LTHREADER: PREDICTION OF LIGAND-RECEPTOR INTERACTIONS USING LOCALIZED THREADING

VINAY PULIM[1]   JADWIGA BIENKOWSKA[1,2] *   BONNIE BERGER[1] *

[1] *Computer Science and Artificial Intelligence Laboratory, MIT*
[2] *Biomedical Engineering Dept., Boston University.*
*Corresponding author

Identification of 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 alignments and integrates statistical and energy scores 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 likely interactions. The significance of predicted interactions is evaluated using the scores for randomized binding surfaces within each family. 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 representatives from different structural classes (all-alpha and all-beta 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. For the TNF-like family our method achieves 70% sensitivity with 55% specificity. This is a dramatic improvement over existing methods. Moreover, LTHREADER predicts several novel potential ligand-receptor cytokine interactions.

Supplementary website: *http://theory.csail.mit.edu/lthreader*

1. Introduction

Proteins are essential for the proper operation of living cells and viruses, performing a wide variety of functions. Most often, they do so by interacting with other proteins. The study of these interactions is extremely important, as many diseases can be traced to undesirable or malfunctioning protein-protein interactions (PPIs). Currently, methods exist for predicting PPIs that have achieved some degree of success, relying mostly on data obtained from high-throughput (HTP) experiments such as yeast-two-hybrid screens.

Receptors are proteins embedded within the cell membrane. Interactions with their extra-cellular ligands occupy a central role in inter-cellular signaling and biological processes that lead to the development and progression of many

2

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 [1].

However, ligand-receptor (L-R) interactions are much more difficult to predict than general PPIs, and methods that work well for PPIs often fail when applied to L-R binding pairs. In particular, the lack of high-throughput experimental data for these interactions makes it difficult to apply existing prediction methods that depend on this information (see Related Work).

We consider the problem of predicting whether a ligand and receptor interact, given only their sequence information and several confirmed L-R PPIs among members of the same structural SCOP family [2]. Even when one or more complex structures are available within an L-R 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.

Our approach is to thread the sequences onto the binding interface of a solved L-R 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 an L-R family; (2) threading the query sequences onto the binding interface; (3) scoring the resulting threaded sequences in order to differentiate between binding and non-binding 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 [3] and scoring them with statistical potentials [4]; correlated mutations [5-8]; and docking methods using physical force fields [9, 10]. 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 [11] 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 [12, 13]. Many of these approaches rely on HTP experimental PPI data itself as a predictor and this information is scarce for L-R pairs.

Contributions. This paper proposes a novel threading algorithm, LTHREADER, which first 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 [3]) are not so successful at aligning interacting residues in sequences with low homology [15], LTHREADER achieves much higher accuracy (see Section 3.1). Given interaction data from gold-standard low-throughput experiments, LTHREADER predicts L-R interactions using statistical and energy scores.

We apply our algorithm to the cytokines, performing significantly better than existing *in silico* methods (see section 3). We investigate two structurally distinct cytokine families: 4-helical bundle cytokines and the TNF-like family belonging to the all-beta 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 simple homology-based methods or general threading techniques. Furthermore, little experimental interaction data exists for cytokine interactions, and the structures for only a few cytokine-receptor complexes have been determined. Therefore, 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.

2. Algorithm

Overview. LTHREADER 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 (Figure 1). In the first stage (Figure 1, Stage 1), from the set (at least two) of template complexes, we determine the residues that are most likely to be involved with L-R binding. We do this by generating a multiple alignment of clusters of interacting residues from each complex and determining the positions that are most conserved. We build a generalized profile for each position in the alignment of interacting residues [16]. In the second stage (Stage 2), the profile is used to identify the most likely location of interacting residues in the query sequences. The locations of the interacting residues in the query sequences define the putative interaction surface. In the third stage, this surface is scored using several methods and an interaction prediction is made using a decision tree classifier (Stage 3). The significance of the classification is then evaluated by

4

estimating the probability of predicting an interaction between the L-R pair using a randomized interaction surface.
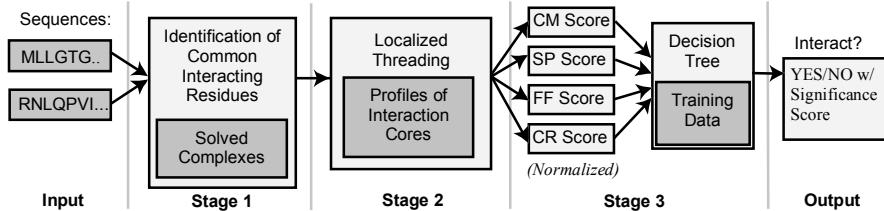
Figure 1: Schematic of LTHREADER. In Stage 3, CM is the compensatory mutation score, SP the statistical potential score, FF the force field score, and CR the conserved residue score.

2.1. *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 L-R pairs that have solved structures for their bound complex and align the ligand and receptor structures separately using POSA [17]. 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 L-R 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 less then 4.5Å. This cutoff is the same as that used by Lu *et al.* [4] to determine statistical potentials for contacting residues.

We define a contact map as a matrix $\mathbf{C}$ such that $c_{i,j} = 1$ if the ith residue of the ligand and the jth residue of the receptor interact, and $c_{i,j} = 0$ 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:

$$dist(\{i_1, j_1\}, \{i_1, j_2\}) = \frac{\sqrt{(i_1 - i_2)^2 + (j_1 - j_2)^2}}{c_{i_1,j_1} c_{i_2,j_2}} \text{, which indicates infinite distance when}$$

any two residues do not interact.

Interacting residue pairs that are separated by a distance, *dist*, less than three are considered members of the same cluster. A cluster in contact map **C** implies 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* (see Figure 2). 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 **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 **C** can vary as long as the requirement that only one cluster can be contained within $\mathbf{C}^{k,l}$ is not violated.
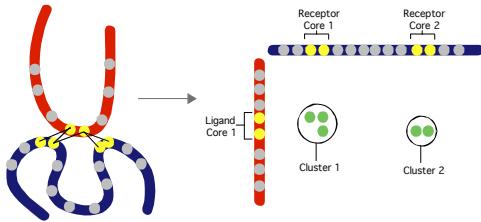
Figure 2: An illustration of how ligand (red) and receptor (blue) cores are derived from a clustering of interactions within the interaction map (at right). The yellow dots correspond to interacting residues and the green dots in the interaction map indicate an interaction. A black line in the cartoon on the left denotes that an interaction occurs between the residues at its endpoints.

Alignment of 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 L-R complexes so that the similarity score of corresponding clusters is maximized. Let **C** be the contact map for the first complex, and **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 **C**, and $\mathbf{D}^{k,l}$ to the corresponding *k,l*-th cluster in **D**. We set the height and width 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 $\mathbf{D}^{k,l}_{p,q}$ be equal to $\mathbf{D}^{k,l}$ offset by *p* along the first dimension of **D** and offset by *q* along the second dimension. Our goal then is to maximize the objective function,

$$f(p_1,...,p_m,q_1,...,q_n) = \sum_{\{k,l\}} sim(\mathbf{C}^{k,l},\mathbf{D}^{k,l}_{p_k,q_l}), \text{ for } 1 \le k \le \text{m and } 1 \le l \le \text{n}$$

6

subject to the following constraints: $-4 \leq p_1,..., p_m \leq 4$ and $-4 \leq q_1,..., q_n \leq 4$. $sim(\textbf{A}, \textbf{B})$ is the measure of similarity between matrices $\textbf{A}$ and $\textbf{B}$ (both of dimension $m \ x \ n$) and is defined by the sum of all entries in the Hadamard product of the two matrices: $sim(\textbf{A}, \textbf{B}) = \sum a_{i,j} b_{i,j}$. Since there are only a few cores in the ligand and receptor (<5 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 [18]. 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.

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 [19].

2.2. *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.

In order to reduce errors, we first limit the search space to the region in the query sequence most likely to contain the core. In the template structure, the interaction cores are localized to specific regions on the sequence delimited by the secondary structure (SS) elements: α-helices (H), β-strands (B) and loops (L). Aligning the predicted (SS) elements (using SABLE [20]) 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 match is 1 and a mismatch -1, between the template and the predicted SS determines the position of the interaction cores in the query sequence.

Next, we predict RSAs for residues in the query ligand and receptor sequences using SABLE. Finally, the generalized profile of the core calculated in the previous stage is used to search the query sequence using the predicted RSAs and SSs [16]. 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:

$$p = \arg\max_{p \in [p_s-5, p_e+5]} \sum_{i=1}^{N} \left(2 \cdot \frac{1}{T} \sum_{t=1}^{T} SEQ(aa_{i+p}, aac_i^t) + \delta(ss_{i+p}, ssc_i) - \frac{sa_{i+p} - \mu_i}{\sigma_i} \right)$$

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^t 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 use BLOSUM62 [21]. We have adopted the relative weights of different score contributions, sequence (SEQ) versus structure (SS and RSA), as previously determined by others [16].

2.3. *Stage 3: Scoring the Interaction Surface*

After the interaction surface is determined for the L-R complex, it is scored using the CM, SP, FF and CR algorithms. The scores are then normalized and integrated using a decision tree classifier. Each is described in detail below.

Correlated Mutations (CM). In order to calculate this score, we first obtain a multiple sequence alignment (MSA) for each L-R sequence S_L, S_R from a set of orthologous species common to both the ligand and receptor. We then compute the Pearson correlation between positions i and j in S_L and S_R respectively, as in [7]. Since we are interested in evaluating the likelihood of interaction, we only sum the correlation scores over all pairs (i,j) within S_L and S_R that are within the putative interaction surface (based on the threading results from stage 2). We assign this sum to the score *CM*.

Statistical Potentials (SP). For each residue pair located in the interaction surface, we assign a pairwise-potential energy from Lu *et al.* [4]. This energy is statistically derived from a set of known pairwise interactions between residues in solved structures. To compute the *SP* score, we sum the potentials corresponding to all interacting residue pairs.

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

8

side-chain packing program [22] 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 [23]. The energy values typically reach a stable minimum after a few steps of minimization. As the last step we compute the energy, *FF*, of the interface surface by applying BALL's AMBER force-field function.

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 [24]. We compute the percentage conservation at each residue position within the ligand and receptor from an MSA. The percentages are averaged over all residues within the putative interaction surface and assigned to the final *CR* score.

Normalization of Scores. The examination of the raw scores of the interaction surface showed that for some receptors the scores are consistently high across all putative ligands. In order to put scores across all receptors and ligands on the same scale, we introduce 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:

$$\|S(L,R)\| = \frac{S(L,R)}{(\sum_{L \in family} S(L,R)) \cdot (\sum_{R \in family} S(L,R))}$$

Decision Tree Classifier. For classification purposes we associate with the pair *L* and *R*, a vector of scores $\boldsymbol{S}_{LR} = (s_1,...,s_4)$ that are generated from each of the scoring methods described in Stage 3 (when applied to *L* and *R*). We then use experimentally determined positive and negative interactions, to train a decision tree DT. DT is then used to classify each pair based on $\boldsymbol{S}_{LR}$. 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.

Significance of the Classifier Predictions. In order to estimate the significance of the predicted interaction for any L-R pair we have implemented the following probabilistic procedure. From all ligands and receptors within a family we create pools of ligand, $\mathbf{P}_L = \bigcup_{l \in family} r_l$, and receptor, $\mathbf{P}_R = \bigcup_{r \in family} r_r$, residues where r_l and r_r belong to the putative binding interface as defined in section 2.2. For each L-R 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 L-R family for the non-randomized surfaces.

3. Results

We applied LTHREADER to cytokine-receptor interactions belonging to all-alpha and all-beta structural families. There are over 100 cytokines and a comparable number of corresponding receptors identified in the human genome. The interactions among cytokines and their receptors play a central role in the etiology of many human diseases and have been the focus of many investigations [1]. Cytokines are a challenging test case for our algorithm due to their low level of sequence similarity, and unavailability of high-throughput PPI data.

Datasets. We chose a subset of cytokines that contained the most solved complexes and had substantial experimental interaction data: the hematopoietins from the SCOP family *long-chain 4-helical bundle* and *TNF-like all-beta cytokines* and their corresponding receptor families.

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 includes 12 ligands and 7 receptors. Our set of template cytokine-receptor complexes consisted of the following structures from the Protein Data Bank (PDB) [25]: 1cd9, 1cn4, 1hwg, 1pvh, and 1p9m. 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.

In the TNF-like family we focused on the *90's loop* binding site on the receptor common to known structural complexes [26]. The TNF-like cytokine dataset includes 13 ligands and 21 receptors. Our template complexes consisted of the five PDB structures: 1d0g, 1oqd, 1oqe, 1xu1 1xu2. The gold standard positive and negative interaction set was taken from the results of the flow-cytometry assays reported in [27]. The training set consisted of 20 positive and 20 negative interactions determined experimentally.

For both families, the set of non-interacting pairs was chosen from the same ligands and receptors as those in the set of known interacting pairs to ensure that the classifier distinguishes complementarily of the interfaces rather than their composition. The detailed list of ligands and receptors in our datasets is available at the supplementary website.

10

3.1. *Alignment of Interacting Residues*

We applied LTHREADER (Sections 2.1 and 2.2) to the 4-helical bundle and TNF-like cytokine datasets. 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. In the case of the TNF-like cytokines, identifying the location of interaction cores in the receptors was more difficult.

When threading the low-similarity cytokine sequences onto the template, we achieved significantly better results with LTHREADER than RAPTOR despite the fact that RAPTOR uses the same structural templates and SS and RSA information. Table 1 shows how successful each algorithm was at identifying the locations of interacting residues. We see that even with low sequence similarity (between 15 to 25%), LTHREADER performs well at identifying interacting residues while RAPTOR struggles. This was not surprising as RAPTOR's accuracy, like most general threaders, decreases as the sequence similarity to the template decreases [15].

Table 1: Comparison of threading accuracy between the RAPTOR and LTHREADER algorithms. We threaded L-R pairs onto other known template complexes (for RAPTOR, L and R were threaded separately) and determined accuracy by the percentage of positively identified interactions out of all interacting pairs in the query complex.

Cytokine Family	% similarity to templates	% accuracy (RAPTOR)	% accuracy (LTHREADER)
4-Helical Ligands	21	35	56
TNF-like Receptors	35	43	63

3.2. *Prediction of Ligand-Receptor Interactions*

As expected, the combination of standalone scoring methods results in higher prediction accuracy than the individual scoring methods, *even when the latter are given correct 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 using 1-fold cross-validation (see Table 2). 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). For 4-helical bundles, the predicted interactions have a significance of 0.62 and for TNF-like cytokines, 0.81, also higher than standalone methods. Individual L-R scores are available at the supplementary website.

We verified that normalizing scores for the interaction surface greatly improved the performance of the method for both the individual and the combined scores (see supplementary website).

11

Table 2: Comparison of single and combined scoring methods using 1-fold cross validation on experimentally confirmed binding and non-binding pairs of ligands and receptors. See section 2.3 for definitions of the CM, SP, FF and CR scoring methods.

Cytokine Family	Scoring Function				
	CM	SP	FF	CR	Combined
4-Helical Bundles					
Sensitivity (%)	58	67	33	50	75
Specificity (%)	93	50	100	64	86
Significance	0.40	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

3.3. *Novel Predictions*

LTHREADER identified previously unidentified cytokine-receptor pairs as likely binding partners. These are osm-lepr, il6-ghr, epo-csf3r, cntf-lepr and lif-prlr out of the 4-helical bundle family and tnfsf1-tnfrsf11a, tnfsf1-tnfrsf11b, tnfsf4-tnfrsf6b, tnfsf4-tnfrsf12a, tnfsf10-tnfrsf1a, tnfsf10-tnfrsf1b, tnfsf13-tnfrsf6b and tnfsf15-tnfrsf1b out of the TNF-like family (see supplementary website for abbreviations).

4. Conclusions

We have shown that more accurate localized threading, and integrating several existing methods for L-R 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 and physical interaction-prediction methods we can achieve high prediction accuracy and statistical significance. However, the success of our approach depends on the availability of structural templates and orthologous sequences. This method helps fill a void in predicting traditionally challenging L-R interactions. We hope to further improve the prediction accuracy with a new scoring function that utilizes randomized surfaces to better separate signal from noise. Given the improved alignments we hope that LTHREADER will enhance structure-based PPI predictors [13] by refining the homology models of the interaction regions. We are currently applying LTHREADER to other L-R families and PPIs in general and will make the program available to the broader community.

12

References

1. Pestka, S., C.D. Krause, and M.R. Walter, *Interferons, interferon-like cytokines, and their receptors.* Immunol Rev, 2004. **202**: p. 8-32.
2. Murzin, A.G., et al., *SCOP: a structural classification of proteins database for the investigation of sequences and structures.* J Mol Biol, 1995. **247**(4): p. 536-40.
3. Xu, J., et al., *RAPTOR: optimal protein threading by linear programming.* J Bioinform Comput Biol, 2003. **1**(1): p. 95-117.
4. Lu, H., L. Lu, and J. Skolnick, *Development of unified statistical potentials describing protein-protein interactions.* Biophys J, 2003. **84**(3): p. 1895-901.
5. Goh, C.S., et al., *Co-evolution of proteins with their interaction partners.* J Mol Biol, 2000. **299**(2): p. 283-93.
6. Olmea, O., B. Rost, and A. Valencia, *Effective use of sequence correlation and conservation in fold recognition.* J Mol Biol, 1999. **293**(5): p. 1221-39.
7. Pazos, F., et al., *Correlated mutations contain information about protein-protein interaction.* J Mol Biol, 1997. **271**(4): p. 511-23.
8. Tan, S.H., Z. Zhang, and S.K. Ng, *ADVICE.* Nucleic Acids Res, 2004. **32**: p. W69-72.
9. Janin, J., *Assessing predictions of protein-protein interaction: the CAPRI experiment.* Protein Sci, 2005. **14**(2): p. 278-83.
10. Summa, C.M., M. Levitt, and W.F. Degrado, *An Atomic Environment Potential for use in Protein Structure Prediction.* J Mol Biol, 2005. **352**(4): p. 986-1001.
11. Aloy, P. and R.B. Russell, *InterPreTS: protein interaction prediction through tertiary structure.* Bioinformatics, 2003. **19**(1): p. 161-2.
12. Qi, Y., J. Klein-Seetharaman, and Z. Bar-Joseph, *Random forest similarity for protein-protein interaction prediction from multiple sources.* Pac Symp Biocomput, 2005: p. 531-42.
13. Singh, R., J. Xu, and B. Berger, *Struct2Net: Integrating StructureInto Protein-Protein Interaction Prediction.* Pac Symp Biocomput, 2006: p. 403-414.
14. Lin, N., et al., *Information assessment on predicting protein-protein interactions.* BMC Bioinformatics, 2004. **5**: p. 154.
15. Xu, J., et al., *Protein threading by linear programming.* Pac Symp Biocomput, 2003: p. 264-75.
16. Przybylski, D. and B. Rost, *Improving fold recognition ....* J Mol Biol, 2004. **341**(1): p. 255-69.
17. Ye, Y. and A. Godzik, *Multiple flexible structure alignment using partial order graphs.* Bioinformatics, 2005. **21**(10): p. 2362-9.
18. Saitou, N. and M. Nei, *The neighbor-joining method: a new method for reconstructing phylogenetic trees.* Mol Biol Evol, 1987. **4**(4): p. 406-25.
19. Kabsch, W., Sander C., *Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features.* Biopolymers, 1983. **22**: p. 2577-2637.
20. Adamczak, R., A. Porollo, and J. Meller, *Combining prediction of secondary structure and solvent accessibility in proteins.* Proteins, 2005. **59**(3): p. 467-75.
21. Henikoff, S. and J.G. Henikoff, *Performance evaluation of amino acid substitution matrices.* Proteins, 1993. **17**(1): p. 49-61.
22. Canutescu, A.A., A.A. Shelenkov, and R.L. Dunbrack, Jr., *A graph-theory algorithm for rapid protein side-chain prediction.* Protein Sci, 2003. **12**(9): p. 2001-14.
23. Kohlbacher, O. and H.P. Lenhof, *BALL.* Bioinformatics, 2000. **16**(9): p. 815-24.
24. Caffrey, D.R., et al., *Are protein-protein interfaces more conserved in sequence than the rest of the protein surface?* Protein Sci, 2004. **13**(1): p. 190-202.
25. Berman, H.M., et al., *The Protein Data Bank.* Acta Crystallogr D Biol Crystallogr, 2002. **58**(Pt 6 No 1): p. 899-907.
26. Hymowitz, S.G., et al., *Triggering cell death: the crystal structure of Apo2L/TRAIL in a complex with death receptor 5.* Mol Cell, 1999. **4**(4): p. 563-71.
27. Bossen, C., et al., *Interactions of tumor necrosis factor (TNF) and TNF receptor family members in the mouse and human.* J Biol Chem, 2006. **281**(20): p. 13964-71.