Biological network evaluation and relation discovery from scientific literature

Chen Li

European Bioinformatics Institute

European Molecular Biology Laboratory

This thesis is submitted to the University of Cambridge

for the degree of *Doctor of Philosophy*

Fitzwilliam College

March, 2014

Chen Li: *Biological network evaluation and relation discovery from scientific literature*, A thesis submitted to the University of Cambridge for the degree of Doctor of Philosophy, ©September 29, 2014.

Supervisor:
Dr. Dietrich REBHOLZ-SCHUHMANN

Thesis Advisory Committee:
Dr. Stephen CLARK, University of Cambridge
Dr. Julio SAEZ-RODRIGUEZ, European Bioinformatics Institute
Dr. Reinhard SCHNEIDER, European Molecular Biology Laboratory

Cambridge, September 29, 2014

This thesis is dedicated to my parents, Xiangrong and my Xixi, who make me a very happy man.

Declaration

This thesis is my own work and includes nothing, which is the outcome of work done in collaboration except as specified in the text. It is not substantially the same as any I have submitted for a degree, diploma or other qualification at any other university; and no part has already been, or is currently being submitted for any degree, diploma or other qualification.

This thesis does not exceed the specified length limit of 300 single-sided pages of double-spaced text as defined by The Biology Degree Committee.

Chen Li
March, 2014

Acknowledgements

First of all, I would like to thank my supervisor, Dietrich Rebholz-Schuhmann, for his support, insight and inspirations. I am lucky to join his group, and received his encouragement.

I want to thank Maria Liakata for sharing her knowledge and collaborating on the projects. Her enthusiasm for the research in biomedical text mining inspires me. I would like to thank Antonio Jimeno-Yepes, who gave me a lot of help at the beginning of the study. I thank Andreas Vlachos, who kindly shared his experience and knowledge about bio-event extraction. I also thank Nigel Collier, for the suggestions during many discussions.

I want to thank my thesis committee, which is kindly attended by Stephen Clark, Julio Saez-Rodriguez and Reinhard Schneider. Especially, I am grateful to Julio, for his help and support. Meanwhile, I thank all the members in his group, which I have been affiliated with in the last two years of my PhD.

I would like to thank all the past members of text mining group in European Bioinformatics Institute, including Christoph for his help in software implementation, Shyamasree for the help in Whatizit indexing, and Ian for database evaluation. I also want to thank Nicolas Le Novère, Stuart Edelstein and other members in the past Compneur group I have worked in.

I have a great family. I cannot thank enough for my aunt's support. I also want to thank my cousin, Lu Li, and Camille for the warm family.

Last but not least, I want to thank my friends in EBI: Leonor, Mar, Rita, Sergio, Stephan, and Weizhong for their friendship.

Biological network evaluation and relation discovery from scientific literature

Chen Li

Even a simple biological phenomenon may introduce a complex network of molecular interactions. A number of manually curated databases store biological networks in various formats. Scientific literature is one of the trustful resources delivering knowledge of these networks. Evaluating biological network with collected evidential statements from the scientific literature aligns the two types of resources and supports knowledge discovery. Hidden molecular relation, which may be pathogenetic but have not been reported due to various reasons, can be discovered or inferred based on evidences collected from scientific literature. Meanwhile, for NLP-based TM, the evaluation delivers extensive knowledge about distribution of entities and reactions in scientific literature.

Biological network evaluation involves several layers of information. The identification and normalisation of biomedical entities is of high importance. The research work as part of the CALBC challenges, benchmarks different biological named entity recognition (NER) solutions leading to the result that a lexical approach in combination with disambiguation solutions trained on gold standard corpora provides state of the results. It also explores methods to automatically harmonise effort of different solutions to generate very large corpus (about one million abstracts).

Event extraction links individual entities with molecular interactions and the main method of collecting evidential statements. We develop PCorral, which assist bio-curators to gain a comprehensive understanding about the entities and the reaction without technological bias, and narrow down selected documents to focus on specific reaction. Later, I develop a precise method for evidential statement collection and hidden relation discovery. The system, named LitWay, is capable of identifying molecular interactions from free-text and flexible for identifying more types of complex reactions, including protein-protein interaction and protein-chemical interaction. The evaluation on BioNLP'13 data proves the system's performance to be state-of-the-art. It can be easily customised for different tasks.

By utilising the methods I developed, I evaluate the curated networks in BioModels Database. The evaluation quantitatively profiles the information of composition and morphology in the scientific literature for diverse networks including metabolic pathway and signalling pathway. The signalling pathways have a better coverage in the scientific literature. The evaluation is also extended to a comprehensive human metabolic network, RECON2, to explore the way of discovering hidden relations.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Key points

- Biological networks are the focus of on-going research for exploring causes of diseases. Increasingly, different types of biological networks are intended to be jointly investigated for reconciling complicated biological systems.

- Aligning curated biological network with knowledge extracted by TM quantitatively characterise existence and morphology of different types of biological networks.

- Curated biological networks can be semantically enriched by their contents being aligned against evidential statements collected from the scientific literature.

- There are six phases towards extracting complete information of networks. Each phase provides a different layer of information for network evaluation.

- The last section of this chapter outlines the structure of the thesis and the contributions of the work.

Biomedical text mining (TM) endeavours to extract unstructured textual information into structured knowledge about biological processes. Then the knowledge can be stored in semantic resources or knowledge-base for easy re-use in the future. Current TM has been pacing toward the goal and mainly focuses on extracting bio-events, which cover various biochemical reactions. However, the gap between structured knowledge of biological network and unstructured texts is still significant. There is no work to systematically align contents and knowledge from two types of data at different levels (entity level, relation level and network level).

The thesis describes an effort of using TM methods to align knowledge of biological networks from two types of data sources. Besides presenting the new TM solutions, the alignments evaluate the TM solutions, analyse the issues of network extraction, and investigate molecular relations in the biological networks. The first type of data sources are structured and curated databases. The data has been checked by curators and fully or partially cross-referenced with semantic resources. Due to high cost of curation, the available data is less than that from the second type of data sources, the scientific literature. The scientific literature is in unstructured text, in which biological networks are described in natural language. Abundant knowledge about biological networks is available from it. However, retrieving them require further improving currently available TM solutions.

The alignment is bidirectional. It could evaluate the TM solutions against the curated biological networks, which helps to identify the issues of the solutions for extracting biological networks. It helps TM researchers to morphologically and quantitatively characterise different types of biological networks in the scientific literature. Conversely, aligning contents in curated databases could link publications or smaller units, e.g. sentences, with corresponding network elements in databases to semantically enrich networks. The alignment results are also valuable for studying molecular activities and detecting unreported relations.

The first chapter will start dissecting the problem by elaborating relationships between biological network and TM, especially based on the categorisation of different types of biological networks. Based on the analysis, it defines the rationality of the work.

1.1 Data for biological networks from different resources

Underneath biological phenomena, there are sophisticated networks of molecular interactions (MI) completing various functions. A biological network is represented in a graph structure composed of nodes that denote biomolecules and edges between the nodes representing the interactions or reactions between the biomolecules. Network representations serve many purposes in bioinformatics, and most importantly, networks are used to judge the functional behaviour of interaction networks on the molecular level. Network representations are used to simulate, analyse and visualise the responses of protein interaction networks, metabolic pathways, specific synapses and even whole systems such as an organ, e.g. the liver or the brain. Studying the topological structure and the functional responses of such systems aims to reveal yet undiscovered mechanisms that could explain or improve specific unfavourable health conditions. In particular, the networks of signalling and metabolic pathways are at the focus of on-going research to explore causes of diseases, and increasingly these different types are investigated jointly for reconciling the outcomes from regulatory or metabolic mechanisms. These developments are a part of the research in Systems Biology, which aims to build large-scale networks of complete living systems.

Following the increasing of interests in biomedical networks and their availability in electronic form, many public repositories have been created for hosting data of bio-networks [1–3]. Electronic representations of networks can serve several purposes: they do not only visualise biological systems but also help to interpret experimental data, to predict reactions between entities, or to infer new gene functions in functional genomics [4].

Mathematical models of biological networks are widely used for simulating processes by combining network topology with quantitative information of reactions between entities. Qualitative models are usually encoded in some logical formalism such as Prolog, while quantitative models are encoded in the standards such as the Systems Biology Markup Language (SBML) [5], CellML [6] and BioPAX [7]. Quantitative models are available in some special repositories [8, 9]. Data population of the repositories, especially open platforms like WikiPathways [10] and BioModels Database [8, 11], are community driven efforts, since encoding of models requires significant contributions from domain experts. Creation of encoded models and later processes of curation and annotation are all extremely time-consuming in order to meet certain quality and reusability. Artificial intelligence technologies of utilising TM have great potential to assist the better and faster processes of creation, curation and annotation [12].

1.1.1 Biomedical networks from the scientific literature

The latest developments in natural language processing (NLP) and biomedical semantics open a way of supporting manual curation, semi-automated curation and semantic enrichment of biological networks [13]. Advances in text mining (TM) show great potential to assist on-going studies of biological networks, ranging from molecular biology [14, 15] to systems biology [16]. In this thesis, analyses of complete biological networks are distinguished from those related to individual biological events or a sets of events. I emphasise the phases that TM has to fulfil to be able to evaluate different types of networks, such as signalling pathways versus metabolic pathways.

Networks are represented as graphs, where entities constitute the nodes. Therefore missions of knowledge retrieval and discovery of biological network, such as network evaluation, have to begin with named entity recognition (NER). Specifically, in network evaluation, NER

- finds participating elements in biomedical pathways, i.e. genes / proteins, chemical entities;
- discovers functions and processes that could be impaired by a specific entity;
- retrieves information denoting molecular properties;
- maps entities to database entries (normalisation).

Relation extraction from the scientific literature attaches another layer of information on top of recognised entities. The information includes

- entity combinations involved in each reaction;
- relation types;

Current TM technologies are not yet fully integrated for gathering evidences from literatures, for example, identifying causative effects of gene-disease associations, which could be a modification of a molecular function of a protein or dysfunction of an anatomical component. No work has been done to align biological networks, such as signalling pathway or metabolic pathway, with knowledge mined from the scientific literature.

The work will require processing and analysing existing data repositories to compare contents against the scientific literature. Such systematic analysis will quantitatively reveal coverage and morphology of pathways in the scientific literature. Furthermore, the TM scientific community has made significant effort in automated identification of protein-protein interactions, but have not investigated the syntactic morphology of molecular relations, like those among proteins/genes and/or chemical entities. The analysis will also help to syntactically distinguish molecular relations between different types of entities, for example, trigger words for protein relation extraction may differ from that of chemical entities. The result may lead to developing novel extraction methods.

In this thesis, based on experiment and statistics, I analyse the aspects that TM has to fulfil to be able to evaluate different types of networks, such as signalling pathways versus metabolic pathways. I present the solutions I developed and discuss how I exploited TM methods to collect evidential statements from the scientific literature to investigate molecular functions and evaluate biological networks, including signalling pathway and metabolic pathway. The study also looks into a comprehensive metabolic network and investigates molecular functions based on evidential statements collected from the scientific literature. The evaluation of biological networks aligns structured knowledge in curated databases with unstructured text. The alignment supports understanding of functional information of entities, and identifying causative entity or reaction in biological phenomenon or disease. Hidden molecular relation, which may be pathogenetic but have not been reported due to various reasons, can be discovered or inferred based on evidences collected from scientific literature.

1.2 Biological network and text mining

First of all, it is necessary to understand the status quo of TM supported network extraction, as different stage of network extraction offers information for different levels of networks. The term ‘biological network’ may be used to refer to various aggregations of relations between biomolecules, diseases or phenotypes etc. In this study, I focus on the biological networks, which are directed graphs consisting of consecutive interactions between biomolecules, such as proteins, genes and chemical substances. This definition includes signalling pathways, metabolic pathways and regulatory networks.

In the era of ‘omics’ data, a desired outcome of many interdisciplinary studies in biology is combinations of semantic resources and high throughput data to produce reliable biological networks. Progress in the field of biological TM supports automatic extraction of such networks from the scientific literature [17–21]. Network extraction is complex and the implementation of a given system may vary (e.g. entity normalisation may happen after relation/event extraction). I distinguish six phases in a typical process as illustrated in a pyramid structure of Fig. 1.1. While the tasks at the top of the pyramid are very challenging, they can deliver information for the simulation of complex biological systems.

- Firstly, biological entities and concepts are tagged in the text; this corresponds to *Entity mentions*.

- During *Entity normalisation*, entity mentions are disambiguated and linked with entries in biological databases or ontologies.

- *Relation extraction* aims to detect biological relationships between any recognised entities (e.g. ‘inhibit’, ‘bind’ etc).

- *Event extraction* identifies the directionality[1] and the polarity[2] of biological interactions/reactions.
- *Network extraction* interconnects extracted events in order to build complete networks depicting biological phenomena, such as the initiation of cell growth as the response to an extracellular signal from ligand binding. Such a network can be modelled as a set of MIs that induce binary signals (on/off) and can be encoded in a logical model.
- *Quantitative modelling* extracts quantitative parameters of biological networks. It enhances networks with kinetic details. The parameters could be mass, concentration of reactants and products, or reaction time etc. They are useful for simulating biological phenomena to gain a better insight about mechanisms of related biological systems.

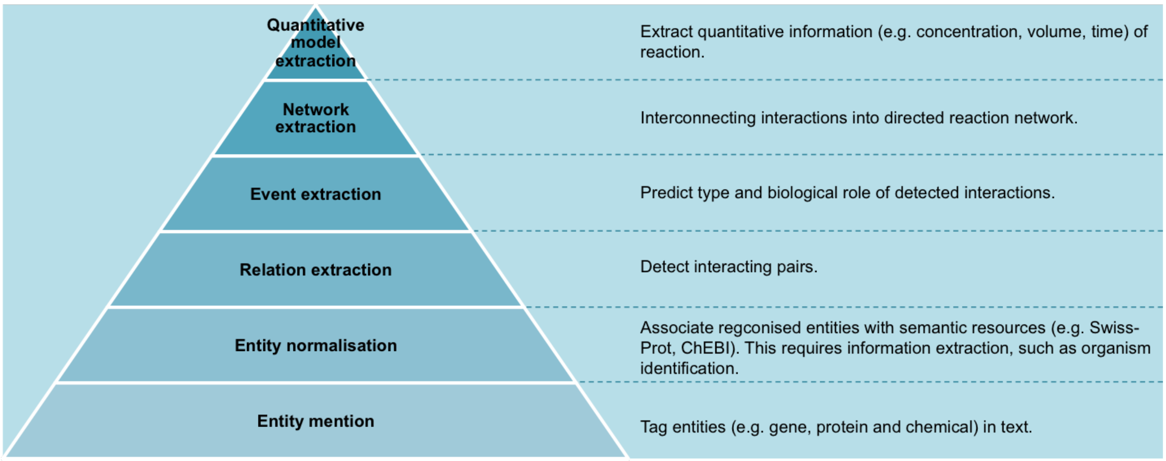

Quantitative model extraction
Extract quantitative information (e.g. concentration, volume, time) of reaction.
Network extraction
Interconnecting interactions into directed reaction network.
Event extraction
Predict type and biological role of detected interactions.
Relation extraction
Detect interacting pairs.
Entity normalisation
Associate regconised entities with semantic resources (e.g. Swiss-Prot, ChEBI). This requires information extraction, such as organism identification.
Entity mention
Tag entities (e.g. gene, protein and chemical) in text.

Fig. 1.1: Six phases of TM for network extraction. As advancing from identifying entities to creating a quantitative model of a reaction network, the TM technology required becomes increasingly more complicated. But it is closer to its initial mission, assisting the analysis of biological systems. Each step relies on the lower level process, and even the further low level, e.g. quantitative model extraction relies on network extraction and entity recognition etc. Relation and event extraction could happen without entity normalisation but the latter is necessary for event interconnection to allow network extraction. Current TM research has reached event extraction and had attempts in network extraction.

The technological maturity at each phase differs significantly. Current TM research has focused on the first four steps of the pyramid with attempts at network extraction. NER is well established as a mature technology in the TM research domain. Selected entity mention solutions recognise entities at a minimum of 80% balanced F-score and even better [22], although the quality of the gold standard corpus (GSC) can lead to a bias in the evaluation and

[1] i.e. which are the agents and which are the targets

[2] i.e. the positive or negative effect of a biochemical reaction. e.g. up-regulation and down-regulation.

cross-corpus [23] or cross-tagger evaluation can lead to variable results. There is also great differentiation between recognition performance for different types of entities, with genes and proteins having the highest recognition rates. Relation extraction and event extraction are currently the TM tasks that receive the most attention in the biological TM research community.

Organised shared tasks and challenges addressing these tasks [15, 24] have gained a lot of prominence in the research community. The corpora and the performances set standards for research in this area. However, the focus in the challenges resides solely with the identification of relations and events between proteins as opposed to relations involving other molecules, such as chemical entities. Meanwhile, researchers have stepped up their efforts on automated [17] or semi-automated [18] network extraction, and, at the same time, very little effort has been spent on extracting quantitative models from the literatures. KiPar [25], KID [26] and KIND [27] are rule-based systems for extracting quantitative information from text. Nonetheless, adding quantitative information to biological networks is complex and challenging as this information is often not available in text but hidden in graphs and tables. BioNLP'13 [28] showed promise in that it is the first time systems were asked to tackle pathway curation as a task, thus investigating current TM capability for automatic pathway extraction.

On occasions, the term *relation* is synonymously used for *event*s. Ananiadou et al. [16] distinguishes a relation from an event in the following way. A relation expresses the existence of an interaction between a pair of entities. For example, in "calcium ions penetrate the site and trigger VAMP, syntaxin and SNAP-25 to bind together in a lock and key fashion" [29], each binary binding among "VAMP", "syntaxin" and "SNAP-25" is the relation between a pair, whereas the formation of the complex composed of three bindings is a biological event. Therefore, an event represents a functionally complete behaviour of biomolecules. Extracting an event requires identification of more complex information than a relation. Essentially, an event have attributes of directionality and polarity, which are extra characteristics than relations have, for example, directionality of protein binding and polarity of regulation. It also should include information about the outcome of a reaction (see Section 4.1.1). To characterise an event fully, one needs to be able to capture various aspects of scientific discourses, such as whether an event is hypothesised or an outcome of a study, or certainty level of the event [30, 31]. In this work, analyses concerning complete biological networks are distinguished from those related to individual biological events or a set of events only.

Even without considering quantitative information, the concept of networks is greater than that of a bag of events. Disregarding the redundancy issue between events, sequence of reactions/events plays an important role in deciding molecular roles within a regulatory mechanism under specific conditions. Under certain circumstances, a network may be topologically and quantitatively dynamic. For example, when other cellular gradients influence a metabolic pathway, the pathway may change over time. These challenges make network extraction conceptually and technologically much more complicated than extracting a bag of events (see Section 4.1.1). However, even obtaining a bag of unique events where coreference between entities and entity mentions have been resolved is far from a trivial task.

1.3 Different types of biological networks

Signalling pathways and metabolic pathways pose different challenges to TM event extraction, since signalling pathways consist of binary protein-protein interactions (PPI) whereas metabolic pathways consist of chemical-protein interactions (CPI). Both networks are often studied independently, although they might be coupled in order to model similar processes at different levels of a functional hierarchy, e.g. insulin signal transduction pathway and regulation of blood glucose.

1.3.1 Signalling pathways

A signalling pathway is a series of PPIs, which transmit signals into a cell. This mechanism is invoked when an extracellular molecule activates a cell surface receptor protein, usually by binding to a receptor's site. A signalling pathway mainly involves processes of protein binding, phosphorylation and localisation. There are three main challenges for automated extraction of signalling pathways (Fig. 1.2).

The first major issue is that only a subset of all interacting entities is directly reported in the scientific literature and in particular as a relation between entities within the same sentence [34]. Some of the entities and events are explained in detail in consecutive sentences, where references to previously mentioned entities are made through coreference mentions (Point **1** in Fig. 1.2). Coreference resolution in the scientific literature is still a complex problem and will be discussed in Section 3.1.3. There are also cases where the entities that are the products of reactions are not directly mentioned in the text. In Fig. 1.2, "Decorin_EGFR" is a multi-protein complex, which is the product of the event,

[3] Part-for-whole metonymy: part of an entity is used to refer to the whole entity.

[4] Predicative metonymy: an association is made between a source and target word based on concomitance. It consists in conveniently making-way of a predicate for a non-standard but related argument.

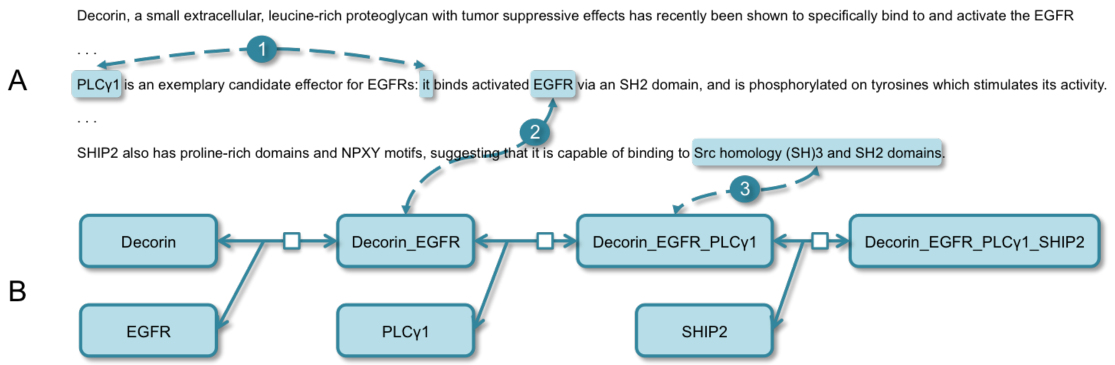

A
Decorin, a small extracellular, leucine-rich proteoglycan with tumor suppressive effects has recently been shown to specifically bind to and activate the EGFR
. . .
1
PLCγ1 is an exemplary candidate effector for EGFRs: it binds activated EGFR via an SH2 domain, and is phosphorylated on tyrosines which stimulates its activity.
. . .
2
SHIP2 also has proline-rich domains and NPXY motifs, suggesting that it is capable of binding to Src homology (SH)3 and SH2 domains.
3
B
Decorin
Decorin_EGFR
Decorin_EGFR_PLCγ1
Decorin_EGFR_PLCγ1_SHIP2
EGFR
PLCγ1
SHIP2

Fig. 1.2: Three hurdles of extracting signalling pathways by TM. **A** is text describing signal transduction stimulated by "Decorin" (From PubMed ID:10209155). The first sentence describes an event of "Decorin" binding "EGFR". The second sentence says that "PLCγ1" binds multi-protein complex "Decorin_EGFR", which was generated by the binding event of the first sentence. In the third sentence, "SHIP2" binds another new complex, which was generated in the second sentence. The diagram of the pathway snippet is visualised in **B**. The three challenges are (1) anaphoric coreference: in this example there is a classic case of pronominal anaphora; (2) part-for-whole metonymy[3] (synecdoche) [32]: the "EGFR" protein is used as a metonym for the entire protein complex resulting from the binding of "Decorin" and "EGFR"; (3) predicative metonymy[4] [33]: in this example "Src homology (SH)3 and SH2 domains" are the binding sites of the protein complex which has resulted from the binding of "Decorin_EGFR" and "PLCγ1". We assume that binding events take protein or protein complexes as their arguments and have a site as a modifier. Thus, in 'SHIP2 is capable of binding to "Src homology (SH)3 and SH2 domains", the binding sites are used in place of the protein complex which is the target of the binding event.

"Decorin...bind to and activate EGFR...". The product of the event is inferred, but it is referred to later in the text, often using referential metonymy. For example, a protein name may be used to refer to an entire protein complex, which includes the protein, where the complex has been implied as the product of previous reactions. Such is the case in Point **2** in Fig. 1.2, where "EGFR" in the sentence, "it binds activated EGFR via an SH2 domain", actually represents the multi-protein complex instead of the original receptor. Thirdly, there can be mention of a part of a molecule, e.g. a binding domain or residue, which denotes a particular protein or complex, previously mentioned or inferred (Point **3** in Fig. 1.2). In this case, the entity normalisation requires semantic information from the entity context and from biological reference ontologies. For example, considering the information, e.g. gene locus and attribute, from Gene Ontology can support disambiguation and improve the normalisation performance [35]. These challenges often cause topological mistakes in the pathway and reduce the automatic and correct interpretation of the generated network. The generation and use of proper nomenclature for the correct representation of macromolecular complexes would be an important step to solving this problem. For example, the protein complex generated by the binding of EGF to EGFR has been called EGF/EGFR [36] or

EGF·EGFR [37], which supports correct extraction results.

1.3.2 Metabolic pathways

Metabolic pathway is a series of enzyme catalysed biochemical reactions, which form or modify many chemicals, as known as metabolites. The reactions may require dietary minerals, vitamins, and other cofactors. Metabolic pathway is important of maintaining the homeostasis within an organism. In contrast to signalling pathways, a metabolic pathway is composed of interactions and reactions among proteins as well as small molecules such as chemicals and enzymes.

Chemical substances are denoted according to the standard nomenclature produced by the International Union of Pure and Applied Chemistry (IUPAC). Organic compounds and inorganic compounds respectively have own nomenclature systems. The Chemical Abstract Service developed a scheme to index chemical substances and assigned each chemical substance an identifier known as CAS registry number. The improved nomenclature eases recognition of reactants and identification of products of a reaction. It is potentially useful for interconnecting extracted events into a network (see Section 4.1.1). Nevertheless, recognition of chemical entities by TM systems is not as advanced as the recognition of protein and gene names, which has been at the heart of challenges and shared tasks in biological TM [15, 24, 38]. This is an additional difficulty to those mentioned above for signalling pathway extraction. Moreover, enzymes in metabolic pathways rarely appear in the same passage, let alone the same sentence [39].

1.4 Outline of the thesis

Identifying entities, including proteins and chemical entities, in text is a fundamental task of the study. In Chapter 2, I look into different biological named entity recognition (NER) solutions, and evaluate them on some community-supported gold standard corpora (GSC). The NER solutions are aligned with terminological resources and the GSCs. The experiment evaluate different NER solutions on several GSC of protein/gene, chemical and oncology. The systematical alignments, between the three resources of NER, i.e. corpora, lexica and taggers, deliver substantial information about how comprehensive each main stream GSCs are, and different approaches' contribution in different aspects of performances, e.g. false positive, false negative etc. Thus, we can know how the NER solutions' performances are influenced by different lexical resources when combining them for more sophisticated network extraction. More importantly, the information overlaps contains knowledge about how the terminological resources, the NER solutions and the GSCs comply with each other.

Chapter 3 focuses on event extraction, which provides another layer of information for biological network evaluation. The chapter specifically analyses the different approaches of event extraction and their limitations for network extraction. Then, it presents two state-of-the-art systems I developed for bio-event extraction, PCorral [40] and LitWay [41]. PCorral is a pipeline aggregating methods from high-recall to high-precision. The combination provides interactive mining of PPIs from the scientific literature allowing curators to skim MEDLINE for PPIs at low overheads. It has been integrated as part of the Whatizit infrastructure. LitWay is a more sophisticated event extraction tool based on a flexible infrastructure. It is designed for investigating precise molecular relationships for evidential statement collection and hidden relation discovery. LitWay is a flexible and modularised system built upon Apache UIMA (Unstructured Information Management applications) [42]. It can be facilitated with a set of syntactic rules, or machine learning (ML) algorithms. It can be easily customised for different tasks, and, on that account, is capable of identifying many kinds of complex MIs, including protein-protein interaction (PPI) and chemical-protein interaction (CPI) from free-text. The current setting of the system has a set of collaborative classifiers as the core. The system can identify different parts of biological events, and construct events. An example about LitWay's configuration file demonstrates the flexibility for other types of event extraction tasks. The evaluation of the pipeline on BioNLP-ST 2013 [28] data shows the system's performance is state-of-the-art.

In Chapter 4, a series of experiments, including using PCorral and LitWay, are conducted to evaluate the biological networks in BMDB. The methods use range from high-recall driven (co-occurrence and tri-occurrence) to high-precision driven (syntactic pattern or machine learning). The experiments quantitatively reveal entity compositions of different types of biological networks. It also characterises syntactic morphology and availability of different biological networks in the scientific literature, especially metabolic network and signalling pathway. This work delivers the first quantitive profiles about bio-molecules, their relations and different types of biological pathways in the scientific literature. At last, a comprehensive human metabolic network, RECON2 [43], is investigated for hidden relations by using LitWay combining statistics. The hypothetical relations and the related molecular activities are analysed with the respect to different cellular components and sub-pathways.

In a brief summary, the following list is the main contributions of this work.

- The study systematically aligns three information resources (GSC, terminology and

NER solutions) for identifying entities in unstructured text.

- The alignment delivers an objective comparison between NER solutions. More importantly, it reveals the knowledge overlap between three resources. ML approaches perform the best on the dataset they have been trained on. On new datasets, ML approaches' performances are even below lexical taggers'. Normalisation relies on lexical approaches.

- The alignment demonstrates that the state-of-the-art performance of the lexical tagger (SwissProt tagger).

- The work presents PCorral, a PPI extraction system. PCorral utilises SwissProt tagger for NER, so its EM component is state-of-the-art. It provides an interactive access for collecting evidential statements of PPI from high-recall to high-precision. The CO2-based component can be used for collecting literature sets. The components based on CO3, even more precise SynP, can be used to determine select explicit statements of PPIs.

- The work presents LitWay, a precise and flexible system for extracting MIs. The evaluation shows that LitWay's performance is state-of-the-art. The feature and the constraint experiments on LitWay show that different bio-events need solutions to have flexible settings for extraction.

- The study aligns the contents of the biological networks in a structured database, BMDB, against evidential statements collected from unstructured text, the scientific literature. The alignment evaluates the TM solutions, and quantitatively profiles the compositions of proteins, chemical entities in different types of the biological networks.

- The study quantitatively profiles the coverages of different molecular interactions (PPIs, PCIs, chemical reactions) in the scientific literature;

- The study investigates hypothesised molecular relations (HMRs) in RECON2 by aligning them with evidential statements from the scientific literature. The investigation is useful for analysing molecular activities in different cellular components and subpathways.

Chapter 2

Entity recognition for biological network extraction

Key points

- Entities are nodes of networks. NER provides the first layer of information about bio-entities' attributes for network evaluation. Meanwhile, relation/event extraction only can be done upon recognised entities.
- Alignments between three types information resources (NERs, GSCs and lexical resources) reveal knowledge overlap between them.
- For EM, ML approaches perform the best on test data, which is in the same set of training data. However, they do not link mentions with entries in structured databases (EN). On new datasets, ML may perform even worse than lexical taggers.
- EN still relies on dictionaries compiled from lexical resources. When normalisation applied, general performances go lower.
- Whatizit-GP7 and SwissProt significantly differ in sizes of the used terminologies. However, their performances are good and similar. This demonstrates that the scientific literature rather makes used of a conserved set of PGN terms.
- Lexical tagger utilising the terms from SwissProt shows competitive performance.
- The alignment is also informative about GSCs against NERs. The PennBio corpus and the FsuPrge corpus are the two corpora where the taggers tend to deliver their best performance. The tagging solutions perform worse on the BioCreative-II corpus than on the FsuPrge corpus.

- Different NER approaches should be used for evaluating different types of biological networks.

NER is a relatively mature technology in comparison with technologies of the other levels of network extraction. It builds a fundamental layer of information for network evaluation from unstructured text. However, different approaches perform differently on different data. Therefore, they need to be characterised for circumstances of evaluating different biological network.

In this chapter, entities extracted by NERs with different approaches are evaluated across publicly available GSCs and lexical resources. The detail about the work has been published in [44], [45] and [46].

2.1 Background

2.1.1 Entity recognition approaches

As entities are nodes of network, results of entity recognition (NER) topologically defines skeletons of extracted networks. There have been dictionary-based solutions or ML-based approaches for NER. Dictionary-based approaches collect lexica, i.e. names, synonyms and acronyms, of protein/gene and chemical from semantic or terminological resources, e.g. UniProt, ChEBI. The lexical information is compiled into controlled vocabulary, or called dictionary, and perform various string matching techniques against a text unit, such as a token. Following the progress in standardised semantic resources, the dictionary or lexical approach is becoming increasingly more compatible with biomedical ontologies [47, 48]. The advantage of dictionary-based approaches is that they do not need to be trained and can theoretically be applied to any scientific text for recognising entity mentions.

ML approaches require annotated corpora as gold standards for NER algorithms to collect characteristics of either text tokens or context. Therefore, NER is sometimes regarded as a sequence labelling task, as token order plays a role in identifying NE components. Hidden Markov Models (HMM) as well as Maximum Entropy Markov Models (MEMMs) have been used to address this [49]. Conditional Random Fields (CRF) is a popular ML alternative to the previous for sequence labelling, often used by biomedical NER, as it combines the advantage of MEMMs in exploiting non independent contextual features of the entity without a label bias problem [50–54]. Another popular ML technique used in NER is Support Vector Machines (SVM) [55, 56]. NER by SVM follows a text classification approach, where each token is given the appropriate NE category based on the morphological characteristics of the named entity and a set of contextual features. These approaches are fast and have a high performance.

One can further distinguish NER solutions into entity mention (EM) and entity normalisation (EN). EM solutions detect text components that make reference to a gene or gene product, while the EN solutions link the recognised entities to data entries in bioinformatics databases. The use of database information can contribute to the semantic disambiguation and homologous analysis of an entity. By crawling cross-linked resources, such as Gene Ontology or PSI-MI [57], network extraction systems can retrieve affinitive entities and interaction patterns. The host organism is usually determined by contextual lookup. As a result, ML solutions, such as SVM [58], can supply EN the assembly of a large amount of features from the context of an entity as opposed to rule-based solutions. GeneTUKit [59] uses SVM and was ranked first in the BioCreAtIvE III GN task according to the Threshold Average Precision (TAP-20) measure. GNer [60] is based on a dictionary compiled from Entrez Gene [61] and BioThesaurus, which it combines with SVMs using a set of extraction rules. GNAT [35] is a hybrid system using a dictionary and CRF for EM of genes and then correlates discovered genes with corresponding species.

Chemical entities play important roles in biological networks. They could be, for example, messengers, in signalling pathway, or diverse metabolites. The identification of chemical entities has not been studied as extensively as the identification of genes and proteins. As a result, fewer resources, i.e. freely available TM systems and corpora, have been made available for the identification of chemical entities and small molecules in comparison to gene and proteins. The Fraunhofer SCAI corpus for chemical compounds comprises 463 MEDLINE abstracts in the training set and 100 abstracts in the test set [62]. Entities in the corpus are IUPAC terms, trivial names, abbreviations, sum formulas, and chemical family names. OSCAR [63, 64] is one of the earliest tools for identifying chemical entities and reaction types in the scientific literature. On the SCAI corpus, ChemSpot [65], combining a dictionary and CRF trained on the IUPAC training corpus, outperformed OSCAR4's F-score by 10.8%.

2.1.2 Challenges, corpora and resources for NER

Challenges and competitions have been introduced to measure different NER solutions against the same benchmark so as to gain an insight into parameters of success for different solutions. The resources created and used for the competitions, e.g. corpora, tools and performance measures, have provided a long-term benefit to the community and have helped move the field forward. Since 2004, the BioCreAtIvE initiative [66] has organised three challenges, and covered topics ranging from gene mention [22, 67], gene normalisation [66, 68] and functional annotation (FA) [69] to PPI [15].

Each approach is usually good at a particular task or aspect. By harmonising different systems together it has been shown that a hybrid system can outperform individual systems [67, 68]. This has lead to the notion that training on harmonised annotations can improve system performance. To this effect, the CALBC (Collaborative Annotation of a Large Biomedical Corpus) initiative attempted to harmonise contributions from the community to automatically annotate a very large corpus (1 million abstracts) [45]. It explored different harmonisation methods based on the semantic groups (protein/gene, chemical, disease and species) and each annotation solutions characteristics. This initiative has also introduced the notion of a silver standard corpus, which is created from the harmonised output of different systems.

There are relatively fewer corpora and community challenges for chemical entities, compared to proteins. Except for SCAI, another corpus for chemical compound has been made available through the collaboration of the European Patent Office and the ChEBI team [70, 71]. The Colorado Richly Annotated Full Text Corpus (CRAFT) [23] contains 67 full text articles annotated with seven different biomedical concepts. Apart from ChEBI entities, it also contains annotations from Cell Ontology, Entrez Gene, Gene Ontology, NCBI Taxonomy, Protein Ontology and Sequence Ontology [72].

NER solutions rely on semantic resources, especially for EN. A number of terminological and semantic resources can potentially benefit NER systems. However, the information in the resources and their cross-linked databases have not been fully exploited for either EN or interaction extraction. Swiss-Prot [73] is a peptide sequence database, which has been widely used as a dictionary for entity recognition, but other information it contains such as location, organism, interaction and function has not been considered so far. Similarly, protein structure in Protein Data Bank [74] provides information such as the binding domain, which is useful for *part of whole metonymy* (Point **3** in Fig. 1.2). Protein family databases, e.g. Pfam [75], are useful to link a generic mention with a specific mention. The entities constituting the essential blocks of metabolism can be linked respectively to entries in databases of chemicals [70, 76–79], enzymes [80], metabolites [81, 82] and drugs [83–85] etc. While TM can benefit from the information in these resources it can also help curate and update them [86]. Research in creating links between entities (e.g. genes, phenotypes, diseases) in different knowledge resources in order to find evidence for disease can also benefit from text mining [72, 87–90].

2.2 IeXML, an in-line annotation format and its harmonisation

The nature of TM tasks is to convert unstructured texts into structured knowledge. Although most scientific literatures, e.g. in MEDLINE, are delivered in XML, the community still do not have a widely supported and standardised format for annotations, which can be easily transferred and understandable by most solutions. The infrastructures like Apache UIMA [42] and GATE [91] encode annotations in memory with offset, however, need extra effort to be converted to formats understandable by other solutions. In BioNLP-ST, several formats of annotations including JavaScript Object Notation (JSON), BioC and self-defined tab-separated format were provided for syntactic parsing result. In this case, large amount of work has to be done to transform annotations in different formats before any further processing. An alignment infrastructure for the systematic assessment of NER solutions should comply with the following requirements:

- A shared annotation format has to be used that fulfils the needs of the alignment task;
- The alignment has to consider the NER boundaries as well as the semantic annotation;
- Any size of annotated corpora has to be processed;
- Meaningful measures have to be implemented into the infrastructure to support the correct interpretation of the results.

The third requirement (any size of annotated corpora) puts constraints on the implementation of the solutions while the other requirements are essential to evaluate annotated corpora against a reference set.

Other annotation formats have been proposed to annotate scientific documents with relevant information, e.g. SciXML, TEI, CES. None of them provides compatible tools, which can be used in the evaluation of annotators. The exchange format for the BioCreative MetaServer (BCMS) [92] is suitable to integrate distributed modules but has not been proposed as a solution for the comparison of large-scale corpora. IeXML is an annotation framework that allows the interoperability of information extraction modules based on the sharing of annotation guidelines for the scientific document.

2.2.1 IeXML and Align-IeXML

Rebholz-Schuhmann D. et al. [93] proposed an in-line XML-tagging annotation format. Accompanying the format, an automatic evaluation of annotators against reference corpora through a Web based infrastructure (called Align-IeXML) was developed. If annotations comply with the IeXML annotation framework, they can be evaluated and receive the result from a web application utilising the compute cluster infrastructure at the EBI. The proposed solution has been used to systematically compare different annotators for genes and proteins against different reference corpora. We can demonstrate the performance differences depending on the reference corpus used. Align-IeXML is being used in the European Support Action CALBC [94] where annotators are systematically evaluated against a large-scale annotated corpus as part of the CALBC challenge.

```xml
<?xml version="1.0" encoding="UTF-8"?>
<ArticleTitle>
  Internal variation in the uptake of
  <e id="UMLS:C0043168:T047">whooping cough</e>
  immunisation within a Health Authority.
</ArticleTitle>
```

Fig. 2.1: IeXML sample annotation (semantic type)

Align-IeXML makes use of the IeXML annotation schema (Fig. 2.1). Currently, it only uses the markup for sentences and entities and uses special annotation guidelines for the representation of nested and overlapping annotation of entities in the text, i.e. in a segment of text, the annotation guidelines propose to tag individual tokens with unique ids and references stretch of an entity as part of the XML tag (see example in annotation guidelines of the CALBC challenge [95]). A standoff annotation format is provided in addition.

IeXML uses a tag *e* to denote an entity in text, and an id attribute to link the entity with the identifier in the existing resource and the semantic category. Fig. 2.1 is an example where an entity *whooping cough* is given an identifier and semantic type from the UMLS, these fields are separated by semicolon. IeXML standoff annotations are supported and facilities exist to turn them into IeXML inline annotation and vice-versa. An entity may be annotated with entries from different resources. In Fig. 2.2, *INS gene* is annotated with a UniProt entry and a UMLS entry. Two entries are separated by a pipe, '|'.

The entities are identified within the knowledge source identified using the *id* attribute in the *e* element. The identifier of a given entity in a given data source is composed of the

```
<e id="Uniprot:P01308:T028:PRGE|UMLS:C1337112:T028:PRGE">INS gene</e>
```

Fig. 2.2: IeXML sample annotation (normalisation)

```
<e id="UMLS:C0222601:T023:1,2|UMLS:C0006142:T191:2,3">
  <w id="1">left</w><w id="2">breast</w>
  <w id="3">cancer</w>
</e>
```

Fig. 2.3: IeXML sample annotation (boundary variance)

namespace of the knowledge source (e.g. UMLS), the identifier of the entity in this source (e.g. C0001403), the semantic type and the semantic group. If multiple identifiers may be assigned to the same text boundary (e.g. in cases of ambiguity), the pipe symbol is used to separate them.

Following the entity identifier, specified above as (namespace:id:semantic type:semantic group), a colon indicates that there is a comma-separated list of token identifiers. The following example illustrates this point: In this example, left breast, i.e. tokens 1 and 2, is identified by UMLS:C0222601:T023, while breast cancer (tokens 2 and 3) is identified by UMLS:C0006142:T191 (Fig. 2.3).

2.2.2 Evaluation against the reference set using Align-IeXML

Comparisons are performed at two dimensions. The first dimension is at the boundary at which the entities are compared; i.e. either the precise boundaries of the entities or more relaxed one like a sentence or a paragraph. The second dimension is the semantics assigned to the entities; e.g. semantic group, semantic type and identifiers provided by different resources.

The tool presents statistics of the alignment at different dimensions. These statistics include comparative figures like precision and recall but as well a frequency sorted list of agreements and disagreements in a pair-wise way.

Furthermore, it is possible to obtain sentences with examples of annotations from the compared corpora providing either a term being annotated or an identifier that provides further context to perform an analysis of the annotations. In addition, boundary disagreement between two annotated corpora can be produced, as seen in Fig. 2.4.

```
E|0|acid phosphatase|2|acid phosphatase|2|6578195_9
E|0|toxins|1|toxins|1|9759608_2
N|0|P ovale|29754317_6
N|0|P malariae|29754317_6
M|0|host|1|host responses|2||responses|2528547_2
L|0|NK 1 1|3|NK 1|2|1|2944955_4
R|0|alpha L fucosidase|3|fucosidase|1|alpha L|3196299_9
```

Fig. 2.4: Example of boundary disagreements

2.2.2.1 Boundary evaluation

Boundaries are either precise entity boundaries, where the tightest set of tokens defining the entities are used, or broader ones like a sentence or a paragraph, where the comparison is closer to text categorisation.

Tight boundaries If we consider tight boundaries, different types of entity matching are available. This includes exact match of the boundaries and relaxation of boundary match [96] that allow us to compare entities under different assumptions, e.g. differences due to function words (cosine) or if the annotations of one system span over a larger piece of text compared to another system (nested) or if there is an overlap between the systems (any).

Examples of agreements and disagreements are available from the tool, as found in Fig. 2.4. In Fig. 2.4 we find entities where the boundaries present: an exact match *acid phosphatase*, cannot be found in the other annotated set *P ovale* or agree on one of the boundary sides *alpha L fucosidase*.

Broad boundaries If we consider broad boundaries, entities are collected from a specified window and provide annotations to a larger span of text like a sentence, a paragraph or a document; e.g. BioCreAtIve II. The XML tag defining this window is used. These alignments allow testing several hypotheses of the underlying annotations and different agreement levels. Statistics of the entity boundaries agreements and disagreements are estimated as can be seen in Fig. 2.5.

2.2.2.2 Semantic evaluation

In addition to boundary alignments, the tool provides a comparison at several levels of semantic annotation from more general to more specific; i.e. the semantic group, the semantic type and the identifier of the annotated entities. These semantic categories cover a broad range of named entity tasks going from named entity recognition (e.g. gene mention) where only the semantic category is required to more specific tasks like named entity resolution

```
Agreement
Frequency: 340 Term: protein
Frequency: 268 Term: infection
Frequency: 225 Term: CD 4
Frequency: 184 Term: binding
Frequency: 173 Term: tumor
Frequency: 154 Term: proteins
...
Disagreement
Frequency: 431 Term: mice
Frequency: 112 Term: beta
Frequency: 99 Term: mouse
Frequency: 83 Term: CTL
Frequency: 83 Term: alpha
```

Fig. 2.5: Example of boundary statistics outputted by Align-IeXML

where the identifier within an existing resource is provided (e.g. gene normalisation).

The boundary and semantic dimension combined, allows using corpora available to evaluate a large set of named entity annotation tasks.

Our framework analyses agreements/disagreements of annotated entities and estimates statistics concerning each semantic level (see Fig. 2.6).

If we consider named entity resolution, concept identifiers normalise annotations of entities in texts; i.e. different strings are linked to the same identifier. Our framework provides mechanisms to show the strings linked to the identifiers (see Fig. 2.6).

Even though, any semantic resource can be used, the UMLS has been integrated into the comparison tool. Semantic categories provided by [97] are used. This means that annotations at several levels can be converted from more specific to more general. This means that given the annotation of an entity with an UMLS CUI we could derive its semantic types and then the semantic groups. If the annotation is done only with the semantic type, the semantic group can be inferred.

2.3 Evaluating GSCs against protein/gene tagging solutions and lexical resources

Proteins are fundamental blocks of living organisms. In the scientific literature, there is no separate nomenclature for protein and gene. In the available solutions, protein/gene names (PGN) are delivered by various lexical resources and terminological resources. GSCs

```
Annotations Group: 71497 chem Chemicals & Drugs
Group: 4832 diso Disorders
Group: 3244 livb Living Beings
Group: 972 phen Phenomena
Group: 603 anat Anatomy
Group: 250 phys Physiology
Id: 504 species:10095 mice|Mice [livb] []
Id: 472 chebi:16541 proteins|protein|Protein [chem] [gene]
Id: 339 disease:c0009450 infectious diseases|Infection|infections [diso] [phen]
Id: 333 disease:c0021311 Infection|infections|Infections|infection [diso] [phen, proc]
Id: 271 disease:c0027651 tumour|tumours|tumors|tumor|Tumour [diso] [phen]
```

Fig. 2.6: Example of group/identifier statistics outputted by Align-IeXML

provide features of contexts. Different solutions exploit the three kinds of resources at different levels. It is known that ML approaches usually perform better on corpora they were trained on. However, studies and comparisons between solutions were usually done against the same corpus. There is no comparison across the three types of resources, and no work has been done to look into the discrepancies by aligning them.

We conduct the experiments to test the mainstream dictionary-based solutions using various terminological resources, and ML-based solutions, on several community-supported GSCs. The investigation delivers valuable information about the characteristics of the differences from, at least, two dimensions: contribution of different types of resources to PGN tagging solutions, and exploitation level of different types of resources by solutions. This gives a further broader view about respective roles of terminology, semantics and NER in biological network extraction. It can also practically guide biologists in choosing suitable TM tools for particular tasks, because, when extracting interesting interaction network, the solution regarded the best do not necessarily perform the same under different task requirements. For example, PGN tagging solutions may perform differently when processing a set of selected articles or an entire literature database, e.g. MEDLINE.

2.3.1 Materials and methods

Gold standard corpora

Gold standard corpora have different characteristics in, for example, size, topic, release date, and annotation guidelines that will impact the performance of the tagging solutions. Table 2.1 contains statistics of the different selected corpora.

The Jnlpba corpus is based on the GENIA corpus [98], which contains annotation for different entity types linked to molecular biology such as PGN, cell type, and cell line, all being compliant with the GENIA ontology.

The BioCreative-II corpus covers human genes and proteins and contains sentences instead of complete MEDLINE abstracts [99]. It makes use of an alternative gene-list in addition to the regular list of genes for example, the GL (= GeneList) contains

(1) P00027967A0207 | 11 31 | secretory HI antibodies,

while the AGL (= Alt Gene List) also includes:

(2) P00027967A0207 | 11 21 | secretory HI,

(3) P00027967A0207 | 20 21 | HI, and

(4) P00027967A0207 | 20 31 | HI antibodies for the same annotation example.

The MEDLINE abstracts in PennBio have been developed with a focus on oncology [100]. The FsuPrge corpus is the largest of all and has a focus on molecular mechanisms such as gene regulation.

Table 2.1: A number of gold standard corpora have been delivered to the public for the evaluation of PGN tagging solutions

Name	Release	# Annot.	# Units	Topic
Jnlpba	2004	6,142	401 abs.	Subset of GENIA
BioCreative-II	2005	5,144	4,171 sent.	Human proteins
PennBio	2006-07	18,148	1,414 abs.	Oncology
FsuPrge	2009	59,483	3,236 abs.	Gene regulatory processes

Tagging solutions

We have tested several tagging solutions, with different underlying approaches. First, we have used state of the art *ML-Tag* approaches for gene mention identification. Second, we have used standard *LexTag* solutions where different terminological resources for PGNs have been integrated with and without disambiguation techniques. Last, we have combined the first and the second type to filter out false positives after dictionary lookup.

The *ML-Tag* approaches comprised Abner (BC1) trained on BioCreative-I, Abner (Jnlpba) trained on Jnlpba, and two taggers trained on BioCreative-II (Banner, “Chang2”) based on Conditional Random Fields (CRF) [52, 101]. The Banner tagger has been downloaded from the distribution site [102]. The Chang2 tagger is a CRF model trained on BioCreative-II data with a set of features as given in [101]: (1) all character *n*-grams of length 2 to 4, (2) inclusion of capitalisation (token starts or ends with a Capital letter, or Capital letters only etc.), (3) length of the tokens (one character, two characters, between three and five), (4) inclusion of digits (1 digit, 2 digits, only digits), and (5) contained punctuation symbol or a Greek character. In addition a contextual window of plus or minus 2 tokens is employed (called “offset conjunctions”). The model is implemented as CRF (using mallet).

For the *LexTag* solutions, we tested publicly available Whatizit modules from UKPMC and other research (cf. Table 2.2) [103, 104]. In addition, we have compared the two latest versions of the Biothesaurus (version 6.0 and 7.0, GP6 and GP7, respectively) as part of the existing solutions to identify the importance of the lexical resource.

SwissProt (SP) is the Whatizit-SwissProt module integrated into different text mining solutions such as EBIMed [105] and PCorral [40]. It uses terms retrieved from the SwissProt subset of UniProtKb obtained in 2007. SP(GP7) is the updated version, which comprises the full selection of SwissProt PGNs from Biothesaurus 7.0. The tagging of genes applies morphological variability to terms, i.e. accepting separators ([- /]), initial capitalisation, and singular-plural variability (“alias matching”), since this morphological variability is very common in the use of gene names [106]. This approach is similar to approximate string matching, e.g. Levenstein distance, but is more specific and thus better suitable for the comparison of PGN terms. Alternative methods include the automatic generation of a large dictionary resource containing all terms exposing the same term variability and then apply exact matching [107].

Finally, all tagged PGNs will be removed that are too unspecific, i.e. those terms that are part of the general English language and would be difficult to attribute to a specific gene or protein. This approach applies to all PGNs that appear in the British National Corpus (BNC) with a frequency rate that is higher than the one of “insulin” (“basic disambiguation” or “BNC disambiguation”). It has been used in all *LexTag* solutions.

The GP7 solution makes use of the full content from Biothesaurus 7.0, applies morphological variability and basic BNC disambiguation only. BioLexicon integrates the full BioLexicon content into the same approach. The two solutions Wh-Ukpmc and Wh-Ukpmc (GP7) implement the same solution as before with Biothesaurus 6.0 and 7.0, respectively, and in addition we apply FP filtering with the Chang2 tagger after the basic BNC disambiguation to increase precision (“false positive filtering”).

Table 2.3 gives an overview on the content used for building the dictionary-based solutions and demonstrates how the lexical resources differ in their size and content. The comparison makes use of exact and alias matching (see above) where a dictionary is incorporated into a lexical tagger (see above) for the matching of terms in the other lexical resource. According to our manual evaluation, GP7 in contrast to GP6 exposes reduced term variability across the resource to be more compliant with the naming standards in the biomedical research community.

The GNAT gene mention tagger and the gene normalisation solution have been tested against the corpora [107]. The GNAT gene normalisation mode has been used, since the gene mention mode of GNAT is based on Banner. The tagger has been applied in different modes using only the dictionary for human genes (GNAT/hum.) or for all species (GNAT/all).

Four GSCs were chosen for the study. JNLPBA [108] is a subset of GENIA corpus released in 2004 with 401 abstracts and 6,142 annotations. BioCreative-II [67] was released as a human PGN recognition challenge in 2005. It has 4,171 sentences and 5,144 annotations. PennBio [100] is an oncological corpus released in 2006 and has 18,148 annotations in 1,414 abstracts. FsuPrge [109] is a gene regulatory processes corpus released in 2009 and has 59,483 annotations in 3,236 abstracts.

2.3.2 Result

Eight tagging systems with different approaches are tested (Table 2.2). Four of them were either trained on several different corpora or integrated with different lexical or terminological resources, so totally thirteen solutions were evaluated on all the GSCs. Firstly, the state of the art *ML-Tag* approaches for gene mention identification have been used. Then, the standard *LexTag*, or dictionary-based, solutions with different terminological resources for PGNs have been integrated with and without disambiguation techniques. Finally, the first and the second type are combined to filter out false positives (FP) after dictionary lookup.

The ML approaches include ABNER (BC1) trained on BioCreative-I, ABNER (JNLPBA) trained on JNLPBA, and another two taggers, Banner [102] and "Chang2" both trained on BioCreative-II by using Conditional Random Fields (CRF) [52, 101]. The Chang2 tagger is based on a CRF model trained on BioCreative-II data. Its features are described in [101]: (1) *n*-grams of characters are with length 2 to 4; (2) character cases are considered (token starts or ends with a capital letter, or capital letters only etc.); (3) the tokens' length is considered (one character, two characters, between three and five); (4) number of digit in token is considered (1 digit, 2 digits, only digits); (5) whether token contains punctuation symbol and Greek character is considered. In addition, a contextual window of plus or minus 2 tokens is employed (called "offset conjunctions"). The CRF model is implemented by using mallet [110].

The *LexTag* solutions used in the experiment include Whatizit [103] modules from UKPMC [104] and other research (see Table 2.2). The two latest versions of the Biothesaurus (version 6.0 and 7.0, GP6 and GP7, respectively) are also tested as part of the

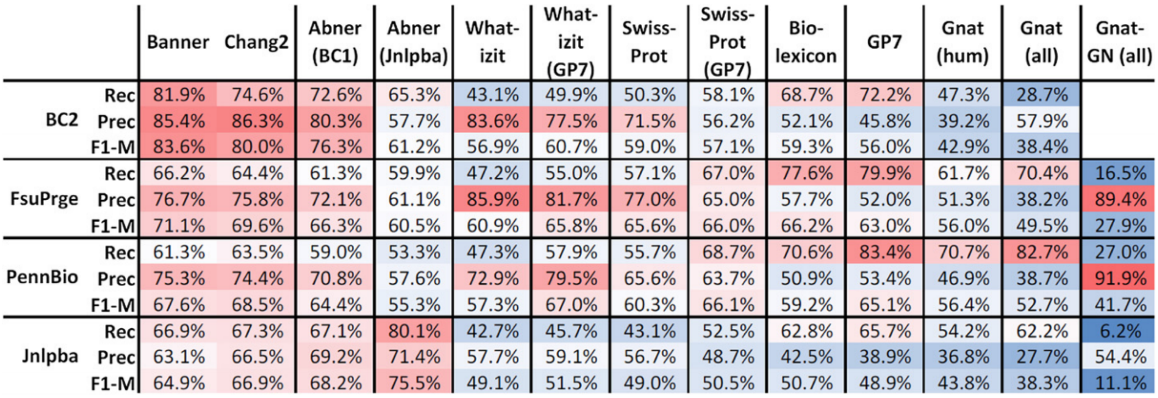
Table 2.2: NER taggers comparison

Tagger name	Tagger type	Lexical resource	# Lexical entries	Id	Training data	FP filter
Banner	ML	-	-	No	BC2	Banner
Chang2	ML	-	-	No	BC2	Chang2
Abner (BC1)	ML	-	-	No	BC1	Abner (BC1)
Abner (Jnlpba)	ML	-	-	No	Jnlpba	Abner (JNLPBA)
SwissProt	Lex	SwissProt	228,893	Yes	-	BNC
SwissProt (GP7)	Lex	GP7	868,050	Yes	-	BNC
BioLexicon	Lex	BioLexicon	653,212	Yes	-	BNC
GeneProt 7.0	Lex	GP7	1,725,500	Yes	-	BNC
Wh-Ukpmc	Lex+ML	SwissProt	228,893	Yes	-	BNC, Chang2
Wh-Ukpmc (GP7)	Lex+ML	GP7	868,050	Yes	-	BNC, Chang2
GNAT (human)	Lex+ML	Human genes		Yes	BC2	-
GNAT(all)	Lex+ML	11 species	80,000	Yes	BC2	-
GNAT-GN (all)	Lex+ML	11 species	80,000	Yes	BC2	-

Table 2.3: Lexical resources comparison

			Tagger							
Match	Corpus	# Entries	SP	[%]	Biolexicon	[%]	SP (GP7)	[%]	GP7	[%]
Exact	SwissProt	228,893			208,069	90.9%	121,369	53.0%	135,018	60.0%
	BioLexicon	653,212	207,976	31.8%			243,573	37.3%	422,477	64.7%
	SP(GP7)	868,050	121,030	13.9%	243,271	28.0%			860,094	99.1%
	GP7	1,725,500	134,275	7.8%	421,520	24.4%	859,536	49.8%		
Alias	SwissProt	228,893			213,009	93.0%	201,633	88.1%	206,047	90.0%
	BioLexicon	653,212	229,759	35.2%			375,550	57.5%	585,205	89.6%
	SP(GP7)	868,050	219,185	25.3%	364,171	42.0%			865,590	99.7%
	GP7	1,725,500	267,947	15.5%	644,115	37.3%	956,314	55.4%		

existing solutions, as the cross comparison identifies the importance of the lexical resource. Integrated as a NER module in Whatizit, SwissProt (SP) builds up its vocabulary by retrieving terms from the SwissProt (Table 2.3). SP(GP7) is the updated version, in which the full set of SwissProt PGNs from Biothesaurus 7.0 is enclosed. The tagging of genes considers morphological variability of terms, such as enclosed separators (hyphen etc.), upper/lower case, and plurality, as this morphological variability is common in the appearance of gene names [106]. Other methods include the automatic generation of a large dictionary resource containing all terms exposing the same term variability and then apply exact matching [107].

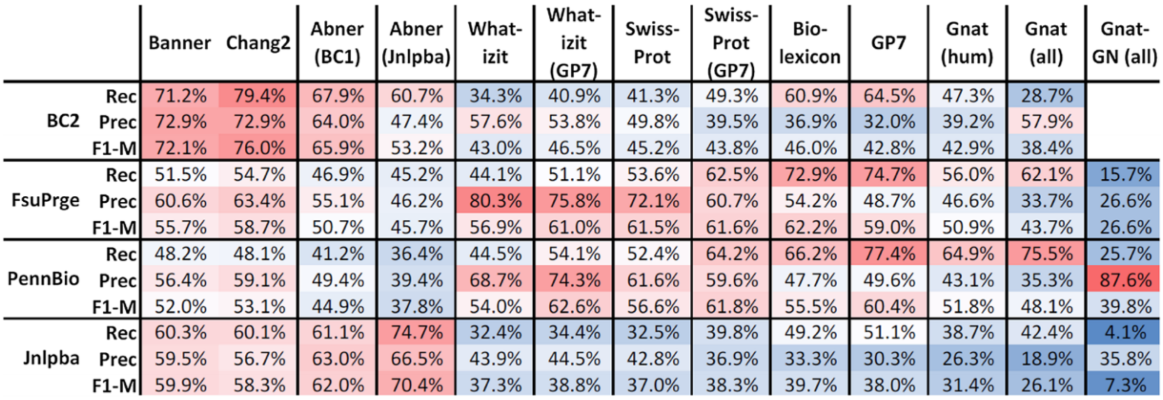


		Banner	Chang2	Abner (BC1)	Abner (Jnlpba)	What-izit	What-izit (GP7)	Swiss-Prot	Swiss-Prot (GP7)	Bio-lexicon	GP7	Gnat (hum)	Gnat (all)	Gnat-GN (all)
BC2	Rec	81.9%	74.6%	72.6%	65.3%	43.1%	49.9%	50.3%	58.1%	68.7%	72.2%	47.3%	28.7%	
	Prec	85.4%	86.3%	80.3%	57.7%	83.6%	77.5%	71.5%	56.2%	52.1%	45.8%	39.2%	57.9%	
	F1-M	83.6%	80.0%	76.3%	61.2%	56.9%	60.7%	59.0%	57.1%	59.3%	56.0%	42.9%	38.4%	
FsuPrge	Rec	66.2%	64.4%	61.3%	59.9%	47.2%	55.0%	57.1%	67.0%	77.6%	79.9%	61.7%	70.4%	16.5%
	Prec	76.7%	75.8%	72.1%	61.1%	85.9%	81.7%	77.0%	65.0%	57.7%	52.0%	51.3%	38.2%	89.4%
	F1-M	71.1%	69.6%	66.3%	60.5%	60.9%	65.8%	65.6%	66.0%	66.2%	63.0%	56.0%	49.5%	27.9%
PennBio	Rec	61.3%	63.5%	59.0%	53.3%	47.3%	57.9%	55.7%	68.7%	70.6%	83.4%	70.7%	82.7%	27.0%
	Prec	75.3%	74.4%	70.8%	57.6%	72.9%	79.5%	65.6%	63.7%	50.9%	53.4%	46.9%	38.7%	91.9%
	F1-M	67.6%	68.5%	64.4%	55.3%	57.3%	67.0%	60.3%	66.1%	59.2%	65.1%	56.4%	52.7%	41.7%
Jnlpba	Rec	66.9%	67.3%	67.1%	80.1%	42.7%	45.7%	43.1%	52.5%	62.8%	65.7%	54.2%	62.2%	6.2%
	Prec	63.1%	66.5%	69.2%	71.4%	57.7%	59.1%	56.7%	48.7%	42.5%	38.9%	36.8%	27.7%	54.4%
	F1-M	64.9%	66.9%	68.2%	75.5%	49.1%	51.5%	49.0%	50.5%	50.7%	48.9%	43.8%	38.3%	11.1%

Fig. 2.7: NER solutions evaluated on GSCs with cosine 98% similarity alignment.

Fig. 2.7 and Fig. 2.8 are the evaluation results of the taggers on the GSCs. The difference between two figures is the evaluation methods. Fig. 2.8 uses exact term boundary matching and Fig. 2.7 applies cosine 98% similarity matching. All ML solutions perform the best on those corpora that they have been trained on and have lower performance against the other corpora. Banner shows the best performance in comparison to all the other PGN taggers only on BioCreative-II (exact and cos98 matching evaluation). Chang2 has higher performance on BioCreative II by exact matching and has slightly lower performance by cosine 98%. It is also trained on BioCreative-II, but profits from optimisation to reach the performance of Banner at the expense of lower precision.

On the FsuPrge corpus, Banner only achieved the best performance when using cosine 98% similarity matching. By exact matching, its performance is lower than four of the six *LexTag* solutions (Fig. 2.8). On PennBio and JNLPBA, Banner is not the best performing solution. Chang2 performs better than Banner although they were trained on the same corpus (only except for PennBio, by exact matching). Abner (BC1), which has been trained on BioCreative-I, shows the best performance against the BioCreative-II test corpus indicating that similar annotation guidelines apply to both corpora.



		Banner	Chang2	Abner (BC1)	Abner (Jnlpba)	What-izit	What-izit (GP7)	Swiss-Prot	Swiss-Prot (GP7)	Bio-lexicon	GP7	Gnat (hum)	Gnat (all)	Gnat-GN (all)
BC2	Rec	71.2%	79.4%	67.9%	60.7%	34.3%	40.9%	41.3%	49.3%	60.9%	64.5%	47.3%	28.7%	
	Prec	72.9%	72.9%	64.0%	47.4%	57.6%	53.8%	49.8%	39.5%	36.9%	32.0%	39.2%	57.9%	
	F1-M	72.1%	76.0%	65.9%	53.2%	43.0%	46.5%	45.2%	43.8%	46.0%	42.8%	42.9%	38.4%	
FsuPrge	Rec	51.5%	54.7%	46.9%	45.2%	44.1%	51.1%	53.6%	62.5%	72.9%	74.7%	56.0%	62.1%	15.7%
	Prec	60.6%	63.4%	55.1%	46.2%	80.3%	75.8%	72.1%	60.7%	54.2%	48.7%	46.6%	33.7%	26.6%
	F1-M	55.7%	58.7%	50.7%	45.7%	56.9%	61.0%	61.5%	61.6%	62.2%	59.0%	50.9%	43.7%	26.6%
PennBio	Rec	48.2%	48.1%	41.2%	36.4%	44.5%	54.1%	52.4%	64.2%	66.2%	77.4%	64.9%	75.5%	25.7%
	Prec	56.4%	59.1%	49.4%	39.4%	68.7%	74.3%	61.6%	59.6%	47.7%	49.6%	43.1%	35.3%	87.6%
	F1-M	52.0%	53.1%	44.9%	37.8%	54.0%	62.6%	56.6%	61.8%	55.5%	60.4%	51.8%	48.1%	39.8%
Jnlpba	Rec	60.3%	60.1%	61.1%	74.7%	32.4%	34.4%	32.5%	39.8%	49.2%	51.1%	38.7%	42.4%	4.1%
	Prec	59.5%	56.7%	63.0%	66.5%	43.9%	44.5%	42.8%	36.9%	33.3%	30.3%	26.3%	18.9%	35.8%
	F1-M	59.9%	58.3%	62.0%	70.4%	37.3%	38.8%	37.0%	38.3%	39.7%	38.0%	31.4%	26.1%	7.3%

Fig. 2.8: NER solutions evaluated on GSCs with exact matching alignment.

2.4 Discussion and conclusion

Identifying entities retrieves the first semantic layer of biological networks from unstructured texts. The alignment between entities and GSC delivers knowledge overlap between the NER solutions and the GSCs. The experiment we conducted compares the NERs with different approaches across publicly available GSCs. The experiment comprehensively aligns three important resources: NER solutions, datasets (GSCs) and lexical resources. This is not merely a comparison between different solutions. More importantly, the alignment shows the knowledge overlap between three types of resources. It also characterises different approaches on different datasets.

For identifying mentions of PG of biological networks (EM) in unstructured text, ML approaches perform the best on the test sets, which are from the same datasets where the training data is from. The ML solutions may perform even worse than lexical tagging solutions on other corpora. The *LexTag* solutions have different profiles for their precision and recall performances, but the F1-measure remains in a very similar range. Meanwhile, ML approaches only play the part of EM, however, does not link the mentions with entries in structured database, which is EN and still relies on dictionaries compiled from lexical resources. When normalisation is applied, general performances go lower.

According to the experiment results, among the evaluated EM solutions, Whatizit-GP7 and SwissProt show similar and actually good results. Whatizit-GP7 has a large terminological resource (1.7 million terms), and is implemented with the FP filtering with Chang2. On another hand, SwissProt has a small number of terms (about 230,000), but the performance is still competitive. This shows that the scientific literature rather makes use of a conserved set of PGN terms.

PCorral [40] presented in next chapter (Section 3.2) is based on SwissProt. Therefore, PCorral is a PPI extraction solution, which EM part is state-of-the-art.

Conversely, the alignment also gives information about GSCs against NERs. For GSCs, the PennBio corpus and the FsuPrge corpus are the two corpora where the taggers tend to deliver their best performance. The tagging solutions perform worse on the BioCreative-II corpus than on the FsuPrge corpus.

Chapter 3

Event extraction for biological network extraction

Key points

- Event extraction is the current focus of TM. It elaborates relationships between molecules and provides another information layer for network evaluation.

- PCorral uses SwissProt tagger, whose performance has been shown to be state-of-the-art. Therefore, PCorral's EM and EN are state-of-the-art.

- PCorral provides an interactive access for collecting evidential statements of PPI from high-recall to high-precision. The CO2-based component can be used to determine the literature set. The components based on CO3, even more precise SynP, can be used to determine the explicit statements of PPIs.

- LitWay is a precise system for event extraction, which gains high precision with a reasonable recall.

- LitWay is implemented with a search-based structured prediction algorithm, SEARN. The evaluation against BioNLP'13 data shows its state-of-the-art performance.

- The feature and the constraint experiments on LitWay show that different bio-events demand solutions to have flexible settings.

- LitWay is a flexible system, which can be configured to extract different types of bio-events.

Event extraction differs from relation extraction, and is the current focus of TM. Besides existence of relation, it also extracts information about directionality and polarity about molecular relations. It provides another information layer for network evaluation.

PCorral [40] and LitWay are two systems of extracting molecular relations with state-of-the-art performances. PCorral's detail has been published in [40]. LitWay is downloadable [41] under GNU General Public License (GPL).

3.1 Background

Rapid improvement of experimental methods, e.g. two-hybrid screening and mass spectrometry for PPI investigation, generated substantial publications about biochemical reactions. Co-occurrence, pattern-based and machine learning are the three dominant approaches for automated bio-event extraction by TM [111]. In the case of co-occurrence based methods the underlying assumption is that a pair of entities is interacting when they appear in a text unit, such as a sentence, a paragraph or an article. More sophisticated approaches to event extraction rely on syntactic parsing to extract grammatical relations. These grammatical relations are aligned with expert defined information extraction templates, or characterised using machine learning algorithms.

3.1.1 Co-occurrence

A co-occurrence approach to event extraction is less computationally intensive, as it does not involve syntactic parsing. Precisely because of this, there is no way to detect the network directionality or even confirm that the entities are actually interacting. The systems [21, 112–114] built upon it can be ran against very large corpora [115, 116]. Co-occurrence based systems can also be combined with a set of trigger words to detect interaction types [117]. The overall number of connections between specific entities can be quantified to reveal entity clusters to represent sub-networks or detect synonyms etc. to reveal the underlying structure of a network [112, 118, 119]. For example, CoPub [118, 119] uses regular expressions to search for a term or a pair/set of terms to infer relations between co-occurring genes, drugs, pathways and diseases. The discovered relation could be a hidden relation of the first type (see Section 4.1.2), if it is novel for the curated databases. Based on an ABC-principle, Frijters et al. [120] inferred that entity A interacted with C when A interacts with B and B interacts with C. The relations discovered in this way could be the candidates of the second type of hidden relations (see Section 4.1.2).

Although most co-occurring entities in text are not really interacting [34], co-occurrence captures the maximum number of true positives within a given unit of text. Co-occurrence based extraction can be used as a baseline, since it reflects the maximum recall that an extraction system can get within a text unit. Meanwhile, due to the agility, co-occurrence can be adopted for quickly filtering articles related to particular concepts [114]. Subsequent processes, which could be more advanced but computationally intensive, such as syntactic parsing, can be run on the data filtered in this way.

3.1.2 Patterns, syntactic parsing and machine learning

Co-occurrence based event extraction suffers from the problems of a bag-of-words approach, which counts the frequency of words encountered in the same context without considering grammatical relations holding between them or even word order. Entities mentioned in the same discourse unit (sentence, paragraph, abstract) are not necessarily interacting.

Furthermore, the presence of a trigger word may not always correspond to an actual interaction between entities and the problem becomes complicated when several potential trigger-words appear in the same text unit. This often results in a large number of false positives (FP), as all the combinations between entities and trigger words are produced. For example, for the sentence, "Phosphorylation of p53 disrupts Mdm2-binding", a system only identifying entities and trigger words will ignore contextual information and erroneously report that p53 binds Mdm2.

In addition, common syntactic patterns such as coordination increase the number of FP results. For example, in the sentence, "Binding of hnRNP H and U2AF65 to respective G-codes and a poly-uridine tract", the two splicing factors and the two binding targets are forming alternatives, respectively, that would lead to four different pairs (hnRNP H → G-codes, hnRNP H → poly-uridine tract, U2AF65 → G-codes and U2AF65 → poly-uridine tract) when using co-occurrence analysis. To accurately acquire the relations, it is necessary to syntactically parse and analyse the sentential structure.

Syntactic parsing is the process of analysing text. It assigns parts of speech to sequential tokens and builds grammatical structure upon tokens. Thus, event extraction systems can determine the existence and attributes of an interaction based on the syntactic characteristics.

Shallow parsing, sometimes called chunking, only identifies the sentence constituents, e.g. noun phrase, verb etc. It does not aim to analyse the grammatical relations between the constituents. By contrast, deep parsing is able to tackle grammatical relations and allow the consideration of semantic relations between constituents [121, 122] as it constructs a more detailed structure based on syntactic grammars.

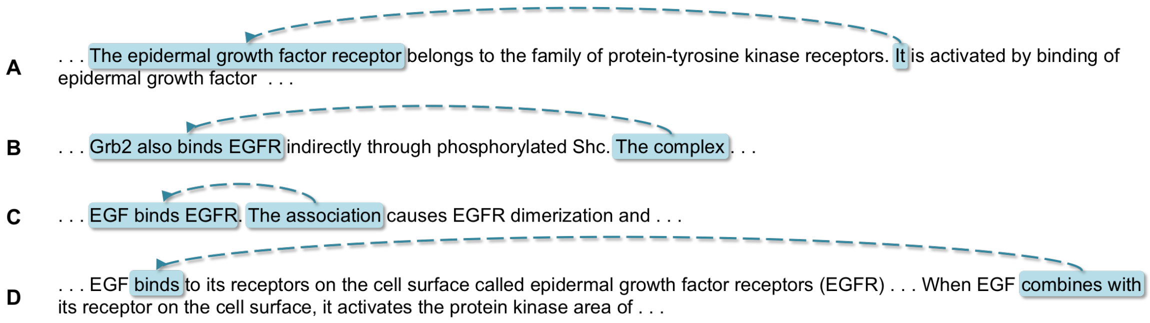
There are many parsers, which correspond to different grammars and schools of thought. Among them, dependency parsing and head-driven phrase structure parsing are two popular approaches. Dependency parsers generate links between words, where one is the head and the other the dependent, whereas phrase structure parsing organises syntax into nested structures, usually syntactic trees. It is possible to generate dependencies from both types of parsers [123–128] while the former have a clear appeal for applications which require grammatical relations, such as event extraction.

Pattern-based systems traverse the extracted syntactic structures and align them with a set of patterns to spot syntactic characteristics of interactions. These patterns are either produced by domain experts or can be obtained automatically from the text. They are usually implemented as regular expressions or as templates, which include part-of-speech (POS) information [129]. Patterns also can be derived automatically from a corpus, using machine-learning, based on a small set of patterns, being collected from sentences with similar characters of POS and/or grammatical relations [130–132]. In either case, the patterns model the characteristic language structure used in the domain and the corpus in question [40], therefore, the approach can achieve high precision [133–136]. Temkin et al. [137] created a set of patterns to extract CPIs by interpreting the output from a context free grammar (CFG) parser.

The interpretation of the type of interaction can vary significantly depending on the context. Thus, contextual information, such as entity types, should be taken into account for determining event types. For example, "a phosphate group attaches on a protein" is an event of type *phosphorylation* if the argument "phosphate group" is considered, rather than an event of *Binding*, which could be triggered by the verb 'attach'. For this reason, machine learning approaches have become popular as they statistically capture the characteristics of context and arguments relating to syntactic structures.

Machine learning approaches, when analysing grammatical relations obtained from parsing, can adopt different strategies as they focus on different aspects of the syntactic structure. There is a trend in combining different parsers so as to make the most of the syntactic structure [138, 139]. The challenges [15, 24, 38] which have taken place in the last few years have produced valuable resources and analyses for the bio-TM community, and at the same time have provided standards for event extraction. In BioNLP'11 [24], the extraction of an event can be achieved by checking three aspects: event triggers, argument linking and argument grouping. Some approaches extract triggers, arguments of simple events and arguments of complex events in three separate stages [140, 141]. More recently, models for joint extraction of event triggers and arguments have been proposed [142–145]. McClosky et al. [145] transforms candidate events, which consist of preliminarily recognised triggers and arguments, and their arguments into dependency trees. Subsequently, it uses a re-ranking dependency parser, which was modified from the MSTParser [146, 147], to parse the event tree.

Variants of event expression in human language always exceeds what a single training annotated source can cover, regardless of its size. To intelligently adapt themselves to arbitrary resources, event extraction systems [144, 148, 149] have used domain adaptation strategies to get supplementary information from unannotated data when training their models.



A . . . The epidermal growth factor receptor belongs to the family of protein-tyrosine kinase receptors. It is activated by binding of epidermal growth factor . . .

B . . . Grb2 also binds EGFR indirectly through phosphorylated Shc. The complex . . .

C . . . EGF binds EGFR. The association causes EGFR dimerization and . . .

D . . . EGF binds to its receptors on the cell surface called epidermal growth factor receptors (EGFR) . . . When EGF combines with its receptor on the cell surface, it activates the protein kinase area of . . .

Fig. 3.1: Examples of anaphoric coreference. The arch in **A** links the pronoun, "it", with the noun phrase, "The epidermal growth factor receptor", in the preceding sentence. The example in **B** cannot be resolved by syntactic approaches to anaphora resolution. "The complex" is a metonymic reference to the multi-protein complex created during the binding event of the previous sentence. Semantic interpretation is required to infer the existence of this multi-protein entity, so that it can be linked to "the complex" (see also the examples in Fig. 1.2). In **C**, "The association" does not refer to a previous mention of an entity or a product of an event. It refers to the previous event in its entirety, namely, "EGF binds EGFR". In **D**, the second event, "combines with", refers to the previous event, signalled by the trigger "bind".

3.1.3 The role of coreference resolution and discourse

Interactions and events may be expressed or denoted in several mostly consecutive sentences. Limiting the unit of event extraction to the sentence level reduces system recall, and this is an issue that current systems for event-extraction are beginning to address. Entity coreference resolution aims to link together mentions of the same entity across different discourse units (usually sentences) [150]. Event coreference is also beginning to receive attention as it caters for the relation between verbs pointing to the same event or entities referring to events. It is expected that entity coreference and event coreference can assist each other [151].

The use of coreference resolution is vital in improving coverage in network extraction as it can allow interlinking events in the network (section 4.1). Fig. 3.1 shows four typical cases of coreference in the biomedical literature. The pronoun "It" in sentence **A** makes reference to "epidermal growth factor receptor" from the previous sentence. This is a classic case of anaphoric coreference. In sentence **B**, "The complex" is a metonymic reference to the multi-protein complex created during the binding event of the previous sentence. Semantic interpretation is required to infer the existence of this multi-protein entity, so that it can be linked to "The complex" (see also the examples in Fig. 1.2). Sentence **C** has an instance of coreference between an entity and an entire event; i.e. "The association" refers to the *binding* event of the preceding sentence. Finally, in the example sentence **D**, the two event triggers "combine" and "bind" represent synonyms and constitute an example of event coreference.

The BioNLP Shared Task series focus on bio-molecular events in scientific literature. They consist of event extraction tasks along with other tasks, which bear the imprint of the trend from core event extraction [38] to coreference [134] and pathway curation [152] (see BioNLP detail in Section 3.1.4). As it was shown that lack of coreference resolution significantly hindered the event extraction performance [38], BioNLP'11 [24] organised a supporting task of identifying coreferential relations between proteins/genes. With all the participating systems favouring precision [134], the best performing solution [153], a modified system from Reconcile [154] based on supervised machine learning, obtained 73.7% precision with 22.2% recall (F-score 34.1%).

Available systems for biomedical coreference resolution focus on the correspondence between entities and their various expressions in text (**A** and **B** in Fig. 3.1), but little attention has been placed on coreference that involves mentions of events (**C** and **D** in Fig. 3.1).

There have been both rule-based [148, 155, 156], machine learning [157, 158] and hybrid approaches [151] to coreference. Miwa et al. [148] analysed anaphoric coreferences with the help of a set of manually created rules that have been determined through parsing the COREF task training data from BioNLP'11. This solution achieved slightly lower precision (69.8%) than the best system when being evaluated against the BioNLP'11 corpus. Its improved recall (53.5%) led to a high F-score performance (60.5%). Yoshikawa et al. [157] extracted events between frequently mentioned entities, then inferred the coreferential relations based on transitivity. It tested two coreference-based models: a pipeline based on SVM classifiers, and a joint Markov Logic Network (MLN). In the evaluation against the

BioNLP'09 event extraction corpus, both models achieved better performance when applying coreference resolution in comparison to other systems.

While many systems classify pairwise coreference independently from entity mention detection, Song et al. [158] proposes a joint learning model with Markov logic, which combined both processes and outperformed other machine learning systems in the CoNLL-2011 shared task [159]. The best overall performance at the CoNLL-2011 shared task was by a rule based system, the Stanford multi-pass sieve [156], which is a collection of deterministic rules incorporating lexical, syntactic, semantic and discourse information. Lee et al. [151] takes this work further by addressing coreferential relations between both entities and events (see an example of the latter in Sentence **D** in Fig. 3.1) simultaneously. They first cluster documents, then extract mentions of both noun and verb phrases within the same document cluster and treat each mention as a singleton cluster. After the application of rule based filters for entity resolution from [156], they iteratively merge singleton clusters and a linear regressor, trained on gold coreference labels, indicates the best merge at each stage. Finally, another component from [156] is used to resolve pronominal coreference.

3.1.4 Challenges and corpora for event extraction

BioCreAtIvE is one of the challenges that have been well recognised and supported by the TM community. Besides the gene mention and gene normalisation tasks, BioCreAtIvE has also benchmarked the functional annotation of gene products (BioCreAtIvE I, task 2) including the identification of the biological role of molecules. The subsequent BioCreAtIvE II protein-protein interaction task [15] organised by IntAct [160], a molecular interaction database, and MINT [161], an experimentally verified protein-protein interaction database, contained four sub-tasks. These include: *Protein Interaction Article Sub-task 1 (IAS)*, *Protein Interaction Pairs Sub-task 2 (IPS)*, *Protein Interaction Sentences Sub-task 3 (ISS)* and *Protein Interaction Method Sub-task 4 (IMS)*. IAS tests a system's accuracy to spot articles containing PPI descriptions, while IPS assesses the system's ability to extract mentioned PPI in an article. ISS further evaluates whether the system is able to pull all the related sentences for a specified PPI pair. To complement the output pairs from systems in ISS with interaction type, IMS was designed to test whether a system can extract description from text for a given interaction type. In 2009, BioCreAtIvE III's PPI task mainly focused on the detection of relevant articles and linking articles to experimental methods.

BioNLP'09 [38] consisted of three challenges: *Core event extraction (GE)*, *Event enrichment* and *Negation and speculation recognition*. BioNLP'11 [24] retained *GE* as the main task, which covered interactions of gene and gene products including:

- *Gene expression*
- *Transcription*
- *Protein catabolism*
- *Phosphorylation*
- *Localization*
- *Binding*
- *Regulation*
- *Positive regulation*
- *Negative regulation.*

BioNLP'11 added another three main tasks: *Epigenetics and Post-translational Modifications (EPI)*, *Infectious Diseases (ID)*, and *Bacteria Track (BB and BI)*. Except for the main tasks, BioNLP'11 offered three supporting tasks, which are *coreference (CO)*, *entity relations (REL)* and *gene renaming (REN)*. *REL* was for the detection of relations between an entity and its related entity, e.g. a protein and the produced multi-protein complex. The related entities could be metonyms of a protein domain or multi-protein complex. Associating them with entities is helpful for identifying reacting entities and product entities, so as to extract the correct topology and reaction order of a network (see Section 4.1.1). *REN* concerned the detection of bacteria gene homonymies and synonymies. BioNLP'11 corpus included additional five full-text articles to the BioNLP'09 data set.

BioNLP'13 [28] consists of six tasks. The GE (Genia event extraction) task aims to construct a knowledge base by combining TM and semantic web technologies, which also demands systems to be able to tackle coreference. CG (Cancer Genetics) concerns the oncological events in biomedical literature. PC (Pathway Curation) tries to investigate the current TM capability for achieving "automatic pathway extraction", and the potential to apply current event extraction solutions for pathway curation. GRO (Corpus Annotation with Gene Regulation Ontology) provides a corpus to be annotated with the Gene Regulation Ontology to tackle the problems existing in semantic search. GRN (Gene Regulation Network in Bacteria) attempts to evaluate the applicability of TM in bacterial gene regulation network extraction. BB (Bacteria Biotopes) identifies the natural location of bacteria. In summary, BioNLP tasks are moving from fundamental event extraction to providing support for systematic analysis of biological networks.

Another widely used biomedical corpus is the AIMed corpus, which consists of 225 abstracts, within which 25 are not directly describing specific interactions. With more and more available corpora, work on comparing them [162] and work on performing cross-corpora evaluation has gained appreciation, as each corpus has its own emphasis and systems trained on one corpus may not necessarily port well to a different one.

The databases of MIs [160, 163, 164] cross-reference database entries with corresponding publications. Corpora for specific topics can be generated based on thematic clusters of publication. The semantic resources or ontologies of MIs [57, 161, 165–170] provide substantial meta-data, in which verbs could potentially populate trigger sets, and interaction types are useful for determining or ranking an extracted relation, for example, insulin can never be involved with an interaction within a cell.

Each corpus, as discussed above, has been designed for a particular task, e.g. PPI. Evaluation against these corpora gives an indication of the capability of a TM system to address the particular task in question. However, good performance on one corpus does not automatically guarantee the same level of performance on other corpora. A system, trained on a specific corpus in the context of a different corpus, needs to consider running a domain adaptation task [171]. Another important consideration is that many systems are designed for or trained on scientific abstracts, which does not guarantee equivalent performance on full papers. Indeed, abstracts contain condensed information but it is the case that most detailed bio-events and information pertaining to networks is found in the body of the articles, which also contains more noise [172]. Several approaches have combined filters to restrict the types of relations and the location of events to consider for extraction [173, 174]. Zoning of scientific articles in the life sciences such as in the case of [31] could also provide a means of filtering and reducing the noise in full papers for the purpose of event extraction.

3.2 PCorral, interactive mining of protein interactions from MEDLINE

Protein-protein interaction (PPI) is almost involved in all the types of biological process including signal transduction, cellular transportation, metabolism etc. Single PPIs can be interconnected to describe protein interaction networks and complex regulatory events forming the core to the genetic regulation.

Currently, a few TM solutions can identify PPIs from the scientific literature on delivery of a specific gene name to initiate the retrieval: two solutions are, for instance, iHOP [175] and PPI finder [176]. This type of solutions allows exploring the identified PPIs, but the user is limited to navigating through many of the already known PPIs that have been identified at a high frequency rate. This is due to the fact that these systems analyse the complete MEDLINE repository; therefore, the selection is not focused on a specific subset of the literature repository for the curation task. Other tools do allow identifying pairs of entities based on a specific MEDLINE query, and thus these tools enable targeting a specific topic, e.g. FACTA [113], but in this case, the relation extraction is not targeting PPIs; therefore, the curator ends up skimming a large number of entity pairs for PPI mentions. Therefore, the available approaches only partially cover the needs that are required for a complete biomedical curation workflow setup, as they either satisfy the needs of the first step only, i.e. collecting related publications, or the third step, i.e. identifying the parts of a specific interaction.

In more detail, collecting evidential statements for PPI in biological network is usually initiated by accumulating information (called “information retrieval” or IR). In this part, no limitation is put on the gathering process to achieve a comprehensive search and to avoid unnecessary biases linked to any restrictions to the size of the data sample. Subsequently, the document collection has to be narrowed down to focus the results to specific information for example to the identification of relations between entities (called “information extraction” or IE).

We have developed PCorral [40] that combines information retrieval (IR) and information extraction (IE) in a single application. It produces results from different extraction methods in a single approach enabling curators to focus on high recall only, or high precision only in the same processing step. The interactive interface of PCorral supports curation work and interactive exploration of the full set of MEDLINE, and curators may integrate the text processing services from Whatizit into their own curation infrastructure.

3.2.1 Co-occurrence, tri-occurrence and syntactic patterns

Fig. 3.2 gives a schematic overview on the infrastructure and workflow of PCorral, which demonstrates its suitability for the biological curation routine work. The front end of PCorral gathers and organises the results in a tabular view (Fig. 3.5). Using a keyword query interface, the user submits his query and retrieves all the relevant documents from MEDLINE, and then all the documents and statements are processed on-the-fly in a short period, and the extracted findings are delivered to the user.

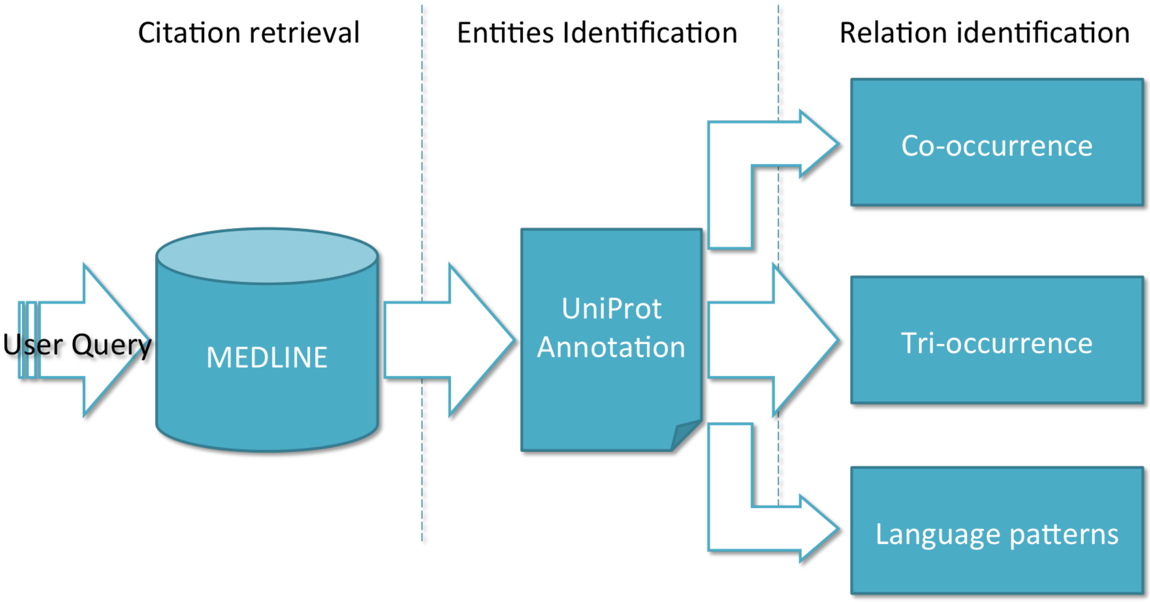


Fig. 3.2: PCorral back end workflow. The processing is split into three main parts: collection of relevant citations querying an index on MEDLINE, identification of gene mentions and normalization to UniProt identifiers and extraction of relations among the identified genes.

The first step in PCorral's workflow consists of collecting publications specified by the user's query; e.g. carotenoid pathway or breast cancer. The articles are retrieved through the MEDLINE index; citations are ranked according to their similarity to the query as determined by Lucene's [177] scoring algorithm. This algorithm identifies which MEDLINE fields, if any, are specified in the query and the syntax of the query, which allows delimiting the terms in the query. Each term is scored according to its relevance to the documents in MEDLINE. The MEDLINE index is the same one used by EBIMed [105] and Whatizit [103], and all the three systems share the same query syntax [178]. The text from the recovered citations is processed to identify sentence boundaries and protein/gene mentions (Whatizit-UniProt), which are then mapped to UniProt identifiers. Basic disambiguation uses the term frequencies from the British National Corpus to distinguish between terms (and entities) that are part of general English in contrast to the specific terminology from UniProtKB [73]. Those terms, which have higher frequency rate than 'insulin', are considered as part of the general English language. They would be difficult to attribute to a specific gene or protein, therefore, will be removed.

PPIs are annotated using three related methods: co-occurrence (CO), tri-occurrence (CO3) and language patterns (SynP). All the three methods solve a specific extraction task (see later in the text) and–according to the specification of the tasks–the results from the three methods form proper subsets of each other: the results from SynP are a subset of the results from CO3, and the same for CO3 in comparison to CO. The first method (CO) is based on co-occurrences and is the same one used in EBIMed. These interactions are based on abstract and sentence level COs. The method delivers the highest recall and is appropriate for exploratory purposes.

The CO3 is more restrictive than the CO method. In addition to two proteins co-occurring in the set, an interaction verb has to be identified from the context of the identified interaction partners. Any triplet of two proteins/genes (PGN) and a verb mention combined in one of the following forms is accepted. In the forms, VP is the verb phrase that represents all the conjugational verb forms and nomVP is the nominalisation of a verb form. Only the pre-selected verbs are considered and, in the case of coordination of two such verbs, both are considered.

- PGN VP PGN
- nomVP PGN PGN
- PGN PGN nomVP

The module that identifies and highlights PPIs searches for phrases that contain a verb or a nominal form describing an interaction like binding or dimerization; the list of verbs is displayed in Table 3.1. The upper set in Table 3.1 comprises all the verbal forms that denote chemical alterations of a protein. The second set of verbs consists of forms that report on interaction and regulation events. "Associate" does not denote any specific binding or transformation event [179].

Table 3.1: List of verbs used in PCorral split into groups defining the interaction type

Verbs denoting protein chemical modification	acetylate, acylate, amidate, brominate, biotinylate, carboxylate, cysteinylate, farnesylate, formylate, "hydrox[iy]late", methylate, demethylate, "myristo?ylate", "palmito?ylate", phosphorylate, dephosphorylate, pyruvate, nitrosylate, sumoylate, "ubiquitin(yl)?ate"
Verbs denoting interaction and regulation events	associate, dissociate, assemble, attach, bind, complex, contact, couple, "(multi|di)meri[zs]e", link, interact, precipitate, regulate, inhibit, activate, "down[-]regulate", express, suppress, "up[-]regulate", block, contain, inactivate, induce, modify, overexpress, promote, stimulate, substitute, catalyze, cleave, conjugate, disassemble, discharge, mediate, modulate, repress, transactivate

If two different verbs have been identified in the context of a gene pair, then both occurrences have been counted. This is also the case for gene pairs that have been identified with syntactical patterns (see later in the text), but this case only occurs at a low frequency.

The approach using syntactical SynPs is more specific, i.e. adds further restrictions to the relation extraction approach in comparison to the solutions of CO and CO3. It extracts PPIs at the highest precision levels but does miss a number of interactions (lower recall). This approach makes use of the following components:

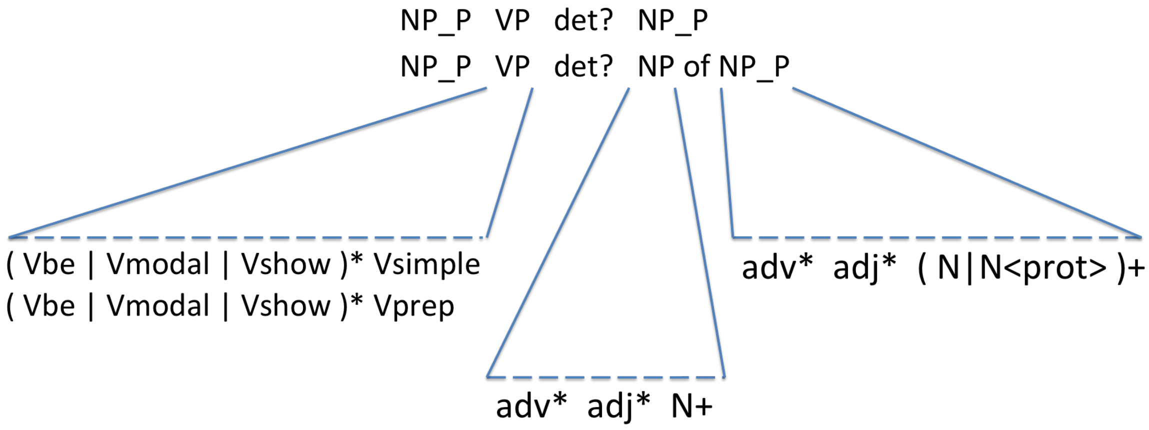


Fig. 3.3: PCorral syntactic patterns. The diagram explains the composition of the SynPs. The VP is composed of several subcomponents that enable the identification of modal verbs (Vmodal), forms of to be (Vbe) and common forms of hedging (Vshown). NP_P is an NP containing a protein mention.

First, one module identifies single adjectives ("adj"), combinations of adjectives and adverbs and the coordination of adverbs. The second module selects the conjugational forms of "to be", also in combination with leading, interleaving and trailing adverbs ("beForm"; see Fig. 3.3). The next module, seeks phrases like "were initially observed" to be combined with "to" and the infinitive of an interaction verb ("shownForm"). In the same sense, modal verbs with optional trailing adverbs, where modal verbs are any of the following: can, could, cannot, do, may, might, must, need, ought, shall, should and would.

Then, the identification of verb phrases is composed of five modules:

- Vsimple covers the verb itself with only optional leading or trailing adverbs.
- Vprep extends Vsimple by a trailing preposition to catch expressions such as "bound to" or "interact with".
- Vbe extends the previous modules by allowing any of the matches produced by the "beForm" stage in front of them and thus targets phrases such as "is regulated" or "are

positively regulated by".

- Vshown allows a match for SynPs that denote expressions like "has been shown" followed by "to" and a match of beForms in front of Vsimple and Vprep. This will tag phrases like "have been shown to be phosphorylated".

- Vmodal works like Vshown but uses a modal verb from the 'shownForm' stage. It will catch phrases like "may be linked to".

Last, the module for noun phrases (NP) identification selects single and multiple nouns in combination with leading adjective modifiers, including coordination of adjective modifier elements leading the sequence of nouns. PGNs are identified as nouns. NPs do not include determiners (e.g. "novel orphan receptor TAK1"). Finally, the module for the PPI syntactical patterns identifies combinations of the previously identified components, such as NP_P VP det? NP_P and NP_P VP det? NP of NP_P, where NP_P is an NP that contains an identified PGN.

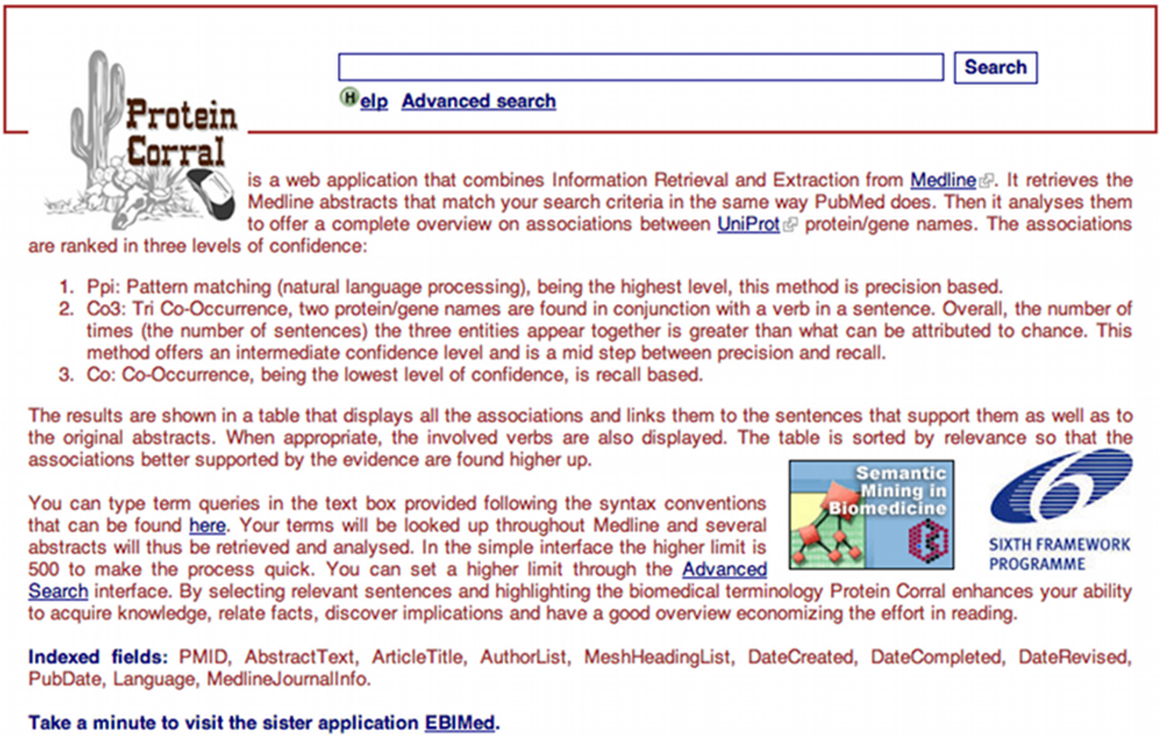
These construction rules for syntactical patterns lead to the selection of structures that are similar to CO3 representations, that form a subset of the CO3 representations and that produce results with highest precision. Similar structures have been proposed by [180]. The syntactical patterns preserve the word order that has been used in the CO3 extraction method, but as additional feature better specifies the verb phrases that are accepted for the extraction of PPIs, and thus generates higher precision results.

Further effort has been spent on the resolution of hedging forms used by authors, i.e. the common use of expressions such as "PGN has been shown to" ('shownForm' syntactical phrase patterns), to increase the recall of the extraction method. In the same vein, the use of syntactical patterns denoting nominalisations improved the recall for the identification of PPIs and follows the representation VP_NP "(of | with | between | through | from)" det? NP_P "(and | with | within | via | through | by)" det? NP_P, where VP_NP is the nominalization of the verb form.

The PPI modules have been assessed using publicly available corpora. Comparative results with a focus to the performance of the different verbs used are available from [179]. The IE pipeline can also be applied as a Whatizit [103] Web service (whatizitProteinInteraction, WhatizitProteinInteractionPMID) for the processing of scientific literature for the identification of PPIs from the text. The system delivers the MEDLINE citations with appended information about the method that identified the PPI, a reference to the matched text and the UniProt identifiers of the related proteins.

3.2.2 Evaluation and result

The simple search of PCorral (Fig. 3.4) interprets a user query to retrieve the documents from MEDLINE that have to be processed. By default, PCorral retrieves the top 500 most relevant citations. Advanced search offers more complex queries to limit or increase the coverage of MEDLINE abstracts for the analysis. In addition, the advanced search allows selecting a specific organism from a predefined list, and this choice restricts the annotation of proteins to those UniProtKB identifiers that belong to the selected organism leading to organism-specific results. EBIMed uses the same approach.



is a web application that combines Information Retrieval and Extraction from Medline. It retrieves the Medline abstracts that match your search criteria in the same way PubMed does. Then it analyses them to offer a complete overview on associations between UniProt protein/gene names. The associations are ranked in three levels of confidence:

1. Ppi: Pattern matching (natural language processing), being the highest level, this method is precision based.
2. Co3: Tri Co-Occurrence, two protein/gene names are found in conjunction with a verb in a sentence. Overall, the number of times (the number of sentences) the three entities appear together is greater than what can be attributed to chance. This method offers an intermediate confidence level and is a mid step between precision and recall.
3. Co: Co-Occurrence, being the lowest level of confidence, is recall based.

The results are shown in a table that displays all the associations and links them to the sentences that support them as well as to the original abstracts. When appropriate, the involved verbs are also displayed. The table is sorted by relevance so that the associations better supported by the evidence are found higher up.

You can type term queries in the text box provided following the syntax conventions that can be found here. Your terms will be looked up throughout Medline and several abstracts will thus be retrieved and analysed. In the simple interface the higher limit is 500 to make the process quick. You can set a higher limit through the Advanced Search interface. By selecting relevant sentences and highlighting the biomedical terminology Protein Corral enhances your ability to acquire knowledge, relate facts, discover implications and have a good overview economizing the effort in reading.

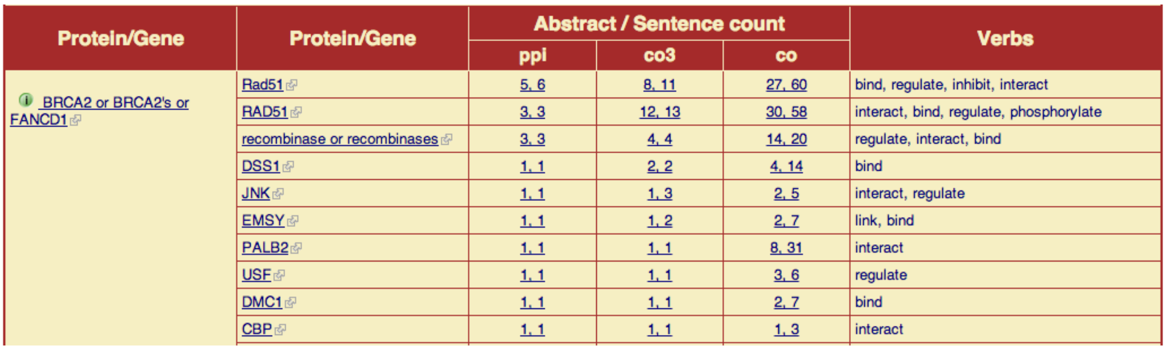
Indexed fields: PMID, AbstractText, ArticleTitle, AuthorList, MeshHeadingList, DateCreated, DateCompleted, DateRevised, PubDate, Language, MedlineJournalInfo.

Take a minute to visit the sister application EBIMed.

Fig. 3.4: PCorral query interface.

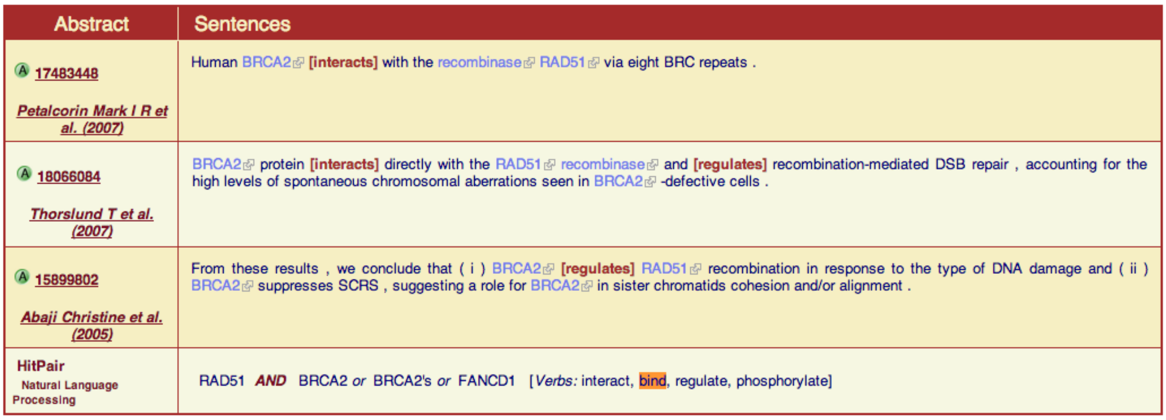
The query interface complies with the document retrieval features that are standards in publicly available search engines, such as PubMed, and follows the specifications of Apache Lucene: e.g. ?AND? and ?OR? queries, keyword mentions and combinations of text features, query language for term and token variability.

Once the citations have been retrieved and fully processed, which may take from only a few seconds up to several minutes (visualized in a progress bar), the interface provides the content as a table containing the extracted PPIs (Fig. 3.5). The list of identified PPI pairs are ranked according to the frequency of the PGN mentions across the whole selected document set, and the most frequent proteins are listed in the top ranks. The related parts of the table show the proteins that the primary protein is interacting with considering the different PPIs extraction methods. The display offers further information such as the frequency counts of abstracts and sentences that make reference to the identified PPIs sorted according to the three methods into different columns. Further information is available for each interaction, as the verb has been identified and displayed that is relevant for the interactions. All the results are interlinked with the underlying biomedical reference databases and also with the MEDLINE documents from which the evidence has been extracted (Fig. 3.6).



Protein/Gene	Protein/Gene	Abstract / Sentence count			Verbs
		ppi	co3	co	
BRCA2 or BRCA2's or FANCD1	Rad51	5, 6	8, 11	27, 60	bind, regulate, inhibit, interact
	RAD51	3, 3	12, 13	30, 58	interact, bind, regulate, phosphorylate
	recombinase or recombinases	3, 3	4, 4	14, 20	regulate, interact, bind
	DSS1	1, 1	2, 2	4, 14	bind
	JNK	1, 1	1, 3	2, 5	interact, regulate
	EMSY	1, 1	1, 2	2, 7	link, bind
	PALB2	1, 1	1, 1	8, 31	interact
	USF	1, 1	1, 1	3, 6	regulate
	DMC1	1, 1	1, 1	2, 7	bind
	CBP	1, 1	1, 1	1, 3	interact

Fig. 3.5: PPI summary table. The screenshot displays in the top ranks those proteins that interact frequently with BRCA2 (using the query “Breast cancer”): amongst all the proteins, RAD51 is most frequently linked to BRCA2 across the selection of documents. The frequency of findings per abstract and per sentence listed for each method is present as well [language pattern (ppi), tri-occurrence (co3) and co-occurrence (co)], including the interaction verbs.



Abstract	Sentences
Ⓐ 17483448 Petalcorin Mark I R et al. (2007)	Human BRCA2 [interacts] with the recombinase RAD51 via eight BRC repeats .
Ⓐ 18066084 Thorslund T et al. (2007)	BRCA2 protein [interacts] directly with the RAD51 recombinase and [regulates] recombination-mediated DSB repair , accounting for the high levels of spontaneous chromosomal aberrations seen in BRCA2 -defective cells .
Ⓐ 15899802 Abaji Christine et al. (2005)	From these results , we conclude that (i) BRCA2 [regulates] RAD51 recombination in response to the type of DNA damage and (ii) BRCA2 suppresses SCRS , suggesting a role for BRCA2 in sister chromatids cohesion and/or alignment .
HitPair Natural Language Processing	RAD51 *AND* BRCA2 *or* BRCA2's *or* FANCD1 [*Verbs:* interact, bind, regulate, phosphorylate]

Fig. 3.6: Example annotation sentences with PPIs. Highlighting of the evidences that allow better identification and curation of the PPIs. Each highlighted protein/gene is linked back to UniProt. Interaction verbs are denoted in square brackets.

In a more comprehensive evaluation, I have analysed which results can be produced from the biomedical literature, namely from MEDLINE abstracts, in comparison with results from full text articles, which are referenced in curated databases. IntAct provides a collection of text from full text articles and the extracted results [181]. These were made available in BioCreative II and can be used for direct comparisons.

In the second evaluation, I have compared the performance of the SynPs considering the different types of verb forms on full text data in comparison with the BioCreative II PPI data set. This evaluation measures the performance of the openly accessible extraction methods against the publicly available benchmark data set.

Table 3.2: Evaluation of COs, CO3, SynP for PPIs on MEDLINE abstracts

Method	Predictions	Correct predictions	Precision (%)	Recall (%)	F-measure (%)
CO	5934	1705	28.73	17.54	21.78
CO3	1461	454	31.07	4.67	8.12
SynP	370	142	38.38	1.46	2.81

Table 3.2 shows the results of running the extraction algorithms on the IntAct text mining corpus [181, 182]. The corpus contains 9,719 manually curated MIs from 1,551 publications. PCorral's extractors are run on the abstracts of the same set of the publications and then compared the extractions of each publication with the same publication's interactions in the corpus. When all the entities of an extracted interaction match those of an interaction from the same document in the corpus, a true positive is counted. With CO method, 17.54% interactions from the corpus are correctly identified, and 28.73% of overall the predictions are correct. The precision increased when the interaction identification was based on CO3, however, with a significant drop on the recall. The extraction based on the SynPs achieved the highest precision, but largely sacrificing the recall.

Table 3.3 shows the results of running the extraction algorithms on the BioCreative II [15] PPI full text sentences. It is found that the recall on full text is higher compared with MEDLINE citations. On the other hand, the precision of MEDLINE information is much higher.

Table 3.3: Evaluation of CO, CO3, SynP for PPIs on the BioCreative II sentences

Method	Predictions	Correct predictions	Precision (%)	Recall (%)	F-measure (%)
CO	52,136	785	1.5	33.2	2.9
CO3	15,823	609	3.8	28.8	6.8
SynP	2078	358	17.2	17.0	17.1

The general performances on both 3.2.2 are relatively low. There are several reasons:

- The entities of many reactions are not occurring in the same sentence. For example, the sentence, “Interaction between two-hybrid positives and different isoforms and variants of PP1c” is associated with 115 IntAct entries. However none of the entries’s both participants are occurring in the sentence.

- Entity names are not aligned with ones used in IntAct database. For example, in the “EBI-1265507” associated sentence, “G9a dimethylates H1K26 and that H1b dimethylated at K26 binds to GST-3MBT”, GST-3MBT cannot be aligned with L3MBT in the IntAct entry.

Since the problem is especially prominent in IntAct corpus, the recall is low when using IntAct corpus. BioCreative II is a corpus particularly designed for extracting PPIs. More reactions’ entities could be found on the same sentences. Then the recall is higher. However, a sentence might contain other unrelated entities. So the precision largely dropped. This problem is compensated by using SynP. So the F-measure in BioCreative II by using SynP is higher than using SynP in IntAct.

Table 3.4: List of verbs that contributed to a correct prediction of related proteins

Verbs	Predictions	Correct predictions	Precision	Recall	F-measure
Regulate	179	12	6.7	0.6	1.0
Contain	286	12	4.2	0.6	1.0
Inhibit	130	9	6.9	0.4	0.8
Mediate	136	7	5.1	0.3	0.6
Activate	165	7	4.2	0.3	0.6
Modulate	31	5	16.1	0.2	0.5
Precipitate	31	4	12.9	0.2	0.4
Express	218	4	1.8	0.2	0.3
Promote	42	3	7.1	0.1	0.3
Induce	110	3	2.7	0.1	0.3
Modify	6	2	33.3	0.1	0.2
Dephosphorylate	8	2	25.0	0.1	0.2
Complex	15	2	13.3	0.1	0.2
Stimulate	41	2	4.9	0.1	0.2
Downregulate	6	2	33.3	0.1	0.2
Methylate	6	1	16.7	0.0	0.1
Substitute	7	1	14.3	0.0	0.1
Assemble	11	1	9.1	0.0	0.1
Block	30	1	3.3	0.0	0.1
Suppress	40	1	2.5	0.0	0.1

CO3 and SynP rely on verbs that have been collected from the research work using different experiments and then published as reference work [179]. I now compare the performance of the different verbs against the content from the corpus to better understand their contributions to the correct predictions (c.f. Table 3.4). Only verbs from Table 3.1 that have contributed to PPI identification in the BioCreative II corpus have been listed in Table 3.4.

Amongst these verbs are the following: upregulate, dissociate, couple, link, overexpress, repress, inactivate, cleave and acetylate. When comparing the list of verbs from Table 3.4 to the proposed verbs from other authors (see Table 3.1), it could be identified that the verbs “downregulate”, “upregulate”, “inactivate” and “stimulate” are relatively less important than “regulate” and “contain”.

The entries in Table 3.4 can be used to optimise the performance of an IE solution, i.e. selection of verbs with a high F-measure to improve the precision/recall ratio of the IE solution and integration of the best performing verbs to improve the overall coverage of the solution. Certainly, more knowledge about the subframe categorisations of the listed verbs will help to further optimise any IE solution and will give contributions to the event identification overall.

3.3 LitWay, a flexible biological event extraction system

Of the three major approaches of event extraction, co-occurrence approach, as a high-recall approach, catches more FP, however, can deliver quantitative information of maximum knowledge, which can be retrieved from articles. On the opposite, syntactic pattern approach, as a high-precision approach, extracts relation with the sacrifice of large amount of true positives. ML approach considers known knowledge about context, syntax etc. and make decision. More importantly, the decision can be tuned according to requirements under different circumstances to achieve a generally good result, or a either precision favoured or a recall favoured result. Biological network evaluation and relation discovery demand a flexible system, which can be used to extract different types of MI and be extended for network extraction. Hence, I developed the LitWay (the **Lit**erature mining system for biological path**Way**) is a configurable pipeline for bio-event extraction.

3.3.1 Modelling event extraction

Extracting biological event is a complex task, which has to correctly recognise entities, their relationships and accurately assemble them together. Each part involves several steps of processes and information from various resources. Appropriate modelling of decomposing the task and defining the workflow mitigates erroneous propagation.

The parts of a biochemical reaction can be defined in several ways. SBML decomposes a reaction to reactants, modifiers and products. However, as discussed in Section 1.2 and Section 1.3.1, it is difficult for TM to get such explicit information of each reaction from text based on current technologies. BioNLP's decomposition about biological or biomedical event is more practical for NLP-based event extraction. In BioNLP schema, *Cause* or *Theme* can be another event, which subsequently constructs a series of consecutive reactions with the previous events. *Trigger* is key word about reaction in text. *Site* is reaction location. Based on BioNLP schema, LitWay extracts bio-event based on the model showed in Fig. 3.7.

3.3.1.1 The system infrastructure of LitWay

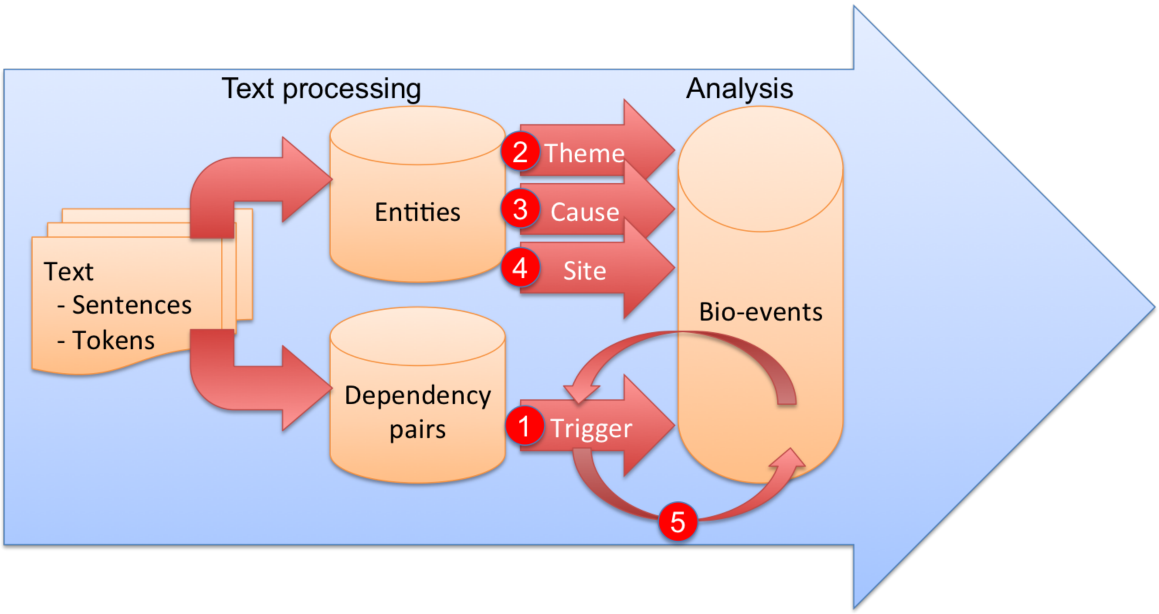


Fig. 3.7: The infrastructure of LitWay pipeline.

LitWay firstly segments text into smaller text units, i.e. sentences and tokens. Sometimes, tokens need to be split into even smaller units, for example, for a token with a hyphen in the middle. The fundamental TM tasks of LitWay adopt the available components in Whatizit. They can be configured to use other tools (See Section 3.3.4). For the BioNLP GE task, LitWay utilises SwissProt tagger for EM, which is similar to PCorral. The syntactic parsing component of LitWay adopts C&C parser [127]. For the BioNLP-ST task, McClosky-Charniak [171] parser is used as it has been provided by the organiser.

LitWay scans through each token in text to identify key words where reactions anchor on. Key words are called *Trigger*s in BioNLP-ST task. The identification is taken charge by a multi-class classifier. Currently, the system is implemented with the classifier based on a search-based structured prediction algorithm, SEARN [183] (see detail in Section 3.3.2). The classifier decides whether the token is a trigger word, and, if it is, which type of reaction the trigger indicates in one prediction. The features currently used for *Trigger* detection are:

- **Text.** The original text of the token.
- **Lemmatised word.** BioLemmatizer [184] is used for token lemmatisation.
- **Stemmed word.** Base or root form of token. It is usually a word's part with inflected or derived parts removed. A number of tools are freely available for the task, for example, the Porter Stemming Algorithm [185].
- **Part-of-speech.** A linguistic category of words.
- **Dependency.** Syntactic dependancies, for example, Stanford dependancies.
- **Bi-gram.** Given token's continuously adjacent tokens in text.
- **Sub-lemma.** For example, in the sentence "critical for nuclear localization and DNA-binding activity", "binding" is the sub-lemma of the token "DNA-binding".
- **Sub-stem.** Similar to sub-lemma, sub-stem is the stemmed, for example, "binding" above.

3.3.1.2 Modelling event extraction

After the pre-process of text including sentencisation, tokenisation, POS-tagging, syntactic parsing, NER etc, LitWay starts the process of recognising relationship among entities or even events. Here I use an example of extracting events in BioNLP-ST to explain the model of event extraction in LitWay.

In the sentence "CpG hypermethylation may account for the absence of IRF-4 expression", two entities are recognised, "CpG" and "IRF-4" (Fig. 3.8). In the experiment, syntactic analysis from the language parser is output in Stanford dependencies [186]. The extracted features deliver the characteristics about the tokens to the classifier, and the classifier tags "hypermethylation" as a trigger of a *methylation* event, "account" as a trigger of a *positive regulation* event, "absence" as a trigger of a *negative regulation* event, and "expression" as a trigger of a *gene expression* event.

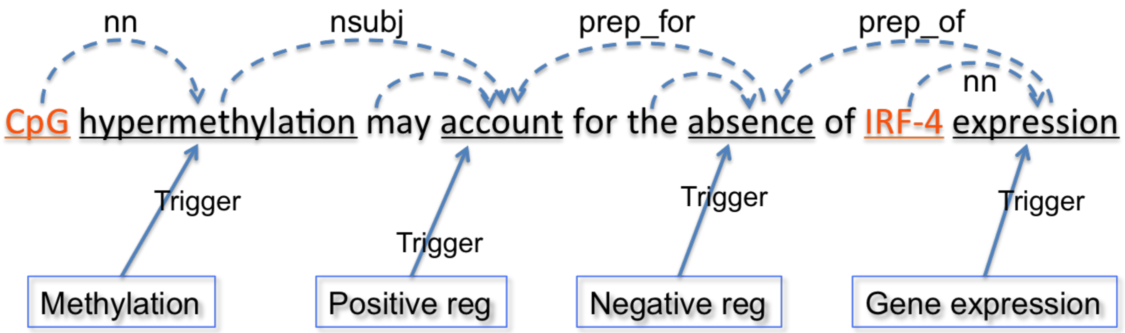


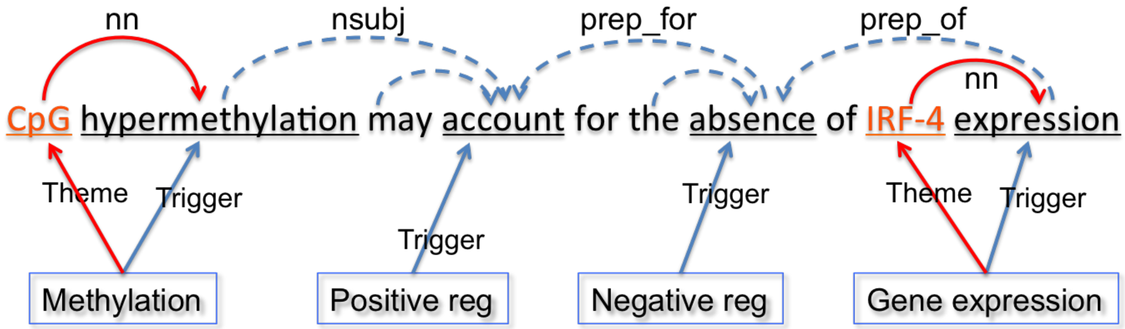
Fig. 3.8: Modelling event extraction (trigger recognition)

Each recognised *Trigger* word is considered as an anchor of a potential event. With the recognised entities and these *Trigger* words, the system tests the entities and assign them to the triggers as their arguments, which literally stand for their bio-relationship described in text. Currently, the event types are categorised into four different groups according to the number of arguments and complexity (Table 3.5). The current categorisation is compatible with BioNLP-ST 2013. The system takes a strategy starting from simpler events. Recursive events are dealt later not only because they are more complex, but also because they may use other events as arguments. Therefore, they are processed at the end when all the other simpler events are recognised..

Table 3.5: The categories of event types

Category	Events	
Simple event	Gene expression, transcription, methylation, protein catabolism, localization	Each event only has one theme, which has to be an entity.
Binding	Binding	Each event has multi-theme and each theme has to be an entity.
Complex event	Protein modification, phosphorylation, ubiquitination, acetylation and deacetylation	Each event has a theme and a cause. The theme only can be an entity, but the cause can be either an entity or another event.
Recursive event	Regulation, positive regulation, negative regulation	Each event has a theme and a cause. Either of them can be an entity or another event.

Theme recognition uses another binary classifier to determine whether an entity or an event is a *Theme*. In Fig. 3.9, “hypermethylation” is recognised as a trigger of a methylation event, a simple event, in the first step. As simple event can only take entity, e.g. protein, as *Theme*, the system calculates the shortest syntactic paths to “CpG” and “IRF-4” respectively and gives them to the classifier. Thus, in Fig. 3.9, “CpG” is recognised as *Theme* of the methylation event. Then, by the same method, the systems identifies “IRF-4” as *Theme* of the gene expression event. There are two simple events extracted and labelled as E1 and E2.

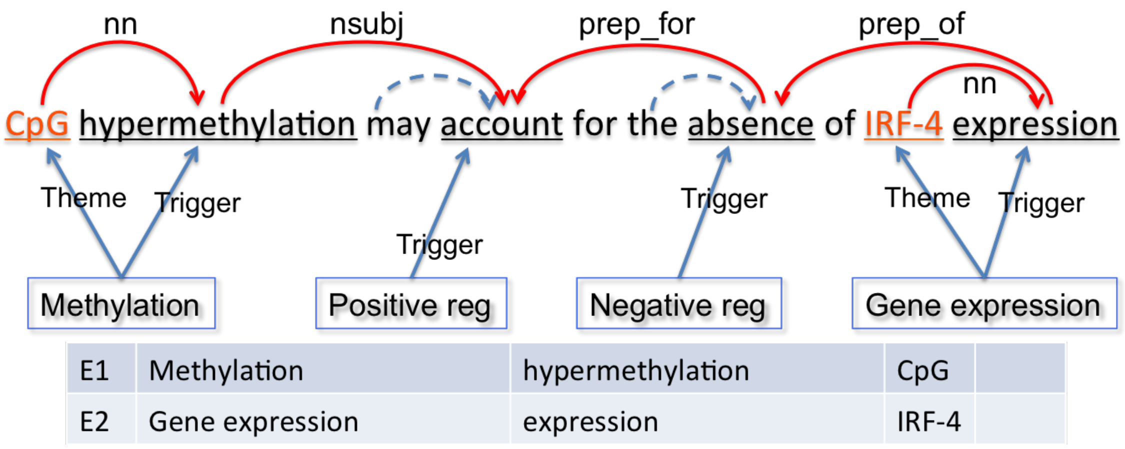


E1	Methylation	hypermethylation	CpG	
E2	Gene expression	expression	IRF-4	

Fig. 3.9: Modelling event extraction (simple events)

In the example, as there is no trigger of *Binding* event detected, the system subsequently starts checking complex events (Fig. 3.10). As binding event is a quite special type of event and requires a separate process due to its multi-*Themes*, the recognition about *Binding* will be discussed lately in this section. In Fig. 3.10, two triggers of complex event were detected, “account” and “absence”. For each trigger, the system uses the same *Theme* classifier to analyse the dependency paths to the two entities in the sentence, and judge whether any of the entities could be the *Theme* of the triggers. In this example, both entities were marked negative as *Themes* of the potential complex events.

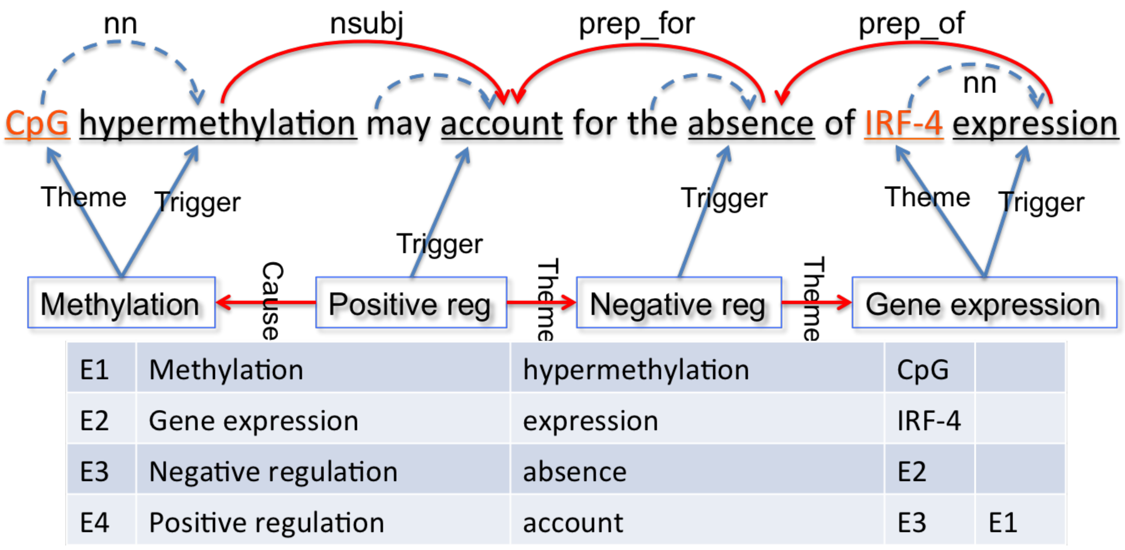
Some complex events could be recursive events, e.g. regulatory events. Recursive events take other events as arguments, either as *Theme* or as *Cause*. Besides regulatory events, many events could be recursive, such as phosphorylation, positive regulation, negative regulation or binding. In the example sentence (Fig. 3.11), there are two triggers of recursive events, positive regulation and negative regulation. The system processes them according to the literal positions from left to right. So far, two events are recognised, E1 and E2. The system firstly checks whether the existing two events could be *Theme* of “account”, and found both negative. Then it checks whether the two events could be *Theme* of negative regulation event with trigger word “absence”, and found the gene expression event, E2 is the *Theme* of the event. So a new event is created and labelled as E3. Because a new event



E1	Methylation	hypermethylation	CpG	
E2	Gene expression	expression	IRF-4	

Fig. 3.10: Modelling event extraction (complex events)

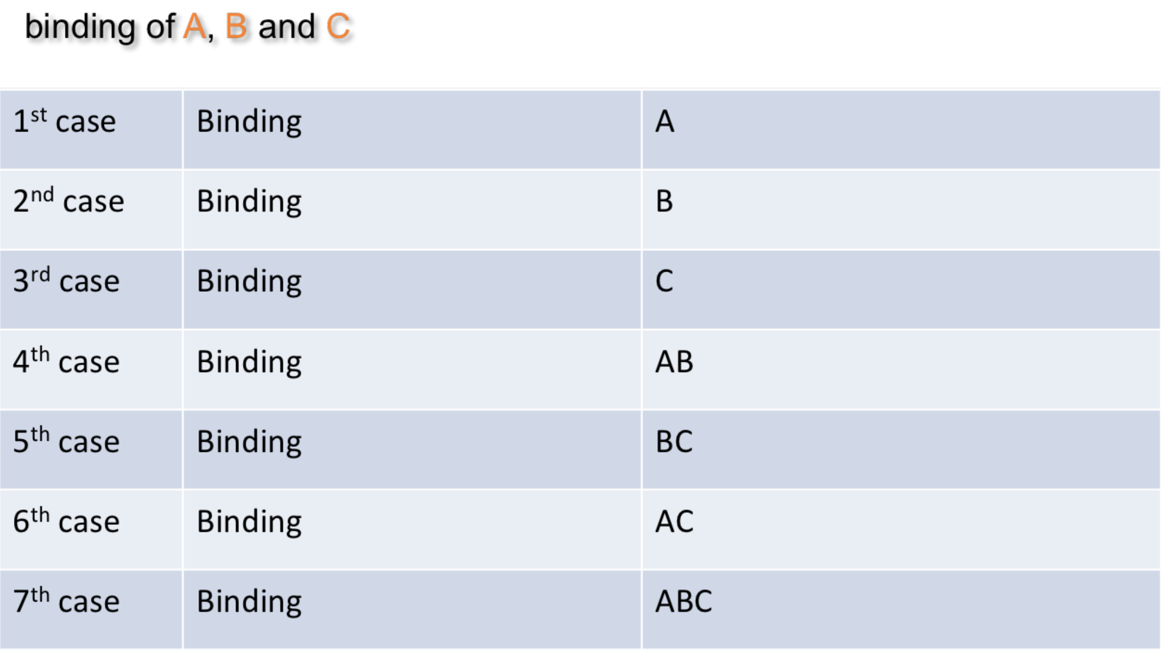
is created, therefore the system has to check all the recursive events again to see whether the newly found event could be *Theme* of any of them, then found E3 is the *Theme* of the positive regulation. Thus another positive regulation event is created and labelled as E4. After all the recursive events' *Themes* are checked and could not produce any new events. The system starts detecting *Cause* for complex events. A *Cause* could be an entity or an event. In the example (Fig. 3.11), the system checks all the complex events and found E1 is the *Cause* of E4.



E1	Methylation	hypermethylation	CpG	
E2	Gene expression	expression	IRF-4	
E3	Negative regulation	absence	E2	
E4	Positive regulation	account	E3	E1

Fig. 3.11: Modelling event extraction (recursive events)

Binding event has to be treated specially, because, for example, a sentence enclosing phrase "binding of A, B and C" has seven different forms of possible binding between the entities, A, B and C (Fig. 3.12). In LitWay, each possible form is created as an independent instance and tested by a binary classifier, which is also based on LIBSVM [187]. Usually, if there is a binding event with wider span, e.g. the 7th case in Fig. 3.12, the system will not consider the binding events with narrower span, e.g. the 1st to the 6th cases. However, the overlap is allowed. For example, the 5th and the 6th can exist at the same time. Thus, the detection creates the permutation of all the possible forms and checks them one by one.



binding of A, B and C

1st case	Binding	A
2nd case	Binding	B
3rd case	Binding	C
4th case	Binding	AB
5th case	Binding	BC
6th case	Binding	AC
7th case	Binding	ABC

Fig. 3.12: Binding event example. For a sentence containing “binding of A, B and C”, seven variant possible binding forms could exist.

The process illustrated from Fig. 3.8 to Fig. 3.12 is based on a very simple example. Nevertheless, the real biological network described in the scientific literature is far more complicated in morphology, grammar and references. Based on current technologies, it is difficult to have a universe system, which can achieve descent performance of detecting all the types of biological events. On another hand, testing a system on different datasets are the way to know the performance. Datasets may be upgraded as well. For example, in BioNLP-ST 2011, a phosphorylation event only has one theme and it has to be an entity. In BioNLP-ST 2013, it can have a theme and a cause, and the cause can be either a protein or another event. Therefore, a flexible system, which can be easily configured and adapted for different tasks with reasonable effort, is more practical (Section 3.3.4).

3.3.2 Search based structured prediction

A biological pathway is a cascade of consecutive reactions, which order is significant for deciding the function of a molecule or even entire system. Naturally, when extracting a bio-event from text, the prediction also influences the down-stream predictions. LitWay is tested with independent classifiers, voted perceptron and support vector machine (SVM). It later adopts a structured prediction algorithm, SEARN [183]. SEARN was used in BioNLP-ST 2011 and showed good performance for structured bio-event prediction [144]. However, BioNLP-ST 2011 challenge is very simplified in comparison with extracting real biological knowledge in scientific literature. It defined nine event types and only had protein as arguments. BioNLP-ST 2013 has more event types and the types' structures are more complicated. Therefore, I implement SEARN and develop a new system with flexibility in mind. The system supports many types of interactions including those between protein-protein and protein-chemical. In addition, the entity type and the reaction types not included at the moment can be added by pre-processing configuration.

The algorithm of SEARN is shown in Algorithm 3.1. Each structured instance $s \in S$ consists of a sequence of T multi-class predictions. The predictions $\hat{y}_1 ... \hat{y}_T$ of the sequence are made according to the current policy h, which is initialised as the optimal policy π at the beginning. For each prediction $\hat{y}_t$, features are extracted from all the previous predictions $\hat{y}_{1:t-1}$) of s and form a feature set Φ_t. Then based on each of other possible actions of the current prediction, the predictions for the rest of the sequence will be made. For each action, the cost will be calculated and the cost vectors of all the possible actions will be appended to the example set E. E will be given to the cost sensitive algorithm (CSCL) to learn a new hypothesis policy h_{new}, which will be incorporated with h by the interpolation parameter β. h will gradually moving away from the optimal policy π. The algorithm can stop after it reaches a threshold or after a pre-setting number of learning.

Algorithm 3.1: SEARN algorithm

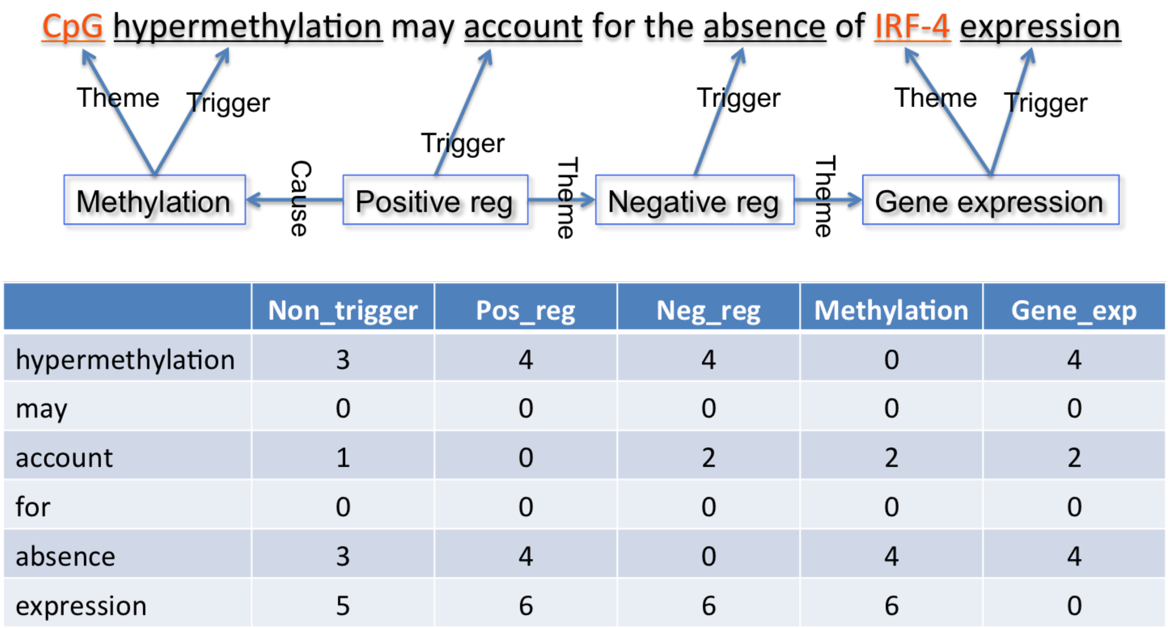
```
Initialise
  Structured instances $S$,
  optimal policy $\pi$,
  cost sensitive learning algorithm $CSCL$
  loss function $\ell$

Train
  current policy $h = \pi$
  while $h$ depends significantly on $\pi$ do
    Examples $E = \emptyset$
```

```
11     for $s$ in $S$ do
12       Predict $h(s) = \hat{y}_1 ... \hat{y}_T$
13       for $\hat{y}_t$ in $h(s)$ do
14         Extract features $\Phi_t = f(s, \hat{y}_{1:t-1})$
15         for each possible action $y_t^i$ do
16           Predict $y'_{t+1:T} = h(s | \hat{y}_{1:t-1}, y_t^i)$
17           Estimate $c_t^i = \ell(\hat{y}_{1:t-1}, y_t^i, y'_{t+1,T})$
18         Add $(\Phi_t, c_t)$ to $E$
19     Learn a classifier $h_{new} = CSCL(E)$
20     $h = \beta h_{new} + (1 + \beta)h$
21   Return policy $h$
```

The task actually needs to recognise a series of bio-events appearing in certain order in text. For sequence labelling tasks, two standard loss functions can be used: whole-sequence loss and Hamming loss. Whole-sequence loss counts a true positive prediction only when the entire sequence is correct. Hamming loss calculates the number of wrong labels. The cost for an action is the difference in loss between taking other action and taking optimal action. There are several methods for calculating cost. [183] lists three of them: Monte-Carlo sampling, single Monte-Carlo sampling and optimal approximation. LitWay uses Hamming loss for loss calculation and optimal approximation for cost calculation. For the same sentence in the previous examples, if use trigger prediction as an example (Fig. 3.13). If the word, "hypermethylation", is not detected as a trigger word, the cost function will count one false negative (FN) in comparison with the gold standard. Besides that, the positive regulation, which trigger word is "account" and has the methylation event as the *Cause*, will be wrong and yield one FP and one FN. The cost for predicting "hypermethylation" as *Non_trigger* will be 3. If it is predicted as a positive regulation, then there will be another FP generated. The cost will be 4. On the analogy of this, only when it is correctly predicted as a methylation trigger word, the cost will be 0. The rest predictions will have *punishing* cost.

Principally, rather than considering each instance independently, SEARN uses features of current and previous predictions as a structured instance. It makes prediction with the consideration of the cost of each possible prediction. Therefore, the learning algorithm used in line 19 has to be cost-sensitive. Daumé III et al. [183] showed that it is possible to use any binary classification algorithm for performing a multi-class cost-sensitive classification.



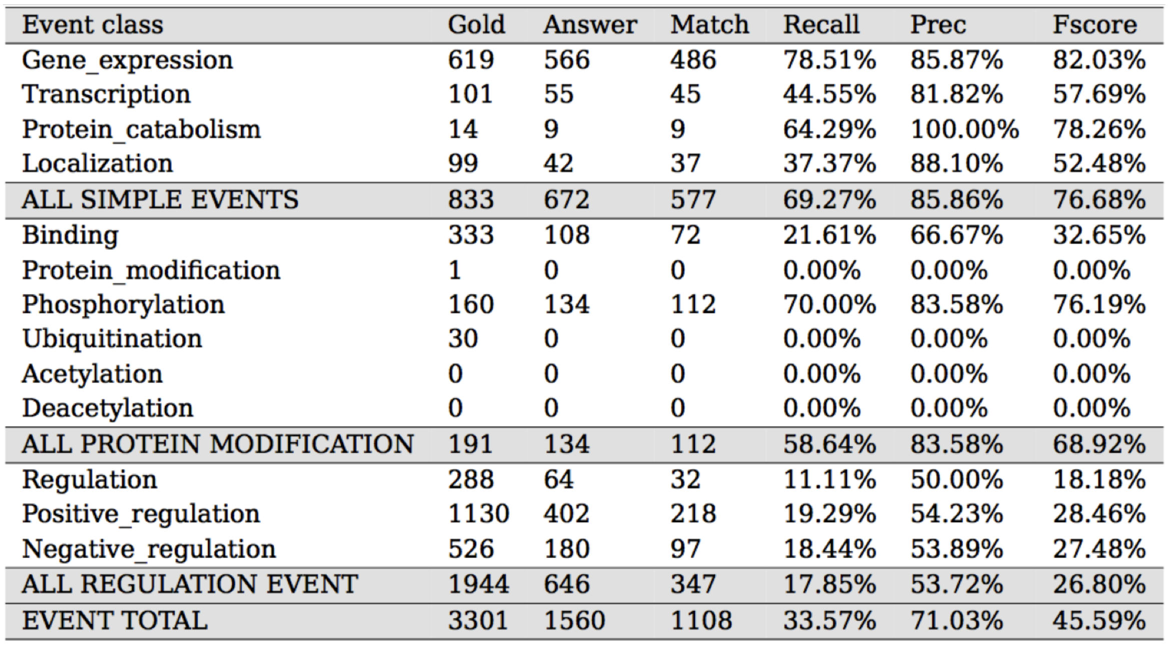
	Non_trigger	Pos_reg	Neg_reg	Methylation	Gene_exp
hypermethylation	3	4	4	0	4
may	0	0	0	0	0
account	1	0	2	2	2
for	0	0	0	0	0
absence	3	4	0	4	4
expression	5	6	6	6	0

Fig. 3.13: Cost calculation in the loss function of LitWay.

3.3.3 BioNLP shared task

The BioNLP Shared Task (BioNLP-ST) is a community-wide challenge incorporating a series of TM tasks for biomedical information retrieval. In comparison with BioCreative, BioNLP-ST more focuses on extracting molecular relationships, which information is more desired for studying molecular functions. Since its first event in 2009, BioNLP-ST has been held in 2011 and 2013. BioNLP-ST 2013 emphasises expressive structured models of extracted information and "high-level" information extraction, for example directionality and polarity of reaction. BioNLP-ST 2013 defines six tasks. They are Genia Event Extraction for NFkB knowledge base (GE), Cancer Genetics (CG), Pathway Curation (PC), Corpus Annotation with Gene Regulation Ontology (GRO), Gene Regulation Network in Bacteria (GRN) and Bacteria Biotopes (BB).

LitWay is tested on GENIA task. GE task is the only task lasted from 2009. Its definition of event structure is closer to reactions in biological network. Each event is associated with an event type, and is explicitly mentioned in the text. Reactant number may vary in each event (Table 3.5). Each reactant has different role, e.g. *Cause*, *Theme*, *Site* etc. Event can play a role in another event. In 2013, it is defined to be much more complicated with much larger data set than it was before. It has more event types and some event types require accurate identification of directionality, polarity, cause, product and location of reaction. Meanwhile, the data set contains much more full-text articles. GENIA task's evaluation



Event class	Gold	Answer	Match	Recall	Prec	Fscore
Gene_expression	619	566	486	78.51%	85.87%	82.03%
Transcription	101	55	45	44.55%	81.82%	57.69%
Protein_catabolism	14	9	9	64.29%	100.00%	78.26%
Localization	99	42	37	37.37%	88.10%	52.48%
ALL SIMPLE EVENTS	833	672	577	69.27%	85.86%	76.68%
Binding	333	108	72	21.61%	66.67%	32.65%
Protein_modification	1	0	0	0.00%	0.00%	0.00%
Phosphorylation	160	134	112	70.00%	83.58%	76.19%
Ubiquitination	30	0	0	0.00%	0.00%	0.00%
Acetylation	0	0	0	0.00%	0.00%	0.00%
Deacetylation	0	0	0	0.00%	0.00%	0.00%
ALL PROTEIN MODIFICATION	191	134	112	58.64%	83.58%	68.92%
Regulation	288	64	32	11.11%	50.00%	18.18%
Positive_regulation	1130	402	218	19.29%	54.23%	28.46%
Negative_regulation	526	180	97	18.44%	53.89%	27.48%
ALL REGULATION EVENT	1944	646	347	17.85%	53.72%	26.80%
EVENT TOTAL	3301	1560	1108	33.57%	71.03%	45.59%

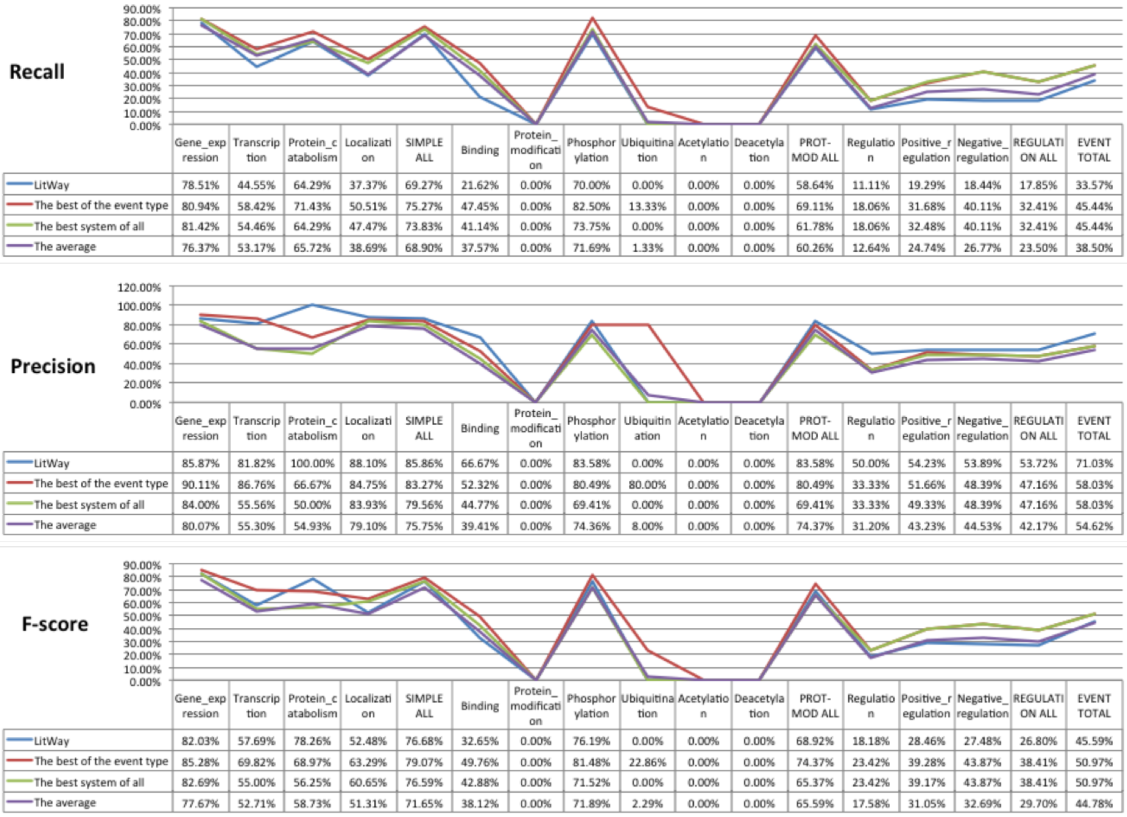
Fig. 3.14: Evaluation of LitWay on GENIA task of BioNLP-ST 2013.

is very strict. For example, if the gene expression event, E2, is wrong or contains error. Then, the negative regulation event, E3, which uses E2 as *Theme*, will be wrong as well. Consequently, the positive regulation event, E4, which uses E3 as *Theme* is wrong too. Under such domino effect, the system will produce three FP and three FN. In such case, making no prediction is even better than doing a wrong one, i.e. making a wrong prediction has higher cost. The structured prediction is suitable for such problem, since it aims to achieve the generally best performance of prediction on each structured instance rather than on a single node. As a structured prediction problem, extract biological networks is also a good test case for structured prediction.

The evaluation shows that LitWay can achieve relatively good performance without coreference (Table 3.14). With competitive recall, it has the best precision for most event types (Fig. 3.15).

3.3.3.1 Constraints and features

Finding key words and assigning arguments are the major and fundamental part of the task. The accuracy at this stage will directly influence the result. Different event types differ from each in instance quantity and appearance in text (Section 4.2.2.2). To investigate the syntactic characteristics of the event types, the experiments applying various constraints and evaluating features are carried out.



	Gene_expression	Transcription	Protein_catabolism	Localization	SIMPLE ALL	Binding	Protein_modification	Phosphorylation	Ubiquitination	Acetylation	Deacetylation	PROT-MOD ALL	Regulation	Positive_regulation	Negative_regulation	REGULATION ALL	EVENT TOTAL
LitWay	78.51%	44.55%	64.29%	37.37%	69.27%	21.62%	0.00%	70.00%	0.00%	0.00%	0.00%	58.64%	11.11%	19.29%	18.44%	17.85%	33.57%
The best of the event type	80.94%	58.42%	71.43%	50.51%	75.27%	47.45%	0.00%	82.50%	13.33%	0.00%	0.00%	69.11%	18.06%	31.68%	40.11%	32.41%	45.44%
The best system of all	81.42%	54.46%	64.29%	47.47%	73.83%	41.14%	0.00%	73.75%	0.00%	0.00%	0.00%	61.78%	18.06%	32.48%	40.11%	32.41%	45.44%
The average	76.37%	53.17%	65.72%	38.69%	68.90%	37.57%	0.00%	71.69%	1.33%	0.00%	0.00%	60.26%	12.64%	24.74%	26.77%	23.50%	38.50%

	Gene_expression	Transcription	Protein_catabolism	Localization	SIMPLE ALL	Binding	Protein_modification	Phosphorylation	Ubiquitination	Acetylation	Deacetylation	PROT-MOD ALL	Regulation	Positive_regulation	Negative_regulation	REGULATION ALL	EVENT TOTAL
LitWay	85.87%	81.82%	100.00%	88.10%	85.86%	66.67%	0.00%	83.58%	0.00%	0.00%	0.00%	83.58%	50.00%	54.23%	53.89%	53.72%	71.03%
The best of the event type	90.11%	86.76%	66.67%	84.75%	83.27%	52.32%	0.00%	80.49%	80.00%	0.00%	0.00%	80.49%	33.33%	51.66%	48.39%	47.16%	58.03%
The best system of all	84.00%	55.56%	50.00%	83.93%	79.56%	44.77%	0.00%	69.41%	0.00%	0.00%	0.00%	69.41%	33.33%	49.33%	48.39%	47.16%	58.03%
The average	80.07%	55.30%	54.93%	79.10%	75.75%	39.41%	0.00%	74.36%	8.00%	0.00%	0.00%	74.37%	31.20%	43.23%	44.53%	42.17%	54.62%

	Gene_expression	Transcription	Protein_catabolism	Localization	SIMPLE ALL	Binding	Protein_modification	Phosphorylation	Ubiquitination	Acetylation	Deacetylation	PROT-MOD ALL	Regulation	Positive_regulation	Negative_regulation	REGULATION ALL	EVENT TOTAL
LitWay	82.03%	57.69%	78.26%	52.48%	76.68%	32.65%	0.00%	76.19%	0.00%	0.00%	0.00%	68.92%	18.18%	28.46%	27.48%	26.80%	45.59%
The best of the event type	85.28%	69.82%	68.97%	63.29%	79.07%	49.76%	0.00%	81.48%	22.86%	0.00%	0.00%	74.37%	23.42%	39.28%	43.87%	38.41%	50.97%
The best system of all	82.69%	55.00%	56.25%	60.65%	76.59%	42.88%	0.00%	71.52%	0.00%	0.00%	0.00%	65.37%	23.42%	39.17%	43.87%	38.41%	50.97%
The average	77.67%	52.71%	58.73%	51.31%	71.65%	38.12%	0.00%	71.89%	2.29%	0.00%	0.00%	65.59%	17.58%	31.05%	32.69%	29.70%	44.78%

Fig. 3.15: Comparison between LitWay and other systems participating GENIA task of BioNLP-ST 2013.

Constraints

Fig. 3.16 shows the performance with different constraints independently applied in the experiment. Six different constraints are tested.

- Concatenate with lemma: Values of other features are concatenated with lemma.
- Remove dependency path to protein: Dependency paths to involved proteins are not used.
- Use dummy word to substitute real entity name: Replace entity names with the same word, which doesn't have to have meaning. In the experiment, all proteins are replaced by "PROTEIN".
- Remove obj, subj in sentence: obj and subj are removed from dependency paths.
- Create equivalent link for "or": Equivalent dependency paths are added on the tokens of both sides of "or".

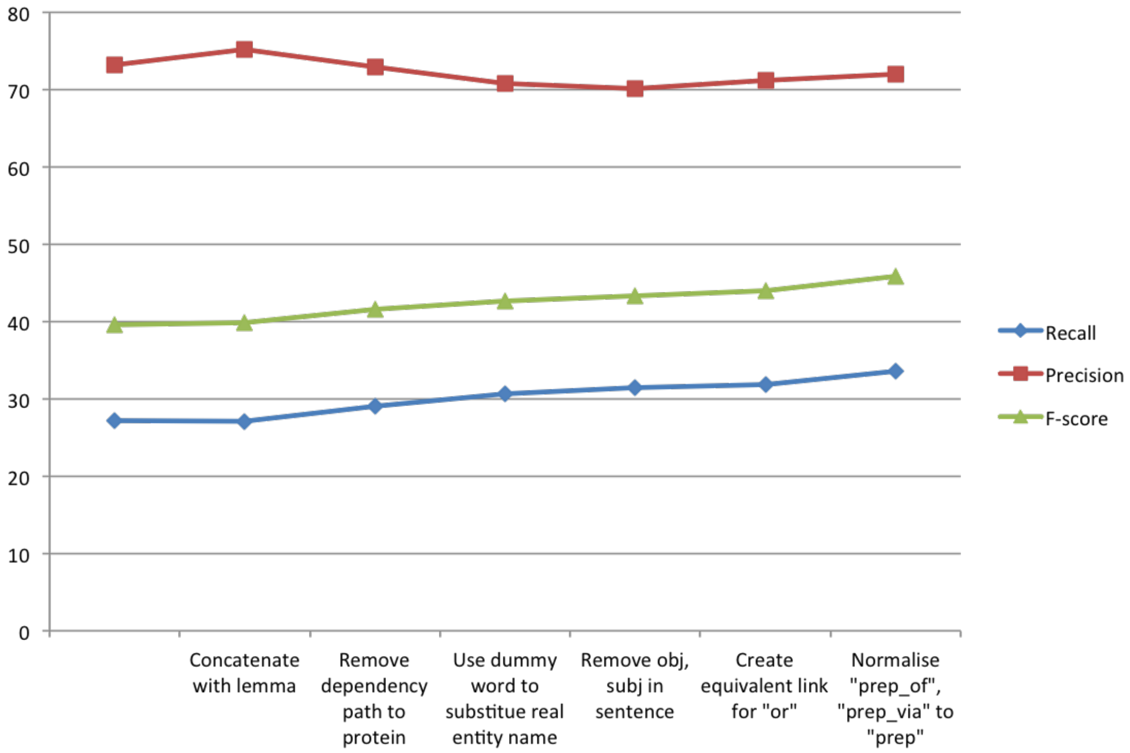


Fig. 3.16: Constraints' influence to the performance of the task.

- Normalise “prep_of”, “prep_via” to “prep”: The same dependency paths are created for the three with “of” or “via” removed.

The precision has obvious increase when other features are concatenated with lemma. This indicates that lemmas play important role in identifying essential words. In a rule-based event extraction system, lemmas have the similar function as a set of controlled vocabulary. It lowers the weight of contextual information for determining triggers, so the precision slightly falls. Nevertheless, the general F-score gains a little increase.

After removing the feature of connecting biological entity, e.g. protein, the performance gains a clear improvement mainly benefitted from the recall. This lifts the restriction on the prediction. Unsurprisingly, the precision sacrifices. Actually, replacing entity word with the same fake name is even better than simply removing them. Although precision suffers bigger loss, the general performance is well compensated by high recall. If the object and subject labelled by syntactic parsing are removed, the performance increases when recall is released by such restrictions. A general performance increase can be gained if syntactic links are created on the both sides of conjunctions, “or” and “and”. Normalising specific syntactic relation with prepositions can also increase the general performance.

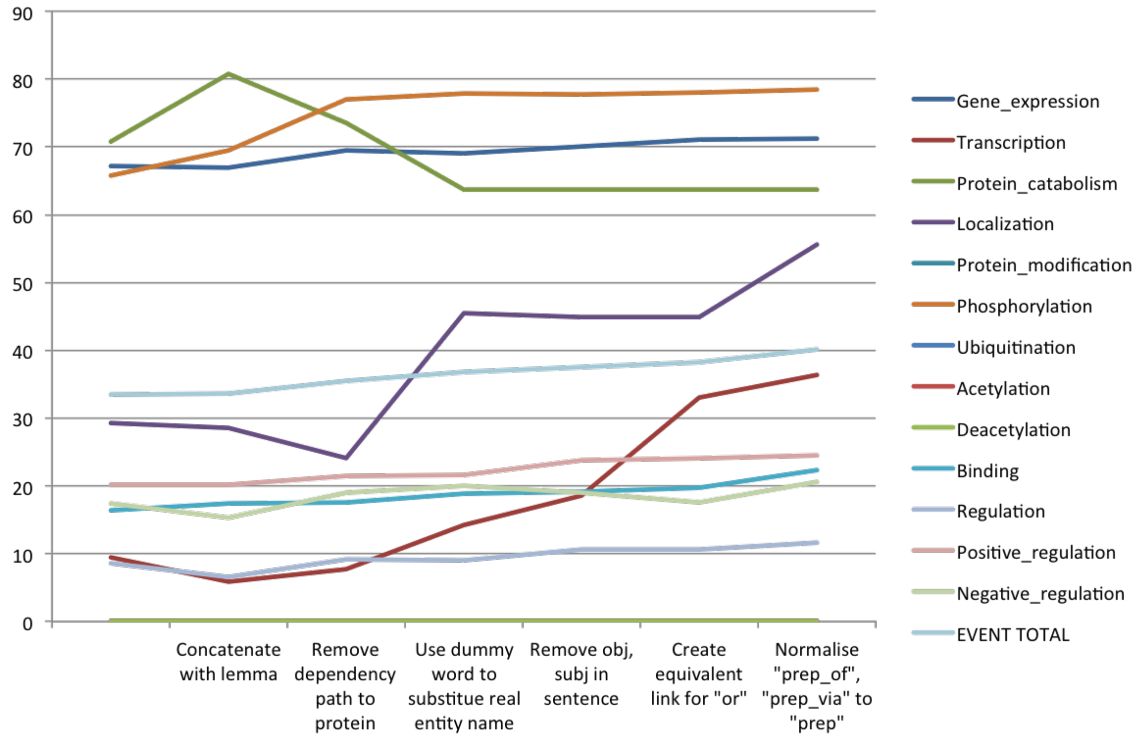


Fig. 3.17: Constraints' influence to the performance of each event type in the task.

Do the constraints deliver the same benefits or restrictions to all the event types? Fig. 3.17 breaks down the previous analysis for different event types. It can be seen that different event types benefit from different textual characteristics delivered by different features. For example, when features are concatenated with lemma, although the F-score increases as discussed in the previous paragraph, its benefit to different event types largely varies. Protein catabolism gains significant increase when transcription suffering big loss. The impacts to the other event types also vary with some performances increasing and some losing. Another constraint causing dramatic opposite effect is to substitute biological entity with the same fake name. Protein catabolism's performance drops significantly. This is probably because that, as an essential process in digestion, protein catabolism is frequently described with a set of enzymes or other substances in text. The weaker contextual information brought by substitutions of such enzymes may cause lower recognisability. Information indicating a localisation event less relies on contextual information, so localisation event gains an increase.

Features

Features and properties of machine learning based systems are usually in black box, which hinders the understanding of the prediction behaviour. Frequently, systems are re-

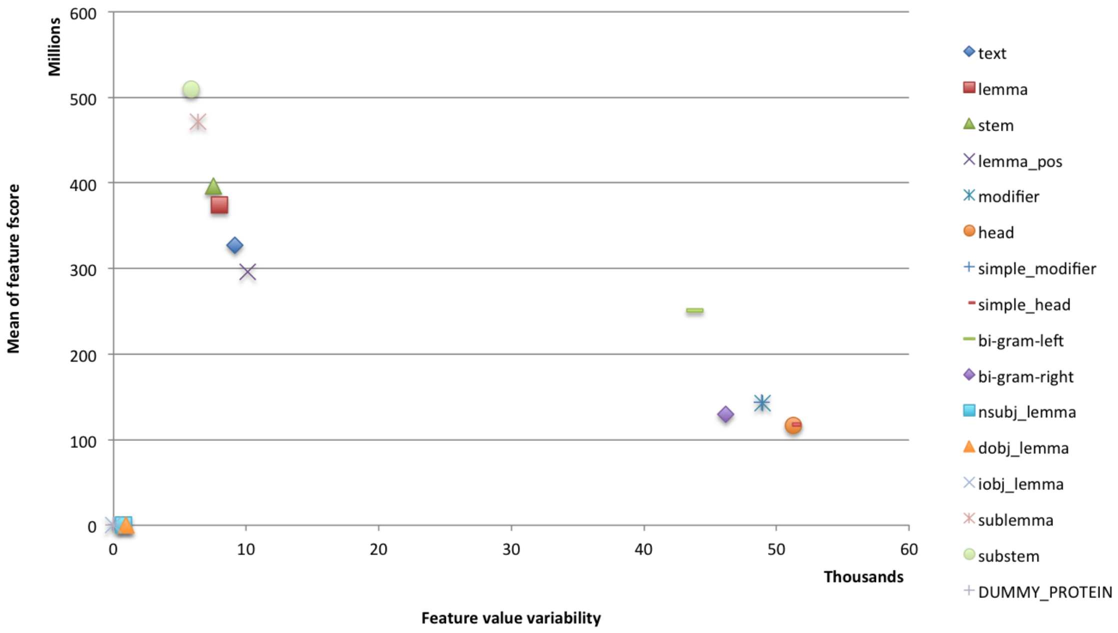


Fig. 3.18: Value variability and influence of the features. The horizontal axis in the above chart is the numbers of the distinct values of the features. The vertical axis is mean of feature's F-score. The experimented features are: text (literal word of token), lemma (lemmatised word of token), stem (stemmed word of token), lemma_pos (combination of lemma and part of speech), modifer (modifier in token's dependency), head (head in token's dependency), simple_modifier (simplified modifier), simple_head (simplified head), bi-gram_left (left token), bi-gram_right (right token), (nsubj_lemma) (lemmatised nsubj tokens), iobj_lemma (lemmatised iobj tokens), sublemma (lemmatised part after hyphen in token), substem (stemmed part after hyphen in token) and DUMMY_PROTEIN (a dummy word indicating existence of co-occuring protein)

peatedly calibrated towards specific tasks to achieve good performance without the insight of the relevant features for bio-event extraction. In this section, a feature selection experiment is carried out on the features of trigger word recognition by using LIBSVM feature selection tool [188]. Trigger recognition is to locate the key word reporting a reaction's existence and type. The features used for the experiment are word, lemmatised word, stemmed word, grammatical dependencies, bi-gram etc (Fig. 3.18 and 3.19). LIBSVM feature selection tool is used to test individual features and calculate corresponding F-scores based on training data. By plotting the feature value variety and F-scores each feature type can get on a cartesian coordinate system, we can know each feature's potential contribution to F-score and how various its value could be.

The horizontal axis in Fig. 3.18 is the number of unique features in each feature type. The vertical axis is the average F-score of all features in each type. The chart again clearly states the importance of lemma in key word identification of event extraction. The feature types in the order of sub-stem, lemma concatenating POS, stem, lemma and original text are the top five. The feature types have the least various values are nsubj_lemma, iobj_lemma, dobj_lemma and DUMMY_PROTEIN. Simplifying modifier does not make big difference. Therefore, modifier and simple_modifier are overlapping on the chart. DUMMY_PROTEIN is a word replacing all protein occurrences. So its value variety is very low and close to the origin of the coordinate system

Fig. 3.19 is similar to Fig. 3.18, however plots all the individual features. This assists building more accurate system by distinguishing the importance of the features. For example, the lemmatised word is usually regarded the most important information for trigger detection. However, bi-gram-left surprisingly has very high F-score. This is because almost all of the protein catabolism occurring as "PROTEIN degradation" in training data have nearly identical forms in the test data. If the features are all listed and tested by using the same method (LIBSVM feature selection tool), it can also be see that “lemma_degradation_lemma_PROTEIN” is the top feature (Table 3.6). Table 3.6 lists all the top 10 features, in which texts, lemmatised words and stemmed words are top feature types. Therefore, for trigger detection, the literal word is the most important than others, such as modifier and *n*-gram.

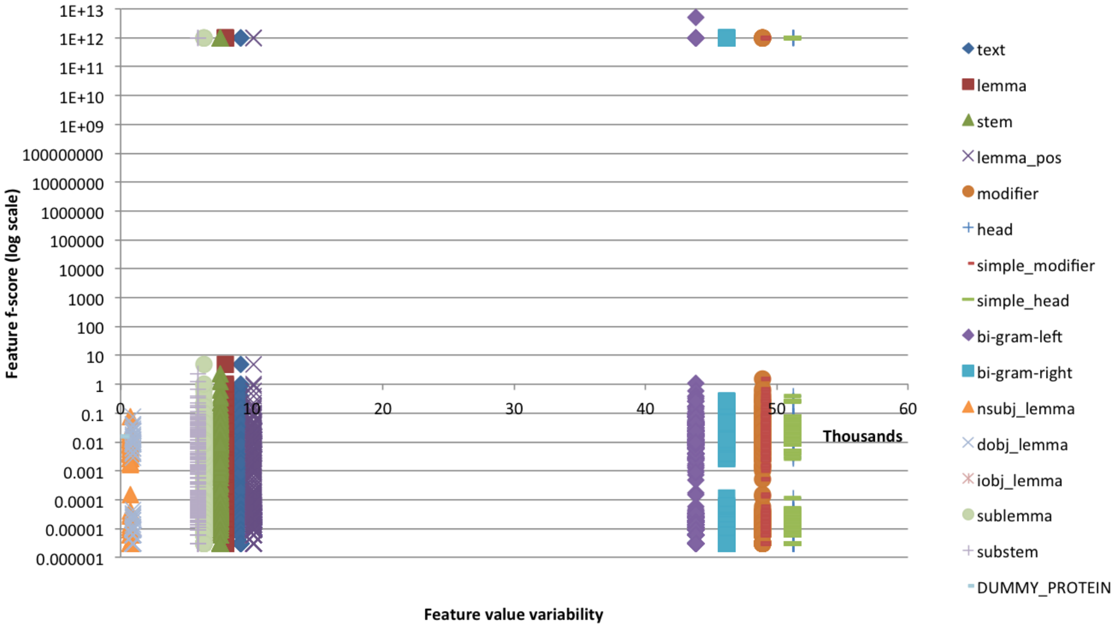


Fig. 3.19: Feature value variability and influence to the result. The horizontal axis in the chart is the numbers of the distinct values of the features. The vertical axis is F-score of feature. It differs from Fig. 3.18 that the vertical is actual F-score of each feature instead of the mean of them.

Table 3.6: Top feature

lemma_degradation_lemma_PROTEIN
text_acetylation
text_modification
text_deacetylation
lemma_modification
lemma_acetylation
lemma_deacetylation
stem_acetyl
stem_modif
stem_deacetyl

3.3.4 A flexibility bio-event extraction system

Biological event extraction is usually a complex pipeline converging different information produced by different tool, and assemble generated result and external information to construct bio-event. The development takes substantial effort, and is difficult to maintain. To evaluate a system, datasets from community-supported challenges are used. However, the datasets may be upgraded. For example, BioNLP-STs GE task has been held in three challenges since 2009. The definitions of the events in the task have largely evolved since the first challenge. Many events are defined to be much more complicated. On another hand, various biological networks have different entity composition and the involved reactions differ from each other in quantity and morphology, which demand tailored settings for different circumstances.

LitWay is built with flexibility in mind. To minimise the effort of adaptation and modification for new requirements, the functions are modularised into the components in an infrastructure. Several unstructured data infrastructures are available for the community and provide good modularisation assistance, such as Apache UIMA [42] and General Architecture for Text Engineering (GATE) [91]. The two can be combined by still are different in many ways [189].

In LitWay, Apache UIMA is used as the scaffold of the system. Information, including bio-entity annotation, syntactic parsing etc., is converged and converted to UIMA annotations before further processing. Output is also as UIMA annotation, thus has potential to be encoded into any formats. Generally, the flexibility of LitWay has two aspects.

First, the components can be easily changed. For example, the system currently uses McClosky-Charniak parser [171] for the syntactic analysis. Different parser has different advantages. In case another parser is needed, the syntactic parsing module can easily incorporate another parser or a combination of parsers [190] into the system.

Another aspect about the system's flexibility is the easy adaptation for extracting different types MI. The system works upon a workflow, which guides its workflow of fetching bio-entities, examining entity types, identifying event argument, and assembling event etc. This workflow is encoded in a configurable xml file. The system reads the configuration file and takes corresponding actions.

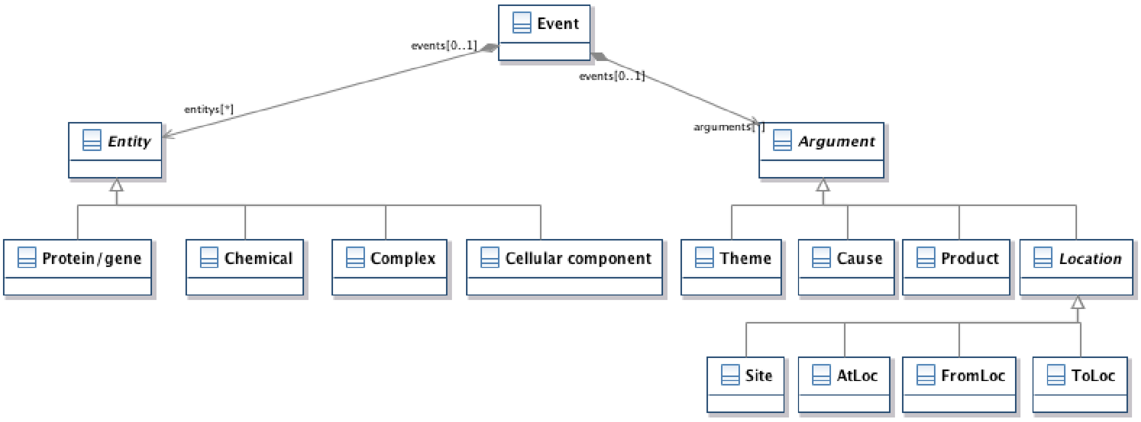


Fig. 3.20: The class diagram of BioNLP-ST pathway curation task.

The system was tested on BioNLP-ST GE task. However, if taking BioNLP-ST pathway curation (PC) as an example, the basic components can be abstracted into two classes, entity and argument, which are the same as GE's (Fig. 3.20). Differing from GE, PC extends the two classes to more types of bio-molecules, which are *protein/gene*, *chemical*, *complex* and *cellular component*. All the arguments in PC, including *Theme*, *Cause* and more, derive from argument class. GE task involves even less types of arguments.

As this can be generalised to many types of biological event extraction tasks, the configuration file is actually created based on this abstract class design. I have defined an XML schema, by following which LitWay is able to read and understand configuration files of tasks as far as the configuration files comply with the schema.

Fig. 3.21 is a snippet of the XML schema defined and used by LitWay. It is correspondent with the abstract classes in the diagram (Fig. 3.20). The top level of the schema is network standing for a biological network. In the first section within the network, entity types are defined. For example, for a PPI network, maybe only one type, protein, needs to be defined. But for other more complicated networks, like the networks in the Pathway Curation task of BioNLP'13, simple chemical, complex and cellular component can be defined in the section. The second section within network is for defining event types. In each event,

```
<xs:element name="network">
  <xs:complexType>
    <xs:sequence>
      <xs:element name="entities">
        <xs:complexType>
          <xs:sequence>
            <xs:element name="entity" type="xs:string" minOccurs="1"/>
          </xs:sequence>
        </xs:complexType>
      </xs:element>
      <xs:element name="events">
        <xs:complexType>
          <xs:sequence>
            <xs:element name="event" minOccurs="1" maxOccurs="unbounded">
              <xs:complexType>
                <xs:sequence>
                  <xs:element name="type" type="xs:string"/>
                  <xs:element name="arguments">
                    <xs:complexType>
                      <xs:sequence>
                        <xs:element name="argument" minOccurs="1" maxOccurs="
                            unbounded">
                          <xs:complexType>
                            <xs:attribute name="name" type="xs:string"/>
                            <xs:attribute name="type" type="xs:string"/>
                            <xs:attribute name="multi" type="xs:boolean"/>
                          </xs:complexType>
                        </xs:element>
                      </xs:sequence>
                    </xs:complexType>
                  </xs:element>
                </xs:sequence>
              </xs:complexType>
            </xs:element>
          </xs:sequence>
        </xs:complexType>
      </xs:element>
    </xs:sequence>
  </xs:complexType>
</xs:element>
```

Fig. 3.21: The snippet of XML Schema of configuration file. The highest level is network. In the first section, entities, within network, entity types could be defined. The second section, events, is for defining event types.

type is the name of the defined event, e.g. phosphorylation etc. arguments may be defined to include Theme, Cause etc. Type of each Theme or Cause may be an entity type defined in entities section or another event type. The complete schema can be downloaded from [191].

Fig. 3.22 is an example of GENIA task's configuration file. In the part of the configu-

```
<entities>
  <entity>Protein</entity>
</entities>
<events>
  <event>
    <type>Gene_expression</type>
    <arguments>
      <argument name="Theme" type="Protein" multi="false" />
    </arguments>
  </event>
  <event>
    <type>Protein_modification</type>
    <arguments>
      <argument name="Theme" type="Protein" multi="false"/>
      <argument name="Cause" type="Protein" multi="false"/>
      <argument name="Cause" type="Gene_expression" multi="false"/>
      <argument name="Cause" type="Transcription" multi="false"/>
```

Fig. 3.22: The snippet of the configuration file of BioNLP-ST GE task.

ration shown in the figure, one entity type is defined, which is Protein. Two event types are defined, Gene_expression and Protein_modification. Gene_expression only has one entity type, which is Protein, as the Theme. Protein_modification has both Theme and Cause. Its Cause take other two events, Gene_expression and Transcription, as arguments. This example can be downloaded from [192].

3.4 Discussion and conclusion

MI is another layer of information for evaluating biological networks. Event extraction solutions with state-of-the-art performances are needed for retrieving MI information from unstructured text.

PCorral is a system designed for extracting PPI from the scientific literature. The EM and EN components of PCorral are state-of-the-art, since they use SwissProt tagger, whose performance has been demonstrated in the previous chapter (Section 2.3). PCorral provides an interactive access for collecting evidential statements of PPI from IR to IE. The IR part is based on CO2 and is used to determine the literature set. The IE part is based on CO3, even more precise SynP, to determine the explicit statements of PPIs. CO2 approach used by PCorral collects a larger set of evidential statements, which are the true superset to the CO3 results. At the same time, CO3 results are again a larger set and the true superset to the SynP's. The result shows that PCorral produces high recall results and also more specific results. It processes request and generates result on the fly, which may not be able to be achieved with LitWay.

Between high-recall and high-precision, LitWay uses ML to extract MI by analysing grammatical characteristics produced by syntactic parsing. The evaluation shows that LitWay is one of the up-to-date systems, which have the best precision among the participating systems of BioNLP'13. Meanwhile, LitWay can be customised for extracting other types of MIs by changing a configuration file.

Different bio-event types need different treatments when their evidential statements are extracted from the scientific literature. One reason is the knowledge discrepancies introduced by different NERs, GSCs and terminologies, which has been proved in Chapter 2. In the next chapter, I will further investigate the problem and explore the answer by evaluating and quantifying the entities and reactions of the biological networks from curated database.

Chapter 4

Biological network evaluation and relation discovery

Key points

- BMDB covers a broad thematic range of biological networks. All networks are curated from the scientific literature and carefully evaluated to be correspondent with the original articles.
- Aligning contents of BMDB quantifies compositions of different entity types and coverages of reactions in biological networks in the scientific literature.
- Differences of quantity and syntactic morphology of proteins and chemical entities in the scientific literature may lead to different strategies of network extraction. The differences have not been investigated by the research community. The experiment shows overlaps between protein and chemical, and between PCI and PPI.
- The entity compositions of the signalling pathways and the metabolic pathways in BMDB are in-line with the empirical knowledge. The signalling pathways involve more proteins, and the metabolic networks involve more chemical entities. However, signalling pathways also have a large portion of chemical entities.
- The entities and the reactions in BMDB are aligned with evidential statements collected from the scientific literature. Generally, PPIs have a better coverage in the scientific literature in comparison with the CPIs. The signalling pathways have more coverage in the literature concerning entities and relations. The metabolic pathways are less covered.

- Network sizes do not have significant correlation with the coverage in the scientific literature.

- Hypothesised molecular relations from RECON2 are evaluated by LitWay. Mention frequencies reflect confidences of hypothesised relations. The result also support analyses of bio-molecular activities in different cellular components and sub-pathways.

Extraction of bio-entities and bio-events provides information for different layer of network. Currently, there is no work linking TM-extracted information with curated biological networks. Aligning curated networks against information retrieved by TM is a way to bridge the gap between structured knowledge and unstructured text. It also delivers the TM community the knowledge about quantity and morphology of the biological networks in the scientific literature.

In this chapter, by exploiting the up-to-date solutions I developed, I evaluate the biological networks from a curated database by knowledge extracted from unstructured text. The detail of a part of the work has been published in [193].

4.1 Biological network extraction and evaluation

An extracted event is a molecule's relationship with another. However, fully understanding such a relation and analysing molecular functions require putting the relation in a context, e.g. a network. A network needs to be topologically accurate with consecutive interactions. Some systems extract related events based on specific entities and connect the events into a network on the basis of common entities [117, 119]. MIs can be catered for by this kind of system, but the sequence of interactions remains relatively unclear. PathwayFinder [18] extracts pathways with the help of user intervention. During extraction, it provides an interface through which a user can influence the result of NER and modify syntactic patterns. GENIES [17] is the only system which is reported to automatically extract a pathway from a full-text article. It uses a grammar consisting of semantic patterns interleaved with syntactic and semantic constraints to identify relevant relationships and to specify target outputs.

Although the technology for extracting complete networks is limited, evaluating networks with event information extracted from the scientific literature still semantically enrich networks for biological study, meanwhile delivers morphological information about reactions and entities in the scientific literature for TM.

4.1.1 Event interconnection

While TM so far has been focusing on event extraction, the potential use of TM in event interconnection is slowly emerging through the notion of event coreference [151, 158]. Slowly but steadily TM is moving from event extraction to network extraction, climbing up the pyramid in Fig. 1.1.

Intuitively, events could be connected together based on common entities. However, solely depending on the common entities would require flawless entity coreference and normalisation. Apart from the diversity of entity names, for example, it is not uncommon that a generic term is used as a metonym of another mentioned term, e.g. glucose versus α-D-glucose. In this case, entities can be heuristically merged according to semantic similarity. This method is also effective for merging similar reactions [151, 158, 159, 194].

In BioNLP [24, 38], an event could appear as a *Theme* or a *Cause* of another event. This reflects the succession and recursion of a biological process. Some approaches statistically link co-referring events together. FAUN (Feature Annotation Using Nonnegative matrix factorization) [195] extracts gene term features and constructs them into a matrix. By using different features and weighing each term, the authors treat it as a clustering problem of functionally related genes, addressed by matrix factorization with non-negative normalisation (NNMF) in FAUN. Bio-LDA (Latent Dirichlet allocation) extracted 13,338 terms for compound, gene, drug, disease, pathways and side effects. Relationships between concepts were hypothesised based on a LDA model [196].

An important aspect of bio-events which causes the problem for entity identification and coreference, and subsequently network extraction, is the fact that many entities (products) resulting from a particular biochemical reaction are omitted or implied in the text and only later referred to, usually by referential metonymy. For example, an issue resulting from this is the difficulty in recognising multi-protein complexes (see Section 1.3.1), which are not specifically named or introduced in the text (e.g. during a binding event) and may be referred to using the name of the proteins they contain. These also point to weaknesses in the usage of the nomenclature, especially for multi-protein complexes, as generated complexes are functionally different entities, but are referred to by the name of the receptor protein. By contrast, products of chemical reactions are unambiguously named. Without properly identifying or inferring the products of biochemical reactions, the generated incomplete network cannot faithfully reflect the mechanism of a pathway, and molecular roles in the pathway.

4.1.2 Hidden relation discovery

If a molecular relation found in the scientific literature using TM is not referenced in a bioinformatics database or knowledge base, it is either judged as a FP or as an unknown relation (incomplete DB). However, such cases may point to the existence of relations, which hold the potential of new discoveries using biomedical TM. While TM cannot and could not replace biological experiments for the discovery of biological relations, it can nevertheless provide evidence about possible MIs and help hypothesise relations.

There are two types of such relations. The first type covers relations that are explicitly mentioned in the publications, but are missing from curated databases. Explicit mentions of relations in text extracted using TM can be used to populate or enrich databases, although manual curation of the results may be required to make sure they comply with database standards. This thesis focuses on the first type of hidden relation.

The second type of relation is not represented as a direct interaction in the text, however, it can be inferred from other relations which are mentioned in the text, in combination with knowledge in existing knowledge resources. Another possibility is to infer such relations by means of combining text with experimental evidence or by finding tentative relations in the text, which could provide evidences for yet uninterpreted experimental findings [197]. For example, after optimising a logical model about the responses of human cells to seven cytokines by using gene expression data, a direct link between IGF1 and Akt without PI3K's activation has been predicted [198]. The prediction questions the current understanding about the necessity of PI3K for activating Akt in the pathway of regulating glucose uptake. In fact, high-throughput data generated a large number of such predictions, and suggested that, in vitro, a protein may play more roles than what has been perceived. While experimental evaluations covering all the predictions are expensive and impractical, evidence automatically collected from the literature can support and prioritise further experiments. In the above example this second type of relation cannot be extracted or predicted explicitly based on NLP and TM methods only. Nonetheless, such relations can be identified through statistical heuristics [196], with the help of unsupervised machine learning methods [195] or with a combination of TM and logical inference [199].

In addition to the above, the MIs or relations currently already in the knowledge bases would strongly benefit from relation discovery for evaluation and annotation purposes. More specifically, biological database curators would like to be able to link curated data with relations in the publications as the latter have more detail and more context illustrated in a more explicit way [86], especially when figures are alongside. The available event extraction solutions can be implemented as software and integrated into routine biological workflows [12].

4.2 Aligning the TM solutions with the contents of the biological networks in BMDB

In this section, a series of experiments, including using PCorral and LitWay, are conducted to evaluate knowledge extracted from unstructured text against the curated biological networks. The methods used range from high-recall driven (co-occurrence and tri-occurrence) to high-precision driven (syntactic pattern or machine learning). The experiments quantitatively reveal entity compositions of different types of biological networks. It also characterises morphology and availability of different types of reactions in the scientific literature, especially those of metabolic networks and signalling pathways.

4.2.1 Method

4.2.1.1 Unstructured text

The abstracts from MEDLINE are extracted, in which protein and chemical entities are tagged by utilising two Whatizit components, SwissProt tagger and ChEBI tagger [103]. The corpus contains 19,961,467 abstracts. The ChEBI tagger utilises Whatizit infrastructure, and searches for chemical entities in text based on terminology from ChEBI database [70].

4.2.1.2 Structured databases of curated biological networks

Choice of curated reaction networks for the experiment is made in terms of network diversity and data accessibility. To easily access and process data, networks have to be encoded in a standard descriptive format. There are several community-supported formats, including SBML, CellML and BioPAX etc. They have similar functions, however differ from each other. SBML describes many different biological networks and is the de facto standard for representing computational models in systems biology today [5].

Meanwhile, the chosen database should cover a wide range of biological network types. Furthermore, the networks of the chosen database should be annotated with related semantic and supplementary information to disambiguate involved biological concepts including molecules and reactions during evaluation.

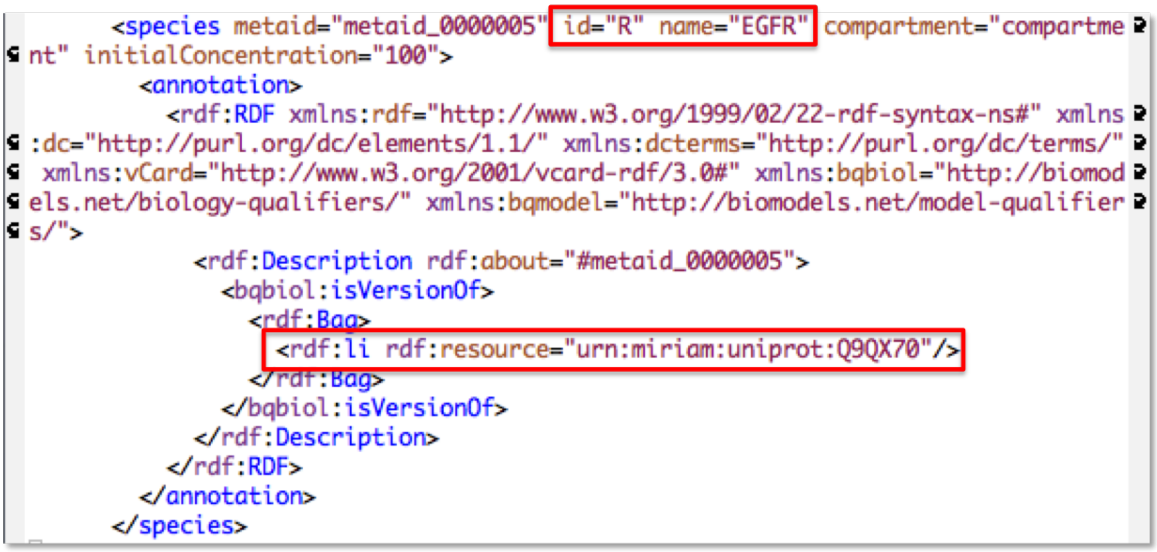
Several public databases are available for the experiment. They are BioModels Database (BMDB) [8, 11], CellML [6], Reactome [2] and KEGG Pathway etc. The experiment use the biological networks from the curated branch of BMDB, because

- All the data is freely available;
- All the networks in BMDB are encoded in a standardised and machine readable format;
- Each network has been manually curated according to the same guidelines [8], on the basis of each individual corresponding publication, to assure the network fully and faithfully represents the reactions described in the publication.
- The database possesses a large amount of manually added annotations which disambiguate involved biological molecules and semantically enrich the reactions and the models. These annotations are linked with the model elements with a set of qualifiers, such as *isPartOf*, to denote the relationships between the annotated elements and the resources used to annotate them.
- It is a standard resource for networks ranging extensively from signal transduction to metabolic pathways and other types of biological reaction networks, like circadian rhythm;
- The database's content is comprehensive and does not emphasise on any particular organisms or species. The biochemical reaction networks in the database include signalling pathway, metabolic pathway, circadian rhythm, cell cycle, response to stimulus, channel activity, and homeostasis etc. The network types in BMDB will be further discussed in Section 4.2.2 with the illustration in Fig. 4.3.

In BMDB, the networks are encoded in SBML along with quantitative information of the reactions, such as concentration, volume, area or mass etc. In scientific publications, quantitative information is frequently presented in figures, which requires figure mining techniques. As this study focuses on quantity and morphology of entities and reactions in the scientific literature rather than simulating them with kinetic information, the quantitative information in BMDB has not been considered.

4.2.1.3 Metadata and annotations in the curated networks

In the curated biological networks, there are two forms of information, metadata and annotation. They are encoded along with the corresponding SBML elements in a scheme relying on the use of RDF, Dublin Core, vCard and the BioModels.net qualifiers (see Fig. 4.1).



```
      <species metaid="metaid_0000005" id="R" name="EGFR" compartment="compartme
nt" initialConcentration="100">
        <annotation>
          <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns
:dc="http://purl.org/dc/elements/1.1/" xmlns:dcterms="http://purl.org/dc/terms/"
 xmlns:vCard="http://www.w3.org/2001/vcard-rdf/3.0#" xmlns:bqbiol="http://biomod
els.net/biology-qualifiers/" xmlns:bqmodel="http://biomodels.net/model-qualifier
s/">
            <rdf:Description rdf:about="#metaid_0000005">
              <bqbiol:isVersionOf>
                <rdf:Bag>
                  <rdf:li rdf:resource="urn:miriam:uniprot:Q9QX70"/>
                </rdf:Bag>
              </bqbiol:isVersionOf>
            </rdf:Description>
          </rdf:RDF>
        </annotation>
      </species>
```

Fig. 4.1: An SBML snippet containing metadata and annotation. The entity in an SBML file represents Epidermal Growth Factor Receptor. In the upper red box, the entity is unambiguously named as EGFR, which is the metadata. In the lower red box, the entity is annotated with UniProt:Q9QX70, which is the annotation. More information about an entity, such as synonyms etc., can be exploited by NER solutions by tracking its annotation in external references.

- **Metadata**

  Nearly every entity comprising an SBML model has an id and a name (see the example in upper red box in Fig. 4.1). They are often named with a protein or a chemical entity's scientific name. In the experiment, id and name of an entity are used as its metadata. According to the SBML specification, neither id nor name is obligated to literally contain a biological term or acronym. Sometimes, they could appear as identifiers for computing purposes, such as *entity1...n*.

- **Annotation**

  Annotations are manually added in the need to disambiguate an entity and enhance its semantic information (see the example in the lower red box in Fig. 4.1). In BMDB, annotations are compliant with an annotation guideline, MIRIAM (Minimum Information Required In The Annotation of Models) [200, 201]. When an annotation is used in the experiment, besides its entry identifier, the names and synonyms of the entry are also retrieved from external resources.

All biological networks in BMDB are available in SBML. In SBML, a metadata set of an element includes the name of a biological entity the element represents. At the same time the metadata may contain other information, such as authorship and modification date etc. Entity names in metadata can be used to directly align with mentions of PGN or chemical names in text. Each element in SBML is also associated with a set of annotations, which link the network element with an accession in external references. By following the references, the element's scientific name, synonyms can also be extracted and used to align with entities in text. The alignments are bidirectional. When aligning an entity name from SBML metadata with a text unit (e.g. a token or a phrase), the text unit will be tagged as the entity's occurrence if the entity name exactly matches the text or matches it with certain similarity, e.g. cosine similarity. On another hand, if an entity in the text is recognised and normalised by a NER solution, and the identifier given by normalisation equals an entry in a database or has certain semantic similarity with the entry, the recognised entity in text and the entry in the database are also a match.

4.2.1.4 Annotation qualifier, entity type and network type

In BMDB, each annotation is linked with a network element by a BioModels.net qualifier (see detail about BioModels.net qualifiers in [202]), which is discriminated in the experiment, because different qualifiers differentiate an element's nature even if it is associated with the same reference in the annotation. For example, in an annotation, "is CHEBI:29108" means that the referred element in the network *is* Calcium, which accession number in *ChEBI* is *29108*. "hasPart CHEBI:29108" means only a part of the referred element is Calcium. The detail about BioModels.net qualifiers can be found on its webpage [202], and will not be described in the thesis. It is important for the experiment to treat the annotations differently according to their qualifiers.

Thus, in this experiment, the elements are only considered as possessing the equivalent attributes as the referred annotations, when the qualifiers is one of *bqbiol:is*, *bqbiol:isVersionOf*, *hasVersionOf*, *isHomologTo*, *isEncodedBy* and *bqbiol:hasPart*. The order of the qualifiers is significant during the process. *bqbiol:is* is with the highest priority, as it explicitly denotes the equivalence between the model element and the biological concept represented by the annotation. *bqbiol:hasPart* is also considered conditionally according to the annotation. For example, if the annotation denotes that the model element "hasPart UniProt:P62158", which means that the represented model element contains human calmodulin protein molecule as part of its macromolecule, then the represented element is considered as a compound protein.

Two types of entities are considered in the experiment. An entity will be considered representing a protein when it is annotated with entries from UniProt, InterPro, Protein Data Bank, Pfam, PROSITE, Microbial Protein Interaction Database or PhosphoSite Residue. An annotation is considered representing a chemical entity when it is from ChEBI, PubChem-substance, PubChem-compound, ChEMBL compound, ChEMBL target, CTD Chemical, or Chemical Component Dictionary.

Based on entity composition, the networks in BMDB are categorised into five groups listed below. PPI networks, chemical reaction networks and CPI networks are grouped based on the networks' involved entity types. Signalling pathways and metabolic pathways are picked out by the annotations and studied as two groups because they are the majority of all the networks in BMDB (Fig. 4.3).

- PPI network. The networks' entities are only PGNs.
- Chemical reaction network. The networks only involve chemical entities.
- CPI network. The networks' entities are a mixture of PGNs and chemical entities.
- Signalling pathway. See definition in Section 1.3.1.
- Metabolic pathway. See definition in Section 1.3.2.

As BMDB is an open source project, which source code and database are available from SourceForge [203]. I create a local MySQL instance replicating the BMDB's metadata and annotations for the experiment.

4.2.1.5 Network size categories

Network size may influence the statistical result and result in biased conclusion. For example, it is not fair to say a 100% recovered network with three entities is better recovered than a 60% recovered network with twenty entities. In the experiment, the networks are analysed upon three respective categories based on the number of involved entities.

Small network has five or less entities involved. Medium network's entity number is between five and twenty. When involving more than twenty entities, network is considered as large network. 5% networks do not have any entities involved (Fig. 4.2), which may be the models representing a biological status, e.g. membrane current (BIOMD0000000020). 25% networks have less than five entities, when 27% networks involve more than twenty entities. The majority entity number of the networks is between five and twenty. The following analysis is also classified based on the same network size category.

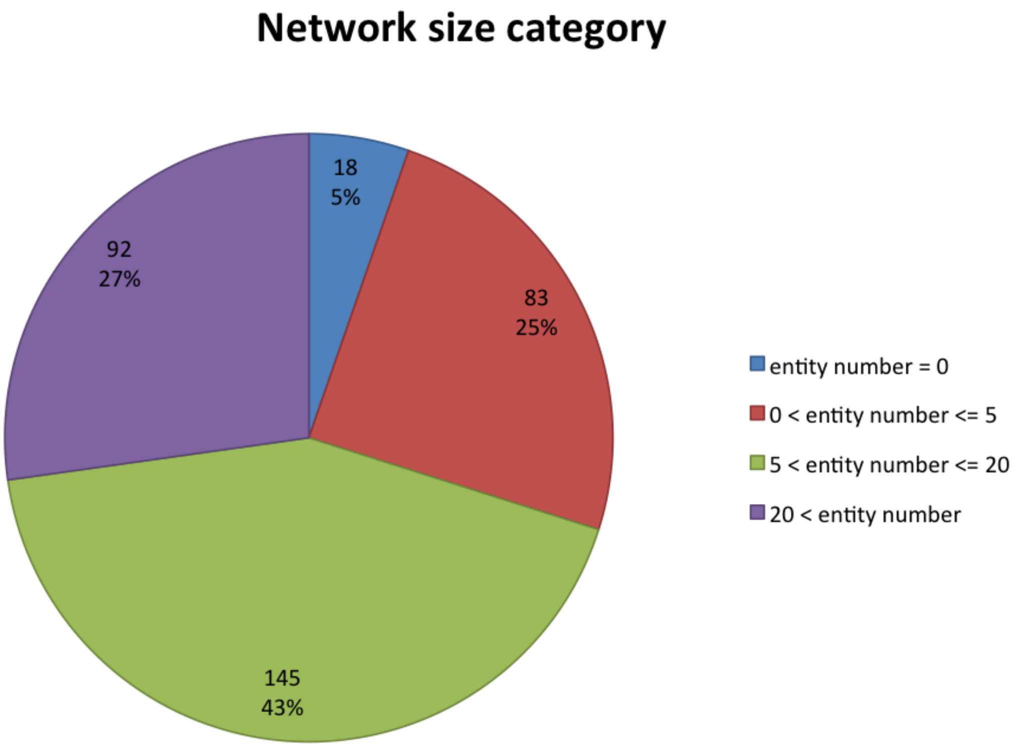


Fig. 4.2: BMDB's network size analysis. The blue sector is the number of the networks not involving entities. The red sector is the number of the networks having 5 or less entities. The green sector is the number of the networks having less than 21 and greater than 5 entities. The purple sector is the number of the networks having more than 20 entities.

4.2.2 Result

Signalling pathways, metabolic pathways and other types of networks, such as regulatory network, are highly integrated in cellular activities. However, analyses usually only focused on independent processes. In this study, each type of reaction networks is characterised for their entity compositions, reaction coverages etc. Then the networks are compared across different types. The characterisation delivers distinct features of each type of the networks.

Protein and chemical entities are major molecules involves in biological networks. Their

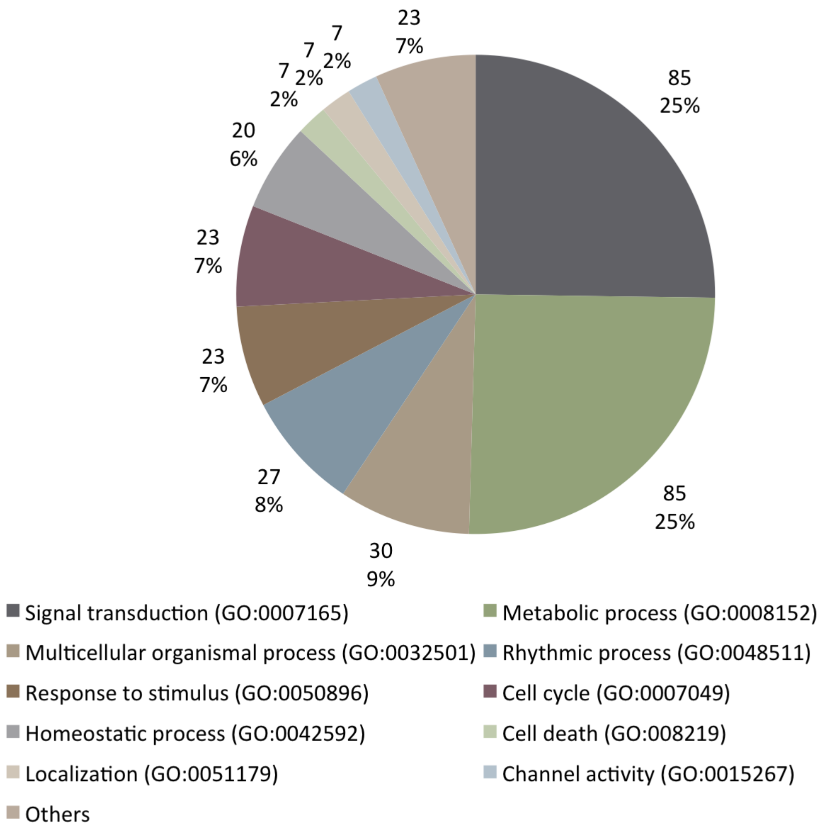


Fig. 4.3: BioModels Database network types. Each sector is the number of a type of biological networks in BMDB. The networks are categorised by using their GO annotations, which are in the parentheses after each type of the biological networks.

NER technologies are relatively mature. In the experiment, the networks from BMDB are categorised into PPI network, CPI network and chemical reaction network. Meanwhile, the signalling pathways and the metabolic pathways are also studied separately because they are the focus of many biological researches. According to its 26th release, half of the biological networks in BMDB are signalling pathways and metabolic pathways (Fig. 4.3). Other networks including cell cycle, circadian rhythm etc. This is calculated by using the manually added GO annotations in the networks. A network is determined as a signalling pathway if its annotation refers to *signal transduction* (GO:0007165) or any descendant processes of *signal transduction*, like *apoptotic signalling pathway* (GO:0097190).

4.2.2.1 Entity composition in BMDB

Different networks involve different types of entities. It has been long presumed that signalling pathways mainly involve proteins and metabolic pathways mainly involve chemical entities. The presumption is mostly based on experience, however, is not supported by any direct evidences. Especially for such networks extracted from the scientific literature, the entity composition directly influences the decision of choosing right NER and event extraction for different biological networks. The networks in BMDB are all extracted from the scientific literature and well checked to be correspondent with the published knowledge. In certain extent, the statistic upon the networks can reflect the entity compositions of different types of networks in the text.

The entities types and the network types can be determined by using the annotations (see the method in Section 4.2.1.3). Then, all the proteins and the chemical entities involved in the networks are distinguished and counted (see how the entity types are determined in Section 4.2.1.4).

In the four charts in Fig. 4.4, each dot represents a reaction network in BMDB. The top left chart illustrates the respective numbers of the proteins and chemical entities occurring in each network. The horizontal axis is the number of the proteins in each network, and the vertical axis is the number of the chemical entities in each network. The top right chart represents the proportions of the proteins and the chemical entities in each network. Its horizontal axis is the percentage of the proteins in a network, and the vertical axis is the percentage of the chemical entities in a network. In both charts, a red dot represents a signalling pathway determined based on its manually added annotation. A green dot represents a metabolic pathway. A blue dot represents a network, which involves biological processes of both signal transduction and metabolic pathway. If a network involves neither a signalling pathway, nor a metabolic pathway, it is represented by a black dot in chart.

The bottom left chart presents the percentage of proteins and chemical entities in each signalling pathway. The bottom right chart shows the percentage of proteins and chemical entities in each metabolic pathway. In the both bottom charts, the horizontal axes are the total number of entities involved in the pathways. The vertical axes are the percentage of the proteins or chemical entities involved in the pathways, which are respectively represented by different colours of dots. An orange dot represents a network, which protein percentage is plotted against the vertical axis. A purple dot represents a network, which chemical percentage is plotted against the vertical axis.

The top left chart in Fig. 4.4 illustrates the respective numbers of proteins and chemical entities in each network. It can be seen that many metabolic pathways do not involve any proteins in the reactions, and many signalling pathways involve relatively less chemical entities in comparison with their proteins. Although signalling pathways and metabolic pathways are integrated and influential to each other in biological processes, however, the

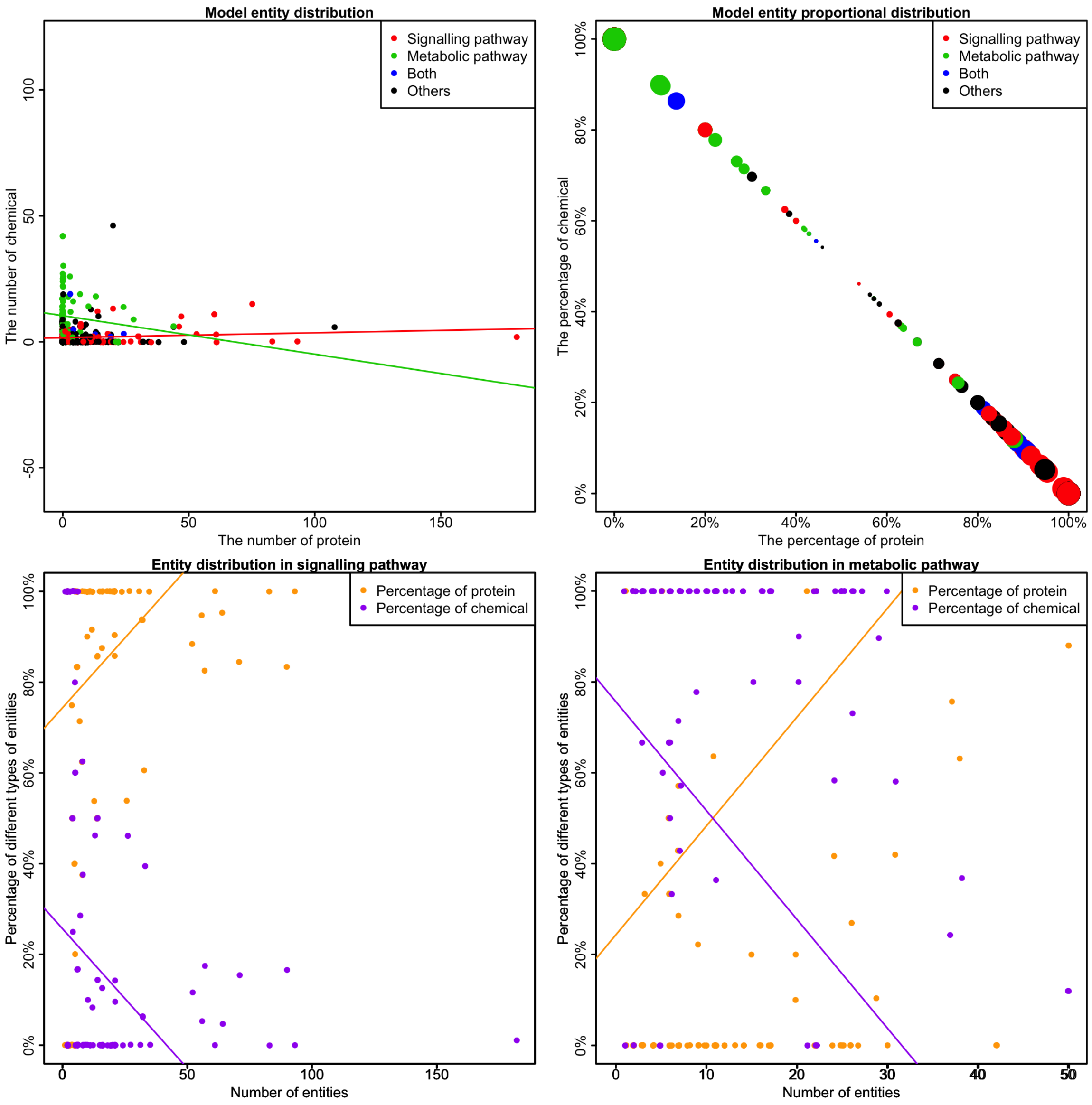


Fig. 4.4: Entity type composition of biological networks in BMDB by entity types. In the four charts, each dot represents a reaction network in BMDB. In the top left chart, the horizontal axis is numbers of proteins in each network, and the vertical axis is the number of the chemical entities in each network. In the top right chart, the horizontal axis is percentage of proteins in a network, and the vertical axis is percentage of chemical entities in a network. In the both top charts, a red dot represents a signalling pathway. A green dot represents a metabolic pathway. A blue dot represents a network involving both signal transduction and metabolic pathway. If a network involves neither signalling transduction, nor metabolic processes, it is represented by a black dot in the charts. In the the both bottom charts, the horizontal axes are total numbers of entities involved in each pathway. The vertical axes are percentage of proteins or chemical entities involved each pathway. An orange dot represents a network, which protein percentage is plotted against the vertical axis. A purple dot represents a network, which chemical percentage is plotted against the vertical axis.

number of the networks involving both systems is limited, and the network sizes are relatively small. By plotting linear regression of all the signalling pathways (the red line) and of all the metabolic pathways (the green line) in BMDB. The trends based on linear regression show that, with network size increasing, signalling pathways involve more proteins. Chemical entity numbers do not have significant increase. The metabolic pathways involve more chemical entities than proteins. When protein numbers in the metabolic pathways increases, chemical entity numbers decline.

Generally speaking, the signalling pathways contain more proteins and the metabolic pathways contain more chemical entities. This is better observed from the top right chart in Fig. 4.4, which presents proportions of proteins and chemical entities in each network. The signalling pathways are gathered close to the bottom right area of the chart, and the metabolic pathways are gathered near the top left area of the chart. The two charts in the bottom are the percentages of the proteins and the chemical entities respectively in comparison to the total number of entities in the pathways. The bottom left chart in Fig. 4.4 presents percentage of proteins and chemical entities in each signalling pathway. From the chart, it can be seen that, for the signalling pathways, the protein numbers are increasing when the network sizes increase. On the opposite, the chemical entity numbers decrease. This means, even smaller signalling pathway, which might describe a part of a big system, still involves chemical entities. Therefore, to extract a biologically meaningful signalling pathway by TM, chemical entities still deserve a certain level of consideration.

The bottom right chart in Fig. 4.4 shows percentage of proteins and chemical entities in each metabolic pathway. From the chart, it is observed that the metabolic pathways contain mostly chemical entities. However, with the network sizes increasing, the percentage of protein in the networks increases. This is natural because bigger metabolic pathways tend to involve related regulatory mechanisms.

The analysis shows that different types of networks demand different treatments during extraction. Based on compositions of participating molecules' types, different types of biological network need discriminative solutions from NER phase. For signalling pathways, PGNs play an important role due to the large involvement of different proteins. In metabolic pathways, more chemical entities are involved. However, the distinction of proteins and chemical entities is not as prominent as it is in signalling pathways.

In a biological reaction network, a reaction participant could be a protein, a chemical entity, or compound of both. Sometimes, it could be a molecule with more complex structure, e.g. a protein complex. Nomenclature for proteins, chemical entities, even compounds are getting better and better established. The examples are the ones mentioned before, UniProt, ChEBI and KEGG etc. Thereafter, NER based on the related terminological and lexical resources is relatively more advanced. However, as being discussed in Section 1.3.1, due to the lack of nomenclature of multi-protein complex, the macromolecules are frequently referred by the name of a part of the entire molecule, for example, EGFR for EGF-EGFR. Without correctly identifying this type of entities, network reaction sequence cannot be faithfully revealed in terms of real reaction in vivo, therefore, cannot truly support large-scale automatic data screening.

To investigate the problem, I analyse the reaction participants, which involve multi-type molecules. Again, all the biological networks are extracted from BMDB, and classified as signalling pathways, metabolic pathways and others in terms of their annotations. A few networks involving both processes are also considered. All the network entities are extracted and analysed based on the qualifiers, which are parts of manually added annotations. As all the annotated entities have been carefully checked by the curators, and associated with the external references considered the most relevant. It is a reasonable assumption that, when an entity is annotated with at least one of the explicit qualifiers (*bqbiol:is*, *bqbiol:isVersionOf* or *bqbiol:hasVersionOf*), the entity has a standard name, which can be used in publication; when an entity is annotated without any explicit qualifiers, but with implicit qualifiers (*bqbiol:hasPart* or *bqbiol:isPartOf*), it is likely to be a macromolecular complex or even a mixture.

Fig. 4.5 differs from Fig. 4.4 as it represents the analysis of entity types based on the qualifiers. In all the charts in Fig. 4.5, a dot represents a reaction network represented by a model. The top left chart illustrates the respective numbers of pure substance and mixture occurred in each network. The horizontal axis is the number of the pure substances in each network, and the vertical axis is the number of the mixtures in each network. The top right chart represents the proportions of the pure substances and mixtures in each network. Its horizontal axis is the percentage of the pure substances in a network, and the vertical axis is the percentage of the mixtures in a network.

In both top charts in Fig. 4.5, a red dot represents a signalling pathway determined based on its manually added annotation. A green dot represents a metabolic pathway. A blue dot represents a network, which involves biological processes of both signal transduction and metabolic pathway. If a network involves neither a signalling pathway, nor a metabolic pathway, it is represented by a black dot in chart.

The bottom left chart presents the percentage of pure substances and mixtures in each signalling pathway. The bottom right chart shows the percentage of pure substance and mixtures in each metabolic pathway. In both bottom charts, the horizontal axes are the total

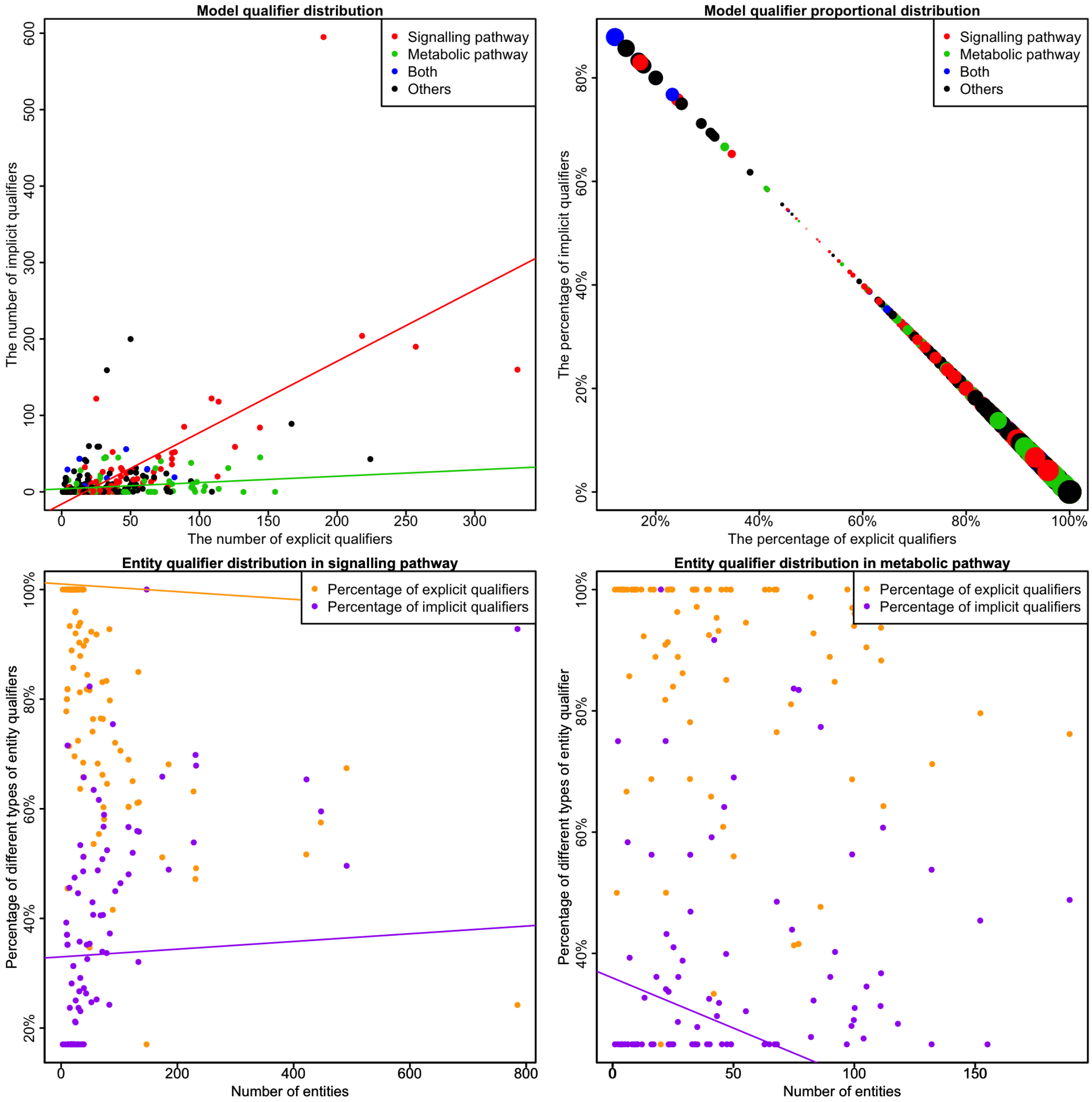


Fig. 4.5: Entity composition of the biological networks in BMDB by qualifiers. In the four charts, each dot represents a reaction network in BMDB. In the top left chart, the horizontal axis is numbers of entities with explicit qualifiers in each network, and the vertical axis is number of entities with implicit qualifiers in each network. In the top right chart, the horizontal axis is percentage of entities with explicit qualifiers in a network, and the vertical axis is percentage of entities with implicit qualifiers in a network. In the both top charts, a red dot represents a signalling pathway. A green dot represents a metabolic pathway. A blue dot represents a network involving both signal transduction and metabolic processes. If a network involves neither signalling transduction, nor metabolic processes, it is represented by a black dot. In the the both bottom charts, the horizontal axes are total numbers of entities involved in each pathway. The vertical axes are percentage of entities with explicit or implicit qualifiers of each pathway. An orange dot represents a network, which percentage of entities with explicit qualifiers shows on the vertical axis. A purple dot represents a network, which percentage of entities with implicit qualifiers shows the vertical axis.

number of entities involved in the pathways. The vertical axes are the percentage of the pure substances or mixtures involved in the pathways. An orange dot represents a network, which pure substance percentage is plotted against the vertical axis. A purple dot represents a network, which mixture percentage is plotted against the vertical axis.

From the top left chart in Fig. 4.5, we can see that the signalling pathways involve more macromolecules or mixtures in comparison with metabolic pathways. This trend is not significant according to the analysis of their percentage in the top right chart, but clearly showed in the bottom left chart. From the bottom left chart, it is observed that the entities with the explicit qualifiers are more than those with the implicit qualifiers in the signalling pathways. With the network size increasing, the percentage of the implicit qualifiers slightly increases. It is similar in the metabolic pathways that the explicit qualifiers are more than implicit qualifiers. However, differing from the signalling pathways, with the metabolic pathways' size getting bigger, the percentage of the implicit qualifiers decreases, which means, in the larger metabolic networks, more macromolecules and mixtures are getting involved into the reactions.

4.2.2.2 Reaction type composition in different networks

The previous section (Section 4.2.2.1) quantitatively analyses the composition of entity types in different biological networks. By using the up-to-date TM technologies, this section evaluates the biological networks from BMDB and quantifies the knowledge about the reactions, and the biological networks in the scientific literature. The analysis focuses on the quantity and the morphology of the reactions and the networks. Firstly, a high-recall approach (CO2) is used to evaluate the curated networks in BMDB and gives statistic about the information coverage of the metabolic pathways and the signalling pathways in scientific literature. Then, by using a more precise method, LitWay (Section 3.3), I evaluate the same set of the biological networks.

The networks in BMDB are grouped into five types according to the entity compositions and the biological functions: PPI networks, chemical reaction networks, CPI networks, signal transduction pathways and metabolic process pathways. The first three categories of the networks are classified based on the involved entity types. When a network only involves proteins, it is considered as a *PPI network*). If it only contains chemical entities, then it is considered as a *chemical reaction network*. If it has both types of entities, then it is considered as a *CPI network*. On another hand, if a network is annotated with *signal transduction pathways* (GO:0007165) or the annotation's descendants, it is considered as a signalling pathway network. If it is annotated with *metabolic process pathways* (GO:0008152) or its

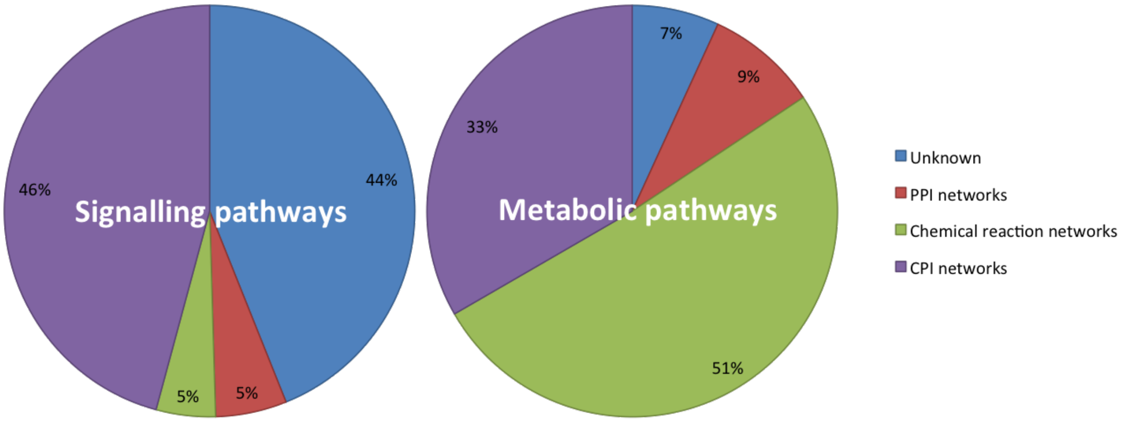


Fig. 4.6: Reaction type distribution of the signalling pathways and the metabolic pathways in BMDB. The left pie chart is the distribution of different reaction types in the signalling pathways. The right chart is the distribution of different reaction types in the metabolic pathways.

descendants, it is considered as a metabolic network. There could be overlap between the first three categories and the last two categories. The analysis about the overlap of the entities can be found in Section 4.2.2.1. The analysis in this section will characterise each type of networks and compare them by trying to recover their entities and reactions from the scientific literature.

Signalling pathways are mainly constructed by PPI. The statistic on the signalling pathways from BMDB shows 5% signalling pathways are purely PPI networks (the left pie chart in Fig. 4.6). The majority of signalling pathways in BMDB are CPI networks, which is 46%. Meanwhile, there is 5% networks are chemical reaction networks. Although the percentage may be altered as 44% networks' type is unknown due to limited annotations, signalling pathways still should not only be focusing on PPI extraction. An assumption about the reason why current text mining is focusing on PPI may be that protein/gene mentions have been long studied and has relatively mature terminological resource, from which lexical tagging may benefit. However, to pace forward to network extraction, CPI involving small molecules as, e.g. messengers, definitely should not be overlooked.

In the right pie chart in Fig. 4.6, CPI networks are 33% of all the metabolic networks in BMDB. More than half of the metabolic networks in BMDB is chemical reaction networks, which has 51%. 9% metabolic networks are PPI-only networks, which may be the networks describing regulatory signals in metabolic processes, for example, the regulation of tyrosine phosphorylation of STAT protein describe in BIOMD0000000167. Thus, statistic on both entity type composition and reaction type composition tells that CPI is much more happening in either signalling pathway or metabolic pathway.

4.2.2.3 Evaluation of different types of networks in BMDB

Two methods (CO2 and LitWay) are used to extract the reactions from the MEDLINE corpus (see detail about the corpus in Section 4.2.1.1) for evaluation the reactions in the networks. During the evaluation, both methods align two different information sources (metadata and annotations, see detail in Section 4.2.1.3) from BMDB with the information extracted from the corpus. When exploiting metadata, ld and name are aligned with extraced information. When using an annotation, the accession id, name and synonyms extracted from external references are aligned with extracted information in the given order. If all the participants of a reaction in BMDB can be aligned with all the participants of a reaction found in the corpus, then the former reaction is considered being recovered by the method. When there are x entities are recovered in a network with y entities, the network's recovery rate P is

$$P = \frac{x}{y} \tag{4.1}$$

The networks are also categorised by the sizes (see detail about categorisation in Section 4.2.1.5). Thus, the experiment calculates percentage of recovered reactions of each network in BMDB across three dimensions: the information sources, the sizes and the methods.

The five figures from Fig. 4.7 to Fig. 4.11 correspond to five types of networks (see detail about the network types in Section 4.2.1.4), which are PPI networks, chemical reaction networks, CPI networks, signalling pathways and metabolic pathways. Each figure shows the recovery rate of the entities and the reactions in the networks of the corresponding type. In each figure, the first row is the recovery rates of the entities. The second row is the recovery rates of the networks by CO2. The third row is recovery rates of network reactions by LitWay. Within each row, the left chart is recovery rates of network elements (entity or reaction) recovered from corpus by using metadata from BMDB. The second chart is recovery rates of network elements recovered by using annotation. The third chart is recovery rates of network elements recovered by using both metadata and annotation. Red box is recovery rates of small networks (five or less entities); green box is recovery rates of medium networks (entity number ranges from five to twenty); blue box is recovery rates of large networks (more than twenty entities).

The coverage of the entities and the reactions in BMDB is also chronologically analysed in the last part of this section.

PPI network

The PPI networks, in which the reactants only consist of proteins, are picked out from BMDB and evaluated in several dimensions: network element (entity and reaction), network size, recovery method and information sources.

When using metadata, the recovery rates of the entities in the small networks range from 0% to 100% (Fig. 4.7) with the mean at around 30%. As introduced in Section 4.2.1.3, when metadata in metadata literally does not have to contain a biological term or acronym, it may be hard to be detected in text. Thus, by using metadata only, the recovery rates decline when network size increases. The trend generally maintains the same in the medium networks and the large networks. However, the recovery rate of the entities increases dramatically when using annotations. The mean of the recovery rates of three sizes of networks are all above 60%, in which medium networks even achieved better rates. When annotations and metadata are combined for the entity recovery, the rates go slightly higher than merely using either of them alone. It proves that semantic resources used in BMDB has a large alignment with terminological resources used by NER solutions. Meanwhile, information from two sources, metadata and annotation, are compensating each other. In all the cases, smaller networks have better recovery rates than bigger ones.

Recovery rates of entities reflect recall of recognised entities from the corpus against the entities in BMDB. Reactions are based on the recovered entities. However, it is common that many reactions take place on a limited number of entities.

The PPI networks are then evaluated by using co-occurrences (CO2) and the more sophisticated method, LitWay (see detail about the system in Section 3.3). Without implementing coreference strategy, a system can only extract relation from the same text unit, which is a sentence in this evaluation. CO2-based recovery senses the maximum amount of relations on sentential level. The three charts in the middle of Fig. 4.7 show that the reaction recovery has the similar trend of the entity recovery. Annotations help improve the recovery rate, and combination of both information resources is better than using either alone. Nonetheless, the difference is not as much as it in entity recovery. Around half of the reactions in the networks can be recovered by using the CO2 approach, and the rates do not have significant difference among different size of networks.

In comparison to the CO2 approach, LitWay recovers fewer reactions in the networks. When using metadata alone, very few relations were recovered. Similarly, annotations helped the recovery. But a number of medium size networks completely cannot by recovered, which is not surprising as LitWay is implemented with stricter linguistic analytic methods.

Chemical reaction network

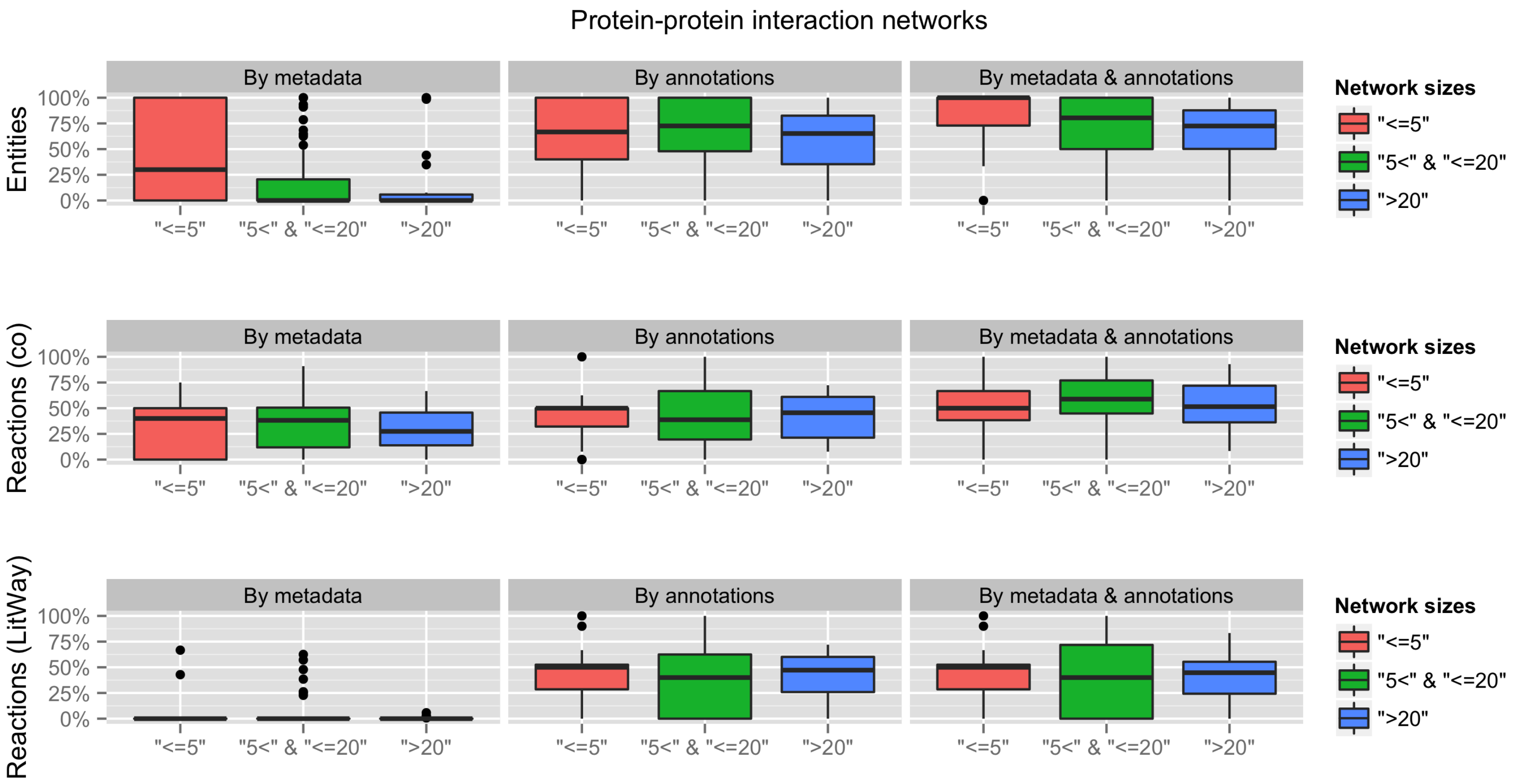


Fig. 4.7: Recovery of the PPI networks in BMDB by different methods. The top chart is percentage of recovered entities in the PPI networks. The middle chart is percentage of reactions recovered by co-occurrences. The bottom chart is percentage of reactions recovered by LitWay. In all the charts, the left sub-charts are percentage of network elements recovered by using metadata; the central sub-charts are percentage of network elements recovered by using annotations; the right sub-charts are percentage of network elements recovered by using metadata and annotations. In all the charts, red boxes are percentage of networks with less than 6 entities; green boxes are percentage of networks with more than 5 and less than 20 entities; blue boxes are percentage of networks with more than 20 entities.

In Fig. 4.8, we still can see the benefit of using annotation in comparison with using metadata alone. Interestingly, although chemical nomenclature is known to be relatively mature, the chemical networks' recovery rate is not as good as PPI's. This is largely due to that fragment of name or conventional name tends to be used when a chemical entity name appears to be very long.

This compensating effect between metadata and annotation is also not as prominent as it in PPI networks. This might indicate that the available semantic information about chemical entities from external resources is more limited than it about proteins and genes. Meanwhile, it is interesting that, by either method, bigger networks have better recovery rate than smaller ones.

The chemical reaction networks are evaluated by using CO2 and LitWay. Surprisingly, by either method, annotations of chemical entities are not helping the recovery of reactions when using CO2 approach. This has two reasons. First, a fragment of a name or conventional name is used when a chemical entity has a very long name. Second, when encoding networks in SBML, many chemical entity metadata is encoded with ambiguous names, acronyms or formula, such as GAP (Glyceraldehyde 3-phosphate) [94]. Named entity recognition without normalisation tagged a big number of such ambiguities, which, however, cannot be aligned with chemical entities' scientific names from semantic resources. When using LitWay to recover the reactions from the networks, due to the stricter grammatical analysis, the general recovery rate is lower than CO2-based.

CPI network

As analysed in the previous section, chemical-protein interaction network is a big portion of all the types of networks. Differing from chemical reaction networks, the general difference between using different methods are similar to it in PPI networks (Fig. 4.9), in which annotations help improving recovery rate and combination of both information resources is better than using either alone. However, the general recovery rate is lower than the PPI networks', and the difference between different methods are not as significant as it is in PPI networks. This is much due to the involvement of chemical entities. For the most of time, medium size networks have the best recovery rate. This demonstrates the available semantic resources have a relatively good coverage of the entities appearing in the scientific literature. This largely benefits from the development of bioinformatic resources of bio-molecules and the fast pace of related curation works.

Signalling pathway

Signalling pathways facilitate the communication of information received at the cell surface to the cellular centre, the nucleus, where signalling through the pathway is able to

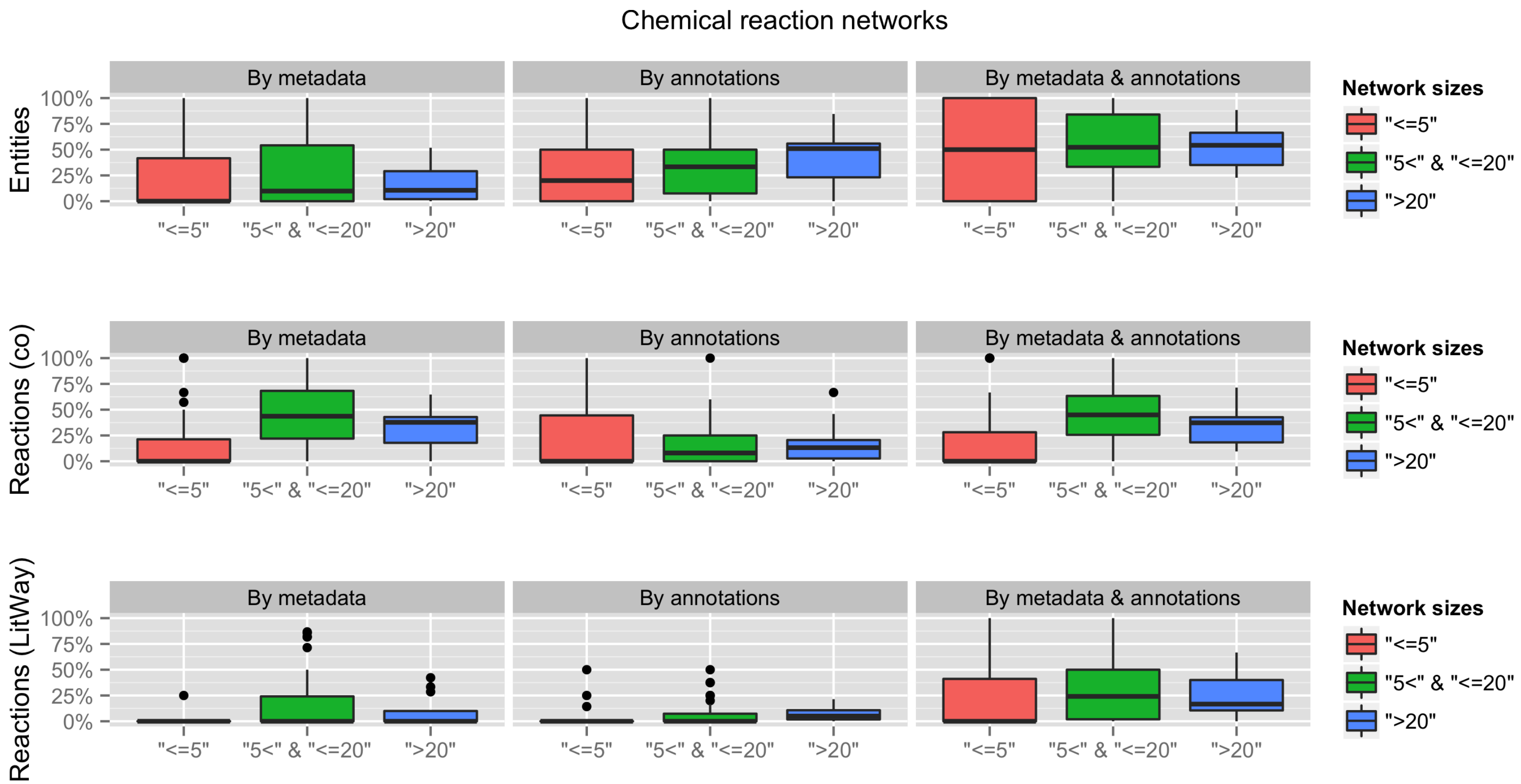


Fig. 4.8: Recovery of the chemical reaction networks in BMDB by different methods. The top chart is percentage of recovered entities in the chemical reaction networks. The middle chart is percentage of reactions recovered by co-occurrences. The bottom chart is percentage of reactions recovered by LitWay. In all the charts, the left sub-charts are percentage of network elements recovered by using metadata; the central sub-charts are percentage of network elements recovered by using annotations; the right sub-charts are percentage of network elements recovered by using metadata and annotations. In all the charts, red boxes are percentage of networks with less than 6 entities; green boxes are percentage of networks with more than 5 and less than 20 entities; blue boxes are percentage of networks with more than 20 entities.

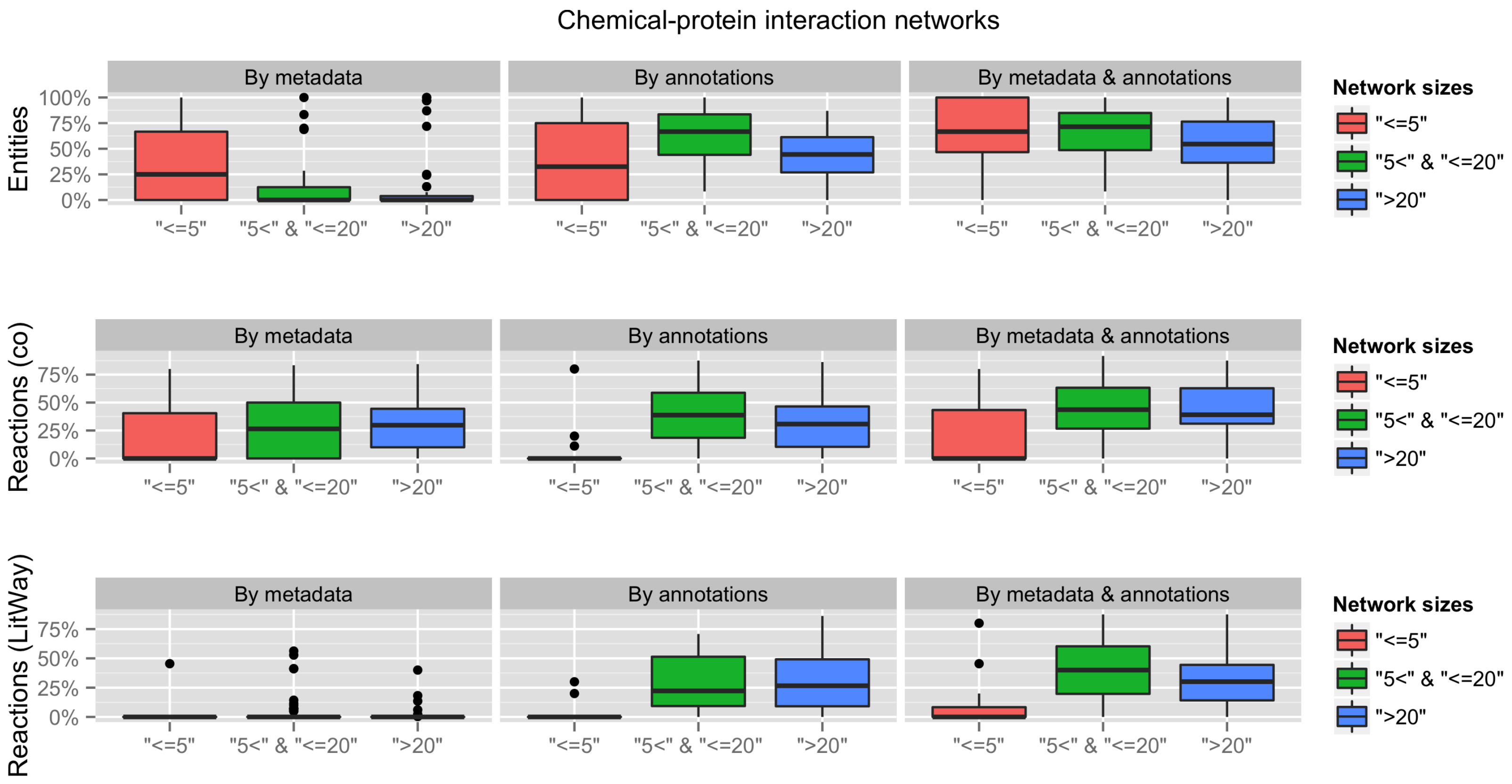


Fig. 4.9: Recovery of the CPI networks in BMDB by different methods. The top chart is percentage of recovered entities in the CPI networks. The middle chart is percentage of reactions recovered by co-occurrences. The bottom chart is percentage of reactions recovered by LitWay. In all the charts, the left sub-charts are percentage of network elements recovered by using metadata; the central sub-charts are percentage of network elements recovered by using annotations; the right sub-charts are percentage of network elements recovered by using metadata and annotations. In all the charts, red boxes are percentage of networks with less than 6 entities; green boxes are percentage of networks with more than 5 and less than 20 entities; blue boxes are percentage of networks with more than 20 entities.

manipulate the behaviour of the cell. The communication embodies a series of biochemical reactions. Fig. 4.6 showed that signalling pathways involve not only PPIs or chemical reactions. The majority of the reactions are actually occurring between proteins and chemical entities, which is different from the empirism. The signalling pathways in BMDB are identified by the GO annotations (the networks annotated with GO:0007165 or its descendant GO terms) and extracted for the experiment.

The entities in the signalling pathways have a relatively good recovery rate by involving annotations (Fig. 4.10). Involving annotations achieved much better recovery rates than using metadata. Despite the poor recovery rate by metadata, the combination of both information still gains a bit improvement, especially for the small networks. This again proves that the information from both metadata and annotation are not fully aligned with each other, and is able to compensate with each other. Therefore, both should be considered during biological network extraction. The sizes of the models do not have clear correlation with the recovery rate, as the result varies. By using metadata, recovery rate decreases when size increases. Using annotation, the medium networks have the best recovery rate. When using both, the small networks gained the best recovery rate.

Based on CO2, using annotations achieved generally better recovery rates besides the small networks, which recovery rate is better when only using metadata. The combination of both information has improvements for all the networks with the different sizes. The network size per se does not have influence on the recovery rate. For example, the large networks get the best recovery rate when only using annotations, which the medium network get the best when using both information. LitWay achieved relatively low recovery rates. The combination of metadata and annotations achieves the best recovery rates.

Metabolic pathway

The metabolic pathways from BMDB are extracted by checking annotations. If a network is annotated with the GO term, metabolic process pathway (GO:0008152), or its descendant terms, then the network is considered as a metabolic network.

The entities in the metabolic pathways are better recovered by using annotations than by using metadata. The combination of both information gets the best entity recovery rate. Network size does not have strong influence on the recovery rate.

CO2 achieved better reaction recovery rate by using metadata than by using annotations. This is because metabolic networks are mainly chemical reaction networks, while many chemical entities are encoded by using ambiguous acronyms or formula, such as GAP (Glyceraldehyde 3-phosphate) or CO (carbon oxide). While the reaction recovery rate by using annotations did not change much by using LitWay, metadata based reaction recovery

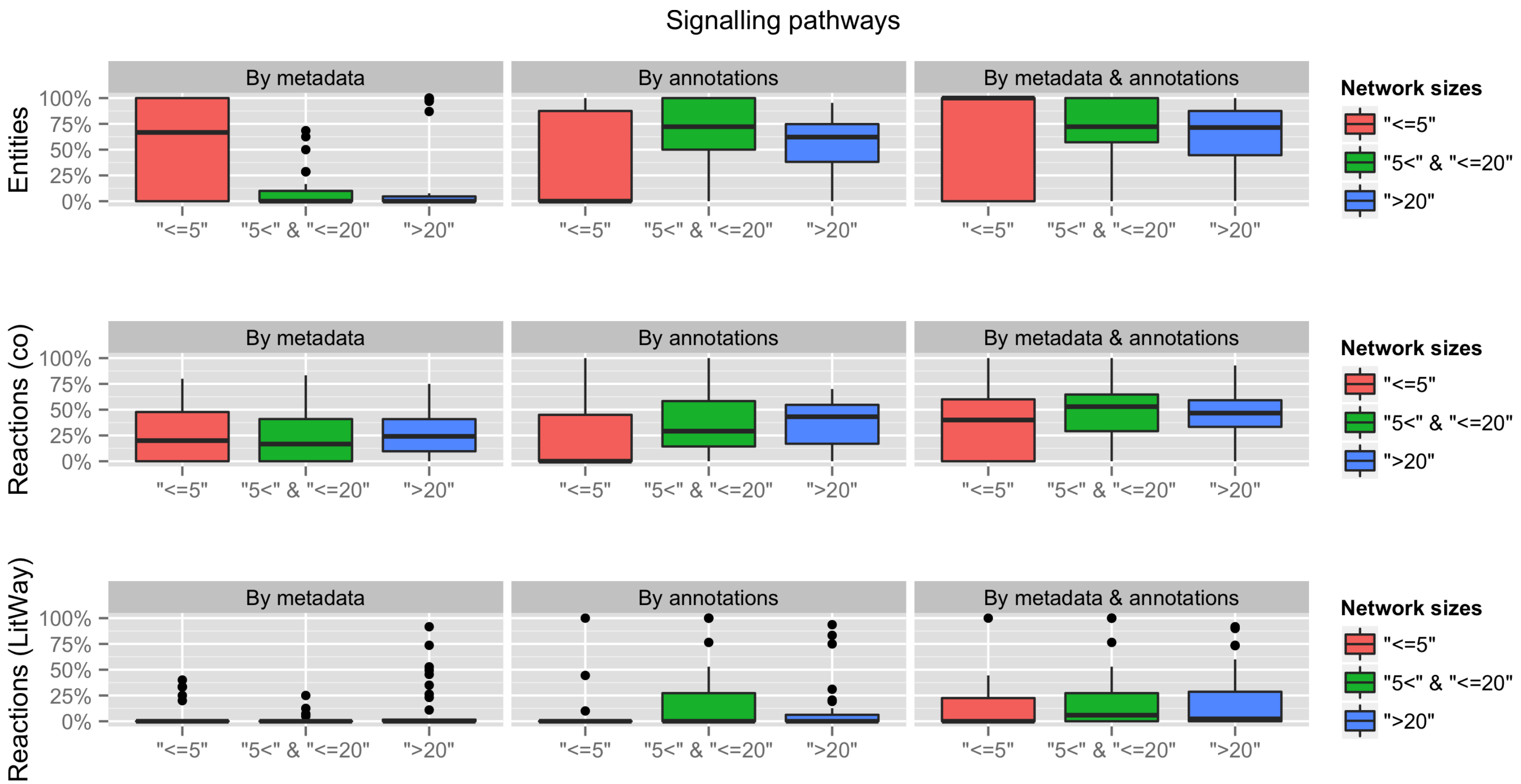


Fig. 4.10: Recovery of the signalling pathways in BMDB by different methods. The top chart is percentage of recovered entities in the signalling pathways. The middle chart is percentage of reactions recovered by co-occurrences. The bottom chart is percentage of reactions recovered by LitWay. In all the charts, the left sub-charts are percentage of network elements recovered by using metadata; the central sub-charts are percentage of network elements recovered by using annotations; the right sub-charts are percentage of network elements recovered by using metadata and annotations. In all the charts, red boxes are percentage of networks with less than 6 entities; green boxes are percentage of networks with more than 5 and less than 20 entities; blue boxes are percentage of networks with more than 20 entities.

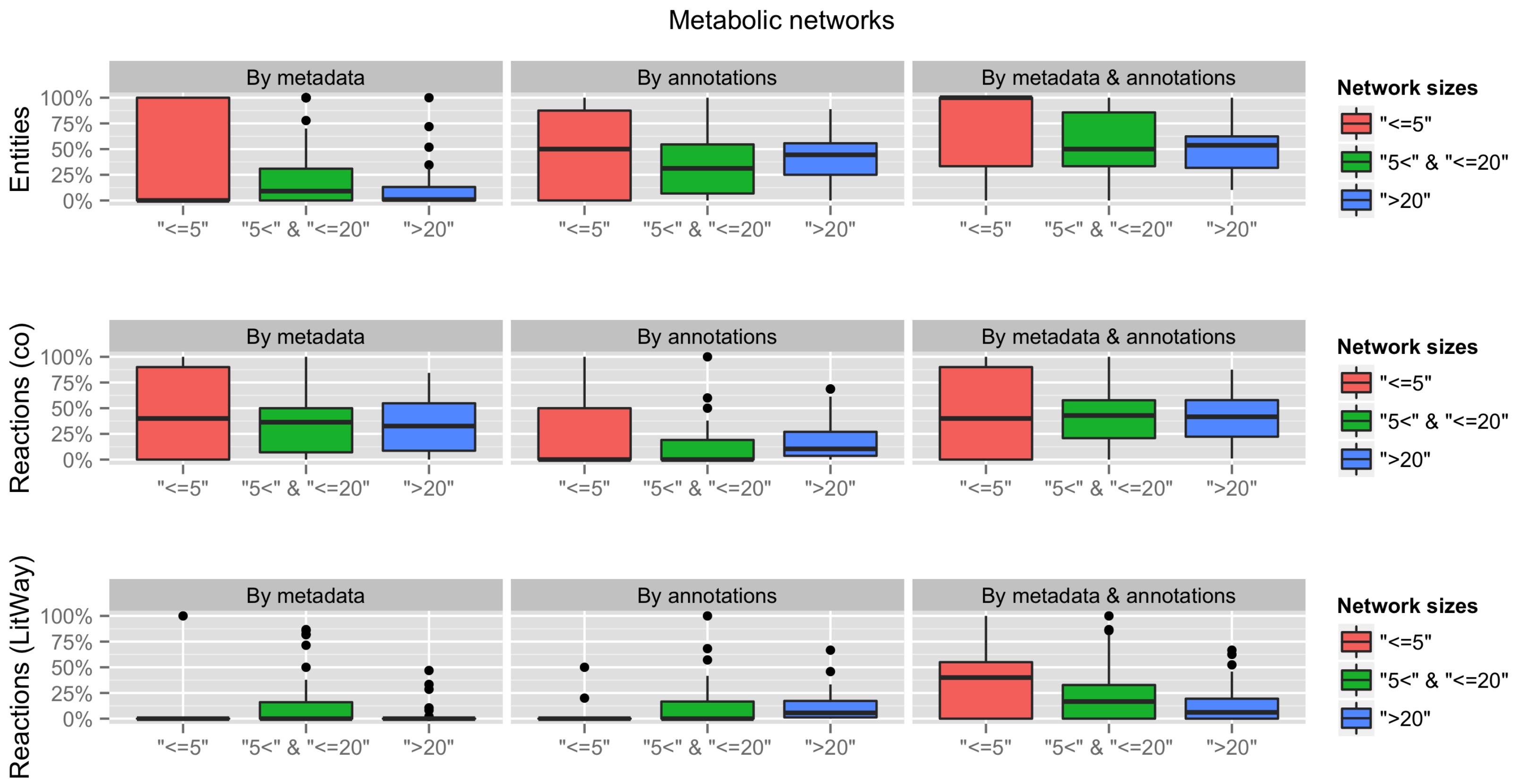


Fig. 4.11: Recovery of the the metabolic networks in BMDB by different methods. The top chart is percentage of recovered entities in the metabolic networks. The middle chart is percentage of reactions recovered by co-occurrences. The bottom chart is percentage of reactions recovered by LitWay. In all the charts, the left sub-charts are percentage of network elements recovered by using metadata; the central sub-charts are percentage of network elements recovered by using annotations; the right sub-charts are percentage of network elements recovered by using metadata and annotations. In all the charts, red boxes are percentage of networks with less than 6 entities; green boxes are percentage of networks with more than 5 and less than 20 entities; blue boxes are percentage of networks with more than 20 entities.

dramatically dropped due to the stricter syntactic analysis.

4.3 Evaluation of a comprehensive human metabolic network

It has been long understood that "one gene-one protein-one function" is an over simplified theory. One gene can be expressed to different protein, e.g. alternative splicing. The same protein also may play different roles in different organisms (e.g. cellular components), or different systems (e.g. pathways). It is similar for chemical entities, one chemical molecule may perform different functions in different environments. Meanwhile, investigating molecular function needs to taking into account all the known properties, necessary constituents, their relationships and even the temporal variation. Systems Biology envisions the inadequacy of studying isolated biological processes and builds more complete biological networks to gain full understanding about diseases, almost all of which are multi-factorial.

RECON2 [43] is so far one of the most comprehensive representations of human metabolism. The network described in its SBML model links the smallest molecular scale to the full cellular level, and contains 1,789 enzyme-encoding genes, 7,440 reactions and 5,063 metabolites. The entities and reactions are distributed over eight cellular components. The big reaction network converged perennial effort from the community, as it is too complicated to be built by any individual researcher.

Metabolism is a set of enzyme-catalysed reactions transforming life-sustaining chemicals within the cells and allowing organisms to grow and reproduce, maintain their structures, and respond to their environments. The sophisticate network adjusts itself for environmental change, however some personal or environmental change may result in diseaseful effect. TM-aided network evaluation collects molecular properties, relationships, locations and environmental changes from published knowledge to assist the network completion and semantical enrichment. I conducted a series of experiments to use different TM methods to recover the network entities and reactions from the scientific literature. Furthermore, alignment between the network and published knowledge reveals unencoded MIs, their roles, and activities in different biological processes and cellular components. High granularity comparison across the processes and locations provide evidences for discovering unreported molecular relations.

4.3.1 Method

The experiment utilises LitWay to scan through the MEDLINE corpus (the same corpus described in Section 4.2.1.1), extract found molecular relations, and store them in a MySQL instance (Fig. 4.12). Extracted relations are stored in *events* and *event_arguments* and associated with the original documents by information stored in the tables, *documents* and *document_events*. Meanwhile, annotations from RECON2 are stored in *networks* and *network_annotations*. The related semantic information, e.g. synonyms, from external reference resources are fetched by Web Services and stored in *semantic_references*. In this way, the analysis can programmatically access centralised information from single database and improve the speed for processing such huge network. The database schema is integrated as part of LitWay. It is generic and can be used for other network analysis.

Entities from RECON2 are all paired with each other. A hypothetical molecular relation (HMR) is created for each pair, if no interaction between them has been coded. All the HMRs are then tested by being aligned with the events extracted from the corpus by LitWay. As a number of entities in RECON2 are redundant (2,626 metabolites are unique in total 5,063 metabolites), the experiment normalises entities according to its annotations at the beginning of the experiment. Each extracted hidden relations is analysed according to the location it taking is place, e.g. cytoplasm, and the potentially related pathways, e.g. aminosugar metabolism or blood group synthesis.

4.3.1.1 Cellular components

As biological processes are highly integrated and influence each other, molecular investigation should always be placed with the consideration of entire biological systems rather than being done individually. Locations where reactions taking place and other biological pathways the molecule involving cross-link different biological pathways and demonstrates versatile roles of the molecule. RECON2 uses *Compartment* in SBML to indicate the cellular components where reactions take place. Seven different cellular components are mentioned in RECON2 for metabolic network (Table 4.1). These are extracted from the SBML file depicting the network. Later analysis will be done with the respect to these cellular components.

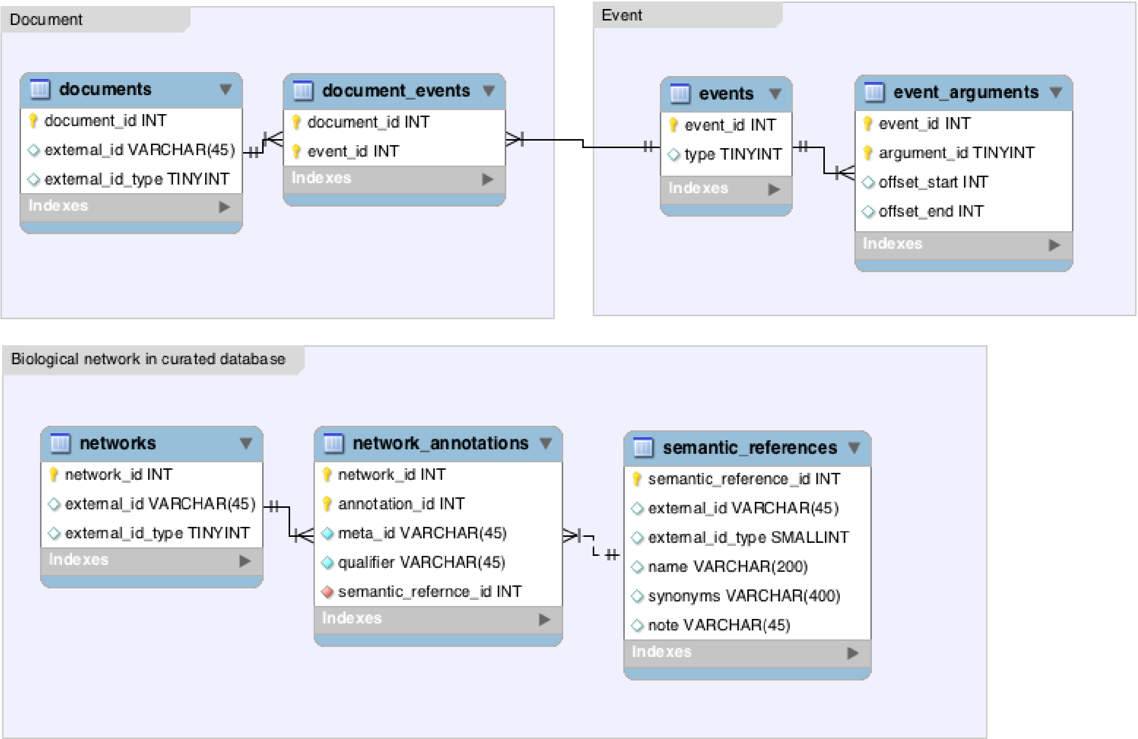


Fig. 4.12: Database schema of LitWay. There are seven tables grouped in three types. The table, documents, has three fields: document_id stores unique identifiers of stored documents; external_id could be an ID from an external publication database, e.g. MEDLINE; external_id_type indicates type of external_id, e.g. MEDLINE or DOI. The table, document_events, references with documents by document_id. Its event_id is unique identifiers of events discovered in documents, and is referenced with event_id in the table, events. The table, events, has type indicating event types (e.g. phosphorylation etc.). The table, event_arguments, is referenced with events by event_id. It also has argument_id as identifiers of arguments (e.g. Theme, Cause etc.). The fields, offset_start and offset_end, indicate annotation offsets in documents. The table, networks, has network_id as identifiers, external_id as identifiers from external databases (e.g. BMDB), and external_id_type for indicating which external database is referenced. The table,network_annotations, is referenced with networks by network_id. The field, annotation_id, is unique identifiers of annotations in networks. The field, meta_id, is identifiers of network elements. The field, quailfier, is for storing qualifiers of semantic_reference_id. The field, semantic_reference_id, is linked with identifiers in the table, semantic_references. The field, external_id, in semantic_references are identifiers from external databases, e.g. UniProt or ChEBI. The field, external_id_type, indicates which database external_id external identifiers are from. The fields, name, synonym and note, are information retrieved from external databases.

4.3.2 Result

4.3.2.1 Entity coverage of the sub-pathways

A complete network like RECON2 offers opportunity to trace up- or downward. As nodes of network, the entities' coverage in the scientific literature directly reflects the coverage

Table 4.1: The cellular components involved in RECON2.

Compartment name
cytoplasm
endoplasmic reticulum
extracellular space
golgi apparatus
lysosome
mitochondrion
nucleus and peroxisome

of the general information about metabolic processes in published knowledge. The evaluation of the entities in RECON2 and its sub-pathways will align the structured information about human metabolism with the semantic information, which may be extracted from the scientific literature, and delivers quantitative and morphological information about human metabolism in unstructured text.

Before RECON2, several models of human metabolism have been created, while each of them only represents a part of the process [204]. In RECON2, 99 pathways are covered and expanded. Each pathway has been studied with different depth. Therefore, their information availability in the scientific literature accordingly varies. Appendix A shows the percentage of the entities in each pathway, which can be recovered from the scientific literature by NER. This represents the maximum portion of the network, which can be extracted by TM.

By using metadata, three pathways of 99 pathways cannot be found related with any abstracts from the corpus, which are O-glycan synthesis, Blood group synthesis, Keratan sulfate synthesis. The largest pathway (exchange/demand reaction) with 741 entities has 34.28% of them recovered from the literature, when the smallest pathway (nucleotide salvage pathway) with 4 entities has 75% recovered. D-alanine metabolism and ROS detoxification have their entities 100% found. With the reinforcement of information from annotations, all the pathways' coverages have been improved. 9 pathways (Cysteine Metabolism, D-alanine metabolism, Lipoate metabolism, Nucleotide salvage pathway, ROS detoxification, Vitamin B12 metabolism, Vitamin B2 metabolism, Vitamin B6 metabolism, and Vitamin E metabolism) have 100% entities found. The largest pathway (exchange/demand reaction) have 79.35% of them recovered, when the smallest pathway (nucleotide salvage pathway) have 100% recovered. The pathway about extracellular transport has the largest number of related abstracts, which is 156,938.

4.3.2.2 Molecular relation discovery by TM

Discovering hidden biological relations is one of the bio-TM tasks favoured by biologists. We have discussed the two types of hidden relations in Section 4.1.2. Based on the relation discovery tools we developed, I focuse on the first type of hidden relation, which has been reported in the scientific literature, however, is missing in structured knowledge-bases. Such relation is not easily accessible for biologists, and does not support automated knowledge retrieval.

Merely based on extracted information, it is very difficult for TM solutions to give a straightforward answer about relation between two biological concepts. Statistics methods are often combined as confident value about extracted relation. The assumption is, when a relation is mentioned in the literature with higher frequency than others, the certainty about the relation is relatively higher. This assumption is a natural extension of term frequency (TF), although it has not considered the inverse document frequency (IDF).

HMRs found by LitWay

All the entities in RECON2 are paired with each other as a hypothetical molecular relation (HMR), if the pair is not encoded in the network. Then each HMR is aligned with the interactions extracted by LitWay from the same corpus used in the previous section (Section 4.2.1.1). If both entities of an HMR were aligned with those of an interaction found by LitWay, the HMR is considered being supported by an evidential statement from the MEDLINE corpus.

Totally 1,115 pairs of molecules are found with corresponding statements by LitWay in the corpus (the same corpus in Section 4.2.1.1) suggesting that they might have direct interaction with each other. Thirty pairs with the highest frequency are listed in Table 4.2. Frequency of mention of a concept indicates the certainty of the public knowledge about the concept.

However, the result about the HMRs may be biased by mentions of common words. For example, Adenosine monophosphate (AMP, CHEBI:60880) can produce Adenosine triphosphate (ATP), a substrate of dephospho-CoA kinase (CHEBI:57328). Although AMP is not directly interacting with the kinase, they are obviously often mentioned in the same text unit.

Meanwhile, in each interaction of a pathway, entities may play different roles, which are not necessarily equally essential for the pathway's function. Furthermore, the same entity may perform different functions in different locations of an organism, e.g. a cell. Therefore, it is more informative to look into the entities, whose interactions are mentioned with high frequency (evidential sufficiency) and with higher number of unique entities (molecular

Table 4.2: The mostly occurring HMR in RECON2.

Entity A		Entity B		Frequency
CHEBI:60880	3'-AMP(2-)	CHEBI:57328	3'-dephospho-CoA(2-)	332
CHEBI:12947	D-galactosyl-N-acylsphingosine	CHEBI:16291	1-alkyl-2-acetyl-sn-glycerol	321
CHEBI:28972	(R)-propane-1,2-diol	CHEBI:17815	1,2-diacyl-sn-glycerol	
CHEBI:29002	(S)-propane-1,2-diol	CHEBI:27458	3α,7α,12α,24-tetrahydroxy-5β-cholestan-26-oyl-CoA	288
CHEBI:52595	1-alkyl-2-acyl-sn-glycerol	CHEBI:37575	thiamine(1+) monophosphate(2-)	265
CHEBI:58601	α-D-glucose 1-phosphate(2-)	CHEBI:57351	4,8,12-trimethyltridecanoyl-CoA(4-)	199
CHEBI:57336	2-methylbutanoyl-CoA(4-)	CHEBI:57335	2-methylacetoacetyl-CoA(4-)	157
CHEBI:53487	all-cis-docosa-7,10,13,16-tetraenoic acid	CHEBI:53488	(7Z,10Z,13Z,16Z,19Z)-docosapentaenoic acid	93
CHEBI:57454	10-formyltetrahydrofolate(2-)	CHEBI:15440	squalene	92
CHEBI:22783	β-D-galactosyl-(1->3)-N-acetyl-D-galactosaminyl group	CHEBI:16250	N-acetyl-β-D-glucosaminyl-(1->3)-N-acetyl-D-galactosaminyl group	85
CHEBI:26174	poly(N-acetyllactosamine)	CHEBI:18225	myo-inositol 1,3-bisphosphate	84
CHEBI:57335	2-methylacetoacetyl-CoA(4-)	CHEBI:16897	D-erythrose 4-phosphate(2-)	83
CHEBI:27979	all-cis-icosa-8,11,14-trienoyl-CoA	CHEBI:17029	chitin	78
CHEBI:326268	1,4-butanediammonium	CHEBI:17587	L-gulono-1,4-lactone	76
HMDB06535	β-1,4-mannose-N-acetylglucosamine	CHEBI:17716	lactose	75
CHEBI:2504	aflatoxin B1	CHEBI:16040	cytosine	74
CHEBI:57337	2-methylcrotonoyl-CoA(4-)	CHEBI:16723	4-methylthio-2-oxobutanoate	55
CHEBI:15598	2-methylcitrate(3-)	CHEBI:57441	2,3-diketogulonate	47
CHEBI:58017	5-O-phosphonato-α-D-ribofuranosyl diphosphate(5-)	hsa:7941	PLA2G7, LDL-PLA2, LP-PLA2, PAFAD, PAFAH	43
CHEBI:58672	L-2-aminoadipate(1-)	CHEBI:57403	prostaglandin I2(1-)	42
CHEBI:33568	adrenaline	CHEBI:28547	D-glucuronate 1-phosphate	41
CHEBI:545959	homovanillic acid	CHEBI:45698	3,3'-diiodo-L-thyronine	40

activity).

Therefore, an insight into evidence frequency of each interaction and molecular activity with the respect to individual entity in different location and sub-pathway can be more informative for biological study. For this reason, further analysis is conducted to look into evidential sufficiency and molecular activity in different cellular components and sub-pathways.

Evidential sufficiency

Frequency of mentioning a concept is regarded as the sufficiency of available evidences about the concept, therefore, indicates the confidence about the concept's existence. Table 4.3 lists the entities, which related HMR have been mentioned with the highest frequencies found by LitWay in the corpus.

Heat map is used to to break down the entities' evidential sufficiency with respect to different cellular components and pathways.

In each grid of the heat map in Fig. 4.13, the spectrum from green to red indicates the frequency of an entity's related HMR in a cellular component from low to high. In such way, we can see how sufficient the evidential statements for each entity's HMRs in a cellular component.

It is observed that the interactions related with D-galactosyl-N-acylsphingosine (CHEBI:12947), α-D-glucose 1-phosphate(2-) (CHEBI:58601), (R)-propane-1,2-diol (CHEBI:28972), 1-alkyl-2-acyl-sn-glycerol (CHEBI:52595) and (S)-propane-1,2-diol (CHEBI:29002) taking place in nucleus are well supported by the evidences. Several entities' related interactions in lysosome are well supported. The interactions happening in extracellular and endoplasmic reticulum and related with 2-methylacetoacetyl-CoA(4-) (CHEBI:57335) and 2-methylbutanoyl-CoA(4-) (CHEBI:57336) are well supported.

In each grid of the heat map in Fig. 4.14, the spectrum from green to red indicates the frequency of an entity's related HMR in a sub-pathway from low to high. The difference between Fig. 4.14 and Fig. 4.13 is that the confidence of each entity's related HMRs is compared in the related sub-pathways (see the full list of the sub-pathways in Figure 2 in [43]) rather than cellular components. This analysis tells how much evidential statements found by LitWay for an entity's relation HMR occurring in the related sub-pathway.

Fig. 4.14 shows that almost all the entities more or less have related reactions in the pathways. Glycerophospholipid metabolism, sphingolipid metabolism, galactose metabolism, exchange/demand reaction, bile acid synthesis, NAD metabolism and urea cycle are particularly influenced by the interactions participated by α-D-glucose 1-phosphate(2-) (CHEBI:58601), D-galactosyl-N-acylsphingosine (CHEBI:12947), 1-alkyl-2-acyl-sn-glycerol

Table 4.3: Entities with the most mentioned HMRs found by LitWay in the scientific literature. The entities in grey are also appearing in Table 4.4.

Entity	Name	Mention number
CHEBI:58601	α-D-glucose 1-phosphate(2-)	18509
CHEBI:52595	1-alkyl-2-acyl-sn-glycerol	18017
CHEBI:29002	(S)-propane-1,2-diol	17976
CHEBI:28972	(R)-propane-1,2-diol	17976
CHEBI:12947	D-galactosyl-N-acylsphingosine	9260
CHEBI:57336	2-methylbutanoyl-CoA(4-)	5271
CHEBI:57335	2-methylacetoacetyl-CoA(4-)	5171
CHEBI:27979	all-cis-icosa-8,11,14-trienoyl-CoA	4865
CHEBI:2504	aflatoxin B_1	3874
CHEBI:26174	poly(N-acetyllactosamine)	2854
CHEBI:22783	β-D-galactosyl-(1→3)-N-acetyl-D-galactosaminyl group	2854
CHEBI:57454	10-formyltetrahydrofolate(2-)	2777
CHEBI:53487	all-cis-docosa-7,10,13,16-tetraenoic acid	2774
CHEBI:57337	2-methylcrotonoyl-CoA(4-)	2600
CHEBI:58524	gibberellin A1(1-)	2446
HMDB06535	β-1,4-mannose-N-acetylglucosamine	2418
CHEBI:326268	1,4-butanediammonium	2418
CHEBI:545959	homovanillic acid	1875
hsa:8854	aldehyde dehydrogenase 1 family, member A2	1790
hsa:55967	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 12	1778
hsa:9536	prostaglandin E synthase	1688
CHEBI:57448	(5Z,9E,12S,14Z)-8,11,12-trihydroxyicosa-5,9,14-trienoate	1688
hsa:79087	ALG12, α-1,6-mannosyltransferase	1688
CHEBI:15876	β-D-galactosyl-1,3-(N-acetyl-β-D-glucosaminyl-1,6)-N-acetyl-D-galactosaminyl group	1681
hsa:145226	retinol dehydrogenase 12 (all-trans/9-cis/11-cis)	1679
hsa:1580	cytochrome P450, family 4, subfamily B, polypeptide 1	1678
hsa:1582	cytochrome P450, family 8, subfamily B, polypeptide 1	1676
CHEBI:57351	4,8,12-trimethyltridecanoyl-CoA(4-)	1675
hsa:6476	sucrase-isomaltase (α-glucosidase)	1670
hsa:7296	thioredoxin reductase 1	1661

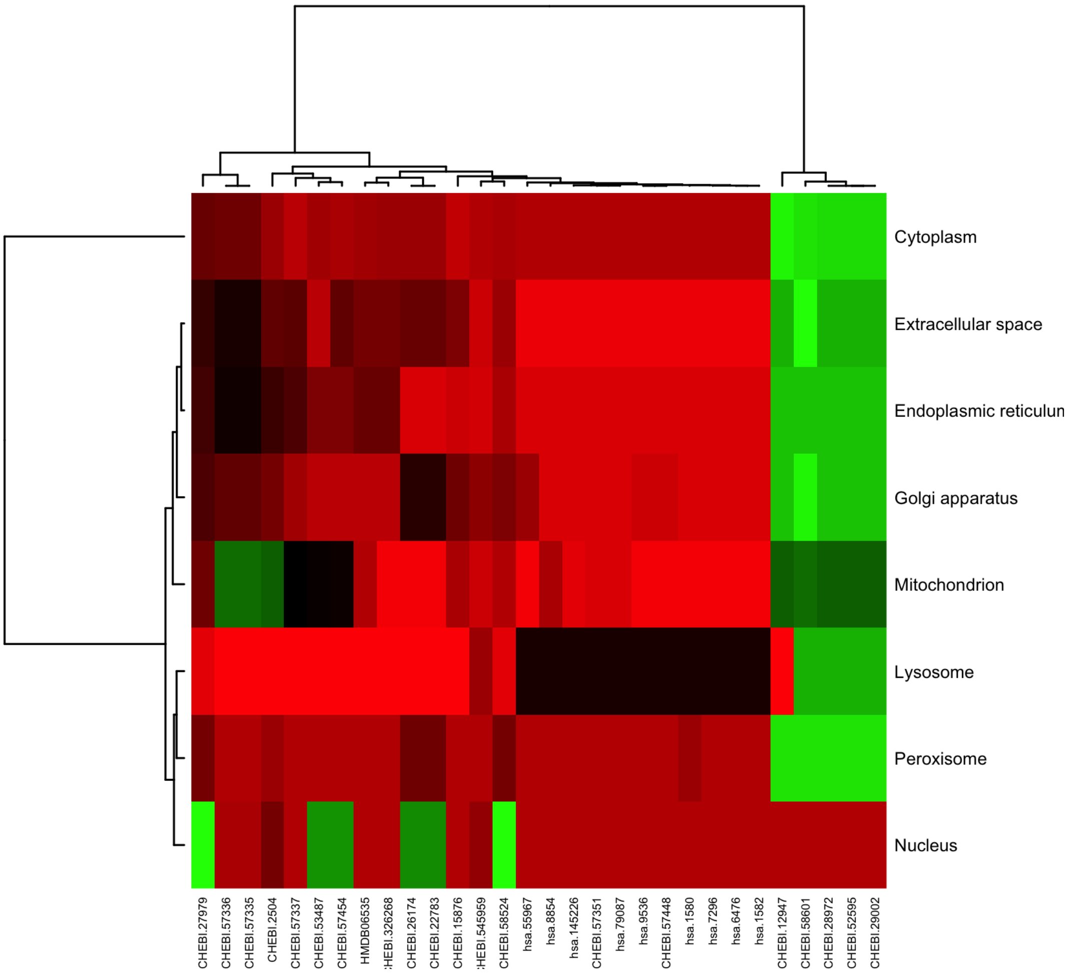


Fig. 4.13: Entities with the most mentioned HMRs occurring in different cellular components found by LitWay. The horizontal axis is the list of entities in Table 4.3. The vertical axis is the different cellular components (Table 4.1). Each grid of the heat map indicates the frequency of an entity's related HMRs in a cellular component found by LitWay. The spectrum from green to red indicates the frequency from low to high.

(CHEBI:52595), (S)-propane-1,2-diol (CHEBI:29002) and (R)-propane-1,2-diol (CHEBI:28972).

Molecular activity

If molecule A involves interacting with more different molecules, in more different cellular components or various pathways, it is reasonable to assume that A is more active and playing more versatile roles than others. Table 4.4 lists the entities found by LitWay that they interact with more different molecules than other, i.e. are involved in more unique HMRs. The first column is the accession of the entity in ChEBI. The second column is the scientific name of the entity extracted from the corresponding external references. The third column is the number of unique HMRs found by LitWay in the corpus. The entities highlighted in grey are those appearing in both Table 4.4 and Table 4.3, which will be presented

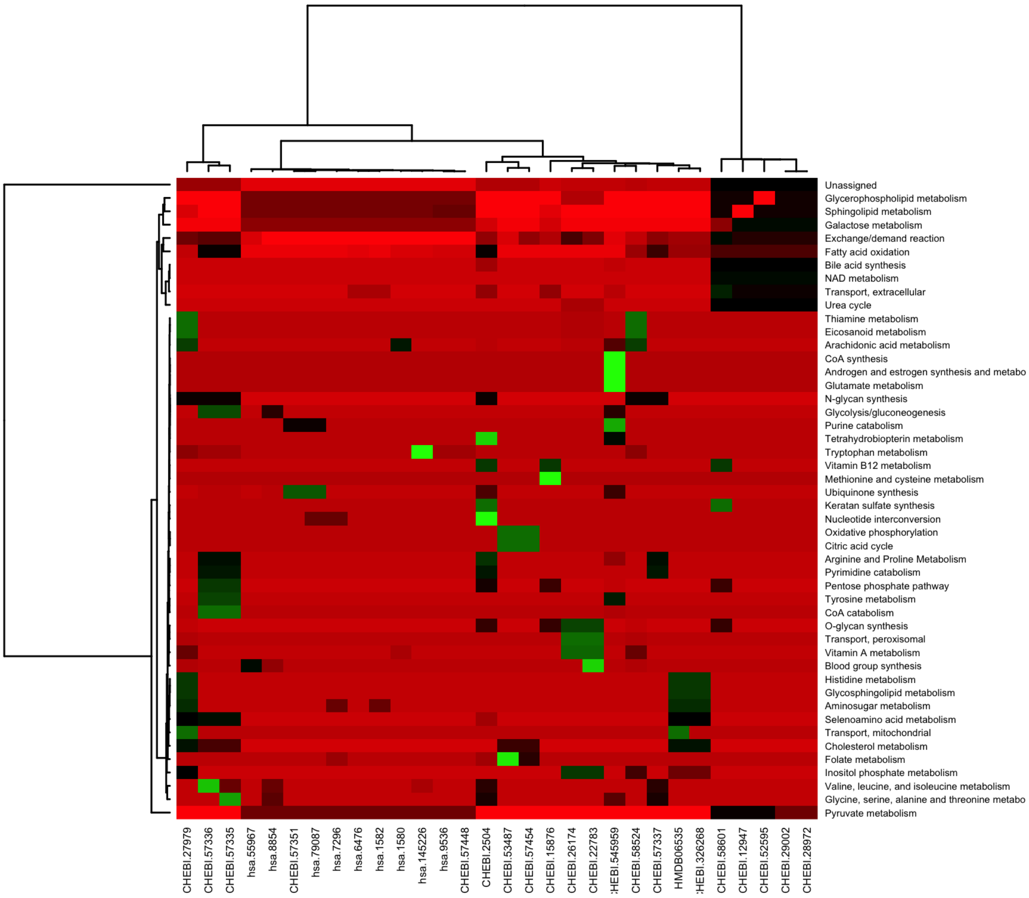


Fig. 4.14: Entities with the most mentioned HMRs in involved in different sub-pathways. The horizontal axis is the list of entities in Table 4.3. The vertical axis is the related sub-pathways in RECON2. Each grid indicates the frequency of an entity's related HMRs in a sub-pathway found by LitWay. The spectrum from green to red indicates the frequency from low to high.

later in this section.

Again, the entities' HMRs are analysed in different cellular components and sub-pathways. In Fig. 4.15, the horizontal axis is the list of the entities, which are those listed in Table 4.4. The vertical axis is the list of locations (Table 4.1). Each grid in the heat map indicates the activity of an entity in a cellular component. From green to red, the spectrum indicates the number of an entity's related unique HMRs from low to high.

From Fig. 4.15, it can be seen that, in peroxisome, the most entities can interact with many different entities. Other locations differ between different entities, and do not have significant discrepancies. 2-aminomuconate(2-) (CHEBI:57937), FADH2(2-) (CHEBI:58307) and methacrylyl-CoA (CHEBI:27754) are active in many different organelles. N-acetylputrescinium (CHEBI:58263), ammonioacetone (CHEBI:58320), methylammonium (CHEBI:59338), dopaminium(1+) (CHEBI:59905), dATP(4-) (CHEBI:61404) and CMP(2_) (CHEBI:60377) are more ac-

Table 4.4: Entities with the highest number of unique HMRs found by LitWay in the scientific literature. The entities in grey are also appearing in Table 4.3.

Entity	Name	Unique HMR number
CHEBI:57336	2-methylbutanoyl-CoA(4_)	175
CHEBI:57335	2-methylacetoacetyl-CoA(4_)	170
CHEBI:26174	poly(N-acetyllactosamine)	163
CHEBI:22783	β-D-galactosyl-(1_3)-N-acetyl-D-galactosaminyl group	161
CHEBI:27979	all-cis-icosa-8,11,14-trienoyl-CoA	159
CHEBI:545959	homovanillic acid	149
CHEBI:57604	3-phosphonato-D-glyceroyl phosphate(4_)	137
CHEBI:60880	3'-AMP(2_)	136
CHEBI:58524	gibberellin A1(1_)	128
CHEBI:58272	3-phosphonato-D-glycerate(3_)	116
CHEBI:2504	aflatoxin B_1	114
CHEBI:58307	FADH2(2-)	97
CHEBI:57937	2-aminomuconate(2-)	97
CHEBI:27754	methacrylyl-CoA	97
CHEBI:57355	4-coumaroyl-CoA(4-)	94
CHEBI:61085	3-sulfino-L-alanine(1-)	73
CHEBI:58017	5-O-phosphonato-α-D-ribofuranosyl diphosphate(5_)	72
CHEBI:57386	octanoyl-CoA(4-)	70
CHEBI:58601	α_-D-glucose 1-phosphate(2_)	69
CHEBI:57315	(R)-3-hydroxybutanoyl-CoA(4_)	64
CHEBI:57450	5(S)-HPETE(1_)	62
CHEBI:57463	leukotriene A4(1_)	61
CHEBI:57453	(6S)-5,6,7,8-tetrahydrofolate(2_)	60
CHEBI:61404	dATP(4-)	59
CHEBI:60377	CMP(2_)	58
CHEBI:59905	dopaminium(1+)	58
CHEBI:59338	methylammonium	58
CHEBI:58320	ammonioacetone	58
CHEBI:58263	N-acetylputrescinium	58
CHEBI:58235	glycocholate	58

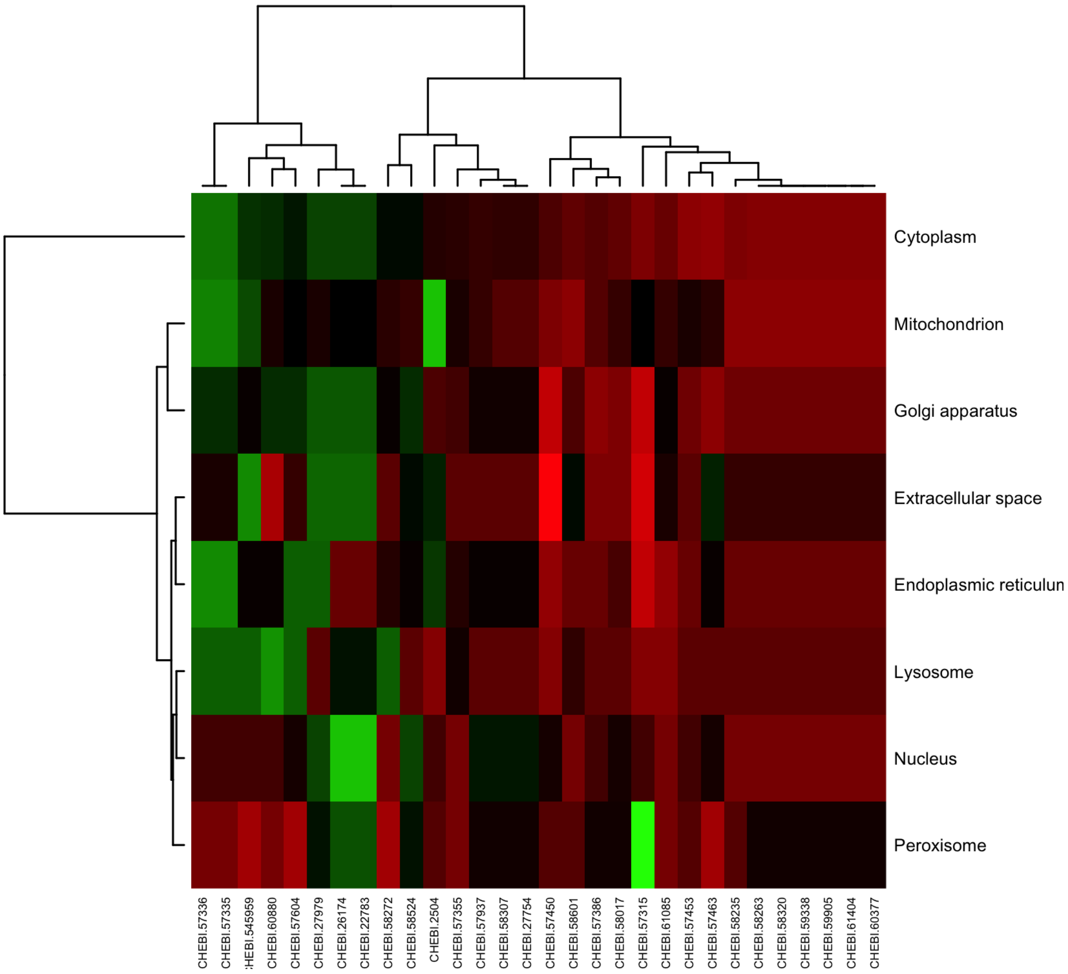


Fig. 4.15: Entities with the highest number of unique HMRs in different cellular components found by LitWay. The horizontal axis is the list of the entities, which are those listed in Table 4.4. The vertical axis is the list of locations (Table 4.1). Each grid represents the number of an entity's unique HMRs in a cellular component found by LitWay. From green to red, the spectrum indicates the number from low to high.

tive in peroxisome than in other organelles. This phenomenon is more obvious for that 2-methylbutanoyl-CoA(4_) (CHEBI:57336), 2-methylacetoacetyl-CoA(4_) (CHEBI:57335) and homovanillic acid (CHEBI:545959) are active in peroxisome in spite of that they have higher number of different interacting entities.

Fig. 4.16's horizontal axis is as same as the horizontal axis in Fig. 4.15, which is the list of the entities in Table 4.4. Its vertical axis is different, and is the list of ninety-nine sub-pathways of RECON2 (see the full list of the sub-pathways in Figure 2 in [43]). Each block in the heat map is the number of unique interactions the entity participating in the sub-pathway. From light colour (green) to dark colour (red), the spectrum indicates the number of unique interactions from low to high.

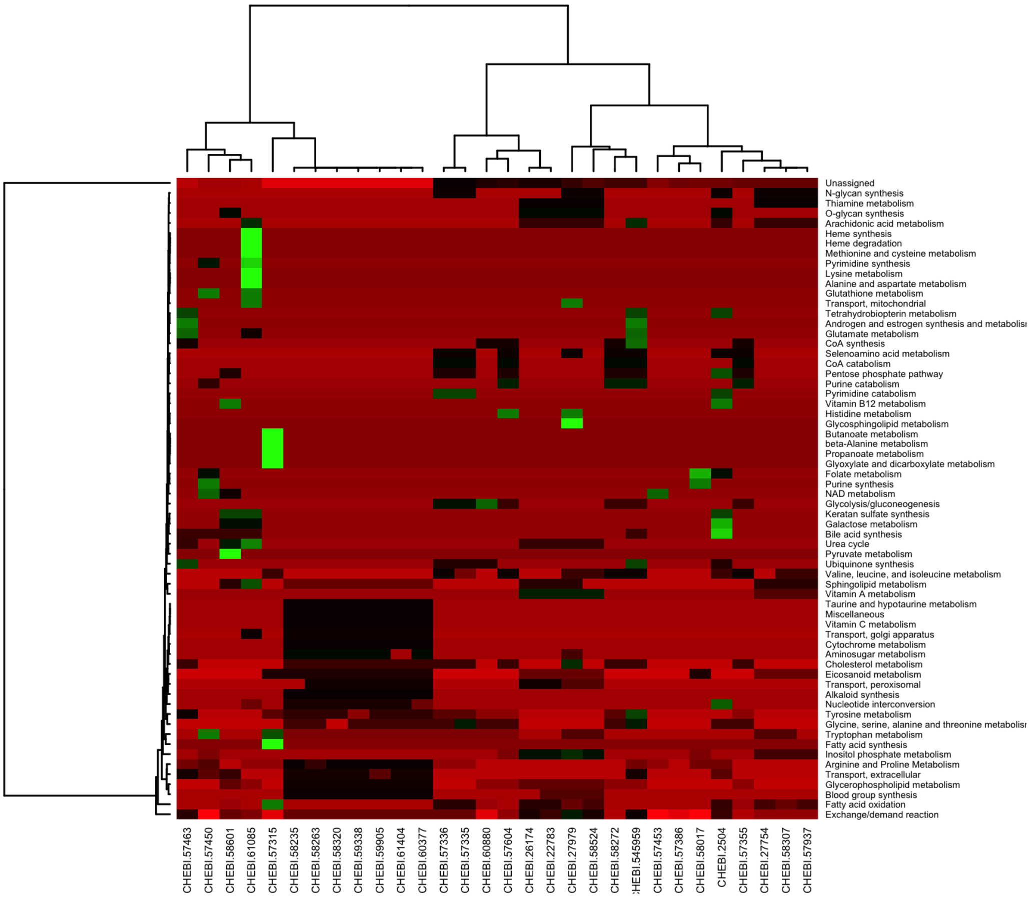


Fig. 4.16: Entities with the highest number of unique HMRs in different sub-pathways found by LitWay. The horizontal axis is the list of the entities, which are those listed in Table 4.4. The vertical axis is the list of the related sub-pathways in RECON2. Each grid represents the number of an entity's related unique HMRs in a sub-pathway found by LitWay. From green to red, the spectrum indicates the number from low to high.

Fig. 4.16 gives a higher granularity about molecular activities. We do not only see how active each entity by visualised spectrum, but also know the generality of its activity in each sub-pathway, which means it can be known that an entity is only active in one biological process or behaves similar in other metabolic processes as well. For example, glycocholate acid (CHEBI:58235) can be seen active in several pathways including vitamin A metabolism, taurine and hypotaurine metabolism and also cholesterol metabolism.

4.4 Discussion and conclusion

Different biological networks involve different entity types and reaction types. For instance, it is empirically thought that signalling pathway mainly contains PPIs and metabolic pathway involves CPIs. Such entity composition and reaction composition is presumed to be reflected in the descriptions in the scientific literature. This knowledge is important for guiding the development of an IR and/or IE system of biological network, which has been proven in the previous chapter. For example, the experiments on feature selection and constraints show that there are significantly diverse effects for different bio-events.

The curated biological networks from BMDB are evaluated by using the methods I developed, which ranges from high-recall to high-precision. The networks cover a comprehensive range of network types and sizes. The entity composition analysis shows that signalling pathways have higher percentage of entities, which is protein, and metabolic network has higher percentage of chemical entities. In certain extent, this is inline with empirical assumptions. However, the distinction is not as absolute as the assumption, and signalling pathways also appear to involve a large number of chemical entities in the text as well.

It is observed that 46% of signalling pathways involve CPIs and 5% of them involve only chemical reactions. Secondly, a large portion of metabolic networks involve protein interactions (33% CPI networks and 9% protein-only networks). These may seem contradicting the presumption. However, as all the networks used for the experiments are carefully curated to be correspondent with the original publications. Therefore, the result reflects the characteristics of such networks in the scientific literature.

A comprehensive comparison of the networks across types, sizes and extraction solutions is conducted. The results represent the recalls of LitWay for extracting different reactions from the scientific literature. As expected, CO2 produces better coverage (high recall) than LitWay, which is in general the case. Only when the entities can be discovered, the reactions can be recovered upon the entities. This is also in general the case, and it does not make a difference when the recovery exploits semantic information ("By Annotation") or metadata ("By Metadata"). Signalling pathways have a very good coverage on entities, but still would only produce about 50% of the relations (CO) or less when LitWays is used. During the experiment, it can be seen that the exploitation of semantic information improves the recalls.

Network size does not show obvious correlation with the coverage in the scientific literature, although, in many cases, medium size networks (entity number between 5 and 20) seem slightly higher. This shows that most available biological networks from the structured database are curated from the published knowledge, which is in unstructured text but has informative description about the corresponding biological processes.

By using LitWay, an experiment is conducted on the big network, RECON2, for discovering molecular relation, which have been reported in the scientific literature but not encoded in RECON2. By using information of metadata and annotation, the alignment between the sub-pathways of RECON2 and extracted entities by TM showed a good overlap between the two sources.

LitWay is used to investigate molecular relations, which have not been encoded in RECON2. The extracted evidential statements are firstly quantified for each discovered reaction, which show the overlap between hypothesised reaction and evidential statements from the scientific literature. Meanwhile, it represents the sufficiency of evidence or, in some way, the confidence of each discovered reaction.

The discovered reactions are then quantified to demonstrate the molecular activities. Some molecules play more active roles than the others. The activities are analysed from two dimensions: cellular components and sub-pathways. It further elaborates the activity of the molecules in the discovered relations.

Chapter 5

Conclusion

Digitally available literatures exponentially increase in last two decades, within which scientific literature is one principle resource of published knowledge. For instance, MEDLINE is updated from Tuesday to Saturday since 2003 and between 2,000 4,000 completed references are added daily [205]. This capacity has been far beyond what manual curation can achieve. Text mining (TM) based on natural language processing (NLP) and machine learning are steadily advancing. However, TM is still rarely used for gathering evidence for causative effects of gene-disease associations, for example, the modification of a molecular function of a protein or the dysfunction of an anatomical component. Meanwhile, there is no work using information mined from scientific literature to align with sophisticate biological networks, such as signalling pathways or metabolic pathways.

This study endeavours investigating TM-aided network evaluation, related challenges, and solutions. I develop new solutions and exploits available solutions to extract evidential statements from the scientific literature to unravel molecular function and evaluate biological networks, including signalling pathways and metabolic pathways. Aligning extracted knowledge with contents of curated biological networks evaluates the TM solutions, and supports understanding functional information of entities, identifying causative entity or reaction in biological phenomenon or disease, cross-linking data entries, examining erroneous information in database, and updating existing knowledge. Hidden molecular relation, which may be pathogenetic but have not been reported due to various reasons, may be discovered or inferred based on evidences collected from the scientific literature. Meanwhile, for NLP-based TM, the evaluation delivers extensive knowledge about distribution of entities and reactions in scientific literature.

A biological network is represented in a graph structure composed of nodes denoting biomolecules and edges between the nodes representing interactions or reactions between the biomolecules. Therefore, identifying entities, including protein and chemical, in text is the fundamental task of biological network extraction. By looking into different biological named entity recognition (NER) solutions [44], I investigate and deliver the knowledge about bio-molecular existence and formation in free-text, and about the most up-to-date solutions of identifying and normalising biomolecules (Chapter 2).

Different approaches of NER solutions have own advantages and disadvantages [44]. ML approaches usually perform well on datasets they been trained on. On the opposite, lexical/dictionary approach's performance is more consistent. For large corpus like MEDLINE, which obtains large update daily, ML approaches may not be the most suitable NER solutions. The evaluation shows that SwissProt tagger integrated in the Whatizit infrastructure [103] possesses state-of-the-art performance.

Event extraction (EE) links individual entities with their relationships. PCorral is a system for identifying molecular relations on sentential level of the entire MEDLINE. It utilises SwissProt tagger, therefore its EM and EN are up-to-date. It is able to extract molecular interactions by co-occurrence, tri-occurrence and syntactic patterns. CO is a larger set and the true superset to the CO3 results, and this is again a larger set and the true superset to the syntactic patterns. PCorral produces high recall results and also more specific results, which assists biologists to gain a comprehensive understanding about the entities and the reaction (called information retrieval) without technological bias, and narrows down selected documents to focus on specific reactions (called information extraction). Its interactive information retrieval is based on-the-fly processing. IntAct corpus has many reactions, which entities are not occurring on the same sentences. The performance seems low. On another hand, when being evaluated on BioCreative II, PCorral's pipeline including NER is used for the evaluation. It does not use the pre-tagged entities in order to test the system's capability of fully automated information extraction without manual efforts.

LitWay is a precise method of bio-EE. It is implemented with a search-based structured prediction algorithm, SEARN, and designed to be flexible and able to identify more types of complex reactions, including protein-protein interaction and protein-chemical interaction. The core of the system uses a set of collaborative classifiers to identify different parts of biological event. The evaluation on BioNLP'13 data proves the system to be one of the up-to-date systems. It can be easily customised for different tasks (Chapter 3).

NER and EE provide information for different layers of the biological network evaluation. Meanwhile, different approaches (CO2, CO3, SynP and ML) covers different aspects of the network evaluation. The evaluation starts with adopting the high-recall approach to evaluate the curated networks in BioModels Database [8] and quantitatively analyse the information composition and morphology of the networks including the metabolic pathways and the signalling pathways in the scientific literature. The work objectively quantifies evidential information of biomolecules, their relations and involving types of biological pathways in scientific literature [193]. Subsequently, diverse TM methods, from high-recall driven (co-occurrence and tri-occurrence) to high-precision driven (syntactic parsing and pattern), are employed for knowledge acquisition and evaluation of biological network [40].

The experiment shows overlaps between PCI and PPI, and between protein and chemical. The overlaps demand different methods and lead to different results in the network extraction by TM. The entity compositions of the signalling pathways and the metabolic pathways in BMDB are in-line with the empirical knowledge. The signalling pathways involve more proteins, and the metabolic networks involve more chemical entities. However, the signalling pathways also have a large portion of chemical entities. Generally, the PPIs have a better coverage in the scientific literature in comparison with the CPIs. The signalling pathways are better analysed and better conserved, i.e. they have more coverage in the literature concerning entities and relations. The metabolic pathways have less coverage in the scientific literature. Meanwhile, network sizes do not have significant correlation with the coverage in the scientific literature.

The experiment of evaluation is extended to investigate a comprehensive human metabolic network, RECON2 [43]. To evaluate the huge network, which consists of 1,789 enzyme-encoding genes, 7,440 reactions and 2,626 unique metabolites, PCorral and LitWay are used for different aspects of the analysis. Firstly, entity coverage in the sub-pathways quantify the maximum information about the network available from the scientific literature. PCorral's methods, from high-recall (CO2) to high-precision (SynP), are used to compare different types of manually curated evidences of reactions with available evidences from the scientific literature.

It has been long accepted that one-gene-one-protein-one-function is an over simplified theory. One gene can be expressed to different proteins, e.g. alternative splicing. One protein can also play different roles in different organisms (e.g. cellular components) or systems (e.g. pathways). I demonstrate using LitWay to collect evidential statements of hypothesised molecular relations and analyse molecular activities in different cellular components and sub-pathways of human metabolism (Chapter 4).

The analysis shows that, in peroxisome, the most entities can interact with many different entities. Other locations differ between different entities, and do not have significant discrepancies. 2-aminomuconate(2-) (CHEBI:57937), FADH2(2-) (CHEBI:58307) and methacrylyl-CoA (CHEBI:27754) are active in many different organelles. N-acetylputrescinium

(CHEBI:58263), ammonioacetone (CHEBI:58320), methylammonium (CHEBI:59338), dopaminium(1+) (CHEBI:59905), dATP(4-) (CHEBI:61404) and CMP(2_) (CHEBI:60377) are more active in peroxisome than in other organelles. This phenomenon is more obvious for that 2-methylbutanoyl-CoA(4_) (CHEBI:57336), 2-methylacetoacetyl-CoA(4_) (CHEBI:57335) and homovanillic acid (CHEBI:545959) are active in peroxisome in spite of that they have higher number of different interacting entities. The analysis also shows that glycocholate acid (CHEBI:58235) are active in several pathways including vitamin A metabolism, taurine and hypotaurine metabolism and also cholesterol metabolism. These observations can be further evaluated in the web-lab experiments in the future.

As discussed before, a network is different from a set of events and more informative. In order to be able to extract biological networks, the current challenge still remain on event extraction. The best performing system in BioNLP achieves less than 60% F-score. Besides the performance limitation of the system, coreference in the text is also a prominent issue. In the future, firstly, the current system, which is based on LitWay infrastructure, will be improved. Then, it will leverage the methods of linguistic discourse, of which coreference is a sub-type. Based on co-referring expressions and context, the system will try to be able to characterise the certainty level of extracted relations/facts. These are essential for synthesising events to produce interconnected networks.

While focusing on improving event extraction technology, the initial objective of text mining should not be overlooked. It could have two folds. Firstly, automatically gathering evidences for causative effects for selected gene-disease associations from published knowledge is an highly demanded use case, which has not been well served. Furthermore, once performance of event extraction and even normalisation reaches certain reliability, text mining should later aim at interlinking data extracted from the scientific literature and semantic databases to automatically construct networks of complicated biological conditions. The interlinked knowledge also has potential to be abstracted into models for simulating the conditions for pharmaceutical purposes, e.g. blood homeostasis and cellular trafficking, which are related with the disease states of diabetes, heart disease and Alzheimer's disease.

References

[1] M. Kanehisa, S. Goto, M. Hattori, K. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa, "From genomics to chemical genomics: new developments in KEGG," *Nucleic Acids Research*, vol. 34, no. suppl 1, pp. D354–D357, 2006.

[2] P. D'Eustachio, "Reactome knowledgebase of human biological pathways and processes.," *Methods in molecular biology (Clifton, N.J.)*, vol. 694, pp. 49–61, 2011.

[3] C. Schaefer, K. Anthony, S. Krupa, J. Buchoff, M. Day, T. Hannay, and K. Buetow, "PID: the pathway interaction database.," *Nucleic acids research*, vol. 37, no. Database issue, pp. D674–D679, 2009.

[4] A. Hodgkin and A. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve.," *The Journal of physiology*, vol. 117, no. 4, pp. 500–544, 1952.

[5] M. Hucka, A. Finney, H. Sauro, H. Bolouri, J. Doyle, H. Kitano, , the rest of the SBML Forum:, A. Arkin, B. Bornstein, D. Bray, A. Cornish-Bowden, A. Cuellar, S. Dronov, E. Gilles, M. Ginkel, V. Gor, I. Goryanin, W. Hedley, T. Hodgman, J. Hofmeyr, P. Hunter, N. Juty, J. Kasberger, A. Kremling, U. Kummer, N. Le Novère, L. Loew, D. Lucio, P. Mendes, E. Minch, E. Mjolsness, Y. Nakayama, M. Nelson, P. Nielsen, T. Sakurada, J. Schaff, B. Shapiro, T. Shimizu, H. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, and J. Wang, "The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models," *Bioinformatics*, vol. 19, no. 4, pp. 524–531, 2003.

[6] C. Lloyd, M. Halstead, and P. Nielsen, "CellML: its future, present and past.," *Progress in biophysics and molecular biology*, vol. 85, no. 2-3, pp. 433–450, 2004.

[7] E. Demir, M. Cary, S. Paley, K. Fukuda, C. Lemer, I. Vastrik, G. Wu, P. D'Eustachio, C. Schaefer, J. Luciano, F. Schacherer, I. Martinez-Flores, Z. Hu, V. Jimenez-Jacinto, G. Joshi-Tope, K. Kandasamy, A. Lopez-Fuentes, H. Mi, E. Pichler, I. Rodchenkov, A. Splendiani, S. Tkachev, J. Zucker, G. Gopinath, H. Rajasimha, R. Ramakrishnan, I. Shah, M. Syed, N. Anwar, O. Babur, M. Blinov, E. Brauner, D. Corwin, S. Donaldson, F. Gibbons, R. Goldberg, P. Hornbeck, A. Luna, P. Murray-Rust, E. Neumann, O. Ruebenacker, O. Reubenacker, M. Samwald, M. van Iersel, S. Wimalaratne, K. Allen, B. Braun, M. Whirl-Carrillo, K.-H. Cheung, K. Dahlquist, A. Finney, M. Gillespie, E. Glass, L. Gong, R. Haw, M. Honig, O. Hubaut, D. Kane,

S. Krupa, M. Kutmon, J. Leonard, D. Marks, D. Merberg, V. Petri, A. Pico, D. Ravenscroft, L. Ren, N. Shah, M. Sunshine, R. Tang, R. Whaley, S. Letovksy, K. Buetow, A. Rzhetsky, V. Schachter, B. Sobral, U. Dogrusoz, S. McWeeney, M. Aladjem, E. Birney, J. Collado-Vides, S. Goto, M. Hucka, N. Le Novère, N. Maltsev, A. Pandey, P. Thomas, E. Wingender, P. Karp, C. Sander, and G. Bader, "The BioPAX community standard for pathway data sharing.," *Nature biotechnology*, vol. 28, no. 9, pp. 935–942, 2010.

[8] C. Li, M. Donizelli, N. Rodriguez, H. Dharuri, L. Endler, V. Chelliah, L. Li, E. He, A. Henry, M. Stefan, J. Snoep, M. Hucka, N. Le Novere, and C. Laibe, "BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models," *BMC Systems Biology*, vol. 4, no. 1, p. 92, 2010.

[9] C. Lloyd, J. Lawson, P. Hunter, and P. Nielsen, "The CellML model repository," *Bioinformatics*, vol. 24, no. 18, pp. 2122–2123, 2008.

[10] WikiPathways, "http://www.wikipathways.org."

[11] C. Li, M. Courtot, N. Le Novère, and C. Laibe, "BioModels.net Web Services, a free and integrated toolkit for computational modelling software.," *Briefings in bioinformatics*, vol. 11, no. 3, pp. 270–277, 2010.

[12] T. Ohta, S. Pyysalo, and J. Tsujii, "From pathways to biomolecular events: Opportunities and challenges," in *Proceedings of BioNLP 2011 Workshop*, BioNLP '11, (Stroudsburg, PA, USA), pp. 105–113, Association for Computational Linguistics, 2011.

[13] L. Jensen, J. Saric, and P. Bork, "Literature mining for the biologist: from information retrieval to biological discovery," *Nat Rev Genet*, vol. 7, no. 2, pp. 119–129, 2006.

[14] M. Krallinger, R. A.-A. Erhardt, and A. Valencia, "Text-mining approaches in molecular biology and biomedicine.," *Drug discovery today*, vol. 10, no. 6, pp. 439–445, 2005.

[15] M. Krallinger, F. Leitner, C. Rodriguez-Penagos, and A. Valencia, "Overview of the protein-protein interaction annotation extraction task of biocreative ii.," *Genome biology*, vol. 9 Suppl 2, no. Suppl 2, p. S4, 2008.

[16] S. Ananiadou, S. Pyysalo, J. Tsujii, and D. Kell, "Event extraction for systems biology by text mining the literature.," *Trends in biotechnology*, vol. 28, no. 7, pp. 381–390, 2010.

[17] 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, no. suppl 1, pp. S74–S82, 2001.

[18] D. Yao, J. Wang, Y. Lu, N. Noble, H. Sun, X. Zhu, N. Lin, D. G. Payan, M. Li, and K. Qu, "PathwayFinder: paving the way towards automatic pathway extraction," in *Proceedings of the second conference on Asia-Pacific bioinformatics-Volume 29*, pp. 53–62, Australian Computer Society, Inc., 2004.

[19] M. Frisch, B. Klocke, M. Haltmeier, and K. Frech, “LitInspector: literature and signal transduction pathway mining in PubMed abstracts,” *Nucleic acids research*, vol. 37, no. suppl 2, pp. W135–W140, 2009.

[20] B. Kemper, T. Matsuzaki, Y. Matsuoka, Y. Tsuruoka, H. Kitano, S. Ananiadou, and J. Tsujii, “PathText: a text mining integrator for biological pathway visualizations,” *Bioinformatics*, vol. 26, no. 12, pp. i374–i381, 2010.

[21] A. Barbosa-Silva, J.-F. Fontaine, E. Donnard, F. Stussi, J. Ortega, and M. Andrade-Navarro, “Pescador, a web-based tool to assist text-mining of biointeractions extracted from pubmed queries,” *BMC bioinformatics*, vol. 12, no. 1, p. 435, 2011.

[22] A. Yeh, A. Morgan, M. Colosimo, and L. Hirschman, “BioCreAtIvE task 1A: gene mention finding evaluation,” *BMC bioinformatics*, vol. 6, no. Suppl 1, p. S2, 2005.

[23] K. Verspoor, K. B. Cohen, A. Lanfranchi, C. Warner, H. L. Johnson, C. Roeder, J. D. Choi, C. Funk, Y. Malenkiy, M. Eckert, *et al.*, “A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools,” *BMC bioinformatics*, vol. 13, no. 1, p. 207, 2012.

[24] J.-D. Kim, S. Pyysalo, T. Ohta, R. Bossy, N. Nguyen, and J. Tsujii, “Overview of bionlp shared task 2011,” in *Proceedings of the BioNLP Shared Task 2011 Workshop*, pp. 1–6, Association for Computational Linguistics, 2011.

[25] I. Spasić, E. Simeonidis, H. L. Messiha, N. W. Paton, and D. B. Kell, “KiPar, a tool for systematic information retrieval regarding parameters for kinetic modelling of yeast metabolic pathways,” *Bioinformatics*, vol. 25, no. 11, pp. 1404–1411, 2009.

[26] S. Heinen, B. Thielen, and D. Schomburg, “KID-an algorithm for fast and efficient text mining used to automatically generate a database containing kinetic information of enzymes,” *BMC bioinformatics*, vol. 11, no. 1, p. 375, 2010.

[27] J.-J. Tsay, B.-L. Wu, and C.-C. Hsieh, “Automatic extraction of kinetic information from biochemical literatures,” in *Fuzzy Systems and Knowledge Discovery, 2009. FSKD’09. Sixth International Conference on*, vol. 5, pp. 28–32, IEEE, 2009.

[28] C. Nédellec, R. Bossy, J.-D. Kim, J.-j. Kim, T. Ohta, S. Pyysalo, and P. Zweigenbaum, “Overview of BioNLP Shared Task 2013,” *ACL 2013*, p. 1, 2013.

[29] V. Benagiano, L. Lorusso, P. Flace, F. Girolamo, A. Rizzi, L. Bosco, R. Cagiano, B. Nico, D. Ribatti, and G. Ambrosi, “VAMP-2, SNAP-25A/B and syntaxin-1 in glutamatergic and GABAergic synapses of the rat cerebellar cortex,” *BMC neuroscience*, vol. 12, no. 1, p. 118, 2011.

[30] M. Miwa, P. Thompson, J. McNaught, D. B. Kell, and S. Ananiadou, “Extracting semantically enriched events from biomedical literature,” *BMC bioinformatics*, vol. 13, no. 1, p. 108, 2012.

[31] M. Liakata, S. Saha, S. Dobnik, C. Batchelor, and D. Rebholz-Schuhmann, “Automatic recognition of conceptualization zones in scientific articles and two life science applications,” *Bioinformatics*, vol. 28, no. 7, pp. 991–1000, 2012.

[32] K. Markert and U. Hahn, "On the interaction of metonymies and anaphora," in *Proceedings of International Joint Conferences on Artificial Intelligence*, pp. 1010–1015, 1997.

[33] D. Stallard, "Two kinds of metonymy," in *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, pp. 87–94, Association for Computational Linguistics, 1993.

[34] H.-W. Chun, Y. Tsuruoka, J.-D. Kim, R. Shiba, N. Nagata, T. Hishiki, and J. Tsujii, "Extraction of gene-disease relations from Medline using domain dictionaries and machine learning.," in *Pacific Symposium on Biocomputing*, vol. 11, pp. 4–15, 2006.

[35] J. Hakenberg, C. Plake, R. Leaman, M. Schroeder, and G. Gonzalez, "Inter-species normalization of gene mentions with GNAT," *Bioinformatics*, vol. 24, no. 16, pp. i126–i132, 2008.

[36] V. Gorgoulis, D. Aninos, P. Mikou, P. Kanavaros, A. Karameris, J. Joardanoglou, A. Rasidakis, M. Veslemes, B. Ozanne, and D. Spandidos, "Expression of EGF, TGF-alpha and EGFR in squamous cell lung carcinomas.," *Anticancer research*, vol. 12, no. 4, p. 1183, 1992.

[37] B. N. Kholodenko, O. V. Demin, G. Moehren, and J. B. Hoek, "Quantification of short term signaling by the epidermal growth factor receptor," *Journal of Biological Chemistry*, vol. 274, no. 42, pp. 30169–30181, 1999.

[38] J.-D. Kim, T. Ohta, S. Pyysalo, Y. Kano, and J. Tsujii, "Overview of BioNLP'09 shared task on event extraction," in *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, pp. 1–9, Association for Computational Linguistics, 2009.

[39] R. Hoffmann, M. Krallinger, E. Andres, J. Tamames, C. Blaschke, and A. Valencia, "Text mining for metabolic pathways, signaling cascades, and protein networks," *Science Signaling*, vol. 2005, no. 283, p. pe21, 2005.

[40] C. Li, A. Jimeno-Yepes, M. Arregui, H. Kirsch, and D. Rebholz-Schuhmann, "PCorral—interactive mining of protein interactions from MEDLINE," *Database: the journal of biological databases and curation*, vol. 2013, 2013.

[41] LitWay Event Extraction System, "https://github.com/li-chen/ee."

[42] Apache UIMA, "http://uima.apache.org."

[43] I. Thiele, N. Swainston, R. M. Fleming, A. Hoppe, S. Sahoo, M. K. Aurich, H. Haraldsdottir, M. L. Mo, O. Rolfsson, M. D. Stobbe, *et al.*, "A community-driven global reconstruction of human metabolism," *Nature biotechnology*, vol. 31, no. 5, pp. 419–425, 2013.

[44] D. Rebholz-Schuhmann, S. Kafkas, J.-H. Kim, C. Li, A. J. Yepes, R. Hoehndorf, R. Backofen, and I. Lewin, "Evaluating gold standard corpora against gene/protein tagging solutions and lexical resources," *Journal of biomedical semantics*, vol. 4, no. 1, p. 28, 2013.

[45] D. Rebholz-Schuhmann, A. J. Yepes, C. Li, S. Kafkas, I. Lewin, N. Kang, P. Corbett, D. Milward, E. Buyko, E. Beisswanger, *et al.*, "Assessment of NER solutions against the first and second CALBC Silver Standard Corpus," *Journal of Biomedical Semantics*, vol. 2, no. 5, pp. 1–12, 2011.

[46] S. Croset, C. Grabmüller, C. Li, S. Kavaliauskas, and D. Rebholz-Schuhmann, "The calbc rdf triple store: retrieval over large literature content," *arXiv preprint arXiv:1012.1650*, 2010.

[47] P. Thompson, J. McNaught, S. Montemagni, N. Calzolari, R. del Gratta, V. Lee, S. Marchi, M. Monachini, P. Pezik, V. Quochi, C. Rupp, Y. Sasaki, G. Venturi, D. Rebholz-Schuhmann, and S. Ananiadou, "The BioLexicon: a large-scale terminological resource for biomedical text mining.," *BMC bioinformatics*, vol. 12, no. 1, p. 397, 2011.

[48] J.-J. Kim and D. Rebholz-Schuhmann, "Improving the extraction of complex regulatory events from scientific text by using ontology-based inference.," *Journal of biomedical semantics*, vol. 2 Suppl 5, p. S3, Oct 2011.

[49] K. Tomanek and U. Hahn, "Semi-supervised active learning for sequence labeling," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, (Stroudsburg, PA, USA), pp. 1039–1047, Association for Computational Linguistics, 2009.

[50] A. McCallum, D. Freitag, and F. C. N. Pereira, "Maximum Entropy Markov Models for Information Extraction and Segmentation," in *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, (San Francisco, CA, USA), pp. 591–598, Morgan Kaufmann Publishers Inc., 2000.

[51] R. McDonald and F. Pereira, "Identifying gene and protein mentions in text using conditional random fields," *BMC Bioinformatics*, vol. 6, no. Suppl 1, p. S6, 2005.

[52] B. Settles, "ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text.," *Bioinformatics (Oxford, England)*, vol. 21, no. 14, pp. 3191–3192, 2005.

[53] C.-N. Hsu, Y.-M. Chang, C.-J. Kuo, Y.-S. Lin, H.-S. Huang, and I.-F. Chung, "Integrating high dimensional bi-directional parsing models for gene mention tagging," *Bioinformatics*, vol. 24, no. 13, pp. i286–i294, 2008.

[54] Y. Sasaki, Y. Tsuruoka, J. McNaught, and S. Ananiadou, "How to make the most of ne dictionaries in statistical NER.," *BMC bioinformatics*, vol. 9 Suppl 11, 2008.

[55] S. Saha, A. Ekbal, and S. Saha, "A supervised approach for gene mention detection," in *Proceedings of the Second International Conference on Swarm, Evolutionary, and Memetic Computing - Volume Part I*, SEMCCO'11, (Berlin, Heidelberg), pp. 425–432, Springer-Verlag, 2011.

[56] L. Shi and F. Campagne, "Building a protein name dictionary from full text: a machine learning term extraction approach.," *BMC bioinformatics*, vol. 6, no. 1, 2005.

[57] J. Kaiser, “Proteomics. public-private group maps out initiatives.,” *Science*, vol. 296, no. 5569, 2002.

[58] M. Neves, J. Carazo, and A. Montano, “Moara: a Java library for extracting and normalizing gene and protein mentions,” *BMC Bioinformatics*, vol. 11, no. 1, p. 157, 2010.

[59] M. Huang, J. Liu, and X. Zhu, “GeneTUKit: a software for document-level gene normalization.,” *Bioinformatics (Oxford, England)*, vol. 27, no. 7, pp. 1032–1033, 2011.

[60] Y. Chen, F. Liu, and B. Manderick, “A machine learning–based system to normalise gene mentions to unique database identifiers,” *International journal of data mining and bioinformatics*, vol. 5, no. 6, pp. 640–660, 2011.

[61] NCBI Entrez Gene, “http://www.ncbi.nlm.nih.gov/gene.”

[62] C. Kolárik, R. Klinger, C. M. Friedrich, M. Hofmann-Apitius, and J. Fluck, “Chemical names: terminological resources and corpora annotation,” in *Workshop on Building and evaluating resources for biomedical text mining (6th edition of the Language Resources and Evaluation Conference)*, vol. 36, 2008.

[63] P. Corbett and P. Murray-Rust, “High-throughput identification of chemistry in life science texts,” in *Computational Life Sciences II*, pp. 107–118, Springer, 2006.

[64] D. M. Jessop, S. E. Adams, E. L. Willighagen, L. Hawizy, and P. Murray-Rust, “OSCAR4: a flexible architecture for chemical text-mining,” *Journal of cheminformatics*, vol. 3, no. 1, pp. 1–12, 2011.

[65] T. Rocktäschel, M. Weidlich, and U. Leser, “ChemSpot: a hybrid system for chemical named entity recognition,” *Bioinformatics*, vol. 28, no. 12, pp. 1633–1640, 2012.

[66] L. Hirschman, A. Yeh, C. Blaschke, and A. Valencia, “Overview of BioCreAtIvE: critical assessment of information extraction for biology,” *BMC bioinformatics*, vol. 6, no. Suppl 1, p. S1, 2005.

[67] L. Smith, L. K. Tanabe, R. J. Ando, C.-J. Kuo, I.-F. Chung, C.-N. Hsu, Y.-S. Lin, R. Klinger, C. M. Friedrich, K. Ganchev, *et al.*, “Overview of BioCreative II gene mention recognition,” *Genome Biology*, vol. 9, no. Suppl 2, p. S2, 2008.

[68] A. A. Morgan, Z. Lu, X. Wang, A. M. Cohen, J. Fluck, P. Ruch, A. Divoli, K. Fundel, R. Leaman, J. Hakenberg, *et al.*, “Overview of BioCreative II gene normalization,” *Genome biology*, vol. 9, no. Suppl 2, p. S3, 2008.

[69] C. Blaschke, E. A. Leon, M. Krallinger, and A. Valencia, “Evaluation of BioCreAtIvE assessment of task 2,” *BMC bioinformatics*, vol. 6, no. Suppl 1, p. S16, 2005.

[70] P. De Matos, R. Alcántara, A. Dekker, M. Ennis, J. Hastings, K. Haug, I. Spiteri, S. Turner, and C. Steinbeck, “Chemical entities of biological interest: an update,” *Nucleic acids research*, vol. 38, no. suppl 1, pp. D249–D254, 2010.

[71] European Patent Office and ChEBI annotated chemical corpus, “http://chebi.cvs.sourceforge.net/viewvc/chebi/chapati/patentsgoldstandard/.”

[72] M. Bada, M. Eckert, D. Evans, K. Garcia, K. Shipley, D. Sitnikov, W. A. Baumgartner, K. B. Cohen, K. Verspoor, J. A. Blake, *et al.*, “Concept annotation in the craft corpus,” *BMC bioinformatics*, vol. 13, no. 1, p. 161, 2012.

[73] U. Consortium *et al.*, “Update on activities at the Universal Protein Resource (UniProt) in 2013,” *Nucleic acids research*, vol. 41, no. D1, pp. D43–D47, 2013.

[74] World Wide Protein Data Bank, “http://www.wwpdb.org/.”

[75] M. Punta, P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, *et al.*, “The Pfam protein families database,” *Nucleic acids research*, vol. 40, no. D1, pp. D290–D301, 2012.

[76] Y. Wang, J. Xiao, T. O. Suzek, J. Zhang, J. Wang, and S. H. Bryant, “PubChem: a public information system for analyzing bioactivities of small molecules,” *Nucleic acids research*, vol. 37, no. suppl 2, pp. W623–W633, 2009.

[77] CAS Registry, “http://www.cas.org/content/chemical-substances.”

[78] Chemical category of MeSH, “http://www.nlm.nih.gov/mesh/.”

[79] Reaxys, “http://www.info.reaxys.com.”

[80] K. Tipton and S. Boyce, “History of the enzyme nomenclature system,” *Bioinformatics*, vol. 16, no. 1, pp. 34–40, 2000.

[81] D. S. Wishart, C. Knox, A. C. Guo, R. Eisner, N. Young, B. Gautam, D. D. Hau, N. Psychogios, E. Dong, S. Bouatra, *et al.*, “Hmdb: a knowledgebase for the human metabolome,” *Nucleic acids research*, vol. 37, no. suppl 1, pp. D603–D610, 2009.

[82] KEGG Compound, “http://www.genome.jp/kegg/compound/.”

[83] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, *et al.*, “ChEMBL: a large-scale bioactivity database for drug discovery,” *Nucleic acids research*, vol. 40, no. D1, pp. D1100–D1107, 2012.

[84] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, *et al.*, “DrugBank 3.0: a comprehensive resource for ‘omics’ research on drugs,” *Nucleic acids research*, vol. 39, no. suppl 1, pp. D1035–D1041, 2011.

[85] KEGG Drug, “http://www.genome.jp/kegg/drug/.”

[86] L. Hirschman, G. A. C. Burns, M. Krallinger, C. Arighi, K. B. Cohen, A. Valencia, C. H. Wu, A. Chatr-Aryamontri, K. G. Dowell, E. Huala, *et al.*, “Text mining for the biocuration workflow,” *Database: the journal of biological databases and curation*, vol. 2012, 2012.

[87] N. L. Washington, M. A. Haendel, C. J. Mungall, M. Ashburner, M. Westerfield, and S. E. Lewis, "Linking human diseases to animal models using ontology-based phenotype annotation," *PLoS biology*, vol. 7, no. 11, p. e1000247, 2009.

[88] A. Oellrich, R. Hoehndorf, G. V. Gkoutos, and D. Rebholz-Schuhmann, "Improving disease gene prioritization by comparing the semantic similarity of phenotypes in mice with those of human diseases," *PloS one*, vol. 7, no. 6, p. e38937, 2012.

[89] T. Ohta, S. Pyysalo, J. Tsujii, and S. Ananiadou, "Open-domain anatomical entity mention detection," in *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*, pp. 27–36, Association for Computational Linguistics, 2012.

[90] S. Pyysalo, T. Ohta, M. Miwa, H.-C. Cho, J. Tsujii, and S. Ananiadou, "Event extraction across multiple levels of biological organization," *Bioinformatics*, vol. 28, no. 18, pp. i575–i581, 2012.

[91] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, G. Gorrell, A. Funk, A. Roberts, D. Damljanovic, T. Heitz, M. A. Greenwood, H. Saggion, J. Petrak, Y. Li, and W. Peters, *Text Processing with GATE (Version 6).* 2011.

[92] F. Leitner, M. Krallinger, C. Rodriguez-Penagos, J. Hakenberg, C. Plake, C.-J. Kuo, C.-N. Hsu, R. Tsai, H.-C. Hung, W. W. Lau, *et al.*, "Introducing meta-services for biomedical information extraction," *Genome Biol*, vol. 9, no. Suppl 2, p. S6, 2008.

[93] D. Rebholz-Schuhmann, H. Kirsch, G. Nenadic, D. Rebholz-Schuhmann, H. Kirsch, and G. Nenadic, "Iexml: towards an annotation framework for biomedical semantic types enabling interoperability of text processing modules," *SIG BioLink, ISMB*, 2006.

[94] D. Rebholz-Schuhmann, A. J. Jimeno-Yepes, E. M. van Mulligen, N. Kang, J. A. Kors, D. Milward, P. Corbett, E. Buyko, K. Tomanek, E. Beisswanger, *et al.*, "The calbc silver standard corpus for biomedical named entities-a study in harmonizing the contributions from four independent named entity taggers.," in *LREC*, 2010.

[95] C. A. Guideline, "http://www.ebi.ac.uk/rebholz-srv/calbc/challenge_guideline.pdf."

[96] R. T.-H. Tsai, S.-H. Wu, W.-C. Chou, Y.-C. Lin, D. He, J. Hsiang, T.-Y. Sung, and W.-L. Hsu, "Various criteria in the evaluation of biomedical named entity recognition," *BMC bioinformatics*, vol. 7, no. 1, p. 92, 2006.

[97] O. Bodenreider and A. T. McCray, "Exploring semantic groups through visual approaches," *Journal of biomedical informatics*, vol. 36, no. 6, pp. 414–432, 2003.

[98] J. Kim, T. Ohta, Y. Tateisi, and J. Tsujii, "Genia corpus—a semantically annotated corpus for bio-textmining," *Bioinformatics*, vol. 19, no. suppl 1, pp. i180–i182, 2003.

[99] L. Tanabe, N. Xie, L. H. Thom, W. Matten, and W. J. Wilbur, "Genetag: a tagged corpus for gene/protein named entity recognition," *BMC bioinformatics*, vol. 6, no. Suppl 1, p. S3, 2005.

[100] S. Kulick, A. Bies, M. Liberman, M. Mandel, R. McDonald, M. Palmer, A. Schein, L. Ungar, S. Winters, and P. White, "Integrated annotation for biomedical information extraction," in *Proc. of the Human Language Technology Conference and the Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL)*, pp. 61–68, 2004.

[101] Y.-M. Chang, C.-J. Kuo, H.-S. Huang, Y.-S. Lin, and C.-N. Hsu, "Analysis and enhancement of conditional random fields gene mention taggers in biocreative ii challenge evaluation.," in *LBM (Short Papers)*, 2007.

[102] BANNER, "http://cbioc.eas.asu.edu/banner/."

[103] D. Rebholz-Schuhmann, M. Arregui, S. Gaudan, H. Kirsch, and A. Jimeno, "Text processing through Web services: calling Whatizit," *Bioinformatics*, vol. 24, no. 2, pp. 296–298, 2008.

[104] J. R. McEntyre, S. Ananiadou, S. Andrews, W. J. Black, R. Boulderstone, P. Buttery, D. Chaplin, S. Chevuru, N. Cobley, L.-A. Coleman, *et al.*, "Ukpmc: a full text article resource for the life sciences," *Nucleic acids research*, vol. 39, no. suppl 1, pp. D58–D65, 2011.

[105] D. Rebholz-Schuhmann, H. Kirsch, M. Arregui, S. Gaudan, M. Riethoven, and P. Stoehr, "EBIMed—text crunching to gather facts for proteins from Medline," *Bioinformatics*, vol. 23, no. 2, pp. e237–e244, 2007.

[106] H. Kirsch, S. Gaudan, and D. Rebholz-Schuhmann, "Distributed modules for text annotation and ie applied to the biomedical domain," *International journal of medical informatics*, vol. 75, no. 6, pp. 496–500, 2006.

[107] J. Hakenberg, C. Plake, L. Royer, H. Strobelt, U. Leser, and M. Schroeder, "Gene mention normalization and interaction extraction with context models and sentence motifs," *Genome Biol*, vol. 9, no. Suppl 2, p. S14, 2008.

[108] J.-D. Kim, T. Ohta, Y. Tsuruoka, Y. Tateisi, and N. Collier, "Introduction to the bio-entity recognition task at JNLPBA," in *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications*, pp. 70–75, Association for Computational Linguistics, 2004.

[109] U. Hahn, E. Beisswanger, E. Buyko, M. Poprat, K. Tomanek, and J. Wermter, "Semantic annotations for biology: a corpus development initiative at the jena university language &amp; information engineering (julie) lab.," in *LREC*, vol. 8, pp. 2257–2261, 2008.

[110] A. K. McCallum, "Mallet: A machine learning for language toolkit," 2002.

[111] A. Skusa, A. Rüegg, and J. Köhler, "Extraction of biological interaction networks from scientific literature," *Briefings in Bioinformatics*, vol. 6, no. 3, pp. 263–276, 2005.

[112] T.-K. Jenssen, A. Lægreid, 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.

[113] Y. Tsuruoka, J. Tsujii, and S. Ananiadou, "FACTA: a text search engine for finding associated biomedical concepts," *Bioinformatics*, vol. 24, no. 21, pp. 2559–2560, 2008.

[114] A. Barbosa-Silva, T. G. Soldatos, I. L. Magalhães, G. A. Pavlopoulos, J.-F. Fontaine, M. A. Andrade-Navarro, R. Schneider, and J. M. Ortega, "LAITOR-Literature Assistant for Identification of Terms co-Occurrences and Relationships," *BMC bioinformatics*, vol. 11, no. 1, p. 70, 2010.

[115] MEDLINE/PubMed, "http://www.ncbi.nlm.nih.gov/pubmed/."

[116] PubMed Central, "http://www.ncbi.nlm.nih.gov/pmc/."

[117] H. Chen and B. M. Sharp, "Content-rich biological network constructed by mining pubmed abstracts," *BMC bioinformatics*, vol. 5, no. 1, p. 147, 2004.

[118] R. Frijters, B. Heupers, P. van Beek, M. Bouwhuis, R. van Schaik, J. de Vlieg, J. Polman, and W. Alkema, "CoPub: a literature-based keyword enrichment tool for microarray data analysis," *Nucleic acids research*, vol. 36, no. suppl 2, pp. W406–W410, 2008.

[119] W. W. Fleuren, S. Verhoeven, R. Frijters, B. Heupers, J. Polman, R. van Schaik, J. de Vlieg, and W. Alkema, "CoPub update: CoPub 5.0 a text mining system to answer biological questions," *Nucleic acids research*, vol. 39, no. suppl 2, pp. W450–W454, 2011.

[120] R. Frijters, M. van Vugt, R. Smeets, R. van Schaik, J. de Vlieg, and W. Alkema, "Literature mining for the discovery of hidden connections between drugs, genes and diseases," *PLoS computational biology*, vol. 6, no. 9, p. e1000943, 2010.

[121] Y. Miyao and J. Tsujii, "Feature forest models for probabilistic HPSG parsing," *Computational Linguistics*, vol. 34, no. 1, pp. 35–80, 2008.

[122] T. Hara, Y. Miyao, and J.-i. Tsujii, "Evaluating the Impact of Re-training a Lexical Disambiguation Model on Domain Adaptation of an HPSG Parser," in *Trends in Parsing Technology*, pp. 257–275, Springer, 2010.

[123] E. Charniak and M. Johnson, "Coarse-to-fine n-best parsing and MaxEnt discriminative reranking," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pp. 173–180, Association for Computational Linguistics, 2005.

[124] R. McDonald, K. Lerman, and F. Pereira, "Multilingual dependency analysis with a two-stage discriminative parser," in *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pp. 216–220, Association for Computational Linguistics, 2006.

[125] M.-C. De Marneffe, B. MacCartney, C. D. Manning, *et al.*, "Generating typed dependency parses from phrase structure parses," in *Proceedings of LREC*, vol. 6, pp. 449–454, 2006.

[126] K. Sagae and J. Tsujii, “Dependency Parsing and Domain Adaptation with LR Models and Parser Ensembles.,” in *EMNLP-CoNLL*, vol. 2007, pp. 1044–1050, 2007.

[127] J. R. Curran, S. Clark, and J. Bos, “Linguistically motivated large-scale NLP with C&amp;C and Boxer,” in *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pp. 33–36, Association for Computational Linguistics, 2007.

[128] D. McClosky and E. Charniak, “Self-training for biomedical parsing,” in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pp. 101–104, Association for Computational Linguistics, 2008.

[129] J.-H. Chiang and H.-C. Yu, “Literature extraction of protein functions using sentence pattern mining,” *Knowledge and Data Engineering, IEEE Transactions on*, vol. 17, no. 8, pp. 1088–1098, 2005.

[130] R. Yangarber, R. Grishman, P. Tapanainen, and S. Huttunen, “Unsupervised discovery of scenario-level patterns for information extraction,” in *Proceedings of the sixth conference on Applied natural language processing*, pp. 282–289, Association for Computational Linguistics, 2000.

[131] R. Yangarber, “Counter-training in discovery of semantic patterns,” in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pp. 343–350, Association for Computational Linguistics, 2003.

[132] H. Uszkoreit, “Learning relation extraction grammars with minimal human intervention: strategy, results, insights and plans,” in *Computational Linguistics and Intelligent Text Processing*, pp. 106–126, Springer, 2011.

[133] K. B. Cohen, K. Verspoor, H. L. Johnson, C. Roeder, P. V. Ogren, W. A. Baumgartner Jr, E. White, H. Tipney, and L. Hunter, “High-precision biological event extraction: Effects of system and of data,” *Computational Intelligence*, vol. 27, no. 4, pp. 681–701, 2011.

[134] J.-D. Kim, N. Nguyen, Y. Wang, J. Tsujii, T. Takagi, and A. Yonezawa, “The genia event and protein coreference tasks of the BioNLP shared task 2011,” *BMC bioinformatics*, vol. 13, no. Suppl 11, p. S1, 2012.

[135] H. Kilicoglu and S. Bergler, “Adapting a general semantic interpretation approach to biological event extraction,” in *Proceedings of the BioNLP Shared Task 2011 Workshop*, pp. 173–182, Association for Computational Linguistics, 2011.

[136] H. Liu, V. Keselj, C. Blouin, and K. Verspoor, “Subgraph matching-based literature mining for biomedical relations and events,” in *Proceedings of the AAAI 2012 Fall Symposium on Information Retrieval and Knowledge Discovery in Biomedical Text. VA, USA: Association for the Advancement of Artificial Intelligence*, 2012.

[137] J. M. Temkin and M. R. Gilder, “Extraction of protein interaction information from unstructured text using a context-free grammar,” *Bioinformatics*, vol. 19, no. 16, pp. 2046–2053, 2003.

[138] M. Miwa, R. Sætre, Y. Miyao, and J. Tsujii, “Protein–protein interaction extraction by leveraging multiple kernels and parsers,” *International journal of medical informatics*, vol. 78, no. 12, pp. e39–e46, 2009.

[139] N. Kang, E. M. van Mulligen, and J. A. Kors, “Comparing and combining chunkers of biomedical text,” *Journal of biomedical informatics*, vol. 44, no. 2, pp. 354–360, 2011.

[140] M. Miwa, R. Sætre, J.-D. Kim, and J. Tsujii, “Event extraction with complex event classification using rich features,” *Journal of bioinformatics and computational biology*, vol. 8, no. 01, pp. 131–146, 2010.

[141] J. Björne, J. Heimonen, F. Ginter, A. Airola, T. Pahikkala, and T. Salakoski, “Extracting complex biological events with rich graph-based feature sets,” in *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, pp. 10–18, Association for Computational Linguistics, 2009.

[142] H. Poon and L. Vanderwende, “Joint inference for knowledge extraction from biomedical literature,” in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 813–821, Association for Computational Linguistics, 2010.

[143] S. Riedel, D. McClosky, M. Surdeanu, A. McCallum, and C. D. Manning, “Model combination for event extraction in BioNLP 2011,” in *Proceedings of the BioNLP Shared Task 2011 Workshop*, pp. 51–55, Association for Computational Linguistics, 2011.

[144] A. Vlachos and M. Craven, “Search-based structured prediction applied to biomedical event extraction,” in *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pp. 49–57, Association for Computational Linguistics, 2011.

[145] D. McClosky, M. Surdeanu, and C. D. Manning, “Event extraction as dependency parsing for BioNLP 2011,” in *Proceedings of the BioNLP Shared Task 2011 Workshop*, pp. 41–45, Association for Computational Linguistics, 2011.

[146] R. McDonald, K. Crammer, and F. Pereira, “Online large-margin training of dependency parsers,” in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pp. 91–98, Association for Computational Linguistics, 2005.

[147] R. McDonald, F. Pereira, K. Ribarov, and J. Hajič, “Non-projective dependency parsing using spanning tree algorithms,” in *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 523–530, Association for Computational Linguistics, 2005.

[148] M. Miwa, P. Thompson, and S. Ananiadou, “Boosting automatic event extraction from the literature using domain adaptation and coreference resolution,” *Bioinformatics*, vol. 28, no. 13, pp. 1759–1765, 2012.

[149] K. Ravikumar, H. Liu, J. Cohn, M. E. Wall, K. Verspoor, *et al.*, “Literature mining of protein-residue associations with graph rules learned through distant supervision,” *Journal of biomedical semantics*, vol. 3, no. Suppl 3, p. S2, 2012.

[150] J. R. Hobbs, "Coherence and coreference," *Cognitive science*, vol. 3, no. 1, pp. 67–90, 1979.

[151] H. Lee, M. Recasens, A. Chang, M. Surdeanu, and D. Jurafsky, "Joint entity and event coreference resolution across documents," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 489–500, Association for Computational Linguistics, 2012.

[152] T. Ohta, S. Pyysalo, R. Rak, A. Rowley, H.-W. Chun, S.-J. Jung, S.-P. Choi, and S. Ananiadou, "Overview of the Pathway Curation (PC) task of BioNLP shared task 2013," in *Proceedings of BioNLP Shared Task 2013 Workshop, Sofia, Bulgaria, August. Association for Computational Linguistics*, 2013.

[153] Y. Kim, E. Riloff, and N. Gilbert, "The taming of reconcile as a biomedical coreference resolver," in *Proceedings of the BioNLP Shared Task 2011 Workshop*, pp. 89–93, Association for Computational Linguistics, 2011.

[154] V. Stoyanov, C. Cardie, N. Gilbert, E. Riloff, D. Buttler, and D. Hysom, "Reconcile: A coreference resolution research platform," 2010.

[155] K. Raghunathan, H. Lee, S. Rangarajan, N. Chambers, M. Surdeanu, D. Jurafsky, and C. Manning, "A multi-pass sieve for coreference resolution," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 492–501, Association for Computational Linguistics, 2010.

[156] H. Lee, Y. Peirsman, A. Chang, N. Chambers, M. Surdeanu, and D. Jurafsky, "Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task," in *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pp. 28–34, Association for Computational Linguistics, 2011.

[157] K. Yoshikawa, S. Riedel, T. Hirao, M. Asahara, and Y. Matsumoto, "Coreference based event-argument relation extraction on biomedical text.," in *Semantic Mining in Biomedicine*, 2010.

[158] Y. Song, J. Jiang, W. X. Zhao, S. Li, and H. Wang, "Joint learning for coreference resolution with Markov logic," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 1245–1254, Association for Computational Linguistics, 2012.

[159] S. Pradhan, L. Ramshaw, M. Marcus, M. Palmer, R. Weischedel, and N. Xue, "CoNLL-2011 shared task: Modeling unrestricted coreference in ontonotes," in *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pp. 1–27, Association for Computational Linguistics, 2011.

[160] S. Kerrien, B. Aranda, L. Breuza, A. Bridge, F. Broackes-Carter, C. Chen, M. Duesbury, M. Dumousseau, M. Feuermann, U. Hinz, *et al.*, "The IntAct molecular interaction database in 2012," *Nucleic acids research*, vol. 40, no. D1, pp. D841–D846, 2012.

[161] L. Licata, L. Briganti, D. Peluso, L. Perfetto, M. Iannuccelli, E. Galeota, F. Sacco, A. Palma, A. P. Nardozza, E. Santonico, *et al.*, “MINT, the molecular interaction database: 2012 update,” *Nucleic acids research*, vol. 40, no. D1, pp. D857–D861, 2012.

[162] S. Pyysalo, A. Airola, J. Heimonen, J. Björne, F. Ginter, and T. Salakoski, “Comparative analysis of five protein-protein interaction corpora,” *BMC bioinformatics*, vol. 9, no. Suppl 3, p. S6, 2008.

[163] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, *et al.*, “The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored,” *Nucleic acids research*, vol. 39, no. suppl 1, pp. D561–D568, 2011.

[164] M. Kuhn, D. Szklarczyk, A. Franceschini, C. von Mering, L. J. Jensen, and P. Bork, “STITCH 3: zooming in on protein–chemical interactions,” *Nucleic acids research*, vol. 40, no. D1, pp. D876–D880, 2012.

[165] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg, “The database of interacting proteins: 2004 update,” *Nucleic acids research*, vol. 32, no. suppl 1, pp. D449–D451, 2004.

[166] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, “BioGRID: a general repository for interaction datasets,” *Nucleic acids research*, vol. 34, no. suppl 1, pp. D535–D539, 2006.

[167] T. Liu, Y. Lin, X. Wen, R. N. Jorissen, and M. K. Gilson, “BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities,” *Nucleic acids research*, vol. 35, no. suppl 1, pp. D198–D201, 2007.

[168] D. Barrell, E. Dimmer, R. P. Huntley, D. Binns, C. O’Donovan, and R. Apweiler, “The GOA database in 2009—an integrated Gene Ontology Annotation resource,” *Nucleic acids research*, vol. 37, no. suppl 1, pp. D396–D403, 2009.

[169] B. Turner, S. Razick, A. L. Turinsky, J. Vlasblom, E. K. Crowdy, E. Cho, K. Morrison, I. M. Donaldson, and S. J. Wodak, “iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence,” *Database: the journal of biological databases and curation*, vol. 2010, 2010.

[170] S. Orchard, S. Kerrien, S. Abbani, B. Aranda, J. Bhate, S. Bidwell, A. Bridge, L. Briganti, F. S. Brinkman, G. Cesareni, *et al.*, “Protein interaction data curation: the International Molecular Exchange (IMEx) consortium,” *Nature methods*, vol. 9, no. 4, pp. 345–350, 2012.

[171] D. McClosky, *Any domain parsing: automatic domain adaptation for natural language parsing*. PhD thesis, Brown University, 2010.

[172] K. B. Cohen, H. Johnson, K. Verspoor, C. Roeder, and L. Hunter, “The structural and content aspects of abstracts versus bodies of full text journal articles are different,” *BMC bioinformatics*, vol. 11, no. 1, p. 492, 2010.

[173] K. Fundel, R. Küffner, and R. Zimmer, "RelEx—Relation extraction using dependency parse trees," *Bioinformatics*, vol. 23, no. 3, pp. 365–371, 2007.

[174] J.-H. Chiang, H.-H. Liu, and Y.-T. Huang, "Condensing biomedical journal texts through paragraph ranking," *Bioinformatics*, vol. 27, no. 8, pp. 1143–1149, 2011.

[175] R. Hoffmann and A. Valencia, "A gene network for navigating the literature," *Nature genetics*, vol. 36, no. 7, pp. 664–664, 2004.

[176] M. He, Y. Wang, and W. Li, "PPI finder: a mining tool for human protein-protein interactions," *PLoS One*, vol. 4, no. 2, p. e4554, 2009.

[177] Apache Luncene, "http://lucene.apache.org."

[178] Whatizit Query Syntax, "http://www.ebi.ac.uk/rebholz-srv/ebimed/help.jsp#querysyntax."

[179] D. Rebholz-Schuhmann, A. Jimeno-Yepes, M. Arregui, and H. Kirsch, "Measuring prediction capacity of individual verbs for the identification of protein interactions," *Journal of biomedical informatics*, vol. 43, no. 2, pp. 200–207, 2010.

[180] 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.

[181] A. Chatr-aryamontri, S. Kerrien, J. Khadake, S. Orchard, A. Ceol, L. Licata, L. Castagnoli, S. Costa, C. Derow, R. Huntley, *et al.*, "Mint and intact contribute to the second biocreative challenge: serving the text-mining community with high quality molecular interaction data," *Genome Biol*, vol. 9, no. Suppl 2, p. S5, 2008.

[182] Sentences for text-mining from IntAct, "ftp://ftp.ebi.ac.uk/pub/databases/intact/current/various/data-mining."

[183] H. C. Daumé III, *Practical Structured Learning Techniques for Natural Language Processing*. PhD thesis, UNIVERSITY OF SOUTHERN CALIFORNIA, 2006.

[184] BioLemmatizer, "http://biolemmatizer.sourceforge.net/."

[185] The Porter Stemming Algorithm, "http://tartarus.org/ martin/porterstemmer/."

[186] M.-C. De Marneffe, B. MacCartney, C. D. Manning, *et al.*, "Generating typed dependency parses from phrase structure parses," in *Proceedings of LREC*, vol. 6, pp. 449–454, 2006.

[187] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[188] Y.-W. Chen and C.-J. Lin, "Combining svms with various feature selection strategies," in *Feature extraction*, pp. 315–324, Springer, 2006.

[189] Combining GATE and UIMA, "http://gate.ac.uk/sale/tao/splitch21.html."

[190] Y. Miyao, K. Sagae, R. Sætre, T. Matsuzaki, and J. Tsujii, "Evaluating contributions of natural language parsers to protein–protein interaction extraction," *Bioinformatics*, vol. 25, no. 3, pp. 394–400, 2009.

[191] LitWay Configuration Schema, "https://github.com/li-chen/ee/blob/configurable/conf/config.xsd."

[192] LitWay GENIA Configuration File, "https://github.com/li-chen/ee/blob/configurable/conf/config_ge.xml."

[193] C. Li, M. Liakata, and D. Rebholz-Schuhmann, "Biological network extraction from scientific literature: state of the art and challenges," *Briefings in bioinformatics*, p. bbt006, 2013.

[194] J. B. Yang, Q. Mao, Q. L. Xiang, I. W. Tsang, K. M. A. Chai, and H. L. Chieu, "Domain adaptation for coreference resolution: an adaptive ensemble approach," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 744–753, Association for Computational Linguistics, 2012.

[195] E. Tjioe, M. W. Berry, and R. Homayouni, "Discovering gene functional relationships using faun (feature annotation using nonnegative matrix factorization)," *BMC bioinformatics*, vol. 11, no. Suppl 6, p. S14, 2010.

[196] H. Wang, Y. Ding, J. Tang, X. Dong, B. He, J. Qiu, and D. J. Wild, "Finding complex biological relationships in recent PubMed articles using Bio-LDA," *PLoS One*, vol. 6, no. 3, p. e17243, 2011.

[197] T. Bekhuis, "Conceptual biology, hypothesis discovery, and text mining: Swanson's legacy," *Biomedical Digital Libraries*, vol. 3, no. 1, p. 2, 2006.

[198] J. Saez-Rodriguez, L. G. Alexopoulos, J. Epperlein, R. Samaga, D. A. Lauffenburger, S. Klamt, and P. K. Sorger, "Discrete logic modelling as a means to link protein signalling networks with functional analysis of mammalian signal transduction," *Molecular Systems Biology*, vol. 5, no. 1, 2009.

[199] L. Tari, S. Anwar, S. Liang, J. Cai, and C. Baral, "Discovering drug–drug interactions: a text-mining and reasoning approach based on properties of drug metabolism," *Bioinformatics*, vol. 26, no. 18, pp. i547–i553, 2010.

[200] N. L. Novere, A. Finney, M. Hucka, U. S. Bhalla, F. Campagne, J. Collado-Vides, E. J. Crampin, M. Halstead, E. Klipp, P. Mendes, *et al.*, "Minimum information requested in the annotation of biochemical models (miriam)," *Nature biotechnology*, vol. 23, no. 12, pp. 1509–1515, 2005.

[201] V. Chelliah, L. Endler, N. Juty, C. Laibe, C. Li, N. Rodriguez, and N. Le Novère, "Data integration and semantic enrichment of systems biology models and simulations," in *Data Integration in the Life Sciences*, pp. 5–15, Springer, 2009.

[202] BioModels.net Qualifiers, "http://co.mbine.org/standards/qualifiers."

[203] BioModels Database Source Code, "http://sourceforge.net/projects/biomodels/."

[204] M. D. Stobbe, S. M. Houten, G. A. Jansen, A. H. van Kampen, and P. D. Moerland, “Critical assessment of human metabolic pathway databases: a stepping stone for future integration,” *BMC systems biology*, vol. 5, no. 1, p. 165, 2011.

[205] MEDLINE Fact Sheet, “http://www.nlm.nih.gov/pubs/factsheets/medline.html.”

Appendix A

Recovery of the pathways in RECON2 by NER

Pathway	Entity number	By name	By name & anno
Alanine and aspartate metabolism	24	66.67%	83.33%
Alkaloid synthesis	9	33.33%	55.56%
Aminosugar metabolism	37	43.24%	56.76%
Androgen and estrogen synthesis and metabolism	64	45.31%	59.38%
Arachidonic acid metabolism	68	33.82%	52.94%
Arginine and Proline Metabolism	55	41.82%	52.73%
Bile acid synthesis	126	20.63%	44.44%
Biotin metabolism	23	26.09%	34.78%
Blood group synthesis	33	0.00%	54.55%
Butanoate metabolism	6	50.00%	66.67%
C5-branched dibasic acid metabolism	9	33.33%	55.56%
Cholesterol metabolism	79	26.58%	73.42%
Chondroitin sulfate degradation	37	2.70%	48.65%
Chondroitin synthesis	45	2.22%	35.56%

Citric acid cycle	31	45.16%	51.61%
CoA catabolism	12	50.00%	75.00%
CoA synthesis	28	42.86%	67.86%
Cysteine Metabolism	6	83.33%	100.00%
Cytochrome metabolism	21	76.19%	85.71%
D-alanine metabolism	6	100.00%	100.00%
Eicosanoid metabolism	250	9.20%	75.20%
Exchange/demand reaction	741	34.28%	79.35%
Fatty acid oxidation	552	7.25%	39.49%
Fatty acid synthesis	123	8.94%	55.28%
Folate metabolism	66	18.18%	33.33%
Fructose and mannose metabolism	26	42.31%	61.54%
Galactose metabolism	17	52.94%	58.82%
Glutamate metabolism	24	58.33%	79.17%
Glutathione metabolism	21	57.14%	71.43%
Glycerophospholipid metabolism	92	27.17%	39.13%
Glycine, serine, alanine and threonine metabolism	57	49.12%	54.39%
Glycolysis/gluconeogenesis	56	44.64%	64.29%
Glycosphingolipid metabolism	22	4.55%	72.73%
Glyoxylate and dicarboxylate metabolism	24	62.50%	75.00%
Heme degradation	7	42.86%	71.43%
Heme synthesis	16	50.00%	62.50%
Heparan sulfate degradation	29	3.45%	55.17%
Histidine metabolism	22	31.82%	36.36%
Hyaluronan metabolism	7	14.29%	28.57%

Inositol phosphate metabolism	45	15.56%	20.00%
Keratan sulfate degradation	65	4.62%	43.08%
Keratan sulfate synthesis	63	0.00%	88.89%
Limonene and pinene degradation	16	37.50%	43.75%
Linoleate metabolism	24	29.17%	37.50%
Lipoate metabolism	5	80.00%	100.00%
Lysine metabolism	41	34.15%	36.59%
Methionine and cysteine metabolism	52	50.00%	55.77%
Miscellaneous	119	43.70%	79.83%
N-glycan degradation	19	15.79%	94.74%
N-glycan synthesis	104	4.81%	26.92%
NAD metabolism	30	30.00%	53.33%
Nucleotide interconversion	147	36.05%	45.58%
Nucleotide salvage pathway	4	75.00%	100.00%
Nucleotide sugar metabolism	9	22.22%	66.67%
O-glycan synthesis	11	0.00%	72.73%
Oxidative phosphorylation	18	44.44%	50.00%
Pentose phosphate pathway	43	37.21%	39.53%
Phenylalanine metabolism	18	66.67%	83.33%
Phosphatidylinositol phosphate metabolism	54	14.81%	35.19%
Propanoate metabolism	23	56.52%	91.30%
Purine catabolism	54	59.26%	75.93%
Purine synthesis	22	40.91%	50.00%
Pyrimidine catabolism	42	50.00%	66.67%
Pyrimidine synthesis	31	45.16%	48.39%
Pyruvate metabolism	36	33.33%	38.89%

R group synthesis	45	4.44%	86.67%
ROS detoxification	12	100.00%	100.00%
Selenoamino acid metabolism	27	59.26%	70.37%
Sphingolipid metabolism	87	16.09%	93.10%
Squalene and cholesterol synthesis	11	36.36%	63.64%
Starch and sucrose metabolism	36	27.78%	75.00%
Steroid metabolism	69	60.87%	66.67%
Stilbene, coumarine and lignin synthesis	4	75.00%	75.00%
Taurine and hypotaurine metabolism	11	45.45%	63.64%
Tetrahydrobiopterin metabolism	27	40.74%	62.96%
Thiamine metabolism	9	44.44%	88.89%
Transport, endoplasmic reticular	158	30.38%	62.66%
Transport, extracellular	561	45.10%	67.38%
Transport, golgi apparatus	83	15.66%	78.31%
Transport, lysosomal	105	50.48%	61.90%
Transport, mitochondrial	216	50.46%	54.17%
Transport, nuclear	65	32.31%	67.69%
Transport, peroxisomal	110	28.18%	47.27%
Triacylglycerol synthesis	18	11.11%	55.56%
Tryptophan metabolism	68	33.82%	44.12%
Tyrosine metabolism	106	38.68%	67.92%
Ubiquinone synthesis	23	30.43%	73.91%
Unassigned	176	27.27%	38.64%
Urea cycle	80	32.50%	41.25%
Valine, leucine, and isoleucine metabolism	48	41.67%	87.50%
Vitamin A metabolism	53	37.74%	96.23%

Vitamin B12 metabolism	8	75.00%	100.00%
Vitamin B2 metabolism	11	63.64%	100.00%
Vitamin B6 metabolism	10	60.00%	100.00%
Vitamin C metabolism	22	50.00%	95.45%
Vitamin D metabolism	29	34.48%	93.10%
Vitamin E metabolism	31	35.48%	100.00%
Xenobiotics metabolism	34	32.35%	55.88%
beta-Alanine metabolism	22	50.00%	63.64%

Appendix B

Acronyms

BioNLP-ST	BioNLP Shared Tasks
BMDB	BioModels Database
CALBC	Collaborative Annotation of a Large Biomedical Corpus
CPI	Chemical-protein interaction
CRAFT	Colorado Richly Annotated Full Text Corpus
CO2	Co-occurrence
CO3	Tri-occurrence
CRF	Conditional Random Fields
ECO	Evidence Code Ontology
EE	Event extraction
EGF	Epidermal growth factor
EGFR	Epidermal growth factor receptor
EM	Entity mention
EN	Entity normalisation
FN	False negtive
FP	False positive
GO	Gene Ontology
GSC	Gold standard corpus
NER	Named entity recognition
HMR	Hypothetical molecular relation
HMM	Hidden Markov Model
IE	Information extraction
IR	Information retrieval

IUPAC	International Union of Pure and Applied Chemistry
JSON	JavaScript Object Notation
MEMM	Maximum Entropy Markov Model
MI	Molecular interaction
ML	Machine learning
NLP	Natural language processing
NP	Noun phrase
PGN	Protein or gene names
PPI	Protein-protein interaction
SBML	Systems Biology Markup Language
SH	Src homology
SP	SwissProt
SSC	Silver Standard Corpora
SVM	Support Vector Machine
SynP	Syntactic pattern
TM	Text mining
UMLS	Unified Medical Language System
VP	Verb phrase

Appendix C

Publications during this work

The following list details work published or submitted for publication during the course of this PhD in a chronological order.

- Samuel Croset, Christoph Grabmüller, **Chen Li**, Silvestras Kavaliauskas, and Dietrich Rebholz-Schuhmann, The CALBC RDF Triple store: retrieval over large literature content, In *Proceedings of SWAT4LS 2010 (Berlin)*, 17 December 2010.

- Dietrich Rebholz-Schuhmann, Antonio Jimeno Yepes, **Chen Li**, Senay Kafkas, Ian Lewin, Ning Kang, Peter Corbett, David Milward, Ekaterina Buyko, Elena Beisswanger, Kerstin Hornbostel, Alexandre Kouznetsov, René Witte, Jonas B Laurila, Christopher JO Baker, Cheng-Ju Kuo, Simone Clematide, Fabio Rinaldi, Richárd Farkas, György Móra, Kazuo Hara, Laura I Furlong, Michael Rautschka, Mariana Lara Neves, Alberto Pascual-Montano, Qi Wei, Nigel Collier, Md Faisal Mahbub Chowdhury, Alberto Lavelli, Rafael Berlanga, Roser Morante, Vincent Van Asch, Walter Daelemans, José Luís Marina, Erik van Mulligen, Jan Kors and Udo Hahn, Assessment of NER solutions against the first and second CALBC Silver Standard Corpus, *Journal of Biomedical Semantics*, **2**(Suppl 5):S11, 2011.

- **Chen Li**, Antonio Jimeno-Yepes, Miguel Arregui, Harald Kirsch, and Dietrich Rebholz-Schuhmann, PCorral–interactive mining of protein interactions from MEDLINE, *Database (Oxford Journals)*, **Database**:bat030, 2013.

- **Chen Li**, Maria Liakata, and Dietrich Rebholz-Schuhmann, Biological network extraction from scientific literature: state of the art and challenges, *Briefings in bioinformatics*, Feburary 2013.

- Dietrich Rebholz-Schuhmann, Senay Kafkas, Jee-Hyub Kim, **Chen Li**, Antonio Jimeno-Yepes, Robert Hoehndorf, Rolf Backofen, and Ian Lewin, Evaluating gold standard corpora against gene/protein tagging solutions and lexical resources, *Journal of Biomedical Semantics*, **4**[1]:28, 2013.