To transfer or not to transfer – Complementing the eukaryotic protein-protein interactome

Robert Pesch and Ralf Zimmer

Institut für Informatik, Ludwig-Maximilians-Universität München, Amalienstr. 17, 80333 München

e-mail: {Robert.Pesch, Ralf.Zimmer}@bio.ifi.lmu.de

Protein interaction networks are important for the understanding of regulatory mechanisms, the prediction of protein complexes and functions. Since most interaction data is available for model organisms, the transfer of interactions to organisms of interest using ortholog relations is a common practice.

In this work, we propose to use a wide range of features to train a Random Forest in order to distinguish between correctly and incorrectly transferred interactions. The performance of our method is estimated with interaction data from *S. cerevisiae* which is currently the most complete eukaryotic protein interaction network. We show that the precision of a direct transfer of interactions from other eukaryotic interaction networks to *S. cerevisiae* is only 0.24. With our additionally introduced filter step using a rich set of features from the target and source network and the ortholog annotations we could increase the average precision to 0.76. Using a final predictor with a reduced feature set to take the sparse annotation information for other species into account we were able to increase the interactome of 97 species from currently 343,704 known interactions to 1,348,092 pair-wise interactions with an expected overall precision of 0.55.

These interaction networks can be used to study conserved functional groups, to explain experimental data, to predict conserved protein complexes, or to assist biologists with the discovery of protein interactions.

1 Introduction

Due to high-throughput screening techniques like yeast two hybrid screens, mass spectrometry and protein microarrays more and more protein interaction data is made publicly available in different databases. Such interaction data can be used to study regulatory networks, to identify protein complexes or to predict the functions of proteins (Zhang, 2009). But still the identification of interactions is a time consuming and expensive process. Therefore, most experiments focus on model organisms like *S. cerevisiae*, *E. coli* and *H. sapiens* and the interaction networks for other species are still sparse (see Table 1). To enrich these networks, interactions can be transferred using orthologs. However, as it was shown, the overlap of transferred interactions and known interactions is small (Gandhi et al., 2006).

Furthermore, false positive rates up to 50 % are reported for yeast two hybrid screens (Rhodes et al., 2005). That makes it necessary to repeat the experiments several times in order to

Species	Common name	Number of proteins	Number of interactions	Average node degree	Number of connected components
S. cerevisiae	Baker's yeast	6,067	174,320	57.46	1
H. sapiens	Human	13,695	91,742	13.39	44
D. melanogaster	Fruit fly	9,557	36,965	7.73	68
S. pombe	Fission yeast	1,996	12,404	12.42	21
C. elegans	Nematode	5,238	11,492	4.38	77
A. thaliana	Mouse-ear cress	2,953	6,285	4.25	116
M. musculus	Mouse	4,065	6,104	3.00	227
P. falciparum	-	1,134	2,219	3.91	22
R. norvegicus	Rat	1,279	1,636	2.56	97
D. rerio	Zebrafish	139	197	2.83	26

Table 1: Protein interaction network overview for the 10 eukaryotic species with the largest number of physical and genetic interactions (after interactions were mapped to UniProt) from the integrated interaction database iRefIndex. For each species the number of proteins, interactions, the average node degree and the number of connected components is given.

identify the true interaction partners. A second issue when working with interaction data is the way how this data is made publicly available. Since there is no common format or central repository, integrating interaction data is commonly performed when inferring information from protein interaction networks as seen in Kim and Tan (2010); Jaeger et al. (2010); Reiss et al. (2006). Nevertheless the integration of these databases is a challenging task since the databases have a heterogeneous structure so that the format, the description and the data structure differ among the databases (Zhang, 2009). Furthermore interactions extracted from the same PubMed abstracts can be represented differently in the databases with respect to the format and protein annotations of the involved interaction partners. These issues result in a small overlap between the interaction sets of the different databases (Turinsky et al., 2010). Finally, the current view on protein interaction networks is static, so that neither spatial nor temporal conditions are considered (Buchanan et al., 2010).

In this work, we show how the quality of a protein interaction transfer can be improved using a wide range of features derived from the interaction partners in the source network, the interaction partners in the target network and the ortholog proteins involved in the transfer. For this purpose a Random Forest (Breiman, 2001) was trained to classify correctly and incorrectly transferred interactions. Furthermore, an integrated protein interaction database was used to have the advantage that interaction data from multiple species is available and that the total set of interactions is larger compared to the individual databases.

Finally, a predictor was used to transfer interactions to 97 eukaryotic species for which ortholog mappings were available.

2 Recent work

Numerous computational approaches have been developed to assist the protein interaction identification process. Since Matthews et al. introduced the term interlog in 2001, many approaches have been developed to transfer interaction data using orthologs (Gandhi et al., 2006; Bork et al.,

2004; De Bodt et al., 2009; Michaut et al., 2008; Yu et al., 2004). In addition to ortholog relations De Bodt et al. (2009) and Michaut et al. (2008) used further features to increase the reliability of an interaction transfer. They compared random protein pairs with known protein interaction partners to define thresholds for example for the similarity of Gene Ontology (Ashburner et al., 2000) (GO) terms, the domain similarity or the gene expression similarity. With these thresholds low confident transferred interactions from different source networks were filtered out. A further common practice is to require a certain sequence similarity between orthologs in order to transfer an interaction. For example Yu et al. (2004) found out that protein interactions can be safely transferred if the joint sequence similarity between the ortholog proteins involved in the transfer is >80 %. Besides that, many different approaches try to predict interactions using structural properties (Ogmen et al., 2005), network topology information (Pao-Yang Chen, 2008), or protein domain information (Luo et al., 2011).

3 Material and methods

3.1 Used databases

We used the protein interaction repository iRefIndex (Razick et al., 2008) as source database, which provides interaction data from multiple protein interaction databases in a common format. From this database physical as well as genetic interactions were transferred using publicly available ortholog mappings. In this work, orthologs from the Orthologs Matrix Project (OMA) (Schneider et al., 2007), InParanoid (Remm et al., 2001) and HomoloGene[1] were used. These databases were used due to their good evaluation results (Altenhoff and Dessimoz, 2009) and the huge coverage of ortholog mappings for different eukaryotic species.

The interaction partners and orthologs were mapped to UniProt (Consortium, 2011) to have on the one hand a common representation of the protein set and on the other hand a rich annotation set including GO terms, synonyms and mappings to external databases like KEGG (Kanehisa et al., 2010).

3.2 Methods

Protein interaction networks were modeled as weighted graph $\mathrm{PPI} = (P, I, \Phi)$ consisting of a set of proteins (P), interactions ($I \subseteq P \times P$) and edge weights Φ (which assigns a weight to each edge $\Phi(i) \rightarrow \mathbb{N}$, $i \in I$). In our case, edges were weighted with confidence values representing the number of publications that support the interaction.

Given an interaction network $PPI_i = (P_i, I_i, \Phi_i)$, a target protein set P_j and an ortholog mapping $O \subseteq P_i \times P_j$ between P_i and P_j, a transferred interaction network consists of $PPI_j = (P_j, I_j, \Phi)$ with $(p_{j,k}, p_{j,c}) \in I_j \iff (p_{i,v}, p_{i,l}) \in I_i, (p_{j,k}, p_{i,v}) \in O, (p_{j,c}, p_{i,l}) \in O$.

A protein $p_{i,h} = (t_{i,h}, g_{i,h}, f_{i,h})$ from a protein set P_i (for example representing the set of proteins for a species) consists of a set of function terms ($t_{i,h} = \{t_{i,h,1}, \ldots\}$), GO terms ($g_{i,h} = \{g_{i,h,1}, \ldots\}$) and family memberships ($f_{i,h} = \{f_{i,h,1}, \ldots\}$). Function terms were derived from protein synonyms by tokenizing, stemming and filtering stop words and to general words resulting in a set of tokens which were most descriptive for the protein.

For the classification of correctly and incorrectly transferred interactions a Random Forest (Breiman, 2001) from the WEKA (Hall et al., 2009) machine learning framework was trained which predicts the outcome class of an instance by using a voting procedure on multiple learned decision trees with different feature sets. Random Forests have shown good evaluation results

[1] http://www.ncbi.nlm.nih.gov/homologene

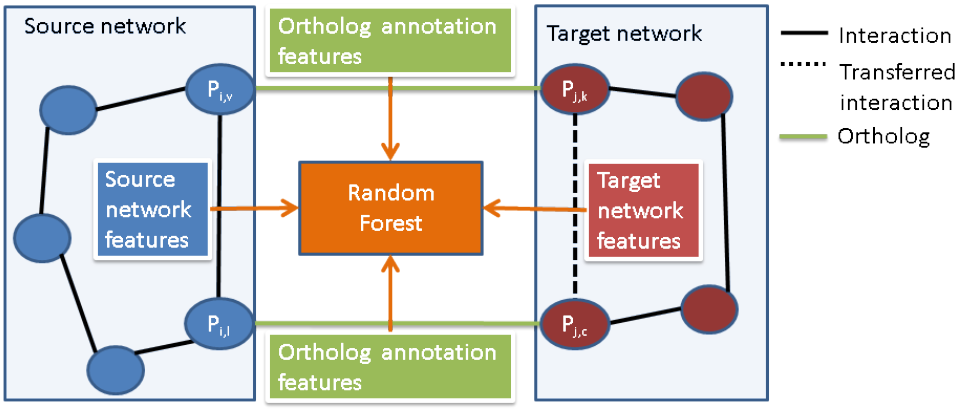

Figure 1: Schematic diagram of the protein interaction classification. Interactions are transferred from a source network to a target network. In order to distinguish between correctly and incorrectly transferred interactions a Random Forest is used for the classification which uses features from the interacting partner in the source network, the interaction partners in the target network and the ortholog proteins.

on similar learning tasks (Caruana and Mizil, 2006) and are more robust against noise than other ensemble machine learning methods (Breiman, 2001).

To assess the quality of the learned model the

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}}, \quad \text{Recall} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}}, \tag{1}$$

and

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2}$$

scores were computed to evaluate the quality of the learned predictor.

3.2.1 Features

As features we used the protein annotations of the interacting partners in the source and the target network and the ortholog proteins from which an interaction was transferred.

GO similarity: The semantic GO term similarity between the interaction partners in the source and target network and the ortholog proteins (Couto et al., 2007).

Network overlap: The overlap of the neighborhood proteins for a given pair of proteins in the source and target network. For this purpose the Jaccard index was computed.

Network GO similarity: The average GO similarity between the pair-wise neighbors of the interaction partners in the networks.

Total support: The number of times an interaction was transferred from all other networks to the target network (as suggested by Mika and Rost (2006) for confidence scoring).

Gene expression correlation coefficient: Given a gene expression time series $x = \{x_1 \dots x_n\}$ and $y = \{y_1 \dots y_n\}$ for two genes g_x and g_y, the Pearson correlation coefficient was computed. The genes were mapped to the respective proteins using the gene expression annotation files.

Sequence identity: The sequence identity of the ortholog proteins.

Token similarity: The inverse likelihood that a different pair of proteins have the same synonym token intersection as the ortholog protein pair.

Domain/Family similarity: Similar to the token similarity the inverse likelihood that two different proteins have the same family and domain annotations in common.

Transitive ortholog: The idea behind this feature is that more conserved orthologs can be traced from a source species to a target species along a phylogenetic tree. For this purpose a phylogenetic tree covering all species with ortholog mapping was used. Given such a tree, a path from a source to a target species was computed by:

1. searching the shortest path between the two species,
2. searching the closest leaf nodes for all inner nodes on the shortest path.

The result is a list of species which are between the target and the source species. An ortholog is defined as transitive consistent if a direct ortholog between the source and the target species can also be reached when going along the pair wise ortholog mappings on the estimated path.

3.3 Experimental settings

The *S. cerevisiae* network was used as gold standard. The training set consisted of 27,410 transferred interactions from all eukaryotic species considered in this study to *S. cerevisiae*. A transferred interaction was assumed to be correct if the interaction could be found in the *S. cerevisiae* network from iRefIndex. 6,934 of the transferred interactions could be validated in the network and the other 20,476 interactions were used as negative set.

The features described in section 3.2.1 were modeled for the protein pairs involved in the transfer. In total 4 proteins were considered for the transfer (two proteins from the source network and two proteins from the target network). The features were modeled between the different protein pairings in the target network, in the source network and between the ortholog proteins (see Figure 1). In Table 2 the mapping between the features and protein pairings is given. In total 19 features were modeled where for the features used for the ortholog proteins one feature for each of the two ortholog protein pairs involved in the transfer was created. For example for the GO similarity one feature was modeled between the interaction partner in the source network, one feature was modeled between the interaction partner in the target network and two features were modeled between the ortholog proteins involved in the transfer. For the gene expression feature the processed expression intensity values from the experiment E-GEOD-5376 in ArrayExpress (Parkinson et al., 2009) were used.

Two experimental settings were constructed to train the Random Forest. One setting in which all features were considered and one setting where only features were used which can be assumed to be available for most of the species. Hence, features containing information about the network structure and the gene expression correlation were excluded in the reduced feature set. The performance of the Random Forest trained with the two feature sets was estimated using a 10-fold-cross validation.

	GO similarity	Network overlap	Network GO similarity	Total support	Edge support	Gene expression	Sequence identity	Token similarity	Domain similarity	Transitive orthologs
Target Network	✓	✓	✓	✓		✓				
Source Network	✓	✓	✓		✓					
Orthologs	✓						✓	✓	✓	✓
Full set	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Reduced set	✓			✓	✓		✓	✓	✓	✓

Table 2: The table lists the features used for the proteins in the target network, the source network, and the orthologs, respectively. It also defines the two sets of features used in the analysis (full set and reduced set).

4 Results and discussion

4.1 Current protein interaction networks

In Table 1 an overview of the protein interaction networks with most interactions is summarized. This interaction data is time and location independent and comes from many different yeast two hybrid screens all showing a high false positive rate. With over 170,000 interactions the most complete eukaryotic interaction network is available for *S. cerevisiae*.

Especially in comparison with the second largest protein interaction network from *H. sapiens* it becomes clear how sparse the networks for the other species still are. The *H. sapiens* network has 2.41 times less interactions and 1.84 more proteins in the network.

Furthermore only the *S. cerevisiae* network consists of exactly one connected component. Therefore, we assume in the following that the *S. cerevisiae* network is almost complete and use this network to evaluate the performance of a protein interaction transfer.

Selecting the *S. cerevisiae* network as gold standard has the disadvantage that protein interactions for higher evolved species are might only be studied because of the prior knowledge that an interlog in *S. cerevisiae* exist. This bias may results in higher accuracies for the interaction transfer from distance species. Nevertheless the *S. cerevisiae* network was used as gold standard because of the high number of available interactions and the availability of curated annotations for most of the proteins.

4.2 Direct protein interaction transfer

The direct transfer achieves only a low precision. Figure 2 shows the precision of the interaction transfers from the interaction networks of Table 1 to *S. cerevisiae* and *H. sapiens*. The overall precision of an interaction transfer to *S. cerevisiae* is 0.24 where most of the interactions were transferred from *H. sapiens* as the second largest interaction network. For a transfer to *H. sapiens* only 8 % of the transferred interactions could be validated.

Given complete interaction data it can be expected that the highest precision can be achieved with a transfer from the phylogenetic closest species. But since the interaction data is sparse and

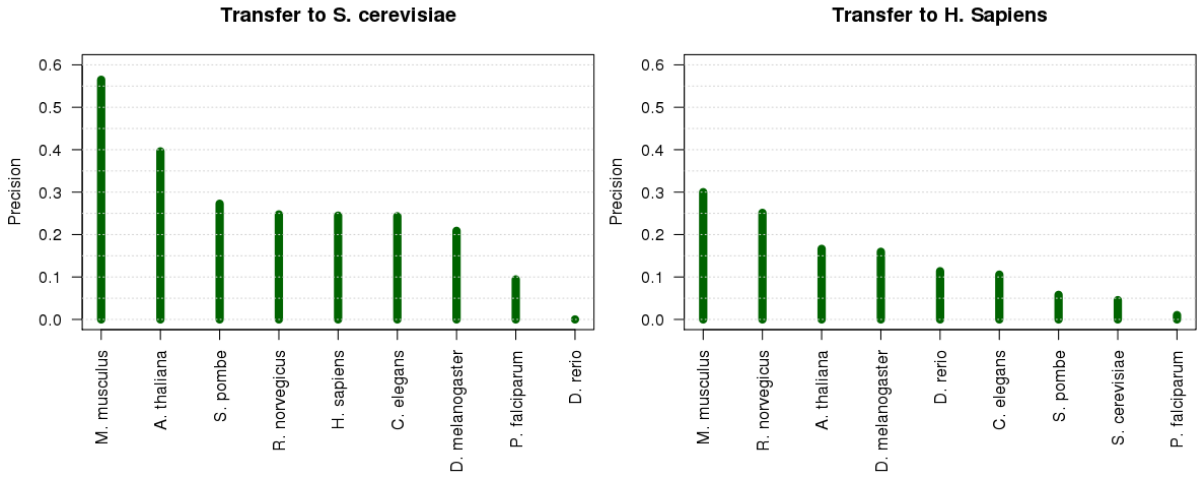

Figure 2: Precision of a direct interaction transfer from 9 eukaryotic species with interaction data to *S. cerevisiae* and *H. sapiens*. The precision ranges between 0.56 and 0 for the two species. The average weighted precision for the transfer to *S. cerevisiae* is 0.24, whereas the average precision for a transfer to *H. sapiens* is 0.08.

interlogs of *S. cerevisiae* were might be used as prior knowledge for the interaction discovery a different order could be observed. Most notable is the performance of a transfer from *M. musculus* to *S. cerevisiae* with an unusually high precision of 0.56. A GO overrepresentation analysis (DAVID, Huang et al. (2009)) showed that some highly conserved processes like DNA-dependent DNA replication, pre-replicative complex assembly, DNA replication initiation and chromosome organization are involved, which might explain the high precision of the interaction transfer.

4.3 Protein interaction filter

Using the features described in Section 3.2.1, a Random Forest with full feature set and one with reduced feature set was trained to distinguish between correctly and incorrectly transferred interactions. A recall of 1 is defined in the case that all correctly transferable interactions were predicted as correct interaction. Thus with a direct transfer a precision of 0.24 and a recall of 1 can be reached, resulting in a F_1 score of 0.39.

In Figure 3 the performance of the two Random Forests is shown by using a 10-fold-cross validation on the entire training set. Combining all features a F_1 score of 0.66 can be reached with a precision for a correct transfer of 0.76 and a recall of 0.58.

The strongest feature is the GO similarity between the interaction partners in the target network with an information gain (Mitchell, 1997) of 0.17. But also the network overlap feature in the target network, which was excluded in the reduced feature set, has a comparable information gain of 0.15. Therefore the F_1 score for the Random Forest with reduced feature set drops to 0.62 with a precision of 0.7 and a recall of 0.56. Compared to the model trained with the full feature set mostly the precision decreases. With a recall of 0.58 and 0.56, respectively, almost half of the correctly transferred interactions with an unfiltered transfer got classified as incorrectly transferred and thus rejected for the transfer. To increase the recall the score threshold for the Random Forest for predicting an interaction as correctly transferred was reduced to 0.3.

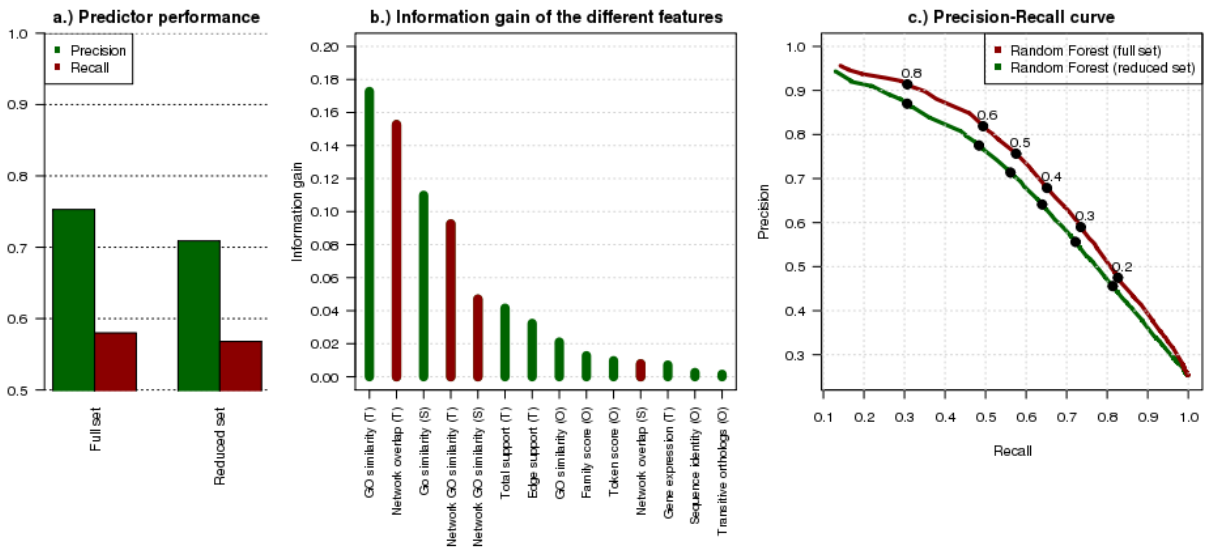

Figure 3: a.) Precision and recall for the Random Forests trained with the full and the reduced feature set. **b.)** Information gain of the different features. The letter T after the feature name stands for the protein pair in the target network, the letter S for the protein pair in the source network and O for the ortholog proteins. For ortholog protein features the average information gain of the two ortholog partners is shown. The features colored in red are exluced in the reduced set. **c.)** Precision-Recall curve for the Random Forests trained with the two feature sets. The black points visualize the score threshold of the Random Forests for predicting an interaction as correctly transferred.

With such a score threshold the precision decreases for the Random Forest with reduced feature set from 0.7 to 0.55, but the recall increases to 0.72 (see Precision-Recall curve in Figure 3).

To assess the quality of the trained predictor, the one with reduced feature set was applied to the transfer using the *H. sapiens* network as gold standard. When transferring interactions directly to *H. sapiens* a precision of 0.08 could be reached (from the 143,378 transferred interactions 11,358 could be validated in the network). With the trained model the precision could be improved to 0.14 (from 52,296 interaction after the filtering, 7,504 could be found in the network). This huge drop in precision was expected compared to the *S. cerevisiae* network transfer, since the *H. sapiens* network is still incomplete as indicated by the number of connected components and the average node degree compared to *S. cerevisiae*. Therefore, the test scenario using the *H. sapiens* network as gold standard is not very meaningful to estimate the absolute precision, but may it is illustrative to access the relative performance in comparison to the direct transfer.

4.4 Comparison with other interaction transfer methods

Comparing the transfer quality against other methods is difficult because of the lack of gold standards and because different publicly available data sets provide transferred interactions for different target species. Nevertheless in Figure 4 the intersections of predicted protein interactions from different data sets and a set of experimentally discovered physical protein interactions from *D. melanogaster* are shown. *D. melanogaster* was chosen since most method predicted interlogs for this species and some experimentally data is available. As dataset Yu et al. (2004)[2] predicted interactions with a joint E-Value threshold of $< 10^{-70}$, all predicted

[2] http://interolog.gersteinlab.org/

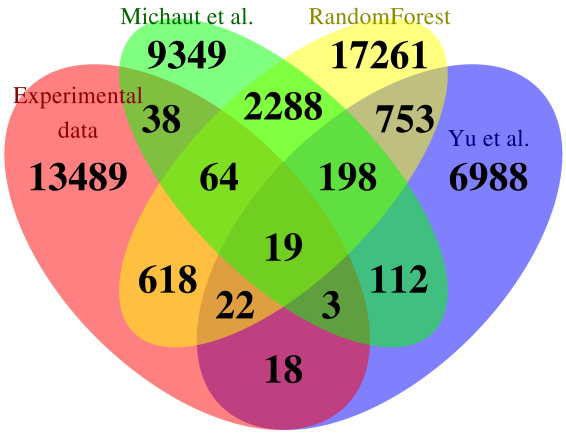

Figure 4: Venn-Diagram of predicted and experimentally discovered physical protein interactions from iRefIndex, Yu et al. (2004), Michaut et al. (2008) and the presented approach using a Random Forest to distinguish between correctly and incorrectly transferred interactions. Proteins were mapped to gene identifier in order to have a common representation of the entities in the different data sources.

physical interactions from Michaut et al. (2008)[3] and all predicted interactions from the method presented here with a threshold of ≥ 0.3 were used.

In general the intersections between the sets are small. The highest precision of 0.03 between the predicted interaction sets and the experimentally discovered set can be reached with the method presented here (from 21.223 interaction 723 could be validated). Between the predicted sets the highest precision can again be reached between the method presented here and the method of Michaut et al. (2008). These low agreements can partly be explained with the fact that different source species were used for the transfer, different ortholog identification methods were applied and that the experimentally discovered data is sparse.

4.5 Enriched protein interaction networks

Using the trained predictor with reduced feature set, interactions were transferred and classified for all eukaryotic species with available ortholog mappings. With a direct interaction transfer the interactome of these 97 species could be increased from currently 343,699 interactions to 4,327,054 interactions. With the additional filter step still 1,348,092 pair-wise interaction are left. Estimated from the transfer to *S. cerevisiae* a precision of 0.55 for this interaction transfer can be expected.

The resulting interactome is shown in Figure 5 for 40 species. The transfer relies on the 1.) availability of ortholog relations 2.) mappings of the orthologs to UniProt entries and 3.) annotations of the UniProt entries. These requirements imply that for some species only few interactions could be transferred. On the other side the interaction networks of *M. musculus*, *D. rerio* and *B. taurus* could be enriched with over 50,000 additional interactions.

5 Conclusion

Using publicly available ortholog and protein interaction data the protein networks of 97 eukaryotic species could be enriched. Transferring interactions directly from one species to another

[3] http://biodev.extra.cea.fr/interoporc/

Figure 5: Enriched protein interactome for 40 species using the trained Random Forest with reduced feature set to distinguish between correctly and incorrectly transferred interactions. The color schema indicates the number of interactions in each interaction network after the transfer. According to the estimated precision for the transfer to *S. cerevisiae* a precision of 0.55 can be expected.

results in a low consistency as measured with the interaction transfer to *S. cerevisiae* which is the most complete eukaryotic interaction network. Therefore, to increase the precision of a transfer, a Random Forest was trained to distinguish between correctly and incorrectly transferred interactions. The model was trained with a rich feature set covering features from the interacting partners in the source and target network and the orthologs from which an interaction was transferred. With this additional classifier the average precision of a transfer to *S. cerevisiae* could be improved from 0.24 to 0.76. Since rich protein annotations are not available from all species, a final predictor with reduced feature set was trained. This classifier was used to enrich the interaction networks of 97 eukaryotic species with an expected precision of 0.55. The so created networks can be used to study conserved functional groups, to explain experimental data, to predict conserved protein complexes, or to assist biologists with the protein interaction discovery.

Due to the increasing speed of protein annotations added to UniProt and the increased usage of yeast two hybrid screens for other species, it can be expected that these networks can be even further enriched in the future and that more features can be used for the transfer so that a higher transfer precision can be achieved.

The transferred networks will be made publicly available via a web service including query and visualization capabilities.

References

Altenhoff, A. M. and Dessimoz, C. (2009). Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Computational Biology*, 5(1):e1000262+.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29.

Bork, P., Jensen, L. J., von Mering, C., Ramani, A. K., Lee, I., and Marcotte, E. M. (2004). Protein interaction networks from yeast to human. *Current Opinion in Structural Biology*, 14(3):292–299.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Buchanan, M., Caldarelli, G., Rios, P. D. L., Rao, F., and Vendruscolo, M., editors (2010). *Networks in Cell Biology*. Cambridge University Press, 1 edition.

Caruana, R. and Mizil, A. N. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168.

Consortium, T. U. (2011). Ongoing and future developments at the universal protein resource. *Nucleic Acids Research*, 39(suppl 1):D214–D219.

Couto, F., Silva, M., and Coutinho, P. (2007). Measuring semantic similarity between gene ontology terms. *Data & Knowledge Engineering*, 61(1):137–152.

De Bodt, S., Proost, S., Vandepoele, K., Rouze, P., and Van de Peer, Y. (2009). Predicting protein-protein interactions in arabidopsis thaliana through integration of orthology, gene ontology and co-expression. *BMC Genomics*, 10(1):288+.

Gandhi, T. K. B., Zhong, J., Mathivanan, S., Karthick, L., Chandrika, K. N., Mohan, S. S., Sharma, S., Pinkert, S., Nagaraju, S., Periaswamy, B., Mishra, G., Nandakumar, K., Shen, B., Deshpande, N., Nayak, R., Sarker, M., Boeke, J. D., Parmigiani, G., Schultz, J., Bader, J. S., and Pandey, A. (2006). Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nature Genetics*, 38(3):285–293.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1):10–18.

Huang, D. W. a. . W., Sherman, B. T., and Lempicki, R. A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1):1–13.

Jaeger, S., Sers, C. T., and Leser, U. (2010). Combining modularity, conservation, and interactions of proteins significantly increases precision and coverage of protein function prediction. *BMC Genomics*, 11(1):717+.

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., and Hirakawa, M. (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research*, 38(Database issue):D355–D360.

Kim, J. and Tan, K. (2010). Discover protein complexes in Protein-Protein interaction networks using parametric local modularity. *BMC Bioinformatics*, 11(1):521+.

Luo, Q., Pagel, P., Vilne, B., and Frishman, D. (2011). DIMA 3.0: Domain interaction map. *Nucleic Acids Research*, 39(Database issue):D724–D729.

Matthews, L. R., Vaglio, P., Reboul, J., Ge, H., Davis, B. P., Garrels, J., Vincent, S., and Vidal, M. (2001). Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Research*, 11(12):2120–2126.

Michaut, M., Kerrien, S., Montecchi-Palazzi, L., Chauvat, F., Cassier-Chauvat, C., Aude, J.-C., Legrain, P., and Hermjakob, H. (2008). InteroPORC: automated inference of highly conserved protein interaction networks. *Bioinformatics*, 24(14):1625–1631.

Mika, S. and Rost, B. (2006). Protein-protein interactions more conserved within species than across species. *PLoS Computational Biology*, 2(7):e79+.

Mitchell, T. (1997). *Machine Learning*. McGraw Hill Higher Education, 1st edition.

Ogmen, U., Keskin, O., Aytuna, A. S., Nussinov, R., and Gursoy, A. (2005). PRISM: protein interactions by structural matching. *Nucleic Acids Research*, 33(suppl 2):W331–W336.

Pao-Yang Chen, Charlotte M. Deane, G. R. (2008). Predicting and validating protein interactions using network structure. *PLoS Computational Biology*, 4(7):e1000118.

Parkinson, H., Kapushesky, M., Kolesnikov, N., Rustici, G., Shojatalab, M., Abeygunawardena, N., Berube, H., Dylag, M., Emam, I., Farne, A., Holloway, E., Lukk, M., Malone, J., Mani, R., Pilicheva, E., Rayner, T. F., Rezwan, F., Sharma, A., Williams, E., Bradley, X. Z. Z., Adamusiak, T., Brandizi, M., Burdett, T., Coulson, R., Krestyaninova, M., Kurnosov, P., Maguire, E., Neogi, S. G. G., Rocca-Serra, P., Sansone, S.-A. A., Sklyar, N., Zhao, M., Sarkans, U., and Brazma, A. (2009). ArrayExpress update–from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Research*, 37(Database issue):D868–D872.

Razick, S., Magklaras, G., and Donaldson, I. M. (2008). iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinformatics*, 9(1):405.

Reiss, D. J., Baliga, N. S., and Bonneau, R. (2006). Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *BMC Bioinformatics*, 7(1):280+.

Remm, M., Storm, C. E., and Sonnhammer, E. L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of Molecular Biology*, 314(5):1041–1052.

Rhodes, D. R., Tomlins, S. A., Varambally, S., Mahavisno, V., Barrette, T., Kalyana-Sundaram, S., Ghosh, D., Pandey, A., and Chinnaiyan, A. M. (2005). Probabilistic model of the human protein-protein interaction network. *Nature Biotechnology*, 23(8):951–959.

Schneider, A., Dessimoz, C., and Gonnet, G. H. (2007). OMA browser–exploring orthologous relations across 352 complete genomes. *Bioinformatics*, 23(16):2180–2182.

Turinsky, A. L., Razick, S., Turner, B., Donaldson, I. M., and Wodak, S. J. (2010). Literature curation of protein interactions: measuring agreement across major public databases. *Database*, 2010:baq026.

Yu, H., Luscombe, N. M., Lu, H. X. X., Zhu, X., Xia, Y., Han, J.-D. D., Bertin, N., Chung, S., Vidal, M., and Gerstein, M. (2004). Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Research*, 14(6):1107–1118.

Zhang, A. (2009). *Protein Interaction Networks: Computational Analysis*. Cambridge University Press, 1 edition.