Extraction of biological interaction networks from scientific literature

Andre Skusa, Alexander Rüegg and Jacob Köhler

Date received (in revised form): 8th June 2005

Andre Skusa
is a PhD student at the NRW International Graduate School in Bioinformatics and Genome Research, Bielefeld University, Germany. His research focus is in reconstruction and analysis of biological interaction networks.

Alexander Rüegg
is a research associate at the Bioinformatics and Medical Informatics Department of the Bielefeld University, Germany. His main research topics are text mining and data integration applied to protein–protein interaction networks.

Jacob Köhler
is a Bioinformatics Principal Investigator at Rothamsted Research, UK. His research interests include data integration, text mining as well as modelling and simulation of biological systems.

Keywords: *network extraction, interaction networks, relation mining*

Abstract

Biology can be regarded as a science of networks: interactions between various biological entities (eg genes, proteins, metabolites) on different levels (eg gene regulation, cell signalling) can be represented as graphs and, thus, analysis of such networks might shed new light on the function of biological systems. Such biological networks can be obtained from different sources. The extraction of networks from text is an important technique that requires the integration of several different computational disciplines. This paper summarises the most important steps in network extraction and reviews common approaches and solutions for the extraction of biological networks from scientific literature.

INTRODUCTION

The extraction of biological networks is an emerging text-mining task, which requires the integration of a wide range of text-mining techniques to support systems biological approaches in modelling, analysis and simulation of biological systems.[1] Furthermore, network extraction is also important for other fields, such as database curation and annotation.[2] Some databases such as TRANSPATH[3] are in fact networks, while others compile interactions between biological entities such as proteins, transcription factors or enzymes and metabolites, eg BIND,[4] DIP[5] and BRENDA.[6] Furthermore, extracted networks can be used to analyse and interpret experimental results, ie to support research and discovery.[7] Another application is to exploit implicit information for generating new knowledge by combining extracted information into a set of hypotheses.[8–12]

The extraction of biological networks requires a combination of several different computational disciplines. Rather than presenting a comprehensive overview about each involved discipline or the whole relation mining field, this paper aims at introducing key aspects and selecting examples that represent the different possible approaches.

Figure 1 introduces the main steps required for reconstructing biological networks from free text and serves also as guideline for the sections on 'Approaches' and 'Tools': first the *texts* to be searched have to be chosen. Then *entities* (eg genes, proteins, metabolites) have to be identified and their (potential) *relations* are to be inferred from the selected texts. Finally, the entities and relations can be combined as nodes and edges into a *network*. The result produced in each step serves as input of the next step. Extracting structured information from unstructured natural language sources cannot yet be expected to produce accurate results that can be used immediately and without further consideration. Therefore, the intermediate *results* of each step also deserve separate *validation* and their performance can be evaluated separately.

The remainder of the paper is organised as follows: the next section ('Approaches') presents (for each step of the workflow in Figure 1) an overview of the different approaches dealing with its respective actions and questions. The section on 'Tools' presents some selected software examples that capture either all network

Andre Skusa,
NRW Graduate School in Bioinformatics and Genome Research,
Bielefeld University,
Postfach 10 01 31,
D-33501 Bielefeld, Germany

Tel: +49 521 106 3955
E-mail:
askusa@cebitec.uni-bielefeld.de

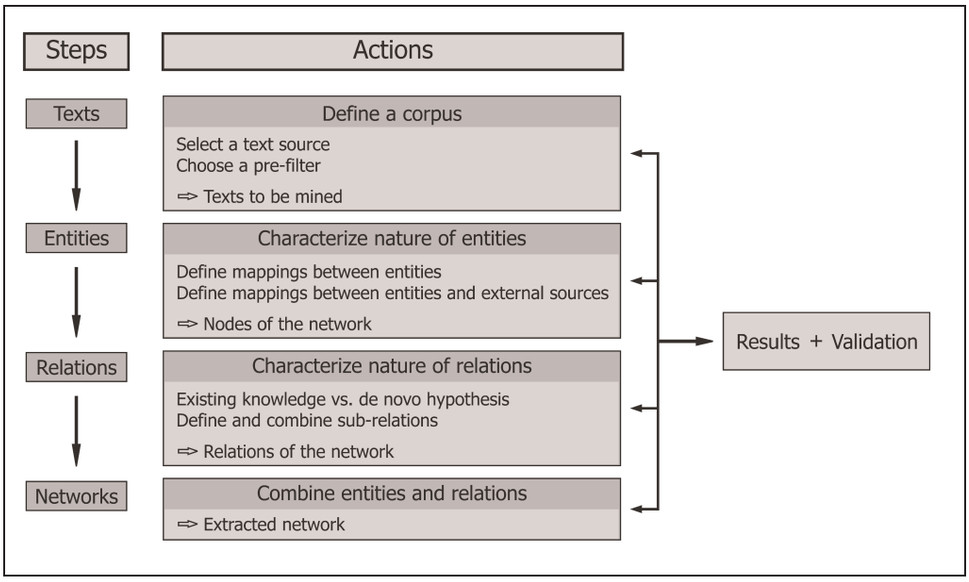

Figure 1: Overview about the main steps and their according actions in network extraction and, following this flow, the paragraphs in section 'Approaches'. Also the software packages described in 'Tools' are chosen for handling tasks appearing in one or more of these steps

extraction steps as an integrated system or one or several individual steps. The paper closes with some final remarks in the conclusion.

APPROACHES

Validation

Evaluation metrics: recall, precision and effectiveness

For each step (Figure 1), the performance depends on the performance of the previous steps. To quantify the performance of text-mining results, three major metrics are normally used: recall, precision and effectiveness.[13] The *recall* is the fraction of correctly identified entities (texts, gene names, protein interactions etc) in the set of *relevant* (ie true positive) entities, whereas the *precision* is the proportion of extracted relevant entities to all entities retrieved. Precision and recall are sometimes also referred to as *specificity* and *sensitivity*. In simple words, the recall shows how *much* of the searched information could be extracted and the precision reflects the *quality* of the method. From this it follows that in order to calculate the recall, usually more information about the searched texts is needed in advance. On the other hand, in order to estimate the precision, one only has to validate a representative subset of the results obtained. For this reason, often the precision is reported without a recall. However, to provide a balanced estimate of the performance of a text-mining approach, both values are combined in the effectiveness measure, which is the reciprocal of the mean of precision and recall.

Texts

The first decision to be made for the extraction of biological networks from scientific literature is the selection of the text sources. One drawback that cannot be avoided is that even if relation mining were 100 per cent successful in retrieving all information from the respective literature, these networks would reflect mostly the current state of the literature, ie they might suffer from both the incompleteness and the biases of the current research efforts in molecular biology and genetics. In effect, networks extracted from scientific literature are not fully connected, and stronger connected subnetworks might stem from research activities concentrating on a couple of interesting genes or substances.[14]

Although in principle any text source can be used for text mining, in practice abstract collections of scientific publications and full text journal publications are normally used. Abstract collections have the advantage of

Selection of appropriate text sources

relatively high information density. Further, they are often already manually annotated and categorised in a structured way that can be exploited for pre-filtering. Whereas MedLine[15] is the largest and most widely used bibliographical resource in the biological domain, other abstract collections and indexing services should also be considered, since MedLine does not necessarily provide the best domain coverage for a specific type of network to be extracted.[16] However, in most text-mining approaches, MedLine is used, which is probably because MedLine is freely available for non-commercial purposes.

Recently, an increasing number of text-mining approaches also utilise full text journal publications,[17–19] and the success of the open access model[20] will remove the financial hurdle for getting hold of a reasonable number of electronic full text publications. Yet dealing with full text publications is also more challenging on a technical level as one has to deal with a range of different formats (pdf, HTML) in which the publications are provided. The more demanding aspect is that the substructure is not always the same. However, since the typical sections of scientific publications (abstract, introduction, methods, results, discussion, figure captions, tables, etc) largely differ in their information density,[21] it is not surprising that those text-mining applications applied on full texts perform best which take the substructure of the paper into account.[22]

Once appropriate text sources have been identified, often the next step is to filter the text sources. In many cases, this is a simple need to reduce the amount of data into a manageable subset: mirroring and indexing all 15 million MedLine abstracts into a local database requires several days on a modern computer.[23] The other reason for filtering is to improve the precision of the subsequent text-mining steps by removing 'obviously' irrelevant text sources. Often, simple methods (keywords, year of publication) are used for filtering. Yet there is the danger that such a simple approach may discard relevant texts. In order to define an organism-specific filter for mice, a naïve filter would only consider abstracts that contain the words 'mouse' or 'mice' or 'mus musculus'. However, such a filter will miss the 18,000 MedLine abstracts with 'murine' as the only word that indicates that they also refer to the same taxonomical entity. In other words, naïve keyword filters may easily miss relevant information and thus already reduce the recall of the whole text-mining process by filtering out relevant texts too early. For such reasons, advanced statistical and machine learning methods can be applied for pre-filtering.[24,25]

In summary, the selection of the text sources and the definition of appropriate filters have a significant influence on subsequent steps: in the worst case, by selecting the wrong text sources or by applying the wrong filters even the best named entity recognition (NER, see 'Entities') and relation mining (see 'Relations') methods are deemed to fail.

Identification of entities in texts

Entities

Before relations can be searched in texts, the entities of the relations have to be identified. Entities represent objects of the real world as, for example, proteins, genes and diseases. Usually these objects do not match simply to one name or symbol in natural language. Thus, many different words or symbols (as synonyms, abbreviations, acronyms or different spellings) have to be considered when a real world entity is searched in texts.

NER is a longstanding NLP (natural language processing) discipline on which a wide range of techniques exists. The different approaches and applications in bioinformatics are very well reviewed by Cohen and Hersh[26] and Krauthammer and Nenadic.[27] In the following, we will outline the basic ideas and principles.

According to Krauthammer and Nenadic,[27] NER consists of three steps: term recognition, term classification and

term mapping, although term classification is not an important step for the purpose of network extraction from scientific literature.

For term recognition, the following approaches can be used:

- *Keywords*: in the simplest case, lists of keywords are used to identify relevant entities.

- *Rules and regular expressions*: for example entities such as fungal gene symbols, *Arabidopsis* gene symbols or enzyme numbers follow a standardised distinct syntax, which can reliably be extracted and identified by regular expressions (ie a string that describes or matches a set of strings, according to certain syntax rules). Yet, unfortunately not all taxonomical entities apply sensible genome nomenclature guidelines.

- *Dictionaries and ontologies*: whereas dictionaries usually are simple term collections, ontologies also store typed relations between the terms, as, for example, 'is a' or 'part of' relations. Terms in ontologies are usually regarded as *concepts*. Entries in dictionaries and concepts of ontologies often contain several synonyms for the same entities. Thus in a dictionary or ontology-based approach the known relations between terms (as, for example, synonym relations) are exploited to identify a searched term in the the text. Apart from manual curation, dictionaries and ontologies can be extracted from free text[28–30] or from scientific databases.[31] Dictionary-based approaches can achieve a balanced precision and recall more than 80 per cent.[32–34] Another advantage of using dictionary-based approaches is that the non-trivial task of term mapping (see below) becomes obsolete, and some dictionary-based approaches can also be used for discriminating between different word senses (mouse as a pointing device versus mouse as an organism).[2,35] Koehler *et al.*[36] present for this purpose an integrated approach where ontologies and databases are mapped in order to perform concept-based term identification and text indexing (see also 'Tools').

- *Machine learning*: one of the most commonly used techniques is machine learning. Here, support vector machines (SVMs)[37,38] as well as hidden Markov models[39,40] are broadly and successfully applied.

Depending on the NER method used, equivalent entities are not always recognised as the same real-world entity, since for most proteins and genes, several synonyms exist. In consequence, relationship mining methods that are developed on top of such NER methods would generate a good deal of redundancy. Such problems can be overcome by selecting an appropriate NER technique, or by subsequent computational or manual linkage of the equivalent entities (term mapping).[41]

At the end of this step, the distinct entities (including in one entity all respective names, synonyms, etc) can be used as the *nodes* of the finally resulting network.

Identification of relations between the recognised entities

Relations

If the entities are defined and localised in the texts, relations between them can be inferred. Usually, the relations to extract are binary. They may or may not be directed or weighted with additional information. Furthermore, it is often required to determine the *type of the relation*,[42] eg whether they link proteins that *interact,* or whether they connect transcription factors that *regulate* genes. Most current efforts in relationship mining deal with protein–protein interactions: yet, also in these cases the different kinds of interactions (activation, binding, etc) need to be characterised.

Relation mining approaches range from applying simple statistical heuristics

(eg by considering co-occurrences of search terms or estimating term frequency distributions) to syntactical and semantical sentence analysis (eg syntactical or semantical parsing) using NLP methods.[43] In *rule-based approaches* a set of additional rules, which for example reflect prior experiences with the considered relation mining task, is added to improve the search.[22] Furthermore, *machine learning* methods can be used, for example, to adapt patterns from text or to discriminate significant words.[19,33]

Extraction of relations by searching for co-occurring terms

One of the most straightforward relation mining approaches is the *co-occurrence search*. The basic assumption here is that for describing a relation between two entities their names usually occur in the same text or part of the text. Thus, for co-occurring entities a relationship can be assumed.

Very basic approaches work with lists of keywords: for example a co-occurrence approach on the sentence level to search for nuclear receptors, their binding proteins and an interaction verb resulted in a precision of 22 per cent when all extracted relations were examined manually.[44]

Another co-occurrence approach is applied in the PubGene database[45] (see also the section on 'Tools') which contains gene–gene relations and was created by searching for pairs of gene names in MedLine abstracts. The extracted relations are weighted by the number of articles in which they were detected. Manual examination of two sets with each 500 randomly selected relations resulted in a precision of 60 per cent for relations found in only one article and 71 per cent for those found in five articles (recall not reported). Further evaluations were conducted by comparing the results with known gene–gene interactions from databases (DIP,[5] OMIM[46]). Between 45 and 51 per cent of the interactions in the database were also found by PubGene.

The performance of co-occurrence searches also depends on the part of the text in which co-occurrences are considered. Ding *et al.*[13] compared recall, precision and effectiveness in single phrases, sentences or the whole abstracts. Interestingly, some *relation types* can best be extracted at the sentence level, whereas others perform better when whole abstracts are considered. Therefore, as a further enhancement, co-occurrence searches can be combined with a set of simple rules that determine the context size and order of the co-occurrence. For example, to extract protein–protein interactions[24] in *Drosophila* the texts were divided into fragments (ie sentences or part of sentences). Then only co-occurrences of protein names and an interaction verb (all taken from predefined lists) possessing the form 'protein A – verb – protein B' are extracted from these fragments.

Scoring the extracted relations and possible relation types can further help to improve the precision. Stephens *et al.*[47] weight each co-occurrence of two gene names in a text by their frequency in the respective text and their inverse frequency in all documents (association score); keywords describing the type of interaction add an additional value. Using this scoring, a search for genes from the same pathway in 5,072 MedLine abstracts resulted in recall and precision rates about 60 and 90 per cent respectively.

Extraction of relations by using natural language parsers

Whereas in co-occurrence approaches only simple rules or patterns are applied to a small set of two or three extracted entities and additional words, NLP[48] techniques parse and analyse the sentences in greater detail. S*hallow parsers* (sometimes referred to as *partial parsers*) are used to identify the syntactic information that is assumed to be the most important. Here, mainly part-of-speech (POS) taggers are used for tagging each word in a sentence with its most likely grammatical function (noun, verb, etc).[48] This can then be used to infer the relations described.[49,50] *Deep parsers* try to reconstruct the complete sentence structure as a tree structure[51,52] and apply a grammar, such as, for example, the combinatory categorial grammar (CCG),[53] which first localises target verbs

to scan afterwards the neighbourhood for the entities of the relations. Generally, full sentence parsers can be distinguished into such reconstructing the syntax or the semantics of a sentence, or a mixture of both. A review by McDonald *et al.*[52] introduces both approaches and mixtures of them and gives an overview on applications in the biomedical text-mining field and the resulting performances, advantages and drawbacks: while syntax-based approaches need no further domain-specific information, they can easily be applied in different domains, but suffer from a lower precision than semantic parsers. For biological relation mining with one exception (Leroy *et al.*[49] reports 90 per cent) no higher precision rates than 83 per cent are reported. The only reported recall was about 47 per cent.[54] Contrarily, semantic grammars apply domain specific resources and thus result in an increased precision (up to 91 and 96 per cent), but are often evaluated in a smaller sample of documents. Consequently, balanced or hybrid approaches have been developed, which try to exploit the benefits of both syntactic and semantic full parsing. The precision of such hybrid systems is high (eg 89[52] or 91 per cent[51]), but the recall is still relatively low (35[52] and 21 per cent[51] respectively).

Inferring hypotheses as new knowledge from text

Comparing NLP approaches with simple co-occurrence assumptions shows that NLP techniques result in some cases in a higher precision, as one could expect from intensive grammar analyses, but at the cost of speed and recall. On the other hand, NLP methods produce knowledge that can be exploited in steps which have to be performed separately when using co-occurrence searches. The POS tagging information can be, for example, used in the NER and the direction or the type of the relation can be easier inferred using the exact structure of the sentence.

Also *machine learning* techniques have been applied to relation mining. An approach that combines dynamic programming and sequencing alignment algorithms as normally used for the comparison of nucleotide sequences is described by Huang *et al.*[18] This approach was applied to 50 full text papers and resulted in a precision/recall of 80.5 and 80.0 per cent. Furthermore, *genetic algorithms* have been used as learning strategy to optimise the set of extracted patterns as well as to train *finite state automatons* for finding new patterns in text.[19] Others use trained classifiers such as SVMs[55] on unprepared[25] or shallow parsed texts[56] to select texts describing an interaction.

Different relation mining strategies were compared in the 'KDD Challenge Cup'.[22] Despite the differences in their approaches, all winning teams have in common that they take the order of words into account rather than considering a text simply as a 'bag of words'. The fact that the winning team applied a purely rule-based approach, and that the other top performing approaches also used a rule-based component in their systems, indicates that machine learning approaches cannot yet compete with rules developed by experts.

Relation mining as described so far can be characterised as reconstructing *established knowledge*, whereas other approaches try to generating *de novo* hypotheses by combining extracted relations. Wren *et al.*[10] and Srinivasan and Libbus[9] both extend and improve the open discovery approach originally proposed by Swanson[8] and Smalheiser in the mid-1980s. The basic assumption is that pairs of terms found in different texts and sharing the same 'intermediate' terms can be linked. An important improvement is to establish a robust and meaningful score for the extracted potential relations. Combining even only a few co-occurrence pairs usually results in a high number of possible implicit links. Wren *et al.*[10] use fuzzy logic methods and compare extracted networks with random networks. Srinivasan and Libbus[9] use combined weights that rank the importance of each identified term (similar to the above-mentioned scoring proposed by Stephens *et al.*[47]). In both

papers hypotheses could be found that have not been reported in a single paper before and that led to new directions for experimental validation. Van Der Eijk *et al.*[12] introduce the associative concept space (ACS) as a metric for weighting the distance between pairs of terms according to the length of the chain of intermediate terms which connect them. Using this method, clusters of functionally related genes could be identified.[57] In Chilibot[11] (see also 'Tools') the whole extracted network is used to generate a network with hypothetical new interactions. However, experimental validation is in most cases still the only way to prove these hypotheses.

As a result of relation mining, *links* of the network to be created can be gained. They might directly consist of a relation between two entities or consist of two or more combined relations.

Networks

Combining extracted entities and relations into a network

Finally, the nodes and links created in the steps 'Entities' and 'Relations' can be integrated into a network. Yet such networks are incomplete and may contain incorrect entities and relations. As already discussed, in each of the different steps a range of methods can be applied that vary significantly in their precision and recall. Therefore, currently only very few approaches are published where networks extracted from texts are used for analysis and further investigations.

Using extracted networks for analysis and validation of experimental results

One possibility of dealing with the uncertainty in the resulting networks is to apply a score that represents the quality of the extracted relations. Such a score can be used as an edge weight to visualise the likelihood of the correctness of relations. New discovered relations then could be drawn in a different way[24] and thus the network visualisation can be used for manual comparison with existing knowledge by experts.[17,45,58,59]

In principle, extracted networks can be used for answering specific biological questions or to provide deeper insights into the general structure of biochemical network topologies. In some cases the resulting network topologies have been investigated. Some topological characteristics of the network can be attributed to the bias of scientific literature (trendy topics and terms resulting in waves of publications on related genes, proteins, etc).[14] Two papers studying the network topology[11,60] report that the distribution of the node degrees is scale-free, ie it follows a power law, meaning that most of the nodes obtain a small connectivity whereas a few nodes (so-called *hubs*) are highly connected (for a good introduction into network theory see Newman[61] and Albert and Barabasi[62]). Again, Chen and Sharp[11] interpret this as the reflection of the intensity of a subject investigated, ie most topics are only scarcely considered and a few are intensively studied. Interestingly, Blaschke and Valencia[60] discovered a correlation between the distance of nodes in the network and their functional relationship. Especially for classifying the correctness of new relations, this could be a helpful measure. In summary, not much work on graph-based analysis of biological networks extracted from text sources exists. So far, topological properties of hypothetical networks have been mainly used for validating and analysing the correctness of the extracted networks.

Rather than analysing the topological properties, the extracted networks can also be used in context with experimental data in order to validate the extracted network as well as to evaluate the experiments. For example Jenssen *et al.*[45] could show that their extracted co-occurrence gene networks reflect biologically meaningful relationships from three large-scale experiments. The resulting PubGene database and tool also allows gene expression data to be analysed in context of extracted networks (see also 'Tools'). Karopka *et al.*[63] apply their extraction approach on lists of gene names from experiments to compare the extracted relations with the experimentally determined ones. Albert *et al.*[44] searched for protein interactions of nuclear receptors and compared these

text-mining results with data from yeast two-hybrid screens. Here they found similarities of the nuclear receptors regarding their connectivities. Also properties of some specific proteins were investigated and could be experimentally validated. Another example for the use of extracted networks is the curation of specific pathways, eg the Wnt pathway.[64]

It is worth mentioning that also most experimental data are far from being complete and unambiguous. Hoffmann and Valencia[65] compared protein networks of the same organism created with different experimental methods and found many differences in the topologies of the networks. Thus, biological interaction networks extracted from texts can be used as additional source for validation. For this purpose tools as Chilibot[11] or iHOP[66,67] can also be used to navigate through the papers constituting an extracted network (see also 'Tools').

Additionally the creation and visual inspection of hypothetical relations can be used to explore new features of the considered entities.[11] For example Wren *et al.*[10] report the discovery of new relationships between cardiac hypertrophy and potential drug targets.

TOOLS

This section introduces selected tools that implement one or more of the approaches discussed for each step of the workflow (Figure 1) in the previous section. Table 1 gives an overview of recently developed and available software.

Examples for integrated applications

Table 1: Selected tools. The columns for methods indicate to which part(s) of the workflow the tool can be used for

No.	Name	Main website	Methods				Availability	Platforms
			Texts	Entities	Relations	Networks		
1	PASTA	http://www.dcs.shef.ac.uk/research/groups/nlp/pasta/	×	×	×	×	Public	Web
2	PathwayAssist	http://www.ariadnegenomics.com/products/pathway.html	×	×	×	×	Commercial	Win
3	Chilibot	http://www.chilibot.net	×	×	×	×	Public	Web
4	E-Utilities	http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html	×				Public	Web, Java
5	TnT	http://www.coli.uni-saarland.de/~thorsten/tnt/		×			Public	Unix
6	CASS	http://www.vinartus.com/spa/		×			Open source	Unix
7	AiSee	http://www.aisee.com/				×	Commercial	Win, Unix
8	PubGene	http://www.pubgene.org/	×	×	×	×	Commercial	Web, Win
9	GraphViz	http://www.graphviz.org				×	Open source	Lin, Win
10	BioNLP	http://www.geneticxchange.com/		×	×		Commercial	Java
11	GATE	http://www.gate.ac.uk/		×	×		Open source	Java
12	ONDEX	http://sourceforge.net/projects/ondex	×	×	×	×	Open source	Lin
13	MedlineR	http://dbsr.duke.edu/pub/MedlineR/	×	×	×		Open source	R
14	Pajek	http://vlado.fmf.uni-lj.si/pub/networks/pajek				×	Public	Win
15	PubMatrix	http://pubmatrix.grc.nia.nih.gov/	×	×	×		Public	Web
16	iHop	http://www.pdg.cnb.uam.es/UniPub/iHOP/	×	×	×	×	Public	Web
17	MedKit	http://metnetdb.gdcb.iastate.edu/medkit/	×				Open source	Java
18	Textomy	http://www.litminer.ca/		×	×		n.a.	n.a.
19	Snowball	http://snowball.tartarus.org/		×			Open source	Lin, Java
20	Qtag	http://www.english.bham.ac.uk/staff/omason/software/qtag.html		×			Different licence	Java
21	NLProt	http://cubic.bioc.columbia.edu/services/nlprot/index.html		×			Public	Lin, Win
22	Ingenuity	http://www.ingenuity.com				×	Commercial	Web
23	Cytoscape	http://cytoscape.org				×	Open source	Java
24	Osprey	http://biodata.mshri.on.ca/osprey/				×	Different licence	Lin, Win

Integrated tools: applying all steps of network extraction

that combine all steps into one system are PIES,[68] SUISEKI,[69] PreBIND,[56] GeneWays[59] and PASTA[70] (the last one is tool no. 1 in Table 1). The commercial software package PathwayAssist (2) also addresses the whole workflow. It uses MedScan[51,71] as module for text mining, which is also available separately and based on NLP techniques. After retrieving PubMed abstracts according to a user-defined query, sentences are filtered out that do not contain at least one concept of a dictionary. The remaining sentences are further processed with a syntactic parser and a semantic interpreter. The resulting relationships can then be visualised and analysed within PathwayAssist. The reported precision is 91 per cent with a recall of 21 per cent.

Chilibot[11] (3) is a web service to construct networks from genes, proteins, drugs and other biological concepts. It uses the E-Utilities (4) service (ESearch and EFetch) at NCBI for retrieval of documents by submitting a query consisting of the pairwise combinations of the user's input terms and their synonyms. Acronyms contained in the user input are automatically resolved to their long-term phrases. Retrieved abstracts containing less than 30 per cent of the acronym's phrase terms are rejected. Sentences from the abstracts that contain two or more query terms and synonyms are further processed by the POS tagger TnT[72] (5) and the shallow parser CASS (6). Following that, the resulting sentences are classified into one of six categories according to the presence/absence of terms indicating special relationships. For visualisation of the extracted relationships AiSee (7) is used in Chilibot. The extracted network can in addition be used for navigating the related literature. The precision of the system was determined to be between 74 and 79 per cent depending on the category and the recall to be about 90 per cent.

PubGene[45] (8) is an integrated system widely used in different projects. It is a commercial tool, but developed in academic research. The basic version described in Jenssen *et al.* [45] uses a dictionary of gene symbols and names collected from HUGO nomenclature database, LocusLink, GDB and GENATLAS to identify genes in Medline. Each gene thereby is represented by its primary gene symbol. With the resulting gene–article–index co-occurrences of pairs of genes in the abstracts are calculated (see also 'Relations' in the previous section). The retrieved network can be enriched with DNA microarray data. The visualisation is done with GraphViz (9).

The systems described so far integrate all parts of the overall workflow. Building blocks of these applications are tools that cover either one task, eg TnT, or many parts, eg BioNLP[73] (10). A public available framework that provides the basic architecture for the development of information extraction applications is GATE[74] (11). In the field of biological relation mining it is used, for example, in PASTA[70] and by Karopka *et al.*[63] GATE includes a set of components, which can be replaced or extended easily as the framework is provided as a Java API. Beside usual modules like a Tokenizer, a Sentence Splitter or a Tagger, components for recognising relations and finding identical entities (Orthomatcher, Coreferencer) are available.

The ONDEX (12) suite is intended for integration of databases, network extraction and graph analysis. Here, a concept-based entity recognition using mapped ontologies is applied in a first step (see also 'Entities') and used for text mining with a co-occurrence search. It is not restricted to PubMed abstracts as texts are imported into a relational database format (PostgreSQL).

The library MedlineR[75] (13) uses the statistical environment and programming language R to define procedures for retrieving articles from NCBI, mapping terms to MeSH and mainly to calculate co-occurrences of terms. The visualisation of the associations is realised through the generation of an output file in the Pajek (14) format.

PubMatrix[76] (15) is, in contrast to MedlineR, a web-based tool intended for interactive querying. To calculate a co-occurrence matrix the user has to define two lists of terms, a search list and a modifier list. The terms of the list, which can be simple keyword lists or gene symbols, are used to create PubMed queries. This is realised by pairwise combining the terms of the different lists. Finally, the resulting matrix contains the frequency of co-occurrences. Another interactive querying tool is the iHOP service (16). It enables the search of genes in a pre-calculated co-occurrence network of genes and proteins (from eight organisms). In contrast to other systems the user retrieves fragments of sentences, which contain relations of the searched gene, and then selects relevant relations that should be added to a user specific literature network.

Specialised tools: applying individual steps of network extraction

Finally, there exist a number of software packages that can be used in each single step of the network extraction workflow (Figure 1): the acquisition of texts can simply be done by using the E-Utilities of NCBI. MedKit[77] (17) is also very useful for this purpose and more powerful. On the other hand more sophisticated methods can be applied to get more appropriate text corpora. Textomy[56] (18), for example, is part of the PreBIND[56] system and uses SVMs for classifying texts.

For identifying entities in text in most systems standard NLP techniques can be applied. In the biomedical domain public available tools have already been used, eg Snowball (19) for stemming or Qtag (20) for part-of-speech tagging. Specialised taggers for biological knowledge also exist under different licensing conditions.

A publicly available system which addresses this task is NLProt[78] (21). It uses different dictionaries, eg a protein names dictionary extracted from Uniprot and a common names dictionary derived from Merriam-Webster, in combination with SVMs. For training the SVMs in the first step each abstract is split into single tokens separated by spaces. From these tokens sample phrases are constructed that are composed of three parts. SVMs then are trained for each of these parts separately. This enables the system to be trained for different purposes, eg one SVM was trained on central words and one for the environment. The system achieves a precision of 75 per cent and a recall of 76 per cent even for novel protein names.

Analysis and visualisation of the generated networks can be supported using specialised biological pathway and network analysis tools, as eg Ingenuity (22), Cytoscape (23), Osprey (24) or ONDEX (12). These tools enable users to analyse experimental data such as gene expression results in context of the biological networks. Ingenuity makes use of a knowledge base, but it could not be determined from the available information in the web whether this database or only part of it has been built using text mining.

More generic applications as, for instance, Pajek (14) are also very useful especially in analysing topological properties of the biological networks. For importing networks as text files the accepted formats of theses tools range from simple tab delimited files to common standards, as, for example, GML or PSI.

CONCLUSIONS

Which of the presented extraction methods performs best obviously depends highly on the specific types of networks to be extracted, and on the typical structure of a publication that contains a relation. For example, protein–protein interactions are often dealt with at the sentence level and achieve a good precision (up to 95 per cent, but low recall in those few cases where the recall is also reported).[18,51,56] The type of networks to be extracted might also determine whether it is sufficient for the actual relation mining to use simple heuristics (as, for example, approaches based on co-occurrences of search terms in the same context) or whether there is a potential benefit in using advanced

methods (such as eg syntactic or semantic parsing of sentences).

Although several systems exist that can be used for certain types of networks (mainly gene–gene and protein–protein interactions), a coherent 'all-in-one' solution for extracting biological networks from text does not exist, nor is it appropriate to address the different types of problems in the same way. Contrariwise, the currently existing approaches and tools provide a set of solid building blocks that can be used to develop customised applications.

Acknowledgments

All authors thank Dion Whitehead for his help on the manuscript. AS and AR thank the Ministry of Research and Education of North Rhine Westfalia (Germany) and the German Federal Ministry of Education and Research respectively for financial support. AS and JK also thank the European Science Foundation (ESF) for the support of AS with a short visit grant. JK gratefully acknowledges support from the Biotechnology and Biological Sciences Research Council of the United Kingdom including that from Grant BBS/B/13640.

References

1. Christopher, R., Dhiman, A., Fox, J. *et al.* (2004), 'Data-driven computer simulation of human cancer cell', *Ann. NY Acad. Sci.*, Vol. 1020, pp. 132–153.
2. Hofmann, O. and Schomburg, D. (2005), 'Concept-based annotation of enzyme classes', *Bioinformatics*, Vol. 21, pp. 2059–2066.
3. Schacherer, F., Choi, C., Gotze, U. *et al.* (2001), 'The TRANSPATH signal transduction database: A knowledge base on signal transduction networks', *Bioinformatics*, Vol. 17, pp. 1053–1057.
4. Bader, G. D., Donaldson, I., Wolting, C. *et al.* (2001), 'BIND – The Biomolecular Interaction Network Database', *Nucleic Acids Res.*, Vol. 29, pp. 242–245.
5. Xenarios, I., Salwinski, L., Duan, X. J. *et al.* (2002), 'DIP, the Database of Interacting Proteins: A research tool for studying cellular networks of protein interactions', *Nucleic Acids Res.*, Vol. 30, pp. 303–305.
6. Schomburg, I., Chang, A., Hofmann, O. *et al.* (2002), 'BRENDA: A resource for enzyme data and metabolic information', *Trends Biochem. Sci.*, Vol. 27, pp. 54–56.
7. Werner, T. (2005), 'The next generation of literature analysis: Integration of genomic analyses into text mining', *Brief. Bioinformatics*, Vol. 6.
8. Swanson, D. R. (1986), 'Fish oil, Raynaud's syndrome, and undiscovered public knowledge', *Perspect. Biol. Med.*, Vol. 30, pp. 7–18.
9. Srinivasan, P. and Libbus, B. (2004), 'Mining MEDLINE for implicit links between dietary substances and diseases', *Bioinformatics*, Vol. 20 (Suppl 1), pp. I290–296.
10. Wren, J. D., Bekeredjian, R., Stewart, J. A. *et al.* (2004), 'Knowledge discovery by automated identification and ranking of implicit relationships', *Bioinformatics*, Vol. 20, pp. 389–398.
11. Chen, H. and Sharp, B. M. (2004), 'Content-rich biological network constructed by mining PubMed abstracts', *BMC Bioinformatics*, Vol. 5, p. 147.
12. Van Der Eijk, C., Van Mulligen, E. M., Kors, J. A. *et al.* (2004), 'Constructing an associative concept space for literature-based discovery', *J. Amer. Soc. Inf. Sci.*, Vol. 55, pp. 436–444.
13. Ding, J., Berleant, D., Nettleton, D. and Wurtele, E. (2002), 'Mining MEDLINE: Abstracts, sentences, or phrases?', in 'Proceedings of the 7th Pacific Symposium on Biocomputing', 3rd–7th January, Hawaii, pp. 326–337.
14. Krauthammer, M., Kra, P., Iossifov, I. *et al.* (2002), 'Of truth and pathways: Chasing bits of information through myriads of articles', *Bioinformatics*, Vol. 18 (Suppl 1), pp. S249–257.
15. Bachrach, C. A. and Charen, T. (1978), 'Selection of MEDLINE contents, the development of its thesaurus, and the indexing process', *Med. Inform. (London)*, Vol. 3, pp. 237–254.
16. Stone, V. L., Fishman, D. L. and Frese, D. B. (1998), 'Searching online and Web-based resources for information on natural products used as drugs', *Bull. Med. Libr. Assoc.*, Vol. 86, pp. 523–527.
17. Friedman, C., Kra, P., Yu, H. *et al.* (2001), 'GENIES: A natural-language processing system for the extraction of molecular pathways from journal articles', *Bioinformatics*, Vol. 17 (Suppl 1), pp. S74–82.
18. Huang, M., Zhu, X., Hao, Y. *et al.* (2004), 'Discovering patterns to extract protein–protein interactions from full texts', *Bioinformatics*, Vol. 20, pp. 3604–3612.
19. Plake , C., Hakenberg, J. and Leser, U. (2005) 'Learning patterns for information extraction from free text', in 'Proceedings of AKKD 2005', Karlsruhe, Germany.
20. Suber, P. (2002), 'Open access to the scientific journal literature', *J. Biol.*, Vol. 1, p. 3.
21. Schuemie, M. J., Weeber, M., Schijvenaars,

B. J. *et al.* (2004), 'Distribution of information in biomedical abstracts and full-text publications', *Bioinformatics*, Vol. 20, pp. 2597–2604.

22. Yeh, A. S., Hirschman, L. and Morgan, A. A. (2003), 'Evaluation of text data mining for database curation: Lessons learned from the KDD Challenge Cup', *Bioinformatics*, Vol. 19 (Suppl 1), pp. i331–339.
23. Oliver, D. E., Bhalotia, G., Schwartz, A. S. *et al.* (2004), 'Tools for loading MEDLINE into a local relational database', *BMC Bioinformatics*, Vol. 5, p. 146.
24. Blaschke, C., Andrade, M. A., Ouzounis, C. and Valencia, A. (1999), 'Automatic extraction of biological information from scientific text: protein–protein interactions', in 'Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology', AAAI Press, Menlo Park, CA, pp. 60–67.
25. Marcotte, E. M., Xenarios, I. and Eisenberg, D. (2001), 'Mining literature for protein–protein interactions', *Bioinformatics*, Vol. 17, pp. 359–363.
26. Cohen, A. M. and Hersh, W. R. (2005) 'A survey of current work in biomedical text mining', *Brief. Bioinformatics*, Vol. 6, pp. 57–71.
27. Krauthammer, M. and Nenadic, G. (2004), 'Term identification in the biomedical literature', *J. Biomed. Inform.*, Vol. 37, pp. 512–526.
28. Yu, H. and Agichtein, E. (2003), 'Extracting synonymous gene and protein terms from biological literature', *Bioinformatics*, Vol. 19 (Suppl 1), pp. i340–349.
29. Rindflesch, T. C., Hunter, L. and Aronson, A. R. (1999), 'Mining molecular binding terminology', in 'Proceedings of the 1999 AMIA Annual Symposium', Bethesda, MD, pp. 127–131.
30. Hatzivassiloglou, V. and Weng, W. (2002), 'Learning anchor verbs for biological interaction patterns from published text articles', *Int. J. Med. Inform.*, Vol. 67, pp. 19–32.
31. Chen, L., Liu, H. and Friedman, C. (2005), 'Gene name ambiguity of eukaryotic nomenclatures', *Bioinformatics*, Vol. 21, pp. 248–256.
32. Hanisch, D., Fluck, J., Mevissen, H. T. and Zimmer, R. (2003), 'Playing biology's name game: Identifying protein names in scientific text', in 'Proceedings of the 8th Pacific Symposium on Biocomputing', 3rd–7th January, Hawaii, pp. 403–414.
33. Ono, T., Hishigaki, H., Tanigami, A. and Takagi, T. (2001), 'Automated extraction of information on protein-protein interactions from the biological literature', *Bioinformatics*, Vol. 17, pp. 155–161.
34. Nenadic, G., Spasic, I. and Ananiadou, S. (2003), 'Terminology-driven mining of biomedical literature', *Bioinformatics*, Vol. 19, pp. 938–943.
35. Ruch, P., Baud, R. and Geissbuhler, A. (2003), 'Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record', *Artif. Intell. Med.*, Vol. 29, pp. 169–184.
36. Koehler, J., Rawlings, C., Verrier, P. *et al.* (2004), 'Linking experimental results, biological networks and sequence analysis methods using ontologies and generalised data structures', *In Silico Biol.*, Vol. 5, p. 5.
37. Shi, L. and Campagne, F. (2005), 'Building a protein name dictionary from full text: A machine learning term extraction approach', *BMC Bioinformatics*, Vol. 6, p. 88.
38. Kazama, J., Makino, T., Ohta, Y. and Tsujii, J. (2002), 'Tuning support vector machines for biomedical named entity recognition', in 'Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain', ACL, Philadelphia, pp. 1–8.
39. Collier, N., Nobata, C. and Tsujii, J. (2000), 'Extracting the names of genes and gene products with a hidden Markov model', in 'Proceedings of COLING 2000', Saarbruecken, Germany, pp. 201–207.
40. Shen, D., Zhang, J., Zhou, G. *et al.* (2003), 'Effective adaptation of hidden Markov model-based named entity recognizer for biomedical domain', in 'Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain', ACL, Saporo, Japan, pp. 49–56.
41. Wren, J. D., Chang, J. T., Pustejovsky, J. *et al.* (2005), 'Biomedical term mapping databases', *Nucleic Acids Res.*, Vol. 33, pp. D289–293.
42. Smith, B., Ceusters, W., Klagges, B. *et al.* (2005), 'Relations in biomedical ontologies', *Genome Biol.*, accepted.
43. Shatkay, H. and Feldman, R. (2003), 'Mining the biomedical literature in the genomic era: An overview', *J. Comput. Biol.*, Vol. 10, pp. 821–855.
44. Albert, S., Gaudan, S., Knigge, H. *et al.* (2003), 'Computer-assisted generation of a protein-interaction database for nuclear receptors', *Mol. Endocrinol.*, Vol. 17, pp. 1555–1567.
45. Jenssen, T. K., Laegreid, A., Komorowski, J. and Hovig, E. (2001), 'A literature network of human genes for high-throughput analysis of gene expression', *Nat. Genet.*, Vol. 28, pp. 21–28.
46. Hamosh, A., Scott, A. F., Amberger, J. S. *et al.* (2005), 'Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes

and genetic disorders', *Nucleic Acids Res.*, Vol. 33, pp. D514–517.

47. Stephens, M., Palakal, M., Mukhopadhyay, S. *et al.* (2001), 'Detecting gene relations from Medline abstracts', in 'Proceedings of the 6th Pacific Symposium on Biocomputing', 3rd–7th January, Hawaii, pp. 483–495.
48. Manning, C. and Schütze, H. (1999), 'Foundations of Statistical Natural Language Processing', MIT Press, Cambridge, MA.
49. Leroy, G., Chen, H. and Martinez, J. D. (2003), 'A shallow parser based on closed-class words to capture relations in biomedical text', *J. Biomed. Inform.*, Vol. 36, pp. 145–158.
50. Sekimizu, T., Park, H. S. and Tsujii, J. (1998), 'Identifying the interaction between genes and gene products based on frequently seen verbs in Medline abstracts', *Genome Inform. Ser. Workshop Genome Inform.*, Vol. 9, pp. 62–71.
51. Daraselia, N., Yuryev, A., Egorov, S. *et al.* (2004), 'Extracting human protein interactions from MEDLINE using a full-sentence parser', *Bioinformatics*, Vol. 20, pp. 604–611.
52. McDonald, D. M., Chen, H., Su, H. and Marshall, B. B. (2004), 'Extracting gene pathway relations using a hybrid grammar: The Arizona relation parser', *Bioinformatics*, Vol. 20, pp. 3370–3378.
53. Park, J. C., Kim, H. S. and Kim, J. J. (2001), 'Bidirectional incremental parsing for automatic pathway identification with combinatory categorial grammar', in 'Proceedings of the 6th Pacific Symposium on Biocomputing', 3rd–7th January, Hawaii, pp. 396–407.
54. Yakushiji, A., Tateisi, Y., Miyao, Y. and Tsujii, J. (2001), 'Event extraction from biomedical papers using a full parser', in 'Proceedings of the 6th Pacific Symposium on Biocomputing', 3rd–7th January, Hawaii, pp. 408–419.
55. Zelenko, D., Aone, C. and Richardella, A. (2003), 'Kernel methods for relation extraction', *J. Machine Learning Res.*, Vol. 3, pp. 1083–1106.
56. Donaldson, I., Martin, J., de Bruijn, B. *et al.* (2003), 'PreBIND and Textomy – mining the biomedical literature for protein–protein interactions using a support vector machine', *BMC Bioinformatics*, Vol. 4, p. 11.
57. Jelier, R., Jenster, G., Dorssers, L. C. *et al.* (2005), 'Co-occurrence based meta-analysis of scientific texts: Retrieving biological relationships between genes', *Bioinformatics*, Vol. 21, pp. 2049–2058.
58. Yao, D., Li, M., Lu, Y. *et al.* (2004), 'PathwayFinder: Paving the way towards automatic pathway extraction', in 'Proceedings of the 2nd Asia-Pacific Bioinformatics Conference (APBC2004)', Chen, Y.-P. P., Ed., Vol. 29, Dunedin, New Zealand, pp. 53–62.
59. Rzhetsky, A., Iossifov, I., Koike, T. *et al.* (2004), 'GeneWays: A system for extracting, analyzing, visualizing, and integrating molecular pathway data', *J. Biomed. Inform.*, Vol. 37, pp. 43–53.
60. Blaschke, C. and Valencia, A. (2001), 'The potential use of SUISEKI as a protein interaction discovery tool', in 'Proceedings of the 12th Genome Informatics Workshop', Universal Academy Press, Tokyo, Japan, pp. 123–134.
61. Newman, M. E. (2003), 'The structure and function of complex networks', *SIAM Rev.*, Vol. 45, pp. 167–256.
62. Albert, R. and Barabasi, A.-L. (2002), 'Statistical mechanics of complex networks', *Rev. Modern Phys.*, Vol. 74, pp. 47–97.
63. Karopka, T., Scheel, T., Bansemer, S. and Glass, A. (2004), 'Automatic construction of gene relation networks using text mining and gene expression data', *Med. Inform. Internet Med.*, Vol. 29, pp. 169–183.
64. Santos, C., Eggle, D. and States, D. J. (2005), 'Wnt pathway curation using automated natural language processing: Combining statistical methods with partial and full parse for knowledge extraction', *Bioinformatics*, Vol. 21, pp. 1653–1658.
65. Hoffmann, R. and Valencia, A. (2003), 'Protein interaction: Same network, different hubs', *Trends Genet.*, Vol. 19, pp. 681–683.
66. Hoffmann, R., Krallinger, M., Andres, E. *et al.* (2005), 'Text mining for metabolic pathways, signaling cascades, and protein networks', *Sci. STKE*, Vol. 2005, p. e21.
67. Hoffmann, R. and Valencia, A. (2004), 'A gene network for navigating the literature', *Nat. Genet.*, Vol. 36, p. 664.
68. Wong, L. (2001), 'PIES, a protein interaction extraction system', in 'Proceedings of the 6th Pacific Symposium on Biocomputing', 3rd–7th January, Hawaii, pp. 520–531.
69. Blaschke, C., Hirschman, L. and Valencia, A. (2002), 'Information extraction in molecular biology', *Brief. Bioinformatics*, Vol. 3, pp. 154–165.
70. Gaizauskas, R., Demetriou, G., Artymiuk, P. J. and Willett, P. (2003), 'Protein structures and information extraction from biological texts: The PASTA system', *Bioinformatics*, Vol. 19, pp. 135–143.
71. Novichkova, S., Egorov, S. and Daraselia, N. (2003), 'MedScan, a natural language processing engine for MEDLINE abstracts', *Bioinformatics*, Vol. 19, pp. 1699–1706
72. Brants, T. (2000), 'TnT – a statistical part-of-speech tagger', in 'Proceedings of the 6th

Applied Natural Language Processing Conference', Seattle, WA.

73. Ng, S.-K. and Wong, M. (1999), 'Toward routine automatic pathway discovery from on-line scientific text abstracts', in 'Proceedings of the 10th Genome Informatics Workshop', Universal Academy Press, Tokyo, Japan, pp. 123–134.
74. Cunningham, H., Maynard, D., Bontcheva, K. and Tablan, V. (2002), 'GATE: An architecture for development of robust HLT applications', in 'Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics', Philadelphia.
75. Lin, S. M., McConnell, P., Johnson, K. F. and Shoemaker, J. (2004), 'MedlineR: An open source library in R for Medline literature data mining', *Bioinformatics*, Vol. 20, pp. 3659–3661.
76. Becker, K. G., Hosack, D. A., Dennis, G. *et al.* (2003), 'PubMatrix: A tool for multiplex literature mining', *BMC Bioinformatics*, Vol. 4, p. 61.
77. Ding, J. and Berleant, D. (2005), 'MedKit: A helper toolkit for automatic mining of MEDLINE/PubMed citations', *Bioinformatics*, Vol. 21, pp. 694–695.
78. Mika, S. and Rost, B. (2004), 'NLProt: Extracting protein names and sequences from papers', *Nucleic Acids Res.*, Vol. 32, pp. W634–637.