The Quality of Protein-Protein Interactions Computationally Extracted from Literature and the Impact on PPI-Applications (Exposé)

Sebastian Arzt arzt@informatik.hu-berlin.de

March 30, 2011

Academic advisors: Prof. Dr. Ulf Leser leser@informatik.hu-berlin.de
Philippe Thomas thomas@informatik.hu-berlin.de
Samira Jaeger sjaeger@informatik.hu-berlin.de

1 Introduction

Proteins interact with each other (e.g. in form of protein complexes) to accomplish a biological function [1]. Therefore, exploring the set of all potential protein-protein interactions, known as the interactome, is essential to understand various biological processes within organisms. This knowledge is fundamental for further studies such as revealing mechanisms behind diseases [2].

One way to detect protein-protein interactions are high-throughput experiments. They detect all pairwise interactions within a set of proteins (e.g. yeast two-hybrid method [3]) or identify the participants of protein complexes (e.g. co-affinity purification method [4]).

A different approach to obtain protein-protein interaction data is the manual curation of scientific literature describing interactions between proteins detected in small-scale experiments. It is assumed that currently most information about protein-protein interactions can only be found in such literature [5]. However, manual curation will not be able to cope with the rapidly increasing number of publications [6] since it is highly time-consuming. Moreover, manual information extraction is an error-prone task resulting in a lower data quality than commonly assumed [7]. Consequently, there is an increasing need for high-quality computational support.

In this work, we exemplarily evaluate whether state-of-the-art Natural Language Procession tools are capable of meeting the requirements. On that account we analyze the quality of protein-protein interactions computationally extracted from scientific literature by comparing them against protein-protein interactions from manually curated biological pathways. From the results we derive which characteristics of the relation extraction (i.e. confidence-score of the classifier, numbers of sentences supporting the relation, full-text or abstract corpus) and which characteristics of the pathways (i.e. size, species, type) induce a high and low coverage, respectively.

Beyond that, we will determine if practical applications benefit from computationally extracted interaction data by comparing the outcomes of a simple protein function prediction method using, on the one hand, interaction data extracted from literature, and on the other hand, data collected from curated protein-protein interaction databases.

2 Related work

In [8] Brosy proposed an approach for evaluating relation extraction by reconstructing curated pathways using protein-protein interactions extracted from the literature. For that account, she initially extracted interactions from a corpus containing about 13 million freely accessible scientific abstracts from PubMed[1]. Subsequently, she compared her data to pathway data from KEGG [9]. Brosy was able to completely reconstruct 2% from the analyzed pathways. However, the overall results show a poor reconstruction quality. From the results she

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

concluded that a potential reason is the limitation on abstracts since, generally, they do not contain full research outcomes.

In [10] Rodríguez-Penagos et al. perform a reconstruction of the regulatory network of E. coli proteobacteria using a rule-based relation extraction of relevant interactions (i.e. transcriptional regulations in E. coli) from abstracts and full-text papers. Their work attempts to estimate the benefit of natural language processing for supporting the manual annotation process. Similar to the approach of Brosy binary relations between pairs of genes/proteins were extracted. RegulonDB [11] which contains manually curated data sets was used as gold standard. The comparison against RegulonDB resulted in an overall f-measure of 57% with a precision and recall of 77% and 45%, respectively.

In [12] Ni et al. study the influence of the quality of protein interaction data on protein function prediction. Therefore, they modify two simple prediction methods, namely the neighbour counting [13] and chi-square method [14]. At first they assign each interaction a quality level depending on the number of methods that detected it (i.e. *high*: at least two different methods, *normal*: one method, with minimal three occurrences, *low*: remaining PPIs). To incorporate the quality levels into the function prediction they modify the scoring functions of each method by weighting the interactions accordingly. To determine the optimal weights for each level they performed a grid search based on a leave-one-out cross-validation. Their results show that weighting positively effects the performance of the function prediction methods. Compared to the unweighted methods the ROC score of the weighted neighbor counting method and weighted chi square method increases from 0.763 to 0.7701 and from 0.7202 to 0.7669, respectively.

3 Materials and methods

3.1 Extraction pipeline

The extraction of protein-protein interactions from the corpora is done in two steps: At first a Named Entity Recognition on gene names is executed. This step utilizes the tool GNAT [15], since it performes well (F-measure: 81,4) on multi-species corpora. Additionally, GNAT performs a Named Entity Normalization by assigning a Entrez Gene [16] identifier to each extracted entity.

Prior to the relation extraction the text will be enriched with linguistic information essential for the relation extraction step. The preprocessing consists of tokenization, sentence detection, part-of-speech tagging and lemmatization.

The second step, the relation extraction, is performed using jSRE [17], a freely available tool for relation extraction based on support vector machines. The extracted interaction data and additional meta-data is stored in a local database. The training corpora for jSRE consists of five corpora [18], namely: AIMed, BioInfer, HPRD50, IEPA and LLL. See Table 1 for the most important characteristics for each corpus.

Corpus	Sentences	Positive pairs	Negative pairs
AImed	1955	1000	4834
BioInfer	1100	2534	7132
HPRD50	145	163	270
IEPA	486	335	482
LLL	77	164	166
Total:	3763	4196	12884

Table 1: Overview of the most important statistics of the training corpora listing the number of sentences and the number of positive/negative interaction pairs per corpus.

3.2 PPI extraction corpus

The corpora used for the extraction of protein-protein interactions are:

- a corpus of full-text publications from PubMed Central[2] provided through open access
- a corpus consisting of freely accessible abstracts obtained from PubMed

3.3 ID-mapping

The Named Entity Normalization performed by GNAT yields an Entrez Gene identifier for each gene. In contrast, our local database integrates the data from pathway databases based on UniProt [19] identifier for each protein. Hence, in order to recover the extracted interactions we have to perform a mapping from UniProt to Entrez Gene identifier. This step has to be executed with care since neither every UniProt identifier can be mapped to Entrez Gene nor this mapping is unique.

3.4 Function prediction

To predict the function of a protein we apply the chi-square method originally proposed by Hishigaki et al. [14]. The method is based on the observation that interacting proteins are likely to have the same function. To infer the function of a novel protein one considers the set of known functions of its neighbors in the interaction network. A protein inherits the function i with the highest score denoted by

$$\chi_i^2 = \frac{(n_i - e_i)^2}{e_i}$$

where n_i and e_i are the frequency of function i among the neighbors and the network-wide expected frequency, respectively.

[2] `http://www.ncbi.nlm.nih.gov/pmc/`

4 Evaluation

The evaluation of the reconstruction is based on curated pathways from the databases KEGG and Reactome [20] (See Table 3). In order to determine precision and recall we first define the sets of true/false positives and false negatives among the predicted interactions.

Let $\mathbf{P} = \{p_1, \ldots, p_n\}$ denote the set of proteins represented by UniProt identifiers. Thus, we represent protein-protein interactions as pairs $(p_1, p_2) \in \mathbf{P} \times \mathbf{P}$. A pathway P_i is interpreted as a set of protein-protein interactions, such that $P_i \subset P \times P$. Moreover, let $p : \mathcal{P}(\mathbf{P} \times \mathbf{P}) \to \mathbf{P}$ determine the proteins involved in a pathway (e.g. $p(\{(p_1, p_2), (p_2, p_3)\}) = \{p_1, p_2, p_3\}$). Additionally, $P_{i|P_j}$ denotes the restriction of P_i on P_j removing all interactions from pathway P_i whose involved proteins are not enclosed in pathway P_j:

$$P_{i|P_j} := \{(p_k, p_l) \mid (p_k, p_l) \in P_i \ \wedge \ p_k, p_l \in p(P_j)\} \tag{1}$$

The protein-protein interactions extracted from full-texts are represented by the relation $E \subset P \times P$, which can be interpreted as a single artificial pathway. Hence, the number of *true-positives* (TP_i), *false-positives* (FP_i) and *false-negatives* (FN_i) among the PPIs of the reconstruction of P_i are defined in the following manner:

$$TP_i := |E_{|P_i} \cap P_i| \tag{2}$$

$$FP_i := |E_{|P_i} \backslash P_i| \tag{3}$$

$$FN_i := |P_i \backslash E_{|P_i}| \tag{4}$$

The corresponding accumulated values for a global evaluation (micro average) of the reconstruction are defined as follows,

$$TP := \sum_{i=1}^{m} TP_i \qquad FP := \sum_{i=1}^{m} FP_i \qquad FN := \sum_{i=1}^{m} FN_i \tag{5}$$

where m is the number of available pathways. The performance of the pathway reconstruction may be optimized by filtering interactions by adjusting the parameters described in Table 2.

Using this evaluation framework we strive for answering the following questions:

- Does the reconstruction based on full texts perform better than using abstracts only?
- For which species does a species-specific reconstruction perform better?
- Does the performance depend on the size of the pathway?
- Which parameter induce the best performance?
- Which are the pathways covered best/worst?

It is noteworthy to mention that false positives among the predicted data may correspond to potential true positive protein-protein interactions. Hence, false positives are still useful to support manual curation.

Parameter	Description
f_s	Filter out interactions mentioned in less than f_s sentences.
f_c	Filter out interactions whose confidence value is lower than f_c.
f_m	Filter out interactions if a protein was multiplied more frequent than f_m times during id-mapping.

Table 2: Overview of the filter parameters to adjust the performance of the pathway reconstruction

Type	Database	PPIs
PPIs	BioGRID	70995
	Database of Interacting Proteins	28310
	Human Protein Reference Database	39097
	IntAct	53594
	Mammalian Protein-Protein Interaction Database	700
	Molecular Interactions Database	38504
	Total	231200
Pathways	KEGG	2550
	Reactome	156174
	Total	158724

Table 3: Number of imported binary interactions per database

4.1 Function prediction

Since we are interested in the impact of the extracted interactions on the quality of function prediction we compare the respective outcomes based on three different interaction networks:

- The first network includes the protein-protein interactions extracted from literature by GNAT/jSRE.
- The second one consists of manually curated interaction data imported from following interaction databases:
  - Database of Interacting Proteins [21]
  - Human Protein Reference Database [22]
  - IntAct [23]
  - MIPS Mammalian Protein-Protein Interaction Database [24]
  - Molecular Interactions Database [25]

  Table 3 lists the number of protein-protein interactions imported from each database.
- To specify the contribution of the extracted protein-protein interactions to the manual curated interaction data the third interaction network combines both sources.

The performance of the function prediction based on these interaction networks is evaluated in the following way: As gold standard, each protein is annotated with its Gene Ontology [26] terms obtained from the UniProt knowledge base. Then, we determine the set P of proteins occurring in the intersection of the extracted and the imported protein-protein interactions. The functions of each protein in P are predicted using the (weighted) chi-square method. The resulting set of annotations per protein is then compared to its gold standard annotation to compute precision, recall and f-measure, accordingly.

The concept of weighting interactions as proposed by Ni et al. is also applicable in our context since the reliability of an extracted interaction depends on the confindence score of jSRE and the number of supporting sentences. To estimate the effect of weighting on the performance we repeat each run with weighting enabled.

References

[1] J. De Las Rivas and C. Fontanillo. Protein–Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks. *PLoS Computational Biology*, 6(6), 2010.

[2] A. Bauer-Mehren, L.I. Furlong, and F. Sanz. Pathway databases and tools for their exploitation: benefits, current limitations and challenges. *Molecular systems biology*, 5(1), 2009.

[3] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America*, 98(8):4569, 2001.

[4] O. Puig, F. Caspary, G. Rigaut, B. Rutz, E. Bouveret, E. Bragado-Nilsson, M. Wilm, and B. Séraphin. The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods*, 24(3):218–229, 2001.

[5] T. Ono, H. Hishigaki, A. Tanigami, and T. Takagi. Automated extraction of information on protein–protein interactions from the biological literature. *Bioinformatics*, 17(2):155, 2001.

[6] L. Hunter and K.B. Cohen. Biomedical Language Processing: Perspective What's Beyond PubMed? *Molecular cell*, 21(5):589, 2006.

[7] M.E. Cusick, H. Yu, A. Smolyar, K. Venkatesan, A.R. Carvunis, N. Simonis, J.F. Rual, H. Borick, P. Braun, M. Dreze, et al. Literature-curated protein interaction datasets. *nature metHods*, 6(1):39–46, 2008.

[8] F. Brosy. Rekonstruktion biologischer Pathways aus sehr großen Korpora. 2009.

[9] M. Kanehisa and S. Goto. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27, 2000.

[10] C. Rodríguez-Penagos, H. Salgado, I. Martínez-Flores, and J. Collado-Vides. Automatic reconstruction of a bacterial regulatory network using Natural Language Processing. *BMC bioinformatics*, 8(1):293, 2007.

[11] H. Salgado, S. Gama-Castro, A. Martínez-Antonio, E. Díaz-Peredo, F. Sánchez-Solano, M. Peralta-Gil, D. Garcia-Alonso, V. Jiménez-Jacinto, A. Santos-Zavaleta, C. Bonavides-Martínez, et al. RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in Escherichia coli K-12. *Nucleic Acids Research*, 32(suppl 1):D303, 2004.

[12] Q.S. Ni, Z.Z. Wang, G.G. Li, G.Y. Wang, and Y.J. Zhao. Effect of the quality of the interaction data on predicting protein function from protein-protein interactions. *Interdisciplinary Sciences: Computational Life Sciences*, 1(1):40–45, 2009.

[13] B. Schwikowski, P. Uetz, and S. Fields. A network of protein–protein interactions in yeast. *Nature Biotechnology*, 18(12):1257–1261, 2000.

[14] H. Hishigaki, K. Nakai, T. Ono, A. Tanigami, and T. Takagi. Assessment of prediction accuracy of protein function from protein-protein interaction data. *Yeast*, 18(6):523–531, 2001.

[15] J. Hakenberg, C. Plake, R. Leaman, M. Schroeder, and G. Gonzalez. Inter-species normalization of gene mentions with GNAT. *Bioinformatics*, 24(16):i126, 2008.

[16] D. Maglott, J. Ostell, K.D. Pruitt, and T. Tatusova. Entrez Gene: gene-centered information at NCBI. *Nucleic acids research*, 35(suppl 1):D26, 2006.

[17] C. Giuliano, A. Lavelli, and L. Romano. Exploiting shallow linguistic information for relation extraction from biomedical literature. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, pages 5–7, 2006.

[18] S. Pyysalo, A. Airola, J. Heimonen, J. Björne, F. Ginter, and T. Salakoski. Comparative analysis of five protein-protein interaction corpora. *BMC bioinformatics*, 9(Suppl 3):S6, 2008.

[19] C.H. Wu, R. Apweiler, A. Bairoch, D.A. Natale, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, et al. The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic acids research*, 34(suppl 1):D187, 2006.

[20] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, GR Gopinath, GR Wu, L. Matthews, et al. Reactome: a knowledgebase of biological pathways. *Nucleic acids research*, 33(suppl 1):D428, 2005.

[21] I. Xenarios, L. Salwinski, X.J. Duan, P. Higney, S.M. Kim, and D. Eisenberg. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic acids research*, 30(1):303, 2002.

[22] TS Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, et al. Human protein reference database—2009 update. *Nucleic acids research*, 37(suppl 1):D767, 2009.

[23] B. Aranda, P. Achuthan, Y. Alam-Faruque, I. Armean, A. Bridge, C. Derow, M. Feuermann, AT Ghanbarian, S. Kerrien, J. Khadake, et al. The IntAct molecular interaction database in 2010. *Nucleic acids research*, 38(suppl 1):D525, 2010.

[24] H.W. Mewes, A. Ruepp, F. Theis, T. Rattei, M. Walter, D. Frishman, K. Suhre, M. Spannagl, K.F.X. Mayer, V. St
"umpflen, et al. MIPS: curated databases and comprehensive secondary data resources in 2010. *Nucleic Acids Research*, 39(suppl 1):D220, 2011.

[25] A. Ceol, A. Chatr Aryamontri, L. Licata, D. Peluso, L. Briganti, L. Perfetto, L. Castagnoli, and G. Cesareni. MINT, the molecular interaction database: 2009 update. *Nucleic acids research*, 38(suppl 1):D532, 2010.

[26] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, et al. Gene Ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.