Hindawi Publishing Corporation
EURASIP Journal on Bioinformatics and Systems Biology
Volume 2007, Article ID 53096, 9 pages
doi:10.1155/2007/53096

Research Article

Extraction of Protein Interaction Data: A Comparative Analysis of Methods in Use

Hena Jose, Thangavel Vadivukarasi, and Jyothi Devakumar

Jubilant Biosys Ltd., #96, Industrial Suburb, 2nd Stage, Yeshwanthpur, Bangalore 560 022, India

Received 31 March 2007; Accepted 8 October 2007

Recommended by Z. Jane Wang

Several natural language processing tools, both commercial and freely available, are used to extract protein interactions from publications. Methods used by these tools include pattern matching to dynamic programming with individual recall and precision rates. A methodical survey of these tools, keeping in mind the minimum interaction information a researcher would need, in comparison to manual analysis has not been carried out. We compared data generated using some of the selected NLP tools with manually curated protein interaction data (PathArt and IMaps) to comparatively determine the recall and precision rate. The rates were found to be lower than the published scores when a normalized definition for interaction is considered. Each data point captured wrongly or not picked up by the tool was analyzed. Our evaluation brings forth critical failures of NLP tools and provides pointers for the development of an ideal NLP tool.

1. INTRODUCTION

Protein interactions represent the social networking that happens within a cell. Understanding these networks provide a snapshot to the regulatory mechanisms that operate within the cellular milieu. The advent of yeast 2 hybrid (Y2H), chromatin IP assay (CHIP assay), microarray, serial analysis of gene expression (SAGE) and two-dimensional polyacrylamide gel electrophoresis (2D-PAGE), and other associated low-throughput as well as high-throughput techniques have accelerated the rate at which data points are added to these networks. This is clearly indicated by the rate at which PubMed grows. PubMed currently has in its repository more than 16 million biomedical articles. The total number of articles published in the year 2005 alone was 666,029, which amounts to more than 1800 records per day [1]. The flood of information is making it increasingly difficult to comprehensively accumulate all known information into building context specific regulatory networks manually. A partial answer to this problem is the creation of databases that enable systematic storing and context specific retrieval of this data. The other essential piece to this puzzle is populating these databases. For this purpose, two different approaches are followed each with its bottlenecks and advantages, namely, manual curation and automated extraction of data using natural language processing (NLP). Here we attempt to evaluate the two methods comparatively and identify the gaps in the data generated.

Manual curation refers to the process wherein data present in the abstracts/full-length articles is manually read by trained personnel and the set of relevant data is extracted and classified into predefined fields. This would be the preferred method if the focus were to be on the quality and comprehensiveness of the data extracted, although time would be a major constraint. The other method used is automated extraction of data using natural language processing technologies. These are fast but the accuracy of the data captured and the data points that are missed out comprise major areas that need to be improved.

The initial years of work in the field of automation was restricted to the identification of protein names, gene names, co-occurrence of words [2, 3]. This evolved to employ different processes such as pattern matching [4], full [5, 6], and partial parsing [7], dynamic programming [8], and rule-based approaches [9] to enhance the performance. Many of the above-mentioned tools are well accepted by their specific niche client community and common standards to evaluate these multiple platforms are needed. The most widely used tools have been discussed in detail in the next section. This technology represented a new wave as it found

direct application in extracting data from biomedical literature including protein interactions, from articles published in MEDLINE [10].

There are a large number of NLP tools available both in the proprietary as well as public domain. Each tool has its reported precision and recall measures. Precision refers to the ability of a tool to retrieve technically accurate interaction details (minimal false positives), and recall measures its ability to retrieve the complete set of interactions from a selected pool of abstracts/full-length articles (minimal false negatives). The precision and recall rates vary widely between different tools. Methodologies used to build some of these tools and their features are described below.

In the public domain, there are multiple tools reported and these include GENIES, BioRAT, IntEX, and Pubminer to name a few.

GENIES utilizes a grammar-based NLP engine for information extraction. It includes substantial syntactic knowledge interleaved with semantic and syntactic constraints. This tool has a reported precision of 96% and recall of 63% [11]. Another tool called BioRAT uses labeling of words according to their parts of speech. A recall of 20.31% and a precision of 55.07% are reported for abstracts with 43.6% recall with 51.25% precision for full-length papers [12].

IntEx is a syntax-driven interaction extractor that tags biological entities with the help of biomedical and linguistic ontologies. IntEx has a reported precision of 45% and recall of 63.64% for abstracts [13].

Another information extraction tool which works on NLP technique is PubMiner. The precision and recall for extracted interaction were 80.2 and 73.9%, respectively [14].

PreBIND searches literature (abstract or title fields) based on protein names (Swiss prot) and gene symbols from RefSeq and SGD databases. Textomy, a support vector machine (SVM) text processing software, forms the core of this tool. This software initially retrieves abstracts from PubMed and assigns a score based on the likelihood of the abstract containing interaction information and identifies the interaction pair. The sentences describing the interaction get highlighted, which makes it easier to analyze the SVM's decision. Textomy also highlights protein names (derived from Swiss-Prot), organism names (derived from MeSH), and interaction phrases (programmed using PERL). PreBIND tool was reported to give a precision of 92% and recall of 92% [15].

Rule-based literature mining system for protein phosphorylation (RLIMS-P) is a text mining tool designed to specifically capture protein phosphorylation information from PubMed abstracts. This tool detects three types of objects from PubMed, namely, agent, theme, and site. RLIMS-P consists of a preprocessor, an entity recognizer, a phrase detector, and a semantic-type classification and relation identifier. These split the text into sentences and words, assign POS tags, detect acronyms and terms, identify phrase, nouns, and verb groups within a sentence, and also identify both verbal and nominal forms. RLIMS-P achieved a precision and recall of 97.9 and 88.0% for extracting protein phosphorylation [9].

MedScan from Ariadne Genomics is a commercially available and widely used tool to extract protein interaction information. This product comprises of a preprocessor, tokenizer, recognizer and syntactic parser, and semantic interpreter [5] all of which together recognize the components and build an interaction event. Reported precision and recall rates were 91% and 21%, respectively [16].

We attempted to analyze the performance and accuracy of two of these tools available in the public domain in comparison to manual curation. A major hurdle we faced in this process was the nonavailability of many of the tools cited in the public domain. Though each of these tools are backed by publications, there are no set of parameters that can be cross compared across these platforms and the reported recall and precision are not generated based on a common set of rules. Also, there is no definition for the sample size to be used for analysis and the spread of content.

Here we have provided the essential elements for an interaction to be termed complete. Also, it has been observed that abstracts are used as a source of protein interaction information. We analyzed the accuracy and completeness of information obtained from abstracts in comparison to the full-length articles as a measure of reliability of abstracts as sources of protein interaction data.

2. METHODS

Selection of articles and abstracts for analysis

A set of 350 articles pertaining to breast cancer were selected and downloaded from PubMed. Interactions were extracted from the manually curated databases, namely, PathArt and IMaps. Two NLP tools were downloaded and the selected sets of articles/abstracts were fed to generate the interaction pool. The interaction sets obtained from both manual curation and NLP were evaluated manually to determine the relevancy of the data and percentage of recall. Evaluation was carried out independently by two different teams to avoid any errors in data interpretation.

2.1. Manual curation

PathArt (proprietary pathway database from Jubilant Biosys Ltd.) is a manually curated database which covers more than 2800 signaling and metabolic pathways across 34 diseases and 20 physiologies extracted from peer-reviewed articles. PathArt captures protein-protein interactions from scientific articles in a pathway perspective. Pathways are classified into disease and physiology groups. Each interaction, in addition to reaction mechanism (activation, inhibition, translocation, etc.) and mode (phosphorylation, acetylation, etc.) gives information on animal model, detection method, and intracellular localization (cytoplasm, membrane, and nucleus). Data is manually entered into PathArt using a curator work bench, a software tool that accepts data in a defined format, and has inbuilt validations for accepting data. This product is well accepted among microarray and drug discovery researchers.

Data from the selected set of full-length articles (350 breast cancer articles) was retrieved from PathArt and used

to validate the interactions extracted using the selected NLP tools.

For obtaining the pool of interactions from abstracts, IMaps (proprietary protein interactions maps database from Jubilant Biosys Ltd.) was used. IMaps is a manually curated database with more than 200 000 protein-protein, protein-RNA, protein-small molecule, and protein-DNA interactions from 17 different organisms.

The curated data from IMaps for the selected set of 350 breast cancer related articles was retrieved and taken up for further analysis.

Guidelines followed for capturing interactions manually and validating interactions derived from NLP tools.

(i) To consider an interaction complete, information on source protein along with its interacting partner, interaction mechanism, evidence statement, and article reference ID are considered mandatory. Additional details captured include organism-related information wherever available.

(ii) In addition to capturing interaction details, information on animal model (cell line, cell type, tissue), reaction (direct or indirect), detection method, disease name, and physiology are also captured wherever available.

(iii) PathArt and IMaps consider the following set of verbs to define an interaction event: accumulation, acetylation, activation, association, bind, cleavage, colocalization, complex formation, deacetylation, deactivation, decrease, degradation, dephosphorylation, dimerize, dissociation, downregulation, efflux, expression, hydrolysis, inactivation, increase, induction, influx, inhibition, interaction, internalization, methylation, phosphorylation, proteolysis, regulation, release, secretion, sensitization, stimulates, synthesis, translocation, ubiquitination, upregulation.

(iv) Interactions are not captured from title of the article, introductory statements, and discussion (the reasons for this is discussed in detail in the later sections). Also, interactions are not captured from references cited.

(v) Entrez Gene standards are used to represent protein names. Those components not present in reference databases such as Entrez Gene, Swiss-Prot are manually annotated by an internal ontology team.

2.2. NLP tools

The following NLP tools were used for analysis.

PreBIND

PreBIND was accessed via the web interface at http://prebind.bind.ca. The selected set of 350 breast cancer related articles were used to generate protein interaction data. Each PubMed reference identifier (one at a time) was pasted on the search page. Results appeared within a few seconds in a new HTML page, along with the corresponding abstract. These results were then copied to a Microsoft Excel file.

RLIMS-P

RLIMS-P was accessed via the web interface at http://pir.georgetown.edu/pirwww/iprolink/rlimsp.shtml. The same set of 350 breast cancer related articles used for PreBIND analysis was used to generate protein interaction data by RLIMS-P. PubMed reference identifiers were pasted on the search page. Result appeared within a few seconds with the respective phosphorylation sites for source and target protein highlighted in the corresponding abstract. Results were copied into an Excel file for analysis.

Data analysis

Results obtained from PreBIND and RLIMS-P where cross verified with data from IMaps. The IMaps data was comparatively analyzed with PathArt to understand the differences in using full-length articles as a source of data versus abstracts.

Calculation of Precision and Recall rate

$$\begin{aligned} \text{Precision} &= \text{TP}/(\text{FP}+\text{TP})*100, \\ \text{Recall} &= \text{TP}/(\text{FN}+\text{TP})*100, \end{aligned} \tag{1}$$

where TP is true positive, FP is false positive, and FN is false negative [9, 17].

3. RESULTS

The present exercise was carried out to comparatively evaluate manual curation and NLP-based technologies with a focus on the advantages and bottlenecks in each of these approaches. Also provided are the pointers to overcome these bottlenecks.

For this, each interaction extracted with selected NLP tool was read and classified as true or false based on the guidelines defined in Section 2. IMaps and PathArt data was taken as the standard set (with precision and recall of 100%) as it was manually curated and quality checked. This was followed by cross comparison with the interaction set from IMaps (at the abstract level) and PathArt^{TM} (at the full-text level) so as to assess the completeness of the data. The focus of this exercise was also to find false-negative and false-positive interactions and the data generated was used to determine the precision and recall rates (Table 1) (Figures 1 and 2).

As depicted in Table 1, IMaps was compared with two different tools: PreBIND and RLIMS-P. In case of PreBIND, all the interactions present in the set of abstracts analyzed were comparatively analyzed. On the other hand, RLIMS-P is a specific tool that detects only phosphorylation events. For an accurate comparison, only phosphorylation events retrieved from IMaps (the number amounting to 119) were taken into consideration.

A total of 350 abstracts were processed through PreBIND as well as manually used IMaps, to extract the pool of protein interactions. These were analyzed for precision and recall as described in Section 2.

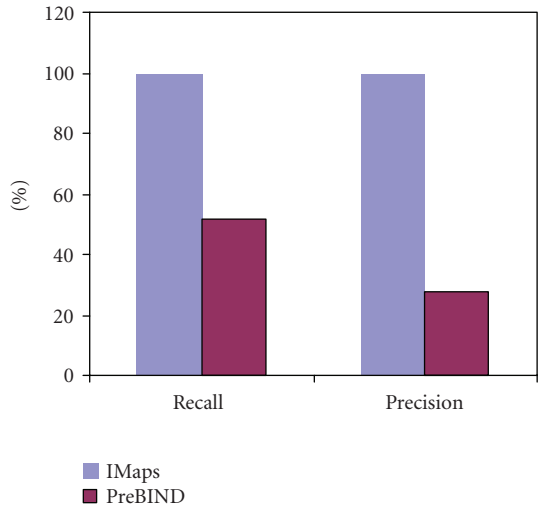

Figure 1: Recall and precision rates for IMaps and PreBIND.

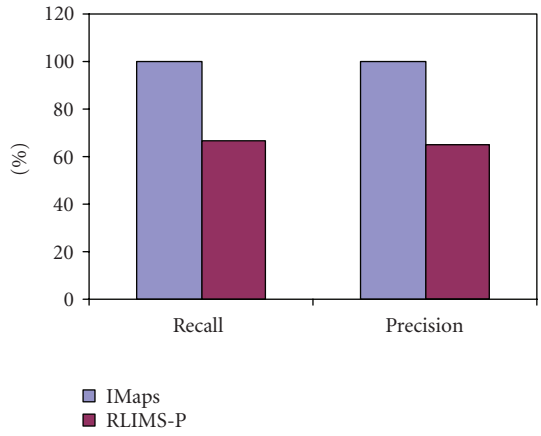

Figure 2: Recall and precision rates for IMaps and RLIMS-P.

A total of 350 abstracts were processed through RLIMS-P as well as manually used IMaps, to extract the pool of protein interactions. These were analyzed for precision and recall as described in Section 2.

4. COMPARATIVE ANALYSIS

The precision and recall rates were found to be lower for all the two NLP tools compared to the scores mentioned in their respective articles (PreBIND 92% and 92% and RLIMS-P 97.9 and 88.0%, resp.). Due to the apparent disparity, we analyzed the set of rules followed in order to classify an interaction as false or true. Our analysis brought into light some of the key points based on which, interactions were treated as true by the selected NLP tools and false by manual curation.

(i) Interactions were taken from introductory and discussion statements.

(ii) Interactions were taken from cited references.

(iii) Interactions were captured from the title of the articles.

All interactions captured from titles, discussions, introductory statements, and back references were classified as false in manual curation.

The detailed analysis carried out revealed several other types of errors apart from those discussed above. These errors observed were grouped under the subheadings of ontology, data misinterpretation, incomplete data capture, and irrelevant data capture and each of these error types is discussed in the following sections with suitable examples.

4.1. Ontology

Ontology refers to a standardized naming convention used to define specific parameters like gene name, cell line, Disease name, and others, for example, Entrez Gene standards for gene names and their corresponding aliases. NLP tools adopt these standards. Despite this, we could find several instances where gene names were inappropriately captured by the selected NLP tools despite the correct isoform being mentioned in the article (example in Table 2). Mapping of genes based on their aliases leads to incorrect component annotation which in turn results in an interaction being wrongly captured (Table 2).

Though the Entrez Gene database and other such resources are taken as standards for gene name annotation, our experience in manual curation has brought out several limitations in this process. Some instances where we fail to obtain corresponding gene name standards are outlined in Table 3.

4.2. Data misinterpretation

Several types of data misinterpretations were observed. One instance was where the tool fails to distinguish between protein and protein reagents that lead to generation of wrong interactions (Table 4). Another instance was where an interaction is drawn between protein and its corresponding siRNA, antibody or specific inhibitor. This might be technically correct, but it would be incorrect to infer it as a physiological process that occurs naturally in a living system since these are reagents used to understand or elicit a physiological effect in vivo/in vitro.

Heterogeneity in the language used by authors to represent data and sentence complexity in many instances leads to wrong representation of interaction data (Table 4). This also results in assigning the wrong interaction verb. In some cases, interactions were retrieved from irrelevant articles/abstracts, for example, PreBIND could derive 42 interactions from an abstract focused on enzyme kinetics (PMID: 7968216).

4.3. Incomplete data capture

The selected NLP tools failed to capture a large number of true interactions whenever an interaction sentence failed to confine to the pattern recognized these tools. In addition, these tools fail to capture interactions which involve mechanisms like complex formation, cleavage, translocation, and so forth due to limited mechanism definition (Table 5).

Table 1: Comparative analyses of precision and recall rates (abstract level data extraction).

Tools	No. of articles	Total No. interactions	True interaction	False positive	False negative	Recall (%)	Precision (%)
IMaps	350	1750	1750	0	0	100	100
PreBIND	350	4637	102	4535	1648	51.50088	27.84407
IMaps	350	119	119	0	0	100	100
RLIMS-P	350	119	64	55	64	66.85393	65.02732

Table 2: Ontology errors: few examples.

Type of error	Tool used	Interaction	Evidence statement	Manual curation	PMID	Comment
Inferring gene names based on aliases	PreBIND	MYC—PS2	Absent	Nil	1899037	Gene pS2 corresponds to TFFI (Trefoil factor 1). In PreBIND few interactions have been tagged to pS2 and remaining to PS2 (Presenilin 2) randomly
Annotation of protein names	PreBIND	PI (serine (or cysteine) proteinase inhibitor, clade A)—PIP (Prolactin-induced protein)	Absent	Nil	7968216	Phosphatidylinositol was wrongly annotated as serine (or cysteine) proteinase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 1 and PI 4-phosphate 5-kinase as prolactin-induced protein.

4.4. *Irrelevant data capture*

Building interaction around irrelevant entities such as saline and buffer, is a major factor which reduces the precision of the tested NLP tools. Also PreBIND tries to bring together any two proteins which cooccur in a sentence, resulting in erroneous interactions (Table 6).

5. NLP AND MANUAL CURATION

The aim of this analysis was to compare results obtained from the selected NLP tools with manual curation, bringing out deficiencies in both with an unbiased view. Manual curation has its own flaws. It is a highly time-consuming process and also requires strict measures to ensure that heterogeneity in data interpretation and capture among multiple curators is effectively weeded out. In our experience, scientific literature is represented in different styles that are highly individualistic. Thus, multiple avenues exist for heterogenous/misinterpretation of data. These can be tackled at two levels, namely, at the level of data entry and quality check. During data entry, errors by curators can be minimized through enforcing strict guidelines as well as by brining standardized platforms for data entry with inbuilt validations. The second level is during quality check where in, validation scripts can be used to retrieve data in bulk after it is entered in to the database and crosschecked. All these are time and manpower intensive. Therefore, manual curation by its nature is not error free if appropriate processes are not followed.

Curator at an average reads 2 to 10 articles in a day which is again based on the expanse of data that needs to be captured. If this were to be coupled to manual quality check, it becomes a highly time-consuming process. Though efficiencies can be built in to the system, it cannot be compared to the speeds achieved by NLP tools. Despite the high quality obtained through manual curation, NLPs represent a more effective and efficient way of data capture

6. MANUAL CURATION USING FULL-LENGTH AND ABSTRACT

Several databases both in public domain as well as proprietary domain use abstract as the sole source of information. Abstracts by definition provide the gist of information published in the paper but do not provide the experimental details to make the information content complete. The major advantages in processing the abstracts include time and free availability. Full-length articles provide access to the scientific content in its entirety along with authors perspective to the work carried out in the form of discussion. The major bottleneck in processing full-length articles is access to all that are cited in PubMed for protein interaction information which is both cost and effort inhibitory.

To understand the difference between the abstract and full-length curation, we compared the data generated from

Table 3: Limitations found in standardization of gene names using Entrez Gene.

Error type	Example
Splice variants	Delta FosB, FBJ murine osteosarcoma viral oncogene homolog B delta, a splice variant of FOSB, is not annotated by Entrez Gene (11854297)
Protein isoforms	STAT1 alpha, an isoform of STAT1 protein, is not annotated by Entrez Gene (14532292)
In case of interactions involving components of a multi-subunit protein	Guanine nucleotide binding protein beta subunit, G beta subunit (8752121)
Rare proteins	Novel gene A1 involved in apoptosis (15480428)
Several components do not have their isoforms annotated across different organisms	CYP2C40, CRYGE are present in mouse and rat and not in human, and CD200R3, CD200R4 are present in mouse and not in human
Interaction involving protein complexes and not individual proteins	T cell receptor complex (9582308) Transcription factor AP1, activator protein 1 (11062239)
Where authors do not mention the specific isoform they are working with and/or mention the entire class of proteins.	Farnesyltransferase (11222387) SMAD, mothers against DPP homolog (11331769)

Table 4: Data misinterpretation: some examples.

Type of error	Tool	Interaction	Evidence statement	Manual curation	PMID	Comment
Wrong interaction	RLIMS-P	STAT3—LEP	Leptin induced time and dose-dependent signal transducer and activator of transcription 3 (STAT3) phosphorylation.	LEP (reference)—phosphorylation (indirect)—STAT3 (homosapiens) [in vitro, MCF-7 cells]	15313931	Sentence complexity leads to reverse representation of the interaction
Wrong interaction	RLIMS-P	MAPK—RAF	The phosphorylation of MAPK by GHRH was prevented by transfection of the cells with dominant-negative Ras or Raf or by pretreatment of cells with Raf kinase 1 inhibitor	GHRH—RAF (homosapiens)—phosphorylation (indirect)—MAPK (homosapiens) [in vitro, MDA-231 cells]	16613992	Data complexity leads to misinterpretation

Table 5: Incomplete data capture: some examples.

Type of error	Tool used	Interaction	Evidence statement	Manual curation	PMID	Comment
Incomplete data capture	PreBIND	Nil	17 beta-estradiol (E2) ablation enhanced expression of TRPM-2 the in MCF-7 human mammary adenocarcinoma cells, indicating that presence of E2 decreased the expression of TRPM2 and TGFB1	Estradiol—downregulate (indirect)—TGFB1, TRPM2 (homosapiens) [in vitro, MCF7 cells]	1899037	PreBIND fails to capture a relevant interactions from abstracts
Incomplete data capture	RLIMS-P	Nil	Heregulin (HRG)-beta1 induced tyrosine phosphorylation of erbB2 and erbB3 receptor heterodimers and increased the association of the dimerized receptors with the 85-kDa subuint of phosphatidylinositol 3-kinase (PI3K)	HRGB1(homosapiens)—phosphorylation (indirect)—ERBB1.ERBB3 (homosapiens) [in vitro, MCF7 cells]	10197638	RLIMS-P failed to capture information on both source and target protein (HRGB1—ERBB2.ERBB3

Table 6: Irrelevant data capture: some examples.

Type of error	Tool used	Interaction	Evidence statement	Manual curation	PMID	Comment
Erroneous interaction	PreBIND	FOS—pS2	c-fos, c-H-ras, and pS2, decrease following E2 ablation.	Nil	1899037	PreBIND tries to bring together any two proteins which cooccur, which results in erroneous set of interactions
Erroneous interaction	PreBIND	PI (serine (or cysteine) proteinase inhibitor, clade A)—MB (Myoglobin)	Nil		7968216	The component name has been captured from cell line MDA- MB -435 cells

Table 7: Comparison of full-length and abstract curation results.

	Full-Length Articles	Abstracts
Sample size taken for analysis	350	350
Interactions derived from	334	294
No. of articles without interactions	16	56
No. of components without organism information (source + target)	46 + 11	368 + 21

PathArt and interaction maps. A quick look through the data indicated that about 40 articles provided interaction information only at the full-text level. A large number of interactions could not be retrieved when abstracts were used as a sole source of information (Table 7). In addition, a large number of interactions derived from abstracts failed to provide information on detection method, interacting domain, cellular localization of the interacting partners, and so forth. Also, abstracts often lack information about the organism in which the study is carried out, thus missing out on vital information on organism specific regulatory networks. With the data presented in the abstract, it is difficult to differentiate in several instances if an interaction studied is structural (direct) or functional (indirect) (Table 8).

We also carried out an analysis to find the extent to which essential interaction details such as organism information were missed out in abstracts. The data obtained is depicted in Table 7. This type of data becomes essential for constructing organism specific interaction networks.

7. DISCUSSION

We present a comparative analysis of two of the publicly available NLP tools (PreBind and RLIMS-P) with manual curation. The next level of analysis provided is between two different manual curation methods developed using different information sources, namely, abstract and full-length articles.

We selected PreBIND as BIND is one of the most widely used public domain protein interaction resources and is built using PreBIND. Also the reported rates of recall and precision are very high for PreBIND. We could compare the results obtained from this tool directly to IMaps as both the systems derive interactions from abstracts. Errors were detected at multiple levels in data retrieved using PreBIND; a large number of valid interactions were missed out (false negatives) and a similarly large number of irrelevant interactions (false positives) were constructed. We had similar experiences with some of the commercially available tools (data not provided). Another major problem encountered is misinterpretation of data. Here, errors were introduced into the interactions as the tested tools were not able to interpret the complexity of natural language used to represent scientific data.

One of the major drawbacks of PreBIND is that it identifies cooccurrences of biomolecules in a sentence as an interaction leading to the generation of erroneous interaction. This can be overcome by adapting it for full-text mining, where there would be clear cut differentiation between interactions and mere cooccurrences of proteins.

The other tool evaluated is RLIMS-P. This is a highly specific tool that identifies and retrieves phosphorylation facts from abstracts. Since the number of verbs that go into defining this niche set of interactions is limited, achieving high recall and precision rates seems a real possibility. This tool also has high precision and recall rates reported in literature.

We formulated a set of guidelines to define an interaction as there are no comparative studies carried out across NLP tools with normalized set interaction definitions in literature. Interactions taken from the title of the paper, introduction, and discussion are categorized as false unless validated by experimental data. Introduction usually provides a preamble to the paper and does not present the original findings represented in the article and the aim of using NLP tools is to mine all the interaction data and each paper presenting an interaction fact would be eventually covered. Repeat mining of the same set of interactions from back/cross references compiled from introductory statements would introduce redundancy

Table 8: Manual curation using full-length and abstract.

Type of error	Full text interaction (evidence statement)	Abstract interaction (evidence statement)	PMID	Comment
Incomplete data capture from abstract	Estradiol—Upregulation (Indirect)-FOS (homosapiens) [Northern Blot] (After Estrogen ablation, there is a 60-70-fold decrease in proliferation associated c-fos oncogene expression)	Estradiol—Upregulation (Indirect)-FOS (homosapiens) (17 beta-estradiol (E2) ablation decreased the expression of c-fos in MCF-7 human mammary adenocarcinoma cells, indicating that presence of E2 induced the expression of c-fos in these cells)	1899037	Abstract failed to provide information on detection method.
Organism information not available in the abstract	TNF (homosapiens)—Upregulation (Indirect)-SOD2 (homosapiens) [Northern Blot] (A 10-fold increase of MnSOD mRNA was observed after exposure to exogenous human TNF for 6 hours)	TNF (-)—Upregulation (Indirect)-SOD2 (homosapiens) [Northern Blot] (Northern blot analysis indicated that following TNF stimulation, the expression of 4-kilobase and 1-kilobase manganese superoxide dismutase mRNAs were 9- to 10-fold induced in MCF7AdrR cells)	7905787	For the interactionbetween TNF and SOD2 the source organism data is present in the full-length article and not in the abstract.
Incomplete data capture	EGF (Reference)—Activation (Direct) EGFR (homosapiens) [Immunoprecipitation] (EGF activated ErbB-2 by binding and activating its receptor EGFR)	EGF (Reference)—Increase—Phosphorylation (Indirect) EGFR (Reference) (Epidermal growth factor (EGF) induced the activation of ErbB-1 in cell lines naturally expressing ErbB-1 protein)	9130710	Information present in the abstract is not sufficient to indicate that the interaction is structural.

into the database and become a hindrance in statistical analysis of interaction data. The discussion part also very often contains statements that would appear as valid interaction facts to an NLP tool but could be mere pointers or inferences drawn by the authors for which there might be no experimental evidence presented in the paper. This could generate potentially large number of unproven interaction data. Thus, the NLP tools can achieve higher precision by attributing different weightage to data retrieved from different sections of the paper and also to interaction facts reiterated across different sections.

The information density is much higher in abstracts. This is attributed to the presence of a large amount of background information and experimental details in full-length articles [18]. Other advantages in using abstracts include their availability in plain text and absence of special characters or super/subscripts. Despite these apparent advantages, we found several instances where in information present in the abstract misses a few if not all interaction facts present in the full-length article. This prompted us to analyze the differences in detail. Our results indicate that several interactions and interaction details are missed out in abstracts. The reasons for this include, the complexity of language used in generating the summary of the entire article as well as lack of experimental details that can lead to data misinterpretation. Thus, NLP tools should be trained to accommodate both full-length articles as well as abstracts based on the intended end application. If building comprehensive networks or understanding the interactions in detail is required, then it would be advisable to use full-length articles as the source of information and on the other hand, if genome wide networks are to be generated where in time becomes a limiting factor, abstracts form an ideal choice.

Organism specific interaction and pathway data are increasingly being recognized as vital to evolutionary studies as well as understanding species specificity in different responses including drug reactions. If the final application of the interactions data derived is for these purposes, then full-text articles should be used for extraction rather than abstracts.

To summarize, machine learning methods are useful as tools to direct interaction and pathway database back-filling; however, this potential can only be realized if these techniques are coupled with human review and entry into a factual database such as PathArt and IMaps. An alternative approach could be to improvise to make each of the steps in data extraction fool proof. For example, most of the NLP tools, while screening through the article, detect interacting components along with interaction mechanism based on a well-defined pattern set. Though a large number of sentences follow this pattern, several cases exist wherein, the complexity of sentence results in incorrect data capture. A probable solution to this could be using large training sets that represent all possible real time complexities in data representation

while designing NLP tools in future. Other areas of improvement include gene mapping, which should be extended from presently used standard databases (Entrez Gene and Swiss-Prot) to manually annotated lists to include alias and isoform mapping deficiencies discussed in the experimental section. Capturing protein-small molecule interactions adds onto the error rate as any nonprotein molecule present within a sentence which conforms to the interaction rules would result in the generation of erroneous interactions. Small databases like CAS or PubChem should be used as reference to identify and annotate protein-small molecule interaction. Limitations exist in coverage of interactions mechanisms that affect recall rates or generate errors in captured interactions. An exhaustive verb list with real time examples built into the training set would be an ideal solution.

The above-suggested modifications are based on the set of analyses carried out by us using two of the NLP tools available in the public domain. This needs to be extended to a larger sample pool of NLP tools. The need of the hour is to develop a consortium of all (public domain if not proprietary) NLP tools for extracting interaction facts so that data obtained from each of these could be analyzed comparatively and interaction repositories could be built by cross validation and complementation.

REFERENCES

[1] L. Hunter and K. B. Cohen, "Biomedical language processing: what's beyond PubMed?" *Molecular Cell*, vol. 21, no. 5, pp. 589–594, 2006.

[2] K. Fukuda, A. Tamura, T. Tsunoda, and T. Takagi, "Toward information extraction: identifying protein names from biological papers," *Pacific Symposium on Biocomputing*, pp. 707–718, 1998.

[3] M. Stephens, M. Palakal, S. Mukhopadhyay, R. Raje, and J. Mostafa, "Detecting gene relations from Medline abstracts," *Pacific Symposium on Biocomputing*, pp. 483–495, 2001.

[4] T. Sekimizu, H. S. Park, and J. Tsujii, "Identifying the interaction between genes and gene products based on frequently seen verbs in medline abstracts," *Genome informatics*, vol. 9, pp. 62–71, 1998.

[5] S. Novichkova, S. Egorov, and N. Daraselia, "MedScan, a natural language processing engine for Medline abstracts," *Bioinformatics*, vol. 19, no. 13, pp. 1699–1706, 2003.

[6] A. Yakushiji, Y. Tateisi, Y. Miyao, and J. Tsujii, "Event extraction from biomedical papers using a full parser," *Pacific Symposium on Biocomputing*, pp. 408–419, 2001.

[7] J. Thomas, D. Milward, C. Ouzounis, S. Pulman, and M. Carroll, "Automatic extraction of protein interactions from scientific abstracts," *Pacific Symposium on Biocomputing*, pp. 541–552, 2000.

[8] M. Huang, X. Zhu, Y. Hao, D. G. Payan, K. Qu, and M. Li, "Discovering patterns to extract protein-protein interactions from full texts," *Bioinformatics*, vol. 20, no. 18, pp. 3604–3612, 2004.

[9] Z. Z. Hu, M. Narayanaswamy, K. E. Ravikumar, K. Vijay-Shanker, and C. H. Wu, "Literature mining and database annotation of protein phosphorylation using a rule-based system," *Bioinformatics*, vol. 21, no. 11, pp. 2759–2765, 2005.

[10] T.-K. Jenssen, A. Lgreid, J. Komorowski, and E. Hovig, "A literature network of human genes for high-throughput analysis of gene expression," *Nature Genetics*, vol. 28, no. 1, pp. 21–28, 2001.

[11] C. Friedman, P. Kra, H. Yu, M. Krauthammer, and A. Rzhetsky, "GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles," *Bioinformatics*, vol. 17, supplement 1, pp. S74–S82, 2001.

[12] D. P. A. Corney, B. F. Buxton, W. B. Langdon, and D. T. Jones, "BioRAT: extracting biological information from full-length papers," *Bioinformatics*, vol. 20, no. 17, pp. 3206–3213, 2004.

[13] S. T. Ahmed, D. Chidambaram, H. Davulcu, and C. Baral, "IntEx: a syntactic role driven protein-protein interaction extractor for bio-medical text," *Association for Computational Linguistics*, pp. 54–61, 2005.

[14] J. Eom and B. Zhang, "PubMiner: machine learning-based text mining for biomedical information analysis," *Genomics & Informatics*, vol. 2, no. 2, pp. 99–106, 2004.

[15] I. Donaldson, J. Martin, B. de Bruijn, et al., "PreBIND and Textomy—mining the biomedical literature for protein-protein interactions using a support vector machine," *BMC Bioinformatics*, vol. 4, no. 1, pp. 11–23, 2003.

[16] N. Daraselia, A. Yuryev, S. Egorov, S. Novichkova, A. Nikitin, and I. Mazo, "Extracting human protein interactions from Medline using a full-sentence parser," *Bioinformatics*, vol. 20, no. 5, pp. 604–611, 2004.

[17] H. Jang, J. Lim, J.-H. Lim, S.-J. Park, K.-C. Lee, and S.-H. Park, "Finding the evidence for protein-protein interactions from PubMed abstracts," *Bioinformatics*, vol. 22, no. 14, pp. e220–e226, 2006.

[18] D. P. A. Corney, B. F. Buxton, W. B. Langdon, and D. T. Jones, "BioRAT: extracting biological information from full-length papers," *Bioinformatics*, vol. 20, no. 17, pp. 3206–3213, 2004.