Localized Methods for Protein Interaction Prediction

by

Vinay Pulim

Submitted to the Department of Electrical Engineering and Computer Science in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

Massachusetts Institute of Technology

June 2008

Author ....
Department of Electrical Engineering and Computer Science
May 23, 2008

Certified by ....
Bonnie Berger
Professor of Applied Mathematics
Thesis Supervisor

Accepted by ......
Terry P. Orlando
Professor of Electrical Engineering
Chairman, Department Committee on Graduate Theses

Localized Methods for Protein Interaction Prediction

by

Vinay Pulim

Submitted to the Department of Electrical Engineering and Computer
Science, on May 23, 2008, in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Abstract

Identification of protein-protein interactions is important for drug design and the treatment of diseases. We propose a novel threading algorithm, LTHREADER, which generates accurate local sequence-structure alignments and integrates various statistical scores and experimental binding data to predict interactions. 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 first 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-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 and shows dramatic improvement in prediction accuracy over existing methods. To further improve alignment accuracy for all PPI families, we also introduce the program CMAPi, a two-dimensional dynamic programming algorithm that, given a pair of protein complexes, optimally aligns the contact maps of their interfaces. We demonstrate the efficacy of our algorithm on complexes from PPI families listed in the SCOPPI database and from highly divergent cytokine families. In comparison to existing techniques, CMAPi generates more accurate alignments of interacting residues within families of interacting proteins, especially for sequences with low similarity.

Thesis Supervisor: Bonnie Berger
Title: Professor of Applied Mathematics

Acknowledgements

First and foremost, I would like to express my deepest gratitude to Bonnie Berger. I consider myself extremely fortunate to have her as my advisor. Her advice and support, both academic and personal, have helped me deal some very difficult times during this decade long journey that began when I worked with her as a UROP student. Her never-ending enthusiasm and belief not just in me, but all of her students is something that I will always cherish.

I would also especially like to thank Jadwiga Bienkowska, who first inspired the research described in this thesis. I could not have asked for a more knowledgeable and enthusiastic collaborator. Her deep understanding of the field and creative problem solving ability helped us deal with many difficult challenges along the way. I also appreciate the amount of time and effort she put into this research even while working fulltime in industry.

I want to thank Jinbo Xu and Rohit Singh for many helpful discussions along the way. Their vast knowledge and experience is far superior to mine. Also, I want to thanks all the members of the computational biology group for their thoughtful feedback during group meetings.

Finally, above all, I want to thank my family for their constant love and support throughout my life. I owe everything to my parents who sacrificed so much to give me and my brother the opportunity to be where we are today. This thesis is as much for them as it is for me. And although my father is no longer here, I know he is sharing in my joy. He has been my constant source of inspiration throughout.

This thesis was supported by Grant Number 1R01GM081871-01A1 from NIH. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of NIH.

Table of Contents

1 Introduction

A genome of tens of thousands of genes can encode hundreds of thousands of unique proteins (the Proteome) that make specific interactions to form multimeric complexes and ultimately protein networks (the Interactome). These protein complexes and networks play fundamental roles in all biological processes including the maintenance of cellular integrity, metabolism, transcription/translation, and cell-cell communication. However for most proteins, in particular human proteins, there is still little knowledge of their specific binding partners. Akin to the complete sequencing of genomes, complete descriptions of interactomes is a fundamental step towards understanding biological life, and is highly relevant medically in the development of therapeutics that could manipulate specific protein-protein interactions (PPIs) in the treatment of disease. Although high-throughput biochemical approaches have been successful at systematically identifying PPIs on genome-wide scales [1-4], computational tools are still required to predict interactions due to the fact that the coverage of many high-throughput studies is still not adequate to investigate large proteomes (i.e. Mammalian proteomes), and fail to detect biologically relevant interactions such as extracellular PPIs and/or interactions involving insoluble membrane proteins.

Existing computational methods for predicting PPIs can be broadly classified into two categories: approaches that use high-throughput experimental PPI data, co-localization, co-expression and functional information; and structure-based homology modeling/threading techniques. Structure-based approaches are becoming feasible as the number of complexes rapidly grows: there has been a 40% increase in the number of complex templates in the 14 months between the two Structural Classification of Proteins database (SCOP) versions (1.65 and 1.69)[5]. In this thesis we integrate these two approaches by 1) developing a new structure-based threading method for protein-protein interfaces, and 2) using high-throughput experimental PPI data in conjunction with machine-learning techniques to predict interactions.

1.1 Computational impact

Accurate alignment of protein-protein interfaces is crucial for the success of structure-based PPI prediction. We introduce the notion of localized threading, which leads to considerably improved accuracy of interface alignments and thus provides a starting point for a generic structure-based PPI prediction approach. In addition, we introduce a high-performance algorithm for aligning contact maps (topological representations of protein-protein interfaces) using a double-dynamic programming approach. Furthermore, we use machine-learning approaches such as decision trees to integrate various methods of scoring putative protein-protein interfaces, which allows us to leverage the complementary information that is provided by statistical potential functions and co-evolution of interacting proteins.

1.2 Medical impact

For most of the ~3000 cell-membrane receptors in the human genome the identity of their ligands is unknown and there is limited structural information regarding ligand-receptor binding. One immediate application of our research is the prediction and identification of novel extracellular ligand-receptor interactions. Many of these receptors are well-characterized oncogenes (i.e. ERBB2, EGFR, FGFR1) and/or have been extensively implicated in human disease. In fact, a number of successful therapeutics have recently been developed that act to inhibit receptor activity (i.e. Herceptin, Erbitux, Etanercept, Avastin, Rituxan) demonstrating that these proteins represent ideal targets for the design of future therapeutics. Interaction between ligands and receptors cannot be studied through high-throughput (HTP) experimental approaches such as yeast-two-hybrid (Y2H). Indeed, a search of the Biomolecular Interaction Network Database (BIND) [6] for HTP protein-protein interactions between transmembrane receptors and extracellular proteins returns only five entries out of ~40,000 non-redundant interactions determined using Y2H. A promising application of this research is the rational design of therapeutics that interfere with or inhibit the binding of ligands to receptors, and thus inhibit their signaling activity.

1.3 Overview

In Chapter 2 of this thesis, we discuss the problem of protein-protein interaction prediction, the challenges involved and prior work in this field. In Chapter 3 of this thesis, we describe our localized threading algorithm in detail and predict interactions within a single family of cytokines. In Chapter 4, we show how we can generalize our algorithm to perform entire genome-scale prediction of PPIs. Finally, in Chapter 5, we draw conclusions and discuss possible improvements.

2 Protein-Protein Interaction (PPI) Prediction

2.1 Introduction

Protein-protein interaction prediction is an emerging area with vast potential to impact systems biology, genomics, molecular biology and therapeutics. Success would vastly improve data mining from genome sequencing, structural proteomics and other large-scale experiments that probe networks. It would also provide leads for experiments and drug design.

The simplest way of utilizing existing protein-protein interaction (PPI) data is based on the idea of "guilt-by-association." If two sequence domain families are known to interact, one can infer that any two proteins containing those domains are likely to interact. However such an inference is an oversimplification of the complex and specific nature of physical interaction between two proteins. For example, as a result of gene duplication during evolution an organism's genome might encode for two or more paralogous proteins that initially shared interaction partners immediately following the duplication event. Eventually, however, evolutionary forces alter the specificities for a particular interaction protein or domain to reduce the number of binding partners, or change the specificity of the partners themselves. Using such refinements organisms do not necessarily need to evolve entirely new interaction domains, but can also limit duplicated domains from cross-reacting with proteins that interact with the original domain. For example, human and other higher eukaryotic genomes contain many large families of ligands that, despite their sequence similarities, interact with only one or very few members from a large receptor family. Furthermore, each interacting ligand and receptor pair can be responsible for specific biological processes. To date, classical biochemical and genetic techniques have not proven effective in determining the exquisite specificity of most receptor-ligand interactions. In fact, the specific ligands for many receptors remain unknown even though structures of complexes from those families are available.

As two proteins cannot be assumed to interact based simply on their sequence homology to known interacting proteins, a more detailed analysis of the putative interaction surfaces is needed. When *structurally homologous* protein pairs do interact, analysis of structural complexes reveals that they interact using the same binding modes. Thus evolution, through expansion of protein families, refines protein binding specificities but conserves the three-dimensional topology of the binding interface. In this thesis, we exploit this observation [7] for evaluation of known protein-protein binding modes using various statistical and machine learning-based scoring functions.

2.2 Challenges

Homology modeling approaches to PPI prediction have had considerable success in predicting general PPIs on a genome scale [8-12], interrogating specific extracellular ligand-receptor families [14-16] and reconstructing and predicting 3D multi-protein complexes [13, 14]. However, homology modeling of the binding interface can be difficult for proteins that share very low sequence similarity but clearly have a common ancestor. The challenges in this case are two-fold. First, alignment of protein-protein interfaces from two different complex structures of distant proteins is difficult and thus constructing a template for a common binding mode is hard. Second, the alignment of protein query sequences to the structural temples is often inaccurate when sequence similarity lies within the "twilight zone" of 15-30% identity [15]. An example of binding interfaces that have similar contact maps of interacting residues but not similar sequences are interactions between long-chain cytokines and their receptors, where sequence identity across the ligands is 12-17%. This thesis addresses challenges (in both sequence-interface alignment and interface recognition) that one encounters using structural information to predict likely protein-protein interactions for sequences with low similarity to those of structural templates.

2.3 Existing knowledge base

2.3.1 Threading

A popular and highly successful approach to large-scale protein structure prediction is threading [20-27]. Protein threading predicts the three-dimensional structure for a new protein (the "query" or "target") by aligning its primary sequence to "templates" for proteins in the Protein Data Bank (PDB) to see if a similar structure can be found. The goodness of one target-template alignment is evaluated using a scoring function. The essential computational components of any threading approach are: template construction, alignment of query sequences to templates, and fold recognition.

2.3.2 Interface template databases

The analysis of binding interfaces relies on the well-established protein domain classification systems applied to all solved structures from the Protein Data Bank (PDB). Hierarchical domain classification systems, such as SCOP [16] and CATH [17], divide the set of known protein structures into families when there is clear evidence of homology typically detectable by sequence similarity. When sequence similarity is limited but evidence of homology is clear, domains are grouped into superfamilies. Proteins are grouped into a common fold category when they share only a similar fold without clear evidence of a common ancestor. Analysis of complexes of known 3D structures [18-20] revealed that there is considerable structural conservation of domain-domain binding interfaces, especially for pairs of domains, A-B and A'-B', where A & A' and B & B' belong to the same families. Significant but lower binding similarities are observed at the superfamily level but at the fold level there is no clear conservation of the binding modes.

Based on the above observations [18], several groups have recently developed databases of binary domain-domain interactions by applying both sequence-based (iPFAM [21]) and sequence-and-structure-based domain identification (SCOPPI [19, 20], PSIBASE [20, 22] and MODBASE [18, 35]) to determine interaction networks among protein domains. Interaction between two domains in the same structure file is defined by the presence of significant inter-domain interaction in the solved structures. The significance is defined by a minimum of five residue-residue contacts at 5Å (PSIBASE) and/or requirement that the buried interface has a surface of at least 600Å^2 (SCOPPI). The binary domain-domain interaction databases, such as iPFAM and PSIBASE, do not distinguish between different structural modes of binding or different types of domain-domain interfaces. The MULTIPROSPECTOR [23] template database has removed the redundancy among similar domain-domain complexes by requiring that at most one chain in the complex has <35% identity to another chain in the database. This latter approach does not take into account the possibility of different binding modes for the same domains nor does it represent conformational variation at the protein-protein interfaces. This is important as 40% of SCOPPI family pairs have at least two different binding modes [19, 20]. Keskin et al. [24] have used geometric features of a single domain binding surface to identify and cluster similar binding surfaces, and later used this clustering and conserved hot-spots of residues to predict PPIs [37, 38]. Recently Shulman-Peleg et al. [25] have developed an algorithm (MAPPIS) for multiple structure alignment of protein-protein interfaces; however, MAPPIS has not yet been applied to generation of interface templates or to prediction of PPIs by alignment of query sequences to templates.

Our research utilizes the SCOPPI clustering of protein-protein interfaces to generate our own interface template database. SCOPPI takes into consideration the case when protein domains in the same family can have different binding modes. The SCOPPI database clusters protein-protein interfaces using a binary encoding of the sequence residues that are involved in the interaction. First a multiple sequence alignment (i.e. with MUSCLE [26]) is generated for each family, and for each interface the contacting residues are mapped onto the multiple alignment. Then the interface is represented by a binary vector and all the interfaces are clustered based on their vector representations.

2.4 Predicting protein-protein interactions

Recently, several groups [14, 23, 27-30] have applied comparative modeling and statistical potentials to predict new PPIs from known structural complexes: InterPrets [28], MULTIPROSPECTOR [23, 31]. They use statistical potential functions to evaluate the homology models of domain-domain interfaces and determine which have favorable interfacial score. Both MULTIPROSPECTOR and InterPrets align the entire query sequence to the template model of the domain. MULTIPROSPECTOR uses a threading scoring function that combines both sequence and structure profiles and pairwise scoring functions to align sequence-to-structure. The common feature of these approaches is the independent alignment of query sequences to domain models for the two putative interaction partners. Our group has applied a similar strategy, using the "DouBLe RAPTOR" (DBLRAP) threading algorithm, to large scale prediction of PPIs [32]. In addition to all-encompassing statistical potential functions to evaluate any protein-protein interface, investigation of very specific protein-protein interfaces such as parallel 2-stranded coiled-coils [33] has indicated that interaction prediction can be greatly improved by careful examination of the geometry of particular residue-residue interactions and optimization of residue-pairing patterns at the binding interface. It should be noted that there do not appear to be published computational approaches geared toward the problem of solely predicting interactions between extracellular ligands and their receptors. The structure-based approach we present here first generates family-specific templates; second, both query sequences are aligned to templates; third, in addition to various standard and learned statistical potentials we use co-evolution to evaluate binding interfaces.

Traditional, structure-based approaches to predict PPIs have also been based on computationally intensive techniques such as protein docking (reviews, [34, 35]). While these methods have had some success in identifying the interaction interface between two proteins that are known to interact, they are not able in general to predict whether two proteins interact. Additionally, the detail of structural information needed for docking methods is much higher than our method requires.

2.5 Gaps to be filled

In recent years, structure-based methods to predict PPIs have received significant interest. However, there remain several challenges that limit the applicability and usefulness of such methods. First, the accuracy of these methods needs improvement. Given two proteins, many of these methods predict, by homology modeling and threading, the putative interaction surface between the two and use it to score the strength of the putative interaction. Both the former step (aligning to a template by threading) and the latter (scoring functions) need improvement. Second, the coverage of current methods is low, i.e., they work only if the input pair of proteins have high sequence similarity to some known template. While there is certainly a trade off between accuracy and coverage, we believe there is scope to improve both (over existing methods). Finally, we believe that structure-based methods to predict PPIs are most useful when used in conjunction with other functional genomic information. Most of the existing methods, as they now stand, are not appropriate for such integration. For example, given an input pair of proteins, many existing methods produce a binary output ("yes/no"). A far more useful output would be a real-valued score or probability which can be used to make fine-tuned decisions by integrating it into a machine learning classification framework. However, this raises the issue of calibration, i.e., how to determine when a score indicates an interaction and when it does not. This issue is closely tied to the choice of classification framework used. Below we describe how our research addresses some of these challenges.

In this thesis, we show that the accuracy of the alignments at the putative interaction interface is critical for proper prediction of the favorable interactions. While homology modeling/threading approaches work well when sequences are similar to their putative templates and have good overall accuracy, they give inaccurate alignments in the putative interaction regions for sequences with low similarity. Our LTHREADER program fills the shortcomings of the current alignment procedures by looking for the best local alignment in the putative interaction regions of the domain sequence. We believe the combination of both structural templates that represent conserved features of interfaces and localized threading of both query sequences onto these templates will generate models of protein-protein interfaces that are most accurate for predicting interactions. A major limitation of PPI prediction from structure is the use of only one type of information to evaluate the complementarity of the binding interface, i.e., the information about the preferential residue-residue contacts. This information is encoded in the statistical potential functions derived from the structures of known complexes. Consequently, in LTHREADER we evaluate several additional sources of information to determine the likelihood of PPIs such as correlated mutations observed at the putative interface and conservation across residue pairs. We also introduce an optimization method to learn pairwise interaction potentials.

2.6 Long-term medical significance

The individual cells of an organism constantly are barraged with hundreds, if not thousands, of extracellular signals that must be appropriately translated by intracellular signaling networks that dictate the appropriate response (i.e. increased transcription/translation, change in cell morphology, apoptosis). Interactions among extracellular ligands and their membrane-bound receptors are thus particularly important in coordinating inter and intra-cellular signals. However, for the majority of the receptors in the human genome the identity of their ligands remains unknown. One reason for the difficulty in identifying interacting ligand-receptors pairs is a large and rapid expansion of their families. Through the course of evolution many large ligand and receptor families, such as different cytokine families [36-38], have emerged in higher eukaryotic organisms. Those families are composed of sequences that share very little similarity and thus pose a challenge to traditional homology modeling.

As we have outlined in the introduction, extracellular ligand-receptor interactions are the focus of many already practiced targeted medicines. Many other therapies altering those interactions are in clinical trials [39]. The more recently discovered cytokines and receptors are also the focus of intensive bio-medical research. For example, it has recently been shown that the interaction between RANKL (a TNF-like cytokine TNSFSF11) and its receptor (TNSFRSF11A) is responsible for melanoma metastasis to bone, and inhibition of this interaction by osteoprotegerin (a soluble TNF receptor-like molecule TNSFRSF11B) prevents metastasis [40]. Thus determination of networks of interactions among families of ligands and receptors is crucial to understanding both the etiology of the disease, as well as possible therapies targeting ligand-receptor interactions. An advantage of our approach is that it may give structural information about the interactions, which can eventually be valuable in drug discovery.

Furthermore, understanding the nature of the signaling networks that act downstream of ligand-receptor interactions to regulate cellular response also remains a fundamental biological challenge. While high-throughput methods have been implemented to study PPIs on genome-wide scales, these approaches remain costly, time-consuming and have limited coverage of large proteomes. Thus we anticipate that improving computational approaches to PPI prediction by incorporating 3D-structure of interacting domains will greatly facilitate efforts to describe both intracellular as well as extracellular signaling networks. Here integration of these structure-based predictions with other data sources will be a powerful tool.

3 Single Family PPI Prediction

3.1 Introduction

Interaction of extra-cellular ligands and their receptors occupy a central role in inter-cellular 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 [36].

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 [16]. As stated in the previous chapter, even when one or more complex structures is 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 1.1 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 4.09Å and 2.75Å 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 1.96Å and 1.73Å respectively).

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 non-binding partners; and (4) evaluating the significance of the predicted interaction scores. We initially focus our efforts on a single family of cytokine ligands and receptors due to their medical significance and low sequence homology.

3.1.1 Related work

Many computational approaches have been applied to prediction of PPIs such as: threading of structural complexes [41-49]; homology modeling and statistical potentials [14, 27-31, 50, 51]; correlated mutations [15, 52-55]; and docking methods using physical force fields [34, 35, 56, 57]. 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 [28] 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 [58, 59]. Many of these approaches rely on HTP experimental PPI data itself as a predictor, and this information is scarce for ligand-receptor pairs.

3.1.2 Contributions

We describe our 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, [49] are not so successful at aligning interacting residues in sequences with low homology [60], LTHREADER achieves much higher accuracy. The improvements were achieved by introducing a concept of localized threading that 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-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. Those methods perform very well when there is a significant sequence similarity among sequences that can be determined by PSI-BLAST [61]. Furthermore, experimental interaction data for cytokines is available only from the 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.

3.2 Materials and methods

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 (Figure 1.2).

In the first stage (Figure 1.2, 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 [62]. 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.

3.2.1 Datasets

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 included 12 ligands and 7 receptors (see Appendix A). Our set of template cytokine-receptor complexes consisted of the following structures from the Protein Data Bank (PDB) [23], listed as *PDB code (ligand-receptor)*: 1cd9 (CSF3-CSF3R), 1cn4 (EPO-EPOR), 1hwg (GH-GHR), 1pvh (LIF-GP130), and 1p9m (IL6-GP130). 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 (see Appendix A).

In the TNF-like family we focused on the *90's loop* binding site on the receptor common to known structural complexes [63]. The TNF-like cytokine data set included 13 ligands and 21 receptors (see Appendix A). Our template complexes consisted of five PDB structures listed as *PDB code (ligand-receptor)*: 1d0g (TNFSF10-TNFRSF10B), 1oqd (TNFSF13B-TNFRSF17), 1oqe (TNFSF13B-TNFRSF13C), 1xu1 (TNFSF13-TNFRSF13B) and 1xu2 (TNFSF13-TNFRSF17). The gold standard positive and negative interaction set was taken from the results of the flow-cytometry assays reported in [64]. The training set consisted of 20 positive and 20 negative interactions determined experimentally (see Appendix A).

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. 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: *M.Musculus, C. Familiaris, B.Taurus, R.Norvegicus, P.Troglodytes and S.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.Trogodytes* orthologs for the 4-helical and TNF-like families respectively. For each protein a multiple sequence alignment (MSA) of orthologs was created using CLUSTALW [65].

3.2.2 Algorithm

Figure 1.2 shows an overview of the LTHREADER algorithm and each stage is described in detail below.

3.2.2.1 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 ligand-receptor pairs that have solved structures for their bound complex and align the ligand and receptor structures separately using POSA [66]. 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 less then 4.5Å. This cutoff is the same as that used by Lu *et al.* [31] to determine statistical potentials for contacting residues.

We define a contact map as a matrix **C** such that $c_{i,j}$ = 1 if the *i*th residue of the ligand and the *j*th residue of the receptor interact, and $c_{i,j}$ = 0 if they do not. Given a contact map **C**, we group together clusters of interacting pairs (non-zero entries of **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 **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:

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

(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 **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* (Figure 1.5). 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.

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 **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 (ligand axis in Figure 1.5) and width (receptor axis in Figure 1.5) 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,

$$f(p_1,\ldots,p_m,q_1,\ldots,q_n) = \sum_{\{k,l\}} sim(\mathbf{C}^{k,l},\mathbf{D}^{k,l}_{p_k,q_l}), \text{ for } 1 \le k \le \text{m and } 1 \le l \le \text{n} \qquad (2)$$

subject to the following constraints: $-4 \le p_1,\ldots,p_m \le 4$ and $-4 \le q_1,\ldots,q_n \le 4$.

sim(**A**, **B**) is the measure of similarity between matrices **A** and **B** (both of dimension $m \times n$) and is defined by the sum of all entries in the Hadamard product of the two matrices: *sim*(**A**, **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 [67]. 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; Figure 1.3). 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 [68].

IRACC (**i**nteracting **r**esidues **acc**uracy)*:* Given a multiple alignment of N complexes IRACC is defined by:

$$iracc = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} iracc_{ij} . \qquad (3)$$

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.

3.2.2.2 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. 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 [69]. 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 secondary structure (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 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 are used to search the query sequences using the predicted RSAs and SSs [62]. 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:

$$p = \arg\max_{p \in [p_s-5, p_e+5]} \sum_{i=1}^{N} \left(2 \cdot \frac{1}{T} \sum_{t=1}^{T} SEQ(aa_{i+p}, aac_i^t) + \delta(ss_{i+p}, ssc_i) - \frac{sa_{i+p} - \mu_i}{\sigma_i} \right) \quad (4)$$

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^t_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 [70]. We have adopted the relative weights of different score contributions, sequence (SEQ) versus structure (SS and RSA), as previously determined by others [62, 71].

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 Ohsen et al. [72] to score profile alignments:

$$score(\,,\,)\log = \varsigma_{i=1}^{20}\varsigma_{j=1}^{20}\;{}_{ij}\;\frac{p_{rel}(i,j)}{p_ip_j} \tag{5}$$

where ⟨ and ® 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 $p_{rel}(i,j)/p_ip_j$ 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.

3.2.2.3 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 $w_{ij} = 1/T\sum_{t=1}^{T} d^t{}_{ij}$, where d^t_{ij} is the Euclidean distance between pair of residues [73] in a complex *t*. The contact pairs in each complex map are used to calculate the total surface complementarity score as defined by:

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

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, force-field energies. Each is described in detail below. The putative binding interface is defined by the 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 pair-wise potential energy. This energy is not calculated from the physical force fields, but instead, is statistically derived from a set of known pair-wise interactions between residues in solved structures. In our case, we use the pair-wise potentials determined by Lu et al. [31]. 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 $X_1,...,X_N$ be the sequences in the MSA for S_L, and $Y_1,...,Y_N$ 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 [54].

$$CM_{ij} = -\frac{1}{N^2} \sum_{k=1}^{N} \sum_{l=1}^{N} \frac{(D_{ikl} - \overline{D_i})(D_{jkl} - \overline{D_j})}{\sigma_i \sigma_j} \tag{7}$$

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. $\overline{D_i}$ is the average sequence similarity at position *i*. $\overline{D_i} = \frac{2}{N(N-1)} \sum_{k=1}^{N} \sum_{l<k}^{N} D_{ikl}$ 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 [74]. 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 $\overline{D_i}$ as above. $CR_{ij} = -\overline{D_i \cdot D_j}$

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 [75] 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 [76] and the AMBER force-fields [77]. 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:

$$\|S(L,R)\| = \frac{S(L,R)}{(\sum_{L \in family} S(L,R)) \cdot (\sum_{R \in family} S(L,R))} \tag{8}$$

Classification. For classification purposes we associate with the pair *L* and *R*, a vector of scores $\boldsymbol{S}_{LR} = (s_1,\ldots,s_4)$ that is 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 [78]. We have used 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 $\boldsymbol{S}_{LR}$. 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.

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, $\mathbf{P}_L = \bigcup_{l \in family} r_l$, and receptor, $\mathbf{P}_R = \bigcup_{r \in family} r_r$, residues where r_l and r_r belong to the putative binding interface. 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 non-randomized surfaces.

3.3 Results

3.3.1 LTHREADER algorithm

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-beta 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 40% gain in sensitivity compared to RAPTOR. For the TNF-like cytokines LTHREADER achieved 70% sensitivity and 55% specificity with 70% gain in sensitivity compared to RAPTOR. These cytokine families are the most challenging test cases due to their low level of sequence similarity, and unavailability of high-throughput PPI data.

3.3.2 Alignment of interacting residues

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. 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 (Figure 1.3). Figure 1.3 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 the Methods. We evaluated the accuracy of aligning contacts using the *IRACC* measure defined by equation 3 in the Methods. Figure 1.4 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 and 1 complexes 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, 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 group's previous investigations with the DBLRAP program [59]. 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 localized profiles were computed for both the query sequences and the template complexes within the core regions and then aligned using log-average scoring [72] (see *Profile-profile alignments*). 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 [61], none of the remaining complexes were amongst the hits. Thus the standard PSI-BLAST cannot generate an alignment for most of the cytokine complexes in our dataset. We cross-threaded the sequences from known structures (Table 7) 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 to 25%), LTHREADER performs well at identifying interacting residues while RAPTOR struggles. This is not surprising as RAPTOR's accuracy, like most standard threaders, decreases as the sequence similarity to the template decreases [49]. We could not compare our threading results with the MULTIPROSPECTOR [51] 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 1). 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 [28] could not even find a confident match to a complex for any of the sequences from the cytokine families.

3.3.3 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 have constructed a generalized sequence profile as described in the Methods. We then have 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 non-interacting ligand-receptor pairs. Below we describe the comparisons of performances 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 have 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 the Methods. In Figure 1.6 we show the distributions of the calculated scores for interacting and non-interacting pairs in the principal components space for both families. From the distributions one can infer that there is some, but not exceptional, degree of clustering within true positive (interacting) or true negative (non-interacting) pairs. Among different machine learning approaches (SVMs, decision trees, regression), we chose to use a decision tree classifier to combine our standalone 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 described in the Methods and structural complexes listed in Table 7. 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 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* 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 1.7. 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.

The LTHREADER integrating multiple complex data and all scores has performance superior to RAPTOR and localized PSI-BLAST. Using RAPTOR and localized PSI-BLAST alignments we have predicted the putative interaction interface and calculated the corresponding surface complementarity scores. We have used the same set of scores (CM, SP, FF, CR) and normalization procedure as for LTHREADER to predict ligand-receptor interactions. We have 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 those of randomized interaction surfaces as described in **Materials and methods**. For 4-helical bundles, the predicted LTHREADER interactions had the overall significance of 0.62 and for TNF-like cytokines, 0.81, also higher than standalone 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% then RAPTOR. Specificity of localized PSI-BLAST is lower then 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 re-computing all of our predictions multiple times, but removing a scoring function during each iteration. The results are shown in Table 6. Removing force field (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 biding interfaces.

3.3.4 Novel predictions

In order to apply LTHREADER to prediction of new ligand-receptor interactions we have trained the classifier 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 1.8 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.

3.3.5 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.

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 ligand-receptor interactions that are traditionally challenging and are important for human diseases. In the case where multiple complex structures are not available, we have shown that a localized PSI-BLAST approach can improve interaction prediction. In the next chapter, we show how we can scale the LTHREADER approach to scale to whole genomes.

4 Genome-Scale PPI Prediction

4.1 Introduction

There has been limited previous work on protein-protein interface alignment. A number of different representations have been used to describe protein structure and thus protein-protein interfaces, from contact maps to all-atom representations. In the previous chapter, we described the LTHREADER program that uses a contact map representation of protein interfaces and generates an accurate alignment of binding interfaces for cytokines. When multiple structural complexes are available for a ligand-receptor family, LTHREADER performs interface alignment in two stages: First, it identifies interaction core regions by clustering contacts within a specified distance threshold; then, it aligns contact maps by maximizing the overlap between the submatrices defined by the core regions. A limitation of this method is that cores are defined before they are aligned which has the potential to generate sub-optimal contact map alignments. The other existing algorithm for interface alignment is MAPPIS, which uses an all-atom representation of protein interfaces and optimizes the alignment of interface regions with similar physico-chemical properties [79]. MAPPIS is useful for certain applications such as function prediction that require recognition of conserved structural patterns of physico-chemical interactions. However, since MAPPIS uses a physical, all-atom based representation of interfaces, it may be sensitive to small differences caused by conformational changes in the interface surface.

In this chapter, we introduce a polynomial-time algorithm for optimal pairwise contact map alignment of protein interfaces (CMAPi) using two-dimensional dynamic programming. For multiple alignment, we apply a neighbor joining algorithm akin to that used for multiple sequence alignment [65]. We evaluate our algorithm on the SCOPPI database [20], which classifies all protein-protein interfaces into similarity classes, and measure its performance according to the percentage of interacting residues aligned correctly (see Methods.) We demonstrate CMAPi produces more accurate alignments than existing methods such as MAPPIS and MUSCLE [79, 80], especially for protein sequences with low similarity. Compared to LTHREADER, CMAPi is faster, automated and as accurate, allowing large-scale application. Moreover, our new approach aligns entire contact maps without having to first identify core regions. Instead, cores are automatically determined by the algorithm as a post-alignment step and then used to generate sequence profiles of the interaction cores. By applying CMAPi to all of the known PPI families in SCOPPI and generating the corresponding profiles (as described previously for LTHREADER), we are able to predict many more interactions.

4.2 Materials and methods

4.2.1 Algorithm

CMAPi finds alignments of similar protein-protein interfaces using a contact map representation. First, we generate optimal pairwise interface alignments and then use a version of the neighbor-joining algorithm to align multiple interfaces.

The contact map representation used in CMAPi contains more information than just the binary values present in the LTHREADER contact maps. A CMAPi contact map is a two-dimensional matrix X indexed by the residues $i \in L$ and $j \in R$ from the interacting proteins L and R. Entry $X_{i,j}$ in contact map X is defined as

$$X_{i,j} = \min_{h_i \in i, h_j \in j}\left(d_{h_i h_j}\right), \tag{1}$$

which is the minimum distance between all heavy atoms, h_i and h_j, of residues i and j. In our contact maps we include all the residues that have at least one contact with the minimum distance less then 10Å. This initial cutoff distance is much more generous then the strict 4.5Å threshold used for defining contacts in single complexes [31, 81]. The more generous initial cutoff allows for alignment of contacts that may pass the strict cutoff in one complex but not others.

Given two contact maps, C and D, our goal is to find the alignment of C and D that maximizes the overlap between interacting residues. Our alignment algorithm uses two-dimensional dynamic programming to optimize the alignment score [82]. We allow for gaps in the maps by assigning a gap penalty that penalizes gap insertions between highly interacting residues. The justification for this penalty is that adjacent residues that are highly interactive should be part of the same interaction core and therefore should not be split.

The first step in the dynamic programming approach is to create a four-dimensional scoring matrix M, where $M_{i,j,k,m}$ is the maximum score at position i,j,k,m ($0 \leq i < width(C)$, $0 \leq j < height(C)$, $0 \leq k < width(D)$, $0 \leq m < height(D)$). Entry $M_{i,j,k,m}$ is then determined from previously solved sub-problems as follows:

$$M_{i,jk,m} = \begin{cases} 0 & \mathit{if}\ ijkm = 0 \\ Max\begin{pmatrix} M_{i-1,jk-1,m} + S(i,j,k,m), \\ M_{i,j-1k,m-1} + S(i,j,k,m), \\ M_{i,jk-1,m} + w_c(C,i), \\ M_{i,jk,m-1} + w_r(C,j), \\ M_{i-1,jk,m} + w_c(D,k), \\ M_{i,j-1k,m} + w_r(D,m) \end{pmatrix} & \mathit{otherwise} \end{cases} \quad (2)$$

where $w_c(X,i)$ is the gap penalty for inserting a gap at column i in contact map X and $w_r(X,j)$ is the gap penalty for inserting a gap at row j in contact map X. In order to ensure that clusters of interacting residues are not split, we assign a high penalty for gap insertions in rows and columns containing a high number of interactions. Specifically, we used the following gap penalty functions:

$$w_c(X,i) = -\sum_j \frac{1}{X_{i,j}^2}, \qquad (3)$$

$$w_r(X,j) = -\sum_i \frac{1}{X_{i,j}^2}$$

$S(i,j,k,m)$ is the similarity score between the interaction at i, j in contact map C and interaction k, m in contact map D. We use the following similarity function:

$$S(i,j,k,m) = \frac{1}{C_{i,j}D_{k,m}}, \qquad (4)$$

Although here we used a simple similarity function based on inter-residue distance within an interaction, one can define a more complex similarity function that incorporates physical and chemical properties of the interacting residues. We note that both the scoring function and gap penalty functions are defined in the same units of inverse square of the distance.

Once all values of M are computed using (2), the optimal alignment of contact maps is determined by backtracking through the scoring matrix as in standard dynamic programming. Movements within the scoring matrix correspond to the following alignment actions:

Change in (*i,j,k,m*)	Contact map *C*	Contact map *D*
(+1, 0, +1, 0)	align column *i*	align column *k*
(0, +1, 0, +1)	align row *j*	align row *m*
(0, 0, +1, 0)	gap at column *i*	
(0, 0, 0, +1)	gap at row *j*	
(+1, 0, 0, 0)		gap at column *k*
(0, +1, 0, 0)		gap at row *m*

An optimal alignment of contact maps is a mapping $A(i,j)=(A(i), A(j))$ of the pair of (i,j) indices from a complex *C* onto the (k, m) pair in a complex *D* where $(k,m)=(A(i),A(j))$.

Multiple alignment of contact maps is accomplished using the same neighbor-joining algorithm as in CLUSTALW but with similarity of contact maps as a distance metric:

$$d_{CD} = \frac{1}{ij} \sum_{i,j,k,m} S(i,j,k=A(i),m=A(j)) \qquad (5)$$

The final step of our algorithm identifies core regions within each of the interface sequences. We consider two consecutive residues in one sequence to be part of the same core if they both interact with the same residue in the second sequence. For a given SCOPPI family consisting of complexes of proteins $\{L\}$ and $\{R\}$, let *i* and *j* denote the aligned positions among all contact maps between the $\{L\}$ and $\{R\}$ proteins. The residue positions *i* and *i*+1 from a set of aligned 'ligand' sequences $\{L\}$ belong to the same interaction core if for any residue *j* from the 'receptor' sequences $\{R\}$, the contact map distance is less then 4.5Å for any complex in the family. A similar definition is applied to define interaction cores in $\{R\}$ sequences. Thus the interaction cores consist of contiguous stretches of aligned residues within the $\{L\}$ and $\{R\}$ protein sequences.

4.2.2 Performance analysis

Since the CMAPi algorithm explores the entire space of possible alignments between contact maps and optimizes the contact map similarity function (4) we can claim that it is optimal for pairwise alignment. In fact, by concatenating the rows in a contact map and creating a one-dimensional sequence of contacts, the CMAPi algorithm can be mapped to a specific case of one-dimensional sequence alignment with a complex, position-dependent gap penalty structure as defined by (3). In the case of multiple alignment of contact maps, although the neighbor-joining method is not optimal, it has been shown in practice to perform nearly as well as an optimal, exhaustive search for multiple sequence alignment [83]. In terms of computational complexity, while single-chain contact map alignment (introduced by [84]), has been shown to be NP-hard by [85], PPI interface alignment is tractable because gap insertions in the two interacting protein sequences defining the contact maps are independent. Furthermore, multiple alignment is also tractable since we are using the polynomial-time neighbor-joining algorithm. Specifically, pairwise alignment is of order $O(k^4)$ where k is the number of interface contacts in a protein complex and multiple alignment is of order $O(k^4m^2)$ given m contact maps. In the SCOPPI database, we found that on average $k \cong 83$ and $m \cong 5$. See Figure 2.2 for the distributions of k and m.

4.3 Results

4.3.1 CMAPi Algorithm

To evaluate the accuracy of our dynamic programming algorithm, we compared our results to those produced by MAPPIS [79] and LTHREADER [81]. We also compared our algorithm to purely sequence based alignments generated by MUSCLE [80] as a baseline for our test. MUSCLE was chosen from among many different sequence alignment algorithms for development of the SCOPPI database. Alignments were scored using the same IRACC function defined in the previous chapter:

$$iracc = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} iracc_{ij} \tag{6}$$

where $iracc_{ij}$ is the alignment accuracy for a pair of template complexes *i,j* and is defined as:

$$iracc_{ij} = \frac{n_{align}(i,j)}{n_{\min}(i,j)} \tag{7}$$

CMAPi generates the most accurate interface alignments for cytokine families, one of the more challenging cases previously investigated by LTHREADER. In this case, since LTHREADER generates different core boundaries than our algorithm, we measured the accuracy of CMAPi and MAPPIS using LTHREADER's core definitions. CMAPi has accuracy almost identical to LTHREADER while not requiring pre-defined interaction cores (see Table 9).

In addition, it shows an improvement over MAPPIS by 5% for 4-helical bundles and 6% for TNF-like cytokines and much higher accuracy than MUSCLE (over 11% for 4-helical bundles and 10% for TNF-like cytokines). The use of structural information by both MAPPIS and CMAPi leads to significantly better alignments when compared to MUSCLE, which only uses sequence information.

We also investigated alignment accuracy as a function of sequence similarity and demonstrated CMAPi's superior performance as sequence similarity declines (see Figure 2.3). To evaluate the influence of sequence similarity on the performance of different algorithms (MUSCLE, MAPPIS CMAPi), we aligned complexes from PPI families listed in the SCOPPI database. Only families containing at least three complexes were chosen to ensure that enough structure information was available to generate alignments. The current release of SCOPPI contains 219 such families. Results from this evaluation are shown in Figure 2.3.

While both methods generate significantly better alignments than MUSCLE at all sequence similarities, CMAPi performs better than both MAPPIS and MUSCLE when sequence similarity is below 75%. Moreover, CMAPi improves alignments over MAPPIS by about 5% for structures that are typically considered for homology modeling, where sequence similarity is 50-70%. This result indicates that CMAPi may also be useful as a first step in building the detailed homology models of protein interfaces from multiple complex structures.

4.3.2 Discussion

We have shown that the alignment accuracy of our CMAPi algorithm is higher than other existing interface alignment algorithms and in particular MAPPIS. Our algorithm is optimal for pairwise alignment of contact maps and near-optimal in practice for multiple alignment, while having polynomial-time complexity. We believe our method generates better alignments of interacting residues due to its use of a contact map representation of protein interfaces instead of the all-atom based representation used by MAPPIS. The all-atom representation is helpful in situations where the fine details of the structure can be predicted with high confidence, such as homology modeling of very similar proteins. However, in the case when fine details cannot be predicted accurately, representations using coarser features, such as contact maps, lead to better predictions. CMAPi is tolerant to conformational changes and thus aligns more of the interaction surface.

5 Future Work

We hope to further improve the prediction accuracy of LTHREADER 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 high computational intensity of FF calculations it is clear that there is no justification to apply FF 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.

In addition, we intend to use the CMAPi alignment algorithm to build profiles for every family of interacting proteins defined in the publicly available SCOPPI database. For each family, we will use the multiple alignment of contact maps corresponding to each PPI complex within the family and generate aligned core regions within each sequence pair. The aligned cores will then be used to derive sequence profiles that will be used for PPI interaction prediction as described in LTHREADER. The improvements in the alignment of interacting residues for sequences with 50-70% similarity indicate that CMAPi could also be helpful in building better homology models of protein-protein interfaces when multiple complexes are available as templates. In this work, we have used a pre-existing classification of protein-protein binding modes provided by SCOPPI. In the future we will investigate if CMAPi can be used to classify protein binding modes based on clustering of similar contact maps.

Tables

	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

Table 1: Comparison of alignment accuracy IRACC (as defined in Methods eq.3) for various alignment methods for the 4-helical bundle and TNF-like cytokine families. 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.

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**

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

Algorithm	**LTHREADER**				
	Scoring Function				
Cytokine Family	**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

Table 3: Comparison of single (CM, SP, FF, CR) and combined (All) scoring methods using leave-one-out cross validation on experimentally confirmed binding and non-binding 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.

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

Table 4: Comparison of sensitivity and specificity for raw vs. normalized scores for standalone and combined 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

Table 5: Performance of LTHREADER, RAPTOR and localized PSI-BLAST integrating all surface scoring methods using leave-one-out cross validation on experimentally confirmed binding and non-binding 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.

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

Table 6: Comparison of predictions done with LTHREADER while removing one score at a time (CM, SP, FF, CR). Predictions were done using leave-one-out cross validation on binding and non-binding pairs of ligands and receptors.

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

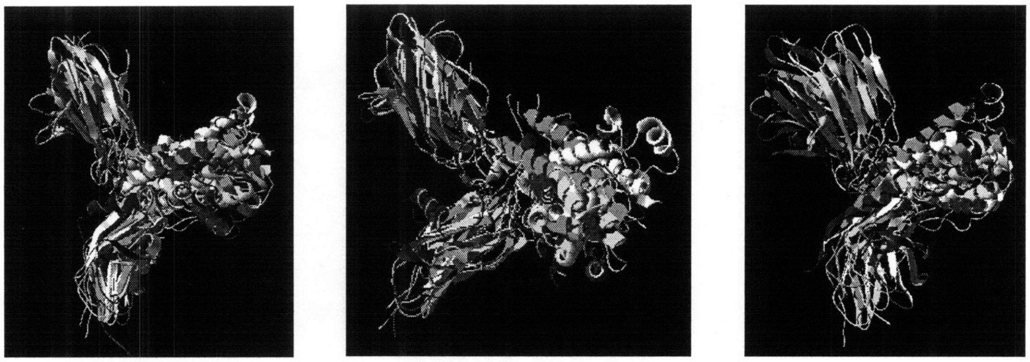
Family	Ligand	Receptor	Significance
4-helical cytokines	Leukemia inhibitory factor	Interleukin 12 receptor	0.65
	Leukemia inhibitory factor	Colony stimulating factor 3 receptor	0.61
	Leukemia inhibitory factor	Erythropoetin receptor	0.59
	Leukemia inhibitory factor	Prolactin receptor	0.57
	Ciliary neurotrophic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

Table 8. Predicted novel ligand-receptor interactions within families of 4-helical bundle and TNF-like cytokines using LTHREADER without the FF component. Significance values for each prediction are also shown.

	4-Helical Bundles	TNF-Like
CMAPi	0.84	0.72
LTHREADER	0.85	0.70
MAPPIS	0.80	0.64
MUSCLE	0.73	0.62

Table 9. Comparison of alignment accuracy, IRACC, for various alignment methods for the 4-helical bundle and TNF-like cytokine families. In both cytokine families, our CMAPi algorithm achieves higher alignment accuracy than MAPPIS and MUSCLE and the same accuracy as LTHREADER.

Figures

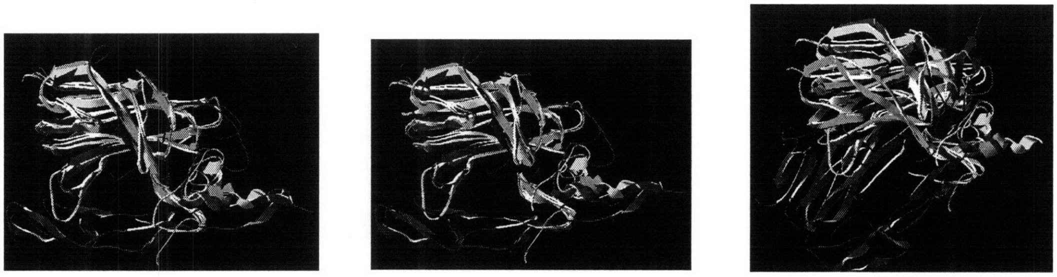


Global Alignment
Global RMSD: 2.58 Å
Interface RMSD: 4.09 Å

Receptor Alignment
Global RMSD: 3.56 Å
Interface RMSD: 2.87 Å

Interface Alignment
Global RMSD: 4.65 Å
Interface RMSD: 1.96 Å

(a)



Global Alignment
Global RMSD: 1.29 Å
Interface RMSD: 2.75 Å

Receptor Alignment
Global RMSD: 1.59 Å
Interface RMSD: 2.70 Å

Interface Alignment
Global RMSD: 3.92 Å
Interface RMSD: 1.73 Å

(b)

Figure 1.1a: Figure 1.1a (left) shows the alignment that minimizes the RMSD for the entire complex. In this case the RMSD over all residues is 2.58Å but the RMSD for the interface residues is even higher at 4.09Å. Figure 1.1a (middle) shows the alignment that minimizes the RMSD for just the receptor residues. This results in a higher overall RMSD of 3.56Å and an RMSD for the interface residues of 2.87Å. Figure 1.1a (right) shows the alignment for the same set of complexes using just the interface residues. In this case the overall RMSD is very high (4.65Å) while the RMSD for the interface is low (1.96Å). Clearly, the "localized alignment" based on just the interface residues (Figure 1.1a, right) 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.

Figure 1.1b: RMSD errors for the entire structure and the interface surface for various structural alignments of the TNF-like cytokine template complexes.

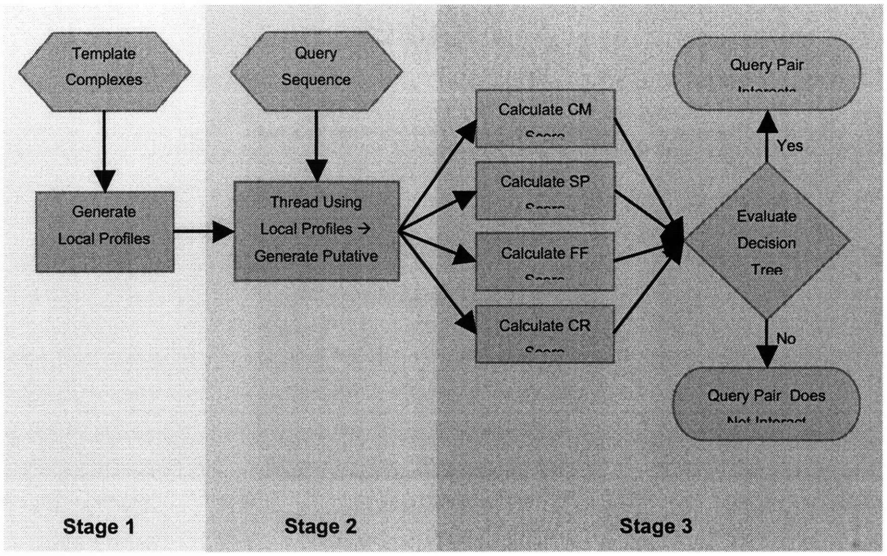
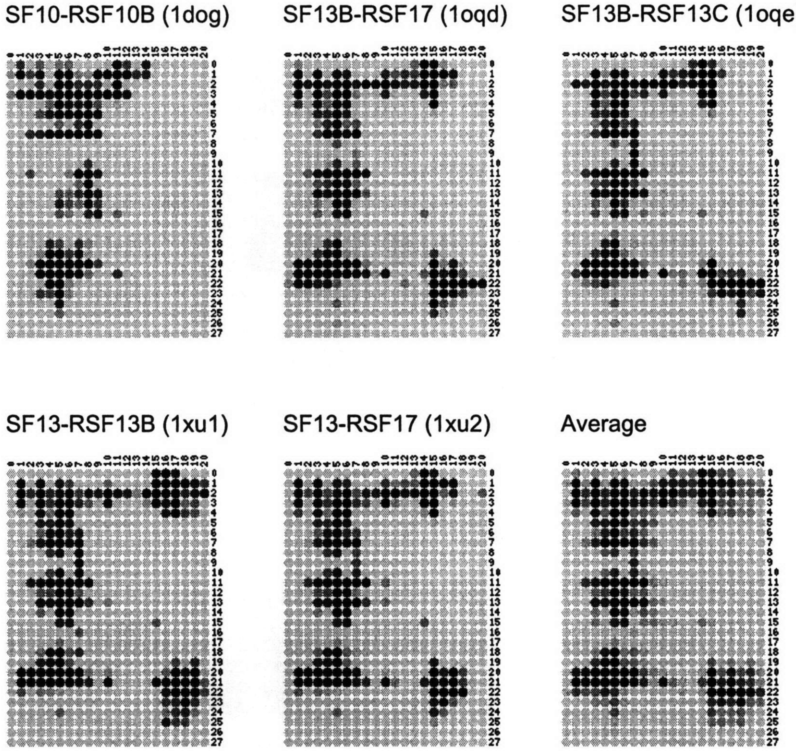


Figure 1.2: Schematic of LTHREADER. In Stage 3, CM is the compensatory mutation score, SP the statistical potential score, FF the force field score, and CR the conserved residue score.



SF10-RSF10B (1dog)

SF13B-RSF17 (1oqd)

SF13B-RSF13C (1oqe)

SF13-RSF13B (1xu1)

SF13-RSF17 (1xu2)

Average

Figures 1.3: Contact maps for residues from complex structures of 4-helical bundle (Figure 1.3, top) and TNF-like cytokines (Figure 1.3, bottom) and their receptors. Only the residue positions from the interaction cores are shown. Columns correspond to residues from receptors and rows, ligands. Horizontal lines de-lineate the interaction core is in the receptor and vertical, in the ligand. The shading of the contacts corresponds to the distance between residues, defined by the shortest distance between any two heavy atoms. The darker 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.

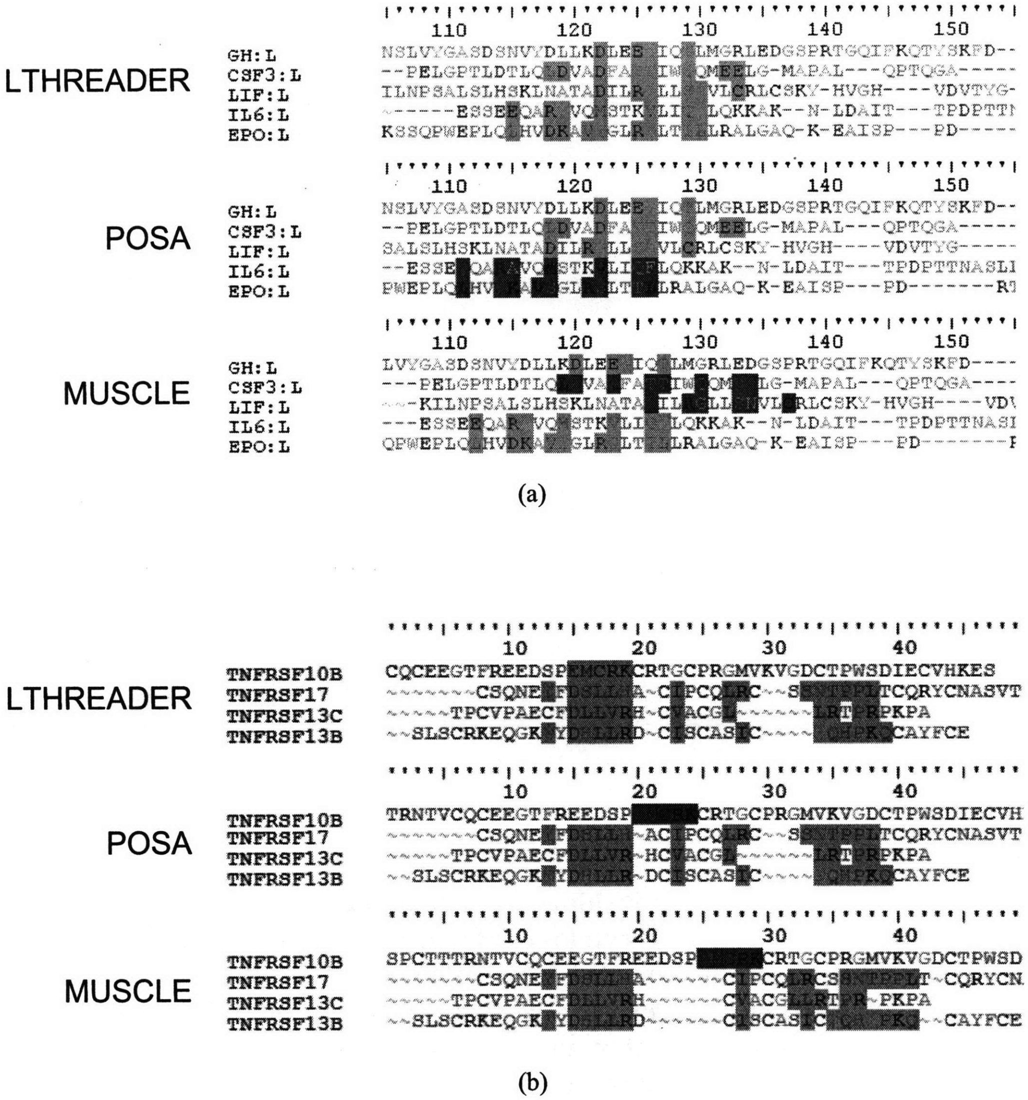


Figure 1.4: Comparison of interface alignments generated by LTHREADER, POSA and MUSCLE. Figure 1.4a shows the alignments of ligands within the 4-helical bundle cytokine family. Figure 1.4b shows the alignments of receptors within the TNF-like cytokine family. Residues involved in an interaction are highlighted in green. Misaligned interacting residues are highlighted in red.

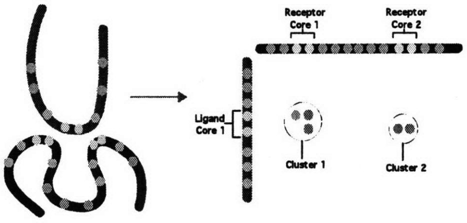


Figure 1.5: An illustration of how ligand (red) and receptor (blue) cores are derived from a clustering of interactions within the interaction map (at right). The yellow dots correspond to interacting residues and the green dots in the interaction map indicate an interaction. A black line in the cartoon on the left denotes that an interaction occurs between the residues at its endpoints.

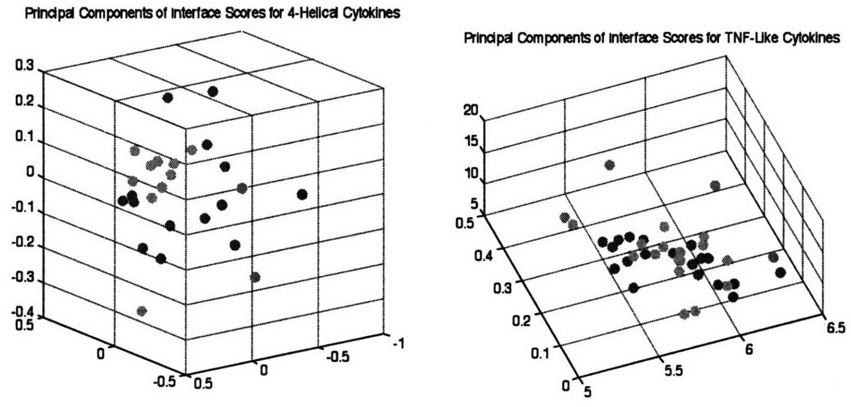


Figure 1.6: Plots of top three principal components of the normalized interface scores for 4-helical bundle and TNF-like cytokine pairs. Green dots correspond to interface scores for interacting pairs and red dots for non-interacting pairs. In the plot for the 4-helical bundles, the interacting pairs cluster fairly well with the exception of three pairs. In the plot for the TNF-like cytokines, non-interacting pairs cluster together in the middle. This clustering indicates that classifiers should be able to perform fairly well on these two families.

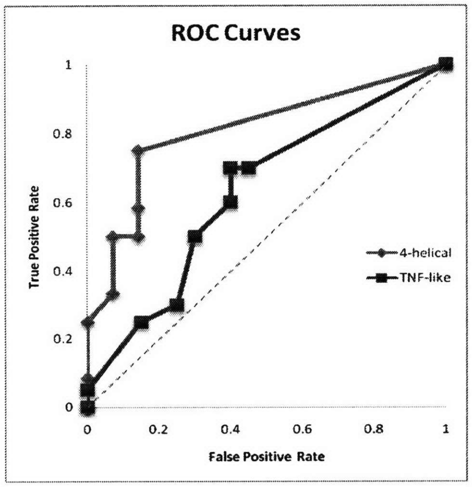
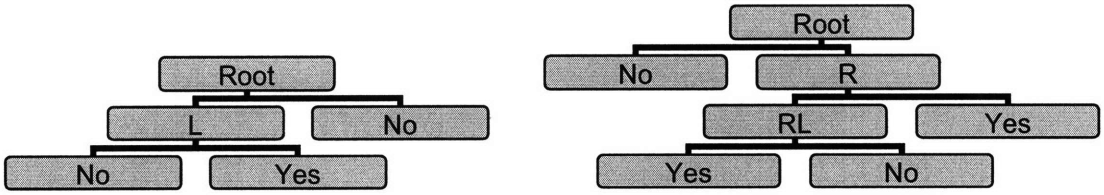


Figure 1.7: ROC curves for the 4-helical bundle and TNF-like families of cytokines. The curves show the change in the true positive and false positive rates as the different significance values are used as a threshold for training the decision-tree classifier.



4-helical bundles

TNF-like

Figure 1.8: 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.

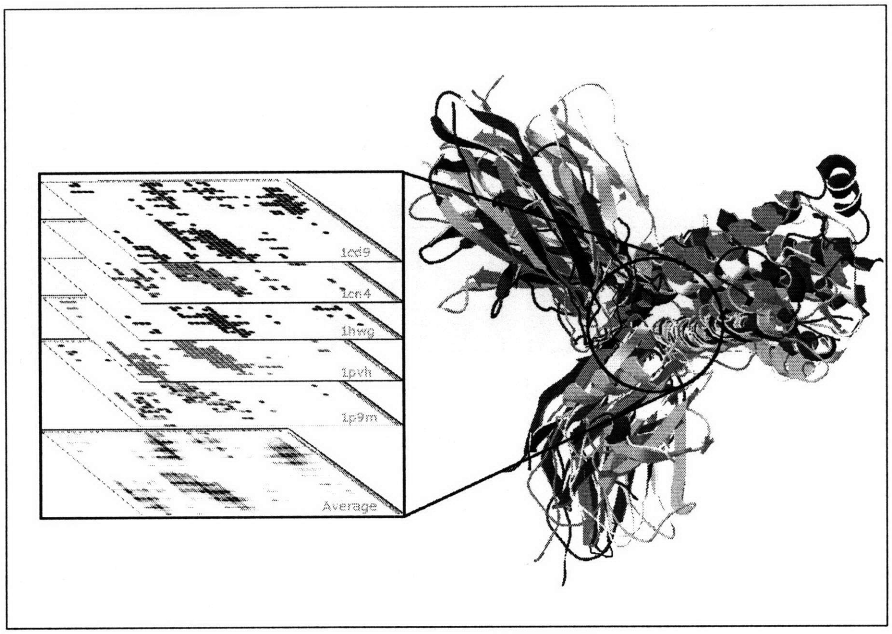


Fig. 2.1. Contact map representation of the aligned interfaces of five 4-helical bundle cytokine complexes. Each color represents a complex and the gray map on the left is the average contact map of the aligned interfaces.

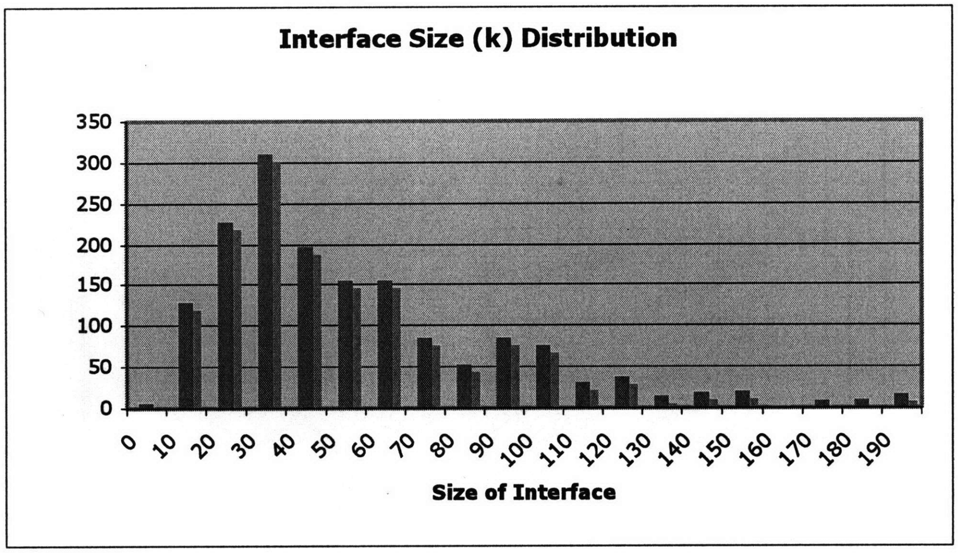


Fig. 2.2a. Distribution of *k* (interface size). The average interface size is 83 contacts.

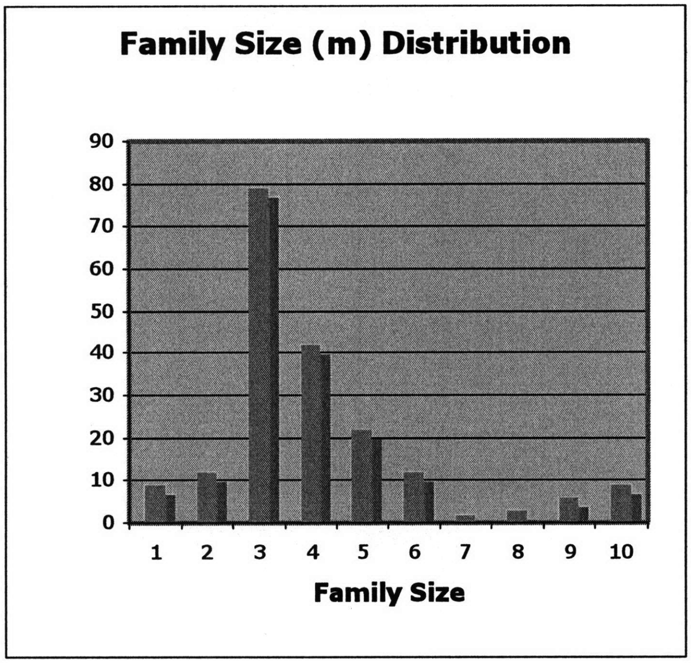


Fig. 2.2b. Distribution of *m* (family size). The average family size is 4.7 complexes.

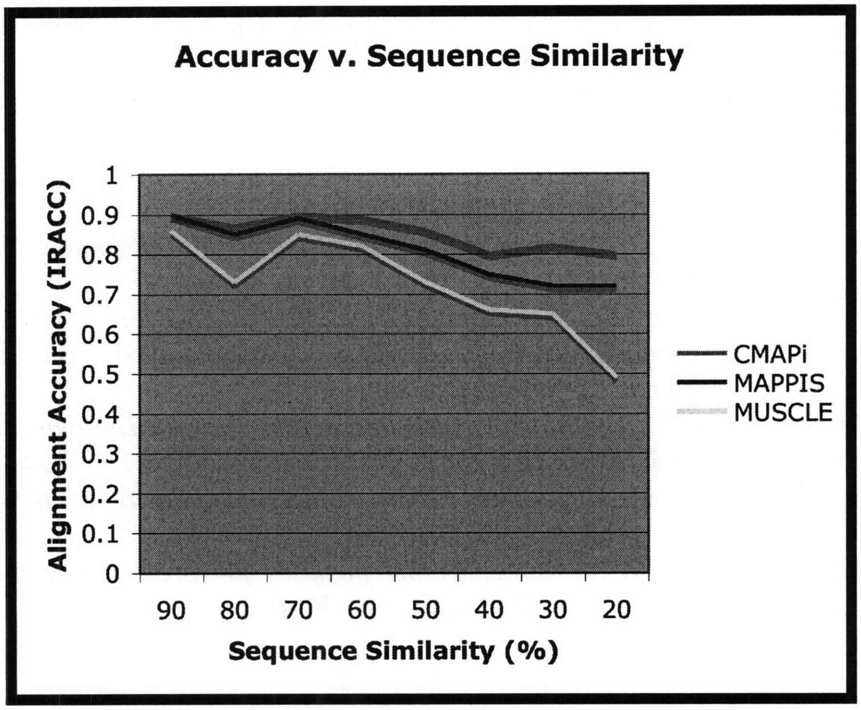


Fig. 2.3. Comparison of alignment accuracy (IRACC) versus the similarity of sequences within complexes using the CMAPi, MAPPIS and MUSCLE algorithms.

References

1. Lin, N., et al., *Information assessment on predicting protein-protein interactions.* BMC Bioinformatics, 2004. **5**: p. 154.

2. Rual, J.F., et al., *Towards a proteome-scale map of the human protein-protein interaction network.* Nature, 2005. **437**(7062): p. 1173-8.

3. Uetz, P., et al., *A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae.* Nature, 2000. **403**(6770): p. 623-7.

4. Giot, L., et al., *A protein interaction map of Drosophila melanogaster.* Science, 2003. **302**(5651): p. 1727-36.

5. Kyu Kim, W., et al., *The Many Faces of Protein-Protein Interactions: A Compendium of Interface Geometry.* PLoS Computational Biology, 2006. **preprint**(2006): p. e124.eor.

6. Alfarano, C., et al., *The Biomolecular Interaction Network Database and related tools 2005 update.* Nucleic Acids Res, 2005. **33**(Database issue): p. D418-24.

7. Gavin, A.C., et al., *Proteome survey reveals modularity of the yeast cell machinery.* Nature, 2006. **440**(7084): p. 631-6.

8. Ben-Hur, A. and W.S. Noble, *Kernel methods for predicting protein-protein interactions.* Bioinformatics, 2005. **21 Suppl 1**: p. i38-46.

9. Deng, M., et al., *Inferring domain-domain interactions from protein-protein interactions.* Genome Res, 2002. **12**(10): p. 1540-8.

10. Tress, M., et al., *Scoring docking models with evolutionary information.* Proteins, 2005. **60**(2): p. 275-80.

11. Wang, H., et al., *Identifying protein-protein interaction sites on a genome-wide scale*, in *Adnavces in Neural Information Processing Systems 17*, L.K. Saull, Y. Weiss, and L. Bottou, Editors. 2005, MIT press: Cambridge, MA. p. 1465-1472.

12. Bock, J.R. and D.A. Gough, *Whole-proteome interaction mining.* Bioinformatics, 2003. **19**(1): p. 125-34.

13. Aloy, P., et al., *Structure-based assembly of protein complexes in yeast.* Science, 2004. **303**(5666): p. 2026-9.

14. Davis, F.P., et al., *Protein complex compositions predicted by structural similarity.* Nucleic Acids Res, 2006. **34**(10): p. 2943-52.

15. 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.

16. 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.

17. Orengo, C.A., F.M. Pearl, and J.M. Thornton, *The CATH domain structure database.* Methods Biochem Anal, 2003. **44**: p. 249-71.

18. Aloy, P., et al., *The relationship between sequence and interaction divergence in proteins.* J Mol Biol, 2003. **332**(5): p. 989-98.

19. Kim, W.K., et al., *The many faces of protein-protein interactions: A compendium of interface geometry.* PLoS Comput Biol, 2006. **2**(9): p. e124.

20. Winter, C., et al., *SCOPPI: a structural classification of protein-protein interfaces.* Nucleic Acids Res, 2006. **34**(Database issue): p. D310-4.

21. Finn, R.D., M. Marshall, and A. Bateman, *iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions.* Bioinformatics, 2005. **21**(3): p. 410-2.

22. Gong, S., et al., *PSIbase: a database of Protein Structural Interactome map (PSIMAP).* Bioinformatics, 2005. **21**(10): p. 2541-3.

23. Berman, H.M., et al., *The Protein Data Bank.* Acta Crystallogr D Biol Crystallogr, 2002. **58**(Pt 6 No 1): p. 899-907.

24. Keskin, O., et al., *A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications.* Protein Sci, 2004. **13**(4): p. 1043-55.

25. Shulman-Peleg, A., et al., *MAPPIS: Multiple 3D Alignment of Protein-Protein Interfaces.* CompLife 2005, 2005: p. 91-103.

26. Barrett, T., et al., *NCBI GEO: mining millions of expression profiles--database and tools.* Nucleic Acids Res, 2005. **33**(Database issue): p. D562-6.

27. Aloy, P. and R.B. Russell, *Interrogating protein interaction networks through structural biology.* Proc Natl Acad Sci U S A, 2002. **99**(9): p. 5896-901.

28. Aloy, P. and R.B. Russell, *InterPreTS: protein interaction prediction through tertiary structure.* Bioinformatics, 2003. **19**(1): p. 161-2.

29. Aytuna, A.S., A. Gursoy, and O. Keskin, *Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces.* Bioinformatics, 2005. **21**(12): p. 2850-5.

30. Davis, F.P. and A. Sali, *PIBASE: a comprehensive database of structurally defined protein interfaces.* Bioinformatics, 2005. **21**(9): p. 1901-7.

31. Lu, H., L. Lu, and J. Skolnick, *Development of unified statistical potentials describing protein-protein interactions.* Biophys J, 2003. **84**(3): p. 1895-901.

32. Singh, R., J. Xu, and B. Berger, *Struct2Net: Integrating structure into protein-protein interaction prediction.* Pac Symp Biocomput, 2006: p. 403-414.

33. Singh, M., B. Berger, and P.S. Kim, *LearnCoil-VMF: computational evidence for coiled-coil-like motifs in many viral membrane-fusion proteins.* J Mol Biol, 1999. **290**(5): p. 1031-41.

34. Mendez, R., et al., *Assessment of blind predictions of protein-protein interactions: current status of docking methods.* Proteins, 2003. **52**(1): p. 51-67.

35. Smith, G.R. and M.J. Sternberg, *Prediction of protein-protein interactions by docking methods.* Curr Opin Struct Biol, 2002. **12**(1): p. 28-35.

36. Pestka, S., C.D. Krause, and M.R. Walter, *Interferons, interferon-like cytokines, and their receptors.* Immunol Rev, 2004. **202**: p. 8-32.

37. Tracey, K.J. and A. Cerami, *Tumor necrosis factor, other cytokines and disease.* Annu Rev Cell Biol, 1993. **9**: p. 317-43.

38. Whicher, J.T. and S.W. Evans, *Cytokines in disease.* Clin Chem, 1990. **36**(7): p. 1269-81.

39. Walsh, G., *Biopharmaceutical benchmarks 2006.* Nat Biotechnol, 2006. **24**(7): p. 769-76.

40. Jones, D.H., et al., *Regulation of cancer cell migration and bone metastasis by RANKL.* Nature, 2006. **440**(7084): p. 692-6.

41. Bowie, J.U., R. Luthy, and D. Eisenberg, *A method to identify protein sequences that fold into a known three-dimensional structure.* Science, 1991. **253**(5016): p. 164-70.

42. Jones, D.T., W.R. Taylor, and J.M. Thornton, *A new approach to protein fold recognition.* Nature, 1992. **358**(6381): p. 86-9.

43. Lathrop, R.H., et al., *A Bayes-optimal sequence-structure theory that unifies protein sequence-structure recognition and alignment.* Bull Math Biol, 1998. **60**(6): p. 1039-71.

44. Lathrop, R.H. and T.F. Smith, *Global optimum protein threading with gapped alignment and empirical pair score functions.* J Mol Biol, 1996. **255**(4): p. 641-65.

45. Bienkowska, J. and R. Lathrop, *Threading Algorithms*, in *Encyclopedia of genetics, genomics, proteomics, and bioinformatics*, M. Dunn, Jorde, L., Little, P., Subramaniam, Editor. 2005, Wiley.

46. Jones, D.T., *GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences.* J Mol Biol, 1999. **287**(4): p. 797-815.

47. Panchenko, A.R., A. Marchler-Bauer, and S.H. Bryant, *Combination of threading potentials and sequence profiles improves fold recognition.* J Mol Biol, 2000. **296**(5): p. 1319-31.

48. Zhang, Y., A.K. Arakaki, and J. Skolnick, *TASSER: an automated method for the prediction of protein tertiary structures in CASP6.* Proteins, 2005. **61 Suppl 7**: p. 91-8.

49. Xu, J., et al., *RAPTOR: optimal protein threading by linear programming.* J Bioinform Comput Biol, 2003. **1**(1): p. 95-117.

50. Pieper, U., et al., *MODBASE: a database of annotated comparative protein structure models and associated resources.* Nucleic Acids Res, 2006. **34**(Database issue): p. D291-5.

51. Lu, L., H. Lu, and J. Skolnick, *MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading.* Proteins, 2002. **49**(3): p. 350-64.

52. Goh, C.S., et al., *Co-evolution of proteins with their interaction partners.* J Mol Biol, 2000. **299**(2): p. 283-93.

53. Lichtarge, O., H.R. Bourne, and F.E. Cohen, *An evolutionary trace method defines binding surfaces common to protein families.* J Mol Biol, 1996. **257**(2): p. 342-58.

54. Pazos, F., et al., *Correlated mutations contain information about protein-protein interaction.* J Mol Biol, 1997. **271**(4): p. 511-23.

55. Tan, S.H., Z. Zhang, and S.K. Ng, *ADVICE: Automated Detection and Validation of Interaction by Co-Evolution.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W69-72.

56. Janin, J., *Assessing predictions of protein-protein interaction: the CAPRI experiment.* Protein Sci, 2005. **14**(2): p. 278-83.

57. 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.

58. 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.

59. Singh, R., J. Xu, and B. Berger, *Struct2Net: Integrating StructureInto Protein-Protein Interaction Prediction.* Pac Symp Biocomput, 2006: p. 403-414.

60. Xu, J., et al., *Protein threading by linear programming.* Pac Symp Biocomput, 2003: p. 264-75.

61. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.

62. Przybylski, D. and B. Rost, *Improving fold recognition without folds.* J Mol Biol, 2004. **341**(1): p. 255-69.

63. 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.

64. 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.

65. Higgins, D.G. and P.M. Sharp, *CLUSTAL: a package for performing multiple sequence alignment on a microcomputer.* Gene, 1988. **73**(1): p. 237-44.

66. Ye, Y. and A. Godzik, *Multiple flexible structure alignment using partial order graphs.* Bioinformatics, 2005. **21**(10): p. 2362-9.

67. 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.

68. Kabsch, W., Sander C., *Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features.* Biopolymers, 1983. **22**: p. 2577-2637.

69. Adamczak, R., A. Porollo, and J. Meller, *Combining prediction of secondary structure and solvent accessibility in proteins.* Proteins, 2005. **59**(3): p. 467-75.

70. Henikoff, S. and J.G. Henikoff, *Amino acid substitution matrices from protein blocks.* Proc Natl Acad Sci U S A, 1992. **89**(22): p. 10915-9.

71. Fischer, D., *Hybrid fold recognition: combining sequence derived properties with evolutionary information.* Pac Symp Biocomput, 2000: p. 119-30.

72. Niklas, v., Ohsen, and Z. Ralf, *Improving Profile-Profile Alignments via Log Average Scoring*, in *Proceedings of the First International Workshop on Algorithms in Bioinformatics*. 2001, Springer-Verlag.

73. Segal, E., et al., *Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data.* Nat Genet, 2003. **34**(2): p. 166-76.

74. 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.

75. 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.

76. Kohlbacher, O. and H.P. Lenhof, *BALL--rapid software prototyping in computational molecular biology. Biochemicals Algorithms Library.* Bioinformatics, 2000. **16**(9): p. 815-24.

77. Ponder, J.W. and D.A. Case, *Force fields for protein simulations.* Adv Protein Chem, 2003. **66**: p. 27-85.

78. Murthy, S., S. Kasif, and S. Saltzberg, *A System for Induction of Oblique Decision Trees.* Journal of Arificial Intelligence Research, 1994. **2**(1): p. 1-33.

79. Shulman-Peleg, A. and M. Shatsky, *MAPPIS: Multiple 3D Alignment of Protein-Protein Interfaces.* Computational Life Sciences: First International Symposium, CompLife 2005, Konstanz, Germany, September 25-27, 2005: Proceedings, 2005.

80. Edgar, R.C., *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucleic Acids Res, 2004. **32**(5): p. 1792-7.

81. Pulim, V., J. Bienkowska, and B. Berger, *LTHREADER: prediction of extracellular ligand-receptor interactions in cytokines using localized threading.* Protein Sci, 2008. **17**(2): p. 279-92.

82. Smith, T.F. and M.S. Waterman, *Identification of common molecular subsequences.* J Mol Biol, 1981. **147**(1): p. 195-7.

83. Gascuel, O. and M. Steel, *Neighbor-joining revealed.* Mol Biol Evol, 2006. **23**(11): p. 1997-2000.

84. Godzik, A., A. Kolinski, and J. Skolnick, *Topology fingerprint approach to the inverse protein folding problem.* J Mol Biol, 1992. **227**(1): p. 227-38.

85. Goldman, D., S. Istrail, and C.H. Papadimitriou, *Algorithmic aspects of protein structure similarity.* Foundations of Computer Science, 1999. 40th Annual Symposium on, 1999: p. 512-521.

Appendix A

4-Helical Bundle Cytokines	
Abbreviation	**Name**
Ligands	
EPO	Erythropoietin
CSF3	Colony stimulating factor 3
GH	Growth hormone
IL6	Interleukin 6
LIF	Leukemia inhibitory factor
OSM	Oncostatin M
CNTF	Ciliary neurotrophic factor
IL23	Interleukin 23
IL12	Interleukin 12
LEP	Leptin
PRL	Prolactin
Receptors	
EPOR	Erythropoietin receptor
CSF3R	Colony stimulating factor 3 receptor
GHR	Growth hormone receptor
GP130	GP130
IL12R	Interleukin 12 receptor
LEPR	Leptin receptor
PRLR	Prolactin receptor

Dataset: Long-chain 4-helical bundles

TNF-Like Cytokines	
Abbreviation	**Name**
Ligands	
TNFSF1	Lymphotoxin alpha (LTA)
TNFSF2	Tumor necrosis factor (TNF)
TNFSF3	Lymphotoxin beta (LTB)
TNFSF4	OX40 antigen ligand (OX40L)
TNFSF5	CD40 antigen ligand (CD40L)
TNFSF6	Fas antigen ligand (FASL)
TNFSF10	TNF-related apoptosis inducing ligand (TRAIL)
TNFSF11	Receptor activator of NF-kappa-B ligand (RANKL)
TNFSF12	TNF-related weak inducer of apoptosis (TWEAK)
TNFSF13	APOL-related leukocyte expressed ligand 2 (APRIL)
TNFSF13B	B-cell activating factor (BAFF)
TNFSF15	TNF ligand-related molecule 1 (TL1A)
Receptors	
TNFRSF1A	Tumor necrosis factor receptor 1 (TNFR1)
TNFRSF1B	Tumor necrosis factor receptor 2 (TNFR2)
TNFRSF4	OX40 antigen (OX40)
TNFRSF5	CD40 antigen (CD40)
TNFRSF6	Fas antigen (FAS)
TNFRSF6B	Decoy receptor 3 (DcR3)
TNFRSF10B	TNF-related apoptosis-inducing ligand receptor 2 (TRAILR2)
TNFRSF11A	Receptor activator of NF-kappaB (RANK)
TNFRSF11B	Osteoprotegerin (OPG)
TNFRSF12	DR3
TNFRSF12A	Type I transmembrane protein Fn14
TNFRSF13B	Transmembrane activator and CAML interactor (TACI)
TNFRSF13C	B-cell activating factor receptor (BAFFR)
TNFRSF17	B-cell maturation factor (BCMA)
LTBR	Lymphotoxin B receptor

Dataset: TNF-like cytokine ligands and receptors

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

Dataset: Template complexes

Comparison of single and combined scoring methods using 1-fold cross validation on experimentally confirmed binding and non-binding pairs of ligands and receptors. YES indicates the method predicted the pair binds, NO that the pair does not bind. Green shading indicates that the prediction agrees with experimental data; red that it does not.

4-Helical Bundle Cytokines							
Ligand	**Receptor**	**Binding?**					
Binding Pairs		Exp	CM	SP	FF	CR	All
EPO	EPOR	YES	NO	YES	YES	NO	YES
CSF3	CSF3R	YES	NO	YES	NO	YES	NO
GH	GHR	YES	YES	YES	NO	YES	YES
IL6	GP130	YES	YES	NO	NO	YES	YES
LIF	GP130	YES	NO	NO	YES	YES	YES
OSM	GP130	YES	YES	YES	NO	NO	YES
CNTF	GP130	YES	YES	YES	YES	YES	YES
CTF1	GP130	YES	YES	YES	NO	NO	YES
IL23	IL12R	YES	YES	YES	NO	YES	YES
IL12	IL12R	YES	YES	NO	NO	YES	YES
LEP	LEPR	YES	NO	NO	YES	NO	NO
PRL	PRLR	YES	NO	YES	NO	NO	NO
Non-Binding Pairs							
IL6	EPOR	NO	NO	NO	NO	NO	NO
IL6	CSF3R	NO	YES	YES	NO	NO	YES
CSF3	EPOR	NO	NO	NO	NO	NO	NO
LEP	GHR	NO	NO	NO	NO	YES	NO
OSM	CSF3R	NO	NO	YES	NO	YES	NO
LEP	GP130	NO	NO	NO	NO	YES	NO
GH	GP130	NO	NO	YES	NO	YES	NO
EPO	GP130	NO	NO	YES	NO	NO	NO
IL12	GP130	NO	NO	YES	NO	NO	NO
IL23	GP130	NO	NO	YES	NO	NO	NO
IL12	EPOR	NO	NO	NO	NO	NO	NO
IL23	EPOR	NO	NO	NO	NO	NO	NO
IL6	PRLR	NO	NO	YES	NO	NO	YES
IL12	LEPR	NO	NO	NO	NO	YES	NO
Sensitivity			58%	67%	33%	50%	75%
Specificity			93%	50%	100%	64%	86%
Significance							0.62

TNF-Like Cytokines							
Ligand	**Receptor**	**Binding?**					
Binding Pairs		Exp	CM	SP	FF	CR	All
SF1	RSF1A	YES	NO	NO	NO	YES	YES
SF1	RSF1B	YES	NO	NO	NO	NO	YES
SF1	RSF14	YES	NO	NO	NO	YES	YES
SF2	RSF1A	YES	NO	NO	NO	YES	YES
SF2	RSF1B	YES	NO	NO	YES	NO	YES
SF3	RSF1A	YES	NO	YES	NO	NO	NO
SF3	RSF1B	YES	YES	YES	NO	YES	YES
SF3	LTBR	YES	YES	NO	NO	NO	YES
SF4	RSF4	YES	NO	NO	NO	YES	YES
SF5	RSF5	YES	NO	YES	NO	NO	NO
SF6	RSF6	YES	NO	NO	YES	YES	YES
SF6	RSF6B	YES	NO	NO	YES	YES	YES
SF10	RSF11B	YES	NO	NO	NO	NO	YES
SF11	RSF11A	YES	NO	YES	NO	NO	NO
SF11	RSF11B	YES	NO	YES	NO	NO	YES
SF12	RSF12A	YES	NO	NO	YES	YES	YES
SF13	RSF17	YES	NO	NO	NO	YES	NO
SF13B	RSF17	YES	NO	NO	NO	YES	NO
SF15	RSF6B	YES	NO	NO	YES	NO	NO
SF15	RSF12	YES	NO	YES	YES	YES	YES
Non-Binding Pairs							
SF1	LTBR	NO	YES	YES	NO	NO	NO
SF1	RSF4	NO	YES	NO	NO	YES	YES
SF1	RSF5	NO	YES	YES	NO	NO	YES
SF2	RSF6	NO	NO	NO	NO	YES	NO
SF2	RSF6B	NO	YES	YES	NO	YES	YES
SF3	RSF11A	NO	YES	YES	NO	NO	NO
SF3	RSF11B	NO	NO	YES	NO	YES	NO
SF3	RSF12A	NO	NO	NO	NO	YES	NO
SF4	RSF17	NO	YES	NO	NO	YES	YES
SF5	RSF6B	NO	YES	NO	NO	NO	NO
SF6	RSF12	NO	NO	YES	NO	YES	YES
SF6	LTBR	NO	YES	YES	YES	NO	NO
SF10	RSF4	NO	YES	YES	NO	YES	NO
SF11	RSF5	NO	YES	YES	NO	NO	NO
SF11	RSF6	NO	YES	YES	YES	YES	YES
SF12	RSF6B	NO	NO	YES	YES	YES	YES
SF13	RSF11A	NO	NO	YES	YES	YES	NO
SF13B	RSF11B	NO	YES	YES	NO	YES	NO
SF15	RSF12A	NO	YES	NO	YES	YES	YES
SF15	RSF17	NO	NO	NO	YES	YES	YES
Sensitivity			10%	30%	30%	55%	70%
Specificity			35%	35%	70%	30%	55%
Significance							0.81

	Raw	Normalized
CM Sensitivity	33%	58%
CM Specificity	57%	93%
Combined Sensitivity	33%	75%
Combined Specificity	50%	86%

Comparison of sensitivity and specificity values for raw versus normalized scores for both the standalone CM and combined methods.

Query Ligand-Receptor Pair	**Template Complex**	**% similarity of query ligand to template ligand**	**% of interacting residue pairs correctly identified (RAPTOR)**	**% of interacting residues correctly identified (LTHREADER)**
GH-GHR	IL6-GP130	23	33	53
GH-GHR	EPO-EPOR	15	30	55
GH-GHR	CSF3-CSF3R	20	31	63
EPO-EPOR	IL6-GP130	24	29	51
EPO-EPOR	GH-GHR	25	44	43
EPO-EPOR	CSF3-CSF3R	18	31	58
IL6-GP130	GH-GHR	22	28	52
IL6-GP130	EPO-EPOR	17	40	63
IL6-GP130	CSF3-CSF3R	21	42	66
CSF3-CSF3R	IL6-GP130	24	51	57
CSF3-CSF3R	EPO-EPOR	19	33	61
CSF3-CSF3R	GH-GHR	25	32	52
Average		**21**	**35**	**56**

Comparison of alignment accuracy between RAPTOR and LTHREADER for 4-helical bundles.

Query Ligand-Receptor Pair	Template Complex	% similarity of query receptor to template receptor	% of interacting residue pairs correctly identified (RAPTOR)	% of interacting residues correctly identified (LTHREADER)
TNFSF13B-TNFRSF17	TNFSF13-TNFRSF13B	23	41	68
TNFSF13B-TNFRSF17	TNFSF10-TNFRSF10B	16	30	55
TNFSF13B-TNFRSF17	TNFSF13B-TNFRSF13C	39	60	82
TNFSF13B-TNFRSF13C	TNFSF13-TNFRSF13B	29	35	65
TNFSF13B-TNFRSF13C	TNFSF10-TNFRSF10B	18	25	45
TNFSF13B-TNFRSF13C	TNFSF13B-TNFRSF17	39	54	66
TNFSF10-TNFRSF10B	TNFSF13-TNFRSF13B	29	56	54
TNFSF10-TNFRSF10B	TNFSF13B-TNFRSF17	16	32	58
TNFSF10-TNFRSF10B	TNFSF13B-TNFRSF13C	18	42	38
TNFSF13-TNFRSF13B	TNFSF13B-TNFRSF17	23	56	74
TNFSF13-TNFRSF13B	TNFSF10-TNFRSF10B	29	40	62
TNFSF13-TNFRSF13B	TNFSF13B-TNFRSF13C	29	45	84
Average		**25**	**43**	**63**

Comparison of alignment accuracy between RAPTOR and LTHREADER for TNF-like cytokines.