Predicting Future Scientific Discoveries Based on a Networked Analysis of the Past Literature

Meenakshi Nagarajan[*,1], Angela D. Wilkins[*,3], Benjamin J. Bachman[*,3], Ilya B. Novikov[3], Shenghua Bao[1], Peter J. Haas[1], María E. Terrón-Díaz[3], Sumit Bhatia[1], Anbu K. Adikesavan[3], Jacques J. Labrie[1], Sam Regenbogen[3], Christie M. Buchovecky[3], Curtis R. Pickering[2], Linda Kato[1], Andreas M. Lisewski[3], Ana Lelescu[1], Houyin Zhang[3], Stephen Boyer[1], Griff Weber[1], Ying Chen[1], Lawrence Donehower[3], Scott Spangler[1*], Olivier Lichtarge[3*]

[1]IBM Almaden Research San Jose, California

[2]The University of Texas MD Anderson Cancer Center Houston, Texas

[3]Baylor College of Medicine Houston, Texas

ABSTRACT

We present KnIT, the Knowledge Integration Toolkit, a system for accelerating scientific discovery and predicting previously unknown protein–protein interactions. Such predictions enrich biological research and are pertinent to drug discovery and the understanding of disease. Unlike a prior study, KnIT is now fully automated and demonstrably scalable. It extracts information from the scientific literature, automatically identifying direct and indirect references to protein interactions, which is knowledge that can be represented in network form. It then reasons over this network with techniques such as matrix factorization and graph diffusion to predict new, previously unknown interactions. The accuracy and scope of KnIT's knowledge extractions are validated using comparisons to structured, manually curated data sources as well as by performing retrospective studies that predict subsequent literature discoveries using literature available prior to a given date. The KnIT methodology is a step towards automated hypothesis generation from text, with potential application to other scientific domains.

Categories and Subject Descriptors

I.2.6 [**Learning**]: Concept Learning and Knowledge Acquisition

General Terms

Algorithms, Experimentation.

Keywords

Text Mining, Scientific Discovery, Hypothesis Generation.

1. INTRODUCTION

Broadening the scientific framework for biological hypothesis generation requires collecting, understanding, and integrating diverse facts. Even many specialized topic areas are becoming too dense to be thoroughly studied and understood by the most well-read experts. For instance, the literature on p53, a single protein with a critical role as a master switch to fight cancer, now exceeds 75,000 published papers. Each one of these papers has observations that may support, contradict or contextualize observations from others. The size, complexity, and growth rate of the literature are even greater for multi-component subjects, such as diseases — the breast cancer literature, for example, now encompasses over 300,000 scientific articles in Medline.

Because biological hypotheses necessarily depend on the limited data, from literature or otherwise, that a single person can read and understand, the challenge in a field with as rich a literature as biology is how to harness the full depth and breadth of the knowledge that already exists in publications so as to formulate logically the best hypotheses consistent with the totality of the data. To do so, we aim: to convert unstructured data sources to structured data via natural language parsing; to represent that knowledge as a network of biological entities; and finally to reason over the resulting graph to identify novel, interesting, and testable hypotheses. Although the literature contains errors and machine translation of text will be incomplete and lack the depth of interpretation that scientists develop over a career, a central hypothesis of this work is that by focusing on only the highest confidence predictions it will be possible to reliably make inferences that lead to new discoveries when tested.

Baylor College of Medicine and IBM Research are engaged in a long-term partnership to create the necessary infrastructure, software, and algorithms to improve scientific discovery in this way. We have focused first on protein-protein interactions (PPIs), which have potential applications in drug discovery and understanding of disease. Our initial, proof-of-principle efforts[45] indicated that new p53 kinases could be predicted using a simple bag-of-words feature space representation combined with a graph diffusion reasoning technique based on labels *manually* curated by experts. In the current paper, we present a more complete, scalable, and general instantiation of the Knowledge Integration Toolkit (**KnIT**) for fully automated discovery.

In what follows, we describe how KnIT discovers PPIs in unstructured documents and then reasons over those interactions to generate hypotheses. The discovery phase comprises extraction and normalization of protein entities followed by extraction of

*These authors contributed equally

KDD '15, August 11-14, 2015, Sydney, NSW, Australia
ACM 978-1-4503-3664-2/15/08.
http://dx.doi.org/10.1145/2783258.2788609

PPIs. The latter task is accomplished by combining a catalog of curated relationship types with rule-based and machine-learning approaches to extract meaningful and accurate PPI triples from parsed text. In the reasoning phase, KnIT can deploy various graph algorithms on protein-network representations to predict new interactions.

We also describe several experiments designed to test KnIT's capabilities. In a time-stamped "retrospective" experiment, we validate predictions based on the literature up to a given date against actual subsequent discoveries. We then show that the discovered triples represent a substantial fraction of known protein-protein interactions, documented in the literature and elsewhere. Finally, we revisit our prior study[45] on phosphorylation of the protein p53 and show how the current version of KnIT avoids the need for manual curation of abstracts and thus can scale to large scientific corpora; in particular, we extend our earlier work to study phosphorylation of *any* protein.

Overall, KnIT embodies an integrated computational approach to formulating novel scientific hypotheses that accelerate laboratory discoveries. KnIT is applicable to practically any industry or domain that is faced with overwhelming data and stagnation in innovation. Information extraction and reasoning components of KnIT are built into Watson Discovery Advisor, an IBM product that aims to accelerate breakthroughs by making connections between different pieces of information. In this work we choose biology as our test bed to demonstrate the capabilities of KnIT.

2. THE HUMAN INTERACTOME

Proteins are the fundamental molecular machinery in the cell, and identifying their functions is a central problem in molecular biology. With ≈ 20,000 distinct proteins in humans[24], and tens of thousands of possible biological process annotations, disease associations, and molecular interactions, the number of experiments that would be required to verify every possible function for every protein is well beyond the capabilities of molecular biology experiments for the foreseeable future.

For example, an important characteristic of any protein is the set of other proteins with which it interacts; there are almost 200,000,000 protein pairs to investigate. To further complicate the issue, the types of relationships that can occur between any two proteins vary widely. To name a few, two proteins may have: a *binding* interaction, in which the two physically interact with each other to carry out a task within the cell; a *positive regulatory* interaction, in which the activity of one protein in the cell increases the amount or activity of another protein; a *negative regulatory* interaction, in which the activity of one protein in the cell decreases the amount or activity of another protein; a *modification* interaction (e.g. phosphorylation or ubiquitination), in which one protein causes a chemical change to the structure, and therefore function, of another protein; an *expression* relationship, in which the quantities of two proteins are frequently observed to be highly correlated or anti-correlated with one another; or a *co-localization* interaction, in which the two proteins are often found physically near each other, but not necessarily touching, in the cell. This type of information is often represented as a multimodal graph of relationships between proteins[21], for example StringDB[13]. Such graphs form a partial representation of the human *interactome*, i.e., the totality of protein interactions in humans.

Knowledge about some of these interaction types is already accessible in a structured data format. For example, expression relationships often come from published microarray data, which measure the levels of thousands of messenger RNAs (each encoding a distinct protein) in specific cells or tissues. Many examples of this type of data are available from the GEO database[2]. Binding interactions can also be identified by large-scale screens; however, these expensive experiments are known to have significant false negative and false positive rates[12]. They too, are documented in a slew of databases[5; 6; 15; 16; 27; 55; 56]. However, combined, these databases contain around 100,000 likely binding interactions, only 10,000 of which are considered to be high quality[13], whereas the total number of binding interactions occurring in the human interactome has been estimated[47] to be as high as 600,000, leaving as many as half a million to be discovered. Other structured data result from human reading of the biomedical literature and manual input into databases, such as the PhosphoSitePlus[23] database of protein modification interactions; this approach requires intensive labor to keep up with the latest developments in the scientific community. In summary, some protein relationship information is accessible, but the quality, coverage, and cost of acquisition of this information must be improved.

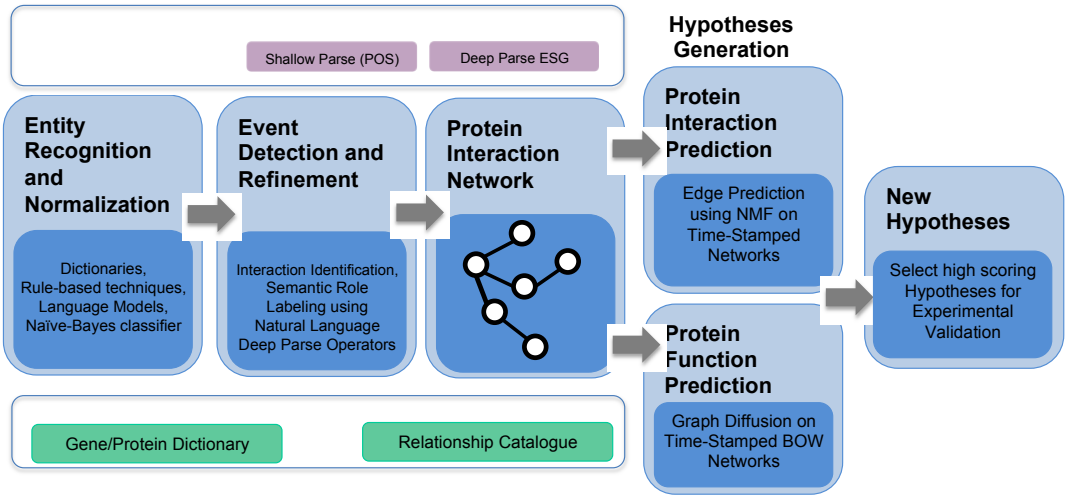

Figure 1 KnIT Extraction and Discovery Pipeline

Natural language processing is a useful tool for the extraction of human interactome data from the biomedical literature. Early text mining mostly entailed identification of entity co-occurrence within publication text. In recent years, co-occurrence has served as a benchmark for comparison with more sophisticated approaches, but is also still used in many database efforts[13]. Some of the early impactful efforts were iHOP[22] and GeneWays[43]. A recent effort has read the available literature (both abstracts and open access articles) and made their analysis available in the EVEX database[52]. Another effort by Microsoft called Literome[40] allows the user to search genes for possible relationships, and it suggests possible intermediate interactions found in the literature. Although these efforts help inform the scientific community, none have taken the next step in proposing new scientific hypotheses.

As the continuously dropping prices of DNA and RNA sequencing bring personalized genomics closer to reality, we gain vast amounts of raw data on the variations between individuals. However, there remains the challenge of contextualizing these variations in order to understand the source of patient symptoms arising from genetic disease, and to potentially predict mutation effects before they are known to cause a disorder. To help do this, we aim to better understand the human genome by documenting – and generating hypotheses upon – the interactome as described by biomedical scientists across tens of millions of published abstracts.

3. PROTEIN-NETWORK EXTRACTION

Biomedical relationship extraction is a well-studied problem[28; 38]. To identify connections between entities such as proteins, drugs and diseases, rule-based or machine-learning-based approaches[9; 49] typically use a knowledge base that catalogs known interactions, such as MetaMap[1] or GeneOntology,[15; 16] together with a text mining approach. Relationships have been mined using co-occurrence techniques in which two entities in a network are linked simply because they are mentioned within the same sentence[25]. Methods based on clustering similar text have also been used to generate networks between entities[4; 42]. In recent times, syntactic parsing of sentences to identify relationships between entities has also gained popularity[14; 37].

KnIT PPI extraction is guided by a catalog of known biological interaction types and relies on natural language parsing of sentences to establish explicit, directional relationships between known entities. Importantly, KnIT identifies not only direct connections between two proteins, but also *indirect* relationships where there is a secondary interaction or biological process at play. As a result, KnIT protein networks are more expressive and better at predicting heretofore unknown PPIs. KnIT relies on a rule-based approach for learning syntactic relations between known entities and known interactions in a labeled dataset. Learned rules are then applied to a Medline corpus for extraction of PPIs.

The KnIT extraction engine is built using IBM InfoSphere® BigInsights™ SystemT[30] – a powerful information extraction system for extracting structured information from unstructured and semi-structured text. SystemT provides us with basic text analytic capabilities such as sentence splitting, token detection, natural language parse of sentences, and so on. Deep-parsing components of SystemT are derived from an English Slot Grammar (ESG) parser[36] that provides core linguistic analysis.

In this section, we will describe the individual components that are central to KnIT PPI Extraction: (1) entity identification and normalization and (2) event recognition and normalization. (An *event* refers to a particular protein-protein interaction mentioned in the text, such as "Plk1 inhibits RSK1".) Figure 1 presents a detailed pipeline of tasks in KnIT PPI extraction.

3.1 Entity Identification and Normalization

Both rule-based[48] and machine-learning-based[20] approaches have been used for protein extraction and normalization. Recent work has also integrated simple rules with machine learning models. For example, Danger et al., integrate a dictionary look-up and a conditional-random-field classifier for recognizing pertinent entity types in protein-protein interaction contexts[10].

KnIT needs to ignore mentions of generic protein families, such as "Histone", and instead focus on mentions of specific proteins such as "Histone H3". This task can be very challenging. For example, given a mention such as "GRK-1,3 and 5 have strong impact", KnIT needs to extract not only GRK-1, but also GRK-3 and GRK-5. Moreover, KnIT needs to normalize proteins to their canonical forms; a key challenge here is that proteins often share synonyms or abbreviations. Thus KnIT must understand mentions of proteins according to their context and map them to standardized identifiers. To this end, KnIT uses a hybrid model for protein extraction and normalization. The protein annotation process involves three steps: candidate generation, candidate selection, and protein normalization.

3.1.1 Protein Candidate Generation

To ensure high recall for the protein extraction task, KnIT relies upon a comprehensive dictionary of human proteins compiled from NCBI[2; 54], the UniProt KnowledgeBase[51], HUGO[17], and CTD[11]. This dictionary has 340,000 entries, where each entry is mapped to one of over 30,000 protein canonical forms. When generating candidates, KnIT employs a combination of dictionary-matching, pattern-matching, and abbreviation-matching rules to maximize recall while maintaining sufficient precision.

Dictionary-matching rules: Because it is infeasible for a dictionary to capture all possible protein synonyms used in the literature, KnIT supports fuzzy matching of dictionary terms:

- Greek letters matches, e.g., "beta(2)-microglobulin" matches "β(2)-microglobulin";
- Bracket independent matches, e.g., "β(2)-microglobulin" matches "β2-microglobulin";
- Space/Hyphen independent matches, e.g., "β2-microglobulin" matches "β2 microglobulin" and "β2microglobulin".

KnIT also accounts for the different levels of ambiguity in the dictionary terms. During dictionary compilation, each dictionary term is semi-automatically categorized into one of three categories: unambiguous terms like "G protein-coupled receptor kinase 2", ambiguous terms like "ATM", and very risky terms like "C1" and "C2". Terms are categorized according to their length and character complexity; the current dictionary contains roughly 316K, 25K, and 1K terms in each category. Each mention that matches some dictionary term is assigned a confidence score of "high", "medium", or "low", depending on the category of the term. Mentions with a confidence score of medium or low are further verified using the context-based protein classifier described in Section 3.1.2 below.

Pattern-matching rules: Typically, some legitimate proteins do not appear in the dictionary at all, e.g., because researchers may not follow naming conventions or because some exploratory entity names are too new to be added to a dictionary. KnIT therefore enriches protein candidate generation using two types of patterns:

- Entities that contain both upper letters and digits with indicative words like "gene" or "protein" in the context.
- Entities comprising a known protein-family identifier followed by specific IDs, e.g., "BRCA-1, 2, and 3" or "Hemoglobin alpha, beta and gamma".

Abbreviation-matching rules: Abbreviations of proteins are commonly used in the literature. Frequently, the abbreviation itself is either ambiguous or does not appear in the dictionary at all. KnIT therefore detects all definition-abbreviation pairs in an abstract and then maps all mentions of an abbreviation in the abstract to its corresponding definition. For example, if the defining text "Stearoyl-CoA Desaturase (SCD)" appears in an abstract and "Stearoyl-CoA Desaturase" is identified as a protein using the dictionary, then all subsequent appearances of SCD are identified as "Stearoyl-CoA Desaturase" automatically. In this manner, all ad hoc abbreviations whose definitions are in the dictionary are correctly identified and disambiguated.

3.1.2 Protein Candidate Selection

Often, the candidate-generation process identifies entities that are false positives, i.e., they are not proteins. For example, SCD is both an abbreviation for "Stearoyl-CoA Desaturase" and an acronym for "Sickle Cell Disease". Eliminating false-positive candidates is a challenging but important task. Intuitively, the context of a mention that refers to a protein should be very different than the context of the mention when it does not refer to a protein. For example, when proteins are mentioned in literature, the surrounding text often discusses protein-specific biological processes such as regulation or phosphorylation. It may also mention other protein names, or explicitly refer to the entity as a protein. In order to exploit the cues in the surrounding context, we trained a naïve Bayes classifier that uses features derived from the neighboring context of a putative protein mention. KnIT uses the sentence containing the discovered mention as the context span and computes the following features:

- All words present to the left and right of the mention;
- The number of protein mentions in the sentence;
- Whether or not the two immediate words to the left and right of the discovered mention are protein names.

Words appearing in the left context and right context are treated separately. To evaluate the classifier, we used a dataset of 2282 positive examples and 2209 negative examples from BioNLP-ST 2013 tasks for event extraction and pathway curation; we used two-thirds of the data for training and tested on the remaining one-third. Our approach achieved an overall classification accuracy of 77.2%.

3.1.3 Entity Normalization

Researchers commonly refer to proteins by different synonyms or aliases, and KnIT must be able to deal with this phenomenon. Mapping a mention back to the correct canonical protein is a hard problem because different proteins often share the same synonym. For example, synonym "D1" maps back to both the Dopamine Receptor D1 (DRD1) and Leiomoden 1 (LMOD1). Whenever KnIT identifies D1 as a protein mention, it also needs to decide precisely which protein is being referenced. Fortunately, the *context* of a protein mention (i.e., the surrounding words) typically contains sufficient clues to identify the correct canonical protein, allowing disambiguation. For example, a mention of "dopamine" or the family name "GPCR" in the abstract would indicate that the abstract was discussing Dopamine Receptor D1 (DRD1).

KnIT adopts the idea of unigram language models from computational linguistics[39] and builds "context models" for proteins. For a protein g and an abstract d, denote by $S_{g,d}$ the set of sentences in d that reference g. The *context* $C_g(s,d)$ of a protein g in a sentence $s \in S_{g,d}$ is the multiset of terms that occur in s. The *context model* P_g of a protein g is defined as a probability distribution over the terms appearing in the protein's context. In order to estimate this distribution, we query all the abstracts in Medline corpus with the canonical name of the protein and select the set D_g of the 100 most relevant abstracts as ranked by TF-IDF score. We then define the combined context for g as $\tilde{C}_g = \bigcup_{d \in D_g} \bigcup_{s \in S_{g,d}} C_g(s,d)$ and the combined context for the corpus as $\tilde{C} = \bigcup_g \tilde{C}_g$; here we take multiset unions, with the latter union being over all proteins in the corpus. Let $V = \{v_1, v_1, \ldots, v_n\}$ be the vocabulary of all terms in the corpus. Then the probability of observing a term v in the context of protein g is defined as:

$$P_g(v) = \frac{(\text{cardinality of } v \text{ in } \tilde{C}_g) + 1}{(\text{cardinality of } v \text{ in } \tilde{C}) + n}$$

The additive factors in the numerator and denominator are used to smooth the probability distribution so that terms not present in $\tilde{C}_g$ are assigned a non-zero probability. We extend this definition to encompass the likelihood of seeing a given context C (i.e., a given multiset of terms) around g by setting $P_g(C) = \prod_{v \in C} P_g(v)$; i.e., we make a simplifying independence assumption *a la* naïve Bayes. Note that if a term v appears k times in the multiset C, then the foregoing product contains k terms equal to $P_g(v)$.

Whenever KnIT encounters an ambiguous protein mention, it uses the context models to identify the most likely canonical protein in a manner similar to query likelihood models in information retrieval[40]. That is, given an ambiguous mention M, its context C_M, and the set G of possible proteins referred to by M, KnIT determines the canonical protein g^* to maximize the likelihood of the context: $g^* = \text{argmax}_g P_g(C_M)$.

3.2 Event Extraction and Normalization

To extract interaction events, KnIT uses a catalog of PPI types together with natural language parsing of sentences. The catalog is maintained as a list structure similar to that used in the BioNLP 2013 pathway extraction task. The goal of extraction is to obtain triples, each comprising:

1. An *event*: an element appearing in the catalog of interactions described in Section 3.2.1 below.
2. An *agent*: a protein that is causally active in an event.
3. A *theme*: a protein that undergoes the effects of an interaction.

In general, depending on the interaction, extracted triples may or may not possess directionality. Binding relationships, for instance, do not represent directional events, whereas phosphorylation is an

example of an interaction that has a clear causal agent and affected theme.

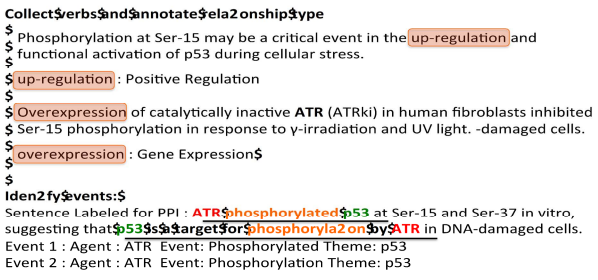

Figure 2 Curating Process Examples

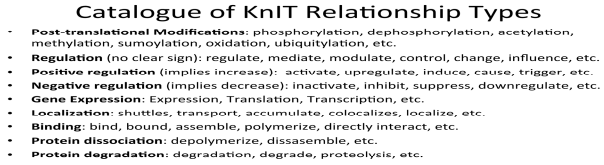

Figure 3 Relationship Types and Synonyms

KnIT extracts two kinds of triples that capture interactions between proteins – direct and indirect. The first is a simple triple where there is a direct relationship between known entities. The two sentences below express direct relationships between known entities Plk1 and RSK1, and between p53 and Notch1.

*In addition, studies on HeLa cells using Plk1 siRNA interference and overexpression showed that phosphorylation of RSK1 increased upon interference and decreased after overexpression, suggesting that **Plk1 inhibits RSK1**.*

*These data indicate that **p53 negatively regulates Notch1** activation during T cell development.*

In addition, KnIT also extracts relationships when the proteins are involved in more complex and indirect interactions. In the examples below we see that KnIT extracts relationships between bcl-2 and p53 that are indirect, since different processes (shown as underlined text) are involved in their interaction.

*Silencing of **Bcl-2** **induced** massive **p53**-dependent apoptosis.*

*Our data suggests that **bcl-2** is able to **modulate** transmembrane trafficking of **p53**.*

3.2.1 Building the KnIT Relationship Catalog

The KnIT relationship catalog lists many types of protein-protein interaction types (some described in Section 2) that were curated using multiple data sources. Post-translational modification (PTM) interactions were extracted from the Gene Ontology[15; 16], a hierarchy developed to represent functional knowledge for proteins. A second set of relationships (binding, expression, etc.) is adapted from previous studies[41] and community tasks[28; 38]. Finally, we included our curated interactions from the training data described in Section 3.2.2 below.

The curated relationship types and their synonyms were considered valid if confirmed in multiple abstracts in PubMed. We also identified and explicitly added variants of an interaction instead of relying on stems, because we found that several verbs can be ambiguous in their stemmed forms. For example, the stem *sulfat* for the PTM sulfation will also match sulfate or sulfates to indicate the presence of chemical compounds. Finally, the interactions in the list were grouped into nine categories; each category name served as the canonical reference term for its associated set of interactions. Figure 3 shows the categories, along with some sample interactions for each; the current KnIT relationship hierarchy contains 71 interactions in total over all categories.

3.2.2 Training data for Event Detection

In order to learn how PPIs are presented in literature, the subject matter experts (SMEs) on the team labeled abstracts containing PTM interactions. Abstracts were collected using a PTM query "phosphorylate". To ensure that the abstracts also mentioned at least one protein, the query also included p53, a commonly studied protein. To further ensure that there were meaningful and diverse interactions present in the abstracts, an abstract was included in the query result only if the corresponding article was referenced by the manually curated PhosphoSitePlus[23] database. Finally, 400 abstracts were selected at random from the query result to use as training data. Of these abstracts, 60% contained mentions of phosphorylation and 40% contained mentions of other PTMs.

The information labeled in the abstracts included (1) verbs or biological interactions that could potentially apply to proteins (for potential insertion into the KnIT interaction catalog), and (2) the agent, theme and event, if present, in each sentence that contained a PPI. If negation was present, e.g., "ATM did *not* phosphorylate p53", then this feature was recorded as well.

Figure 2 shows examples of interactions collected for cataloguing and a sentence showing two PPIs labeled for Agent, Interaction and Theme instances.

3.2.3 Learning Patterns for Event Extraction

Using the sentences labeled for PPIs (see Section 3.2.2), KnIT learns rules over natural language parse trees for event detection. Each sentence is run through the IBM ESG Parser and the dependency parse structure is examined for syntactic connections between the labeled interaction, agents and themes. Here we show a few examples of rules learned.

(a) The dependency parse structure of a sentence "ATR phosphorylated p53 at Ser-15", reveals that the verb of interest, *phosphorylated*, is connected to a known agent and theme via *nsubj* and *dobj,* (noun-subject and direct-object) relationships. This connection is identified as a rule for semantic role labeling of agents and themes. In more detail, the parse structure is:

```
root(ROOT, phosphorylated)
nsubj(phosphorylated, ATR)
dobj(phosphorylated, p53)
prep_at(phosphorylated, Ser-15)
```

(Here `prep_at` means that the arguments are connected by the preposition "at".) For test sentences that express syntactically similar relationships between known interactions and protein entities, following the *nsubj* and *dobj* connections will result in identification of agents and themes for that interaction. KnIT always relies on the catalog for identifying interactions; the learned rules assign "agent" or "theme" semantic labels to protein entities.

(b) KnIT extracts complex rules where secondary interactions are involved. For example a sentence such as "ATM-mediated phosphorylation of p53 at serine 15." contains an indirect interaction, *mediated*, between the agent and the main interaction, *phosphorylation*, i.e., "ATM-mediated" is an adjective modifying "phosphorylation":

```
amod(phosphorylation, ATM-mediated)
prep_of(phosphorylation, p53)
```

A rule is learned to allow for any secondary interaction (from the KnIT relationship catalog) to act as a connector between an agent and the interaction.

(c) KnIT learns rules where agents and themes are not simply proteins but rather processes or events involving a protein. For example, in the sentence "Our data suggests that bcl-2 is able to modulate transmembrane trafficking of p53." the interaction *modulate* is connected to an entity *p53* via a process that is captured in the phrase *"transmembrane trafficking of p53"*.

Parse structures for every PPI in the labeled training data are examined in the above manner to collect a set of rules that describe how interactions and entities might relate to one another. Rules that have a support of at least three PPIs in the training data are selected for inclusion into KnIT.

Although the rules are learned over interactions that are PTMs, the various sentence structures involved are very similar to those for the other interactions in the KnIT catalog, so that the learned rules are applicable to a wide variety of interactions. Although the set of rules is not exhaustive and will not identify every possible PPI, the high quality of the rules yields precision and recall sufficient to create a network well suited to PPI prediction.

KnIT also learns two kinds of negation patterns. In the first pattern, negation is associated with the verb and is readily available in the dependency parse tree. For example in the sentence, "ATM did not phosphorylate p53" the dependency parse tree shows a connection between the interaction '*phosphorylate*' and negation word '*not*':

```
root(ROOT, phosphorylate)
neg(phosphorylate, not)
nsubj(phosphorylate, ATM)
aux(phosphorylate, did)
dobj(phosphorylate, p53)
```

More complex representations of negation such as an event "was absent", "rarely detected" need special treatment. The SMEs helped curate a list of expressions that are prevalent in PubMed in expressing negative interactions. As with event extraction, negations are predicted by learning rules that connect the known interaction and negation-indicating phrases.

3.2.4 Online Event Extraction

Here we discuss the end-to-end KnIT Extraction pipeline where the learned rules are applied to all 24 million abstracts in Medline, extracting protein entities and their interactions to construct PPI networks. Let us revisit Figure 1 for an overview.

1. KnIT ingests one abstract at a time and performs a sentence split on the title and abstract body. All further extractions in KnIT are at the sentence level.
2. KnIT applies the entity detection and normalization procedure, as described in Section 3.1, to identify all protein mentions and their canonical forms.
3. Every sentence is then checked for the presence of an interaction term from the KnIT relationship catalog (Section 3.2.1). Only sentences that contain an interaction and at least two protein entities proceed to the next step.
4. KnIT performs a natural language parse of the sentence using the IBM parser.
5. The parse tree is then examined using the rules learned between interactions and entities (Section 3.2.3) in order to assign semantic role labels to agent and theme entities.
6. Sentences with interactions, agents and themes are considered as valid triples for the network and are examined for the presence of negation attached to the interaction.

We applied the KnIT extraction engine to a Medline corpus of almost 24 million artifacts. All analysis was restricted to title and abstracts. We were able to extract a total of 171,798 unique triples from 245,833 text occurrences across 167,590 unique abstracts where the agents and themes matched normalized protein entities, i.e. there was no ambiguity in what these entities were.

Examples of interactions between ATM and p53 that can be extracted by KnIT using the learned rules are as follows.

*However, IR-activated, **ATM**-mediated **phosphorylation** of **p53** at serine 15 (human) or 18 (mouse) [Ser15(h)/18(m)], and apoptosis occurred in myoblasts but was impaired in myotubes.*

*Nuclear accumulation and **ATM**-dependent **phosphorylation** of **p53** on serine 15 were also observed.*

*Upon treatment with 5-Aza-CdR, **ATM** activation was clearly associated with **P53 phosphorylation** at Ser(15), which was directly responsible for 5-Aza-CdR modified P21(Waf1/Cip1) expression.*

*Moreover, knockdown of **ATM** by siRNA significantly **reduced** **p53** phosphorylation and stabilization..*

ATM** and ATR cannot **phosphorylate** **p53** on Ser-20. **(negation)

***ATM** **phosphorylates** the **p53 tumor suppressor** on a site (Ser15) that regulates transcription activity.*

4. GENERATION OF HYPOTHESES

After extracting (agent, event, theme) triples, KnIT then can use a variety of graph-reasoning techniques to predict heretofore unknown PPIs. There are many possible reasoning techniques[8; 29; 50; 57; 58]. To date, we have primarily investigated two different methods.

Non-negative matrix factorization (NMF): This technique can use the discovered triples directly to simultaneously predict many kinds of PPIs. Originally introduced by Paatero and Tapper in 1994[39], NMF can, given a matrix X, determine (two) smaller matrices that approximate X when multiplied together. If X has dimensions $m \times n$, NMF calculates smaller W and H matrices of

dimensions m x k and k x n, respectively, where the parameter k << min(m, n) determines the granularity of the approximation. In our case, $X_{i,j}$ would have a binary value indicating whether or not KnIT found a connection between proteins i and j, i.e., whether i and j appeared jointly in some triple. There are multiple possible objective functions available for determining "optimal" values W^* and H^* of W and H, but we follow Paatero and Tapper and simply minimize the Frobenius norm of the factorization error: $W^*, H^* = \text{argmin}_{W,H} \|X - WH\|_F$, where $\|A\|_F = \left(\sum_{i,j} a_{ij}^2\right)^{1/2}$.

Matrix factorization is essentially a compression algorithm, where row i of W, corresponding to protein i in the original interaction matrix X, is a small "meta-feature" vector representing the nature of the relationships that row i (protein i) of X has with the columns of X (the other protein), which each have their own meta-feature vector in the H matrix, as shown in Figure 5A. The dot product of the i^{th} row of W with the j^{th} column of H yields a "corrected" value for $X_{i,j}$, which may indicate a new connection if the original value was 0 and the corrected value is not. Berry et al. review many different regularization and parallelization methods[3] that appeared since the first publication, notably Lee and Seung's 1999 paper that rekindled the popularity of NMF [7; 18; 19; 26; 31; 32]. A range of NMF applications already exist in biology, such as for biological image processing[59] and microarray data analysis. The latter produces meta-gene expression levels for each sample, which can then be used for clustering and classification[53]. Applying NMF to PPI networks, we expect it to determine latent variables that indicate the types of interactions each protein takes part in.

BOW-GID: This technique was proposed in Spangler et al.[46] to which we refer the reader for details. Briefly, an event and theme are fixed, and the goal is to find previously undiscovered agents. KnIT constructs a similarity network of proteins based on a bag-of-words (BOW) analysis[44]. Each protein that is a known agent is labeled with a weight of 1 and all other proteins are labeled with a weight of 0. KnIT then uses a graph information diffusion[33] (GID) technique to propagate the weights from known agents to the unknown agents. Candidates with the highest diffused weights are considered to be the most likely to be undiscovered agents.

5. COMPUTATIONAL EXPERIMENTS

In this section, we describe several experiments designed to test the scope and accuracy of the KnIT discovery pipeline. We first describe a retrospective study using the NMF technique, then test KnIT's ability to discover PPIs that are found in sources other than the biomedical literature, and finally show how we can automate, scale up, and generalize the BOW-GID study in Spangler, et al.[46]

5.1 Retrospective Predictions

Our retrospective study was designed to test whether KnIT could predict additional connections that are absent from the network because they are not documented in the literature, could not be parsed, or remain undiscovered. We applied NMF in order to predict, using papers published before a given date, PPIs discovered in papers after that date. The results provide a consistent internal measure of data quality and a realistic assessment of the likelihood KnIT can accurately predict future discoveries. Subgraphs were constructed for each interaction type. The most recent ~20% of edges was removed and used as a test set while the earlier 80% formed the training set for the factorization model. Eight separate datasets, one for each of the most common interaction types, was used, and their results are shown in Figure 5B. The areas under the ROC curves show that each subnetwork was significantly more accurate than chance.

Table 1 – Statistics on the networks created based on each of the 8 most frequently observed relationship types. E/N is the edge to node ratio.

RelationshipType	Edges	Nodes	E/N
RegulationPositive	54278	8296	6.5
RegulationNegative	44091	7812	5.6
Regulation	28717	7063	4.1
Phosphorylation	20991	4280	4.9
Binding	10832	5284	2.1
Expression	5918	2459	2.4
Localization	1335	1526	0.9
Ubiquitination	1155	909	1.3

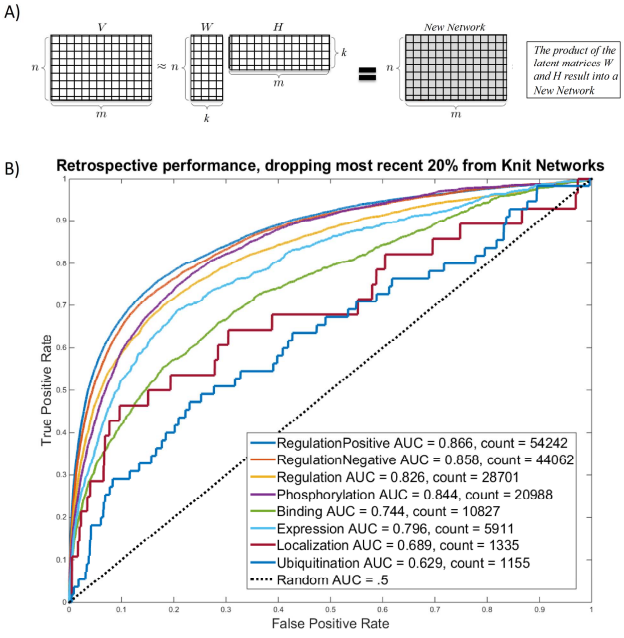

Figure 5. Hypothesis generation on KnIT network. A) Nonnegative matrix factorization was applied for B) retrospective validation in the KnIT network

Notably, the models from networks with more edges did better, suggesting that ultimate performance depends on the richness of the information available in the network to start with. These data demonstrate that the automated KnIT extraction and prediction pipeline can identify new hypotheses that prove correct, i.e., foretell future observations on a large scale in time-stamped experiments designed to mimic realistic conditions.

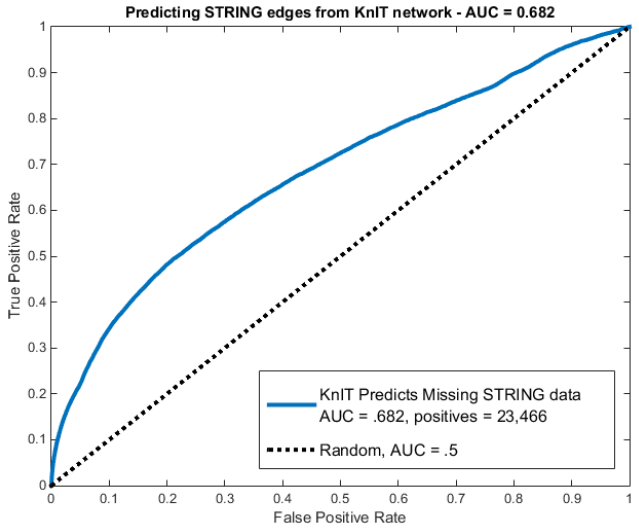

Figure 6. Prospective analysis where the positive set are interactions currently found in StringDB.

5.2 External Validation

In order to examine our ability to mine text for relationships, we compared the state of the KnIT extracted network to StringDB [13]. StringDB is a long-running effort to gather and combine a multitude of information sources on connections between proteins. Particularly useful is its sub database specific to protein-protein interactions with experimental evidence. The data sources, such as BioGRID, BIND, and DIP [13], manually collect information extracted from multiple experimental techniques (yeast two-hybrid experiments, mass spectrometry, genetic interactions, phage display, etc.). Because this information is often taken from large-scale screens, the interactions are unavailable to a literature-mining-only approach. Despite this, we would expect KnIT relationships to have significant correlation with StringDB.

KnIT's extracted network contains 118,864 edges and 11,375 nodes, while StringDB contains 136,734 edges and 13,893 nodes. Although each node represents a protein, not all proteins occur in the underlying data sources, and so only 8,994 nodes are shared between them. Within this shared set of nodes, the KnIT extracted network contains 107,388 edges while StringDB contains 85,954. Had the KnIT extracted network been composed of random edges in this space, only 260 edges would be expected to be shared, but instead, 10,279 edges are in common ($p<0.0001$). This shows that KnIT is significantly aligned with existing structured data, while also finding over 100,000 relationships that were not previously documented by StringDB.

Next, we will show that the information appearing in StringDB that is not detected by KnIT extracted network could potentially be predicted using the factorization method used in section 5.1. To do so, we apply factorization to the KnIT network to predict interactions unique to StringDB. The positive test set was composed of edges that are in StringDB but not detected by our KnIT reading on the literature (75675 interactions), while the negative set is all non-existent edges connecting pairs of proteins that do not have a connection in either network. The NMF analysis of Section 5.1 was repeated, but using the entire KnIT network to train the NMF model. Predicting these edges, the model had a performance, as measured by AU(ROC) (Figure 6) of 0.68. These data show that edge-prediction applied to KnIT relationships can discover true relationships absent from the literature with a significant enrichment over what would be expected by chance.

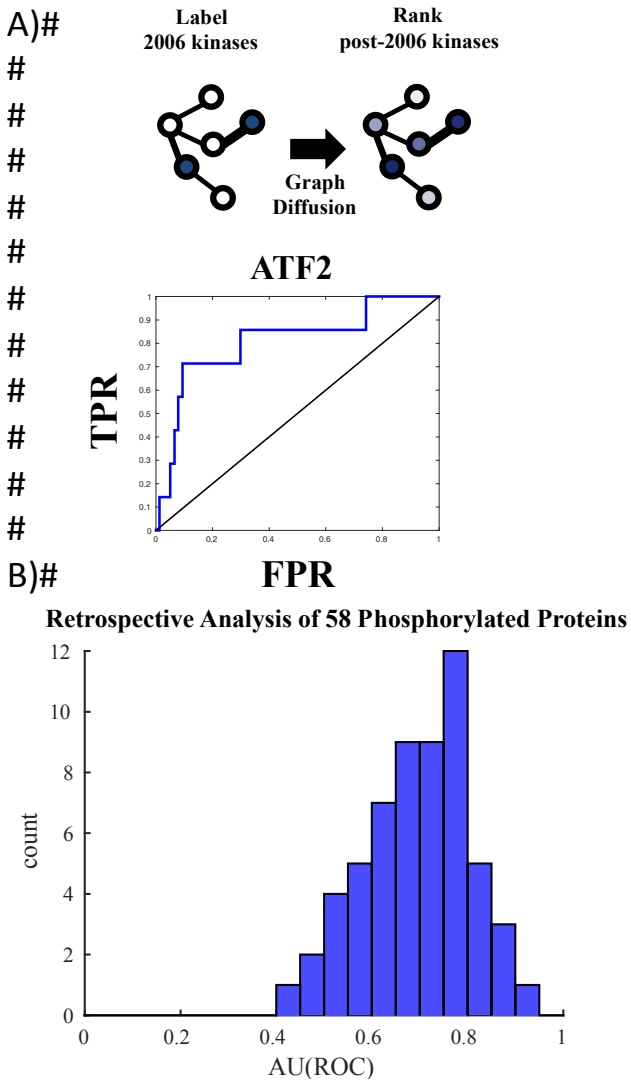

Figure 7 A) Retrospective analysis for kinases phosphorylating B) 58 specific protein targets.

5.3 Automating/Expanding a Prior Analysis

In our final experiment, we tested whether the fully automated KnIT system could eliminate the need for manual, expert-based annotation. Specifically, in a prior study[46], the BOW-GID technique was used to predict p53 kinases, proteins that add phosphate molecules to the tumor-suppressor protein p53. In that retrospective study, abstracts available in 2003 were used to predict subsequent discoveries in the literature. Remarkably, the system identified 7 out of 9 subsequently discovered p53 kinases. This proof-of-principle experiment demonstrated the prediction of novel facts based on the literature, but still required SMEs to manually identify kinases in the abstracts, and in particular to identify kinases that phosphorylated p53 (so that they could receive an initial weight of 1 in the graph diffusion analysis). The goals of our current experiment are (1) to see whether the prior, painstaking manual annotation by experts can be replaced by the new automated triple-extraction technology, thereby allowing KnIT to handle much larger corpora, and (2) to show the potential

of such scaling by expanding the prior study to identify kinases that target proteins in general, not just p53.

In more detail, we asked whether the kinases known up to a particular year for a given protein could predict kinases discovered after that year for that same protein, using the (now automated) BOW-GID method. We selected a set of 58 phosphorylated proteins with at least 4 kinases known in 2006 and 4 kinases discovered after 2006. We created a BOW similarity network of 332 kinases based on the literature through 2006, used the triples discovered by KnIT to assign a weight of 1 to known kinases, and then ranked the unknown kinases by their post-diffusion score. Figure 7 shows that 68% of the protein test sets achieved an AU(ROC) > 0.6, and 56% an AU(ROC) > 0.7. The p53-specific result was AU(ROC) = 0.71 compared to AU(ROC) = 0.84 in the prior study, reflecting both the larger search space and the use of an automatically generated initial labeled network rather than one created manually by a human expert. This study shows that, using KnIT in 2006, we automatically modeled the kinase-kinase network and predicted a number of novel kinases found later in publication. Thus, KnIT is effective for predicting novel connections on a large scale without manual intervention.

6. DISCUSSION AND CONCLUSION

This study presents a complete end-to-end discovery pipeline that transforms unstructured data from the biological literature into new scientific hypotheses, significantly enriched for truth. In the process, we show that we can automatically extract relationships from text. These relationship triples are pooled into networks that are reasoned over by diffusing weights or by generating novel edges via matrix completion. Time-stamped studies rigorously assessed the predictive power to foretell discoveries made after a given date, based on what was known prior to that date. We also tested predictions against controls that do not appear in the literature. Finally, we applied KnIT knowledge extraction via triples to obviate the need for manual annotation of abstracts by SMEs when automatically reasoning via BOW-GID. These results point to the quality and predictive power of both the triples and of the network-based reasoning algorithms that we apply to them. KnIT has shown good performance in prospective analysis and, in practice, already suggests thousands of hypotheses that have the potential to efficiently guide biological discoveries.

Baylor College of Medicine and IBM Research are working together to accelerate scientific discoveries, with the long-term goal of taking on the challenge of personalized medicine. In order to understand individual genetic variation, we must be able to contextualize mutations with detailed annotation of genes and proteins. However, to make this practical on a large-scale, data must be structured, and reasoning automated. This study represents the early stages of our effort. In the future, our team will focus on a wider area of biological entities, such as drugs, treatments, and diseases, in order to develop an all-inclusive network of biomedical literature knowledge. The next steps will include a collection of algorithms validated with biological experiments to identify new hypotheses based on the entire network. We will also develop techniques to prioritize the large number of hypotheses generated based on likelihood, value and risk. Such analyses will yield a set of hypotheses that will fill gaps in science, accelerate the pace of discovery and, in turn, improve our understanding of disease to better guide patient care.

7. ACKNOWLEDGMENTS

This work was supported by the Robert and Janice McNair Foundation, DARPA (N66001-14-1-4027), National Science Foundation (NSF DBI-1356569, NSF DBI-0851393), National Institutes of Health (NIH-GM079656, NIH-GM066099), and was supported in part by the IBM Accelerated Discovery Lab.

8. REFERENCES

[1] Aronson, A.R. and Lang, F.M., 2010. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc 17*, 3 (May-Jun), 229-236.

[2] Ashburner, M., et al., 2000. Gene ontology: tool for the unification of biology. *Nat Genet 25*, 1 (May), 25-29.

[3] Berry, M.W., et al., 2007. Algorithms and applications for approximate nonnegative matrix factorization. *Comp Statistics & Data Analysis 52*, 1, 155-173.

[4] Brohee, S. and Van Helden, J., 2006. Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics 7*, 1, 488.

[5] Cancer Genome Atlas, N., 2012. Comprehensive molecular characterization of human colon and rectal cancer. *Nature 487*, 7407 (Jul 19), 330-337.

[6] Cancer Genome Atlas Research, N., 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature 455*, 7216 (Oct 23), 1061-1068.

[7] Catral, M., et al., 2004. On reduced rank nonnegative matrix factorization for symmetric nonnegative matrices. *Linear Algebra and its Applications 393*, 107-126.

[8] Chinnasamy, A., et al., 2006. Probabilistic prediction of protein-protein interactions from the protein sequences. *Comput Biol Med 36*, 10 (Oct), 1143-1154.

[9] Cohen, A.M. and Hersh, W.R., 2005. A survey of current work in biomedical text mining. *Brief Bioinform 6*, 1 (Mar), 57-71.

[10] Danger, R., et al., 2014. Towards a Protein–Protein Interaction information extraction system: Recognizing named entities. *Knowledge-Based Systems 57*, 104-118.

[11] Davis, A.P., et al., 2011. The Comparative Toxicogenomics Database: update 2011. *Nucleic Acids Res 39*, Database issue (Jan), D1067-1072.

[12] Edwards, A.M., et al., 2002. Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends Genet 18*, 10 (Oct), 529-536.

[13] Franceschini, A., et al., 2013. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res 41*, (Jan), D808-815.

[14] Fundel, K., et al., 2007. RelEx—Relation extraction using dependency parse trees. *Bioinformatics 23*, 3, 365-371.

[15] Gene Ontology, C., 2008. The Gene Ontology project in 2008. *Nucleic Acids Res 36*, Database issue (Jan), D440-444.

[16] Gene Ontology, C., 2010. The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res 38*, Database issue (Jan), D331-335.

[17] Gray, K.A., et al., 2013. Genenames.org: the HGNC resources. *Nucleic Acids Res 41*, (Jan), D545-552.

[18] Guillamet, D., et al., 2001. A weighted non-negative matrix factorization for local representations IEEE, I-942-I-947 vol. 941.

[19] Hamza, A.B. and Brady, D.J., 2006. Reconstruction of reflectance spectra using robust nonnegative matrix factorization. *IEEE Transactions on Signal Processing 54*, 9, 3637-3642.

[20] Hatzivassiloglou, V., et al., 2001. Disambiguating proteins, genes, and RNA in text: a machine learning approach. *Bioinformatics 17 Suppl 1*, suppl 1, S97-106.

[21] Heath, L.S. and Sioson, A.A., 2009. Multimodal networks: structure and operations. *IEEE/ACM Trans Comput Biol Bioinform 6*, 2 (Apr-Jun), 321-332.

[22] Hoffmann, R. and Valencia, A., 2004. A gene network for navigating the literature. *Nat Genet 36*, 7 (Jul), 664.

[23] Hornbeck, P.V., et al., 2015. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res 43*, Database issue (Jan), D512-520.

[24] International Human Genome Sequencing, C., 2004. Finishing the euchromatic sequence of the human genome. *Nature 431*, 7011 (Oct 21), 931-945.

[25] Jenssen, T.K., et al., 2001. A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet 28*, 1 (May), 21-28.

[26] Jia, Y.W.Y. and Turk, C.H.M. ,2004 Fisher non-negative matrix factorization for learning local features. *In Proc. Asian Conf. on Comp. Vision*, 27-30.

[27] Jupe, S., et al., 2012. Reactome - a curated knowledgebase of biological pathways: megakaryocytes and platelets. *Journal of Thrombosis and Haemostasis : JTH,* 10(11), 2399–2402.

[28] Kim, J.D., et al., 2011. Overview of BioNLP shared task 2011 *Association for Computational Linguistics*, 1-6.

[29] Kuchaiev, O., et al., 2009. Geometric de-noising of protein-protein interaction networks. *PLoS Comput Biol 5*, 8 (Aug), e1000454.

[30] Laura, C., et al., 2010. SystemT: an algebraic approach to declarative information extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, 128-137.

[31] Lee, D.D. and Seung, H.S., 1999. Learning the parts of objects by non-negative matrix factorization. *Nature 401*, 6755 (Oct 21), 788-791.

[32] Lee, D.D. and Seung, H.S., 2000. Algorithms for Non-negative Matrix Factorization. In *NIPS*, 556-562.

[33] Lisewski, A.M. and Lichtarge, O., 2010. Untangling complex networks: risk minimization in financial markets through accessible spin glass ground states. *Physica A 389*, 16 (Aug 15), 3250-3253.

[34] Manning, C.D., et al., 2008. *Introduction to Information Retrieval*. Cambridge University Press Cambridge.

[35] Manning, C.D. and Schütze, H., 1999. *Foundations of statistical natural language processing*. MIT press.

[36] Mccord, M.C. and Bernth, A., 2010. Using slot grammar. *IBM TJ Watson Res. Center, Yorktown Heights, NY, IBM Res. Rep. RC23978.*

[37] Miyao, Y., et al., 2009. Evaluating contributions of natural language parsers to protein-protein interaction extraction. *Bioinformatics 25*, 3 (Feb 1), 394-400.

[38] Nédellec, C., et al., 2013. Overview of BioNLP shared task 2013. *Proceedings of the BioNLP Shared Task 2013 Workshop*, 1–7.

[39] Paatero, P. and Tapper, U., 1994. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics 5*, 2, 111-126.

[40] Poon, H., et al., 2014. Literome: PubMed-scale genomic knowledge base in the cloud. *Bioinformatics 30*, 19 (Oct), 2840-2842.

[41] Pyysalo, S., et al., 2007. BioInfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics 8*, 50.

[42] Quan, C., et al., 2014. An unsupervised text mining method for relation extraction from biomedical literature. *PLoS One 9*, 7, e102039.

[43] Rzhetsky, A., et al., 2004. GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *J Biomed Inform 37*, 1 (Feb), 43-53.

[44] Salton, G. and Mcgill, M.J., 1986. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc.

[45] Scott, S., et al., 2014. Automated hypothesis generation based on mining scientific literature. In *Proceedings of the 20th ACM SIGKDD*, New York, New York, USA, 1877-1886.

[46] Scott, S., et al., 2014. Automated hypothesis generation based on mining scientific literature. In *Proceedings of the 20th ACM SIGKDD*, New York, USA, 1877-1886.

[47] Stumpf, M.P., et al., 2008. Estimating the size of the human interactome. *Proc Natl Acad Sci U S A 105*, 19 (May 13), 6959-6964.

[48] Tanabe, L. and Wilbur, W.J., 2002. Tagging gene and protein names in biomedical text. *Bioinformatics 18*, 8 (Aug), 1124-1132.

[49] Tikk, D., et al., 2010. A comprehensive benchmark of kernel methods to extract protein-protein interactions from literature. *PLoS Comput Biol 6*, 7, e1000837.

[50] Tuncbag, N., et al., 2011. Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM. *Nat Protoc 6*, 9 (Sep), 1341-1354.

[51] Uniprot, C., 2013. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res 41*, Database issue (Jan), D43-47.

[52] Van Landeghem, S., et al., 2013. Large-scale event extraction from literature with multi-level gene normalization. *PLoS One 8*, 4, e55814.

[53] Wang, H., et al., 2013. Predicting protein-protein interactions from multimodal biological data sources via nonnegative matrix tri-factorization. *J Comput Biol 20*, 4 (Apr), 344-358.

[54] Wheeler, D.L., et al., 2003. Database resources of the National Center for Biotechnology. *Nucleic Acids Res 31*, 1 (Jan 1), 28-33.

[55] Wishart, D.S., et al., 2009. HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res 37*, Database issue (Jan), D603-610.

[56] Xie, Z., et al., 2010. hPDI: a database of experimental human protein-DNA interactions. *Bioinformatics 26*, 2 (Jan 15), 287-289.

[57] You, Z.H., et al., 2013. Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis. *BMC bioinformatics 14 Suppl 8*, S10.

[58] Zhang, Q.C., et al., 2012. Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature 490*, 7421 (Oct 25), 556-560.

[59] Zitnik, M., et al., 2013. Discovering disease-disease associations by fusing systems-level molecular data. *Sci Rep 3*, 3202.