Computational Prediction and Prioritization of Receptor–Ligand Pairs

Ernesto IACUCCI

Jury:
Prof. dr. ir. Yves Jean-Luc Moreau, promoter
Prof. dr. Catherine Verfaillie, co-promoter
Prof. dr. ir. Bart De Moor
Prof. dr. Jan Aerts
Prof. dr. Jesse Davis
Prof. dr. Pierre Rouze (University of Ghent)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor in Engineering

March 2013

Preface

Over the past few years, I have had the chance to meet and work with many interesting and gifted people. I wish to thank them for their help and friendship here. Francesca, Roland, Shi, Fabian, Leo, Georgios, Sylvain, Dusan, Tunde, Alejandro, Minta, Nico, Yusef, Charalampos, Ryo, Arnault, Peter, Jiqiu, Raf, Lieven, Pieter, Amin, Olivier, Annaleen, Pooya, Sarah, Tim, Karien, Thomas, Daniela, Joanna, Wouter, Dries, Philip, Kim, Rocco, Marco, Samuel, and Liesbeth are all names I will not forget soon.

I would also like to that the jury Catherine Verfaillie, Bart De Moor, Jan Aerts, Pierre Rouze, Jesse Davis and my supervisor, Yves Jean-Luc Moreau, I hope you all enjoyed the science.

I would like to thank the staff at ESAT. Maarten, Liesbeth, Inge, Steven, Mimi, Ida, Pela, John, Elsy, Elian, Ilsa, Lut and Evelyn have all made the ESAT experience special for me.

Finally, I would like to thank my family and friends who are found in various physical places but always in my thoughts. Arlene, Thomas, Shamika, Melanie, Alicia, Fearghal, Frank, Paul, Anya, Tobias, Claudia, Julien, Lorena, Maria, Donato, Silvia, Sofie, Ana-Luisa, Chris, Nadine, Marlies, Karin, Marijke, Jens, Maarten, Roxanna, Yann, Diana and Emilia have always encouraged me on my journey.

Thanks to my nieces and nephews Michael, Sofia, Alessandro, and Olivia for providing all the hugs and a zio needs. Thanks to Settimio, Angela, Anna, Frankie, Gina, and Mary for lending me out to the world for a few years, we're always together in spirit.

Ernesto Iacucci
Brussels, March 2013

Abstract

Regulation of cellular events is initiated, often, via extracellular signaling when a circulating protein ligand interacts with one or more membrane-bound protein receptors. Identification of receptor–ligand pairs is thus an important and difficult task to address as this form of interaction is transient and not well studied. In order to address this problem, we collect the most readily available data from repositories (expression, domain, pathway, sequence, and text-based), and apply a high through-put analysis to this problem.

We have worked on the receptor–ligand pairing problem in three main studies. In our first study, using an LS-SVM classifier, we show that we are able to more aptly match members of the chemokine and *tgfβ* families than a previously published method (Gertz et al. 2003). Notably, we are able to achieve an increase in recall of 0.76 over the 0.44 for the matching of receptor–ligands in the tgfβ family. In our subsequent study, we benchmarked several machine learning techniques, and essayed several parameters, on the receptior–ligand interaction prediction task. We found that we could reach a balanced accuracy of 0.84. In our final work, we produce a publicly available database of our results with respect to a text-based *in silico* prediction workflow. The resulting database, contains several key findings, particularly predictions in the GPCR family with a balanced accuracy of 0.96.

The receptor–ligand prediction task is an essential one, as the challenge of predicting such pairs is an important issue in wet-labs, biotech, and pharmaceutical companies. Through several studies, we have determined the most appropriate methodology to predict the receptor–ligand pairs and have made available high-quality predictions at our ReLianceDB website, a tool to aid in performing effective and targeted research.

Nomenclature

BioGrid	Database of protein and genetic interactions
DIP	Database of Interacting Proteins
DT	Decision Tree
FN	False Negatives
FP	False Positives
HPRD	Human protein reference database
IntAct	Molecular interaction database
KEGG	Kyoto encyclopedia of genes and genomes
KNN	K-Nearest Neighbors
MINT	Molecular Interaction
MIPS	Mammalian protein-protein interaction database
NBC	Naïve Bayes Classifier
PPI	Protein-Protein Interaction
QDA	Quadratic Discriminant Analysis
RF	Random Forest
STRING	Search Tool for the Retrieval of Interacting Genes
TN	True Negatives
TP	True Positives

Contents

Contents

List of Figures …… 6

1 Introduction ……... 7

1.1 Motivation……..…… 8

1.2 Receptors and Ligands…… 9
1.2.1 Receptors……..... 9
1.2.2 Ligands…… 9
1.2.3 Receptor–Ligand Binding……. 10
1.2.4 Protein-Protein Interactions……. 12

1.3 Computational Prediction…… 16
1.3.1 Machine Learning…… 17
1.3.2 Performance Measures..…… …...…… 19
1.3.2.1 Sensitivity……..…….… 20
1.3.2.2 Specificity……..…….…20
1.3.2.3 Balanced Accuracy ..…….…. 20
1.3.2.4 Precision ……..…….…. 20
1.3.2.5 Recall…. ……..…….…. 20
1.3.2.6 F-Measure ……..…….… 20
1.3.2.7 Receiver Operating Characteristic……..…. 20

1.4 Data Sources……..…….. 21

1.5 Structure of Thesis and Personal Contribution…… 22

2 Computational Approaches to PPI Prediction…… 24

3 Study 1: LS-SVM Prediction of Receptor–Ligand Binding…… 39

4 Study 2: Benchmarking of Machine Learning Methods
on the Receptor–Ligand Binding Prediction Task……. 49

5 Study 3: Text-based Prediction of Receptor–Ligand Binding…… 67

6 Discussion and Conclusions…… 77

Curriculum Vitae 85

List of Publications 86
Bibliography 88

List of Figures

1.1 Figure #1 Schematic view of the receptor-ligand pairing prediction task…… 16

1.2 Figure #2 Overview of thesis chapters .....…… 23

1.3 Figure #3 Family Analysis of ReLiance Training Data and Results ….. 81

Chapter 1

Introduction

This chapter introduces the basic concepts of receptor–ligand biology and the computational approaches used to predict their interactions. Section 1 outlines the motivation behind this work and the hypothesis under which we are working. Section 2 outlines the biology behind receptor–ligand Interactions. Section 3 outlines the computational approaches used in the thesis to make the predictions. Section 4 outlines the data sources used in this work and section 5 outlines the structure of the rest of the thesis.

1.1 Motivation

The work presented here is a study of the computational prediction and prioritization of receptor–ligand pairs. The hypothesis being advanced here is that the prediction of receptor–ligand pairings can be advanced through machine learning techniques, particularly when combined with literature information for *in-silico* discovery.

Essentially, the challenge of predicting receptor–ligand pairs is an important issue in wet-labs, biotech, and pharmaceutical companies. The broad range data available to bioinformaticians are represented in our work (expression, domain, sequence, etc). We would like to provide valuable insight into the most appropriate setting (experimental design, feature selection, and similarity measures) to make the interaction predictions for use by any lab/institution working with an interesting set of genes and who wish to know which are the best receptor–ligand candidates to follow-up on?

The current state of knowledge in this field is restricted to physical interaction experiments (see below) as well as docking studies of protein complexes (see below). The first approach is limited in that it does not work for transient interaction which is the basis of many receptor–ligand interactions. The second approach is limited in that the proteins must be collected and purified in high quantity; this is, generally, not possible for membrane bound proteins such as receptors. As the

current options are not feasible, we produce a novel line of research which uses disparate data sources to prioritize candidates for the research community. We use the only, literature validated, dataset of known protein ligand–receptor interactors (see below) and apply various machine learning methods to address this problem. We look both at the chemokine and *tgfb* families from this dataset for our first study. We then look at the whole dataset (which contains disparate families) and assess how well we can use machine learning techniques to learn from this data. Later on, we seek to make novel *in-silico* discoveries using methods trained on this data. We then take our trained classifier and apply it to all terms which co-occur with the entire training dataset members and prioritize those candidates. The final outcome of this work is that there now exist a prioritized receptor–ligand database (generated from a classifier trained on the only known literature validated dataset and applied to the latest list of co-occurring terms) for which wet lab scientists can now consult for their proteins of interest and may validate at their own volition.

The dataset (see below) used was picked because it is the, to our knowledge, the best dataset available. This is due to the fact that the dataset is experimentally validated and publicly available and is the largest known to exist. The features are chosen because they are ubiquitous, well understood by bioinformaticians, and very well curated.

Through several studies, we have determined the most appropriate methodology, in our experimental setting, to predict the receptor–ligand pairs and have made available high-quality predictions at our ReLianceDB website, a tool to aid in performing effective and targeted research.

1.2 Receptors and Ligands

Receptors and their corresponding ligands play a central role in the execution of several biological processes. Typically, ligands circulating in the bloodstream will interact with a membrane bound receptor. The receptor will, thereupon, change its conformation into an active state so as to allow intracellular players to propagate its signal. Here we examine the receptors, ligands, and the phenomenome of receptor–ligand binding.

1.2.1 Receptors

Receptors are divided into two main types: transmembrane and intracellular. Transmembrane receptors are those which are found in the cell membrane and have an extracellular domain which interacts with the ligand(s). These receptors have a transmembrane domain which keeps them integrated into the cell membrane via secondary structures such as alpha helixes. These receptors also have an

intracellular domain which mediates the signaling to the cellular machinery (Congreve and Marshall, 2010b).

Intracellular receptors are globular proteins which are found in the cytoplasm but more commonly in the nucleus (nuclear receptors). Nuclear receptors contain a ligand binding domain as well as a DNA binding domain. Once activated, nuclear proteins bind to DNA to influence transcriptional activity. The ligands which interact with Intracellular receptors must pass through the cell membrane and thus should be small and hydrophobic (Cascio, 2004a).

1.2.2 Ligands

Ligands are usually small molecules which initiate a cascade of events within the cell via their binding to a receptor. They may be a chemical substance or a small protein (peptide) produced in the body. For the purposes of the work presented here, it should be stated that we are discussing exclusively protein ligands.

Ligands may be classified with relation to their activity vis-à-vis the receptor. Ligands classified as *Full agonists* are those which activate the receptor to produce a full biological response. Partial agonists, as the name suggestions, induce a response which is less than the full agonist, yet still present. *Inverse agonists* inhibit the receptor from performing its biological response. The action of *antagonist* ligands is to occupy the receptor, yet not induce any positive response from them. This action prohibits the binding of the other classes of ligands and their subsequent biological effects (Milligan, 2003).

Understanding ligands in terms of these classifications is important as one can immediately see the potential for regulation of receptor activity at the biological function level. Furthermore, in terms of computational prediction, one can appreciate that these different classes add complexity to the elucidation of receptor–ligand pairs as the relationship between features may differ greatly with the relationship between biological realities. For example, the expression profile between an interacting receptor–ligand pair may be correlated, however, the ligand may be a full agonist or an antagonist resulting in vastly different biological responses.

1.2.3 Receptor–Ligand Binding

The activity of a receptor–ligand pairing has often been compared to a lock and key mechanism. This analogy is accurate in that it emphasizes the selective binding of a ligand to its specific receptor. It should be clarified that a ligand may interact with

more than one receptor and vice-versa. In the field of graph theory, this would be described as a "few to few" relationship in a bipartite graph.

Another departure from the lock and key metaphor is the principle of *induced fit*. While a key and a lock are manufactured exactly to compliment each other, a receptor and a ligand are not. As a ligand and receptor come into contact, the two proteins which are constantly shifting in conformation, induce a complementary state in one another through stabilizing residues (intermolecular forces). The residues are charged and occupy key geometric positions, allowing for the necessary intermolecular bonding to achieve binding between the receptor and the ligand (Congreve and Marshall, 2010a; Cascio, 2004b; Kumar and Thompson, 1999).

As the biological environment in which receptors find themselves is highly packed with various proteins, the ability of the cell to create successful receptor–ligand complexes will be influenced by the concentration of both the ligand in extracular matrix and the number of receptors on the cell membrane. Regulation of the receptor activity may therefore occur if the cell produces fewer receptors (receptor desensitization) and if the receptors are internalized (receptor sequestration) (Boulay et al., 1994).

Receptor–ligand pairings are a specific form of protein-protein interactions (PPI), very different from cellular PPI investigation. A PPI which occurs in the cell is often characterized by stable complex formation. Conversely, receptor–ligand interaction occurs in the extracellular environment and is often transient in nature. Furthermore, existing PPI methods are wrought with false positives and false negatives (true positives are values which are reported as true which are actually false, likewise false negatives are values which are reported as false but are actually true). The Database of Ligand–Receptor Partners (DLRP) is unique in that it is a literature-curated database derived from experimentally validated data. The use of other PPI databases can be used for this prediction task, however, the quality of the results will be lower as they are derived from noisier high through-put data.

Disruptions of receptor─ligand interaction are a source of many pathologies. For instance, during the normal process of glucose regulation in the body, insulin interacts with its receptor and a endocrine signal is then propagated until the body begins to store glucose as glycogen (Vind et al., 2012). A disruption in this process takes place when the receptor is no longer present at the cell surface in sufficient amounts (due to cell endocytosis of the receptor or due to a hereditary defect) to accommodate the physiological demand and hyperglycemia ensues (Vind et al., 2012).

The experimental methods for determining receptor─ligand interaction vary greatly. The principle methods for detecting interaction range from radioactive to fluorescence assisted assays. A radioreceptor assay (RRA) involves the use of a

labeled ligand and an analyte which must complete for interaction with the receptor. The last step in the process involves separating the bound and unbound ligand to determine the fraction of the ligand bound to the receptor (Sadee et al., 1982; Villiger et al., 1981). Another radioactivity based assay is the scintillation proximity assay (SPA) where an immobilized receptor is mixed with the labeled ligand. An emission of light results from the close proximity (10 μm) of the receptor and the ligand and is thus registered as an interaction event (Ferrer et al., 2003; Sen et al., 2002; Gobel et al., 1999).

As radioactivity based assays present a health concern, other methods present a safer alternative. For example, assays based on fluorescence, have been developed to produce a colored result when measuring interaction. One example is fluorescence energy transfer (FRET) where energy transfer between the interacting receptor and ligands cause a fluorescent emission (Milligan, 2004; Pope et al., 1999). Another example of a fluorescence based assay would be the use of enhanced green fluorescence protein (EGFP) fused protein (Ilien et al., 2003). Briefly, this method makes use of a receptor fused to the EGFP which emits light when the receptor comes into close proximity to the labeled ligand.

Another noteworthy method, similar to fluorescence based techniques, would be bioluminescent assays. As is the case with FRET, bioluminescence resonance energy transfer (BRET), makes use of light emission to detect interaction. The main difference between BRET and FRET would be that BRET emissions are the result of an enzymatic reaction whereas FRET relies on the energy transfer between donor and acceptor molecules (Milligan, 2004; Pfleger and Eidne, 2003; Boute et al., 2002; Issad et al., 2002; Angers et al., 2000).

Building upon the success of FRET and BRET technologies (described above), other non-radioactive methods have been developed to detect receptor─ligand binding. For example, AlphaScreen (Wilson et al., 2003; Rouleau et al., 2003; Ullman et al., 1996) and flow cytometry (Edwards et al., 2004; Simons et al., 2004; Sklar et al., 2002; Bohn, 1980) make use of bead-immobilized receptors or ligands in a media which contains reagents which allow for a light emitting reaction upon interaction of the receptor and ligand.

1.2.4 Protein–Protein Interactions

The ‘omics’ era has presented tremendous opportunity for high throughput investigation into important questions facing the research community. Many investigative strategies of implementing data mining techniques in combination with high through put experiment have accomplished much. Several of these high-through-put experimental methods, yeast two hybrid systems--Y2H (Ito et al., 2001), pull-down assays (Vikis and Guan, 2004), tandem affinity purification (Puig et al., 2001), mass spectrometry (Gavin et al., 2002; Puig et al., 2001), microarrays (Stoll

et al., 2005) and phage display (Willats, 2002), have all generated enormous datasets yet they are incomplete and are composed of many false positives and false negatives.

A protein–protein interaction (PPI) is a physical interaction of two proteins. These interactions may be determined by methods listed above. For example, a yeast two hybrid system involves the measuring of a conjugated florescence of interacting fusion proteins (Ito et al., 2001). This system is simple and allows for high through-put screening but results in many false positives. Conversely, mass spectrometry is a complicated and expensive method which allows for the essay of two proteins at a time. Under this setting interacting proteins are broken into ionized fragments and their mass-to-charge ratio is measured by a mass spectrometer. A spectra output of peaks is then generated and must be interpreted in order to find out which proteins composed the starting sample (Gavin et al., 2002).

More recent advances in PPI investigation has lead to some notable methods which are more capable to measure PPI, particularly receptor–ligand interactions. One such method is the Avidity-based EXtracelllar Interaction Screen (AVEXIS) which makes use of a library of recombinant proteins of the extracellular domain of receptors and ligands to create baits and preys which are later used for interaction screenings. A similar method which makes use of ubiquitin to determine PPI is Membrane Yeast Two Hybrid (MYTH) (Snider et al., 2010). Briefly, this method fuses candidate partners (such as receptors and ligands) to the C and N terminus of a split-ubiquitin protein. The ubiquitin moieties are brought together when there is an interactin between the two candidates. This triggers a deubiquitinating enzyme to recognize this molecule and cleave an attached transcription factor. The transcription factor then turns on the expression of a reporter gene. Another notable method would be Membrane-SPINE (Strep-protein interaction experiment) (Muller et al., 2011) which makes use of conventional strep-tag purification but allows for its extension to membrane proteins by introducing a cross-linking step (carried out with formaldehyde). The cross-linked proteins are then identified through mass spectrometry and immunoblot analysis.

Several databases exist to store information about validated or predicted protein–protein interactions. They include the Munich Information Center for Protein Sequences--MIPS database (Mewes et al., 2004), the Molecular Interactions--MINT database (Zanzoni et al., 2002) the IntAct database (Kerrien et al., 2007), the Database of Interacting Proteins--DIP (Xenarios et al., 2000), the Biomolecular Interaction Network Database--BIND (Bader et al., 2001; Xenarios et al., 2000), and the BioGRID database (Stark et al., 2006; Xenarios et al., 2000). Some, like the Yeast Proteome Database (Xenarios et al., 2000; Hodges et al., 1999) contain interactions which are a combination of protein interaction and literature data.

The field of general PPI prediction is crowded with very different approaches, all aimed at providing high quality results. Some of the approaches make use of relating the features of proteins in order to determine if there is an interaction. For example, (Shen et al., 2007) make use of a conjoint triad sequence feature in order to predict PPI using an SVM approach. Building on this, (Chang et al., 2010) combine this approach with protein surface information (accessible surface area) in order to produce results which improve upon those from surface identification alone.

The D-Star method (Tan et al., 2006), makes use of domain data by finding over-represented motifs which co-occur in interacting protein pairs. The rational behind this approach was then extended by (van Dijk et al., 2010) who select relevant motifs from those determine in D-Star. Similarly, (Yu et al., 2010) make use of sequence data as they mine amino-acid triplets to find which are present in interacting proteins. The data is transformed into a fearture vector which also takes into account amino acid charges.

The Search Tool for the Retrieval of Interacting Genes (STRING) (Szklarczyk et al., 2011) database combines several lines of evidence to determine protein–protein association information. The two primary means of assigning associations comes from either analysis of genomic information or transferring associations between organisms ('interlog' information). A confidence score is assigned to all associations in order to provide a gage of the quality of the association. This score is probabilistically derived based on the benchmarking of groups of associations against the manually currated associations found in KEGG database (Punta et al., 2012).

Throughout these studies an important theme emerges from the work which must be addressed. The question of how does the biological relevance of the data play a role in the development and training of the machine learning algorithms used in this work. Essentially, the training of the machine learning algorithms are based on the structure discerned from the training data which can then be applied to new examples to determine if they are interacting or non-interacting. While it's not entirely possible to directly resolve from where all of the structure in the data is derived, it is possible to provide guidance based on the state of the art understanding in PPI.

The concept of gene duplication provides the best explanation of the structure present in the data. The phenomenon of gene duplication is a well studied area of research (Pereira-Leal et al., 2007) and has implications in PPI (O'Brien et al., 2005) as well as Protein Interaction networks (PIN) (2011). Gene duplication occurs when a gene undergoes a duplication event, this is generally followed by a divergence between the original and the copy due to mutations arising from evolutionary processes (Lynch and Katju, 2004). A consequence of gene duplication is the conservation or loss of protein interaction partners of the copied gene (Ihmels et al.,

2004). Considering the protein interaction network (PIN) of an organism, these gene duplication events contribute serious complexity to the PIN as a whole. The addition of duplicated genes within the network creates hubs and increases the number of edges. These types of conserved interactions form important redundant architecture for network connectivity (Yang et al., 2011; Lovell and Robertson, 2010; Lockless and Ranganathan, 1999).

The receptor–ligand pairs which are the subject of the work presented in this thesis are very much a part of the PPI network and have evolved under similar conditions Looking at the examples given in the Gertz et al. work (2003) and the comparison work done in Iacucci et al., two families are examined in depth in terms of prediction performance. These "families" have very different evolutionary histories due to the fact that they have been annotated in very different manners. The cytokine family has been annotated with respect to the signaling function of the family members and not in any way annotated due to evolutionary relatedness (Locksley et al., 2001). The TGF-beta family, in contrast, has been annotated due to the evolutionary relationship present in the transforming growth factor-beta superfamily (with such members as TGF-B1, TGF-B2, TGF-B3 which are iso-forms of one another). Characteristic of TGF-beta (and present in the iso-forms) is the presence of an N-terminal signal peptide, a latency associated peptide region, and a C-terminal region which undergoes protein modification during the protein's maturation (Herpin et al., 2004).

Parallel to the way in which our methods uses the structure found in the data to make predictions, many algorithms use similarly based clustering strategies in order to find local dense sub-networks in PIN data. These strategies include HUNTER (Chin et al., 2010), CORE (Leung et al., 2009), HCPIN (Wang et al., 2011),COACH (Wu et al., 2009), NWE (Maruyama and Chihara, 2011). The motivation behind these network studies is to find functional modules in the PIN. As explained above, these functionally related modules should have arisen through shared evolutionary history (as suggested via gene duplication events). Nowhere is this point better demonstrated in studies in which "bridge nodes" are taken into account. Bridge nodes in a graph represent nodes which are connected to multiple clusters. Evolutionarily, these nodes can be explained as heavily connected genes which have duplicated and maintain conserved edges. Studies and algorithms which are based on the special consideration of bridge nodes are hub-duplication (Asur et al., 2007) and DECAFF (Li et al., 2008). These algorithms will ignore the bridge nodes and then mine the rest of the PIN to look for densely connected regions which are more likely to have functionally related members. These regions are more likely to be functionally related as they should also be evolutionarily related. As discussed above, this concept also plays a role in the receptor-ligand prediction task. Therefore, the success of these methods are based on a similar rationale as the ones discussed in the receptor-ligand prediction task.

1.3 Computational Prediction

Computational prediction allows for the facilitation of high-dimensional problems which may contain complex and enigmatically related data. The high through-put era of computational biology has added further difficulty as the problems inherent to this field often involve thousands of data points which cannot be processed without computational resources. The prediction of receptor–ligand pairs is one such problem for which we apply methods for the prediction and the prioritization of candidate pairs. Below we will discuss the machine learning methods and performance measures which we have applied to this problem.

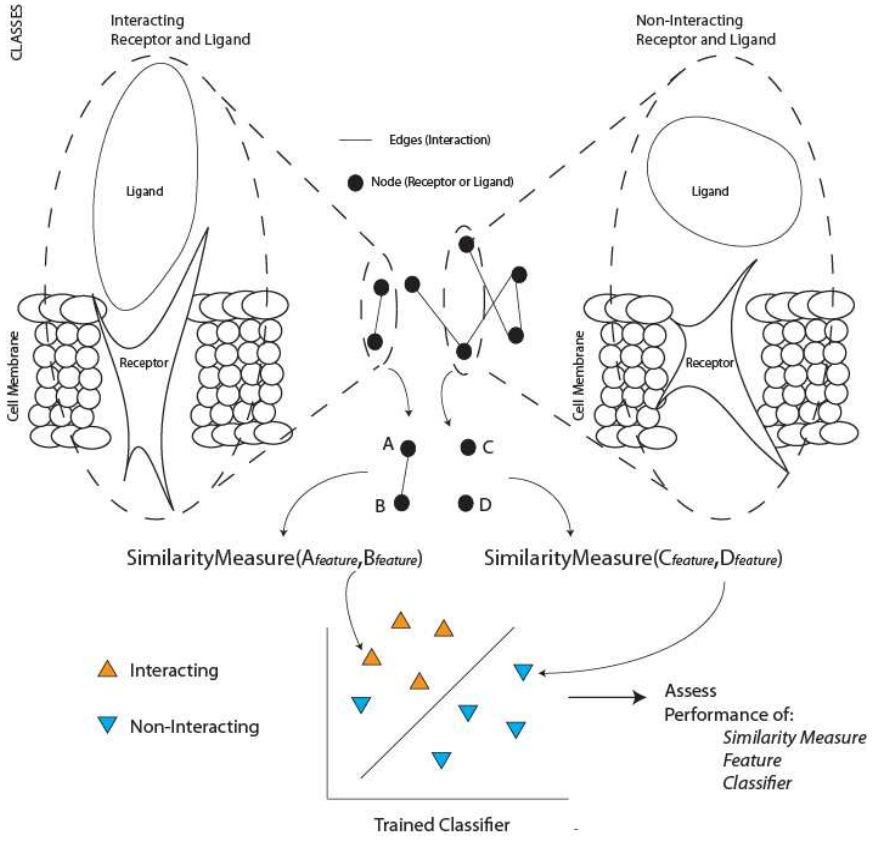


Figure 1 ***Schematic view of the receptor-ligand pairing prediction task*** On top is displayed a schematic view of an interacting (left) and a non-interacting receptor-ligand pair (right). Below is represented the workflow for training a classifier using (i) features from a genomic data source, (ii) a similarity measure, and (iii) a classification algorithm.

1.3.1 Machine Learning

Identification of receptor–ligand pairs is an important and specific form of protein–protein interaction (PPI) prediction. We consider that all protein pairs may be assigned to one of two classes, the *interacting* class and the *non-interacting* class (see Figure 1).

Interacting receptor–ligand pairs often share common or related features such as expression profile and phylogenetic history. Using attribute-value data to predict classes (such as interacting or non-interacting) is an active area of research in the domain of machine learning (Cooper et al., 1997). Generally, the process of learning from the features to predict classes, in a "supervised" context, is a process which starts with partitioning the data. Data is partitioned into training and test data. Training data is used by the classifiers to construct rules based on known feature measurements of members of the classes. Once a classifier is built, its performance can be assessed using the test data.

There exists a variety of classifiers, each with various strengths and weaknesses to consider. Some of these classifiers have been built with individual purposes in mind and, in noteworthy manner, some are related. Here we review the methods used in this thesis according to the aforementioned criteria.

A K-Nearest Neighbor classifier (KNN) is a method by which objects are classified based on their proximity to the closest training example, as found in the feature space (Plewczynski, 2009; Raymer et al., 1997). KNN is one of the simplest machine learning methods as labeling a new example requires a majority vote of its *k* nearest neighbors. On one hand, KNN is very simple to implement and understand, yet on the other hand one must have enough training examples which cover the input space well enough in order to have a well trained model. This model is also very sensitive to noise and can be slow in execution if the training set is very large. However, it is suitable for online learning as training data simply consists of adding new examples.

Related to linear discriminate analysis, a quadratic discriminant analysis classifier (QDA) differs in that it uses a quadratic surface to separate the measurements of the objects being classified (Lu and Luo, 2008). QDA separates classes based on a quadratic surface. This surface is chosen to maximize the ratio of between-class to within-class variance. The method requires Gaussian distribution of the classes and it is very sensitive on deviations from this assumption.

Based on Bayes's Theorem, Naïve Bayes Classifiers (NBC) are simple classifiers which assume conditional independence of features when assigning classes to objects, which renders it less sensitive to the curse of dimensionality (Murakami and Mizuguchi, 2010). The NBC is trained by learning statistical parameters from the

given data. In addition, it has been shown both in practice and theoretically that Naïve Bayes can produce reliable classification even in some cases where independence assumption is considerably violated.

A Decision Tree (DT) is a predictive model used to map observations of features into a tree structure so as to be able to make conclusions about the class assignments (Darnell et al., 2007). Each node in the tree searches for a feature which best separates all examples trained at that node. A trained classification tree predicts membership of the new instance given values of one or several of its attributes sequentially, by testing them against rules embedded in nodes of previously learned hierarchical structure. A disadvantage to using decision trees is that the topology of the tree is highly influenced by the training data, in order to avoid this, the use of Random Forests have become popular. Advantages of using decision trees are that they are fast to train and evaluate.

Random Forests belongs to group of ensemble methods. Random Forests (RF) are a collection of decision trees generated from randomly sampling the features in the training set. Essentially, it is collection of decision trees built on different bootstrap samples extracted from training data. An additional randomness is injected to algorithm by limiting the choice of candidate variables for each split in each decision tree to randomly chosen subset, typically of size proportional to logarithm or square root of total number of variables. Assignments are made when all trees "vote" for the one class or another and the class assignment is subsequently made to the most popular class. The advantage of using random forests over decision trees is that the data is less sensitive to a "lucky split" of the data (a split of the training and test data where, by chance, the training data for the classifier contains informative examples which result in unrealistically high testing results), as the training data is sampled several times. The disadvantages to using random forests is that they are slow to train and evaluate (Qi et al., 2005).

Support Vector Machines are class of methods for nonlinear classification and regression based on the statistical learning theory and structural risk minimization. Using a kernel function they map observations to new, often infinite dimensional, feature space where optimal (margin-maximizing) separating hyperplane is then constructed (Bleakley and Yamanishi, 2009; Jacob and Vert, 2008; Nagamine and Sakakibara, 2007; Suykens et al., 2001). Due to this, the support vector machines are able of inferring arbitrary shaped decision boundaries and of handling highly dimensional data. Disadvantages of using kernels is that they must be completely retrained when presented with new training examples and that various parameters of the model are laborious to tune.

Proteins, such as receptors and ligands, have numerous features. For example, the proteins are encoded by genes which have expression measurements which vary with time or condition (these measurements are known as expression profiles). A

ligand “A” and its corresponding receptor “B” have individual feature measurements, such as a vector representation of their expression profiles (as the two proteins interact, it’s intuitive that the two profiles should be similar as, in order to perform their biological function, they should be present at the same time in the cell). Applying a similarity measure to the two individual features will generate a summary feature for the pair (intuitively, the measure should be higher for interacting protein pairs) which can then be used to train a classifier (see Figure 1).

The data used in the learning task must be separated in order to train the classifier, validate the tuned parameters, and test the performance of the classifier. Thus the data is separated into training data, validation data, and testing data. The training data is used to train the classifier to create rules for the classifier to use when it encounters new examples. The validation data is used to tune certain training parameters used in the classifier. For example, as the Random Forest classifier has a variable which is the number of decision trees to use in the ensemble. The testing data is used to measure the trained classifiers' performance with new examples.

In study two and three, the data is split into two-thirds training and one-third testing with validation data being undertaken separately due to the fact that validation is handled internally by the implementations of the machine learning software used. The parameter turning was done only on the training set and not on the testing data. In our settings we opted to use repeated random sub-sampling of the data. As the orginal dataset was bias (1% positive examples vs 99% negative examples), it was necessary to balance the dataset. Balancing the dataset involves assuring that training examples contain an equal number of postive and negative examples. This was accomplished by randomly sampling from both the positive examples and negative examples in a 1:1 ratio. As there were 449 positive examples, two-thirds (299) of the positive examples were used as training data and another 299 of the negative examples were used in the training data As a result, it is also necessary to test the data under balanced and unbalanced settings.

1.3.2 Performance Measures

In order to assess the “goodness” of our results we have used several performance measures. The measures make use of the number of true positives (*TP*), true negatives (*TN*), false positives (*FP*), and false negatives (*FN*) which were made in the predictions. The use of balanced accuracy is explained by the need to accommodate an imbalanced dataset. Essentially, when either the negative or the positive class of examples is disproportional high, balanced accuracy will correct for the imbalanced.

1.3.2.1 Sensitivity

$$Sensitivity = \frac{TP}{TP + FN}$$

1.3.2.2 Specificity

$$Specificity = \frac{TN}{TN + FP}$$

1.3.2.3 Balanced Accuracy

$$BalancedAccuracy = \frac{TP + TN}{(TP + TN + FP + FN)}$$

1.3.2.4 Precision

$$\Pr ecision = \frac{TP}{TP + FP}$$

1.3.2.5 Recall (see 1.2.2.1 Sensitivity)

1.3.2.6 *F*-Measure

$$F - measure = 2 * \frac{precision * recall}{precision + recall}$$

1.3.2.7 Receiver Operating Characteristic

Analysis of results can often be subjective in situations where there is no clear selection criteria in terms of sensitivity or specificity. The receiver operating characteristic (ROC) is a useful tool in this regard. Briefly, ROCs are constructed by plotting the sensitivity (along the x-axis) against 1-specifity (along the y-axis). The resulting curve corresponds to the trade-off between the two dimensions when choosing from a range of parameter function (most commonly a threshold for the classifier outcome).

The use of the ROC is generally executed by examining the area under the curve (AUC). AUC values can be interpreted as follows, an AUC of 0.50 indicate that the outcomes plotted are close to random (as would be the case if the points were plotted along a diagonal line from the bottom left corner to the upper left corner of the plot). An AUC value of 1.0 is a perfect classification as this would represent a 100% sensitivity and a 100% specificity. An AUC of below 0.50 would be considered worse then random and (if extremely low) may indicate that the labels are erroneously switched. AUC values between 0.5 and 1.0 indicate that predictions are better than random.

1.4 Data Sources

Several sources are considered to extract protein data that can be relevant for our prediction task. In particular, we retrieve expression profiles, curated pathway information, and sequence data. While many more data sources are publicly available, we have selected the data sources that are representative of the various forms of evidence currently available and that have already been used for receptor-ligand prediction (Iacucci et al., 2011b; Gertz et al., 2003).

Phylogenetic: complete protein sequences are retrieved for seven species (Rattus norvegicus, Mus musculus, Homo sapiens, Pan troglodytes, Canis familiaris, Cavia porcellus, and Bos taurus) from EnsEMBL build 51 (Hubbard et al., 2009; Hodges et al., 1999). Sequences are then aligned using ClustalW (Thompson et al., 1994b)., 1994) to detect orthology. The protein sequence and the six orthologous sequences that possibly exist are edited (as per procedure outlined in (Iacucci et al., 2011b)). Eventually, and for each protein, all pair-wise alignment scores between these seven sequences are used to build the phylogenetic vector.

Expression: the gene expression profiles are retrieved from the GNF human expression atlas (Su et al., 2004). Each profile contains 79 values corresponding to the 79 conditions considered by Su et al.. The profiles are normalized prior to analysis (means are set to zero, and standard deviations to one).

Domain: the domain information is retrieved from the InterPro database (Hunter et al., 2009), through EnsEMBL. Only the domains present in at least one of the 210 receptors and ligands considered are kept for further analysis. Each protein is therefore represented by a sparse binary vector of size 224, corresponding to the 224 relevant InterPro domains.

Motif: The motif data is created using the Toucan toolbox to look for putative motifs in the upstream sequence of the genes. Each protein is then represented by a vector of size 674. Each value represents the score of the corresponding motif for the given protein (i.e., the confidence one can have that this motif is an active motif).

Pathway: The pathway data is retrieved from the Kegg Pathway database (Ogata et al., 1999). Similarly to the domain data source, only the pathways in which at least one of the 210 receptors and ligands is involved are used to build the final profiles. In the end, proteins are represented by sparse binary vectors of size 314.

Golden Standard: While many PPI resources are available, all except one have no specific scope geared towards receptor–ligand interactions. The exception, used in the work presented in this thesis is DIP which has a carefully curated subset of ligand–receptor partners which forms the bases of its Database of Ligand–Receptor Partners (DLRP) (Graeber and Eisenberg, 2001). The database was constructed using experimentally determined ligand−receptor cognate pairs through a literature review. The database contains 314 proteins, 210 of which are used in our experiments (study 2 and 3) as they contain all the necessary data for the execution of our work. Study 1 uses only the members of the database which are contained in the chemokine and tgfb families as this study is a comparison to that of Gertz et al. which restricts the analysis to only these 2 families.

1.5 Structure of Thesis and Personal Contribution

The work presented here is a study of the computational prediction and prioritization of receptor–ligand pairs. The thesis is structured as a combination of three related studies on the prediction of receptor–ligand interaction prediction.

In chapter 2, we present a review of PPI and the concepts behind several methods which are used to make PPI predictions.

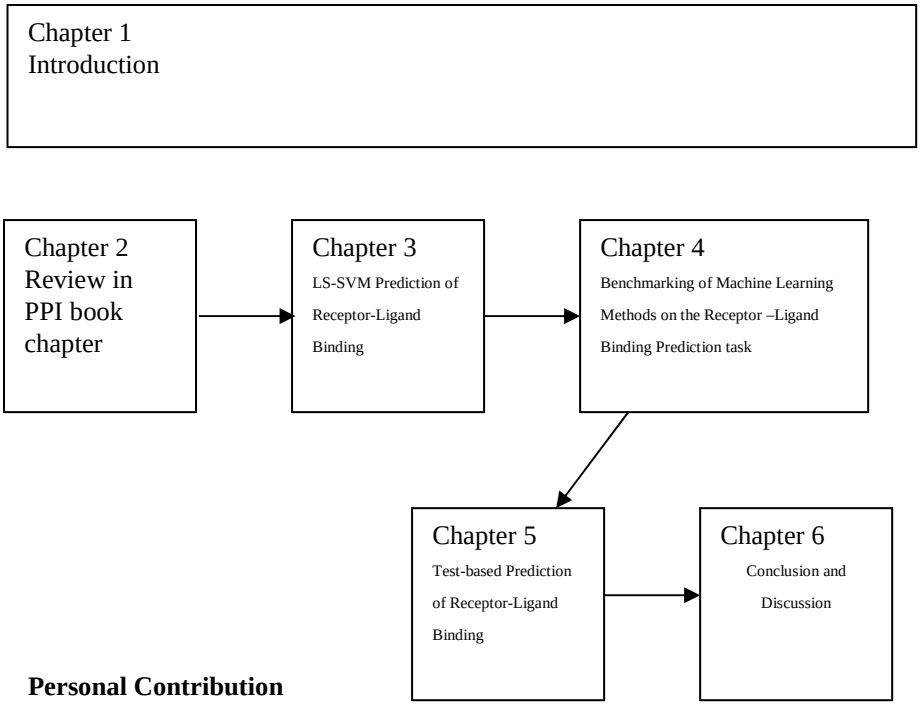
In study #1 (chapter #3), using a LS-SVM classifier, we show that we are able to more aptly match members of the chemokine and *tgfβ* families than a previously published method (Gertz et al. 2003). Notably, we are able to achieve an increase in recall of 0.76 over the 0.44 for the matching of receptor–ligands in the *tgfβ* family.

In study #2 (chapter #4), we benchmarked several machine learning techniques, and essayed several parameters, on the receptor–ligand interaction prediction task. We found that we could reach a balanced accuracy of 0.84.

In study #3 (chapter #5), we produce a publicly available database (ReLianceDB) of our results with respect to a text-based *in silico* prediction workflow. The resulting database, contains several key findings, particularly predictions in the GPCR family with a balanced accuracy of 0.96.

Chapter 6 is a discussion summarizing the merits of our approaches and what perspective has been gained through these works. We conclude with a discussion on possible improvements and future directions in the field.

Figure 2: Overview of thesis chapters



Personal Contribution

Chapter 1: The PhD candidate worked in collaboration with his co-author for this book chapter.

Study 1: The PhD candidate is wholely responsible for all work presented here. The second author provided extremely helpful guidance and discussion concerning the experimental design and was consulted when necessary in the revision process.

Study 2: The PhD candidate is responsible for designing of the experiment, data processing, technical troubleshooting, analysis of experimental results, and writing of the paper. The second and third authors provided data collection, experimental design ideas, implementation of matlab routines, extremely helpful discussion concerning the whole project.

Study 3: The PhD candidate is responsible for designing of the experiment, data processing, analysis of experimental results, and writing of the paper. The second, third, and fourth authors provided data collection, implementation of matlab routines, figures, website design, and extremely helpful discussion concerning the whole project.

Chapter 2

Computational Approaches to PPI Prediction

This chapter represents a review of PPI prediction techniques. As the work presented in this thesis centres around receptor─ligand pairing, which is a specific form of PPI, it is a logical objective to use a review as a starting point for this thesis. The subsequent chapters cover the novel experimentation carried out to make receptor─ligand pairing predictions.

Publication Information: This work is published at: Iacucci E., Xavier de Souza S., Moreau Y., "Computational Approaches to Elucidating Transient Protein-Protein Interactions, Predicting Receptor-Ligand Pairings", Chapter in Protein-Protein Interactions - Computational and Experimental Tools, (Cai W., ed.), InTech (Rijeka, Croatia), 2012, ISBN: 978-953-51-0397-4

Personal Contribution: The PhD candidate worked in collaboration with his co-author for this book chapter.

13

Computational Approaches to Elucidating Transient Protein-Protein Interactions, Predicting Receptor-Ligand Pairings

Ernesto Iacucci[1], Samuel Xavier de Souza[2] and Yves Moreau[1]
[1]*K.U.Leuven*
[2]*Universidade Federal do Rio Grande do Norte*
[1]*Belgium*
[2]*Brazil*

1. Introduction

Protein-protein interactions (PPI) are one of the most important biological events which occur in the cell. As PPIs regulate almost all biological processes in the cell, aberrations in PPI may cause severe health problems. One specific area of PPI is receptor-ligand interactions. These interactions are transient yet account for a large part of cell-to-cell communication. As PPI is an important area of research, many groups have proposed methods to make computational predictions of PPI.

The basis of the majority of these methods rely largely on the phylogenetic profile analysis of candidate interactors. These methods determine the similarity of the phylogenetic history of a protein *A* and its putative protein partner *B*, examining the most accurate measure of similarity between the phylogenetic histories of *A* and *B* in order to predict interaction. As interacting proteins should co-adapt as they are under the same evolutionary pressures, it is self-evident that interacting receptors and ligands should be identifiable by application of the same methodology.

While several methods, described below, make use of phylogenetic information to predict protein-protein interaction (PPI), more contemporary work has been conducted in the area of data fusion and kernel learning. We describe one method [Iacucci et al. 2011] in detail which does both. In this work, the existing line of phylogenetic research is extended by using phylogenetic data to construct a kernel to train a least square support vector machines (LS-SVM) in order to classify candidate receptors and ligands as *interacting* or *non-interacting*.

In this chapter, we discuss the plethora of various methods for determining protein-protein interactions. In addition, we evaluate the application of LS-SVMs to the sub-problem of receptor-ligand interaction prediction.

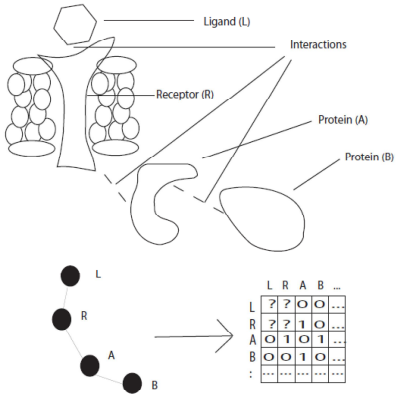


Fig. 1. The Receptor Ligand Schematic. Schematic of receptor-ligand and protein-protein interaction model. Top image is a representation of in-vivo interaction of proteins, receptors, and ligands while bottom image is the graph representation from which a PPI adjacency matrix may be derived. (Figure published in Iacucci *et al.* 2010)

2. Current computational approaches for predicting protein-protein interaction

During the past decade, many methods for prediction of interaction between proteins have been studied due to the crucial role that these interactions have in the understanding of the diverse cellular mechanisms of life forms. Many of these methods involve experimental analysis of specific protein pairs in a smaller scale or, in current high throughput methods [Uetz et al. 2000, Giot et al. 2003], a large amount of protein interactions. The later can be used to detect many interactions with reasonable sensitivity but rather low specificity. Another, relatively inexpensive, way to predict protein-protein interactions does not include wet lab analysis, using instead a variety of computational approaches. These approaches can complement experimental wet lab techniques and are often supported by either the hypothesis of protein co-evolution [Tan et al. 2004, Tillier et al. 2006, Izarzugaza et al. 2006], structural similarities [Gong et al. 2005, Ogmen et al. 2005] or amino-acids sequence conservation [Pitre et al. 2006].

While the entire genomes of many species are already completely sequenced, the interactone of these life forms is often many orders of magnitude larger and yet far from being fully mapped [Claverie et al. 2001, Rubin et al. 2001]. High throughput experimental techniques will certainly help to create this mapping and computational approaches can complement their results identifying false positive interactions, and therefore improving the specificity of these experimental techniques. Apart from the experimental techniques, computational methods are themselves a powerful and affordable alternative to contribute to interactome mapping.

Several computational approaches have been developed in recent years. Many of them are freely available as web tools offering a variety of services to biologists and bioinformatics that range from prediction of interactions between of proteins in pairs or in batch mode, through browsing of consolidated large scale analysis, up to visualization of binding sites and physical interactions in 3-dimensional images.

The methodologies of these many different approaches vary, but they all seem to be supported by the following findings: (a) evidences in favor of the hypothesis of protein co-evolution and the similarities observed in the phylogenetic trees of these proteins; and (b) datasets of already known protein-protein interactions verified by experimental techniques. Co-evolutionary methods find protein pairs with the highest co-evolutionary signal. This information is powerful to predict which members of interacting protein families are associated structurally or functionally although it is not specific enough to predict whether or not two protein families interact. On the other hand, methods supported by verified protein-protein interactions make use of the structural or amino-acid sequence similarities of interacting proteins partners to predict interaction between query protein pairs. This makes such methods more suitable to predict physical interactions rather than functional relationships.

We have reviewed 6 methods and their web tools for predicting protein-protein interactions. Three of them, supported by the protein co-evolution hypothesis, are: TSEMA [Izarzugaza et al. 2006], ADVICE [Tan et al. 2004], Codep [Tillier et al. 2006]. The other three, supported by datasets of verified interactions, are: PIPE [Pitre et al. 2006], PSIbase [Gong et al. 2005], and PRISM [Ogmen et al. 2005]. In the next Sections, we describe each one of these two types of methods.

2.1 Current co-evolutionary methods

Many studies of the problem of predicting protein-protein interactions investigate the similarity of the phylogenetic history of the interaction partners. Many examples of interaction between proteins have presented signs of co-evolution in such a way that members of different interacting protein families present similarity between their phylogenetic trees [Fryxell 1996, Goh et al. 2000, van Kesteren et al. 1996, Moyle et al. 1994, Pazos and Valencia 2001]. The core of co-evolutionary methods is based on measures of similarity for the phylogenetic trees of interacting protein partners.

There are several measures for similarity between phylogenetic trees. The trees can be compared directly [Goh et al. 2000], via distance matrices [Moyle et al. 1994, Goh and Cohen 2002, Ramani and Marcotte 2003, Gertz et al. 2003], or using multiple sequence alignments [Tillier et al. 2006]. In the following Sections, we present three co-evolutionary methods: TSEMA and ADVICE, which uses distance to compare the phylogeneitc trees, and Codep, which computes the correlation between co-evolving partners from their multiple sequence alignments.

2.1.1 Interactive prediction of protein pairing between interacting families TSEMA

TSEMA is a method and web tool to predict mappings between two families of homologous proteins. The probed protein families can either be inputted using the Newick format or in a format comparable with ClustalW, which is used to build the trees. The distances for all

pairs of proteins within both families are extracted from their phylogenetic trees by summing the length of the branches separating each pair of proteins in the trees. The algorithm of TSEMA finds the mapping between the two sets proteins which maximizes the matching between the sets of distances using a modified implementation of the Ramani and Marcotte's Monte Carlo Metropolis method [Ramani and Marcotte 2003].

Availability: http://tsema.bioinfo.cnio.es/

2.1.2 Automated Detection and Validation of Interaction by Co-Evolution – ADVICE

ADVICE predicts and validate protein-protein interactions using observed co-evolution between proteins. The web tool retrieves orthologous sequences of a list of input protein sequences and compute the similarities among the proteins evolutionary histories. The tool also provides visualization for the resulting network of co-evolved proteins.

The ADVICE algorithm infers interaction based on the correlation between distance matrices constructed from the evolutionary history using orthologous sequences of top 10 species. The tool uses BLAST [Altschul et al. 1990] to search the orthologous sequences from Swiss-Prot and TrEMBL databases [Boeckmann 2003]. The distance matrices are constructed using only pairs of orthologous sequences occurring together in the same species. By default, only the orthologous sequences of the top 10 species, based on the BLAST E-value, are used to construct the matrices, excluding those species where more than one orthologous sequence of the input sequence is found. The actual distance matrices are build from the respective multiple sequence alignments using ClustalW [Thompson et al. 1994]. The algorithm then calculates the correlation between pairs of matrices measuring the Pearson's correlation coefficients, which has values between -1, implying 100% anti-correlation, and 1,which representing 100% evolutionary history similarity, being velues above 0.8 good indicators of interaction and values below 0.3 a good cut-off value to detect potential spurious interaction.

Availability: http://advice.i2r.a-star.edu.sg

2.1.3 Maximizing co-evolutionary interdependencies to discover interacting proteins – Codep

Codep and the other co-evolutionary methods find proteins with the highest co-evolutionary signals, independent of physical or functional interaction. The main difference of Codep is that it uses multiple sequence alignments directly rather than distances obtained from the sequences. The user inputs two phylogenetic trees with orthologous sequences. The algorithm maximizes interdependency based on the maximal mutual information. It does this by fixing one of the multiple sequence alignments and varying the order of the other via exhaustive search or via simulated annealing.

The rationale to use directly multiple sequence alignments instead of the distance matrices, which provides a faster way to calculate correlation, is that character-state methods in the field of phylogenetic analysis are more powerful than distance method and some information can be lost in transforming character-state data into distance matrices.

Availability: http://www.uhnresearch.ca/labs/tillier/

2.2 Methods based on verified interactions

Another promising computational approach to predict new protein-protein interactions is to look at the physical structure and the conservation of amino-acid sequences in partners of interactions that are already reliably known to exist. Then, use the gathered information to find correlation with query protein partners of a probed interaction. Many methods apply this approach, which have delivered powerful tools for finding new interactions [Pitre et al. 2006] and even to corroborate with the protein co-evolution hypothesis [Kim et al. 2004]. In the next three Sections we describe three of these methods: PIPE, which compares amino-acid subsequences between probed protein partners and partners of verified protein interactions from a database; and PSIbase and PRISM, both which compare structural characteristics of probed and verified interactions.

2.2.1 Protein-Protein Interaction Prediction Engine – PIPE

PIPE is a computational tool that can effectively identify protein-protein interactions among *S. cerevisiae* protein pairs. It relies on previously determined *S. cerevisiae* protein interactions compiled from the DIP [Salwinski et al. 2004] and MIPS [Mewes et al. 2002] databases to construct a graph where the nodes are proteins and the edges represent the relationship of interacting proteins.

The working principle of the PIPE algorithm to probe interaction between the pair of proteins A-B is to compare sliding subsequences of amino-acids of size w from A to subsequences of the same size of all proteins in the graph of known interactions; then compare sliding subsequences of B to the neighbors of all matches of A. If protein pair C-D are connected in the graph, representing a verified interaction, and if A has subsequence matches with C and B has matches with D, then the pair A-B is more likely to present interaction. The accumulation of all matches of subsequence comparisons presented in form of a matrix indicates a predicted interaction when the higher values in this matrix is above a given threshold of M matches.

The algorithm has three tuning parameters: w, M, and S_{PAM}, which is the threshold value that indicates a match between two subsequences of amino-acids. The author of PIPE chose to fix w in 20, and tune the other two parameter either by trial and error or by statistical evaluation.

PIPE is reported to have success rate comparable to biochemical techniques, with a sensitivity of 61% , specificity of 89%, and overall accuracy of 75%. The main disadvantages of PIPE is its heavy computational burden and its limitation to yeast proteins.

Availability: http://pipe.cgmlab.org

2.2.2 Protein Structural Interactome Map – PSIMAP

PSIMAP is a map that describes the information about domain-domain and protein-protein interactions known to exist in the Protein Data Bank of structures. It is based on the principle that interaction between protein structures is conserved as closely as protein structures themselves [Park et al. 2001, Aloy and Rossell, 2002; Aloy et al. 2003]. It that predicts if domains or proteins structures interact calculating if every possible pair of

structures has an Euclidean distance below a certain threshold. There are three different methods to do this: Full Atom Contact (FAC); Sample Atom Contact (SAC); and Bounding Box Contact (BBC). FAC is the most accurate, whereas SAC and BBC [Dafas et al. 2004] are faster methods.

PSIMAP extract the molecular interaction information of proteins from the PDB. It associates this information to domains using the Structural Classification of Proteins (SCOP) to assign the domains to the structures.

Availability: http://psimap.org and http://psibase.kaist.ac.kr/

2.2.3 Protein Interactions by Structural Matching – PRISM

The PRISM tool allows the user to explore protein interfaces and predict protein-protein interactions by comparing the structure of query proteins to those of a structurally and evolutionarily subset of biological and crystal interactions present in the Protein Data Bank (PDB) [Berman 2000]. Interfaces are defined as the set of residues forming the region of the structure through which two different protein chains bind to each other. This set consists the contacting residues between the chains and the neighboring residues up to a certain distance threshold.

The interfaces in PRISM were obtained from all higher complexes of proteins available in the PDB [Keskin et al. 2004]. From the 49512 interfaces extracted form the PDB, 8205 clusters were obtained using a sequence order-independent computer vision-based algorithm to structurally compare the interfaces. From these 8205 clusters, PRISM considers only 158 template interfaces (Oct/2011) that were found to have evolutionary hotspots [Keskin et al. 2005].

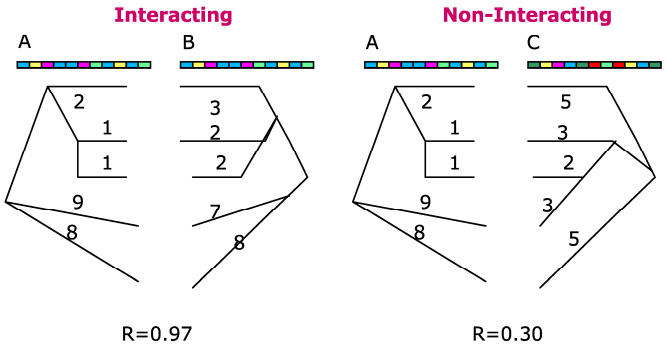


Fig. 2. Phylogenetic Analysis of Proteins

As proteins *A* and *B* are interacting proteins, they share a similar phylogenetic history and thus their phylogenetic profiles are highly correlated (R=0.97). Proteins A and C are non-interacting and are thus not strongly correlated (R=0.30).

PRISM algorithm compares the 158 template interfaces to a target dataset of 18698 structures obtained from passing the structures extracted form the PDB through a 50% sequence identity filter, splitting multimeric proteins into constituent chains, and counting homologous chains only once. The user can also probe a protein structure that is originally not present in the target dataset. To compare target proteins to template interfaces PRISM algorithm do as follow: (a) extract target protein surfaces; (b) compare the target surface with all interface complementary partners from the template dataset using MULTIPROT [Shatsky et al. 2004] in order to detect common geometrical cores in a sequence-order-independent way; (c) check for the presence of hotspots in the target structure. The final prediction score is calculated weighting the structural match ratio and the hotspot match ratio.

Availability: http://prism.ccbb.ku.edu.tr/prism/

3. Phylogenetics and beyond, how multiple kernel learning can improve predictions of receptor-ligand pairings

As seen in the sections above, there are several groups which have used phylogenetic analysis to predict PPI. Here we examine the use of multiple kernel learning in the task of PPI prediction. Kernel learning provides the ability to utilize directly and indirectly related data (such as expression measures, domain content, etc.) and perform classification in high dimensional space. When different data sources are used, separate kernel classifiers can be built and the combined output used to provide a final result.

One of the first groups to look at predicting PPI using multiple data sources was Bhardwaj et al. (2003). They use both phylogenetic information as well as expression data to make their predictions. The use of both data sources were proved, in their work, to provide results with greater accuracy than with using phylogenetic analysis alone. Co-expression is a logical source of information for use in this setting as proteins which interact for the purpose of performing a common function are likely to be co-expressed as they will need to be present at the same time in the cell [Bhardwaj et al. 2003, Grigoriev et al. 2001].

The idea of combining expression and phylogenetic information to predict PPI is clearly a step on a path which leads one to consider a wider variety of data integration. Other data sources include domain information as domains are known to interact and it is clear that this data would provide additional insight into the task of protein-protein interaction. Combining the above mentioned data sources can be carried out by using multiple kernel learning.

To examine the utility of multiple kernel learning with respect to this task, it is necessary to cite an example in which it performs better than other settings. One such example exists when one looks at the work of Gertz et al. (2003) and compare it with the work presented in Iacucci et al. (2011). Both groups look at the receptor-ligand prediction task and apply computational methods to the same dataset. The datasets consist of members of the chemokine and tgfβ ligand families with their respective receptor families. In the case of Gertz et al (2003), distances matrices are created for the families and are matched according to their similarity. Using a Metropolis Monte Carlo optimization algorithm, the Gertz et al. (2003) group explored and scored possible matches between the two matrices, until they reached optimal solutions. A limitation of this approach is that it relied on phylogenetic distance information alone.

Contrary to the work of Gertz et al (2003), the work presented by Iacucci et al 2011 proposes that the integration of multiple data sources results in more accurate matches. This work involved the creating of a combined kernel classifier to carry out the learning task. While other kernel-based works have been applied to the PPI task [Kim et al. 2010, Miwa et al. 2009], the work of Iacucci et al (2011) is unique as they apply multiple kernel learning to the receptor-ligand problem. More specifically, they apply the least-squares support vector machines (LS-SVM) method based on the conclusions by Suykens et al. (2001) which shows this implementation to be robust.

The ability of Iacucci et al. (2011) to predict candidate receptor-ligand pairs has been show to outpace that of Gertz et al. (2003) on the same dataset. This work involves using multiple data sources (expression, phylogenetic, and protein-domain content information), computing separate kernels for each data type, creating LS-SVM classifiers and combining the results to predict receptor-ligand pairs. The specifics of these steps will be discussed below.

3.1 Data sources

Several choices for data sources can be considered when addressing the PPI prediction task. While the studies, mentioned above, which use phylogenetic information rely on sequence data, other sources are available. Such sources include domain content data and expression data.

The phylogenetic data used in the Iacucci et al. (2011) study was derived through several steps. First, candidate receptor and ligand sequences were retrieved for seven species (*Rattus norvegicus*, *Mus* musculus, *Homo sapiens*, *Pan troglodytes*, *Canis familiaris*, *Cavia porcellus*, and *Bos taurus*) from ensemble build 51 [Hubbard et al. 2009]. Following this, the sequences were aligned using ClustalW [Thompson et al. 1994] Once aligned, the sequences were edited so as to eliminate the positions which were not conserved across the seven orthologous sequences. Finally, the pair-wise alignment score was then taken for each possible species to species comparison between the edited orthologous sequences (as seven species are used, a total of 21 pair-wise comparisons for each candidate are created). The distance scores form a phylogenetic vector which was then used to create the phylogenetic kernel.

The expression data used in the Iacucci et al. (2011) work was taken from the well-known GNF human expression atlas (79 tissues) [Su et al. 2004], the data was normalized (values were mean-zeroed and the standard deviation was set to one) and was further transformed into the expression kernel.

For the Iacucci et al. (2011) work, the domain content of each candidate protein (receptor or ligand) was taken from the Interpro Database [Hunter et al. 2009]. A vector for each candidate protein was created where the presence of a protein domain was indicated with a '1' and the absence of a domain was indicated by a '0'. This data was then transformed to create the domain content kernel.

The "Golden Standard" for the verification of the Gertz et al (2003) and the Iacucci (2011) et al. work is based on the Database of Ligand-Receptor Partners (DLRP) [Graeber et al. 2001]. This dataset is an experimentally derived dataset where known receptor-ligand pairs are stored. The information found here was used to train the LS-SVM described below. In addition, it was also used as the "Golden Standard" to determine which predictions, by both

groups, were true positives and false positive as well as false negatives and true negatives. These values were then used to calculate specificity and sensitivity of each groups' predictions to ultimately determine which approach provided better results.

3.2 Kernel creation and the LS-SVM

The creation of the kernels and the training of the least-squares support vector machine (LS-SVM) in the work presented by Iacucci et al. (2011) required multiples steps. First, the data sources, discussed above, were used to create data matrices (phylogenetic, expression, and domain content) which were then used to create three kernels for each receptor-ligand family. Following this, the LS-SVMs were trained using the three kernels to predict outcomes for receptor-ligand pairs known from the DLRP "Golden Standard".

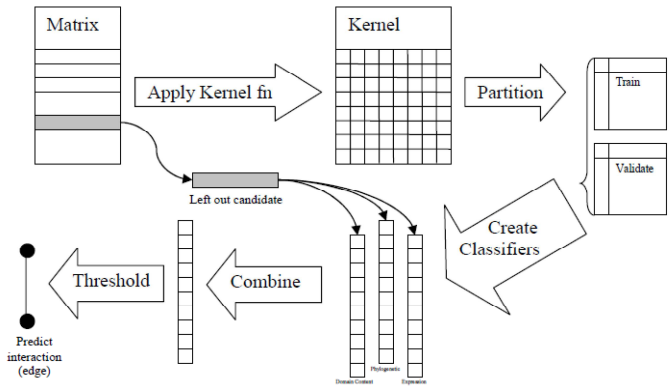


Fig. 3. Work flow of the combined kernel classifier

Data was partitioned into training and validation sets and parameters were tuned using a five fold validation strategy. The final output of the classifiers was achieved by a leave one out strategy. The classifier values were combined for a final result and a threshold was applied to determine which values are predicted edges (Figure published in Iacucci *et al.* 2011).

The kernel function used by Iacucci et al. (2011) measures the similarity between two proteins *A* and *B* (K(*A*,*B*)), one a candidate receptor *A* and the other a candidate ligand *B*. The LS-SVM classifier produced by Iacucci et al. (2011) is a binary predictor which assigns new examples in "interacting" or "non-interacting" classes. Creating the kernels from the various data matrices involved trials with different kernel functions, with linear functions ultimately being found to give the best performance in all cases. Data was partitioned into training and validation sets and parameters were tuned using a five fold validation strategy The final output of the classifiers was achieved by a leave-one-out strategy. The classifier values were scaled (minimum set to zero, maximum set to one). The values were then

combined, as defined in (1), for a final result. Figure 3 provides an overview of the workflow as described above.

$$g_{comb}(x) = \frac{g_{phylo}(x) + g_{\exp}(x) + g_{dom}(x)}{3} \quad (1)$$

3.3 Results and discussion

The comparison of the phylgenetic based method of Gertz et al. (2003) and the combined kernel classifer method of Iacucci et al. (2011) provides a clear perspective on the advantages of multiple kernel learning in the PPI prediction task. As both groups use the same dataset and have results which can be summarized and contrasted using recall, precision, and the *F*-measures.

The Iacucci et al. (2011) predictions for the tgfβ family accurately reconstructed over 76% of the supported edges (0.76 recall and 0.67 precision) of the know DLRP receptor-ligand pairs. In this case, the combined kernel classifier was able to relatively improved upon the Gertz et al. (2003) work by a factor of approximately two as the Gertz et al. (2003) work reconstructs 44% of the supported edges (0.44 recall and 0.53 precision) of the know DLRP receptor-ligand pairs. Comparing *F*-measures, we see that the combined kernel classifer method improved upon that of Gertz et al. (2003) significantly as the Iacucci et al. (2011) method has an *F*-measure of 0.71 while that of Gertz et al. (2003) has a value of 0.48.

The Iacucci et al. (20011) predictions for the chemokine family accurately reconstructed over 65% of the supported edges (0.65 recall and 0.23 precision) of the know DLRP receptor-ligand pairs. In this case, the combined kernel classifier was able to relatively improved upon the Gertz et al. (2003) work by a factor of approximately three as the Gertz et al. (2003) work reconstructs 22% of the supported edges (0.22 recall and 0.37 precision) of the know DLRP receptor-ligand pairs. Comparing *F*-measures, we see that the combined kernel classifer method improved upon that of Gertz et al. (2003) significantly as the Iacucci et al (2011) method has an *F*-measure of 0.33 while that of Gertz et al. (2003) has a value of 0.27.

Qualitatively, the performance of the Iacucci et al (2011) method also seems to be matching the performance of Gertz et al. (2003), as the novel interaction of CCR1 with SCY11 [Gao et al. 1996] reported in their work is also discovered using Iacucci et al (2011) method.

The comparison of the results of the two methods discussed here support the notion that kernel learning presents a useful methodology for elucidating receptor-ligand pairings. The benefits of the combined kernel classifier method over the Gertz et al. (2003) method are clear. Foremost in the advantages are the ability to predict multiple ligands for one receptor, which represents an necessary feature for receptor-ligand research. Also, as the classifier output is continuous, the results can be considered to be prioritized, this presents a major convenience to researchers as often the set of candidate ligands are large and financial and time resources to validate few.

4. Conclusion

The task of PPI prediction is a difficult and important area of bioinformatics research. As the number of possible interacting protein pairs in the cell is huge, wet-lab experimentation

validation of all of them is essentially impossible. In addition to being time consuming, in-vivo validation costs are also a consideration. Having a computational method for predicting PPI is therefore a necessary tool for researchers.

Several groups have addressed the PPI prediction task. While several have used phylogenetics to solve the problem, others have used physical protein structures and amino-acid sequence information to assist in making the predictions. We have reviewed these methods and discussed the key differences among them.

Methods, which rely on the physical structure and the conservation of amino-acid sequences in partners of interactions that are already reliably known to exist, also give researchers additional insight to function prediction as the methods are based on known examples. The drawback of these methods is that one has to have a known example for a comparison, which is not always the case when researching candidate receptor-ligand pairs.

Methods which rely on phylogenetic histories to determine PPI are based on a well-establish rational which holds that as interacting proteins co-evolve, there phylogenetic histories should be similar. This explains why the methods which rely on phylogenetic information are largely based on measures of similarity for the phylogenetic trees of interacting protein partners.

The advantage of using multiple kernel learning to predict PPI is apparent when using multiple sources of data. Many of the methods, mentioned above, rely on an ever growing amount of publicly available data. The ever expanding amount of high throughput data which continues to become available to the bioinformatics community represents an excellent opportunity to enhance the kernel classifier method presented in Iacucci et al. (2011).

A practical advantage of using multiple data sources allows one to extend the method as new and higher quality sources become available. For example, if better micro-array dataset becomes available in the future, it is an advantage to be able remove the existing expression-based kernel with one derived from the new dataset without having to the retrain a global classifier. Likewise, if additional data sources become available, adding an additional sub-classifier based on the new data source would take less time to train than adding the data source and retraining the global classifier.

Looking forward many exciting challenges remain to be addressed in this field. While the task of PPI is daunting and complex, the work reviewed above demonstrates that it is also rich with opportunities for improvement and further development.

5. Acknowledgments

Funding: The authors would like to acknowledge support from:

- Research Council KUL:
  ProMeta, GOA Ambiorics, GOA MaNet, CoE EF/05/007 SymBioSys en KUL PFV/10/016 SymBioSys , START 1, several PhD/postdoc & fellow grants
- Flemish Government:
  FWO: PhD/postdoc grants, projects , G.0318.05 (subfunctionalization), G.0553.06 (VitamineD), G.0302.07 (SVM/Kernel), research communities (ICCoS, ANMMM, MLDM); G.0733.09 (3UTR); G.082409 (EGFR)
  IWT: PhD Grants, Silicos; SBO-BioFrame, SBO-MoKa, TBM-IOTA3

FOD:Cancer plans
IBBT
- Belgian Federal Science Policy Office: IUAP P6/25 (BioMaGNet, Bioinformatics and Modeling: from Genomes to Networks, 2007-2011) ;
-EU-RTD: ERNSI: European Research Network on System Identification; FP7-HEALTH CHeartED

6. References

Aloy, P., & Russell, R. B. (2002). Interrogating protein interaction networks through structural biology. Proceedings of the National Academy of Sciences of the United States of America, 99(9), 5896-901. doi:10.1073/pnas.092147999

Aloy, P., Ceulemans, H., Stark, A., & Russell, R. B. (2003). The relationship between sequence and interaction divergence in proteins. Journal of molecular biology, 332(5), 989-98. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/14499603

Berman, H. M. (2000). The Protein Data Bank. Nucleic Acids Research, 28(1), 235-242. doi:10.1093/nar/28.1.235

Bhardwaj N, Lu H: Correlation between gene expression profiles and protein-protein interactions within and across genomes. Bioinformatics 2005, 21:2730-2738.

Bleakley K, Yamanishi Y: Supervised prediction of drug-target interactions using bipartite local models. Bioinformatics 2009, 25:2397-2403.

Boeckmann, B. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Research, 31(1), 365-370. doi:10.1093/nar/gkg095

Claverie, J. M. (2001). Gene number. What if there are only 30,000 human genes? Science (New York, N.Y.), 291(5507), 1255-7. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11233450

Dafas, P., Bolser, D., Gomoluch, J., Park, J., & Schroeder, M. (2004). Using convex hulls to extract interaction interfaces from known structures. Bioinformatics (Oxford, England), 20(10), 1486-90. doi:10.1093/bioinformatics/bth106

Fryxell, K. J. (1996). The coevolution of gene family trees. Trends in Genetics, 12(9), 364-369. doi:10.1016/S0168-9525(96)80020-5

Gao JL, Sen AI, Kitaura M, Yoshie O, Rothenberg ME, Murphy PM, Luster AD: Identification of a mouse eosinophil receptor for the CC chemokine eotaxin. Biochem Biophys Res Commun 1996, 223:679-684.

Ge H, Liu Z, Church GM, Vidal M: Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. Nat Genet 2001, 29:482-486.

Gertz J, Elfond G, Shustrova A, Weisinger M, Pellegrini M, Cokus S, Rothschild B: Inferring protein interactions from phylogenetic distance matrices. Bioinformatics 2003, 19:2039-2045.

Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y. L., et al. (2003). A protein interaction map of Drosophila melanogaster. Science (New York, N.Y.), 302(5651), 1727-36. doi:10.1126/science.1090289

Goh, C. S., Bogan, A. A., Joachimiak, M., Walther, D., & Cohen, F. E. (2000). Co-evolution of proteins with their interaction partners. Journal of molecular biology, 299(2), 283-93. doi:10.1006/jmbi.2000.3732

Gong, S., Yoon, G., Jang, I., Bolser, D., Dafas, P., Schroeder, M., Choi, H. H., et al. (2005). PSIbase: a database of Protein Structural Interactome map (PSIMAP). Bioinformatics (Oxford, England), 21(10), 2541-3.

doi:10.1093/bioinformatics/bti366
Graeber TG, Eisenberg D: Bioinformatic identification of potential autocrine signaling loops in cancers from gene expression profiles. Nat Genet 2001, 29:295-300.
Grigoriev A: A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast Saccharomyces cerevisiae. Nucleic Acids Res 2001, 29:3513-3519.
Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Clarke L et al.: Ensembl 2009. Nucleic Acids Res 2009, 37:D690-D697.
Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L et al.: InterPro: the integrative protein signature database. Nucleic Acids Res 2009, 37:D211-D215.
Izarzugaza, J. M. G., Juan, D., Pons, C., Ranea, J. a G., Valencia, A., & Pazos, F. (2006). TSEMA: interactive prediction of protein pairings between interacting families. Nucleic acids research, 34(Web Server issue), W315-9. doi:10.1093/nar/gkl112
Jacob L, Vert JP: Protein-ligand interaction prediction: an improved chemogenomics approach. Bioinformatics 2008, 24:2149-2156.
Keskin, O., Ma, B., & Nussinov, R. (2005). Hot regions in protein--protein interactions: the organization and contribution of structurally conserved hot spot residues. Journal of molecular biology, 345(5), 1281-94. doi:10.1016/j.jmb.2004.10.077
Keskin, O., Tsai, C.-J., Wolfson, H., & Nussinov, R. (2004). A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications. Protein science : a publication of the Protein Society, 13(4), 1043-55. doi:10.1110/ps.03484604
Kim S, Yoon J, Yang J, Park S: Walk-weighted subsequence kernels for protein-protein interaction extraction. BMC Bioinformatics 2010, 11:107.
Kim, W. K., Bolser, D. M., & Park, J. H. (2004). Large-scale co-evolution analysis of protein structural interlogues using the global protein structural interactome map (PSIMAP). Bioinformatics (Oxford, England), 20(7), 1138-50. doi:10.1093/bioinformatics/bth053
Mewes, H. W., Frishman, D., Güldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., et al. (2002). MIPS: a database for genomes and protein sequences. Nucleic acids research, 30(1), 31-4. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=99165&tool=pmcentrez&rendertype=abstract
Miwa M, Saetre R, Miyao Y, Tsujii J: Protein-protein interaction extraction by leveraging multiple kernels and parsers. Int J Med Inform 2009, 78:e39-e46.
Moyle, W. R., Campbell, R. K., Myers, R. V., Bernard, M. P., Han, Y., & Wang, X. (1994). Co-evolution of ligand-receptor pairs. Nature, 368(6468), 251-5. doi:10.1038/368251a0
Nagamine N, Sakakibara Y: Statistical prediction of protein chemical interactions based on chemical structure and mass spectrometry data. Bioinformatics 2007, 23:2004-2012.
Ogmen, U., Keskin, O., Aytuna, a S., Nussinov, R., & Gursoy, a. (2005). PRISM: protein interactions by structural matching. Nucleic Acids Research, 33(Web Server), W331-W336. doi:10.1093/nar/gki585
Park, J., Lappe, M., & Teichmann, S. A. (2001). Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. Journal of molecular biology, 307(3), 929-38. doi:10.1006/jmbi.2001.4526
Pazos, F., & Valencia, A. (2001). Similarity of phylogenetic trees as indicator of protein-protein interaction. Protein engineering, 14(9), 609-14. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11707606

Pitre, S., Dehne, F., Chan, A., Cheetham, J., Duong, A., Emili, A., Gebbia, M., et al. (2006). PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs. BMC bioinformatics, 7, 365. doi:10.1186/1471-2105-7-365

Rubin, G. M. (2001). The draft sequences. Comparing species. Nature, 409(6822), 820-1. doi:10.1038/35057277

Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., & Eisenberg, D. (2004). The Database of Interacting Proteins: 2004 update. Nucleic acids research, 32(Database issue), D449-51. doi:10.1093/nar/gkh086

Sato T, Yamanishi Y, Kanehisa M, Toh H: The inference of protein-protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. Bioinformatics 2005, 21:3482-3489.

Shatsky, M., Nussinov, R., Wolfson, H., Guigó, R., & Gusfield, D. (2002). Algorithms in Bioinformatics. (R. Guigó & D. Gusfield, Eds.) (Vol. 2452, pp. 235-250). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/3-540-45784-4

Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke,M.P., Walker,J.R., Hogenesch,J.B: A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A 2004, 101:6062-6067.

Suykens JA, Vandewalle J, De MB: Optimal control by least squares support vector machines. Neural Netw 2001, 14:23-35.

Tan, S.-H., Zhang, Z., & Ng, S.-K. (2004). ADVICE: Automated Detection and Validation of Interaction by Co-Evolution. Nucleic acids research, 32(Web Server issue), W69-72. doi:10.1093/nar/gkh471

Thompson JD, Higgins DG, Gibson TJ: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994, 22:4673-4680.

Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic acids research, 22(22), 4673-80. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=308517&tool=pmcentrez&rendertype=abstract

Tillier, E. R. M., Biro, L., Li, G., & Tillo, D. (2006). Codep : Maximizing Co-Evolutionary Interdependencies to Discover Interacting Proteins, 831(December 2005), 822- 831. doi:10.1002/prot

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., et al. (2000). A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature, 403(6770), 623-7. Macmillian Magazines Ltd. doi:10.1038/35001009

van Kesteren, R. E., Tensen, C. P., Smit, A. B., van Minnen, J., Kolakowski, L. F., Meyerhof, W., Richter, D., et al. (1996). Co-evolution of ligand-receptor pairs in the vasopressin/oxytocin superfamily of bioactive peptides. The Journal of biological chemistry, 271(7), 3619-26. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8631971

Chapter 3

Study 1: LS-SVM Prediction of Receptor–Ligand Binding

As in the previous chapter we have reviewed PPI prediction methods, we now turn our attention to the specific task of making receptor─ligand pairing predictions. This chapter represents contemporary work, conducted in the area of data fusion and kernel learning. In this work, the existing line of research is extended by using disparate data sources to construct a kernel to train a least square support vector machines (LS-SVM) in order to classify candidate receptors and ligands as interacting or non-interacting. The objective of this chapter is is demonstrate an improvement over the work of Gertz et al. This chapter contains work which is based on three data sources (phylogenetic, domain, and expression), in the subsequent chapter, we extend this to six data sources and several more machine learning techniques.

Publication Information: This work is published at: Iacucci E., Ojeda F., De Moor B., Moreau Y., "Predicting Receptor–ligand Pairs through Kernel Learning", *BMC Bioinformatics*, , vol. 12, no. 336, 2011, pp. 1-8.

Personal Contribution: The PhD candidate is wholely responsible for all work presented here. The second author provided extremely helpful guidance and discussion concerning the experimental design and was consulted when necessary in the revision process.

Iacucci *et al. BMC Bioinformatics* 2011, **12**:336
http://www.biomedcentral.com/1471-2105/12/336

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Predicting receptor-ligand pairs through kernel learning

Ernesto Iacucci, Fabian Ojeda, Bart De Moor and Yves Moreau*

Abstract

Background: Regulation of cellular events is, often, initiated via extracellular signaling. Extracellular signaling occurs when a circulating ligand interacts with one or more membrane-bound receptors. Identification of receptor-ligand pairs is thus an important and specific form of PPI prediction.

Results: Given a set of disparate data sources (expression data, domain content, and phylogenetic profile) we seek to predict new receptor-ligand pairs. We create a combined kernel classifier and assess its performance with respect to the Database of Ligand-Receptor Partners (DLRP) 'golden standard' as well as the method proposed by Gertz *et al.* Among our findings, we discover that our predictions for the tgfβ family accurately reconstruct over 76% of the supported edges (0.76 recall and 0.67 precision) of the receptor-ligand bipartite graph defined by the DLRP "golden standard". In addition, for the tgfβ family, the combined kernel classifier is able to relatively improve upon the Gertz *et al.* work by a factor of approximately 1.5 when considering that our method has an *F*-measure of 0.71 while that of Gertz *et al.* has a value of 0.48.

Conclusions: The prediction of receptor-ligand pairings is a difficult and complex task. We have demonstrated that using kernel learning on multiple data sources provides a stronger alternative to the existing method in solving this task.

Background

Regulation of cellular events is initiated, often, via extracellular signaling. Extracellular signaling occurs when a circulating ligand interacts with one or more membrane-bound receptors. Identification of receptor-ligand pairs is thus an important and specific form of protein-protein interaction (PPI) prediction. While the problem of predicting PPI has been highly studied, little effort has been placed on the sub-problem of predicting receptor-ligand interactions.

A tremendous amount of research has been applied to the problem of predicting PPI. Foremost in the field has been prediction via phylogenetic profile analysis. Generally, this type of investigation studies the similarity of the phylogenetic history of a protein *A* and its putative protein partner *B*. Broadly speaking, the assortment of these types of phylogenetic studies examine the most accurate measure of similarity between the phylogenetic histories of *A* and *B*. Findings from these studies support the idea that proteins which interact have similar phylogenetic profiles, as these proteins should co-adapt as they are under the same evolutionary pressures [1,2].

Bhardwaj *et al.* [3] make use of the phylogenetic information strategy while introducing expression data to predict PPI. Their findings support the idea that integrating gene expression profile and phylogenetic information increases the accuracy of predictions than phylogenetic analysis alone. The rational of using co-expression as an indicator of PPI originates from the observation that proteins which interact for the purpose of performing a similar function are likely to be co-expressed as they will need to be present at the same time to carry out their common biological activity [4,5].

The notion of combining expression and phylogenetic information to predict PPI is clearly a step in a direction which leads us to consider a wider variety of data integration. Here we propose a framework in which other sources of data (such as domain content) can be applied to a kernel solution to the problem. The rational behind incorporation of domain content information is as

* Correspondence: yves.moreau@esat.kuleuven.be
SCD-ESAT, Department of Electrical Engineering, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, Leuven, 3001, Belgium

BioMed Central

© 2011 Iacucci et al; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

follows: as certain domains are known to interact, it is self evident that this data would provide additional insight into the problem of determining receptor-ligand pairs.

One of the groups which have tackled the receptor-ligand prediction task is Gertz *et al.* [6]. In their work, they match members of the chemokines and tgfβ ligand families with their respective receptor families. They used distance matrices of the receptors and ligands families to measure similarity between the groups. Through a Metropolis Monte Carlo optimization algorithm, they explored and scored possible matches between the two matrices, until they reached optimal solutions. While their work was successful, they rely only on phylogenetic distance matrices, here we propose the integration of multiple data sources to help make more accurate matches.

We look into the use of creating a combined kernel classifier to carry out this learning task. While many kernel-based machine learning techniques have been applied to the PPI task [7,8], it has hitherto never, to our knowledge, been used on the receptor-ligand problem. Kernel learning provides the means to utilize enigmatically related data (such as expression measures, domain content, etc.) and perform classification in higher dimensional space via kernel methods. In our work, we apply the least-squares support vector machines (LS-SVM) method based on the conclusions by Suykens *et al.* [9] which shows this implementation to be robust. As different data sources are used, separate LS-SVM kernel classifiers were built and the combined output used to provide a final result.

While the task addressed here is the predictions of successful protein ligand-receptor pairings, a related area of research is the protein-chemical interaction prediction task for which kernels have, sometimes, been applied. For example, Nagamine *et al.* [10] approach this task through the use of a SVM trained on vector representations of protein-chemical pairs. Building on this, Jacob *et al.* [11] demonstrate the utility of using hierarchical kernels to match proteins with chemical ligands in a similar learning task. This line of research was then further advanced by Bleakley *et al.* [12] who introduce the use *of bipartite local models* which use kernels to successfully predict several reported drug-target interactions.

We first describe our combined kernel classifier method to predict receptor-ligand pairings. We then present the bipartite-graph we derive from our findings and compare it to our "Golden Standard" and to results previously published by Gertz *et al.* [6]. Following this, we interpret the performance of our method with respect to this comparison. To conclude, we discuss the benefits and limitations of our method and possible future directions for this work.

Methods

1. Problem Formulation

Our objective is to predict candidate receptor-ligand pairs; more specifically, we seek to create a method to identify known pairs as well as to determine putative pairs for further research. Our method involves using multiple data sources (expression, phylogenetic, and protein-domain content information), computing separate kernels for each data type, creating LS-SVM classifiers and combining the results to predict receptor-ligand pairs.

2. Data Sources

For the datasets used, our setting is as follows, candidate receptor and ligand sequences were retrieved for seven species (*Rattus norvegicus, Mus musculus, Homo sapiens, Pan troglodytes, Canis familiaris, Cavia porcellus*, and *Bos taurus*) from ensemble build 51 [13]. The sequences were then aligned using ClustalW [14]. Once aligned, the sequences were edited so as to eliminate the positions which had the 5% lowest substitution scores across the seven orthologous sequences. The pair-wise alignment score was then taken for each possible species to species comparison between the edited orthologous sequences (as seven species are used, a total of 21 pairwise comparisons for each candidate is created). The distance scores form a phylogenetic vector [2] which will then be used to create the phylogenetic kernel.

The expression for the candidates was taken from the well-known GNF human expression atlas (79 tissues) [15], the data was normalized (values were mean-zeroed and the standard deviation was set to one) and was further transformed into the expression kernel.

The domain content of each candidate protein (receptor or ligand) was taken from the Interpro Database [16]. A vector for each candidate protein was created where the presence of a protein domain was indicated with a'1' and the absence of a domain was indicated by a '0'. This data was then transformed to create the domain content kernel.

3. Kernels and LS-SVM Classifier

The above mentioned data matrices (phylogenetic, expression, and domain content) were used to create three kernels for each receptor-ligand family. LS-SVMs [9] were trained using the three kernels to predict outcomes for receptor-ligand pairs known from our Database of Ligand-Receptor Partners (DLRP) "Golden Standard".

Our kernel function measures the similarity between two proteins A and B (K(A,B)), one a candidate receptor (A) and the other a candidate ligand (B). Our LS-SVM classifier is a binary predictor which assigns new examples in "interacting" or "non-interacting" classes. Creating the kernels from these matrices involved trials with different kernel functions (radial based function, linear, and

polynomial), linear functions being found to give the best performance in all cases. A combined kernel approach was also considered but empirical results determined that a combined classifier approach was preferable. The regularization parameters for the LS-SVMs were tuned using a two tier grid search which, at first, uniformly ranged from 10^{-6} to 10^{6} in 10^{1} unit steps followed by a second finer search with $10^{0.1}$ unit steps. For each candidate, data was partitioned into training and validation sets and parameters were tuned using a 5-fold validation strategy (300 random partitions of the data were performed). The final output of the classifiers was achieved by a leave-one-out strategy. The classifier values were scaled (minimum set to zero, maximum set to one) and combined, as defined in (1), for a final result. Figure 1 provides an overview of the workflow as described above.

$$f_{comb}(x) = \frac{f_{phylo}(x) + f_{\exp}(x) + f_{dom}(x)}{3} \quad (1)$$

4. Construction of the Receptor-Ligand Bipartite Graph

We take as our "Golden Standard" the receptor-ligand dataset from the The Database of Ligand-Receptor Partners (DLRP) [17]. In this dataset, cytokines and interleukins (as well as other ligands) are taken and paired with their corresponding receptor partners. These interactions are then represented in an adjacency matrix where an interaction is represented as a '1' and lack of interaction is represented as a '0'. These are the values we are ultimately trying to predict.

In order to compare the pairings predicted by the combined kernel classifier, we compared the known bipartite receptor-ligand graph (constructed from the known DLRP values) with the predictions from [6] and from the combined kernel classifier. As the combined classifier values are continuous and known values (from DLRP) are binary, it is necessary to determine a threshold value t to distinguish between the two classes. The thresholds for each ligand are considered and evaluated as follows. Edges between receptor and ligands are assigned based on the decision function defined in (2), the predicted edge set is then compared to the "*golden standard*" and the precision and recall are calculated. The optimal threshold t is then determined by taking the average classifier value of the maximal "*F-measure*" (3) for each ligand. The bipartite graph is then constructed using the edge set resulting from the optimal

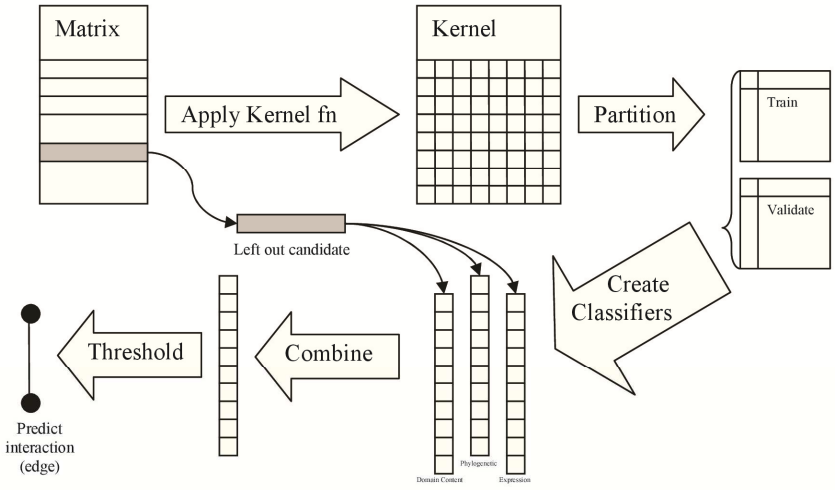


Figure 1 Work flow of the combined kernel classifier. For each candidate, Data was partitioned into training and validation sets and parameters were tuned using a 5 fold validation strategy. The final output of the classifiers was achieved by a leave one out strategy. The classifier values were combined for a final result and a threshold was applied to determine which values are predicted edges in the receptor-ligand bipartite graph.

Table 1 Chemokine receptor-ligand predictions

	Gertz et al.(2003)		Iacucci *et al.*	
Receptor	**Supported**	**Unsupported**	**Supported**	**Unsupported**
CCR3	SCYA24	–	SCYA26 SCYA7 SCYA5 SCYA15 SCYA11 SCYA13	SCYA3 SCYA17 SCYB6 SCYA4 SCYA21 SCYB5 IL8 SCYA27
CCR1	SCYA2 SCYA8 SCYA13 SCYA7	SCYA11 SCYA1	SCYA3 SCYA7 SCYA8 SCYA5 SCYA15 SCYA13	SCYA26 SCYA17 SCYB6 SCYA4 SCYA21 SCYB5 IL8 SCYA11 SCYA27
CCR5	SCYA3 SCYA4 SCYA5	SCYA14 SCYA15 SCYA23	SCYA3 SCYA4 SCYA5	SCYA26 SCYA17 SCYB6 SCYA21 SCYA15 SCYB5 IL8 SCYA11 SCYA27
IL8RA	SCYB6 SCYB5	GRO1 GRO2 GRO3 PPBP	SCYB6 IL8	SCYA17 SCYA4 SCYA21 SCYA15 SCYA13
CCR4	–	SCYA26	SCYA17	SCYA3 SCYA26 SCYA7 SCYA8 SCYB6 SCYA21 SCYB5 SCYA11 SCYA13 SCYA27
CCR2	–	SCYA21 SCYA19	–	SCYA3 SCYB6 SCYA4 SCYA21 SCYA5 SCYA15 SCYB5 SCYA11
CCR8	SCYA17	SCYA22	–	SCYA3 SCYA26 SCYA7 SCYA8 SCYB6 SCYB5 IL8 SCYA11 SCYA27

Table 1 Chemokine receptor-ligand predictions *(Continued)*

CXCR3, GPR9	–	SDF1	SCYA21 SCYA11	SCYA3 SCYA26 SCYA17 SCYA7 SCYA8 SCYB6 SCYA4 SCYA15 SCYB5 IL8 SCYA13 SCYA27
IL8RB	IL8	–	SCYB6 SCYB5 IL8	SCYA3 SCYA17 SCYA7 SCYA8 SCYA4 SCYA21 SCYA5 SCYA15 SCYA11 SCYA13 SCYA27
BLR1, CXCR5	–	MIG SCYB10 SCYB11	–	–
CCBP2, CCR9	SCYA25	–	–	SCYB6 SCYA21
CCR6	–	SCYB13	–	SCYA26
CXCR4	–	SCYA27	–	–
CCR7	–	SCYA20	–	SCYA3 SCYA26 SCYA8 SCYA5 SCYA15 SCYB5

Table of chemokine receptors and their predicted ligand partners, both supported and unsupported partners are shown in both Gertz *et al.* (2003) and the Iacucci *et al.* predictions.

threshold.

$$E_{pq} = \begin{cases} 1 \; f_{comb} \geq t \\ 0 \; f_{comb} < t \end{cases} \tag{2}$$

$$F - measure = \frac{2 \times precision \times recall}{precision + recall} \tag{3}$$

Results and Discussion

The known tgfβ receptor-ligand set used by Gertz *et al.* [6] consists of 18 known matches. Gertz *et al.* [6] predicted 15 edges, 8 of which were supported and 7 of which were unsupported. In contrast, our tgfβ set consists of 79 known edges, and our approach predicts 90 edges, 60 of which were correct and 30 of which were unsupported. The detailed pairings for this family are shown in Table 1.

We discover that our predictions for the tgfβ family accurately reconstruct over 76% of the supported edges (0.76 recall and 0.67 precision) of the receptor-ligand bipartite graph defined by the DLRP. In addition, the combined kernel classifier is able to relatively improved upon the Gertz *et al.* [6] work by a factor of approximately two as the Gertz *et al.* [6] work reconstructs 44% of the supported edges (0.44 recall and 0.53 precision) of the receptor-ligand bipartite graph defined by the DLRP. For this family of receptors and ligands, there exists an advantage in our approach to make predictions as we reconstruct more known edges and introduce less noise. Comparing *F*-measures, we see that our method improved upon that of Gertz *et al.* [6] significantly as our method has an *F*-measure of 0.71 while that of Gertz *et al.* [6] has a value of 0.48.

The known chemokine receptor-ligand set used by Gertz *et al.* [6] consists of 63 known matches. Gertz *et al.* [6] predicted 38 edges, 14 of which were supported and 24 of which were unsupported. In contrast, our chemokine set consists of 53 known matches, and our approach predicts 98 edges, 22 of which were correct and 76 of which were unsupported. Our classifier was

Table 2 Tgfβ recptor-ligand predictions

	Gertz et al (2003)		Iacucci *et al*	
	Supported	**Unsupported**	**Supported**	**Unsupported**
TGFBRII	Tgfb2 Tgfb3 Tgfb1	–		
BMPRIa	–	Gdf5	BMP8 BMP4 BMP2 BMP15 BMP3 BMP6 BMP10	Tgfb2 Tgfb3 Tgfb1 INHBA INHIA INBB INHC
AMHR	–	–	–	–
BMPRIB	Bmp2 Bmp4	–	BMP8 BMP4 BMP2 BMP15 BMP3 BMP6 BMP10	Tgfb2 Tgfb3 Tgfb1 INHBA INHA INBB INHC
ACTRIIa	ActivinBA	–	BMP7 BMP5 BMP8 BMP4 BMP2 BMP15 BMP3 BMP6 BMP10 INHBA INHA INHBB INHBC	Tgfb3 Tgfb2 Tgfb1
ACTRIIb	ActivinBB	–	BMP7 BMP5 BMP8 BMP4 BMP2 BMP15 BMP3 BMP6 BMP10 INHBA INHA INHBB INHBC	Tgfb2 Tgfb3 Tgfb1
SAX	–	Bmp3	–	–
TKVR	–	Bmp10	–	–
ACTRII	–	DPP	–	–
TGFBRI	–	Bmp7 Bmp6	–	–
BMPRII	–	Gdf8	BMP8 BMP4 BMP2 BMP15 BMP3 BMP6 BMP10	Tgfb2 Tgfb3 Tgfb1 INHBA INHA INBB INHC

Table of tgfβ receptors and their predicted ligand partners, both supported and unsupported partners are shown in both Gertz *et al.* (2003) and the Iacucci *et al.* predictions.

constructed using ligands which have at least two receptor partners as this greatly improved the precision (0.67 recall and 0.12 precision when all the ligands are used in the classifier). The detailed pairings for this family are shown in Table 2.

We also find that our predictions for the chemokine family accurately reconstruct over 65% of the supported edges (0.65 recall and 0.23 precision) of the receptor-ligand bipartite graph defined by the DLRP. In addition, the combined kernel classifier is able to relatively improved upon the Gertz *et al.* [6] work by a factor of approximately three as the Gertz *et al.* [6] work reconstructs 22% supported edges (0.22 recall and 0.37 precision) of the receptor-ligand bipartite graph defined by the DLRP. While the precision of the Gertz *et al.* [6] is higher, the recall of our method is about three fold higher. Comparing *F*-measures, we see that our method improved upon that of Gertz *et al.* [6], slightly as our method has an *F*-measure of 0.33 while that of Gertz *et al.* [6], has a value of 0.27.

Qualitatively, the performance of our method also seems to be matching the performance of Gertz *et al.* [6], as the novel interaction of CCR1 with SCY11 [18] reported in their work is also discovered using our method.

The overall results presented here support the notion that kernel learning presents a useful methodology for elucidating receptor-ligand pairing. Using disparate data sources, we propose a combined kernel classifier which is able to reconstruct the majority of known edges in the chemokine and tgfβ receptor-ligand bipartite graphs. In order to evaluate our pairings, we consider the bipartite graph which we construct from our results (see Figure 2). The success of the results are summarized by two performance measures; the recall and the precision of the edges predicted in the tgfβ and chemokine bipartite graphs. The relative performance of each method examined here is evaluated using the *F*-measure.

The combined classifier performs better using the tgfβ family of receptors and ligand than using the chemokine family of receptors and ligands. This can be attributed to two reasons. Firstly, the tgfβ has more positive examples than the chemokine family to train with. Secondly, the tgfβ family is more evolutionarily related while the chemokine family is related by function. Thus, it is more difficult to learn with data from the chemokine family as there is less evolutionarily related structure inherent to the data for the LS-SVM to learn with.

The benefits of the combined kernel classifier method are clear. Foremost in the advantages are the ability to predict multiple ligands for one receptor, which

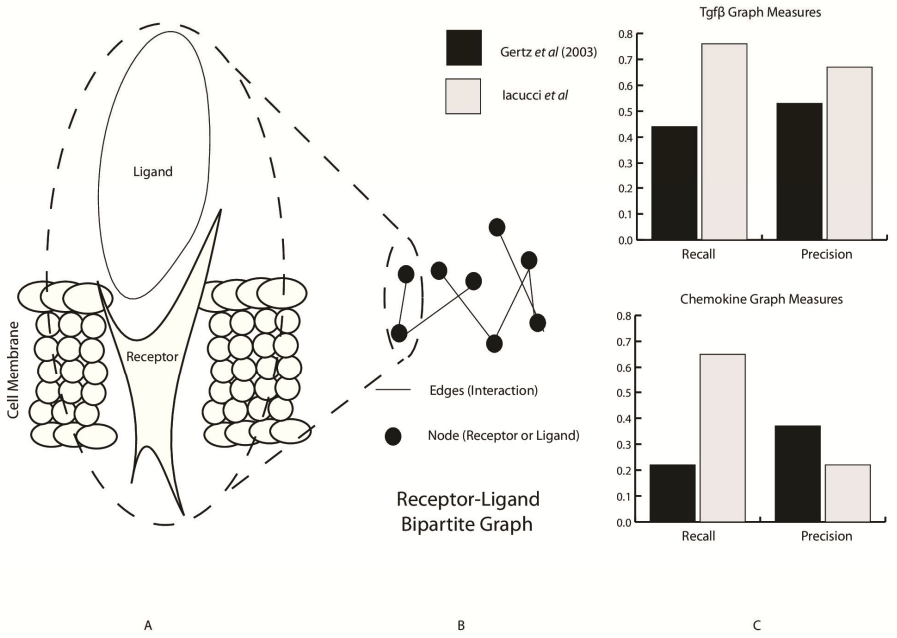


Figure 2 Schematic of receptor-ligand bipartite graph and performance measures. (a) Schematic of in-vivo receptor-ligand interaction as found interacting in the cell membrane (b) Bipartite graph schematic of the receptor-ligand interaction network (c) Performance measures of tgfβ and chemokine bipartite graph construction across Gertz *et al.* (2003) and Iacucci *et al.* methods.

represents an imperative feature for receptor-ligand research. In addition, as the classifier output is continuous, the results can be considered to be prioritized, this presents a major convenience to researchers as often the set of candidate ligands are large and resources to validate few. The major limitation of the method rests in the need to have training examples for receptor-ligands which one is trying to predict. This is particularly true for predicting the pairing in the chemokine dataset as when we consider only ligand candidates with two or more receptor pairings, the precision performance of our method improves (0.79 recall and 0.31 precision) (see Table 1).

The advantage of using the three sub-classifiers instead of a global classifier which combines all features is two fold. The first reason is that the data sources used here are disparate and heterogeneous. A global classifier would require a mapping step which may introduce some noise. The second reason is that using separate sub-classifiers would allow for removal and addition of sub-classifiers. For example, if a better micro-array dataset becomes available in the future, it would be an advantage to be able remove the existing expression-based kernel with one derived from the new dataset without having to the retrain a global classifier. Also, if additional data sources become available, adding an additional sub-classifier based on the new data source would take less time to train than adding the data source and retraining the global classifier.

A practical advantage of using three sub-classifiers in our work became apparent when considering the performance of the individual classifiers versus that of the combined kernel classifier. More specifically, the combined kernel classifier performed equally as well or better than any of the individual classifiers. In the case of the chemokine family, the performance of all three individual classifiers was not nearly as good as the combined kernel classifier. In the case of the tgfβ family, the

expression classifier performed nearly as well as the combined kernel classifier (see Additional File 1, Table S1).

Conclusions

The prediction of receptor-ligand pairings is a difficult and complex task. We have demonstrated that using multiple data sources provide an advantage over single data sources in solving this task. The use of multiple data sources allows us to extend our method as new data becomes available. Among our main contributions we count the ability of our method to prioritize candidate pairs, which represents an imperative feature for receptor-ligand research. As *in-vivo* validation is costly and time consuming, it's important that researchers have a ranking of a, potentially, large number of candidates. In addition, we provide a method which has high recall (0.76 and 0.67) and improved *F*-measures when compared to Gertz *et al.* [6] (0.71 for Iacucci *et al.* vs 0.48 for Gertz *et al.* [6] when evaluationg the tgfβ family and 0.33 for Iacucci *et al.* vs 0.27 for Gertz *et al* [6] when evaluating the chemokine family). Thus, the method is reliable in so far that it will retrieve a large portion of the true positives while not introducing too much noise. As more high throughput data becomes available, we expect to extend the current methodology to accommodate it.

Additional material

Additional file 1: Classifier Performance Measures. The performance of the individual kernel classifiers are displayed in addition to the combined kernel classifier and the Gertz *et al.* (2003) method.

Acknowledgements

The authors also give thanks to Dr. Shi Yu, Dr. Léon-Charles Tranchevent, and Dr. Anneleen Daemen for their thoughtful suggestions.

Funding: The authors would like to acknowledge support from:

*Research Council KUL: ProMeta, GOA Ambiorics, GOA MaNet, CoE EF/05/007 SymBioSys, START 1, several PhD/postdoc & fellow grants

*Flemish Government:

o FWO: PhD/postdoc grants, projects, G.0318.05 (subfunctionalization), G.0553.06 (VitamineD), G.0302.07 (SVM/Kernel), research communities (IC-CoS, ANMMM, MLDM); G.0733.09 (3UTR); G.082409 (EGFR)

o IWT: PhD Grants, Silicos; SBO-BioFrame, SBO-MoKa, TBM-IOTA3

o FOD:Cancer plans

* Belgian Federal Science Policy Office: IUAP P6/25 (Bio-MaGNet, Bioinformatics and Modeling: from Genomes to Networks, 2007-2011);

* EU-RTD: ERNSI: European Research Network on System Identification; FP7-HEALTH CHeartED

The scientific responsibility is assumed by its authors.

Authors' contributions

EI set up the experiments, analyzed the data and wrote the paper. FO participated in designing the study and provided valuable insight and advice. YM and BDM supervised the project. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Received: 21 February 2011 Accepted: 11 August 2011
Published: 11 August 2011

References

1. Izarzugaza JM, Juan D, Pons C, Ranea JA, Valencia A, Pazos F: **TSEMA: interactive prediction of protein pairings between interacting families.** *Nucleic Acids Res* 2006, **34**:W315-W319.
2. Sato T, Yamanishi Y, Kanehisa M, Toh H: **The inference of protein-protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships.** *Bioinformatics* 2005, **21**:3482-3489.
3. Bhardwaj N, Lu H: **Correlation between gene expression profiles and protein-protein interactions within and across genomes.** *Bioinformatics* 2005, **21**:2730-2738.
4. Grigoriev A: **A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast Saccharomyces cerevisiae.** *Nucleic Acids Res* 2001, **29**:3513-3519.
5. Ge H, Liu Z, Church GM, Vidal M: **Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae.** *Nat Genet* 2001, **29**:482-486.
6. Gertz J, Elfond G, Shustrova A, Weisinger M, Pellegrini M, Cokus S, Rothschild B: **Inferring protein interactions from phylogenetic distance matrices.** *Bioinformatics* 2003, **19**:2039-2045.
7. Kim S, Yoon J, Yang J, Park S: **Walk-weighted subsequence kernels for protein-protein interaction extraction.** *BMC Bioinformatics* 2010, **11**:107.
8. Miwa M, Saetre R, Miyao Y, Tsujii J: **Protein-protein interaction extraction by leveraging multiple kernels and parsers.** *Int J Med Inform* 2009, **78**: e39-e46.
9. Suykens JA, Vandewalle J, De MB: **Optimal control by least squares support vector machines.** *Neural Netw* 2001, **14**:23-35.
10. Nagamine N, Sakakibara Y: **Statistical prediction of protein chemical interactions based on chemical structure and mass spectrometry data.** *Bioinformatics* 2007, **23**:2004-2012.
11. Jacob L, Vert JP: **Protein-ligand interaction prediction: an improved chemogenomics approach.** *Bioinformatics* 2008, **24**:2149-2156.
12. Bleakley K, Yamanishi Y: **Supervised prediction of drug-target interactions using bipartite local models.** *Bioinformatics* 2009, **25**:2397-2403.
13. Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Clarke L, *et al*: **Ensembl 2009.** *Nucleic Acids Res* 2009, **37**: D690-D697.
14. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22**:4673-4680.
15. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB: **A gene atlas of the mouse and human protein-encoding transcriptomes.** *Proc Natl Acad Sci USA* 2004, **101**:6062-6067.
16. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, *et al*: **InterPro: the integrative protein signature database.** *Nucleic Acids Res* 2009, **37**:D211-D215.
17. Graeber TG, Eisenberg D: **Bioinformatic identification of potential autocrine signaling loops in cancers from gene expression profiles.** *Nat Genet* 2001, **29**:295-300.
18. Gao JL, Sen AI, Kitaura M, Yoshie O, Rothenberg ME, Murphy PM, Luster AD: **Identification of a mouse eosinophil receptor for the CC chemokine eotaxin.** *Biochem Biophys Res Commun* 1996, **223**:679-684.

doi:10.1186/1471-2105-12-336
Cite this article as: Iacucci *et al.*: Predicting receptor-ligand pairs through kernel learning. *BMC Bioinformatics* 2011 **12**:336.

Supplementary Table 1: Classifier Performance Measures:

	Measure	*Domain*	*Expression*	*Phylogenetic*	*Combined*	*Gertz*
Chemokine	Recall	0.34	0.39	0.36	0.64	0.22
Family	Precision	0.17	0.11	0.06	0.22	0.37
	F-Measure	0.23	0.17	0.10	0.33	0.27
Tgfβ	Recall	0.59	0.79	0.65	0.76	0.44
Family	Precision	0.75	0.64	0.61	0.66	0.53
	F-Measure	0.66	0.70	0.63	0.71	0.48

The performance of the individual kernel classifiers are displayed in addition to the combined kernel classifier and the Gertz *et al*. (2003) method.

Chapter 4

Study 2: Benchmarking of Machine Learning Methods on the Receptor–Ligand Binding Prediction Task

This chapter represents work which deals with the receptor–ligand prediction task with respect to the various machine learning methods available to make the PPI prediction. Beyond the work presented in the pervious chapter, we now use additional data sources, machine learning techniques, and extend our dataset to the whole DLRP dataset. The objective of this study is to provide the best knowledge under our setting (such as data sources, parameters, and workflow), necessary to perform the *in-silico* prediction carried out in the subsequent chapter where we use literature based data to provide a realistic candidate list.

Publication Information: This work is under review at BMC Bioinformatics: Iacucci E., Popovic E., Tranchevent L-C., De Moor B., Moreau Y., "Critical Assessment of Receptor–Ligand Predictions", Internal Report 11-207, ESAT-SISTA, K.U.Leuven (Leuven, Belgium),

Personal Contribution: The PhD candidate is responsible for designing the experiment, data processing, technical troubleshooting, analysis of experimental results, and writing the paper. The second and third authors provided data collection, experimental design ideas, implementation of matlab routines, and extremely helpful discussion concerning the whole project.

Critical assessment of machine learning methods for receptor-ligand predictions

E. Iacucci [1], D. Popovic[1], L.-C. Tranchevent[1], B. De Moor[1], and Y. Moreau[1,†]

EI: ernesto.iacucci@gmail.com
DP: Dusan.Popovic @esat.kuleuven.be
LCT: Leon-Charles.Tranchevent @esat.kuleuven.be
BDM: bart.demoor@rec.kuleuven.be
YM: yves.moreau@east.kuleuven.be

[1]ESAT-SCD / IBBT Future Health Department, KU Leuven, Leuven, Belgium

ABSTRACT

† Corresponding author: yves.moreau@esat.kuleuven.be

ABSTRACT

Background: Intercellular communication is mediated by the interaction of circulating ligands and cellular receptors. These interactions initiate important biological processes such as tissue development, homeostasis, and stress response. Identification of receptor-ligand pairs is thus an important task, facilitated by computational prediction. We consider that all protein pairs may be assigned to one of two classes, the *interacting* class and the *non-interacting* class. All proteins have measurements derived from various data sources such as expression profiles and sequence information. The association between the measurements of two proteins can be captured by various similarity measures. As several data sources, similarity measures, and classifiers exist, it's not clear which combination will function best at addressing the interaction prediction task described above. In this study, we benchmark the performance of several popular classifiers across several different data sources and similarity measures to find the best combination for this prediction task.

Results: Through benchmarking of Database of Ligand-Receptor Partners (DLRP) protein receptor-ligand dataset across several classifiers we find the best performing classifier to be the random forest with a sensitivity of 0.87, a specificity of 0.80, and a balanced accuracy of 0.84. The most informative data source is found to be Kegg as its removal from the trials causes a drop in performance of about 15%. The most useful similarity measure, when the random forest is used, is found to be the absolute cosine function.

Conclusions: The results from this work suggest several key findings. First, there exists a major advantage in balancing the training data, particularly with the lowest performing classifiers, suggesting that they are sensitive to imbalanced data. Second, there is a high amount of mutual dependence between some data sources, suggesting that they are not all necessary for the interaction prediction task. Finally, the best performing data source is Kegg, although it seems that combining the other sources of information also works well.

1 INTRODUCTION

Intercellular communication is mediated by the interaction of circulating ligands and cellular receptors. These interactions initiate important biological processes such as tissue development, homeostasis, and inflammation (Kenakin and Miller, 2010; Zhang et al., 2009; Jensen et al., 2008; Ganguli et al., 1998; Sugg et al., 1995). Problems associated with aberrations in the receptor-ligand signaling process can lead to serious pathologies such as cancer, diabetes, and Murk Jansen Metaphysyeal Chrondrodysplasia (Kenakin and Miller, 2010; Traynelis et al., 2010; Jensen et al., 2008; Schipani et al., 1995; Dingledine et al., 1988). Identification of receptor-ligand pairs is thus an important task, facilitated by computational approaches (Iacucci et al., 2011; Gertz et al., 2003).

Proteins, such as receptors and ligands, have numerous measurements drawn from various data sources. For example, proteins are encoded by genes which have expression measurements which vary with time or condition (expression profiles). A ligand *A* and it's corresponding receptor *B* have individual measurements, such as a vector representation of their expression profiles (as the two proteins interact, it's intuitive that the two profiles should be similar as, in order to perform their biological function, they should be present at the same time in the cell). Applying a similarity measure to the two individual profiles will generate a similarity measure for the pair (intuitively, this measure should be higher for known interacting protein pairs) which can then be used to train a classifier (see Figure 1).

There exists a variety of classifiers. For example, *k*-Nearest Neighbor classifiers (kNN) are a method by which objects are classified based on their proximity to the closest training example, as found in the feature space. A Decision Tree (DT), another classification method, is a predictive model used to map observations of features into a tree structure so as to be able to make conclusions about the class assignments. Each node in the tree searches for a feature which best separates all examples trained at that node (Darnell et al., 2007). Random forests (RF) classifiers are a collection of decision trees generated from randomly sampling the features in the training set. Assignments are made when all trees "vote" for the one class or another, and the class assignment is subsequently made to the most popular class (Qi et al., 2005).

The general task of predicting protein-protein interaction (PPI) using machine learning techniques has already been examined in the past (Guarracino, 2010). For instance, Qi *et al.*, propose the use of random forests followed by a modified *k*-Nearest Neighbor algorithm to predict whether proteins pairs do or do not interact. Similarly, Gurracino *et al.*, have used kNN to predict PPI. Their results reveal that the algorithm performs with high prediction accuracy.

In this study, we propose a benchmark of machine learning techniques in the context of receptor-ligand prediction. These benchmarks are usually very complex since various techniques and parameters have to be taken into account. In addition, and as several data sources and similarity measures exist, it's not clear which combination will function best at addressing the receptor-ligand interaction prediction task described above. In the cur-

rent study, we benchmark the performance of six popular classifiers across six different data sources and four similarity measures to identify the best performing combination.

2 METHODS

In the first step, data is retrieved from different data sources. This data usually links proteins to various features and describes what is currently known about these proteins (section 2.1). In a second step, similarity matrices are computed using several similarity measures (section 2.2). In a third step, classification algorithms are benchmarked using a combination of similarity matrices from distinct data sources (sections 2.3 and 2.4).

2.1 Data sources

Several sources are considered to extract protein data that can be relevant for our prediction task. In particular, we retrieve expression profiles, curated pathway information, and sequence data. In addition, the sequence data is analyzed using different strategies as described below. While many more data sources are publicly available, we have selected the data sources that are representative of the various forms of evidence currently available and that have already been used for receptor-ligand prediction (Iacucci et al., 2011; Gertz et al., 2003).

Gold Standard: the validation dataset is based on the experimentally validated dataset from the Database of Ligand-Receptor Partners (DLRP) "Golden Standard. In the version we use (retrievedin July 2009), there are 449 positive interactions between 210 receptors and ligands.

Phylogenetic: complete protein sequences are retrieved for seven species (*Rattus norvegicus*, *Mus musculus*, *Homo sapiens*, *Pan troglodytes*, *Canis familiaris*, *Cavia porcellus,* and *Bos taurus*) from EnsEMBL build 51 (Hubbard et al., 2009). Sequences are then aligned using ClustalW (Thompson et al., 1994) to detect orthology. The protein sequence and the six orthologous sequences that possibly exist are edited (as per procedure outlined in (Iacucci et al., 2011)). Eventually, and for each protein, all pair-wise alignment scores between these seven sequences are used to build the phylogenetic vector.

Expression: the gene expression profiles are retrieved from the GNF human expression atlas (Su et al., 2004). Each profile contains 79 values corresponding to the 79 conditions considered by Su *et al*.. The profiles are normalized prior to analysis (means are set to zero, and standard deviations to one).

Domain: the domain information is retrieved from the InterPro database (Hunter et al., 2009), through EnsEMBL. Only the domains present in at least one of the 210 receptors and ligands considered are kept for further analysis. Each protein is therefore represented by a sparse binary vector of size 224, corresponding to the 224 relevant InterPro domains.

Motif: The motif data is created using the Toucan toolbox to look for putative motifs in the upstream sequence of the genes. Each protein is then represented by a vector of size 674. Each value represents the score of the corresponding motif for the given protein (*i.e.*, the confidence one can have that this motif is an active motif).

Paralogy: information about paralogy is computed using the NCBI BLAST (Altschul et al., 1997) between all possible protein pairs. A given protein is then represented by its 210 BLAST alignment scores (sequence similarities to all proteins including self).

Pathway: The pathway data is retrieved from the Kegg Pathway database (Ogata et al., 1999). Similarly to the domain data source, only the pathways in which at least one of the 210 receptors and ligands is involved are used to build the final profiles. In the end, proteins are represented by sparse binary vectors of size 314.

2.2 Similarity measures

Similarity matrices are built from the original data matrices by computing, for each pair, the similarity between the associated profiles. This is done to transform the granularity of these records of from single proteins to the pairs that we aim to predict. Thus, this is essentially feature construction step. The secondary goal of this procedure is to avoid the dominance of those original data matrices which have many variables over others which have fewer variables, when estimating the classification function. Below are the four similarity measures that have been applied. In the case of mutual information, the values are unit scaled beforehand. In the case of the Jaccardi coefficient, usage is restricted to binary vectors (*i.e.*, domain and pathway data sources) due to the nature of this measure.

Cosine:

$$\cos(\theta) = \frac{A \bullet B}{\| A \| \| B \|} \qquad \text{(Eq.1)}$$

Mutual Information:

$$\mathrm{I}(A,B) = \sum_{b \in B} \sum_{a \in A} p(a,b) \log\left(\frac{p(a,b)}{p_1(a) p_2(b)} \right) \qquad \text{(Eq.2)}$$

Absolute Correlation Coefficient:

$$| C(A,B) | = | \frac{\mathrm{cov}(A,B)}{\sigma_A \sigma_B} | \qquad \text{(Eq.3)}$$

1- Jaccardi Coefficient:

$$1 - J(A,B) = 1 - \frac{| A \cap B |}{| A \cup B |} \qquad \text{(Eq.4)}$$

2.3 Classification

Generally, the process of learning from features to predict classes, in a "supervised" context, is a process which starts with partitioning the data. Data is partitioned into training and test data. Training data is used by the learning algorithm to produce an inferred function called a classifier (Cooper et al., 1997). Once a classifier is built, its performance can be assessed using the test data. We use the algorithms below to train our classifiers.

Quadratic Discriminant Analysis (QDA): related to linear discriminate analysis, quadratic discriminant analysis differs in that it uses a quadratic surface to separate the measurements of the objects being classified. Once the surface has been determined, new examples are classified based on where they are positioned relative to the boundary defined by the surface (Lu and Luo, 2008).

K-Nearest Neighbor classifiers (kNN): kNN is a method by which objects are classified based on the labels of the k closest training examples, as found in the feature space. KNN is one of the simplest machine learning methods as it requires only a majority voting of its k nearest neighbors. As the value of k can vary, it is often determined through cross-validation (Plewczynski, 2009; Raymer et al., 1997). In the current study, we have determined a value of 10 to be optimal for our setting (data not shown).

Naïve Bayes classifiers (NBC): based on Bayes's Theorem, NBCs are simple classifiers which assume independence of the features. Parameters for the Bayes model may be generated using maximum likelihood. As only a small amount of training data is necessary to estimate the parameters, it's particularly useful in settings where only a small amount of data is available (Murakami and Mizuguchi, 2010).

Decision Trees (DT): a decision tree is a predictive model used to map observations of predictors into a tree structure so as to be able to make conclusions about the outcomes. Here, we use a particular type of decision tree called a *classification tree*, where the predicted outcome is the class to which the observation belongs (Darnell et al., 2007).

Random Forests (RF): Random forests are an ensemble method which is constituted of many decision trees. Essentially, several decision trees are generated from randomly drawn subsets of the training data. New examples are classified using these trees and then a majority vote is taken from the results of the individual classifiers (Qi et al., 2005).

Noisy Or (NO): We propose and construct a *noisy or* classifier based on a noisy or score which assigns values to receptor-ligand pairs as follows:

$$\text{Score}=1-\Pi(1-S_j)$$

Where S_j is the value from the individual data source. Two means are then calculated, one from the scores of the positive examples and another from the negative examples. New pairs are then assigned a positive or negative label based on their distance from these means.

2.4 Benchmark design

We have designed two benchmark setups to tackle different questions. Both procedures are however very similar. The dataset is always split using the same scheme (2 thirds for training, and the remaining third for testing; and splitting is repeated 100 times using Monte Carlo simulations (MC)), as described in (Molinaro et al., 2005) and named Monte-Carlo cross-validation. Furthermore, as the number of positive examples in our

dataset is much smaller than the number of negatives, we train each classifier in two ways. We first define a balanced setting, in which as many negative than positive examples are used for training. We then compare to the imbalanced setup in which the original distribution between the positive and negative examples is kept. Finally, we test the trained classifiers with left-out data (test sets) over the 100 MC runs. Performance is estimated using sensitivity, specificity, and balanced accuracy. Values are averaged over the 100 MC runs and standard deviation is used to estimate the stability of the methods. As we are dealing with an imbalanced dataset with over 99% of negative cases (non-interacting pairs), we define the best performing model as being the model that yields the highest balanced accuracy.

For the first of our benchmark, we try all possible combinations of classifiers and similarity measures in order to identify which ones are the most appropriate. Using the results of the first benchmark, we propose a second benchmark to assess the contribution of each individual data source. We choose the best classifier (Random Forest) and the best performing similarity measures (for each data source) to build seven classifiers (one with all the data sources, and six with each time one of them removed).

3 RESULTS AND DISCUSSION

Our benchmarks reveal several interesting findings. Significant among the findings are the answers to our initial questions concerning which are the best choice of similarity measure, classifier, and data sources. In addition, we observe mutual dependence between information sources and advantages when balancing the training data. Below, we detail our findings and discuss their significance.

Similarity measure performance

The performance of the four similarity measures (see Figure 2) brings attention to a few key issues. First, it's clear that across all classifiers and for each data source, there does not exist one similarity measure which performs, in all cases, better than the others. This holds true for both the imbalanced and the balanced scenarios. This indicates that selecting the correct similarity measure is not a trivial consideration, but our results suggest which might be used given a particular classification algorithm and data source. Furthermore, measures such as the absolute correlation coefficient and cosine are frequently chosen, which implies that they are suitable in most cases. Thus, we do not expect any major improvement when one chooses one instead of another. Lastly, the Jaccardi coefficient is the best performing measure for the pathway data source for four of the six classifiers (NB, QDA, kNN, and NO). In contrast, for the remaining two classifiers (DT and RF), the correlation coefficient performs better for the same pathway data.

Classifier performance

The overall best performing classifier is the random forest (see Table 1) (with a sensitivity of 0.874, a specificity of 0.800, and a balanced accuracy of 0.837) while decision tree is a close second (with a sensitivity of 0.839, a specificity of 0.79, and a balanced accuracy of 0.813). The ability of the random forest to outperform the decision tree is intuitive as random forests are a collection of decision trees. Looking across all the classifiers, there is a clear advantage in balancing the training set. The worst performing classifier is the Noisy Or (with a sensitivity of 0.605 , a specificity of 0.84031, and a balanced accuracy of 0.723). The lack of performance for naïve Bayes may be explained by the fact that the data sources used are mutually dependent (see below) and naïve Bayes assumes independence of the features.

Data source performance

When assessing data source performance it becomes clear that not all are equally useful. Every feature is, in turn removed and the performance is assessed using the best classifier (RF) and the respective best similarity measure (see Figure 3). We observe that the removal of Kegg results in the largest drop in performance in both sensitivity and specificity. The difference of the means between the original run that uses all data sources and the run made without Kegg is in fact statistically significant (paired *t*-test value $p<0.001$). None of the other comparisons are significant. Indeed, it seems that the best performance, in terms of specificity, occurs when all data sources are used with the random forest. As the Kegg data source was the most informative, we examined the performance of the classifiers without Kegg and found that there was a similar drop in performance across all classifiers (see supplementary table 1).

The only data source that does not seem to add to the performance of the method is the motif feature. In fact, removing the motif source causes a slight increase in performance, in terms of specificity, for the random forest classifier. We also observe that some of the features (Blast, motif, and expression) may be mutually dependent as removing any of these sources does not alter the specificity and sensitivity values. This is important as it indicates that pairing may be successfully predicted even when there are missing values for some of the data sources.

Future directions

The prediction of receptor-ligand pairings is a difficult and complex task. While our benchmark dataset is still relatively small as compared to the whole proteome, our work here leads us to understand that there are many open questions remaining concerning the general PPI prediction task. For example, how would other similarity measures (such as the Tanimoto coefficient) perform in our benchmark? Would other hierarchical classifier algorithms work better than the random forests? Are there other data sources for the proteins which we could take into consideration for construction of features? Another issue to investigate in the future could be the effect of dimensionality reduction (within one data source) on the performance of the classifier. We intend to address these questions and more as we continue our research in this area.

Conclusion

The results from this work suggest several key findings. First, there exists a major advantage in balancing the training data, particularly with the lowest performing classifiers, suggesting that they are sensitive to imbalances in training data. Second, there is a high mutual dependence between some data sources, suggesting that they are not all necessary for the prediction task. Finally, the best performing data source is Kegg, although it seems that combining the other sources of information also works.

FIGURES AND TABLES

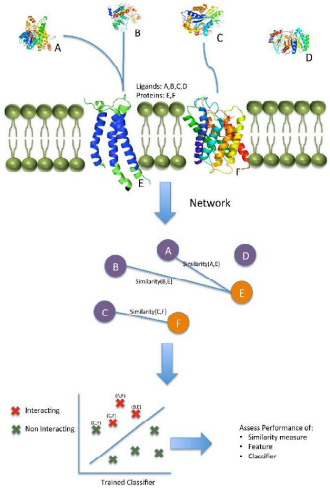


Figure 1. *Schematic view of the receptor-ligand pairing prediction task*. On top is displayed a schematic view of an interacting (left) and a non-interacting receptor-ligand pair (right). Below is represented the workflow for training a classifier using (i) features from a genomic data source, (ii) a similarity measure, and (iii) a classification algorithm.

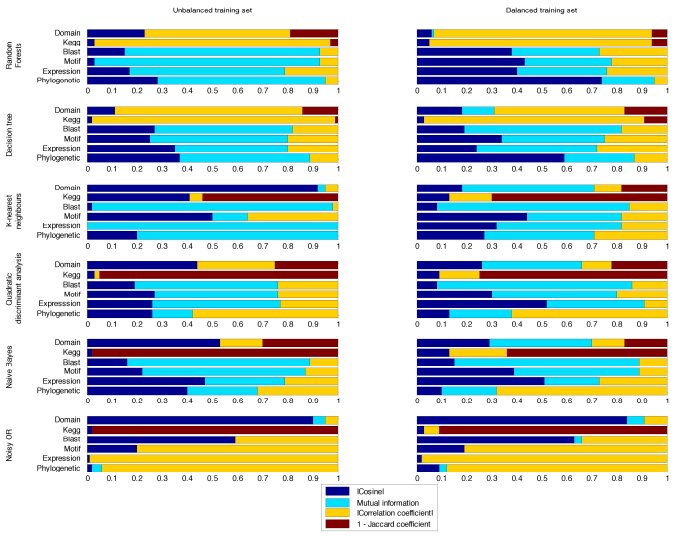


Figure 2. *Comparison of four similarity measures*. The stacked bars demonstrate the best similarity measures for the balanced and imbalanced scenarios across all the classifiers. The proportion of runs in which a particular similarity measure performed best is indicated by the colour of the bar. Notice that the Jaccard coefficient is only available for the domain and pathway sources (top two lines of each plot).

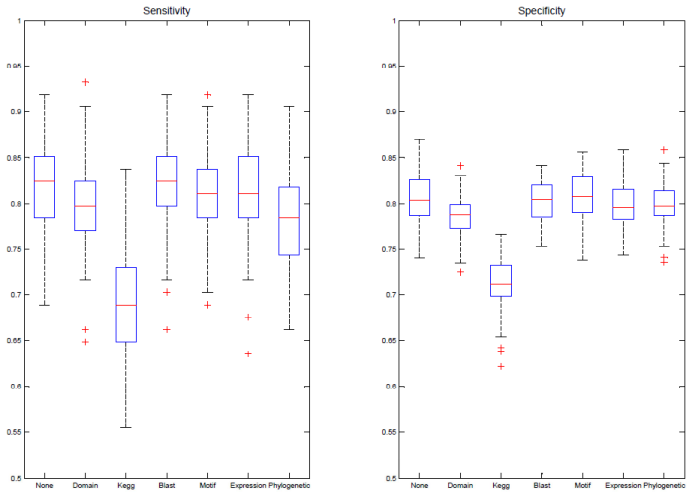


Figure 3. *Performance of individual data sources in random forest.*
Based on the best performing algorithm, random forests, benchmarking of the data sources is undertaken. Box plots show the average performance when single data sources are removed (balanced case only). The data sources listed under the ticks are the data sources which are removed in the run.

		Balanced accuracy		Sensitivity		Specificity	
		mean	std	mean	std	mean	Std
Imbalanced	Random forests	0.66565	0.02132	0.33243	0.04273	0.99887	0.00046
	Decision Trees	0.70993	0.02209	0.42459	0.04439	0.99526	0.00139
	KNN	0.57161	0.01516	0.14365	0.03029	0.99956	0.00025
	QDA	0.58209	0.01889	0.17959	0.04279	0.98459	0.00932
	Naive Bayes	0.54702	0.01618	0.10203	0.03666	0.99200	0.00672
	Noisy Or	0.72235	0.02553	0.57216	0.08042	0.87254	0.07054
Balanced	Random forests	**0.83667**	0.01787	**0.87365**	0.03905	**0.79969**	0.02190
	Decision Trees	0.81328	0.01735	0.83932	0.03959	0.78723	0.03005
	KNN	0.79697	0.01945	0.80149	0.04285	0.79246	0.02575
	QDA	0.78381	0.02020	0.78324	0.04716	0.78437	0.02999
	Naïve Bayes	0.78243	0.01963	0.78405	0.05020	0.78082	0.03511
	Noisy Or	0.72299	0.02514	0.60568	0.12107	0.84031	0.12872

Table 1. Performance of the classifiers. Classifiers performance is assessed in terms of accuracy, specificity and sensitivity. Performance is also assessed under balanced and unbalanced data settings. In bold are the values for the best performing classifier (random forest in the balanced setting).

Altschul,S.F. et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25, 3389-3402.

Cooper,G.F. et al. (1997) An evaluation of machine-learning methods for predicting pneumonia mortality. *Artif. Intell. Med.*, 9, 107-138.

Darnell,S.J., Page,D. and Mitchell,J.C. (2007) An automated decision-tree approach to predicting protein interaction hot spots. *Proteins*, 68, 813-823.

Dingledine,R. et al. (1988) Amino acid receptors and uptake systems in the mammalian central nervous system. *Crit Rev. Neurobiol.*, 4, 1-96.

Ganguli,S.C. et al. (1998) Protean effects of a natural peptide agonist of the G protein-coupled secretin receptor demonstrated by receptor mutagenesis. *J. Pharmacol. Exp. Ther.*, 286, 593-598.

Gertz,J. et al. (2003) Inferring protein interactions from phylogenetic distance matrices. *Bioinformatics.*, 19, 2039-2045.

Guarracino,M.R.a.N.A. (2010) Predicting protein-protein interactions with k-Nearest Neighbors classification algorithm. pp. 139-150.

Hubbard,T.J. et al. (2009) Ensembl 2009. *Nucleic Acids Res.*, 37, D690-D697.

Hunter,S. et al. (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res.*, 37, D211-D215.

Iacucci,E. et al. (2011) Predicting receptor-ligand pairs through kernel learning. *BMC. Bioinformatics.*, 12, 336.

Jensen,R.T. et al. (2008) International Union of Pharmacology. LXVIII. Mammalian bombesin receptors: nomenclature, distribution, pharmacology, signaling, and functions in normal and disease states. *Pharmacol. Rev.*, 60, 1-42.

Kenakin,T. and Miller,L.J. (2010) Seven transmembrane receptors as shapeshifting proteins: the impact of allosteric modulation and functional selectivity on new drug discovery. *Pharmacol. Rev.*, 62, 265-304.

Lu,J. and Luo,L. (2008) Prediction for human transcription start site using diversity measure with quadratic discriminant. *Bioinformation.*, 2, 316-321.

Molinaro,A.M., Simon,R. and Pfeiffer,R.M. (2005) Prediction error estimation: a comparison of resampling methods. *Bioinformatics.*, 21, 3301-3307.

Murakami,Y. and Mizuguchi,K. (2010) Applying the Naive Bayes classifier with kernel density estimation to the prediction of protein-protein interaction sites. *Bioinformatics.*, 26, 1841-1848.

Ogata,H. et al. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, 27, 29-34.

Plewczynski,D. (2009) kNNsim: k-nearest neighbors similarity with genetic algorithm features optimization enhances the prediction of activity classes for small molecules. *J. Mol. Model.*, 15, 591-596.

Qi,Y., Klein-Seetharaman,J. and Bar-Joseph,Z. (2005) Random forest similarity for protein-protein interaction prediction from multiple sources. *Pac. Symp. Biocomput.*, 531-542.

Raymer,M.L. et al. (1997) Predicting conserved water-mediated and polar ligand interactions in proteins using a K-nearest-neighbors genetic algorithm. *J. Mol. Biol.*, 265, 445-464.

Schipani,E., Kruse,K. and Juppner,H. (1995) A constitutively active mutant PTH-PTHrP receptor in Jansen-type metaphyseal chondrodysplasia. *Science*, 268, 98-100.

Su,A.I. et al. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. U. S. A*, 101, 6062-6067.

Sugg,E.E. et al. (1995) CCK-A receptor selective antagonists derived from the CCK-A receptor selective tetrapeptide agonist Boc-Trp-Lys(Tac)-Asp-MePhe-NH2 (A-71623). *J. Med. Chem.*, 38, 207-211.

Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22, 4673-4680.

Traynelis,S.F. et al. (2010) Glutamate receptor ion channels: structure, regulation, and function. *Pharmacol. Rev.*, 62, 405-496.

Zhang,F. et al. (2009) Anti-receptor for advanced glycation end products therapies as novel treatment for abdominal aortic aneurysm. *Ann. Surg.*, 250, 416-423.

Chapter 5

Study 3: Text-based Prediction of Receptor–Ligand Binding

This chapter represents the logical conclusion of our previous two studies. Having determined the best workflow for making receptor–ligand predictions, we now apply a trained Random Forest classifier on new text-based examples. While the objective in the previous chapters are a course of reseach in order to determine the "best practices" under our setting, the objective here is to now make useful, *in-silico* predictions. It is for this reason for which we create here a literature based candidate list and feed it through our trained classifer to make meaningful results for the receptor─ligand research community. We assess our results in terms of predicting known examples as well as the qualitative features of our prioritized list. Finally, we store our results in our online database, ReLianceDB.

Publication Information: This work is under review at ECCB12 (Proceedings published in the journal *Bioinformatics*) Iacucci E., Tranchevent L-C., Popovic D., Pavlopoulos G., De Moor B., Schneider R., Moreau Y., "*ReLianceDB: A machine learning and literature-based prioritization of Receptor─Ligand pairings*", ESAT-SISTA, K.U.Leuven (Leuven, Belgium),

Personal Contribution: The PhD candidate is responsible for designing of the experiment, data processing, analysis of experimental results, and writing of the paper. The second, third, and fourth authors provided data collection, implementation of matlab routines, figures, website design, and extremely helpful discussion concerning the whole project.

BIOINFORMATICS

Vol. 28 ECCB 2012, pages i569–i574
doi:10.1093/bioinformatics/bts391

ReLiance: a machine learning and literature-based prioritization of receptor—ligand pairings

Ernesto Iacucci[1,*], Léon-Charles Tranchevent[1], Dusan Popovic[1], Georgios A. Pavlopoulos[1], Bart De Moor[1], Reinhard Schneider[2] and Yves Moreau[1,*]

[1]IBBT Future Health Department/ESAT-SCD, KU Leuven, Kasteelpark Arenberg 10, 3001, Heverlee-Leuven, Belgium and [2]Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Campus Limpertsberg, 162 A, avenue de la Faïencerie, 1511 Luxembourg, Germany

ABSTRACT

Motivation: The prediction of receptor—ligand pairings is an important area of research as intercellular communications are mediated by the successful interaction of these key proteins. As the exhaustive assaying of receptor—ligand pairs is impractical, a computational approach to predict pairings is necessary. We propose a workflow to carry out this interaction prediction task, using a text mining approach in conjunction with a state of the art prediction method, as well as a widely accessible and comprehensive dataset.

Among several modern classifiers, random forests have been found to be the best at this prediction task. The training of this classifier was carried out using an experimentally validated dataset of Database of Ligand-Receptor Partners (DLRP) receptor—ligand pairs. New examples, co-cited with the training receptors and ligands, are then classified using the trained classifier. After applying our method, we find that we are able to successfully predict receptor—ligand pairs within the GPCR family with a balanced accuracy of 0.96. Upon further inspection, we find several supported interactions that were not present in the Database of Interacting Proteins (DIPdatabase).

We have measured the balanced accuracy of our method resulting in high quality predictions stored in the available database ReLiance.

Availability: http://homes.esat.kuleuven.be/~bioiuser/ReLianceDB/index.php

Contact: yves.moreau@esat.kuleuven.be; ernesto.iacucci@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

The 'omics' era has presented tremendous opportunities for high-throughput investigations into important questions facing the research community. Many investigative strategies of implementing data mining techniques in combination with high throughput experiments have accomplished much. Several of these high-throughput experimental methods, yeast two-hybrid systems—Y2H (Ito *et al.*, 2001), pull-down assays (Vikis and Guan, 2004), tandem affinity purification (Puig *et al.*, 2001), mass spectrometry (Gavin *et al.*, 2002; Puig *et al.*, 2001), microarrays (Stoll *et al.*, 2005) and phage display (Willats, 2002), have all generated enormous datasets, yet they are incomplete and are composed of many false positives and false negatives.

Several databases exist to store information about validated or predicted protein–protein interactions (PPI). They include the Munich Information Center for Protein Sequences—MIPS database (Mewes *et al.*, 2004), the Molecular Interactions—MINT database (Zanzoni *et al.*, 2002) the IntAct database (Kerrien *et al.*, 2007), the Database of Interacting Proteins—DIP (Xenarios *et al.*, 2000), the Biomolecular Interaction Network Database—BIND (Bader *et al.*, 2001) and the BioGRID database (Stark *et al.*, 2006). Some, like the Yeast Proteome Database (Hodges *et al.*, 1999) contain interactions which are derived from wet lab results, as well as those curated from literary sources.

Currently, there exists thousands of candidate receptors and ligands and potentially hundreds of thousands of interactions. As the exhaustive assaying of every possible receptor—ligand pairs is impractical, a computational approach to the prediction task is necessary. For example, Gertz *et al.* (2003) created a receptor—ligand matching algorithm for the *chemokine* and *tgfβ* families. Later, we more aptly matched the *tgfβ* family with kernels (Gertz *et al.*, 2003; Iacucci *et al.*, 2011) with an increase in recall of 0.76 over the 0.44 obtained from the results of Gertz *et al.* Following this, we have benchmarked several machine learning techniques, and assayed several parameters, on the receptor—ligand interaction prediction task (Iacucci *et al.*, submitted for publication). The results of this work show that we can obtain a balanced accuracy of 0.84 in this prediction task.

Having used a 'golden standard' (Graeber and Eisenberg, 2001) to determine which is the best machine learning technique to apply to this problem, we now seek to make *in silico* predictions. The starting point of this novel course of research is the widely applied and powerful field of text mining. In this article, we present a strategy that takes into account text mining information in conjunction with a popular machine learning algorithm, the random forest (Qi *et al.*, 2005), which integrates several data sources such as domain, expression and phylogenetic-based evidence.

We build our new candidate list by searching for genes that are co-cited with the receptors and ligands from the DLRP database (Graeber and Eisenberg, 2001). We then make predictions using our trained classifier and evaluate the results in terms of known pairs, as well as in terms of distribution of the co-citations in our ranked list of predictions.

*To whom correspondence should be addressed.

© The Author(s) 2012. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from http://bioinformatics.oxfordjournals.org/ at Ghent University on March 5, 2013

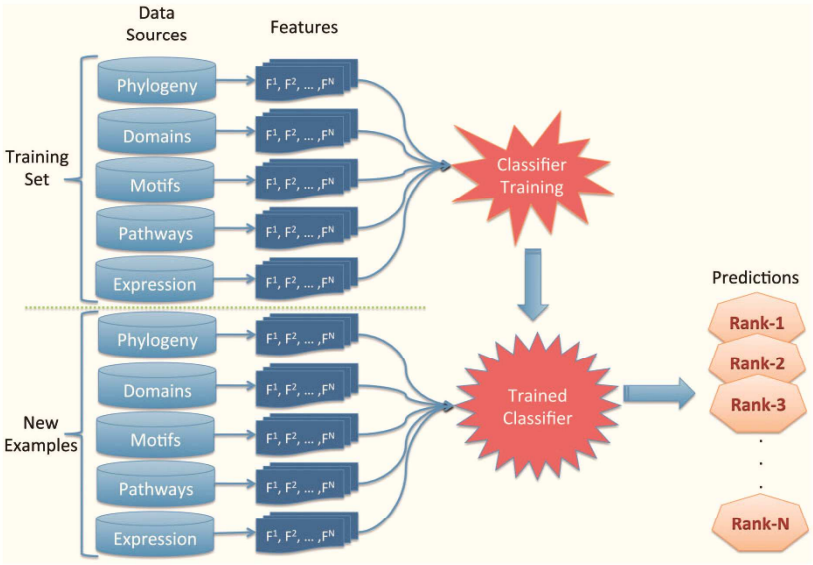


Fig. 1. Family analysis. Candidates from the new examples where mapped to Gene Ontology and a search was performed for classifications containing the term 'receptor' with more than five members. The three classifications resulting from this search criteria were 'Peptide Receptor Activity, G-protein Coupled', 'Transmembrane Receptor Tyrosine Kinase Activity' and 'Cytokine Receptor Activity'. We then used the DIP database as a baseline for calculations of sensitivity, specificity and balanced accuracy for each classification

2 METHODS

2.1 Overview

We propose a pertinent process (Fig. 1) to carry out this prediction task, using a state-of-the-art prediction method and an accessible and comprehensive dataset. Among several classifiers, random forests have previously been identified to work best for this prediction task (Iacucci *et al.*, submitted for publication). The training of this classifier was carried out using an experimentally validated dataset of receptor—ligand pairs (Graeber and Eisenberg, 2001). New examples are then classified using the trained classifier. The process begins with the collection of candidate genes to enter as new examples to our trained classifier. The new examples may be derived in a variety of ways (an up-regulated gene list, genes that contain a specific protein domain, protein-array experimentation). In our setting, we look at genes that are co-cited with the receptors and ligands from the DLRP database. The DLRP was constructed using experimentally determined ligand-receptor cognate pairs through a literature review. The database contains 314 proteins, 210 of which are used in our training set.

2.2 Creation of the candidate list

We created a candidate list by taking the 210 receptors and the ligands in the DLRP database (Graeber and Eisenberg, 2001) and finding all the co-cited genes [using the text mining track of the STRING database (Von Mering *et al.*, 2007)] for which all the information from all the data sources was available. In total, 483 candidates form this list.

2.3 Data sources

The data were collected and processed as reported in Iacucci *et al.* (submitted for publication). Briefly, profiles associated to genes were retrieved from various databases [i.e. Kegg, Interpro, Toucan, (Aerts *et al.*, 2005; Hunter *et al.*, 2009)]. These profiles are vectors of measurements, which represent the candidate ligand or receptor. We create a feature value for each candidate receptor–ligand pair by taking the two vectors and applying a pairwise similarity measure (cosine, mutual information, absolute correlation coefficient, jaccardi coefficient) to them. This value is now the feature value that corresponds to the similarity between each ligand and receptor. The feature values which arise from the use of the various sources are then inputted into our random forest.

These profiles contained between 79 (expression data) and 674 features (domain data) depending on the data source being considered. For the phylogenetic vector, complete protein sequences were retrieved for seven species (*Rattus norvegicus*, *Mus musculus*, *Homo sapiens*, *Pan troglodytes*, *Canis familiaris*, *Cavia porcellus* and *Bos taurus*) from EnsEMBL build 51 (Hubbard *et al.*, 2009). Sequences were then aligned using ClustalW (Thompson *et al.*, 1994) to detect orthology. The gene expression profiles were retrieved from the Genomics Institute of the Novartis Research Foundation (GNF) human expression atlas (Su *et al.*, 2004). Each profile contains 79 values corresponding to the 79 conditions considered by (Su *et al.*, 2004). The domain information is retrieved from the InterPro database (Hunter *et al.*, 2009), through EnsEMBL. Only the domains present in at least one of the 210 receptors and ligands considered are kept for further analysis. The motif data was created using the Toucan toolbox to search

Downloaded from http://bioinformatics.oxfordjournals.org/ at Ghent University on March 5, 2013

for putative motifs in the upstream sequence of the genes. Each protein is then represented by a vector of size 674. Each value represents the score of the corresponding motif for the given protein. The pathway data were retrieved from the Kegg Pathway database (Aerts *et al.*, 2005). Only the pathways in which at least one of the 210 training example receptors and ligands is involved are used to build the final profiles. This means that, in the case of Kegg, candidate proteins are represented by sparse binary vectors of size 314.

2.4 Classifier

Following our previous work, we selected the method that provides the best balanced accuracy (Iacucci *et al.*, submitted for publication). This method was found to be the random forest with a balanced accuracy of 0.84 (Iacucci *et al.*, submitted for publication). We used the best performing similarly measures for each data source (domain, phylogenetic—absolute correlation coefficient; expression, kegg, motif—absolute cosine). Our classifier was trained using the DLRP dataset described previously (Graeber and Eisenberg, 2001). We applied Matlab 2010a implementation of random forest (class TreeBagger) with the number of trees set to 1000 and the number of variables to select at random for each decision split set to default value (square root of number of variables). To achieve stability in the prioritized lists, the algorithm has been trained and tested 200 times, after which the new predictions were assigned by averaging scores. The running time for this procedure was ∼2h.

The candidate list was inputted into the trained classifier (trained and calibrated as described above). Each possible receptor—ligand pair was classified as interacting or noninteracting and was ranked according to a score assigned by algorithm—that is, the probability of observation belonging to particular class given as fraction of observations of that class in predicted leafs across the ensemble. The resulting list is of size 116 403 with 7521 positive predictions, 108 882 negative predictions and 7958 predictions for which co-citations exist.

2.5 Co-citation and receptor family analysis

Co-citation values were downloaded from text mining track of the STRING database (Von Mering *et al.*, 2007). The normalized co-citation score is the total number of co-citations for the members of a bin divided by the connectivity score for the members of a bin and then scaled to the maximum value across all bins. The co-citation analysis was performed by comparison of the ranked results with the co-citation score. In order to assess the biological relevance of our work in terms of individual receptor families, we map our candidates to Gene Ontology (GO) classifications and examine the performance of our predictions to individual receptor families. Gene Ontology analysis was carried out by mapping the candidate list to gene ontology classifications that contained the word 'receptor' and had more than five members. Three classifications resulted from this search criteria ('Peptide Receptor Activity, G-protein Coupled', 'Transmembrane Receptor Tyrosine Kinase Activity' and 'Cytokine Receptor Activity'). The predictions from our method were then assessed using the members of each of these classifications by comparing our predictions with the interactions reported in the DIP (Xenarios *et al.*, 2000) database for these classifications as it is known to contain experimentally validated receptor—ligand pairings (Graeber and Eisenberg, 2001).

3 RESULTS AND DISCUSSION

Evaluating the results of our predictions can be challenging, as the overall objective of our work is to predict novel receptor—ligand pairs, yet the merits of this work must be measured using known interactions. In order to address this, we look at the qualitative aspect of the co-citation profile of the ranked predictions. Furthermore, we look at the receptor classifications of Gene Ontology and assess our ability to make predictions in terms of sensitivity, specificity and balanced accuracy with respect to the known DIP interactions for these families. Finally, we assess the predictions one-by-one by examining the top ten predictions in the different quarters of the ranked predictions.

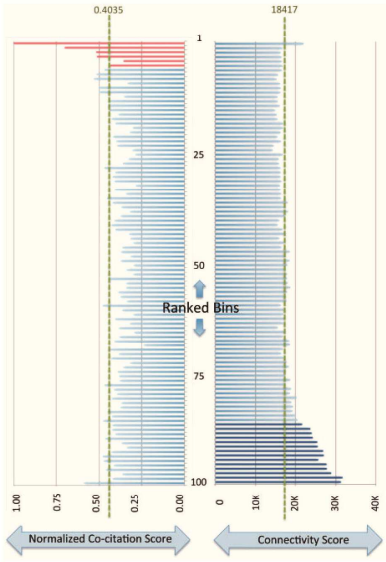


Fig. 2. Histogram. The prioritized list resulting from our workflow is binned into 100 ranked bins of size 1164. On the left we see the ranked bins that have a length corresponding to their normalized co-citation score. The normalized co-citation score is the total number of co-citations for the members of a bin divided by the connectivity score for the members of a bin and then scaled to the maximum value across all bins. The bins colored in red correspond to the bins that contain pairs, which are called a positive by our classifier. On the right, we see the ranked bins that have a length corresponding to their connectivity score. The connectivity score is the total number of edge degree (number of predicted interactors in the genome) of each of the members of the bin. The dark blue bins correspond to bins, which contain members with higher connectivity than average. The green dashed lines correspond to the average value across all bins

3.1 Co-citation analysis

Assuming co-citation should positively correlate with correct pairing of candidate pairs, we examine the predicted prioritized candidate list in terms of co-citation. The overall shape of the prioritized candidate list (Fig. 2) suggests that the performance of our workflow is consistent with an accurate classification strategy. Looking at Figure 2, we see the 100 ranked bins containing the co-citation values for the prioritized list. Looking at the top six bins, we find that there is a peak in co-citation, suggesting an overall positive

Downloaded from http://bioinformatics.oxfordjournals.org/ at Ghent University on March 5, 2013

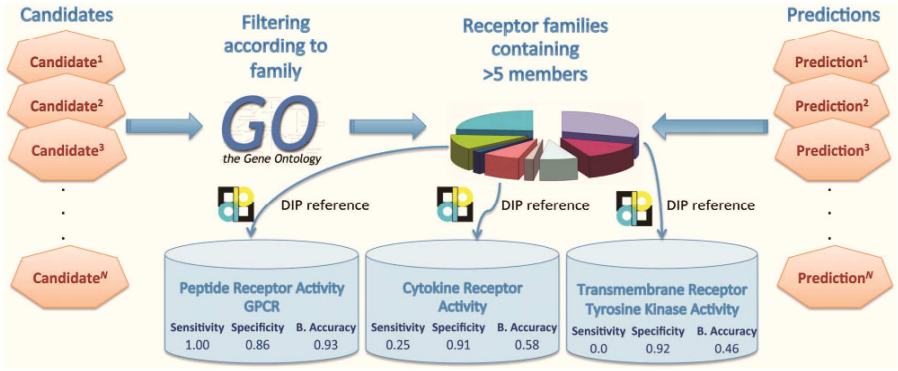


Fig. 3. Workflow. The trained classifier is provided with new examples (genes which are co-cited with the receptors and ligands from the DLRP database) and makes predictions based on its ability to distinguish between interacting and noninteracting pairs. The predictions are ranked by the random forest score provided by the class

prediction area. Indeed, if we examine the first quarter of the bins, we see that there is an enrichment of co-cited terms as they are higher than the average level (value 0.4035), which one would expect to see if the co-citations were uniformly distributed (P-value <0.05 when applying a t-test between the first and second quarters). Looking at the second and third quarters, we see a depletion of co-cited terms in those bins, whereas the fourth quarter contains levels of co-citation close to the average level. This plot suggests that our classifier is able to learn the structure in the data of interacting receptors and ligands, as well as noninteracting pairs, thus explaining the enriched and depleted areas of the plot (P-value <0.05 when applying a t-test between the first and second quarters). The bins toward the end of the ranking, which measured at about the average level, suggest that a low classifier score is assigned to pairs that show structure related to neither interacting nor noninteracting pairs. In fact, we examined these pairs and found that they were highly connected relative to the rest of the list, suggesting that the classifier could better identify pairs with where the candidate ligand and receptor had fewer, specific interactions and not more promiscuous proteins.

3.2 Receptor Family analysis

In order to find areas of high competency in our prioritization, we map the candidates to Gene Ontology [using DAVID (Dennis, Jr. *et al.*, 2003)] and perform a search for classifications containing the term 'receptor' and found three classifications with more than five members (Fig. 3). The three classifications resulting from this search criteria were 'Peptide Receptor Activity, G-protein Coupled', 'Transmembrane Receptor Tyrosine Kinase Activity' and 'Cytokine Receptor Activity'. We then used the DIP database as a baseline for calculations of sensitivity, specificity and balanced accuracy for each classification. More specifically, we take the candidates from our experiment that are mapped to these classifications and compare our positive predictions with those reported in the DIP database. We find that we are able to successfully predict receptor—ligand pairs within the G-protein coupled receptors (GPCR) family with a balanced accuracy of 0.96. The GPCR family represents the most important group of current drug targets because 40% of all modern medicinal drugs are GPCR related (Overington *et al.*, 2006) (e.g. imatinib, cetirizine, hydroxyzine and acebutolol). This is no surprise as they are key agents in several diseases (Overington *et al.*, 2006) (e.g. WHIM syndrome, Retinitis pigmentosa and Cryptorchidism). We searched our results for novel interactions that were made with members of this family, which did not exist in the DIP database. We examined the top 10 predictions (Table 1) from our prioritization involving members of the GPCR family found several supported interactions such as those between TCF7-CTNNB1 (Kerrien *et al.*, 2007; Mewes *et al.*, 2004; Xia *et al.*, 2006), LEF1-CTNNB1 (Bader *et al.*, 2001; Goel *et al.*, 2012; Mewes *et al.*, 2004; Stark *et al.*, 2006) and ANGPT2-F2R (Kerrien *et al.*, 2007). In addition, we find high STRING (Von Mering *et al.*, 2007) scores for the following interactions CCL22-CCR1, CXCL13-CCR1 and CCL22-CX3CR1. Notably, we found that two of these novel *in silico* predictions (CXCL13-CCR1 and CCL22-CX3CR1) showed experimental evidence of interaction (Booth *et al.*, 2008; Hoglund *et al.*, 2011).

3.3 Qualitative analysis

We examine the 10 co-cited predictions made at the top of each quarter of our ranking and find that the several of the predictions made at the top of the 1st quarter were supported (Table 2). Among these interacts are those between CD3G-CD3D (Goel *et al.*, 2012; Kerrien *et al.*, 2007; Mewes *et al.*, 2004; Stark *et al.*, 2006), B2M-CALR (Goel *et al.*, 2012; Kerrien *et al.*, 2007; Mewes *et al.*, 2004; Stark *et al.*, 2006), CDC14A-CDC7 (Kerrien *et al.*, 2007), PDGFRB-GRB7 (Goel *et al.*, 2012; Mewes *et al.*, 2004) and SMAD7-ACVRL1 (Kerrien *et al.*, 2007). The second and third quarters (which were below the threshold of a positive call in the algorithm) contain predictions that were not supported.

Downloaded from http://bioinformatics.oxfordjournals.org/ at Ghent University on March 5, 2013

Table 1. *In Silico* GPCR predictions: top ten predictions made in the GPCR family of receptors and ligands

In Silico GPCR Predictions				
Rank	**Query protein**	**Predicted partner**	**Score**	**Evidence of interaction**
1	CD27	CX3CR1	0.789	
2	TCF7	CTNNB1	0.780	INTNETDB, MIPS, INTACT
3	CD27	CCR1	0.723	
4	CCL22	CCR1	0.721	High STRING prediction: 0.964
5	LEF1	CTNNB1	0.720	BIND, BIOGRID, HPRD, MIPS
6	CCR1	CSF1	0.716	
7	CXCL13	CCR1	0.694	High STRING prediction: 0.983 Experimental (Booth *et al.*, 2008)
8	EDAR	CX3CR1	0.680	
9	ANGPT2	F2R	0.667	INTACT
10	CCL22	CX3CR1	0.638	High STRING prediction: 0.945 experimental (Hoglund *et al.*, 2011)

Table 2. Top 10 co-cited predictions: the top 10 co-cited predictions with co-citation

TOP 10, first quarter				
Rank	**Query protein**	**Predicted partner**	**Score**	**Evidence of interaction**
1	CD3G	CD3D	0.903	BIOGRID, HPRD, INTACT, MIPS
2	CRK	ALK	0.894	
3	AC003958.6.1	TNFSF4	0.890	
4	B2M	CALR	0.879	BIOGRID, HPRD, MIPS
5	CDC14A	CDC7	0.876	INTACT
6	DCN	SMAD7	0.870	
7	PDGFRB	GRB7	0.867	HPRD, MIPS
8	SMAD7	ACVRL1	0.866	INTACT
9	TNFSF4	IL18	0.866	
10	WDR48	ERBB2	0.865	

The remaining, fourth quarter contained three supported interactions (Supplementary Tables 1, 2 and 3). The interactions present in the fourth quarter are due to the highly connected nature of the interacting receptors and ligands that are assigned low scores (as described above).

3.4 ReLiance database

The ReLiance database is available at: http://homes.esat.kuleuven.be/~bioiuser/ReLianceDB/index.php. Among the information accessible from the database are the putative protein partner, the score for our prediction, as well as other pertinent information (hugo names, swissprot ids, PDB entries, etc.). The database entries are annotated and enriched on-the-fly using the reflect API (Pafilis *et al.*, 2009).

A mouse-click function is applied to the highlighted proteins and genes to generate informative pop-up windows hosting information where the specific bioentities from public databases are summarized. Thus, links to the synonyms, the complete sequence of the longest transcript, domains from the Simple Modular Architecture Research Foundation (SMART) (Letunic *et al.*, 2006) database, the PDBsum (Laskowski, 2001) structure, the interactions from STITCH (Kuhn *et al.*, 2008), the known subcellular location and the source organism are shown, respectively. Most of these features on the pop-up are hyperlinked to related database entries.

In addition, Medline abstracts concerning the bioentity are offered by the iHOP (Hoffmann and Valencia, 2004) service. Similar functionality can be offered by OnTheFly (Pavlopoulos *et al.*, 2009) service for a more targeted search as selected results can be stored locally and then annotated. This way, in terms of data integration and identifier updates, the database can always be up to date and supported by the Reflects and OnThFly's dictionary.

Networks showing the interaction partners can dynamically be explored by the Medusa application (Pavlopoulos *et al.*, 2011) that runs as an applet on our site. Force-directed layout algorithms provide intuitive layouts, whereas the color scheme encodes certain information. The 'purple' node represents the query protein, the 'red' nodes correspond to the predicted partner proteins which are also co-cited to the query protein and 'yellow' nodes correspond to the predicted partner proteins that are not co-cited with the query protein.

4 CONCLUSIONS AND FUTURE DIRECTIONS

Our workflow to carry out the receptor—ligand pairing prediction task provides several findings. Using a text mining approach in conjunction with a state of the art prediction method, as well as a widely accessible and comprehensive dataset, we have produced a prioritized list that is consistent with a successful classification scenario. In addition, we find that we are able to successfully predict receptor—ligand pairs within the GPCR family with a balanced accuracy of 0.96. We introduce ReLiance, our database for predicted receptor—ligand pairings which provides high data integration and visualization capabilities.

Funding: This work was supported by Research Council KUL (ProMeta, GOA Ambiorics, GOA MaNet CoE EF/05/007); SymBioSys en KUL PFV/10/016 SymBioSys; START 1 (several PhD/postdoctorate and fellow grants); Flemish Government: FWO: [PhD/postdoctorate grants, projects G.0318.05 (subfunctionalization), G.0553.06 (VitamineD), G.0302.07 (SVM/Kernel), research communities (ICCoS, ANMMM, MLDM); G.0733.09 (3UTR)]; G.082409 (EGFR) IWT: PhD Grants, Silicos; SBO-BioFrame, SBO-MoKa, TBM-IOTA3 FOD: Cancer plans, IBBT; Bel- gian Federal Science Policy Office: IUAP P6/25 (BioMaGNet, Bioinformatics and Modeling: from Genomes to Networks, 2007- 2011); EU-RTD: ERNSI: European Research Network on System Identification; FP7-HEALTH; CHeartED

Conflict of Interest: None declared.

REFERENCES

Aerts,S. *et al.* (2005) TOUCAN 2: the all-inclusive open source workbench for regulatory sequence analysis. *Nucleic Acids Res.*, **33**, W393–W396.

Bader,G.D. *et al.* (2001) BIND–The Biomolecular Interaction Network Database. *Nucleic Acids Res.*, **29**, 242–245.

Booth,G. *et al.* (2008) Gene expression profiles at different stages of collagen-induced arthritis. *Autoimmunity*, **41**, 512–521.

Downloaded from http://bioinformatics.oxfordjournals.org/ at Ghent University on March 5, 2013

Dennis,G, Jr. *et al.* (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.*, **4**, 3.
Gavin,A.C. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.
Gertz,J. *et al.* (2003) Inferring protein interactions from phylogenetic distance matrices. *Bioinformatics*, **19**, 2039–2045.
Goel,R. *et al.* (2012) Human Protein Reference Database and Human Proteinpedia as resources for phosphoproteome analysis. *Mol. Biosyst.*, **8**, 453–463.
Graeber,T.G and Eisenberg,D. (2001) Bioinformatic identification of potential autocrine signaling loops in cancers from gene expression profiles. *Nat. Genet.*, **29**, 295–300.
Hodges,P.E. *et al.* (1999) The Yeast Proteome Database (YPD): a model for the organization and presentation of genome-wide functional data. *Nucleic Acids Res.*, **27**, 69–73.
Hoffmann,R. and Valencia,A. (2004) A gene network for navigating the literature. *Nat. Genet.*, **36**, 664.
Hoglund,R.A. *et al.* (2011) Expression and functional activity of chemokine receptors in glatiramer acetate-specific T cells isolated from multiple sclerosis patient receiving the drug glatiramer acetate. *Hum. Immunol.*, **72**, 124–134.
Hubbard,T.J. *et al.* (2009) Ensembl 2009. *Nucleic Acids Res.*, **37**, D690–D697.
Hunter,S. *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res.*, **37**, D211–D215.
Iacucci,E. *et al.* (2011) Predicting receptor—ligand pairs through kernel learning. *BMC Bioinformatics*, **12**, 336.
Ito,T. *et al.* (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.
Kerrien,S. *et al.* (2007) IntAct–open source resource for molecular interaction data. *Nucleic Acids Res.*, **35**, D561–D565.
Kuhn,M. *et al.* (2008) STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Res.*, **36**, D684–D688.
Laskowski,R.A. (2001) PDBsum: summaries and analyses of PDB structures. *Nucleic Acids Res.*, **29**, 221–222.
Letunic,I. *et al.* (2006) SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res.*, **34**, D257–D260.
Mewes,H.W. *et al.* (2004) MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res.*, **32**, D41–D44.
Overington,J.P. *et al.* (2006) How many drug targets are there? *Nat. Rev. Drug Discov.*, **5**, 993–996.
Pafilis,E. *et al.* (2009) Reflect: augmented browsing for the life scientist. *Nat. Biotechnol.*, **27**, 508–510.
Pavlopoulos,G.A. *et al.* (2009) OnTheFly: a tool for automated document-based text annotation, data linking and network generation. *Bioinformatics*, **25**, 977–978.
Pavlopoulos,G.A. *et al.* (2011) Medusa: A tool for exploring and clustering biological networks. *BMC Res. Notes*, **4**, 384.
Puig,O. *et al.* (2001) The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods*, **24**, 218–229.
Qi,Y. *et al.* (2005) Random forest similarity for protein-protein interaction prediction from multiple sources. *Pac. Symp. Biocomput.*, **2005**, 531–542.
Stark,C. *et al.* (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, **34**, D535–D539.
Stoll,D. *et al.* (2005) Protein microarrays: applications and future challenges. *Curr. Opin. Drug Discov. Dev.*, **8**, 239–252.
Su,A.I. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, **101**, 6062–6067.
Thompson,J.D. *et al.* (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
Vikis,H.G. and Guan,K.L. (2004) Glutathione-S-transferase-fusion based assays for studying protein-protein interactions. *Methods Mol. Biol.*, **261**, 175–186.
Von Mering,C. *et al.* (2007) STRING 7–recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.*, **35**, D358–D362.
Willats,W.G. (2002) Phage display: practicalities and prospects. *Plant Mol. Biol.*, **50**, 837–854.
Xenarios,I. *et al.* (2000) DIP: the database of interacting proteins. *Nucleic Acids Res.*, **28**, 289–291.
Xia,K. *et al.* (2006) IntNetDB v1.0: an integrated protein-protein interaction network database generated by a probabilistic model. *BMC Bioinformatics*, **7**, 508.
Zanzoni,A. *et al.* (2002) MINT: a Molecular INTeraction database. *FEBS Lett.*, **513**, 135–140.

Downloaded from http://bioinformatics.oxfordjournals.org/ at Ghent University on March 5, 2013

TOP 10, 2nd Quarter				
Rank	**Query Protein**	**Predicted Partner**	**Score**	**Evidence of Interaction**
1	PXN	SLC2A4	0.33381	---
2	TNFRSF1A	STAT2	0.33378	---
3	ARF1	P4HB	0.33367	---
4	RUNX3	ZFPM2	0.33356	---
5	PPARG	OLR1	0.33347	---
6	BMP4	CTSK	0.33342	---
7	SPARC	BMP4	0.33326	---
8	PDYN	GHRH	0.33325	---
9	MAP2	BCR	0.33318	---
10	CCL22	IKZF1	0.33318	---

Supplementary Table 1: Top 10 Co-cited Predictions: The **t**op ten co-cited predictions with co-citation in the 2nd Quarter.

TOP 10, 3rd Quarter				
Rank	**Query Protein**	**Predicted Partner**	**Score**	**Evidence of Interaction**
1	TWIST1	MMP13	0.20414	---
2	TRAF3	IRAK4	0.20409	---
3	COMT	BCR	0.20407	---
4	INHBE	GDF9	0.20376	---
5	TBP	RET	0.20366	---
6	NRG1	NOD2	0.20357	---
7	TGFB1	TNNI3	0.20348	---
8	ETV1	ELF3	0.20346	---
9	IRS1	CARD11	0.20336	---
10	CLTA	NEDD8	0.20327	---

Supplementary Table 2: Top 10 Co-cited Predictions: The **t**op ten co-cited predictions with co-citation in the 3rd Quarter.

TOP 10, 4th Quarter				
Rank	**Query Protein**	**Predicted Partner**	**Score**	**Evidence of Interaction**
1	IL18	NOD2	0.12548	---
2	CD44	PLP1	0.12548	---
3	PENK	MME	0.12547	---
4	CD5	LCK	0.12544	BIOGRID, DIP, HPRD, MIPS
5	COMT	NR2F2	0.12544	---
6	CD5	ODF1	0.1254	---
7	IFNG	CD3G	0.1252	---
8	IRF8	IL18	0.12514	MIPS
9	CSK	LCK	0.12502	HPRD, MIPS
10	GALNS	CCL5	0.12493	---

Supplementary Table 3: Top 10 Co-cited Predictions: The **t**op ten co-cited predictions with co-citation in the 4th Quarter.

Chapter 6

Discussion and Conclusions

The task of receptor–ligand pairing prediction is a difficult and important area of bioinformatics research. The advancement of this task has been outlined in this thesis. In our first study we discuss the improvement of an LS-SVM approach over the state-of-the-art method. This work was restricted to making predictions on two receptor–ligand families and using three data sources. In our second study, we discuss the benchmarking of various machine learning techniques on a larger dataset with additional data sources. Our third study centres around making text-based *in silico* predictions using the trained classifier from study number three.

As the number of possible interacting protein pairs in the cell is large, wet-lab experimentation validation of all of them is essentially impossible. In addition to being time consuming, *in-vivo* wet-lab validation costs are also a consideration. Having a computational method for predicting receptor–ligand pairs is therefore a necessary tool for researchers.

Here we discuss the achievements and merits of the approaches discussed in studies one through three. In our first study, using an LS-SVM classifier, we show that we are able to more aptly match members of the chemokine and *tgfβ* families than a previously published method (Gertz et al., 2003). Notably, we are able to achieve an increase in recall of 0.76 over the 0.44 for the matching of receptor–ligands in the *tgfβ* family. In our subsequent study, we benchmarked several machine learning techniques, and assayed several parameters, on the receptor–ligand interaction prediction task. We found that we could reach a balanced accuracy of 0.84. In our final work, we produce a publicly available database of our results with respect to a text-based *in silico* prediction workflow. The resulting database, contains several key findings, particularly predictions in the GPCR family with a balanced accuracy of 0.96 (Iacucci et al., 2011a; Gertz et al., 2003).

LS-SVMs as an Improved Classifier for Receptor–Ligand Pairing Predictions

The comparison of the phylogenetic based method of (Gertz et al., 2003) and the combined kernel classifer method of (Iacucci et al., 2011b) provides a clear perspective on the advantages of multiple kernel learning in the PPI prediction task. As both groups use the same dataset and have results which can be summarized and contrasted using recall, precision, and the *F*-measures.

Our (Iacucci et al. (2011)) predictions for the *tgfβ* family accurately reconstructed over 76% of the supported edges (0.76 recall and 0.67 precision) of the known DLRP receptor–ligand pairs. In this case, the combined kernel classifier was able to improve upon the Gertz et al. (2003) work by a factor of approximately two as the Gertz et al. (2003) work reconstructs 44% of the supported edges (0.44 recall and 0.53 precision) of the know DLRP receptor–ligand pairs. Comparing *F*-measures, we see that the combined kernel classifer method improved upon that of Gertz et al. (2003) significantly as the combined kernel classifier method has an *F*-measure of 0.71 while that of Gertz et al. (2003) has a value of 0.48.

The task of learning from the structure present in the data is limited in many respects, foremost being the heterogeneity of the available data. In our first study, we examine two receptor–ligand families, *tgfβ* and the chemokines, which are of size 28 and 50 respectively. In this study two LS-SVM classifiers were built and the results were examined in terms of precision and recall. The *tgfβ* classifier was able to make predictions with a recall of 0.76 and a precision of 0.67. The chemokine classifier was able to make predictions with a recall of 0.79 and a precision of 0.31. These results can be explained in terms of the quality of the training data. As the *tgfβ* are an evolutionarily related family, the training data has a structure from which the classifier can learn. In contrast, the chemokine family are disparate in terms of evolutionary relationship and thus provide poorer training data in terms of related structure. As precision is a measure of true positives over all positive calls, it stands to reason that the *tgfβ* classifier was able to produce a very high quality result, given high quality training data.

The merits of this approach are many. Foremost are the ability of the classifier to predict multiple ligands for one receptor, which represents an important option for receptor–ligand research. In addition, as the classifier output is continuous, the results can be considered to be prioritized, this presents a major convenience to researchers as often the set of candidate ligands is large and resources to validate few. In contrast, PPI investigation has traditionally been restricted to binary interaction prediction.

The advantage of using the three sub-classifiers instead of a global classifier which combines all features is two fold. The first reason would be that the data sources used here are disparate and heterogeneous. A global classifier would require a

mapping step which may introduce some noise. The second reason is that using separate sub-classifiers would allow for removing and adding of sub-classifiers. For example, if a better microarray dataset becomes available in the future, it would be an advantage to be able remove the existing expression-based kernel with one derived from the new dataset without having to the retrain a global classifier. Also, if additional data sources become available, adding an additional sub-classifier based on the new data source would take less time to train than adding the data source and retaining the global classifier.

The major limitation of this method rests in the need to have training examples for receptor–ligands which one is trying to predict. This is particularly true for predicting the pairing in the chemokine dataset as when we consider only ligand candidates with two or more receptor pairings, the precision performance of our method improves (0.79 recall and 0.31 precision).

Benchmarking of Machine Learning Methods for the Receptor–Ligand Prediction Problem

This work represents a radical extension of the previous study. While the first study dealt with only three data sources (expression, domain, and phylogenetic), this study makes use of three additional data sources (motif, blast, and Kegg). In addition, while the first study centred around only the LS-SVM, this study benchmarks naïve Bayes, decision trees, KNN, quadratic discriminant analysis, noisy or, and the random forest. Lastly, this study makes use of the whole DLRP dataset in lieu of the two families used in the first study.

The merits of our second study stem from the three main findings of this work. First, there exists a major advantage to balancing the training data, particularly with the lowest performing classifiers, suggesting that they are sensitive to imbalanced data in this setting. Second, there is a high mutual dependence between some features (and consequently between data sources), suggesting that they are not all necessary for the prediction task. Finally, as the best performing feature was the Kegg membership, though it seems that the performance of the other features combined also works very well.

The limitations of this work are few yet fierce. First, the nature of this benchmarking requires heavy computational resources as the number of randomizations necessary to carry out this work are many. Second, the dataset must be sufficiently large to satisfy the requirements of the various methods. For example, the random forest is generated via a sampling procedure which makes use of a large amount of training data. The last limitation is the inability of the methods to accommodate additional data sources once the classifier is trained. Introducing a new data source requires retraining and re-benchmarking of the dataset which takes additional time.

The ReLianceDB Creation

As the ultimate goal of the receptor–ligand prediction work described in studies 1 and 2 is to help biologists make new discoveries using a computationally derived candidates list, we create and make public a high quality ranking in our database ReLianceDB. We build our new candidate list by searching for genes which are co-cited with the receptors and ligands from the DLRP database (Graeber and Eisenberg, 2001) used in the second study. We then make predictions using our trained classifier (trained as per the second study) and evaluate the results in terms of known pairs as well as in terms of distribution of the co-citations in our ranked list of predictions.

The merits of this work are best described by the resulting list. After applying our method, we find that we are able to successfully predict receptor─ligand pairs within the GPCR family with a balanced accuracy of 0.96. Upon further inspection, we find several supported interactions that were not present in the DIP database. In addition to the quality of the results, this study also allows biologist to access the list as it is hosted on the web at our ReLianceDB website.

The limitations of this work are few but noteworthy. First, the classifier is trained using a well curated but limited dataset. This means that the classifier can easily identify new examples where the candidate ligand and receptor have fewer, specific interactions but not more promiscuous proteins. The second limitation of this work is that it can not easily accommodate changes in the data sources nor the classifier parameters as that would mean retraining and re-benchmarking of the work.

It is important to understand the biological significance of the predictions made in the ReLiance database. Which families are represented in the training set and what are the families for which our results show the strongest predictions? We apply the Pfam family classifications to our training set and find the proportion in which each family is represented. We show in figure 3 that the families are generally equally distributed in the training examples. After consulting the predictions found in ReLiance database we see that the FG-GAP repeat and integrin alpha families show the best performance, in terms of balanced accuracy, when validating with the MIPS database (Mewes et al., 2004). The FG-GAP repeat and integrin alpha family are important to researchers working with integrins as both are components of integrins. This family is responsible for such processes as signal transduction and cell adhesion. From a structural biology point of view, this family is characterized by a series of heterodimers of alpha and beta subunits. Noteable examples of integrin receptors are the fibronectin and laminin receptors which are responsible for cell motility and adhesion.

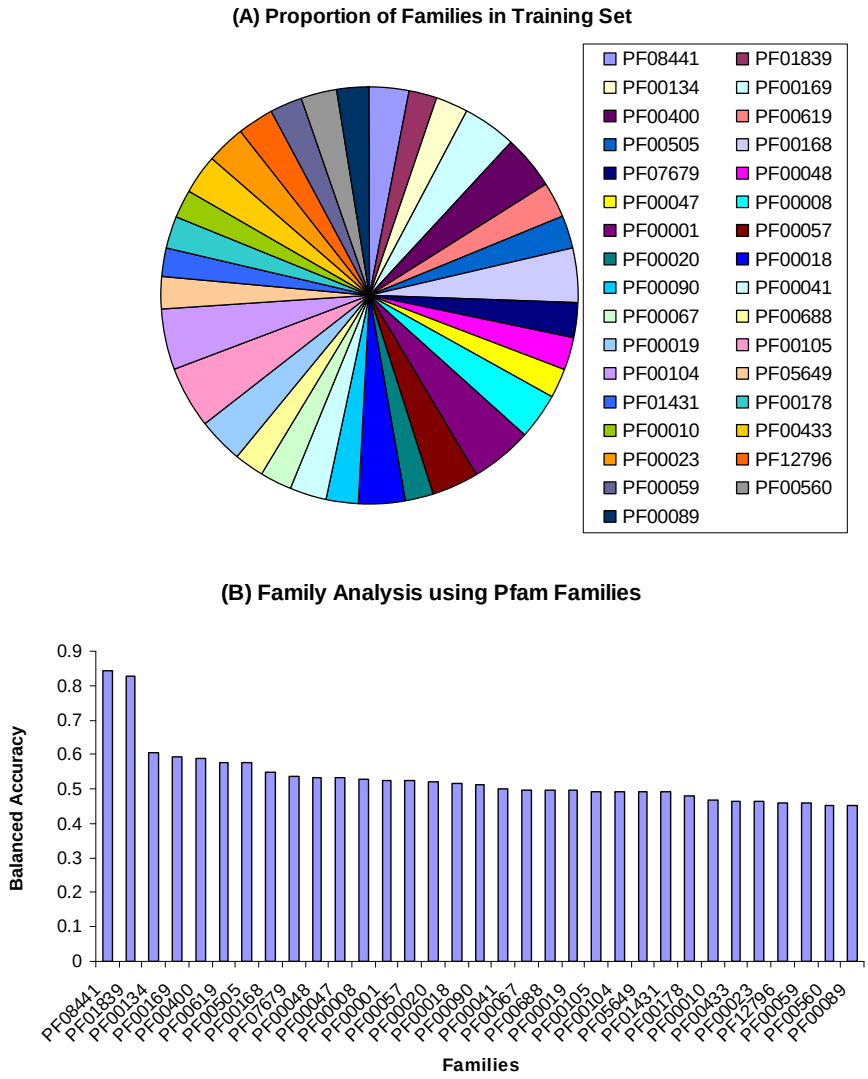


Figure 3 ***Family Analysis of ReLiance Training Data and Results*** On top (A) a pie chart depiction of the proportion of Pfam families used in the training set. On bottom (B), a bar chart of the Balanced accuracy of the predictions of the ReLiance database when MIPS values are used to validate the predicted interactions.

Similar to the validations carried out in study 3, we examine the predicted interactions between integrins and non-integrins in our results. In total, there were 80 positive predictions made for these types of interactions and manual investigations of the literature for some of these predictions revealed few supported interactions. The first investigation followed the same logic as the validations made in study 3 where the top 10 interactions (PELO-ITGA5, ITGA3-PELO, ITGA4-PELO, IBSP-ITGA6, ITGA6-PELO, ITGA6-CHAD, ITGA3-CD44, CD44-ITGA5, FAP-ITGAV, ITGA5-IGF2R) on the ranked list were investigated. The second investigation was carried out on the 15 interactions (IBSP-ITGA6, KRIT1-ITGAV, NISCH-ITGA5, CYR61-ITGA5, BCAR1-ITGA4, ITGA4-NOD2, FAP-ITGAV, ITGAV-SPON1, ITGA3-COL7A1, ITGA6-COL7A1, ITGA3-MMP15, IBSP-ITGAV, ITGA6-CD34, ITGA3-CD44, CD44-ITGA5) with significant co-citation scores. Of the above listed interactions, only one interaction (CD44-ITGA5) was supported by evidence of direct interaction (Fullar et al., 2012) while two others (ITGA6-CHAD and CYR61-ITGA5) showed weaker evidence ranging from co-expression to indirect complex formation (Volkmer et al., 2012; Manque et al., 2011; Reuter et al., 2009). Other predicted interactions might have been co-cited but the studies provided no evidence of interaction. While not strong results, these findings are not surprising as they are consistent with the validations present in study 3.

The final study into the literature aided prediction of pairing of receptor-ligand candidates is a novel approach to *in-silico* predictions. While we have examined the resulting predictions list in terms of family predictions as well as against known PPI databases, we also wish to consider the evidence for extracellular ligands and membrane bound protein interaction. We therefore turn our attention to the characterization of the interactions produced by the ReLiance database. Of the positive predictions, we see that 3.07% of the predicted partners are between proteins which contain extracellular export signal domain (ES) and a transmembrane domain (TM). When examining the negative predictions, we see that 2.69% of the predicted partners are between proteins which contain an ES and TM. While the results of this characterization do not show a striking difference in the ratio of extracellular/transmembrane partner pairings, it should be mentioned that the presence of a transmembrane domain does not indicate the capacity to propagate a receptor signal nor does the extracellular localization of a protein make it a potential ligand.

Conceptual improvements

Methods, which rely on the physical structure and the conservation of amino-acid sequences in interacting partners that are already reliably known to exist, also give researchers additional insight to function prediction as the methods are based on known examples. The main drawback of these methods is that one has to have a known example for a comparison, which is not always the case when researching

candidate receptor–ligand pairs. A possible improvement to the concepts described in our work would be a workflow which adds a physical structure check. This step would involve the modeling of the predicted receptor–ligand pair and the fitting of these structures into know interacting homologous pairs.

Technical improvements

Methods which rely on phylogenetic histories to determine PPI are based on a well-establish rational which holds that as interacting proteins co-evolve, their phylogenetic histories should be similar (Iacucci et al., 2011a; Sato et al., 2006; Tillier et al., 2006; Sato et al., 2005; Kim et al., 2004; Gertz et al., 2003; Thompson et al., 1994a). This explains why the methods which rely on phylogenetic information are largely based on measures of similarity for the phylogenetic trees of interacting protein partners. A technical improvement to our work would be to increase the number of sequences used for the phylogenetic vector. To date, not enough high quality sequences are available for use to extend the phylogenetic vector to include more species.

The advantage of using multiple kernel learning to predict PPI is apparent when using multiple sources of data. Many of the methods, mentioned above, rely on an ever growing amount of publicly available data. The ever expanding amount of high throughput data which continues to become available to the bioinformatics community represents an excellent opportunity to enhance the kernel classifier method presented in Iacucci et al. (2011). We foresee that as these data sources become available, their inclusion into our work will increase the performances described above.

Conclusions

The prediction of receptor–ligand pairings is a difficult and complex task. While PPI interaction for stable interacting partners has been examined in other works, we present methods for the prediction of transient receptor–ligand interactions. The analysis presented in our work demonstrates that the interactions present in our ReLianceDB support the idea that we are specifically capturing receptor–ligand interactions. When inspecting this data we find that the top of our prioritization have supported receptor–ligand interactions while the bottom of our priorization show highly connected partners with non-specific PPI binding.

Our work here leads us to understand that there are many open questions remaining concerning this prediction task. For example, how would other similarity measures (such as the Tanimoto coefficient) perform in our benchmark? Would other hierarchical classifier algorithms work better than the Random Forests? Are there

other protein features which we could take into consideration when performing our benchmarking? This approach can, of course, be expanded beyond the set of genes prioritized in our work. By having a well defined training set and a benchmarking set-up similar to the one used here, it is possible to prioritize other candidates in a manner similar to the one show in our work.

As the hypothesis being put forward in this thesis holds that receptor-ligand pairings can be advanced through the use of machine learning techniques, the next step in this research line would be *in-vitro*/*in-vivo* validation. Given the list of interactions predicted in the ReLiance database, wet-lab scientists now have the opportunity to validate the results. As more candidates are validated, the number of positive and negative known interactions will increase and can be then used to retrain a new future classifier which will benefit from more examples. A limitation of this approach is that there needs to be a way to generate new candidates. Here we use a literature based strategy to generate candidates. In other experimental settings, it may make sense to use other strategies such as looking at a set of co-expressed genes or proteins which have been identified to participate in a certain biological process. The take-away message from this thesis is that the successful methods presented here are only a starting point for these types of investigations. The work carried out in this thesis is the culmination of the most representative data and methods present in bioinformatics today (under our experimental setting). The data can be improved and expanded (Urquiza et al., 2012; Kshirsagar et al., 2012) in the future and new methods (Browne et al., 2010; Agarwal et al., 1993) will surely arise and similar work to ours will need to be implementing to continue to improve predictions. Likewise, the results from these types of studies will need to be validated in the wet lab. The wet lab validation of the results will provide improved and enlarged training data which will also contribute to an overall benefit to this type of work. This is already what has been seen in the general field of PPI (Reynes et al., 2010) and will no doubt continue over time.

Looking forward, many exciting challenges remain to be addressed in this field. While the task of PPI is daunting and complex, the work reviewed above demonstrates that it is also rich with opportunities for improvement and further development.

Curriculum Vitae

Ernesto Iacucci was born in Montreal (Canada). He completed a Bachelors at McGill University in Computer Science and Biochemistry where he completed his honours lab rotation at the Biotechnology Research Institute (BRI) under the supervision of Dr. Andre Nantel where his project concerned the microarray analysis of the growth of *Candida albicans*. Remaining at McGill and under the supervision of Dr. Hans Zingg and Dr. T. Perkins, he then completed a Masters in Physiology (quantitative) where his thesis comprised of the creation of 2 novel statistical methods for determining the significance of Gene Ontology classifications which he would later publish. He then moved to Belgium where he obtained his PhD at the KU Leuven (University of Leuven) in Bioinformatics at the electrical engineering department under the patronage of Dr. Yves Jean-Luc Moreau.

List of Publications

Publications - *International Journal Papers*

- Iacucci E., Ojeda F., De Moor B., Moreau Y., "Predicting Receptor-ligand Pairs through Kernel Learning", BMC Bioinformatics, vol. 12, no. 336, pp. 1-8.
- Charalampos N. Moschopoulos, Georgios A. Pavlopoulos, Ernesto Iacucci, Jan Aerts, Spiridon Likothanassis, Reinhard Schneider and Sophia Kossida., "Which algorithm is better for predicting protein complexes?" BMC Res Notes. (*In press*)
- Ernesto Iacucci, Hans H. Zingg, Theodore J. Perkins, "Methods for Determining the Statistical Significance of Enrichment or Depletion of Gene Ontology Classifications under Weighted Membership" Frontiers in Genetics (*In press*)
- Iacucci E., Tranchevent L.-C., Popovic D., Pavlopoulos G. A., De Moor B., Schneider R., Moreau Y., "A Bioinformatics E-dating Story: Computational Prediction and Prioritization of Receptor-Ligand Pairs", Internal Report 12-84, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 2012. BMC Bioinformatics, (Accepted)
- Ernesto Iacucci, Dusan Popovic, Leon-Charles Tranchevent, Bart De Moor, and Yves Moreau, "Critical Assessment of Receptor-Ligand Predictions Methods" BMC Bioinformatics (s*ubmitted*).
- Ernesto Iacucci, Leon-Charles Tranchevent, Dusan Popovic, Bart De Moor, and Yves Moreau, "ReLianceDB: A machine learning and literature-based prioritization of Receptor─Ligand pairings" Bioinformatics (*accepted*)
- Ernesto Iacucci, Dusan Popovic, Marijke Bauters, Léon-Charles Tranchevent, Georgios A. Pavlopoulos, Guy Froyen, Bart De Moor, Yves Moreau, Beyond Enrichment Scores: Novel Strategies for Predicting Epigenetically Modified Regions, (*submitted, Genome Biology*)

Publications - *International Conference Papers*

- Iacucci E., Moreau Y., "Towards Better Receptor-Ligand Prioritization: How Machine Learning on Protein-Protein Interaction Data Can Provide Insight Into Receptor-Ligand Pairs.", in Proc. of the 20th International Conference on Artificial Neural Networks (ICANN 2010), Thessaloniki, Greece, Sept. 2010, pp. 267-271. (*In press*)

- Ernesto Iacucci, Dusan Popovic, Leon-Charles Tranchevent, Georgios Pavlopoulos, Bart De Moor, and Yves Moreau, “Prioritization of Epigenetically Modified DNA Regions” SETN 2012 (*Accepted*)

Publications - *Book Chapter*

- Iacucci E., Xavier de Souza S., Moreau Y., "Computational Approaches to Elucidating Transient Protein-Protein Interactions, Predicting Receptor-Ligand Pairings", in Chapter Protein Interaction , (Cai W., ed.), InTech (Rijeka, Croatia), 2011, pp. 1-16.

Bibliography

(2011) Evidence for network evolution in an Arabidopsis interactome map. *Science*, 333, 601-607.

Agarwal,A. et al. (1993) Synthesis, cloning and expression of a synthetic gene for high potential iron protein from Chromatium vinosum. *Biochem. Biophys. Res. Commun.*, 197, 1357-1362.

Angers,S. et al. (2000) Detection of beta 2-adrenergic receptor dimerization in living cells using bioluminescence resonance energy transfer (BRET). *Proc. Natl. Acad. Sci. U. S. A*, 97, 3684-3689.

Asur,S., Ucar,D. and Parthasarathy,S. (2007) An ensemble framework for clustering protein-protein interaction networks. *Bioinformatics.*, 23, i29-i40.

Bader,G.D. et al. (2001) BIND--The Biomolecular Interaction Network Database. *Nucleic Acids Res.*, 29, 242-245.

Bleakley,K. and Yamanishi,Y. (2009) Supervised prediction of drug-target interactions using bipartite local models. *Bioinformatics.*, 25, 2397-2403.

Bohn,B. (1980) Flow cytometry: a novel approach for the quantitative analysis of receptor--ligand interactions on surfaces of living cells. *Mol. Cell Endocrinol.*, 20, 1-15.

Boulay,G. et al. (1994) Short-term desensitization of the angiotensin II receptor of bovine adrenal glomerulosa cells corresponds to a shift from a high to a low affinity state. *Endocrinology*, 135, 2130-2136.

Boute,N., Jockers,R. and Issad,T. (2002) The use of resonance energy transfer in high-throughput screening: BRET versus FRET. *Trends Pharmacol. Sci.*, 23, 351-354.

Browne,F. et al. (2010) A knowledge-driven probabilistic framework for the prediction of protein-protein interaction networks. *Comput. Biol. Med.*, 40, 306-317.

Cascio,M. (2004a) Structure and function of the glycine receptor and related nicotinicoid receptors. *J. Biol. Chem.*, 279, 19383-19386.

Cascio,M. (2004b) Structure and function of the glycine receptor and related nicotinicoid receptors. *J. Biol. Chem.*, 279, 19383-19386.

Chang,D.T., Syu,Y.T. and Lin,P.C. (2010) Predicting the protein-protein interactions using primary structures with predicted protein surface. *BMC. Bioinformatics.*, 11 Suppl 1, S3.

Chin,C.H. et al. (2010) A hub-attachment based method to detect functional modules from confidence-scored protein interactions and expression profiles. *BMC. Bioinformatics.*, 11 Suppl 1, S25.

Congreve,M. and Marshall,F. (2010a) The impact of GPCR structures on pharmacology and structure-based drug design. *Br. J. Pharmacol.*, 159, 986-996.

Congreve,M. and Marshall,F. (2010b) The impact of GPCR structures on pharmacology and structure-based drug design. *Br. J. Pharmacol.*, 159, 986-996.

Cooper,G.F. et al. (1997) An evaluation of machine-learning methods for predicting pneumonia mortality. *Artif. Intell. Med.*, 9, 107-138.

Darnell,S.J., Page,D. and Mitchell,J.C. (2007) An automated decision-tree approach to predicting protein interaction hot spots. *Proteins*, 68, 813-823.

Edwards,B.S. et al. (2004) Flow cytometry for high-throughput, high-content screening. *Curr. Opin. Chem. Biol.*, 8, 392-398.

Ferrer,M. et al. (2003) A fully automated [35S]GTPgammaS scintillation proximity assay for the high-throughput screening of Gi-linked G protein-coupled receptors. *Assay. Drug Dev. Technol.*, 1, 261-273.

Fullar,A. et al. (2012) Tumor cell and carcinoma-associated fibroblast interaction regulates matrix metalloproteinases and their inhibitors in oral squamous cell carcinoma. *Exp. Cell Res.*, 318, 1517-1527.

Gavin,A.C. et al. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415, 141-147.

Gertz,J. et al. (2003) Inferring protein interactions from phylogenetic distance matrices. *Bioinformatics.*, 19, 2039-2045.

Gobel,J., Saussy,D.L. and Goetz,A.S. (1999) Development of scintillation-proximity assays for alpha adrenoceptors. *J. Pharmacol. Toxicol. Methods*, 42, 237-244.

Graeber,T.G. and Eisenberg,D. (2001) Bioinformatic identification of potential autocrine signaling loops in cancers from gene expression profiles. *Nat. Genet.*, 29, 295-300.

Herpin,A., Lelong,C. and Favrel,P. (2004) Transforming growth factor-beta-related proteins: an ancestral and widespread superfamily of cytokines in metazoans. *Dev. Comp Immunol.*, 28, 461-485.

Hodges,P.E. et al. (1999) The Yeast Proteome Database (YPD): a model for the organization and presentation of genome-wide functional data. *Nucleic Acids Res.*, 27, 69-73.

Hubbard,T.J. et al. (2009) Ensembl 2009. *Nucleic Acids Res.*, 37, D690-D697.

Hunter,S. et al. (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res.*, 37, D211-D215.

Iacucci,E. et al. (2011a) Predicting receptor-ligand pairs through kernel learning. *BMC. Bioinformatics.*, 12, 336.

Iacucci,E. et al. (2011b) Predicting receptor-ligand pairs through kernel learning. *BMC. Bioinformatics.*, 12, 336.

Ihmels,J., Bergmann,S. and Barkai,N. (2004) Defining transcription modules using large-scale gene expression data. *Bioinformatics.*, 20, 1993-2003.

Ilien,B. et al. (2003) Fluorescence resonance energy transfer to probe human M1 muscarinic receptor structure and drug binding properties. *J. Neurochem.*, 85, 768-778.

Issad,T., Boute,N. and Pernet,K. (2002) A homogenous assay to monitor the activity of the insulin receptor using Bioluminescence Resonance Energy Transfer. *Biochem. Pharmacol.*, 64, 813-817.

Ito,T. et al. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. U. S. A*, 98, 4569-4574.

Jacob,L. and Vert,J.P. (2008) Protein-ligand interaction prediction: an improved chemogenomics approach. *Bioinformatics.*, 24, 2149-2156.

Kerrien,S. et al. (2007) IntAct--open source resource for molecular interaction data. *Nucleic Acids Res.*, 35, D561-D565.

Kim,W.K., Bolser,D.M. and Park,J.H. (2004) Large-scale co-evolution analysis of protein structural interlogues using the global protein structural interactome map (PSIMAP). *Bioinformatics.*, 20, 1138-1150.

Kshirsagar,M., Carbonell,J. and Klein-Seetharaman,J. (2012) Techniques to cope with missing data in host-pathogen protein interaction prediction. *Bioinformatics.*, 28, i466-i472.

Kumar,R. and Thompson,E.B. (1999) The structure of the nuclear hormone receptors. *Steroids*, 64, 310-319.

Leung,H.C. et al. (2009) Predicting protein complexes from PPI data: a core-attachment approach. *J. Comput. Biol.*, 16, 133-144.

Li,M. et al. (2008) Modifying the DPClus algorithm for identifying protein complexes based on new topological structures. *BMC. Bioinformatics.*, 9, 398.

Lockless,S.W. and Ranganathan,R. (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, 286, 295-299.

Locksley,R.M., Killeen,N. and Lenardo,M.J. (2001) The TNF and TNF receptor superfamilies: integrating mammalian biology. *Cell*, 104, 487-501.

Lovell,S.C. and Robertson,D.L. (2010) An integrated view of molecular coevolution in protein-protein interactions. *Mol. Biol. Evol.*, 27, 2567-2575.

Lu,J. and Luo,L. (2008) Prediction for human transcription start site using diversity measure with quadratic discriminant. *Bioinformation.*, 2, 316-321.

Lynch,M. and Katju,V. (2004) The altered evolutionary trajectories of gene duplicates. *Trends Genet.*, 20, 544-549.

Manque,P.A. et al. (2011) Trypanosoma cruzi infection induces a global host cell response in cardiomyocytes. *Infect. Immun.*, 79, 1855-1862.

Maruyama,O. and Chihara,A. (2011) NWE: Node-weighted expansion for protein complex prediction using random walk distances. *Proteome. Sci.*, 9 Suppl 1, S14.

Mewes,H.W. et al. (2004) MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res.*, 32, D41-D44.

Milligan,G. (2004) Applications of bioluminescence- and fluorescence resonance energy transfer to drug discovery at G protein-coupled receptors. *Eur. J. Pharm. Sci.*, 21, 397-405.

Milligan,G. (2003) Constitutive activity and inverse agonists of G protein-coupled receptors: a current perspective. *Mol. Pharmacol.*, 64, 1271-1276.

Muller,V.S. et al. (2011) Membrane-SPINE: an improved method to identify protein-protein interaction partners of membrane proteins in vivo. *Proteomics.*, 11, 2124-2128.

Murakami,Y. and Mizuguchi,K. (2010) Applying the Naive Bayes classifier with kernel density estimation to the prediction of protein-protein interaction sites. *Bioinformatics.*, 26, 1841-1848.

Nagamine,N. and Sakakibara,Y. (2007) Statistical prediction of protein chemical interactions based on chemical structure and mass spectrometry data. *Bioinformatics.*, 23, 2004-2012.

O'Brien,K.P., Remm,M. and Sonnhammer,E.L. (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.*, 33, D476-D480.

Ogata,H. et al. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, 27, 29-34.

Pereira-Leal,J.B. et al. (2007) Evolution of protein complexes by duplication of homomeric interactions. *Genome Biol.*, 8, R51.

Pfleger,K.D. and Eidne,K.A. (2003) New technologies: bioluminescence resonance energy transfer (BRET) for the detection of real time interactions involving G-protein coupled receptors. *Pituitary.*, 6, 141-151.

Plewczynski,D. (2009) kNNsim: k-nearest neighbors similarity with genetic algorithm features optimization enhances the prediction of activity classes for small molecules. *J. Mol. Model.*, 15, 591-596.

Pope,A.J., Haupts,U.M. and Moore,K.J. (1999) Homogeneous fluorescence readouts for miniaturized high-throughput screening: theory and practice. *Drug Discov. Today*, 4, 350-362.

Puig,O. et al. (2001) The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods*, 24, 218-229.

Punta,M. et al. (2012) The Pfam protein families database. *Nucleic Acids Res.*, 40, D290-D301.

Qi,Y., Klein-Seetharaman,J. and Bar-Joseph,Z. (2005) Random forest similarity for protein-protein interaction prediction from multiple sources. *Pac. Symp. Biocomput.*, 531-542.

Raymer,M.L. et al. (1997) Predicting conserved water-mediated and polar ligand interactions in proteins using a K-nearest-neighbors genetic algorithm. *J. Mol. Biol.*, 265, 445-464.

Reuter,J.A. et al. (2009) Modeling inducible human tissue neoplasia identifies an extracellular matrix interaction network involved in cancer progression. *Cancer Cell*, 15, 477-488.

Reynes,C. et al. (2010) Designing focused chemical libraries enriched in protein-protein interaction inhibitors using machine-learning methods. *PLoS. Comput. Biol.*, 6, e1000695.

Rouleau,N. et al. (2003) Development of a versatile platform for nuclear receptor screening using AlphaScreen. *J. Biomol. Screen.*, 8, 191-197.

Sadee,W. et al. (1982) [3H]diprenorphine receptor binding in vivo and in vitro. *Eur. J. Pharmacol.*, 81, 431-440.

Sato,T. et al. (2006) Partial correlation coefficient between distance matrices as a new indicator of protein-protein interactions. *Bioinformatics.*, 22, 2488-2492.

Sato,T. et al. (2005) The inference of protein-protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. *Bioinformatics.*, 21, 3482-3489.

Sen,S. et al. (2002) Development of a scintiplate assay for recombinant human alpha(2B)-adrenergic receptor. *Anal. Biochem.*, 307, 280-286.

Shen,J. et al. (2007) Predicting protein-protein interactions based only on sequences information. *Proc. Natl. Acad. Sci. U. S. A*, 104, 4337-4341.

Simons,P.C. et al. (2004) Real-time analysis of ternary complex on particles: direct evidence for partial agonism at the agonist-receptor-G protein complex assembly step of signal transduction. *J. Biol. Chem.*, 279, 13514-13521.

Sklar,L.A. et al. (2002) Flow cytometric analysis of ligand-receptor interactions and molecular assemblies. *Annu. Rev. Biophys. Biomol. Struct.*, 31, 97-119.

Snider,J. et al. (2010) Split-ubiquitin based membrane yeast two-hybrid (MYTH) system: a powerful tool for identifying protein-protein interactions. *J. Vis. Exp.*

Stark,C. et al. (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, 34, D535-D539.

Stoll,D. et al. (2005) Protein microarrays: applications and future challenges. *Curr. Opin. Drug Discov. Devel.*, 8, 239-252.

Su,A.I. et al. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. U. S. A*, 101, 6062-6067.

Suykens,J.A., Vandewalle,J. and De,M.B. (2001) Optimal control by least squares support vector machines. *Neural Netw.*, 14, 23-35.

Szklarczyk,D. et al. (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.*, 39, D561-D568.

Tan,S.H. et al. (2006) A correlated motif approach for finding short linear motifs from protein interaction networks. *BMC. Bioinformatics.*, 7, 502.

Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994a) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22, 4673-4680.

Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994b) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22, 4673-4680.

Tillier,E.R. et al. (2006) Codep: maximizing co-evolutionary interdependencies to discover interacting proteins. *Proteins*, 63, 822-831.

Ullman,E.F. et al. (1996) Luminescent oxygen channeling assay (LOCI): sensitive, broadly applicable homogeneous immunoassay method. *Clin. Chem.*, 42, 1518-1526.

Urquiza,J.M. et al. (2012) Using machine learning techniques and genomic/proteomic information from known databases for defining relevant features for PPI classification. *Comput. Biol. Med.*, 42, 639-650.

van Dijk,A.D. et al. (2010) Sequence motifs in MADS transcription factors responsible for specificity and diversification of protein-protein interaction. *PLoS. Comput. Biol.*, 6, e1001017.

Vikis,H.G. and Guan,K.L. (2004) Glutathione-S-transferase-fusion based assays for studying protein-protein interactions. *Methods Mol. Biol.*, 261, 175-186.

Villiger,J.W., Boas,R.A. and Taylor,K.M. (1981) A radioreceptor assay for opiate drugs in human cerebrospinal fluid and plasma. *Life Sci.*, 29, 229-233.

Vind,B.F. et al. (2012) Hyperglycaemia normalises insulin action on glucose metabolism but not the impaired activation of AKT and glycogen synthase in the skeletal muscle of patients with type 2 diabetes. *Diabetologia*.

Volkmer,J.P. et al. (2012) Three differentiation states risk-stratify bladder cancer into distinct subtypes. *Proc. Natl. Acad. Sci. U. S. A*, 109, 2078-2083.

Wang,J. et al. (2011) A fast hierarchical clustering algorithm for functional modules discovery in protein interaction networks. *IEEE/ACM. Trans. Comput. Biol. Bioinform.*, 8, 607-620.

Willats,W.G. (2002) Phage display: practicalities and prospects. *Plant Mol. Biol.*, 50, 837-854.

Wilson,J. et al. (2003) A homogeneous 384-well high-throughput binding assay for a TNF receptor using alphascreen technology. *J. Biomol. Screen.*, 8, 522-532.

Wu,M. et al. (2009) A core-attachment based method to detect protein complexes in PPI networks. *BMC. Bioinformatics.*, 10, 169.

Xenarios,I. et al. (2000) DIP: the database of interacting proteins. *Nucleic Acids Res.*, 28, 289-291.

Yang,M. et al. (2011) Coevolution study of mitochondria respiratory chain proteins: toward the understanding of protein--protein interaction. *J. Genet. Genomics*, 38, 201-207.

Yu,C.Y., Chou,L.C. and Chang,D.T. (2010) Predicting protein-protein interactions in unbalanced data using the primary structure of proteins. *BMC. Bioinformatics.*, 11, 167.

Zanzoni,A. et al. (2002) MINT: a Molecular INTeraction database. *FEBS Lett.*, 513, 135-140.