In *Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics*
pp. 46--53, Detroit, MI, June 2005.

Using Biomedical Literature Mining to Consolidate the Set of Known Human Protein–Protein Interactions

Arun Ramani, Edward Marcotte
Institute for Cellular and Molecular Biology
University of Texas at Austin
1 University Station A4800
Austin, TX 78712
arun@icmb.utexas.edu
marcotte@icmb.utexas.edu

Razvan Bunescu, Raymond Mooney
Department of Computer Sciences
University of Texas at Austin
1 University Station C0500
Austin, TX 78712
razvan@cs.utexas.edu
mooney@cs.utexas.edu

Abstract

This paper presents the results of a large-scale effort to construct a comprehensive database of known human protein interactions by combining and linking known interactions from existing databases and then adding to them by automatically mining additional interactions from 750,000 Medline abstracts. The end result is a network of 31,609 interactions amongst 7,748 proteins. The text mining system first identifies protein names in the text using a trained Conditional Random Field (CRF) and then identifies interactions through a filtered co-citation analysis. We also report two new strategies for mining interactions, either by finding explicit statements of interactions in the text using learned pattern-based rules or a Support-Vector Machine using a string kernel. Using information in existing ontologies, the automatically extracted data is shown to be of equivalent accuracy to manually curated data sets.

1 Introduction

Proteins are often considered in terms of their networks of interactions, a view that has spurred considerable effort in mapping large-scale protein interaction networks. Thus far, the most complete protein networks are measured for yeast and derive from the synthesis of varied large scale experimental interaction data and in-silico interaction predictions (summarized in (von Mering et al., 2002; Lee et al., 2004; Jansen et al., 2003)). Unlike the case of yeast, only minimal progress has been made with respect to the human proteome. While some moderate-scale interaction maps have been created, such as for the purified TNFa/NFKB protein complex (Bouwmeester et al., 2004) and the proteins involved in the human Smad signaling pathway (Colland et al., 2004), the bulk of known human protein interaction data derives from individual, small-scale experiments reported in Medline. Many of these interactions have been collected in the Reactome (Joshi-Tope et al., 2005), BIND (Bader et al., 2003), DIP (Xenarios et al., 2002), and HPRD (Peri et al., 2004) databases, with Reactome contributing 11,000 interactions that have been manually entered from articles focusing on interactions in core cellular pathways, and HPRD contributing a set of 12,000 interactions recovered by manual curation of Medline articles using teams of readers. Additional interactions have been transferred from other organisms based on orthology (Lehner and Fraser, 2004).

A comparison of these existing interaction data sets is enlightening. Although the interactions from these data sets are in principle derived from the same source (Medline), the sets are quite disjoint (Figure 1) implying either that the sets are biased for different classes of interactions, or that the actual number of interactions in Medline is quite large. We suspect both reasons. It is clear that each data set has a different explicit focus (Reactome towards core cellular machinery, HPRD towards disease-linked genes, and DIP and BIND more randomly distributed). Due to these biases, it is likely that many interactions from Medline are still excluded from these data sets. The maximal overlap between interaction data sets is seen for BIND: 25% of these interactions are also in HPRD or Reactome; only 1% of Reactome interactions are in HPRD or BIND.

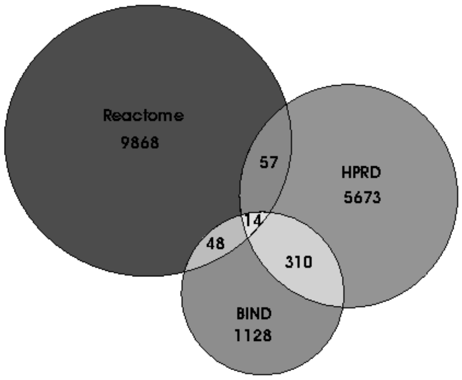

Figure 1: Overlap diagram for known datasets.

Medline now has records from more than 4,800 journals accounting for around 15 million articles. These citations contain thousands of experimentally recorded protein interactions, and even a cursory investigation of Medline reveals human protein interactions not present in the current databases. However, retrieving these data manually is made difficult by the large number of articles, all lacking formal structure. Automated extraction of information would be preferable, and therefore, mining data from Medline abstracts is a growing field (Jenssen et al., 2001; Rzhetsky et al., 2004; Liu and Wong, 2003; Hirschman et al., 2002).

In this paper, we describe a framework for automatic extraction of protein interactions from biomedical literature. We focus in particular on the difficult and important problem of identifying interactions concerning human proteins. We describe a system for first accurately identifying the names of human proteins in the documents, then on identifying pairs of interacting human proteins, and demonstrate that the extracted protein interactions are comparable to those extracted manually. In the process, we consolidate the existing set of publically-available human protein interactions into a network of 31,609 interactions between 7,748 proteins.

2 Assembling existing protein interaction data

We previously gathered the existing human protein interaction data sets ((Ramani et al., 2005); summarized in Table 1), representing the current status of the publically-available human interactome. This required unification of the interactions under a shared naming and annotation convention. For this purpose, we mapped each interacting protein to LocusLink (now EntrezGene) identification numbers and retained only unique interactions (i.e., for two proteins A and B, we retain only A–B or B–A, not both). We have chosen to omit self-interactions, A–A or B–B, for technical reasons, as their quality cannot be assessed on the functional benchmark that we describe in Section 3. In most cases, a small loss of proteins occurred in the conversion between the different gene identifiers (e.g., converting from the NCBI 'gi' codes in BIND to LocusLink identifiers). In the case of Human Protein Reference Database (HPRD), this processing resulted in a significant reduction in the number of interactions from 12,013 total interactions to 6,054 unique, non-self interactions, largely due to the fact that HPRD often records both A-B and B-A interactions, as well as a large number of self interactions, and indexes genes by their common names rather than conventional database entries, often resulting in multiple entries for different synonyms. An additional 9,283 (or 60,000 at lower confidence) interactions are available from orthologous transfer of interactions from large-scale screens in other organisms (orthology-core and orthology-all) (Lehner and Fraser, 2004).

3 Two benchmark tests of accuracy for interaction data

To measure the relative accuracy of each protein interaction data set, we established two benchmarks of interaction accuracy, one based on shared protein function and the other based on previously known interactions. First, we constructed a benchmark in which we tested the extent to which interaction partners in a data set shared annotation, a measure previously shown to correlate with the accuracy of functional genomics data sets (von Mering et al., 2002; Lee et al., 2004; Lehner and Fraser, 2004). We used the functional annotations listed in the KEGG

Dataset	Version	Total Is (Ps)	Self (A-A) Is (Ps)	Unique (A-B) Is (Ps)
Reactome	08/03/04	12,497 (6,257)	160 (160)	12,336 (807)
BIND	08/03/04	6,212 (5,412)	549 (549)	5,663 (4,762)
HPRD*	04/12/04	12,013 (4,122)	3,028 (3,028)	6,054 (2,747)
Orthology (all)	03/31/04	71,497 (6,257)	373 (373)	71,124 (6,228)
Orthology (core)	03/31/04	11,488 (3,918)	206 (206)	11,282 (3,863)

Table 1: **Is** = **I**nteraction**s**, **Ps** = **P**rotein**s**.

(Kanehisa et al., 2004) and Gene Ontology (Ashburner et al., 2000) annotation databases. These databases provide specific pathway and biological process annotations for approximately 7,500 human genes, assigning human genes into 155 KEGG pathways (at the lowest level of KEGG) and 1,356 GO pathways (at level 8 of the GO biological process annotation). KEGG and GO annotations were combined into a single composite functional annotation set, which was then split into independent testing and training sets by randomly assigning annotated genes into the two categories (3,800 and 3,815 annotated genes respectively). For the second benchmark based on known physical interactions, we assembled the human protein interactions from Reactome and BIND, a set of 11,425 interactions between 1,710 proteins. Each benchmark therefore consists of a set of binary relations between proteins, either based on proteins sharing annotation or physically interacting. Generally speaking, we expect more accurate protein interaction data sets to be more enriched in these protein pairs. More specifically, we expect true physical interactions to score highly on both tests, while non-physical or indirect associations, such as genetic associations, should score highly on the functional, but not physical interaction, test.

For both benchmarks, the scoring scheme for measuring interaction set accuracy is in the form of a log odds ratio of gene pairs either sharing annotations or physically interacting. To evaluate a data set, we calculate a log likelihood ratio (LLR) as:

$$LLR = ln\frac{P(D|I)}{P(D|\neg I)} = ln\frac{P(I|D)P(\neg I)}{P(\neg I|D)P(I)} \quad (1)$$

where $P(D|I)$ and $P(D|\neg I)$ are the probability of observing the data D conditioned on the genes sharing benchmark associations (I) and not sharing benchmark associations ($\neg I$). In its expanded form (obtained by applying Bayes theorem), $P(I|D)$ and $P(\neg I|D)$ are estimated using the frequencies of interactions observed in the given data set D between annotated genes sharing benchmark associations and not sharing associations, respectively, while the priors $P(I)$ and $P(\neg I)$ are estimated based on the total frequencies of all benchmark genes sharing the same associations and not sharing associations, respectively. A score of zero indicates interaction partners in the data set being tested are no more likely than random to belong to the same pathway or to interact; higher scores indicate a more accurate data set.

Among the literature-derived interactions (Reactome, BIND, HPRD), a total of 17,098 unique interactions occur in the public data sets. Testing the existing protein interaction data on the functional benchmark reveals that Reactome has the highest accuracy (LLR = 3.8), followed by BIND (LLR = 2.9), HPRD (LLR = 2.1), core orthology-inferred interactions (LLR = 2.1) and the non-core orthology-inferred interaction (LLR = 1.1). The two most accurate data sets, Reactome and BIND, form the basis of the protein interaction–based benchmark. Testing the remaining data sets on this benchmark (i.e., for their consistency with these accurate protein interaction data sets) reveals a similar ranking in the remaining data. Core orthology-inferred interactions are the most accurate (LLR = 5.0), followed by HPRD (LLR = 3.7) and non-core orthology inferred interactions (LLR = 3.7).

4 Framework for Mining Protein–Protein Interactions

The extraction of interacting proteins from Medline abstracts proceeds in two separate steps:

1. First, we automatically identify protein names

using a CRF system trained on a set of 750 abstracts manually annotated for proteins (see Section 5 for details).

2. Based on the output of the CRF tagger, we filter out less confident extractions and then try to detect which pairs of the remaining extracted protein names are interaction pairs.

For the second step, we investigate two general methods:

- Use co-citation analysis to score each pair of proteins based on the assumption that proteins co-occurring in a large number of abstracts tend to be interacting proteins. Out of the resulting protein pairs we keep only those that co-occur in abstracts likely to discuss interactions, based on a Naive Bayes classifier (see Section 6 for details).

- Given that we already have a set of 230 Medline abstracts manually tagged for both proteins and interactions, we can use it to train an interaction extractor. In Section 7 we discuss two different methods for learning this interaction extractor.

5 A CRF Tagger for Protein Names

The task of identifying protein names is made difficult by the fact that unlike other organisms, such as yeast or E. coli, the human genes have no standardized naming convention, and thus present one of the hardest sets of gene/protein names to extract. For example, human proteins may be named with typical English words, such as ”light”, ”map”, ”complement”, and ”Sonic Hedgehog”. It is therefore necessary that an information extraction algorithm be specifically trained to extract gene and protein names accurately.

We have previously described (Bunescu et al., 2005) effective protein and gene name tagging using a Maximum Entropy based algorithm. Conditional Random Fields (CRF) (Lafferty et al., 2001) are new types of probabilistic models that preserve all the advantages of Maximum Entropy models and at the same time avoid the label bias problem by allowing a sequence of tagging decisions to compete against each other in a global probabilistic model.

In both training and testing the CRF protein-name tagger, the corresponding Medline abstracts were processed as follows. Text was tokenized using white-space as delimiters and treating all punctuation marks as separate tokens. The text was segmented into sentences, and part-of-speech tags were assigned to each token using Brill's tagger (Brill, 1995). For each token in each sentence, a vector of binary features was generated using the feature templates employed by the Maximum Entropy approach described in (Bunescu et al., 2005). Generally, these features make use of the words occurring before and after the current position in the text, their POS tags and capitalization patterns. Each feature occurring in the training data is associated with a parameter in the CRF model. We used the CRF implementation from (McCallum, 2002). To train the CRF's parameters, we used 750 Medline abstracts manually annotated for protein names (Bunescu et al., 2005). We then used the trained system to tag protein and gene names in the entire set of 753,459 Medline abstracts citing the word “human”.

In Figure 2 we compare the performance of the CRF tagger with that of the Maximum Entropy tagger from (Bunescu et al., 2005), using the same set of features, by doing 10-fold cross-validation on Yapex – a smaller dataset of 200 manually annotated abstracts (Franzen et al., 2002). Each model assigns to each extracted protein name a normalized confidence value. The precision–recall curves from Figure 2 are obtained by varying a threshold on the minimum accepted confidence. We also plot the precision and recall obtained by simply matching textual phrases against entries from a protein dictionary.

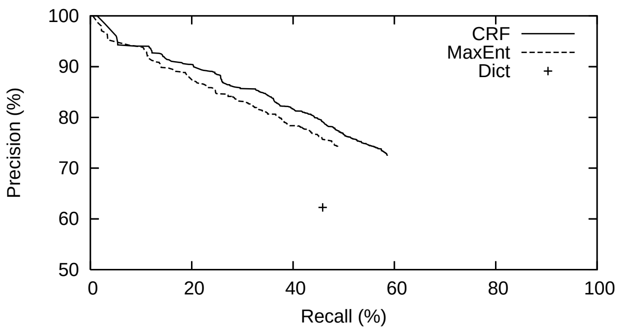

Figure 2: Protein Tagging Performance.

The dictionary of human protein names was assembled from the LocusLink and Swissprot databases by manually curating the gene names and synonyms (87,723 synonyms between 18,879 unique gene names) to remove genes that were referred to as 'hypothetical' or 'probable' and also to omit entries that referred to more than one protein identifier.

6 Co-citation Analysis and Bayesian Classification

In order to establish which interactions occurred between the proteins identified in the Medline abstracts, we used a 2-step strategy: measure co-citation of protein names, then enrich these pairs for physical interactions using a Bayesian filter. First, we counted the number of abstracts citing a pair of proteins, and then calculated the probability of co-citation under a random model based on the hypergeometric distribution (Lee et al., 2004; Jenssen et al., 2001) as:

$$P(k|N,m,n) = \frac{\binom{n}{k}\binom{N-n}{m-k}}{\binom{N}{m}} \qquad (2)$$

where N equals the total number of abstracts, n of which cite the first protein, m cite the second protein, and k cite both.

Empirically, we find the co-citation probability has a hyperbolic relationship with the accuracy on the functional annotation benchmark from Section 3, with protein pairs co–cited with low random probability scoring high on the benchmark.

With a threshold on the estimated extraction confidence of 80% (as computed by the CRF model) in the protein name identification, close to 15,000 interactions are extracted with the co-citation approach that score comparable or better on the functional benchmark than the manually extracted interactions from HPRD, which serves to establish a minimal threshold for our mined interactions.

However, it is clear that proteins are co-cited for many reasons other than physical interactions. We therefore tried to enrich specifically for physical interactions by applying a secondary filter. We applied a Bayesian classifier (Marcotte et al., 2001) to measure the likelihood of the abstracts citing the protein pairs to discuss physical protein–protein interactions. The classifier scores each of the co-citing abstracts according to the usage frequency of discriminating words relevant to physical protein interactions. For a co-cited protein pair, we calculated the average score of co-citing Medline abstracts and used this to re-rank the top-scoring 15,000 co-cited protein pairs.

Interactions extracted by co-citation and filtered using the Bayesian estimator compare favorably with the other interaction data sets on the functional annotation benchmark (Figure 3). Testing the accuracy of these extracted protein pairs on the physical interaction benchmark (Figure 4) reveals that the co-cited proteins scored high by this classifier are indeed strongly enriched for physical interactions.

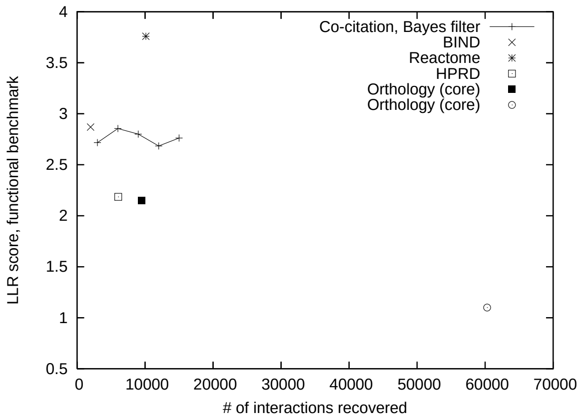

Figure 3: Accuracy, functional benchmark

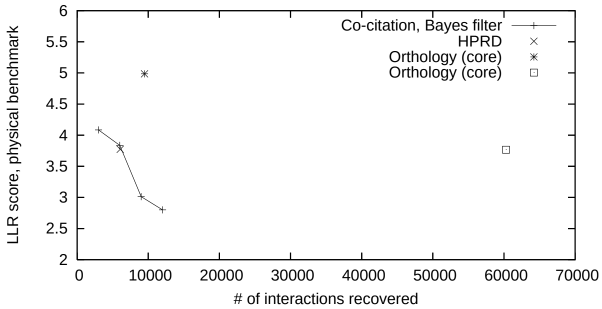

Figure 4: Accuracy, physical benchmark

Keeping all the interactions that score better than HPRD, our co-citation / Bayesian classifier analysis yields 6,580 interactions between 3,737 proteins. By combining these interactions with the 26,280 interactions from the other sources, we obtained a final set of 31,609 interactions between 7,748 human proteins.

7 Learning Interaction Extractors

In (Bunescu et al., 2005) we described a dataset of 230 Medline abstracts manually annotated for proteins and their interactions. This can be used as a training dataset for a method that learns interaction extractors. Such a method simply classifies a sentence containing two protein names as positive or negative, where positive means that the sentence asserts an interaction between the two proteins. However a sentence in the training data may contain more than two proteins and more than one pair of interacting proteins. In order to extract the interacting pairs, we replicate the sentences having n proteins ($n \geq 2$) into C_2^n sentences such that each one has exactly two of the proteins tagged, with the rest of the protein tags omitted. If the tagged proteins interact, then the replicated sentence is added to the set of positive sentences, otherwise it is added to the set of negative sentences. During testing, a sentence having n proteins ($n \geq 2$) is again replicated into C_2^n sentences in a similar way.

7.1 Extraction using Longest Common Subsequences (ELCS)

Blaschke *et al.* (Blaschke and Valencia, 2001; Blaschke and Valencia, 2002) manually developed rules for extracting interacting proteins. Each of their rules (or frames) is a sequence of words (or POS tags) and two protein-name tokens. Between every two adjacent words is a number indicating the maximum number of intervening words allowed when matching the rule to a sentence. In (Bunescu et al., 2005) we described a new method ELCS (Extraction using Longest Common Subsequences) that automatically learns such rules. ELCS' rule representation is similar to that in (Blaschke and Valencia, 2001; Blaschke and Valencia, 2002), except that it currently does not use POS tags, but allows disjunctions of words. Figure 5 shows an example of a rule learned by ELCS. Words in square brackets separated by '|' indicate disjunctive lexical constraints, i.e. one of the given words must match the sentence at that position. The numbers in parentheses between adjacent constraints indicate the maximum number of unconstrained words allowed between the two (called a *word gap*). The protein names are denoted here with PROT. A sentence matches the rule if and only if it satisfies the word constraints in the given order and respects the respective word gaps.

```
- (7) interaction (0) [between | of] (5) PROT (9) PROT (17) .
```

Figure 5: Sample extraction rule learned by ELCS.

7.2 Extraction using a Relation Kernel (ERK)

Both Blaschke and ELCS do interaction extraction based on a limited set of matching rules, where a rule is simply a sparse (gappy) subsequence of words (or POS tags) anchored on the two protein-name tokens. Therefore, the two methods share a common limitation: either through manual selection (Blaschke), or as a result of the greedy learning procedure (ELCS), they end up using only a subset of all possible anchored sparse subsequences. Ideally, we would want to use all such anchored sparse subsequences as features, with weights reflecting their relative accuracy. However explicitly creating for each sentence a vector with a position for each such feature is infeasible, due to the high dimensionality of the feature space. Here we can exploit an idea used before in string kernels (Lodhi et al., 2002): computing the dot-product between two such vectors amounts to calculating the number of common anchored subsequences between the two sentences. This can be done very efficiently by modifying the dynamic programming algorithm from (Lodhi et al., 2002) to account only for anchored subsequences i.e. sparse subsequences which contain the two protein-name tokens. Besides restricting the word subsequences to be anchored on the two protein tokens, we can further prune down the feature space by utilizing the following property of natural language statements: whenever a sentence asserts a relationship between two entity mentions, it generally does this using one of the following three patterns:

- **[FI] F**ore–**I**nter: words before and between the two entity mentions are simultaneously used to express the relationship. Examples: 'interaction of $\langle P_1 \rangle$ with $\langle P_2 \rangle$', 'activation of $\langle P_1 \rangle$ by $\langle P_2 \rangle$'.

- **[I] I**nter: only words between the two entity mentions are essential for asserting the relationship. Examples: '$\langle P_1 \rangle$ interacts with $\langle P_2 \rangle$', '$\langle P_1 \rangle$ is activated by $\langle P_2 \rangle$'.

- **[IA] I**nter–**A**fter: words between and after the two entity mentions are simultaneously used to express the relationship. Examples: $\langle P_1 \rangle$ – $\langle P_2 \rangle$ complex', '$\langle P_1 \rangle$ and $\langle P_2 \rangle$ interact'.

Another useful observation is that all these patterns use at most 4 words to express the relationship (not counting the two entities). Consequently, when computing the relation kernel, we restrict the counting of common anchored subsequences only to those having one of the three types described above, with a maximum word-length of 4. This type of feature selection leads not only to a faster kernel computation, but also to less overfitting, which results in increased accuracy (we omit showing here comparative results supporting this claim, due to lack of space).

We used this kernel in conjunction with Support Vector Machines (Vapnik, 1998) learning in order to find a decision hyperplane that best separates the positive examples from negative examples. We modified the **libsvm** package for SVM learning by plugging in the kernel described above.

7.3 Preliminary experimental results

We compare the following three systems on the task of retrieving protein interactions from the dataset of 230 Medline abstracts (assuming gold standard proteins):

- **[Manual]**: We report the performance of the rule-based system of (Blaschke and Valencia, 2001; Blaschke and Valencia, 2002).

- **[ELCS]**: We report the 10-fold cross-validated results from (Bunescu et al., 2005) as a precision-recall graph.

- **[ERK]**: Based on the same splits as those used by ELCS, we compute the corresponding precision-recall graph.

The results, summarized in Figure 6, show that the relation kernel outperforms both ELCS and the manually written rules. In future work, we intend to analyze the complete Medline with ERK and integrate the extracted interactions into a larger composite set.

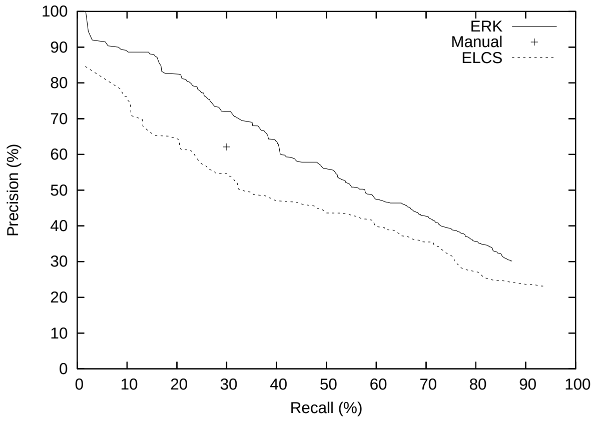

Figure 6: PR curves for interaction extractors.

8 Conclusion

Through a combination of automatic text mining and consolidation of existing databases, we have constructed a large database of known human protein interactions containing 31,609 interactions amongst 7,748 proteins. By mining 753,459 human-related abstracts from Medline with a combination of a CRF-based protein tagger, co-citation analysis, and automatic text classification, we extracted a set of 6,580 interactions between 3,737 proteins. By utilizing information in existing knowledge bases, this automatically extracted data was found to have an accuracy comparable to manually developed data sets. More details on our interaction database have been published in the biological literature (Ramani et al., 2005) and it is available on the web at http://bioinformatics.icmb.utexas.edu/idserve. We are currently exploring improvements to this database by more accurately identifying assertions of interactions in the text using an SVM that exploits a relational string kernel.

9 Acknowledgements

This work was supported by grants from the N.S.F. (IIS-0325116, EIA-0219061), N.I.H. (GM06779-01), Welch (F1515), and a Packard Fellowship (E.M.M.).

References

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

G. D. Bader, D. Betel, and C. W. Hogue. 2003. Bind: the biomolecular interaction network database. *Nucleic Acids Research*, 31(1):248–250.

C. Blaschke and A. Valencia. 2001. Can bibliographic pointers for known biological data be found automatically? protein interactions as a case study. *Comparative and Functional Genomics*, 2:196–206.

C. Blaschke and A. Valencia. 2002. The frame-based module of the Suiseki information extraction system. *IEEE Intelligent Systems*, 17:14–20.

T. Bouwmeester, A. Bauch, H. Ruffner, P. O. Angrand, G. Bergamini, K. Croughton, C. Cruciat, D. Eberhard, J. Gagneur, S. Ghidelli, and et al. 2004. A physical and functional map of the human tnf-alpha/nf-kappa b signal transduction pathway. *Nature Cell Biology*, 6(2):97–105.

Eric Brill. 1995. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, 21(4):543–565.

Razvan Bunescu, Ruifang Ge, Rohit J. Kate, Edward M. Marcotte, Raymond J. Mooney, Arun Kumar Ramani, and Yuk Wah Wong. 2005. Comparative experiments on learning information extractors for proteins and their interactions. *Artificial Intelligence in Medicine (special issue on Summarization and Information Extraction from Medical Documents)*, 33(2):139–155.

F. Colland, X. Jacq, V. Trouplin, C. Mougin, C. Groizeleau, A. Hamburger, A. Meil, J. Wojcik, P. Legrain, and J. M. Gauthier. 2004. Functional proteomics mapping of a human signaling pathway. *Genome Research*, 14(7):1324–1332.

K. Franzen, G. Eriksson, F. Olsson, L. Asker, P. Liden, and J. Coster. 2002. Protein names and how to find them. *International Journal of Medical Informatics*, 67(1-3):49–61.

L. Hirschman, J. C. Park, J. Tsujii, L. Wong, and C. H. Wu. 2002. Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, 18(12):1553–1561.

R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. J. Krogan, S. Chung, A. Emili, M. Snyder, J. F. Greenblatt, and M. Gerstein. 2003. A bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302(5644):449–453.

T. K. Jenssen, A. Laegreid, J. Komorowski, and E. Hovig. 2001. A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, 28(1):21–28.

G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. R. Gopinath, G. R. Wu, L. Matthews, and et al. 2005. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research*, 33 Database Issue:D428–432.

M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori. 2004. The kegg resource for deciphering the genome. *Nucleic Acids Research*, 32 Database issue:D277–280.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of 18th International Conference on Machine Learning (ICML-2001)*, pages 282–289, Williamstown, MA.

I. Lee, S. V. Date, A. T. Adai, and E. M. Marcotte. 2004. A probabilistic functional network of yeast genes. *Science*, 306(5701):1555–1558.

B. Lehner and A. G. Fraser. 2004. A first-draft human protein-interaction map. *Genome Biology*, 5(9):R63.

H. Liu and L. Wong. 2003. Data mining tools for biological sequences. *Journal of Bioinformatics and Computational Biology*, 1(1):139–167.

Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. 2002. Text classification using string kernels. *Journal of Machine Learning Research*, 2:419–444.

E. M. Marcotte, I. Xenarios, and D. Eisenberg. 2001. Mining literature for protein-protein interactions. *Bioinformatics*, 17(4):359–363.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

S. Peri, J. D. Navarro, T. Z. Kristiansen, R. Amanchy, V. Surendranath, B. Muthusamy, T. K. Gandhi, K. N. Chandrika, N. Deshpande, S. Suresh, and et al. 2004. Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Research*, 32 Database issue:D497–501.

A. K. Ramani, R. C. Bunescu, R. J. Mooney, and E. M. Marcotte. 2005. Consolidating the set of know human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biology*, 6(5):r40.

A. Rzhetsky, I. Iossifov, T. Koike, M. Krauthammer, P. Kra, M. Morris, H. Yu, P. A. Duboue, W. Weng, W. J. Wilbur, V. Hatzivassiloglou, and C. Friedman. 2004. Geneways: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *Journal of Biomedical Informatics*, 37(1):43–53.

Vladimir N. Vapnik. 1998. *Statistical Learning Theory*. John Wiley & Sons.

C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork. 2002. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887):399–403.

I. Xenarios, L. Salwinski, X. J. Duan, P. Higney, S. M. Kim, and D. Eisenberg. 2002. Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Research*, 30(1):303–305.