and is also available at the supplemental material Web site.

For both families, the set of noninteracting pairs were chosen from the same ligands and receptors as those in the set of known interacting pairs to ensure that the classifier distinguishes complementarity of the interfaces rather than their composition. For each sequence we identified a set of orthologs from the available genomic databases. Since cytokines belong to families that were greatly expanded and diversified in mammalian evolution, we included the sequences from the following genomes: *Mus Musculus*, *Canis Familiaris*, *Bos Taurus*, *Rattus Norvegicus*, *Pan Troglodytes*, and *Sus Scrofa*. We initially addressed the challenge of calculating correlated mutation scores by insisting that ligands and receptors from the same family have the same set of orthologs. We thus had to omit *S. Scrofa* and *P. Troglodytes* orthologs for the 4-helical and TNF-like families, respectively. For each protein an MSA of orthologs was created using CLUSTALW (Higgins and Sharp 1988).

Algorithm

Figure 2 shows an overview of the LTHREADER algorithm and each stage is described in detail below.

Stage 1: Generation of localized profiles for interaction cores

In this stage, we assume that, if a set of ligands and receptors have similar structures and binding orientation, then their corresponding interface surfaces will have good alignment. We first examine the ligand–receptor pairs that have solved structures for their bound complex and align the ligand and receptor structures separately using POSA (Ye and Godzik 2005). Then, clusters of interacting residues are identified within these complexes and mapped to their corresponding ligand and receptor sequences, thus delimiting core regions of interaction within each sequence. Given a set (minimum two) of complexes, the positions of the cores are then optimized to ensure that the locations of the interactions contained in the clusters overlap as much as possible between complexes. Finally, generalized profiles are computed for each residue in the core regions of all pairs of ligand–receptor sequences.

Clustering of residue interactions. For two interacting domains in a complex structure we define the interface residues as those in contact with residues from the other domain. We define two residues to be in contact if the distance between any two of their heavy atoms is <4.5 Å. This cutoff is the same as that used by Lu et al. (2003) to determine statistical potentials for contacting residues.

We define a contact map as a matrix $\mathbf{C}$ such that $c_{ij} = 1$ if the i th residue of the ligand and the j th residue of the receptor interact, and $c_{ij} = 0$ if they do not. Given a contact map $\mathbf{C}$, we group together clusters of interacting pairs (non-zero entries of $\mathbf{C}$) by using a simple index-based distance function to determine inclusion. The distance between two interacting pairs $\{i_1, j_1\}$ and $\{i_2, j_2\}$ in $\mathbf{C}$, where i_1 and j_1 are the ligand and receptor indices, respectively, for the first interacting pair, and i_2 and j_2 , for the second pair, is defined as follows:

$$\text{dist}(\{i_1, j_1\}, \{i_2, j_2\}) = \frac{\sqrt{(i_1 - i_2)^2 + (j_1 - j_2)^2}}{c_{i_1, j_1} c_{i_2, j_2}} \quad (1)$$

which indicates infinite distance when any two residues do not interact. Using k-neighbor joining clustering we identify contact clusters in a contact map. We chose $k = 3$ where k is defined by the distance measure dist in Equation 1. This choice of k clusters residues that are spatial nearest neighbors on the same side of a β -strand or α -helix as these secondary structures are defined by periodicities of i , $i + 2$ and i , $i + 3$. Interacting residue pairs that are separated by a distance, dist , less than four are considered members of the same cluster. A cluster in contact map $\mathbf{C}$ implies

Table 6. Comparison of predictions done with LTHREADER while removing one score at a time (CM, SP, FF, CR)

Family	All-CM	All-SP	All-FF	All-CR	All
4-Helical					
Sensitivity (%)	58	83	75	75	75
Specificity (%)	64	71	86	71	86
TNF-like					
Sensitivity (%)	60	35	60	50	70
Specificity (%)	30	55	70	50	55

Predictions were done using leave-one-out cross validation on binding and nonbinding pairs of ligands and receptors.

Figure 7. Decision tree classifiers for 4-helical bundles and TNF-like cytokine families. Hyperplanes for 4-helical bundles: Root: CM = 0.760; L: SP = 0.777. Hyperplanes for TNF-like: Root: CR = 0.924; R: CM = 0.406, RL: CR = 0.942. “Yes” indicates that an interaction is predicted. “No” indicates that no interaction is predicted.

a corresponding sub-matrix whose non-zero entries are members of that cluster. Note that cluster edges delimit a contiguous sequence stretch on both the ligand and receptor sequences, referred to as a “core” (Fig. 8). Thus we can define a notation for indexing a cluster by the index of its corresponding cores in the ligand and receptor.

Given contact map $\mathbf{C}$, we denote $\mathbf{C}^{k,l}$ as the sub-matrix containing the cluster indexed by the k th core in the ligand and the l th core in the receptor. The size and position of $\mathbf{C}^{k,l}$ within $\mathbf{C}$ can vary as long as the requirement that only one cluster can be contained within $\mathbf{C}^{k,l}$ is not violated.

Alignment of clusters for a pair of ligand–receptor complexes. The next step of our algorithm optimizes the length and location of cores within a pair of ligand–receptor complexes so that the similarity score of corresponding clusters is maximized. Let $\mathbf{C}$ be the contact map for the first complex, and $\mathbf{D}$ be the contact map for the second complex. Let m be the number of cores in the ligands for both complexes, and n the number of cores in the receptor for both complexes. Let $\mathbf{C}^{k,l}$ refer to the k , l -th cluster in $\mathbf{C}$, and $\mathbf{D}^{k,l}$ to the corresponding k , l -th cluster in $\mathbf{D}$. We set the height (ligand axis in Fig. 8) and width (receptor axis in Fig. 8) of both sub-matrices to the maximum of the height and width of each sub-matrix. (Note that this accounts for the rare case when two clusters in one complex map to a single larger cluster in another.)

The precise alignment of the interaction cores is the goal of the following optimization procedure. For the k , l -th cluster we fix the starting position of $\mathbf{C}^{k,l}$, but allow the starting position of $\mathbf{D}^{k,l}$ to vary. Let $\mathbf{D}_{p,q}^{k,l}$ be equal to $\mathbf{D}^{k,l}$ offset by p along the first dimension of $\mathbf{D}$ and offset by q along the second dimension. Our goal then is to maximize the objective function,

$$f(p_1, \dots, p_m, q_1, \dots, q_n) = \sum_{\{k,l\}} \text{sim}(\mathbf{C}^{k,l}, \mathbf{D}_{p_k, q_l}^{k,l}), \quad (2)$$

for $1 \leq k \leq m$ and $1 \leq l \leq n$ subject to the following constraints: $-4 \leq p_1, \dots, p_m \leq 4$ and $-4 \leq q_1, \dots, q_n \leq 4$.

$\text{sim}(\mathbf{A}, \mathbf{B})$ is the measure of similarity between matrices $\mathbf{A}$ and $\mathbf{B}$ (both of dimension $m \times n$) and is defined by the sum of all entries in the Hadamard product of the two matrices: $\text{sim}(\mathbf{A}, \mathbf{B}) = \sum a_{i,j} b_{i,j}$. Since there are only a few cores in the ligand and receptor (less than five in most cases), we use a brute-force iteration over all possible values of the offset variables p, q in order to maximize f .

Multiple alignment of interaction cores. The above method allows us to find the location of cores in the ligand and receptor sequences that maximizes the overlap of interacting residues between a pair of complexes. For more than two complexes in the training set, we extend the pairwise-alignment method in a way that optimizes their multiple alignment using a variant of

the neighbor-joining method of Saitou and Nei (1987). At each step of the neighbor-joining procedure, we create a new contact matrix from the union of the Hadamard products of the contact matrices from each complex. The final result is a contact matrix representing the interaction surface common to all complexes (referred to as the average map; Fig. 3). From the multiple alignment of core regions, we construct a generalized profile from relative solvent accessibility (RSA), secondary structure (SS), and sequence at each interaction core position. RSA and SS values are calculated using DSSP (Kabsch and Sander 1983).

IRACC (interacting residues accuracy). Given a multiple alignment of N complexes $IRACC$ is defined by:

$$IRACC = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^N iracc_{ij}, \quad (3)$$

where $iracc_{ij}$ is the alignment accuracy for a pair of template complexes i, j and is defined as: $iracc_{ij} = \frac{n_{align}(i, j)}{n_{min}(i, j)}$. $n_{align}(i, j)$ is the number of aligned interacting residue positions between two complexes and $n_{min}(i, j)$ is the minimum number of interacting residue positions in complexes i and j , the maximal number of contacts that can be aligned.

Stage 2: Threading of query sequences onto the template

In this stage we determine which residues in the query sequence pair would be part of the putative interaction surface by threading their sequences onto a template complex. To do this, we devise a localized threading algorithm that aligns sequences to the generalized profile of the interaction cores. The interaction cores can be localized in the sequence relative to secondary structures such as β -strands, α -helices, or coil regions.

In order to reduce errors, we first limit the search space to the region in the query sequence most likely to contain the interaction cores by using predicted SS from SABLE (Adamczak et al. 2005). In the template structure the interaction cores are localized to specific regions on the sequence delimited by the secondary structure elements: α -helices (H), β -strands (B), and loops (L). Aligning the predicted SS elements to the template structure elements identifies the likely positions of interaction cores. Specifically the alignment of secondary structure tags, where tag = (HLHLBLB...) and a score for a

Table 7. Template complexes from the 4-helical bundle and TNF-like cytokine families

4-Helical bundle cytokine complexes		
Ligand	Receptor	PDB ID
CSF3	CSF3R	1cd9
EPO	EPOR	1cn4
GH	GHR	1hwg
IL6	GP130	1p9m
LIF	GP130	1pvh
TNF-like cytokine complexes		
Ligand	Receptor	PDB ID
TNFSF10	TNFRSF10B	1dog
TNFSF13	TNFRSF13B	1xu1
TNFSF13	TNFRSF17	1xu2
TNFSF13B	TNFRSF13C	1oqe
TNFSF13B	TNFRSF17	1oqd

Table 8. Predicted novel ligand–receptor interactions

Family	Ligand	Receptor	Significance
4-Helical cytokines	Leukemia inhibitory factor	Interleukin 12 receptor	0.65
	Leukemia inhibitory factor	Colony stimulating factor 3 receptor	0.61
	Leukemia inhibitory factor	Erythropoietin receptor	0.59
	Leukemia inhibitory factor	Prolactin receptor	0.57
	Ciliary neurotrophic factor	Erythropoietin receptor	0.51
TNF-like cytokines	OX40 antigen ligand (OX40L)	TNF receptor, member 9	0.58

Interactions are within families of 4-helical bundle and TNF-like cytokines using LTHREADER without the FF component. Significance values for each prediction are also shown.

match is 1 and a mismatch -1 with 0 gap penalty, between the template and the predicted SS determines the position of the interaction cores in the query sequence.

Second, we predict RSAs for residues in the query sequence pair, again using SABLE. Finally, the generalized profile of the core calculated in the previous stage is used to search the query sequences using the predicted RSAs and SSs (Przybylski and Rost 2004). The search is performed by sliding a window of length equal to that of the core along the query sequence. The position, p , at which the window best matches the profile defines the location of the putative core. We search for interaction cores (ICs) within five residues before and after a predicted SS element that contains the core to account for SS prediction errors. We define p_s and p_e to be the start and end position, respectively, of a predicted SS element within the query sequence. We compute p , the position of the predicted IC within the query sequence restricted to positions between $p_s - 5$ and $p_e + 5$ as follows:

$$p = \arg \max_{p \in [p_s - 5, p_e + 5]} \times \sum_{i=1}^N \left(2 \cdot \frac{1}{T} \sum_{t=1}^T \text{SEQ}(aa_{i+p}, aac'_i) + \delta(ss_{i+p}, ssc_i) - \frac{sa_{i+p} - \mu_i}{\sigma_i} \right) \quad (4)$$

where aa_{i+p} is the amino acid, ss_{i+p} is the predicted SS, and sa_{i+p} is the RSA of the residue at position $i + p$ in the query sequence. μ_i and σ_i are the mean and standard deviation, respectively, of the RSA at position i within the IC multiple alignment, and ssc_i is the SS of the core position and aac'_i is the amino acid from the template complex structure t . δ is an indicator function for equality. N is the length of the IC multiple alignment profile and T is the total number of complex structures used as templates. For the sequence similarity matrix, SEQ, we will use BLOSUM62 (Henikoff and Henikoff 1992). We have adopted the relative weights of different score contributions, sequence (SEQ) versus structure (SS and RSA), as previously determined by others (Fischer 2000; Przybylski and Rost 2004).

Profile–profile alignments. An alternative to the above method of threading the query sequence onto the template is to use PSI-BLAST to compute sequence profiles of the query sequences and the template sequences and then perform a profile–profile alignment. In our tests, we use the log-average scoring method of von Öhsen and Zimmer (2001) to score profile alignments:

$$\text{score}(\alpha, \beta) = \log \sum_{i=1}^{20} \sum_{j=1}^{20} \alpha_i \beta_j \frac{p_{rel}(i, j)}{p_i p_j} \quad (5)$$

where α and β are amino acid frequency vectors at two different profile positions, p_{rel} is the probability distribution of related amino acid pairs, and p_i is the background amino acid probability distribution. The value of $p_{rel}(i, j)/p_i p_j$ can be derived from the BLOSUM matrix and is equivalent to $2^{(BLOSUM(i, j)/2)}$. Only the sequence profiles corresponding to core regions are aligned and the search space within the query sequence is limited by using predicted SS values from SABLE as described above.

Stage 3: Scoring the interaction surface

After the interaction surface is determined for the ligand–receptor complex, it is scored and normalized as follows. Each contact from the aligned contact map calculated in Stage 1 is characterized by w_{ij} , the residue–residue distance averaged over

the set of T complexes $w_{ij} = 1/T \sum_{t=1}^T d'_{ij}$, where d'_{ij} is the Euclidean distance between pair of residues (Davis et al. 2006) in a complex t . The contact pairs in each complex map are used to calculate the total surface complementarity score as defined by:

$$S = \min_{t \in T} \sum_{\{i, j\} \in C(t)} \frac{1}{w_{ij}} S_{ij} \quad (6)$$

where $C(t)$ is the contact map of complex t defined by the interaction cores, and S_{ij} is the score of the pair. In our studies of the cytokine families we included the following measures of different properties of the putative binding interface between proteins: statistical potentials, correlated mutations, residue conservation, and force-field energies. Each is described in detail below. The putative binding interface is defined by the

Figure 8. An illustration of how ligand (red) and receptor (blue) cores are derived from a clustering of interactions within the interaction map (at right). The yellow dots correspond to interacting residues and the green dots in the interaction map indicate an interaction. A black line in the cartoon on the left denotes that an interaction occurs between the residues at its endpoints.

alignment of query sequences to complex templates generated by LTHREADER.

Statistical potentials (SP). For each residue pair located in the interaction surface, we assign a pairwise potential energy. This energy is not calculated from the physical force fields but, instead, is statistically derived from a set of known pairwise interactions between residues in solved structures. In our case, we use the pairwise potentials determined by Lu et al. (2003). To compute the SP score, we calculated the weighted sum of the potentials corresponding to all interacting residue pairs as defined by Equation 6.

Correlated mutations (CM). In order to calculate this score, we first obtain a multiple sequence alignment (MSA) for each ligand–receptor sequence S_L , S_R from a set of orthologous species common to both the ligand and receptor. Let $X_1, \dots, X_N$ be the sequences in the MSA for S_L , and $Y_1, \dots, Y_N$ be the sequences in the MSA for S_R . We then compute the Pearson correlation between positions i and j in S_L and S_R , respectively, as in Pazos et al. (1997).

$$CM_{ij} = -\frac{1}{N^2} \sum_{k=1}^N \sum_{l=1}^N \frac{(D_{ikl} - \bar{D}_i)(D_{jkl} - \bar{D}_j)}{\sigma_i \sigma_j} \quad (7)$$

Here, D_{ikl} is the similarity between the residues at position i in sequences X_k and X_l , and D_{jkl} is the similarity between the residues at position j in sequences Y_k and Y_l . $\bar{D}_i$ is the average sequence similarity at position i . $\bar{D}_i = \frac{2}{N(N-1)} \sum_{k=1}^N \sum_{l < k}^N D_{ikl}$ and σ_i is its standard deviation. We use the BLOSUM62 similarity matrix to compute D . Since we are interested in evaluating the likelihood of interaction, we only sum the correlation scores CM_{ij} over all pairs (i, j) within S_L and S_R that are within the putative interaction surface.

Conserved residues (CR). This is a sequence-based scoring method for determining whether the conservation across species of the interacting residues in the threaded complex plays a predictive role. It is based on the assumption that residues that are contained within an interaction region are less likely to mutate than those outside of the region (Caffrey et al. 2004). We compute the conservation score at each residue position within the ligand and receptor from an MSA. The conservation score at the position i in the alignment is defined by average sequence similarity $\bar{D}_i$ as above. $CR_{ij} = -\bar{D}_i \cdot \bar{D}_j$.

AMBER force fields (FF). This score is equal to the calculated energy of the putative interface surface within the threaded complex. We use the SCWRL 3.0 side-chain packing program (Canutescu et al. 2003) to first determine the coordinates for all the side-chain atoms in the ligand and receptor. Second, we fix atom positions for all residues that do not belong to the interface. Third, allowing the flexibility of interacting residues we perform 20 steps of conjugated gradient minimization using the molecular dynamics package BALL (Kohlbacher and Lenhof 2000) and the AMBER force fields (Ponder and Case 2003). The energy values typically reach a stable minimum after few steps of minimization. As the last step we compute the energy, FF, of the interface surface by applying the AMBER force-field function using BALL. The force fields are calculated only among the residues within the putative interaction surface and are not weighted by the averaged distance as are other scores. These calculations produce detailed all-atom interface models.

Normalization of scores. In order to put scores across all receptors and ligands on the same scale, we introduced the

following formula to determine new normalized values for the scores. For each pair of ligand L and receptor R from the family we have the raw score $S(L, R)$ calculated by one of the above methods $S = \{CM, SP, FF, CR\}$. The normalized scores are then given by:

$$\|S(L, R)\| = \frac{S(L, R)}{\left(\sum_{L \in \text{family}} S(L, R)\right) \cdot \left(\sum_{R \in \text{family}} S(L, R)\right)} \quad (8)$$

Classification. For classification purposes we associate with the pair L and R , a vector of scores $S_{LR} = (s_1, \dots, s_d)$ that are generated from each of the scoring methods described above (when applied to L and R). We then use experimentally determined positive and negative interactions, to train a decision tree DT. We have used the publicly available DT software OC1 (Murthy et al. 1994), which uses information gain as a cost function and the oblique mode, as opposed to axis-parallel, of partitioning the attribute space (the score space). DT is then used to classify each pair based on S_{LR} . We used decision trees because they provide a very intuitive understanding of the contributions and relative strengths of the different scoring variables used for prediction.

Randomized interaction surfaces. In order to estimate the significance of the predicted interaction for any ligand–receptor pair we have implemented the following probabilistic procedure. From all ligands and receptors within a family we create pools of ligand, $\mathbf{P}_L = \bigcup_{l \in \text{family}} r_l$, and receptor, $\mathbf{P}_R = \bigcup_{r \in \text{family}} r_r$, residues where r_l and r_r belong to the putative binding interface. For each ligand–receptor pair we generate 100 randomized interaction surfaces by grafting onto the interaction cores amino acids randomly drawn from pools $\mathbf{P}_L$ and $\mathbf{P}_R$. We then score and classify them to determine f , the frequency at which the randomized surfaces are predicted to interact. $1 - f$ is the significance of predicted interactions within the ligand–receptor family for the nonrandomized surfaces.

Acknowledgments

We thank Jinbo Xu for providing assistance with RAPTOR and suggesting the use of PSI-BLAST for localized threading. Also, thanks to Andrew Macdonnell, Rohit Singh, and Jinbo Xu for many helpful discussions and computational assistance.

References

- Adamczak, R., Porollo, A., and Meller, J. 2005. Combining prediction of secondary structure and solvent accessibility in proteins. *Proteins* **59**: 467–475.
- Aloy, P. and Russell, R.B. 2002. Interrogating protein interaction networks through structural biology. *Proc. Natl. Acad. Sci.* **99**: 5896–5901.
- Aloy, P. and Russell, R.B. 2003. InterPreTS: Protein interaction prediction through tertiary structure. *Bioinformatics* **19**: 161–162.
- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25**: 3389–3402. doi: 10.1093/nar/25.17.3389.
- Aytuna, A.S., Gursay, A., and Keskin, O. 2005. Prediction of protein–protein interactions by combining structure and sequence conservation in protein interfaces. *Bioinformatics* **21**: 2850–2855.
- Berman, H.M., Battistuz, T., Bhat, T.N., Bluhm, W.F., Bourne, P.E., Burkhardt, K., Feng, Z., Gilliland, G.L., Iype, L., Jain, S., et al. 2002. The Protein Data Bank. *Acta Crystallogr. D Biol. Crystallogr.* **58**: 899–907.
- Bienkowska, J. and Lathrop, R. 2005. Threading algorithms. In *Encyclopedia of genetics, genomics, proteomics, and bioinformatics* (eds. M. Dunn et al.), pp. 3555–3568. J. Wiley, Hoboken, NJ.

- Bossen, C., Ingold, K., Tardivel, A., Bodmer, J.L., Gaide, O., Hertig, S., Ambrose, C., Tschopp, J., and Schneider, P. 2006. Interactions of tumor necrosis factor (TNF) and TNF receptor family members in the mouse and human. *J. Biol. Chem.* **281**: 13964–13971.
- Bowie, J.U., Luthy, R., and Eisenberg, D. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253**: 164–170.
- Caffrey, D.R., Somaroo, S., Hughes, J.D., Mintseris, J., and Huang, E.S. 2004. Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci.* **13**: 190–202.
- Canutescu, A.A., Shelenkov, A.A., and Dunbrack Jr., R.L. 2003. A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci.* **12**: 2001–2014.
- Davis, F.P. and Sali, A. 2005. PIBASE: A comprehensive database of structurally defined protein interfaces. *Bioinformatics* **21**: 1901–1907.
- Davis, F.P., Braberg, H., Shen, M.Y., Pieper, U., Sali, A., and Madhusudhan, M.S. 2006. Protein complex compositions predicted by structural similarity. *Nucleic Acids Res.* **34**: 2943–2952. doi: 10.1093/nar/gkl353.
- Fischer, D. 2000. Hybrid fold recognition: Combining sequence derived properties with evolutionary information. *Pac. Symp. Biocomput.* **5**: 119–130.
- Goh, C.S., Bogan, A.A., Joachimiak, M., Walther, D., and Cohen, F.E. 2000. Co-evolution of proteins with their interaction partners. *J. Mol. Biol.* **299**: 283–293.
- Henikoff, S. and Henikoff, J.G. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.* **89**: 10915–10919.
- Higgins, D.G. and Sharp, P.M. 1988. CLUSTAL: A package for performing multiple sequence alignment on a microcomputer. *Gene* **73**: 237–244.
- Hymowitz, S.G., Christinger, H.W., Fuh, G., Ultsch, M., O'Connell, M., Kelley, R.F., Ashkenazi, A., and de Vos, A.M. 1999. Triggering cell death: The crystal structure of Apo2L/TRAIL in a complex with death receptor 5. *Mol. Cell* **4**: 563–571.
- Janin, J. 2005. Assessing predictions of protein-protein interaction: The CAPRI experiment. *Protein Sci.* **14**: 278–283.
- Jones, D.T. 1999. GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.* **287**: 797–815.
- Jones, D.T., Taylor, W.R., and Thornton, J.M. 1992. A new approach to protein fold recognition. *Nature* **358**: 86–89.
- Kabsch, W. and Sander, C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**: 2577–2637.
- Kohlbacher, O. and Lenhof, H.P. 2000. BALL—rapid software prototyping in computational molecular biology. Biochemicals algorithms library. *Bioinformatics* **16**: 815–824.
- Lathrop, R.H. and Smith, T.F. 1996. Global optimum protein threading with gapped alignment and empirical pair score functions. *J. Mol. Biol.* **255**: 641–665.
- Lathrop, R.H., Rogers Jr., R.G., Smith, T.F., and White, J.V. 1998. A Bayes-optimal sequence-structure theory that unifies protein sequence-structure recognition and alignment. *Bull. Math. Biol.* **60**: 1039–1071.
- Lichtarge, O., Bourne, H.R., and Cohen, F.E. 1996. An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**: 342–358.
- Lu, L., Lu, H., and Skolnick, J. 2002. MULTIPROSPECTOR: An algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins* **49**: 350–364.
- Lu, H., Lu, L., and Skolnick, J. 2003. Development of unified statistical potentials describing protein-protein interactions. *Biophys. J.* **84**: 1895–1901.
- Mendez, R., Leplae, R., De Maria, L., and Wodak, S.J. 2003. Assessment of blind predictions of protein-protein interactions: Current status of docking methods. *Proteins* **52**: 51–67.
- Murthy, S., Kasif, S., and Saltzberg, S. 1994. A system for induction of oblique decision trees. *J. Artif. Intell. Res.* **2**: 1–32.
- Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. 1995. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**: 536–540.
- Olmea, O., Rost, B., and Valencia, A. 1999. Effective use of sequence correlation and conservation in fold recognition. *J. Mol. Biol.* **293**: 1221–1239.
- Panchenko, A.R., Marchler-Bauer, A., and Bryant, S.H. 2000. Combination of threading potentials and sequence profiles improves fold recognition. *J. Mol. Biol.* **296**: 1319–1331.
- Pazos, F., Helmer-Citterich, M., Ausiello, G., and Valencia, A. 1997. Correlated mutations contain information about protein-protein interaction. *J. Mol. Biol.* **271**: 511–523.
- Pestka, S., Krause, C.D., and Walter, M.R. 2004. Interferons, interferon-like cytokines, and their receptors. *Immunol. Rev.* **202**: 8–32.
- Pieper, U., Eswar, N., Davis, F.P., Braberg, H., Madhusudhan, M.S., Rossi, A., Marti-Renom, M., Karchin, R., Webb, B.M., Eramian, D., et al. 2006. MODBASE: A database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res.* **34**: D291–D295. doi: 10.1093/nar/gkj059.
- Ponder, J.W. and Case, D.A. 2003. Force fields for protein simulations. *Adv. Protein Chem.* **66**: 27–85.
- Przybylski, D. and Rost, B. 2004. Improving fold recognition without folds. *J. Mol. Biol.* **341**: 255–269.
- Qi, Y., Klein-Seetharaman, J., and Bar-Joseph, Z. 2005. Random forest similarity for protein-protein interaction prediction from multiple sources. *Pac. Symp. Biocomput.* 531–542.
- Saitou, N. and Nei, M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**: 406–425.
- Singh, R., Xu, J., and Berger, B. 2006. Struct2Net: Integrating structure into protein-protein interaction prediction. *Pac. Symp. Biocomput.* **11**: 403–414.
- Smith, G.R. and Sternberg, M.J. 2002. Prediction of protein-protein interactions by docking methods. *Curr. Opin. Struct. Biol.* **12**: 28–35.
- Summa, C.M., Levitt, M., and Degrad, W.F. 2005. An atomic environment potential for use in protein structure prediction. *J. Mol. Biol.* **352**: 986–1001.
- Tan, S.H., Zhang, Z., and Ng, S.K. 2004. ADVICE: Automated detection and validation of interaction by co-evolution. *Nucleic Acids Res.* **32**: W69–W72. doi: 10.1093/nar/gkh471.
- von Ohsen, N. and Zimmer, R. 2001. Improving profile-profile alignments via log average scoring. In *Proceedings of the First International Workshop on Algorithms in Bioinformatics*. Springer-Verlag, London.
- Winter, C., Henschel, A., Kim, W.K., and Schroeder, M. 2006. SCOPPI: A structural classification of protein-protein interfaces. *Nucleic Acids Res.* **34**: D310–D314. doi: 10.1093/nar/gkj099.
- Xu, J., Li, M., Kim, D., and Xu, Y. 2003a. RAPTOR: Optimal protein threading by linear programming. *J. Bioinform. Comput. Biol.* **1**: 95–117.
- Xu, J., Li, M., Lin, G., Kim, D., and Xu, Y. 2003b. Protein threading by linear programming. *Pac. Symp. Biocomput.* **8**: 264–275.
- Ye, Y. and Godzik, A. 2005. Multiple flexible structure alignment using partial order graphs. *Bioinformatics* **21**: 2362–2369.
- Zhang, Y., Arakaki, A.K., and Skolnick, J. 2005. TASSER: An automated method for the prediction of protein tertiary structures in CASP6. *Proteins (Suppl. 7)* **61**: 91–98.