BIOINFORMATICS
Vol. 17 Suppl. 1 2001
Pages S74–S82

GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles

Carol Friedman[1,2], Pauline Kra[2], Hong Yu[2], Michael Krauthammer[2] and Andrey Rzhetsky[2,3]

[1]Computer Science Dept, Queens College CUNY, Flushing, NY, 11367, USA, [2]Department of Medical Informatics, Columbia University, New York, 10032, USA and [3]Genome Center, Columbia University, New York, 10032, USA

Received on January 31, 2001; revised and accepted on March 30, 2001

ABSTRACT

Systems that extract structured information from natural language passages have been highly successful in specialized domains. The time is opportune for developing analogous applications for molecular biology and genomics. We present a system, GENIES, that extracts and structures information about cellular pathways from the biological literature in accordance with a knowledge model that we developed earlier.

We implemented GENIES by modifying an existing medical natural language processing system, MedLEE, and performed a preliminary evaluation study. Our results demonstrate the value of the underlying techniques for the purpose of acquiring valuable knowledge from biological journals.

Contact: friedman.carol@dmi.columbia.edu

INTRODUCTION

Recently the fields of molecular biology and medicine have enjoyed an explosive development; as a result, individual researchers find it difficult to keep up with all the new, relevant information. Several knowledge-based systems have been developed that capture and organize information, such as that in the domain of molecular interactions (Chen *et al*., 1997; Karp *et al*., 1999; Selkov *et al*., 1997). Most of these systems represent a single well-defined area, such as metabolic pathways for one bacterial species; even so, populating and maintaining the knowledge bases (e.g., Ashburner *et al.*, 2000; Baker *et al.*, 1999; Kanehisa & Goto, 2000)) requires enormous work. For example, the articles listed in PubMed as related to the *cell cycle* number *hundreds of thousands*; thus, manual identification and entry of information into a knowledge base is not practical.

We are developing a system called GeneWays to perform massive automated extraction of information from research literature and automated maintenance of a knowledge base that contains comprehensive information about signal-transduction pathways, about the diseases associated with the pathways, and about the drugs that affect them. GeneWays contains many modules (see Figure 1 for an overview). One module performs natural language processing (NLP), but the details of the NLP module are not shown in Figure 1. However that module is itself composed of a number of components: a tagger

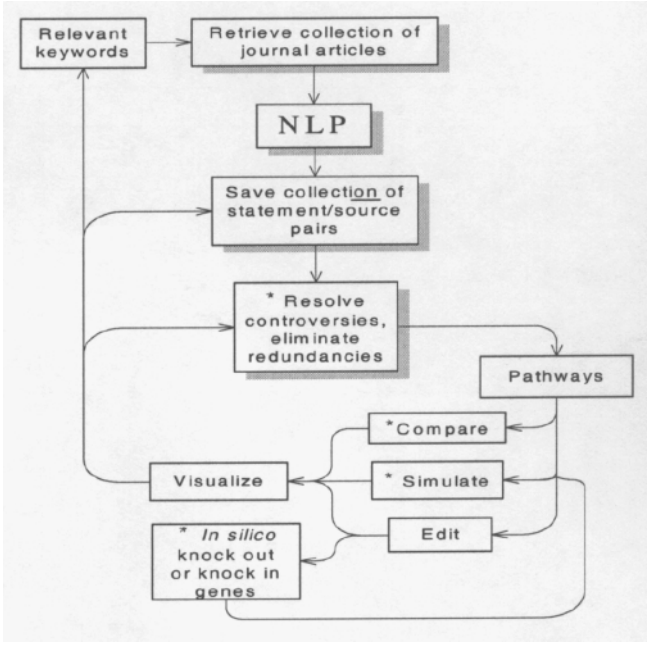

Fig. 1. An outline of GeneWays, the system that embeds an NLP component consisting of GENIES and other NLP components (Krauthammer *et al*., 2000; Hatzivassiloglou *et al*., 2001). Components that are starred are not yet implemented.

that processes HTML tags and performs part of speech tagging, a term tagger (Krauthammer *et al*., 2000) that identifies genes and proteins, a semantic disambiguation component (Hatzivassiloglou *et al*., 2001), and a component called GENIES (GENomics Information Extraction System) that extracts and structures information related to molecular pathways. GeneWays starts with a literature search to identify at least one known gene that is associated with the biological system of interest. Next, it does an automated iterative search through reference databases (such as MEDLINE), followed by automated download of complete journal articles. At each iteration GENIES captures regulatory pathways by parsing the literature to identify known genes situated in the regulatory hierarchy immediately "above" and "below" the original gene. The iteration is repeated for each of the new genes. The collected redundant and potentially controversial data are compared, "weighted," and "cleaned" by the domain experts. Networks assembled in this way are then edited, visualized, and modeled.

In this paper we focus on GENIES. It is based on an adaptation of an existing NLP system, called MedLEE (Friedman *et al*., 1994), which has been used successfully in medicine since 1995. In the Background Section, we discuss related work and present an overview of MedLEE. The Methods Section describes our modifications to MedLEE that resulted in GENIES. In the Evaluation Section, we report on our preliminary evaluation and present results, and in the Discussion Section we discuss their significance.

BACKGROUND

There are a number of projects aimed at automatic extraction of biological knowledge from electronic texts (e.g., see Iliopoulos *et al*., 2001; Park *et al*., 2001; Yakushiji *et al*., 2001, for recent overviews).

One type of system primarily identifies gene or protein names in biological texts, (Fukuda *et al*., 1998; Jenssen & Vinterbo, 2000; Krauthammer *et al*., 2000) a task that is critically important for subsequent recognition of interactions among molecular entities. Some of these systems are rule-based whereas others use external knowledge sources. GENIES exploits a term tagging component (Krauthammer *et al*., 2000) that identifies gene and protein names in text by using both rules and external knowledge sources.

Another type of system extracts both functional relations and molecular entities from text. Four such systems depend heavily on recognition of noun phrases that surround verbs of interest. Sekimizu and colleagues (Sekimizu *et al*., 1998) extract relations associated with seven different verbs (*activate, bind, interact, regulate, encode, signal*, and *function*) found in Medline abstracts. Their system finds noun phrases in a sentence that contains one of the specified verbs, and determines which are the most probable subject and object. The system's precision ranged from 67.8% to 83.3%, depending on the particular verb used. Rindflesch and colleagues (Rindflesch *et al*., 1999; Rindflesch *et al*., 2000) report on two different systems; one finds noun phrases in MEDLINE abstracts related to the process of *binding* of substances and the other, EDGAR, identifies relationships between genes, and drugs in cancer therapy. Both systems use a part-of-speech tagger, NLP techniques developed for the Specialist language-processing system (McCray *et al*., 1996), the Unified Medical Language System (UMLS; Humphreys *et al*., 1998), GenBank (Benson *et al*., 2000), and other knowledge sources and contextual rules to identify noun phrases that have an appropriate semantic type. To identify candidate noun phrases associated with the *binding* relation, the first of the two analyzes those to the left and right of the verb *bind*. A formal evaluation found a recall of 72% and a precision of 79%. In contrast, EDGAR is more complex; it first identifies candidate noun phrases based on semantic classification, and then attempts to identify interactions of drugs, genes, and cells. The task of identifying predications is based on a partial parser that is in early stages of development. Blaschke and associates (Blaschke *et al*., 1999) extract protein interactions from Medline articles without relying on linguistic knowledge.

Other systems employ more complex NLP. Hafner and colleagues (Hafner *et al*., 1994) built a prototype system to populate a knowledge base of experimental processes and analytic techniques that are described in the Materials and Methods sections of biological-research papers. This team undertook a parsing experiment, which involved simplification of a sample set of sentences that contained the verbs *measure, determine, compute*, and *estimate*. They developed a grammar that contained semantic phrases intended for parsing the simplified sentences. This system is at an early stage of development, and recognizes only simplified sentences that conform *exactly* to the grammar.

Several systems were developed using technology developed for the Message Understanding Conferences (MUCs). Thomas and associates (Thomas *et al*., 2000) report on Highlight, a system that uses part-of-speech tagging and partial parsing of certain syntactic structures, such as noun phrases. It also uses discourse analysis to identify co-referring noun phrases, and then uses domain-specific patterns to map relevant information to templates that contain slots for specific information. For example, one pattern looks for a noun phrase followed by a verb, a particle, and another noun phrase. The system captures only the subset of protein interactions associated with the verb phrases *interact with, associate with*, and *bind to*. A

given template is ranked according to a measure of confidence that it is filled correctly, depending on factors such as confidence that each noun phrase is a protein, number of times the relation occurs, and modality associated with the relation. Highlight's overall recall ranged from 29% to 58%; its precision ranged from 69% to 77%. Humphreys and colleagues (Humphreys *et al*., 2000) developed two information extraction systems named EMPathIE and PASTA. The first, EMPathIE, captures enzyme interactions; the second, PASTA, captures information concerning the role of aminoacids in protein molecules. These systems are designed similarly to the one described by Thomas and associates. The Humphreys' systems have an overall recall of 77% and a precision of 94% for extracting information about interactions.

Two systems that identify protein interactions use more comprehensive parsing techniques than partial parsing. Park and colleagues (Park *et al*., 2001) capture protein interactions using a part of speech tagger plus rules to identify unknown words, a regular grammar to gather information about neighbors of keywords, and a parser that uses a type of grammar formalism called CCG (combinatory categorical grammar) to scan the neighbors to the left and right of the interaction to evaluate candidate noun phrases as legitimate arguments of the interaction and to obtain a parse of the sentence. In addition to part of speech tagging, the CCG grammar also requires a CCG lexicon, which assigns CCG categories to word entries. The recall and precision of the system was reported to be 48% and 80% respectively. Yakushiji and colleagues (Yakushiji *et al*., 2001) also describe a system that obtains a full analysis of the sentences. This system not only captures relationships between substances but also between events, which is a more complex task because it identifies the dependencies or sequences of the events. A full analysis of the sentence is used to map the sentence to more regularized form called an argument structure; this process involves identification of the underlying verb, its subject and object arguments as well as modifiers. Two preprocessors are used to reduce ambiguity in the syntactic parsing stage. One preprocessor identifies and semantically classifies noun phrases that are technical terms; these are treated as atomic units in the parsing stage. The second preprocessor uses local constraints instead of part of speech tagging to reduce lexical ambiguity. After the parsing stage, a domain-specific rule-based component is used to map the regularized form to frames representing substances, events, and their relationships. A preliminary evaluation showed that 23% of the relationships were extracted uniquely and another 24% were extracted with ambiguity. Measures for precision were not given.

Our system, GENIES is similar to Hafner's system in using a semantic grammar, but it also includes substantial syntactic knowledge interleaved with semantic and syntactic constraints; it works with the original complex (rather than simplified) sentences. It is similar to Park's and Yakushiji's systems in that it always attempts to obtain a complete parse in order to achieve high precision; however, if a sentence cannot be parsed exactly according to the grammar rules, GENIES uses alternative strategies, such as segmenting and partial parsing to achieve improved recall. GENIES is also similar to Yakushiji's system because, unlike the other related systems, it also captures relations between interactions, which is more complex than capturing direct binary interactions between two proteins. Moreover, the subject or object of an interaction does not have to be a protein; it may be a process, such as *myogenesis*; a tissue, such as *T-cells* or a relationship between interactions. For example, in *Pax-3 may mediate activation of myod*, the object of *may mediate* is the interaction *activation of myod*. GENIES is capable of extracting complex nested chains of interactions, as we illustrate in Section 4. GENIES is different from related systems in several other ways. First, it parses *complete journal articles*, rather than only abstracts. Second, rather than extracting only binding- or enzyme-related interactions, GENIES semantically classifies and captures a complete set of interactions and relationships between biological molecules. Currently, it recognizes about 125 different verbs that are important in this field, and partitions them into 14 broader semantic classes; we plan further expansion as we identify missing interactions and semantic classes. GENIES also handles nominalized and agentive forms of verbs, such as *inhibition* and *inhibitor*, which occur frequently in this domain. Third, it assigns semantic features to verbs of interest, such as the number of arguments expected and the argument order, as described in the Methods Section.

MEDLEE OVERVIEW

MedLEE consists of several modular components divided according to functionality. Figure 2 shows the programming components as rectangles, and the knowledge sources as ovals. Its developers designed MedLEE to facilitate application to other domains; which they can accomplish by creating new domain-specific knowledge sources while leaving the programming components as they are.

The first component, the *preprocessor*, delineates the sections of the report, and identifies individual sentences. It does lexical lookup to identify and categorize single and multiword phrases in each sentence, and to determine the target output forms. Its output is (1) a list of elements, where each element is either a single word or a list of words that constitutes an atomic phrase, and (2) a separate list that contains the categories and target forms.

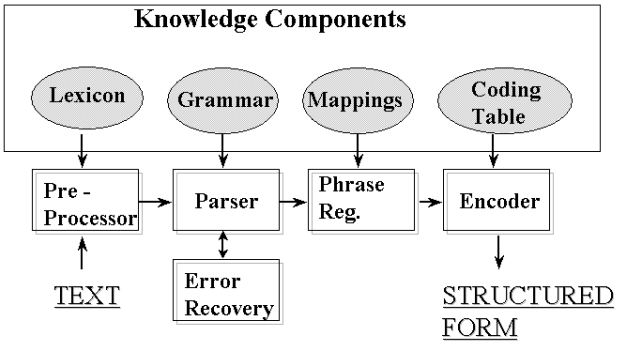

Fig. 2. Overview of the components used by MedLEE. The ovals represent knowledge components and the rectangles programming components.

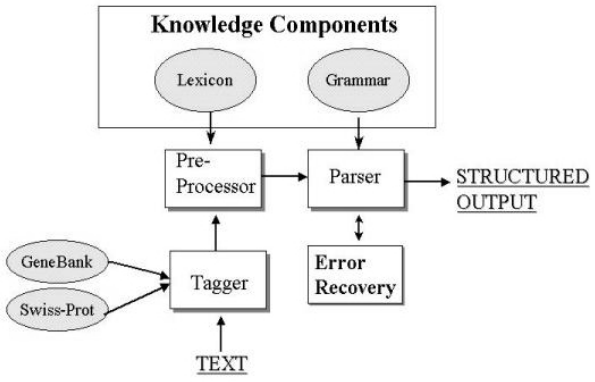

Fig. 3. Current architecture of GENIES. There are two internal knowledge sources: a lexicon and a grammar; three processing components: a preprocessor, parser, and error recovery component; a plug-in component term tagger that utilizes two external knowledge sources (GenBank (Benson *et al*., 2000) and SwissProt (Bairoch & Apweiler, 2000)).

For example, a sentence such as *severe radiating chest pain* would be represented as **[severe, radiating, [chest, pain]]**.

The second component is the *parser*. It uses the categories assigned to the words of the sentence and follows the grammar rules to recognize well-formed structures as well as to generate target forms. The target output is in the form of a primary frame consisting of a **type–value** pair followed optionally by other frames, which represent modifiers of the primary frame. In each frame, **type** represents the type of information, and **value** represents the value. Thus, for our sample sentence, the output would be **[problem, chest pain, [degree, severe], [descriptor, radiating]]**, where **degree** is a **type**, and **severe** is a **value**. Similarly, **problem** is a **type**, and **chest pain** is a **value**. There are two modifier frames denoting the information types **degree** and **descriptive**; they have the values **severe** and **radiating**, respectively.

If a parse of the sentence is not obtained, but the sentence contains relevant clinical information, the *error-recovery* component is activated. This component uses various strategies to break up the sentence into segments and to parse the segments. In effect, this component relaxes the initial strict parsing and lexical requirements, which achieve high specificity when possible.

The third and fourth components are not used by GENIES. The *compositional-regularization* component composes phrases that have been separated in the sentence. For example, this component would combine the output forms for *chest* and *pain* in the output form for *radiating pain was experienced in chest* so that the primary information is **chest pain**. The fourth component is the *encoding* component, which maps the target forms generated by the previous phase into a specified coded controlled vocabulary. For example, if the controlled vocabulary is the UMLS, the value **chest pain** is mapped to the corresponding UMLS code **C0008031**. Use of a coded vocabulary facilitates retrieval and subsequent access to the extracted information.

Numerous evaluations of MedLEE have been carried out independent of the system developers (Elkins *et al*., 2000; Hripcsak *et al*., 1995; Hripcsak *et al.*, 1998; Jain & Friedman, 1997; Jain *et al*., 1996; Knirsch *et al*., 1998). MedLEE performed well, and the evaluators concluded that the system was safe for use in real-world clinical applications. (MedLEE is currently used in the production mode at the New York Presbyterian Hospital in the New York City.)

METHODS

GENIES Overview

We implemented GENIES by combining three existing processing components of MedLEE with our own newly developed component, *term tagger* (Figure 3):

- **Term Tagger:** This plug-in component currently uses BLAST techniques, specialized rules, and external knowledge sources to identify and tag genes and proteins in the text articles
- **Preprocessor:** This MedLEE component determines sentences, words, and phrases, and performs lexical lookup.
- **Parser:** This MedLEE component uses a grammar consisting of semantic patterns interleaved with syntactic and semantic constraints to identify relevant relationships and to specify target output forms.

- **Error recovery:** This MedLEE component uses various strategies to parse segments of a sentence.

Target Structure

GENIES' basic output is similar to MedLEE's. It consists of frames, where each frame is a list beginning with the elements **type, value**, possibly followed by additional frames. The output for information associated with objects and their properties is slightly different from that associated with actions. For example, the output for a protein object *Il-2* is a type–value frame **[protein, Il-2]**. If the object has a modifier, it is represented as a nested frame; for example, the output for *activated Il-2* is **[protein, Il-2, [state, active]]**. In this example, *activated* is interpreted to be a state with a target value **active**.

Unlike objects, actions, in addition to the **type–value** pair, have ordered arguments. For example, the output for *Raf-1 activates Mek-1* has a subject and complement frame as follows: **[action, activate, [protein, Raf-1], [protein, Mek-1]]**. This representational form can model information that is both complex and nested, because actions have subject or complement arguments that are not only other objects but also may be actions or processes. For example, the output for the phrase *mediation of sonic hedgehog-induced expression of Coup-Tfii by a protein phosphatase* is shown below with indentation added for ease of comprehension:

[action,promote,[geneorprotein,phosphatase],
[action,activate,[geneorprotein,sonic hedgehog],
[action,express,X,[geneorprotein,Coup-Tfii]]]]

In this phrase, there are three actions: **promote**, which is the target output form for *mediation*; **activate**, which is the target form for *induced*; and **express**, which is the target form for *expression*. The agent of the primary action **promote** is an object that is a gene or protein that has the value **phosphatase**. The complement of **promote** is a nested action **activate**, which has an agent that is a gene or protein object **sonic hedgehog**. **Activate** also has a complement, which is another nested action **express** with an unknown subject represented by **X**, and a complement which is an object consisting of **Coup-Tfii**.

Another case of a nested action occurs when a substance is modified by a relative clause or another type of sentential clause containing an interaction. For example, in *Anergic alloantigen-specific human T-cells contain phosphorylated Cbl that coimmunoprecipitated with Fyn*, the protein Cbl and the adjoining relative clause *that coimmunoprecipitated with Fyn* would be represented as:

[protein,Cbl,[state,phosphorylated],
[action,attach,[protein,Cbl,[state,phosphorylated]],
[protein,Fyn]]]

Table 1. Semantic classes associated with actions, processes, and other relations.

Class	Actions and Processes
activate	hasten, incite, up-regulate
attach	bind, form complex, add
breakbond	sever, cleave, dephosphorylate
cause	based on, due to, result in
contain	contain, container
createbond	methylate, phosphorylate
generate	express, produce, overexpress
inactivate	repress, suppress, down-regulate
modify	mutate, modify
process	myogenesis, apoptosis, cell cycle
react	interact, react
release	disassemble, discharge
signal	regulate
substitute	replace, substitute

This nested action **attach** corresponds to the target form of *coimmunoprecipitated with*; it has a subject **phosphorylated Cbl** that is the same as the outer host *phosphorylated Cbl*. The representation illustrated above would itself be nested because it would occur as the complement of the relation **contain**; the subject would be the frame representing *anergic alloantigen-specific human T-cells*.

Semantic Categories

The bulk of our work to create GENIES was the development of a grammar and a lexicon. We started with establishing semantic categories for the extraction system – this task in its turn required development of an ontology for the signal transduction domain (Rzhetsky *et al*., 2000). The semantic categories identify relevant information to extract and are used by the lexicon and the grammar. A portion of the categories overlap with MedLEE's (for example, **certainty** (e.g., *no*), and **connective** (e.g., *after*)), but most of the categories are specific to genomic functional information (for example, types of substances and biological interactions). Table 1 contains descriptions and examples of the semantic classes associated with interactions and certain relations. Examples of semantic classes associated with objects are **amino acid, cell, complex, domain, DNA region, gene, protein, site, small molecule, species, state,** and **substance**.

Term Tagger

As explained, the term tagger (Krauthammer *et al*., 2000) identifies genes and proteins by using BLAST techniques, special rules, and external knowledge sources, such as GenBank (Benson *et al*., 2000) and Swiss-Prot (Bairoch & Apweiler, 2000). The term tagger is a plug-in component and GENIES is designed to operate with or without it.

When the tagger identifies a term, it encloses it in an XML tag. If the tagger is not used, the names of genes and proteins must be incorporated into the lexicon. The tagger significantly increases the flexibility of the system in the quickly changing environment (new gene and protein names appear weekly!): it eliminates the lexical effort that would be required to update the lexicon frequently, and it allows for a more tailored treatment of specialized textual terms.

Preprocessor

The preprocessor separates the article into sentences and the sentences into single words or atomic multiword phrases. Words and atomic phrases are identified via tags or lexical lookup. If a phrase has a tag, the preprocessor records the information associated with the tag and bypasses lexical lookup. Lexical lookup is used for terms in the domain that are not identified by the tagger. Currently, there are about 530 entries in the lexicon that correspond to phrases associated with interactions and relations.

The lexicon developed for GENIES contains several informational categories that overlap with MedLEE for words associated with certainty, degree, and quantitative information, as well as categories associated with functional words such as prepositions and conjunctions. An entry for an object or property in GENIES is similar in form to an entry for a clinical term in MedLEE: it specifies only the semantic category and target form. In contrast, the entries associated with interactions and certain relations contain additional information that is needed for accuracy. Interactions are associated with a particular number of arguments, and the arguments appear in a certain order; as such this additional information is specified in the lexical entries. For example, *activates* is associated with two arguments, whereas *transcribes* is associated with one. In *X activates Y*, the first argument of activate, *X*, is the agent, and the second argument, *Y*, is the complement. The relation *attributable to* also has two arguments, but the situation is reversed: In *X is attributable to Y*, the target form is cause, the second argument, *Y*, is the agent and the first argument, *X*, is the complement.

Syntactic classifications are also maintained in the lexicon for entries associated with interactions. The classifications are associated with verbal forms: **v, vp, ved, ven, ving, vn, vor** (e.g., *activate, activates, activated, activated, activating, activation, and activator*); the parser uses these categories to constrain the grammar patterns associated with interactions.

Parser

The parser uses grammar rules to recognize well-formed patterns and to generate target output. A grammar rule consists of (1) specification of semantic and syntactic components, (2) specification of the target form if the rule is successful, and (3) constraints that ensure that the components are well formed. We developed our grammar manually by observing typical semantic and syntactic co-occurrence patterns in sample texts. The grammar is a definitive clause grammar (DCG) (Pereira & Warren, 1980). The initial implementation recognized simple binary actions, such as *X activates Y*; we then incrementally added the ability to handle more complex structures, such as nested actions, modifiers of objects and actions, relations between actions, relative clauses, and conjunctions. The following is output obtained for the sentence *phosphorylated* **Cbl** *coprecipitated with* **CrkL***, which was constitutively associated with the* **C3G**:

[action,attach,[protein,Cbl,[state,phosphorylated]],
[protein,CrkL,[action,attach,[protein,CrkL],
[protein,C3G]]]]

In the example, there is one nested action and one primary action. Both actions are **attach**, corresponding to the target output form associated with *coprecipitated with* and *associated with*; the agent and complement in the nested action are the proteins **CrkL** and **C3G** respectively. The outer action has an agent, which is the protein **Cbl**, and a complement, which is the protein **CrkL** that contains a nested action. Note that, for the protein **Cbl**, GENIES also captured the state, **phosphorylated**. The grammar rules for GENIES are the same in form as, but are substantially different in content from, the rules used by MedLEE (although there are overlapping rules). Both grammars contain relations that connect two interactions, such as *during* and *and*. For example, the output for the phrase *Cbl tyrosine phosphorylation during induction of anergy* is a relation **during** that connects two actions **phosphorylate** and **activate**, both of which have unspecified agents:

[rel,during,[action,phosphorylate,X,[protein,Cbl],
[site,[aminoacid,tyrosine]]],
[action,activate,X,[state,anergy]]]

EVALUATION

We performed a pilot evaluation study. We compared GENIES' output to that obtained manually by an expert. Independently of system developers, the expert chose an article from *Cell*, (Maroto *et al*., 1997), read it, and highlighted those sentences that he judged to contain information relevant to signal transduction pathways. In addition, he noted the binary and nested relations of interest for each highlighted sentence. We also used GENIES to extract signal

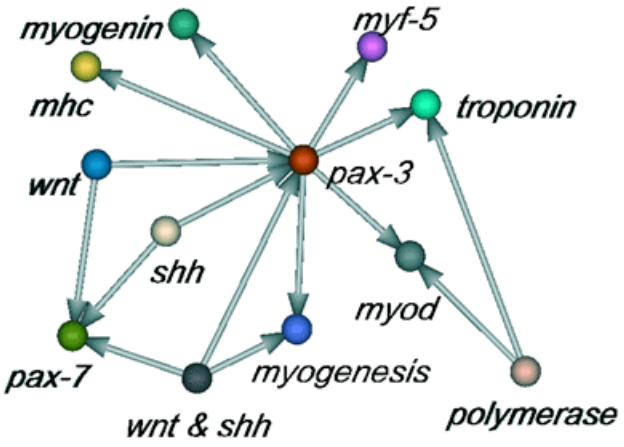

Fig. 4. Molecular interactions that were extracted by GENIES from the test Cell article were visualized by the authors with the program CUtenet. (Koike & Rzhetsky, 2000)

transduction-related information from the same article *automatically*, and to obtain the structured output. We then calculated the recall and precision measurements. Recall was computed as the number of correct relations extracted by GENIES divided by the number obtained by the expert, and precision was computed as the number of correct relations extracted by GENIES divided by all relations extracted by GENIES. Measurements for binary and nested relations were computed separately.

Results of evaluation

The article contained 7,790 words and took 1.3 minutes to process on a 500 MHZ PC with 128 MB RAM. The expert identified 51 binary relations; GENIES correctly extracted 27 (53%) stemming from the same sentences. GENIES' precision was thus 100%. Many of the relations were redundant: in the whole article only 19 relations were unique. Of the 19, GENIES retrieved 12 (63%; Figure 4). Thirteen of the relations identified by the expert contained nesting; GENIES captured 8; 7 were correct and 1 was incorrect (54% sensitivity; 88% precision). GENIES identified 30 binary relations not noted by the expert for a total of 57. The expert evaluated the additional relations, and judged that only two were incorrect. Therefore, we evaluated the precision of GENIES (when considering all binary information) for extracting binary relations as 96% (55/57).

DISCUSSION

Our pilot evaluation was based on only one article; The article was chosen by the expert, rather than by the system developers. The work that the expert performed was considerable (a few hours) and in general it is rather difficult to find volunteers for such an evaluation. To address this problem, we are currently developing tools to assist the expert in recording and editing interactions. Ideally, we would like to evaluate the system with a large number of articles (containing several hundred relations), although that would require an extraordinary amount of work. We have subsequently processed 140 complete journal articles in preparation for a second more comprehensive evaluation.

GENIES processes complete articles, whereas other systems process abstracts only. Processing complete articles has clear-cut advantages, although the evaluation effort is substantially more time consuming. Complete articles contain more interactions than do abstracts (only *7 of the 19* unique interactions that the expert identified were mentioned in the abstract of the *Cell* paper (Maroto *et al.*, 1997)), and therefore more information, than the abstracts. Another advantage is that complete articles contain redundant information; if an interaction is not extracted from one part of the article by the system, it may still be extracted from another portion of the article. In our evaluation, the expert noted that there were 51 interactions overall, but that only 19 were unique. An additional benefit is that recognition of a redundant interaction may justify increased confidence in the correctness of the interaction.

We analyzed the errors that occurred and found that two types caused the majority of the problems. One type was due to an incomplete lexicon. For example, *expands* was not in the lexicon, causing an incorrect interpretation of *Ectopic expression of SHH expands myod expression in the paraxial mesoderm of either chick*. Future extension of the lexicon is likely to reduce this type of error. The second type of error was caused when a well-formed pattern that was covered by the grammar was interrupted in the middle by information not in the grammar rule or in the lexicon; sometimes this was due to non-critical information, (e.g. *directly or indirectly* in *Pax-3 directly or indirectly activates Myod expression*), and sometimes to lexical omissions (.e.g. in *Inhibition of hedgehog signaling in the paraxial mesoderm of zebrafish reduces myod expression*, the phrases *paraxial mesoderm* and *zebrafish* were not in the lexicon, thereby interrupting *inhibition of hedgehog signaling reduces myod expression*). A few sentences were not interpreted correctly because they were very complex. For example in the output for *Wnt or SHH signals alone are insufficient to induce high level expression of either pax-3 or pax-7*, the system incorrectly represented that either Wnt signals were insufficient to induce expression or SHH signals were insufficient. Other types of errors that occurred infrequently were due to incorrect tagging and lack of a discourse component.

One of our goals in developing GENIES was to achieve a high accuracy. Our preliminary results for precision,

96% (binary relations) and 88% (nested relations) respectively made us optimistic that we would achieve our goal. This measure compares favorably with the evaluation results of the related systems discussed in Section 2. The results for precision also are comparable to those that we obtained for MedLEE in the medical domain (Friedman & Hripcsak, 1998). Because MedLEE is a mature system, it has been evaluated numerous times, often independently of the system developers. Our results in the GENIES evaluation demonstrate that similar methods can be used to extract accurate information associated with molecular pathways, but that syntactic knowledge is more useful in this domain than in the medical domain (data not shown).

Future work in improving GENIE will consist of (1) extending the lexicon; (2) adding more patterns to the grammar; (3) improving partial parsing techniques; (4) integrating the medical domain with the molecular domain to capture molecular pathways, diseases, drugs, and their relationships; this task should be relatively straightforward in that GENIES and MedLEE use the same processing engine; only the grammars and lexicons differ and need to be combined. Future work involving the other NLP components of GeneWays that affect GENIES will involve (5) establishing a method for the unique identification of synonymous forms of substances; (6) improving the term tagger; (7) improving the term disambiguation component and linking it to GENIES; (8) adding a discourse component; (9) adding a filter to weed out irrelevant text, such as references, and to recognize special structures, such as titles and captions; (10) resolving conflicting hypotheses within one article and between articles.

SUMMARY

We have demonstrated that it is possible to apply the general information-extraction system MedLEE, previously applied to the domain of clinical records to the domain of literature associated with molecular information. Our pilot evaluation demonstrated high precision (96%) and satisfactory recall (63%). To build GENIES, we created a new knowledge model, (Rzhetsky *et al*., 2000), new semantic categories, a new lexicon, a new grammar, and new representational frames. We also made basic changes to use specialized processes and external knowledge sources to identify genes and proteins, and we incorporated more syntactic and semantic features into the grammar, particularly for verbs of interest. We will continue to refine, improve, and evaluate GENIES because it demonstrated its effectiveness for acquiring worthwhile knowledge from journal articles.

ACKNOWLEDGEMENTS

This publication was supported in part by grants LM06274 from the National Library of Medicine and by the Columbia CAT supported by the NYS Science and Technology Foundation

REFERENCES

Ashburner M., Ball C. A., Blake J. A., Botstein D., Butler H., Cherry J. M., Davis A. P., Dolinski K., Dwight S. S., Eppig J. T., Harris M. A., Hill D. P., Issel-Tarver L., Kasarskis A., Lewis S., Matese J. C., Richardson J. E., Ringwald M., Rubin G. M., and Sherlock G. (2000). Gene ontology: tool for the unification of biology. *The Gene Ontology Consortium. Nat. Genet,* **25,** 25-9.

Bairoch A., and Apweiler R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.,* **28,** 45-8.

Baker P. G., Goble C. A., Bechhofer S., Paton N. W., Stevens R., and Brass A. (1999). An ontology for bioinformatics applications. *Bioinformatics,* **15,** 510-20.

Benson D. A., Karsch-Mizrachi I., Lipman D. J., Ostell J., Rapp B. A., and Wheeler D. L. (2000). GenBank. *Nucleic Acids Res.,* **28,** 15-8.

Blaschke C., Andrade M. A., Ouzounis C., and Valencia A. (1999). Automatic extraction of biological information from scientific text: protein-protein interactions. *Ismb,* 60-7.

Chen R. O., Felciano R., and Altman R. B. (1997). RIBOWEB: linking structural computations to a knowledge base of published experimental data. *Ismb,* **5,** 84-7.

Elkins J. S., Friedman C., Boden-Albala B., Sacco R. L., and Hripcsak G. (2000). Coding neuroradiology reports for the Northern Manhattan Stroke Study: a comparison of natural language processing and manual review. *Comput. Biomed. Res.,* **33,** 1-10.

Friedman C., Alderson P. O., Austin J. H., Cimino J. J., and Johnson S. B. (1994). A general natural-language text processor for clinical radiology. *J. Am. Med. Inform. Assoc.,* **1,** 161-74.

Friedman C., and Hripcsak G. (1998). Evaluating natural language processors in the clinical domain. *Methods Inf. Med.,* **37,** 334-44.

Fukuda K., Tamura A., Tsunoda T., and Takagi T. (1998). Toward information extraction: identifying protein names from biological papers. *Pac. Symp. Biocomput.,* 707-18.

Hafner C. D., Baclawski K., Futrelle R. P., Fridman N., and Sampath S. (1994). Creating a knowledge base of biological research papers. *Ismb,* **2,** 147-55.

Hatzivassiloglou V., Duboue P. A., and Rzhetsky A. (2001). Disambiguating proteins, genes, and RNA in text: a machine learning approach. *Ismb,*(accepted).

Hripcsak G., Friedman C., Alderson P. O., DuMouchel W., Johnson S. B., and Clayton P. D. (1995). Unlocking clinical data from narrative reports: a study of natural language processing. *Ann. Intern. Med.,* **122,** 681-8.

Hripcsak G., Kuperman G. J., and Friedman C. (1998). Extracting findings from narrative reports: software transferability and sources of physician disagreement. *Methods Inf. Med.,* **37,** 1-7.

Humphreys B. L., Lindberg D. A., Schoolman H. M., and Barnett G. O. (1998). The Unified Medical Language System: an

informatics research collaboration. *J. Am. Med. Inform. Assoc.,* **5,** 1-11.

Humphreys K., Demetriou G., and Gaizauskas R. (2000). Two applications of information extraction to biological science journal articles: enzyme interactions and protein structures. *Pac. Symp. Biocomput.,* 505-16.

Iliopoulos I., Enright A. J., and Ouzounis C. (2001). TEXTQUEST: Document Clustering of MEDLINE Abstracts For Concept Discovery In Molecular Biology. *Pacif. Symp. Biocomp.,* **6,** 374-383.

Jain N. L., and Friedman C. (1997). Identification of findings suspicious for breast cancer based on natural language processing of mammogram reports. *Proc. AMIA Annu. Fall Symp.,* 829-33.

Jain N. L., Knirsch C. A., Friedman C., and Hripcsak G. (1996). Identification of suspected tuberculosis patients based on natural language processing of chest radiograph reports. *Proc. AMIA Annu. Fall Symp.,* 542-6.

Jenssen T. K., and Vinterbo S. (2000). A set-covering approach to specific search for literature about human genes. *Proc. AMIA Symp.,* 384-8.

Kanehisa M., and Goto S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.,* **28,** 27-30.

Karp P. D., Riley M., Paley S. M., Pellegrini-Toole A., and Krummenacker M. (1999). Eco Cyc: Encyclopedia of Escherichia coli genes and metabolism. *Nucleic Acids Res.,* **27,** 55-58.

Knirsch C. A., Jain N. L., Pablos-Mendez A., Friedman C., and Hripcsak G. (1998). Respiratory isolation of tuberculosis patients using clinical guidelines and an automated clinical decision support system. *Infect. Control Hosp. Epidemiol.,* **19,** 94-100.

Koike T., and Rzhetsky A. (2000). A graphic editor for analyzing signal-transduction pathways. *Gene,* **259,** 235-244.

Krauthammer M., Rzhetsky A., Morozov P., and Friedman C. (2000). Using BLAST for identifying gene and protein names in journal articles. *Gene,* **259,** 245-252.

Maroto M., Reshef R., Munsterberg A. E., Koester S., Goulding M., and Lassar A. B. (1997). Ectopic Pax-3 activates MyoD and Myf-5 expression in embryonic mesoderm and neural tissue. *Cell,* **89,** 139-48.

McCray A. T., Razi A. M., Bangalore A. K., Browne A. C., and Stavri P. Z. (1996). The UMLS Knowledge Source Server: a versatile Internet-based research tool. *Proc. AMIA Annu. Fall Symp.,* **,** 164-8.

Park J. C., Kim H. S., and Kim J. J. (2001). Bidirectional Incremental Parsing for Automatic Pathway Identification with Combinatory Categorial Grammar. *Pacif. Symp. Biocomp.,* **6,** 396-407.

Pereira F. C. N., and Warren D. (1980). Definite clause grammars for language analysis – a survey of the formalism and comparison with augmented transition networks. *Artificial Intelligence,* **13,** 231-278.

Rindflesch T. C., Hunter L., and Aronson A. R. (1999). Mining molecular binding terminology from biomedical text. *Proc. AMIA Symp.,* 127-31.

Rindflesch T. C., Tanabe L., Weinstein J. N., and Hunter L. (2000). EDGAR: extraction of drugs, genes and relations from the biomedical literature. *Pac. Symp. Biocomput.,* **,** 517-28.

Rzhetsky A., Koike T., Kalachikov S., Gomez S. M., Krauthammer M., Kaplan S. H., Kra P., Russo J. J., and Friedman C. (2000). A knowledge model for analysis and simulation of regulatory networks. *Bioinformatics,* **16,** 1120-1128.

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.,* **9,** 62-71.

Selkov E., Galimova M., Goryanin I., Gretchkin Y., Ivanova N., Komarov Y., Maltsev N., Mikhailova N., Nenashev V., Overbeek R., Panyushkina E., Pronevitch L., and Selkov E., Jr. (1997). The metabolic pathway collection: an update. *Nucleic Acids Res.,* **25,** 37-8.

Thomas J., Milward D., Ouzounis C., Pulman S., and Carroll M. (2000). Automatic extraction of protein interactions from scientific abstracts. *Pac. Symp. Biocomput.,* 541-52.

Yakushiji A., Tateisi Y., Miyao Y., and Tsujii J. (2001). Event Extraction from Biomedical Papers Using a Full Parser. *Pacif. Symp. Biocomp.,* **6,** 408-419.