Characterization of Protein Interactions

Robert Küffner, Timo Duchrow, Katrin Fundel, Ralf Zimmer

Institut für Informatik, Ludwig-Maximilians-Universität München,
Amalienstrasse 17, 80333 München, Germany

Abstract. Available information on molecular interactions between proteins is currently incomplete with regard to detail and comprehensiveness. Although a number of repositories are already devoted to capture interaction data, only a small subset of the currently known interactions can be obtained that way. Besides further experiments, knowledge on interactions can only be complemented by applying text extraction methods to the literature. Currently, information to further characterize individual interactions can not be provided by interaction extraction approaches and is virtually nonexistent in repositories.
We present an approach to not only confirm extracted interactions but also to characterize interactions with regard to four attributes such as activation vs. inhibition and protein-protein vs. protein-gene interactions. Here, training corpora with positional annotation of interacting proteins are required. As suitable corpora are rare, we propose an extensible curation protocol to conveniently characterize interactions by manual annotation of sentences so that machine learning approaches can be applied subsequently. We derived a training set by manually reading and annotating 269 sentences for 1090 candidate interactions; 439 of these are valid interactions, predicted via support vector machines at a precision of 83% and a recall of 87%. The prediction of interaction attributes from individual sentences on average yielded a precision of about 85% and a recall of 73%.

1 Introduction

The discovery or extension of molecular pathways and disease models requires the detailed knowledge on molecular interactions and their properties. Databases already capture many thousands of interactions between molecules [1,2,3], sometimes organized as pathways [3,4,5]. Most interactions were derived from large scale experiments, lacking additional details, e.g. to distinguish activation from inhibition. On the other hand, the bulk of the knowledge on interactions resides in the literature and can be accessed systematically only by automated extraction techniques. A number of such approaches have been published (a brief review can be found in [6]) but they usually do not predict any additional details on interactions. As common in the field, interactions are extracted from sentences that in turn are derived from publication abstracts as provided by Medline. We subdivide the extraction of interactions from sentences into the following steps for which we provide novel solutions:

1. We present a novel curation protocol (section 2.2) for a positional annotation of the interacting proteins. Manual annotation and systematic curation protocols are necessary as suitable training corpora are rare, e.g. as provided by the LLL challenge [7] dataset on procaryotic gene interactions.
2. Following this protocol, 269 sentences including 1090 possible interactions have been carefully read and annotated to derive a training data set (section 3.1). The large number of possible interactions is due to the fact that sentences tend to be long and frequently contain more than two proteins, and therefore $\binom{n}{2}$ co-occurrences for n proteins, and it might be difficult to decide which of the respective pairs of proteins actually interact and in which way.
3. To distinguish interactions from co-occurrences, we first identify the relevant part of sentences via RelEx [6]. Subsequently, each co-occurrence is evaluated in turn to predict interactions using support vector machines (Section 2.4).
4. Sentences frequently provide additional information to characterize individual interactions. Here we aimed to derive four attributes from the texts (Table 1): (a) directed vs. nondirected, (b) activation vs. inhibition, (c) immediate vs. long range and (d) protein-protein vs. protein-gene.

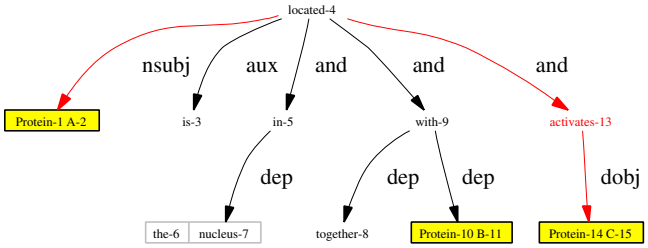
2 Methods

2.1 Preparation of data

In order to create a training set we compiled a list of PubMed abstracts likely to contain protein interactions. To this end, some preprocessing and preselection is required. In the context of this paper we were interested in human protein interactions, so we first screened Medline abstracts for human proteins via ProMiner [8]. The protocol to derive mappings between Medline abstracts and human proteins has been described in detail in [9]. Then, so called *interaction paths* have been derived via RelEx [6] based on dependency parse trees contructed by the Stanford Lexicalized Parser [10]. RelEx extracts chains of dependencies (*paths*) that connect two proteins to create candidate interactions (Figure 1). Thus, paths should contain the semantic dependencies and the corresponding subset of words from the original sentence necessary and sufficient to describe the relationship for each pair of protein entitites. This allows to subdivide the extraction of interactions from texts into (a) extracting the corresponding path and (b) to distinguish protein interactions from mere co-occurrences based on features from the path. The second subproblem will be described below (section 2.4). The following protocol has been applied to select sentences with an increased probability to contain descriptions of interactions:

1. We screened molecular biology databases [1,2,3] for PubMed references.
2. Sentences are selected according to one of the following two criteria. Variant 1 selects sentences matching HPRD interactions denoted as (PubMedId, protein 1, protein 2). Here, sentences from abstracts defined by PubMedId were selected that include both proteins. As we encountered some difficulties with this criterion (compare section 3.1) most sentences have been selected by variant 2 that simply requires sentences to contain at least two proteins.

3. Valid sentences were further required to contain at least one RelEx path and an interaction keyword (such as 'activate', 'formylation' etc). For this purpose we compiled a list of some 300 keywords, which we consider almost exhaustive. We randomly selected 4500 Medline abstracts that satisfy the above criteria as the source for an inital training data set.
4. From each of the selected abstracts only one sentence was selected at random. This constraint intends to avoid bias from abstracts referring to a particular interaction several times or containing many proteins.



Protein A is located in the nucleus together with protein B and activates protein C.

Fig. 1. Dependency parse tree of an example sentence as contructed by the Stanford Lexicalized Parser [10]. Arrows represent dependencies between terms. Proteins (yellow boxes) and noun phrase chunks containing several words are combined into larger nodes. The sentence contains one interaction keyword (*activates*) and one corresponding dependency path extracted by RelEx [6] that is marked in red. The path correctly maps *activates* to {A, C}, but not B.

2.2 Manual annotation

A simple textual annotation form is generated for each sentence selected in section 2.1. Proteins have already been detected via ProMiner [8] during sentence selection. Pairs of detected proteins yield candidate interactions that are manually annotated by five different attributes (Table 1). We use five labels that denote different levels of confidence to describe each attribute thereby providing some flexibility for the annotation of difficult cases. Figure 2 shows the annotation of a sample sentence. In addition to the five confidence labels the curator can indicate additional hints .

In the following we will introduce the concept of ***hints*** that are used to safeguard the selection of meaningful training contexts. During the development of our annotation protocol we had to ensure that results from curation are suitable for a subsequent classification/prediction setting. We need to keep in mind that

```
7838715.2.5 Chemical-1 sequencing-2 and mass-4 spectral-5 analysis-6 of
tryptic-8 peptides-9 derived-10 from the-12 purified-13 polypeptides-14
identifies-15 the-16 ARF6-17 complex-18 as a-20 heterodimer-21 of the-23
retinoid-24-X-25 receptor-26-alpha-27      (RXR-29-alpha-30)-31 and the-33
murine-34 peroxisome-35-proliferator-36-activated-37-receptor-38-gamma-39
(PPAR-41-gamma-42)-43.
27 39 interacting=5 18 21
27 39 directed   =1
27 39 activating =3
27 39 immediate  =5
27 39 expression =1
```

Fig. 2. Annotation of entry 7838715.2.5 (PubMedId, <1=title, 2=abstract>, SentenceNo), an undirected, immediate protein-protein interaction. Two proteins have been detected by ProMiner [8], thus a single annotation slot below the sentence has been generated. Here, names ending at token positions 27 and 30 as well as 39 and 42, respectively, are consecutive synonyms referring to the same entity and thus do not yield additional candidate interactions. Each interaction slot is defined by the token positions of the two proteins as denoted by the first two columns of integers. The third column specifies the attribute that is to be labelled. The attribute value is manually entered into the fourth column, here already filled in. Further columns are reserved for hints (token *complex-18* or token *heterodimer-21*), required to be present on paths for training classifiers.

potentially not all the words from a sentence might be avaliable to a classifier, e.g. features might be generated from RelEx paths only. At the same time, we had to ensure that the curation process is independent from feature generation/classification as the exact specifications of RelEx or other underlying tools might be subject to change. Frequently, the decision if a particular label should be attributed (e.g. *expression*) depends on the presence of an essential term (e.g. *gene*) as in the sentence *The gene coding for A is regulated by B.* By denoting the keyword *gene* as a hint for the decision protein-protein vs. protein-gene interaction this sample would be valid only if the keyword *gene* is part of the respective set of features, or path, as the assertion of the attribute *expression* would not be possible based on the second part of the sentence (*A is regulated by B*) alone. In the classification setting, instances are removed from the training and classification pools if they lack features annotated as hints.

2.3 Generation of features

Features are generated for all sentences chosen by our selection protocol (section 2.1). Our approach is to define generic *feature sources* that are applied to each candidate interaction (i.e. (PubMedId, position protein 1, position protein 2)). Each feature source generates features that are added to a global feature list for this candidate. This makes it possible to combine several feature sources

with each other to define a feature space. Protein names are excluded to avoid overfitting. Features are derived from words stemmed by the Porter [11] stemmer.

Bag-of-words (BOW) creates features from all words in a sentence.

Bag-of-words-path (BOW-path) only creates features for a subset of the words in a sentence, i.e. for a path determined by RelEx. Given a sentence and a pair of proteins (candidate interaction), a subset of paths from the set of all paths for the given sentence are selected that contain the proteins. This feature source also uses hints entered into the curation forms. If no hints are given all applicable paths are selected. If hints are defined specifically for an attribute only those paths are admitted that contain at least one of the hints.

2.4 Classification procedure

Besides predicting protein interactions from co-occurrences we also predict the type of interaction with respect to four attributes: (a) directed vs. nondirected, (b) activating vs. inhibiting, (c) immediate vs. long range and (d) protein-protein vs. protein-gene. Training and predictions for the latter 4 attributes are performed even if a candidate interaction is annotated or predicted as invalid.

For learning, a reduced set of labels is constructed by combining 1+2 as well as 4+5. The prediction of interactions is a two class problem and has been realized by training a single SVM classifier. The other four attributes each constitute three class problems, e.g. *activating* (1+2), vs. *inhibiting* (4+5) vs. *not specified* (3). A three class problem can be reduced to a set of two class problems using the one-versus-rest (OVR) strategy. Two binary SVM classifiers are constructed for each class vs. the other classes, i.e. 1+2 vs. 3+4+5 and 4+5 vs. 1+2+3. No classifiers were constructed for *not specified* vs. *rest*, though, so that two classifiers are required for each of the three class problems. Thus, a total of nine classifiers are required for the five attributes. To combine the outputs of the two classifiers for a specific attribute we use the following rule: *not specified* is predicted if a new sample is located on the side of the negative training samples with regard to the decision hyperplane for both classifiers. Otherwise, the class is selected that corresponds to the maximum value of the SVM decision functions of the two respective classifiers.

All training and classification using support vector machines has been performed using svmlight [12]. We used the default parameters (linear kernels), except that the cost-ratio for training errors on positive samples has been set to the ratio of the corresponding class sizes, i.e. #negative examples / #positive examples.

3 Results

3.1 Construction of a test set

In order to increase the probability that selected sentences indeed describe interactions, we first used variant 1 of our sentence selection protocol (Section

Attribute	label 1	label 2	label 3	label 4	label 5
interacting	no=**661**	**0**	**0**	**37**	yes=**392**
directed	undirected=**186**	**4**	**3**	**6**	directed=**240**
activating	inhibiting=**36**	**0**	**280**	**10**	activating=**113**
immediate	indirect=**101**	**13**	**33**	**64**	direct=**228**
expression	protein-protein=**258**	**32**	**44**	**9**	protein-gene=**96**

Table 1. Attribute labels and their distribution in the training data. Label 3 indicates that an attribute is not specified in the given sentence. Intermediate labels 2 and 4 indicate that the annotation has been attributed with only moderate confidence by the curator. The 661 samples labeled as not interacting are assigned label 3 for the other attributes (not counted here).

2.2) to select 50 abstracts and one sentence from each abstract. Thereby, sentences are selected that were likely sources for interactions derived by HPRD [3]. We manually labelled these sentences and analyzed the results with regard to the five interaction attributes. This analysis showed that about 90% of the selected sentences described interactions. Unfortunately, the analysis also showed that the distribution of attribute labels was significantly imbalanced towards protein-protein interactions based almost exclusively (>90%) on the keywords *binds*, *interacts* and *complex*. Most sentences did not provide any information on activation/inhibition, expression or directed interactions. This indicates that the curation protocol employed by HPRD is selective with regard to immediate protein-protein interactions and we could not expect to derive a balanced distribution of attribute labels this way.

Further curation thus focused on the second variant of our sentence selection protocol. In total, 269 sentences have been annotated yielding attribute labels for 1090 instances of candidate interactions. The overall distribution of labels is shown in Table 1. The *interacting* and *directed* attributes were most straightforward to annotate. Only few instances were labelled with moderate confidence (labels 2 and 4) whereas label 3 (not specified) was virtually absent. Table 1 also shows that certain attributes are less frequent in free text interactions especially striking for *inhibition*, but still noticeable in the case of *long-range* and *protein-gene* interaction.

3.2 Evaluation of classifiers

Evaluation of performance for different classifiers was carried out on a set of 1090 annotated training instances defined by a sentence identifier and both interaction partners. For training and prediction, both strong (labels 1 and 5) and moderate (labels 2 and 4) confidence annotations were included. A stratified 10-fold cross validation has been repeated 10 times (i.e. 10*10) for different random splits. The performance estimates (Table 2 and Figure 3) show that attributes

with a larger number of examples yield a better performance. On average, precison is higher than recall, so predicted interactions and interaction attributes are reliable while some annotations could not be recovered. We also compared the performance with regard to the different options for feature generation (Table 3). The performance increased significantly when specific features were generated for dependency paths. Table 3 also compares the influence of hints on the performance. Here, hints showed a significant increase in performance (+5.5% in f-measure) only if features from the RelEx paths were included. The influence of hints was hardly noticable if only the simple bag-of-words feature source has been used.

Classifier	Accuracy	Precision	Recall	F-measure
interacting	94.1	82.7	87.2	84.9
not directed	97.6	90.7	81.8	86.0
directed	95.1	84.6	67.2	74.9
inhibiting	99.1	75.4	63.6	69.0
activating	97.7	85.1	q.0	73.0
long range	97.4	79.0	49.2	60.6
immediate	94.8	83.3	78.6	80.9
protein-protein	95.5	86.9	81.2	83.9
protein-gene	97.8	86.1	65.3	74.3
overall	96.9	85.4	73.3	78.9

Table 2. Cross-validation performance on a data set of 1090 candidate interactions. Mean measures have been calculated via microaveraging. The *overall*-performance was calculated as the mean of all classifiers except *interacting.*

Protocol	Precision	Recall	F-measure
bag-of-words (BOW)	35.5	68.3	46.7
BOW + hints	36.1	69.0	47.4
BOW + path	78.2	82.3	79.4
BOW + path + hints	82.7	87.2	84.9

Table 3. The prediction performance of the classifier co-occurrence vs. interaction has been compared with regard to different feature sources and the utilization of hints.

In the following (see also Figure 4), a few examples will be mentioned where classification has been mislead by lexical subtleties or incorrect parse trees. In the

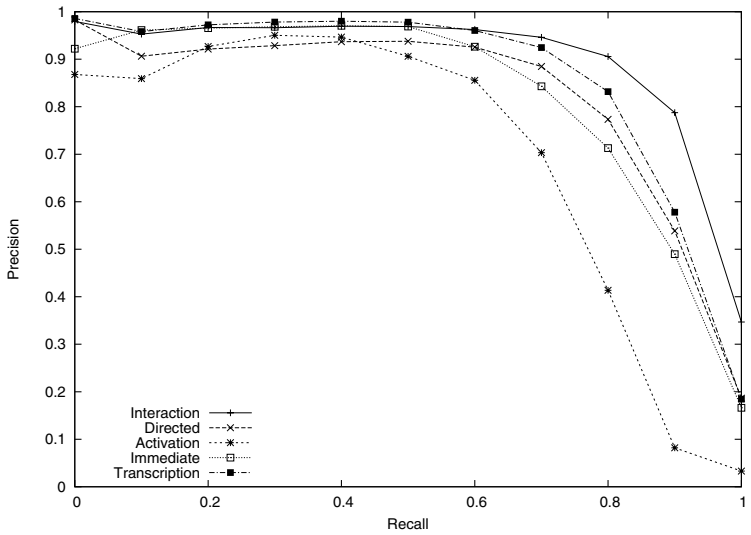
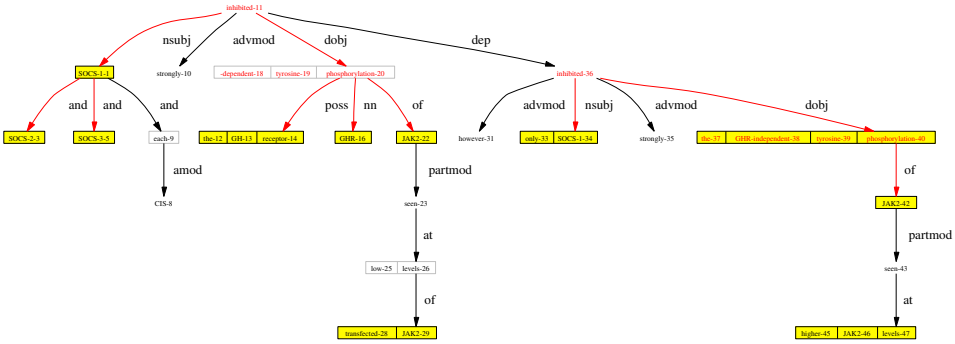


Fig. 3. Precision of attributes at 11 recall values. Performance estimates have been combined for the four attributes that require 2 classifiers, i.e. except for the simple attribute *interaction.*

sentence *A induces methylation of B* the word *induce* does not refer to induction of a gene as it would be the case if the term *methylation* would have been absent. This is different from the sentence *A induces methylating B.* As currently only word stems are considered for features and frequent stop words (such as 'of') are discarded both sentences yield the same set of features. Another difficult case is represented by *A inhibits signalling downstream of B* where a direct relationship (in the causal sense) between A and B is not necessarily implied. Some problems arise from incorrect dependency trees, e.g. in *A activates B but not C* the negation refers to B according to the parser [10]. Future improvements will also need to focus on multiple negations and to consider specific negations such as *A-null mice* or *A(-/-) mice.*

4 Discussion

The construction of advanced causal network models requires specific annotation (called attributes throughout this paper) on protein interactions such as activating vs. inhibiting or protein-protein vs. protein-gene. Such details on interactions are not available from current databases or text extraction approaches in a systematic and comprehensive way. We propose to alleviate this problem with a two step strategy for the extraction and characterization of molecular interactions from free texts. Starting from sentences we narrow down to the context or *path* comprising the actual assertions on a given candidate interaction. We presented



SOCS-1, SOCS-2, SOCS-3, and CIS each strongly inhibited the GH receptor (GHR)-dependent tyrosine phosphorylation of JAK2 seen at low levels of transfected JAK2; however, only SOCS-1 strongly inhibited the GHR-independent tyrosine phosphorylation of JAK2 seen at higher JAK2 levels.

Fig. 4. Dependency graph of a misclassified interaction. Here the interaction between each of the SOCS and GHR-16 are incorrectly classified as inhibiting. However, the text describes no direct inhibiting interaction between SOCS and GHR, but SOCS inhibits the GHR dependent phosphorylation of JAK2-22.

two major contributions: (1) a systematic and convenient curation protocol for the positional curation of candidate protein interactions including the manual annotation of a training set and (2) a protocol for training and evaluation of classifiers for the accurate prediction of interactions and four interaction attributes (Table 1).

Candidate interactions are annotated according to three levels of confidence: not specified, moderate and high confidence (Table 1). The introduction of the moderate confidence level helped to speed up the curation process as it was especially applicable to difficult examples. Without this level of confidence, several examples would have been annotated as *not specified,* so it also helped to improve recall during curation. We also introduced *hints*, i.e. labelling of special words essential for capturing a particular meaning of a given interaction. Hints are used to ensure that interaction paths can be excluded from classifier training if essential terms have been lost during preprocessing. We showed that the annotation of hints did not introduce a significant bias into classification (Table 3). As an additional advantage, hints capture information on why curators made particular decisions. In our experience the proposed curation protocol was simple to learn and use and categorized curator decisions appropriately.

We then constructed classifiers for the five attributes. These demonstrated good cross validation performance for predicting interactions (as opposed to mere co-occurrence of proteins) as well as other attributes. On average, precision was higher than recall, indicating that the manual annotation could not always be recovered automatically from the given sentences. At the same time we noticed that attribute performance was positively correlated with the abun-

dance of available annotation. This indicates that an enlargement of our current dataset, possibly selective with regard to the underpopulated attributes, will be beneficial. Our method itself is generic, so that an extension to accommodate additional attributes would be simple although additional manual annotation would be required to provide the necessary training data.

References

1. G. D. Bader, D. Betel, and C. W. Hogue, "Bind: the biomolecular interaction network database," *Nucleic Acids Res*, vol. 31, no. 1, pp. 248–50, 2003.
2. 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 Res*, vol. 30, no. 1, pp. 303–5, 2002.
3. S. Peri, J. D. Navarro, T. Z. Kristiansen, R. Amanchy, V. Surendranath, B. Muthusamy, T. K. Gandhi, K. N. Chandrika, N. Deshpande, S. Suresh, B. P. Rashmi, K. Shanker, N. Padma, V. Niranjan, H. C. Harsha, N. Tal reja, B. M. Vrushabendra, M. A. Ramya, A. J. Yatish, M. Joy, H. N. Shivashankar, M. P. Kavitha, D. R. Menezes, M. a nd Choudhury, N. Ghosh, R. Saravana, S. Chandran, S. Mohan, C. K. Jonnalagadda, C. K. Prasad, C. Kumar-Sinha, K. S. Deshpande, and A. Pandey, "Human protein reference database as a discovery resource for proteomics," *Nucleic Acids Res*, vol. 32, no. Database issue, pp. D497–501, 2004.
4. M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. . Araki, and M. Hirakawa, "From genomics to chemical genomics: new developments in kegg," *Nucleic Acids Res*, vol. 34, no. Database issue, pp. D354–7, 2006.
5. M. Krull, S. Pistor, N. Voss, A. Kel, I. Reuter, D. Kronenberg, H. Michael, K. Schwarzer, A. Potapo v, C. Choi, O. Kel-Margoulis, and E. Wingender, "Transpath: an information resource for storing and visualizing signaling pathways and their pathological aberrations," *Nucleic Acids Res*, vol. 34, no. Database issue, pp. D546–51, 2006.
6. K. Fundel, R. Küffner, and R. Zimmer, "Relex - a new approach for relation extraction using dependency parse trees," *manuscript in preparation*, 2006.
7. C. Nedellec, "Learning language in logic - genic interaction extraction challenge," *Proceedings of the ICML05 workshop: Learning Language in Logic (LLL05)*, 2005.
8. D. Hanisch, K. Fundel, H. T. Mevissen, R. Zimmer, and J. Fluck, "Prominer: rule-based protein and gene entity recognition," *BMC Bioinformatics*, vol. 6 Suppl 1, p. S14, 2005.
9. R. Küffner, K. Fundel, and R. Zimmer, "Expert knowledge without the expert: integrated analysis of gene expression and literature to derive active functional contexts," *Bioinformatics*, vol. 21 Suppl 2, pp. ii259–ii267, 2005.
10. D. Klein and C. D. Manning, "Fast exact inference with a factored model for natural language parsing," *Advances in Neural Information Processing Systems 15 (NIPS 2002)*, 2002.
11. M. Porter, "An algorithm for suffix stripping," *Program*, vol. 14 (3), pp. 130–137, 2003.
12. T. Joachims, *Learning to Classify Text Using Support Vector Machines*. Dissertation, Kluwer.