BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–9

Using Lexical Chaining to Rank Protein-Protein Interactions in Biomedical Text

Xiaofen He [a] *and Chrysanne Di Marco* [a*]

[a] *School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada*

ABSTRACT

Biomedical information extraction is becoming an increasingly important application of Computational Linguistics research. We propose a method for analyzing full-text articles on protein interactions that takes a discourse-based approach to provide a means of ranking the biological validity of such interactions. Specifically, we use *lexical chaining*—strings of semantically related words—as an indicator of the validity of the protein interactions appearing in the same textual context.
Contact: cdimarco@uwaterloo.ca

1 INTRODUCTION

Each living cell is rich in proteins that continuously interact with each other. Knowledge about the identities and functions of interacting proteins contributes significantly to the understanding of biological processes by providing insight into the roles of important genes, elucidating relevant pathways, and facilitating the identification of potential drug targets for use in developing novel therapies.

A large volume of protein-protein interactions has been identified, and information about such interactions is now readily available in online databases such as BIND (Bader *et al*., 2001). However, the information stored in current databases does not allow us to rank the biological validity of the interactions—it may be the case that interactions occurring under laboratory conditions do not actually occur in the living cell. A researcher trying to establish the quality of the interactions identified in a database could read the details of the experiments in each related scientific article, but this is labourious and time-consuming. If the number of relevant papers is high, it will be difficult or even impossible for a researcher to manually process all the articles to assess the value of the interactions. For example, a text query in BIND for interactions of the single protein Cdc42 will retrieve 512 records, far too many to be easily read and analyzed by manual methods—there is a clear need for an automated information extraction system to assist researchers in analyzing the online literature to better judge the quality of biomolecular interactions.

1.1 Natural Language Processing and information extraction

Natural Language Processing (NLP) techniques are now widely used in biomedical information extraction (IE). The general approach to using natural language methods in automated information extraction involves a detailed analysis of basic grammatical features (e.g., identifying each word's part-of-speech) and then a shallow analysis of deeper syntactic structure using targetted grammatical rules to identify simple syntactic patterns ('templates') or basic grammatical units (e.g., noun phrases, verb phrases) within the sentence.

Representative approaches to extracting information from biomedical texts include: using the frequency of "discriminating words" to score paper abstracts to determine whether the paper is about protein interactions (Marcotte *et al*., 2001); simple-template–based parsing of sentences to build networks of protein interactions (Blaschke *et al*., 1999); and a general-purpose information-extraction engine using both symbolic and statistical Computational Linguistic techniques to build a database of protein interactions (Thomas *et al*., 2000). However, these approaches are inherently limited: they currently target only paper abstracts, they deal with only a single sentence at a time, and they use simplified methods of linguistic analysis. As a consequence, these current approaches to biomedical information extraction miss a great deal of the detailed information on protein interactions that is contained in the text.

Potentially a great deal of additional information on protein interactions could be extracted from scientific articles if we were able to analyze the entire text of the article to derive detailed linguistic information such as lexical meanings, syntactic structure, semantic content, and discourse structure. However, the present-day state of Computational Linguistics is still not sufficiently advanced to handle these difficult problems even for restricted sublanguages and certainly not for the very large corpora needed for useful biomedical information extraction. Previous systems have attempted to finesse these difficulties by using a method of text analysis that approximates full syntactic processing and that takes a heuristic approach to semantic analysis based on the recognition of interactions between proteins and other molecules in the form of templates

* The authors should be regarded as joint First Authors.

matching specific linguistic patterns (cf. Thomas *et al*. 2000; Pustejovsky *et al*. 2002).

In this paper, we propose a method for extracting information on protein-protein interactions from online biological literature that aims to obtain more-detailed knowledge than previous systems and that uses both more-sophisticated Computational Linguistic methods and computationally tractable algorithms capable of handling large corpora. We base our method on the inherent biological characteristic of protein-protein relationships, namely that interacting proteins will tend to have similar biological functions:

> …Although proteins from different groups of biological functions can still interact with each other, it has been shown that the degree to which interacting proteins are annotated with the same functional category is a measure of quality for the predicted interactions (von Mering *et al.* 2002).

We may reasonably expect then to find biological terms in the context surrounding a protein interaction that indicate the common functions of these proteins. If we can determine such terms by an automated method of linguistic analysis, we would have an additional means of discovering evidence in the literature that the interaction is indeed biologically valid.

The idea of using semantically related strings of words to determine the topic structure of text is known as *lexical chaining* (Morris and Hirst, 1991), a method that fulfils our dual criteria of being both discourse-based and computationally efficient. We propose to use lexical chains to retrieve additional information on protein interactions by finding the biological terms in the passage surrounding an interaction that form the theme structure of the text. Our method requires readily available linguistic and biomedical resources: an online lexical thesaurus (e.g., WordNet; Fellbaum *et al*. 1998) and shallow syntactic parsers, as well as biological and medical ontologies (e.g., Unified Medical Language System http://www.nlm.nih.gov/research/umls), which provide semantic and conceptual knowledge. By constructing the lexical chains related to protein interactions, we will not only extract additional important information about interactions from the literature, but we hypothesize that we will also be able to use the strength of the chains to rank the apparent quality of the interactions.

2 BACKGROUND AND RELATED WORK

2.1 What is Lexical Chaining?

The notion of lexical chaining was first introduced by Morris and Hirst (1991), and derives from the concept of textual cohesion. The linguistic study of textual cohesion shows that a text or discourse is not just a set of sentences, each on some random topic; rather, the sentences and phrases of any sensible text tend to 'stick together' by various means to form a unified whole. There are a number of forms of textual cohesion, such as grammatical cohesion (reference, substitution, ellipsis, conjunction) and lexical cohesion (i.e., semantically related words). Lexical cohesion arises from semantic relationships between words, and is the most frequent and most easily identifiable type of cohesion. Halliday and Hasan (1976) classifed lexical cohesion into two categories, reiteration and collocation. Reiteration includes not only repetition and reference, but also superordinates, subordinates, synonyms, and hypernyms/hyponyms. Collocation is defined as semantic relationships between words that often co-occur in the same lexical contexts. As an illustration, the following passage shows several types of lexical cohesion.

(1) John has a Jaguar.

(2) He loves the car.

(3) John works in the garage taking care of his Jaguar.

In this passage, the word *Jaguar* in sentence (1) and sentence (3) represents the simplest form of reiteration: repetition; *Jaguar* and *car* form a part-whole relationship that falls into the category of systematic semantic collocation; *car* and *garage* have a nonsystematic semantic relationship. Lexical cohesion occurs only between two terms, but may lead to sequences of related words. A *lexical chain* may then be defined as a sequence of related words in the text, spanning a topical unit of the text, be it short (adjacent words or sentences) or long (entire text). In the passage above, a lexical chain would be {*Jaguar*, *car*, *garage*, *Jaguar*}. In general, each document will contain many lexical chains, each of which forms a portion of the cohesive structure of the document.

Lexical chains are important for computational text understanding not only because they provide a context for resolving word ambiguity, but also because they indicate the discourse structure of the text. Since lexical chaining was introduced in 1991, it has been successfully used in a number of Information Retrieval and Natural Language Processing applications, such as term weighting, malapropism detection, hypertext generation, and text summarization. In this paper, we argue that lexical chains can be used in detailed information extraction from biological literature, specifically, the assessment of the biological validity of protein-protein interactions.

2.2 How to determine lexical chains

Generally speaking, lexical chains can be computed by grouping sets of words that are semantically related (words that have relationships such as identities, synonyms, and hypernyms/hyponynms). In terms of actual computing procedures, most lexical-chaining algorithms can be summarized by the following three steps:

1. Select a set of candidate words (i.e., all noun instances).

2. For each candidate word, find an appropriate chain relying on a relatedness criterion among members of the chains.
3. If such a chain is found, insert the word in the chain; otherwise a new chain is created.

The difficult, and computationally costly, part of this process is that each candidate word must be assigned to exactly one lexical chain, and the words must be grouped in such an optimal way that these groupings create the longest/strongest lexical chains. In our research, we will adapt the lexical-chaining algorithm developed by Silber and McCoy (2002). Their linear-time algorithm was based on the complete method implemented by Barzilay and Elhadad (1997) that runs in exponential time. Silber and McCoy's method uses WordNet, an online lexical database (to be discussed below) as the knowledge source for the lexical semantic relationships used in constructing the lexical chains. The algorithm implicitly builds all possible 'metachains' for each sense of a word in WordNet; a single metachain represents all possible lexical chains for that core meaning. For each noun in the document, for every sense of the noun in WordNet, the noun sense is placed into every metachain for which it has an identity, synonym, or hypernym relation with that sense. After each noun has been inserted into one or more metachains, the next step is to find the single metachain for each noun that the noun contributes to most, based on the type of relation and distance factors. For example, identity and synonymy are considered equally strong contributors to a lexical chain over a passage of three sentences, but hypernymy is considered less strong over the same distance. When the algorithm completes, each noun will belong to only one metachain, with all its occurrences in other metachains having been removed. When all the nouns have been processed, the optimal lexical chains will remain.

2.3 Discourse analysis using lexical chaining

WordNet (Miller *et al*., 1990; Fellbaum *et al*., 1998) is an online lexical database that organizes lexical concepts according to various semantic relations. Words (nouns, verbs, adjectives, and adverbs) are each organized into 'synonym sets', known as *synsets*, each of which represents the lexical concept underlying a group of words that are synonymic or near-synonymic in a given context. Synsets can be related by various lexical semantic relations: synonymy, antonymy, hyponymy/hypernymy (subclass/superclass, also known as the *IsA* relation), and meronymy (also known as holonymy, representing various types of *part-whole* relationships). For example, the word *board* has two different senses, a piece of lumber or a group of people who have a directive role. In WordNet, there would therefore be two distinct synsets for these different concepts, {*board*, *plank*} and {*board*, *committee*}. There is also another sense of *board*, providing room and meals for regular payment, which would be represented by yet another synset, this one having only a single member. Synsets, it should be noted, do not necessarily represent exact synonymy, which is rare, but rather will usually describe a weakened form of near-synonymy by designating words that may be used interchangeably in a given sentence, i.e., without perceptibly altering the truth value of the sentence. A side-effect of this requirement that members of a synset be interchangeable is that nouns, verbs, adjectives, and adverbs must be partitioned in WordNet so that synsets in each syntactic grouping contain only words of their syntactic part-of-speech. As we will describe below, we will use a broader definition of synonymy, suited to the purpose of our study, and this will have a consequent effect on the organization of a protein-related WordNet-like lexical database.

Similarly, the other lexical semantic relations used in WordNet may be broadened for our purposes. Antonymy in WordNet has a precise meaning as a lexical relation between word forms rather than a semantic relation between word meanings. In our context, however, the concept of antonymy is better-expressed as a broader contrast in functional meaning between biological terminology. We will discuss this below in more detail.

Finally, the remaining two lexical relations in WordNet, hyponymy/hypernymy and meronymy will be applied in their existing form. Hyponymy/hypernymy will be used to describe a semantic relation of subordination/superordination (subclass/superclass) between individual word meanings. For example, *cytokinesis* is a subclass (specialization of) *cell-division*. Meronymy will be used to describe a relation between concepts (synsets) whereby a concept X is a meronym of another concept Y if each word in the synset represented by X can be described as a part of a corresponding term in synset Y. For example, if we have synsets *organelle* and *biological-structure*, with *organelle* containing *nucleus*, and *biological-structure* containing *cell*, then, given that the nucleus is a part of the cell, the synset *organelle* is thus a meronym of *biological-structure*.

Our research methodology relies on a WordNet-like concept taxonomy as the basis for automated discourse analysis in the form of lexical chaining to extract information about protein-protein interactions in biomedical literature. Specifically, we adapt existing algorithms for information extraction and lexical-chaining to provide additional knowledge on the biological validity of these interactions.

3 A LEXICAL CHAINING ALGORITHM FOR RANKING PROTEIN INTERACTIONS

3.1 The algorithm

We were able to adapt two existing algorithms as the basis of our lexical-chaining algorithm for analyzing biomedical texts related to protein interactions. The first algorithm is Silber and McCoy's lexical chainer; the second is the general algorithm currently used in biomedical information extraction (e.g., Thomas *et al*. 2000; Pustejovsky *et al*. 2002). Following the

- **Preprocessing:**
  1. Tokenize input text
  2. Tag each token with appropriate part-of-speech
- **Step 1: Find protein-interaction templates**
  for each sentence
  1. Group sequences of terms into phrases (e.g., noun phrase, verb phrase) using cascaded finite state machines
  2. Determine phrases referring to protein entities or interaction events using pattern-matching techniques
  3. Link coreferring phrases (phrases that refer to same protein)
  4. Construct protein-interaction templates
- **Step 2: Find protein-interaction lexical chains**
  for each paragraph
  for context of current protein-interaction template
  - **Step 2a**
    for each biological-term instance
    for each sense of the biological term
    Compute all scored metachains
  - **Step 2b**
    for each biological-term instance
    for each metachain to which the term belongs
    Keep word instance in the metachain to which it contributes most
    Update the scores of each other metachain
- **Step 3: Compute rankings of protein interactions according to strength of their lexical chains**

Fig. 1. A lexical chaining algorithm for protein interaction texts

style of Thomas *et al*. and Pustejovsky *et al*., the initial stages of our algorithm use similar methods of shallow syntactic analysis in the form of tokenization, part-of-speech tagging, and recognition of phrasal units by cascaded finite-state machines. Simple grammatical 'templates' of protein-protein interactions are then constructed using statistical pattern-matching techniques. A typical template would be: a noun phrase, followed by a verb, a particle, then another noun phrase, as in *A interacts with/binds to/associates with B* (Thomas *et al*. 2000, p. 6).

We then integrate Silber and McCoy's lexical chainer with this parsing framework to obtain strings (the 'chains') of semantically related words which indicate the topic structure of the passage of text surrounding a protein interaction. To do this, we modify the original lexical-chaining algorithm to build chains that are composed of biologically significant terms, specifically those related to protein functions. The resulting algorithm is shown in Figure 1.

3.2 Scoring lexical chains using Hirst and St.-Onge's algorithm

In Silber and McCoy's algorithm, a critical component involves determining the relatedness of words making up a lexical chain. Initially, a noun is put into a metachain if it is in some way related to the sense with which the metachain is indexed. Subsequently, the degree to which the word contributes to the metachain must be measured in order to decide which metachains will be kept. In order to do this, we need a means of measuring the semantic relatedness of words.

There are various WordNet-based word similarity measurements (e.g., Hirst and St.-Onge 1997; Jiang and Conrath 1997; Banerjee and Pedersen 2002). In this paper we adopt Hirst and St.-Onge's measure because it is a simple and effective method easily used with a manual form of corpus analysis. Hirst and St.-Onge adapted Morris and Hirst (1991)'s semantic distance algorithm, which used Roget's thesaurus, for use with WordNet. Their method views semantic relationships between words in terms of a graph, and correlates semantic relatedness between words with the nature of the corresponding path between concepts in the graph. Semantic relatedness is then determined based on the path shape and distance between concepts using the relations connecting them in the WordNet taxonomy.

The Hirst and St.-Onge measure classifies WordNet relations as having direction (upward, downward, or horizontal), and then establishes the relatedness between two concepts A and B by finding a path that is neither too long nor that changes direction too often. Three kinds of relations are defined: extra-strong (between a word and its repetition), strong (between two words connected by a WordNet relation), and medium-strong (when the link between the synsets of the words is longer than one and satisfies certain restrictions).

As an example, two words are strongly related if one of the following holds:

1. They are members of the same synset (e.g., *human* and *person*).
2. They are associated with two different synsets connected by the antonymy relation (e.g., *human* and *object*).
3. One of the words is a compound (or a phrase) that includes the other, and there is any kind of link at all between the synsets associated with each word (e.g., *school* and *private school*).

Two words are said to be in a medium-strong relation if there exists an 'allowable' path connecting the synsets associated with each word. An allowable path involves certain patterns of links between synsets that may vary among upward (hypernymy and meronymy), downward (hyponymy and holonymy), and horizontal (antonymy).

In Hirst and St.-Onge's scheme, the strength of a lexical chain is based both on its length and the types of relationships among its members. Extra-strong relations have the highest weight, next in weight are strong relations, and lowest are medium-strong relations. Unlike extra-strong and strong relations, medium-strong relations have varied weights according to the following formula (Hirst and St.-Onge 1997, p. 308):

weight = C – path length – k ∗ number of changes of direction (where C and k are constants[1].)

The overall strength ('score') of a lexical chain may then be taken to be the sum of weights assigned to each pair of semantic relations in the chain.

4 EXPERIMENT: MANUAL RANKING OF A SAMPLE CORPUS OF PROTEIN INTERACTION ARTICLES

4.1 The corpus

We applied our lexical-chaining algorithm for protein interaction texts and method for ranking lexical chains in an initial manual study. We selected 15 articles focussing on the identification of protein-protein interactions in yeast and analyzed these by hand, first to determine the total number and nature of lexical chains in the contexts surrounding the mention of protein interactions, then to test our ranking method on a sampling of the chains. In choosing these articles, we aimed to represent a variety of the research techniques used in studying protein-protein interactions. In this way, we hoped to find a good sampling of the kinds of biological terms likely to occur in protein-interaction contexts and which would ultimately be included in our protein-related version of WordNet.

4.2 Constructing a protein-related extension of WordNet

We followed our algorithm as given in Figure 1 in analyzing by hand the lexical chains in the contexts surrounding protein-protein interactions in our sample corpus. By "context", we mean a passage of text within a single paragraph that 'talks about' a particular protein-protein interaction. In an automated analysis, we would have to rely primarily on overt discourse cues such as coreferential expressions, rhetorical markers, and lexical meanings to determine the topic structure of a text. In a manual study, we used such cues but also used a deeper understanding of the semantic content of the technical material. At this stage, as our primary goal was to collect and classify information on the regularities in biological terminology that appear in protein-protein interaction contexts, it seemed sound methodology to rely on our own human natural language processing ability; in further work, when we plan an analysis of a much larger corpus, we will adhere to the limitations inherent in automated processing, i.e., the necessarily partial linguistic analysis provided by syntactic templates and cascaded finite state machines.

We modelled the basic structure of our concept taxonomy for biological terms relevant to protein interactions on the existing concept structure in WordNet. For example, in WordNet, the concept *assembly* has in its synset the terms *construction* and *building*, while its superclasses comprise the more-general concepts *construction, building* and the most-generic concept *activity*. We based our taxonomy on the topmost concept *biological-activity* then created three hypernyms of this generic concept based on the primary activities involving the cell: *cell-death*, *cell-maintenance*, and *cell-development*. the concept *cell-maintenance* has as its synset {*cell-construction*, *cell-building*} and is then specialized further to *cell-assembly*. From these primary concepts, we built up the hierarchical structure by selecting biologically significant concepts in the contexts surrounding protein interactions and adding them into the taxonomy based on relations involving synonymy, antonymy, and hypernymy/holonymy.

Because of the specialized nature of our lexical-chaining analysis, we adapted the meanings of the classical lexical semantic relations to be more tuned to our needs. Synonymy between terms, rather than being a strict meaning relation based on truth-condition–preserving substitution, is more usefully interpreted as a relation between terms having a close similarity in biological function. So, for example, the concepts of *defect*, *mutant*, and *mutation* will all be in the same synset. Antonymy, as well, seems more appropriately defined for our needs as a contrast in terms of biological function; thus, the concepts *cell-death* and *cell-growth* will be antonyms of one another. A portion of a WordNet-like concept taxonomy for protein-related terms is given in Figure 2.

4.3 Enumerating lexical chains

Using our concept taxonomy, we constructed lexical chains for a sample corpus of protein-interaction articles using a manual version of our algorithm. The main difference between the formal algorithm and our manual analysis is that we counted protein interactions which could be easily recognized by a human reader but that might be beyond the capability of an automated system relying on a template-based method for recognizing interactions. The articles we analyzed were all concerned with protein interactions in yeast and covered a range of the experimental techniques used to detect protein interactions. In order to keep our sample corpus size manageable yet still obtain a good number of protein interactions, we focussed on articles that were specifically about finding novel interactions, rather than just detailed studies of specific interactions.

For each article, we recorded the number of protein interactions in each of the following categories, based on the number and nature of their lexical chains (biological terms in the examples below are shown in boldface):

1. Bare mention of a protein interaction with no additional biologically related terms.

[1] C has value 8 and k has value 1. (Graeme Hirst, personal communication) In our examples, we set the weight of a strong relation to be 7 (i.e., assuming a path length of one and no changes of direction). However, we set an extra-strong relation to have a weight of 10, to reflect the special status of repetition.

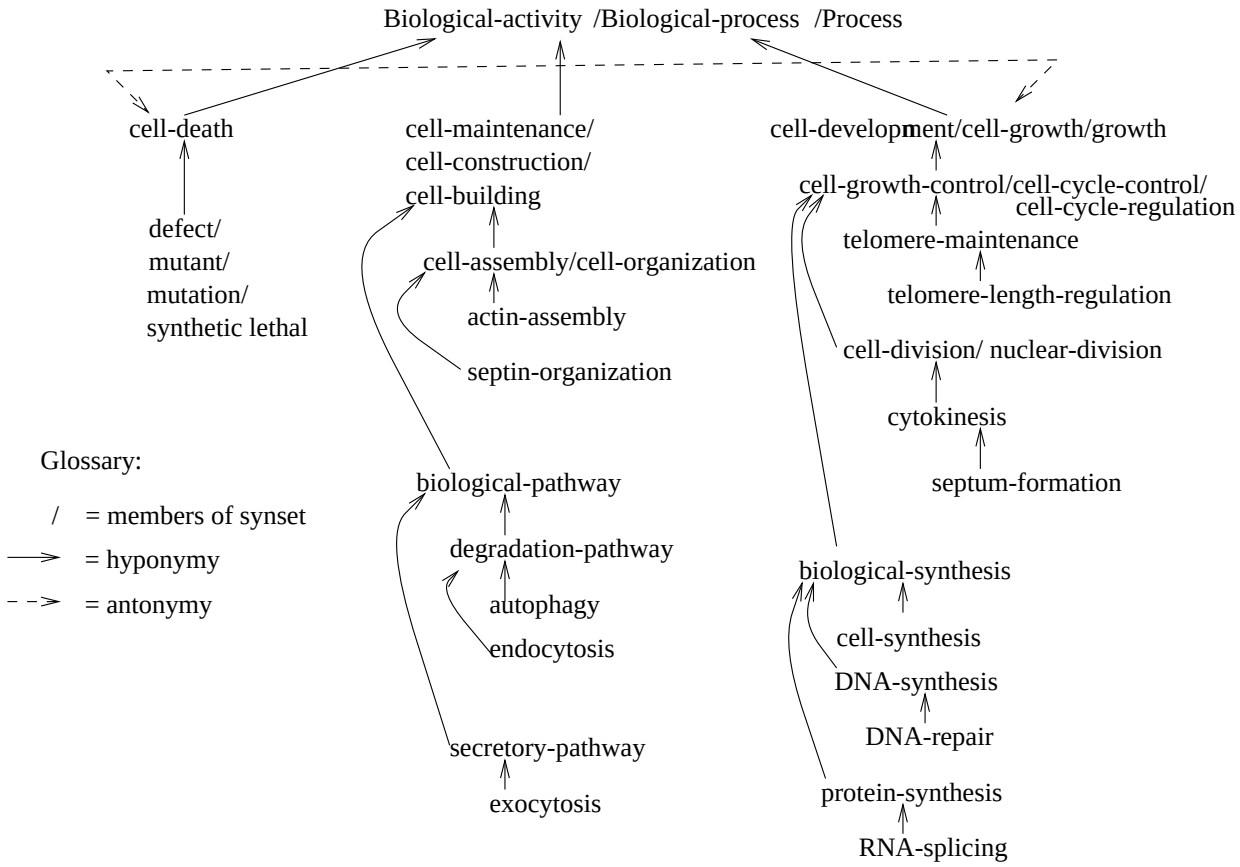


Fig. 2. A portion of a WordNet-like concept taxonomy for protein-related terms

(4) For example, two proteins of unknown function, YGR010W and YLR328W (77% identical), were observed to interact with each other. (Uetz *et al*. 2000, p. 625)

2. Single-term occurrence of a biological term in a protein-interaction context.
   (5) A two-hybrid interaction between Cla4 and Msb2 suggests that Msb2 is part of the Cdc42 regulatory **pathway**. (Drees *et al*. 2001, p. 558)
3. Single-theme lexical chain in which all the terms are semantically related to one another.
   (6) We found that Msb2 interacts with Bni4, a protein that targets chitin deposition to sites of polarized **growth** by linking chitin synthase to septins (DeMarini et al., 1997). Msb2 might coordinate cell wall **growth** with other Cdc42-regulated **processes**. (Drees *et al*. 2001, pp. 558–559)
4. Multiple-theme lexical chains in which each chain forms a distinct string of semantically related words.[2] (in the example below, the different-themed lexical chains are shown in bold and italic.)
   (7) In the present work we show that the N-terminal region comprising amino acids 1–252 of Cdc13p interacts with Pol1p, Sir4p, Zds2p and Imp4p. Moreover, *CDC13*-deleted yeast cells expressing Cdc13p lacking the N-terminal 1–252 amino acids region or Cdc13p with point **mutations** in this region caused **defects** in progressive *cell growth* and in *cell cycle control*. These cells also have **defects** in *telomere length regulation*. Thus, we conclude that the N-terminal region of Cdc13p is involved in *telomere maintenance*, *telomere length regulation* and *cell growth control* through its interaction with its binding proteins. (Hsu *et al*. 2004, p. 512)

[2] We accepted examples in this category in which one chain was a 'null chain' (i.e., a single term) as long as there was at least one other 'real chain' of two or more terms on a distinctly different theme.

In addition, we recorded examples of protein interactions that were hedged (i.e., the authors expressed uncertainty about the validity of the interaction), negative (i.e., of the form *protein A does not interact with protein B*), and too difficult for the lay reader to analyze. The results of our enumeration are shown in Table 1[3]. As may be observed in these results, there was a wide range of distribution across the articles in the types of protein interactions they contained. The bulk of the protein-interaction instances were 'bare mentions', i.e., simply stated, as might be expected from the reporting style of most of these articles. Many of the articles did however include explanations about the nature of the protein interaction, and it is these descriptive passages which were picked up as single-theme and multiple-theme lexical chains.

4.4 Sample rankings of lexical chains

We propose that one way to assess the biological validity of a protein-protein interaction mentioned in a scientific article is to use the strength of the lexical chains in the surrounding context as a measure of the quality of the interaction. Hirst and St.-Onge's scoring algorithm gives us one such measure. We applied this algorithm to the sample passages in (4) through (7) to arrive at the following results[4] (summarized in Figure 3):

Both passages (4) and (5) contain no lexical chains so receive scores of zero. This does not necessarily mean that the protein interactions they describe are not valid; rather, it indicates that the supporting evidence for the quality of the interaction is weak (at least insofar as this fragment of the article is concerned). Example (6) contains a strong singly-themed lexical chain, as evidenced by its score (17), and this serves to indicate a correspondingly strong biological quality of the interaction. The last passage, example (7), contains two strong lexical chains, with a consequent very high overall score (50)[5] indicating that the protein interactions described herein have strongly supportive biological evidence.

5 DISCUSSION AND FUTURE WORK

We have outlined a method for biomedical information extraction that makes use of the lexical-chaining structure in scientific articles to determine strings of biologically related words in protein-interaction contexts, and hence the biological significance of these interactions. Our immediate task is to improve the means of ranking protein interactions based on their related lexical chains. The current ranking scheme includes only two factors: the number of lexical chains related to an interaction in a given context in a single article and the strength of each individual lexical chain. If related lexical chains for the same protein-protein interaction could be detected across a corpus of articles, the evidence for the validity of that interaction would definitely be strengthened. More-specific protein-related terms in the concept taxonomy would also enhance the scoring of the biological significance of the lexical chains. Most importantly, a large-scale corpus study using an automated version of our algorithm will be needed to evaluate the effectiveness of our method.

[3] The articles are listed in the order in which they were read.

[4] The weights we assign to medium-strong relations are derived from Hirst and St.-Onge's formula. For example, in example (6), the medium-strong relation in the lexical chain (i.e., {growth, processes}) is computed as follows: According to our concept taxonomy in Figure 2, the path between *growth* and *process* is a single-link relation in the upward direction (hypernymy), therefore:
weight = (8 – path length – number of changes of direction) = (8 – 1 – 0) = 7

[5] In fact, the score for this example would be increased if we added in the relationship of antonymy between the two lexical chains, but for reasons of simplification, we have not done so.

ACKNOWLEDGEMENT

We would like to thank Prof. Alan Davidson of the Department of Biochemistry at the University of Toronto for valuable discussions and helpful pointers to relevant resources.

REFERENCES

Bader, G.D., Donaldson, I., Wolting, C., Ouellette, B.F., Pawson, T., and Hogue, C.W. (2001) BIND—The Biomolecular Interaction Network Database, *Nucleic Acids Res.*, **29**, 242–245.

Banerjee, S., and Pedersen, T. (2002) An adapted Lesk algorithm for word sense disambiguation using WordNet, *Proceedings, Fourth International Conference on Computational Linguistics and Intelligent Text Processing*.

Barzilay, R., and Elhadad, M. (1997) Using lexical chains for text summarization, *Proceedings of the Intelligent Scalable Text Summarization Workshop (ISTS-97)*, Madrid, Spain.

Blaschke, C., Andrade, M.A., Ouzounis, C., and Valencia, A. (1999) Automatic extraction of biological information from scientific text: Protein-protein interactions, *International Conference on Intelligent Systems for Molecular Biology*.

Cagney, G., Uetz, P., and Fields, S. (2001) Two-hybrid analysis of the *Saccharomyces cerevisiae* 26S proteasome, *Physiological Genomics*, **7**, 27–34.

Drees, B.L., Sundin, D., Brazeau, E., Caviston, J.P., Chen, G-C., Guo, W., Kozminski, K.G., Lau, M.W., Moskow, J.J., Tong, A., et al. (2001) A protein interaction map for cell polarity development, *Journal of Cellular Biology*, **154**(3), 549–571.

Fellbaum, C., (editor) (1998) *WordNet: An electronic lexical database*, The MIT Press.

Flores, A., Briand, J-F., Gadal, O., Andrau, J-C., Rubbi, L., Van Mullem, V., Boschiero, C., Goussot, M., Marck, C., Carles, C., et al. (1999) A protein-protein interaction map of yeast RNA polymerase III, *Proc. Natl. Acad. Sci. USA*, **96**, 7815–7820.

Fromont-Racine, M., Mayes, A.E., Brunet-Simon, A., Rain, J-C., Colley, A., Dix, I., Decourty, L., Joly, N., Ricard, F., Beggs, J.D., and Legrain, P. (2000) Genome-wide protein

Article	Bare mention	Single term	Single-theme lexical chain	Multiple-theme lexical chain	Hedge	Negative mention	Too difficult to analyze	Total
Drees *et al*.	25	18	14	25	(8)	–	–	82
Uetz *et al*.	5	3	4	2	(1)	(1)	–	14
Lu *et al*.	1	1	3	–	(1)	–	(1)	5
Cagney *et al*.	12	1	1	–	(1)	–	(1)	14
Fromont-Racine *et al*.	5	2	2	2	–	(1)	(2)	11
Hsu *et al*.	9	11	5	4	–	(14)	(2)	29
Printen and Sprague	20	2	3	3	(5)	(5)	(2)	28
Ito *et al*.	1	–	2	2	–	–	–	5
Schwikowski *et al*.	4	1	1	–	–	–	–	6
Ho *et al*.	11	11	7	2	–	(1)	–	31
Zheng *et al*.	16	5	3	3	(5)	(5)	(5)	27
Flores *et al*.	7	4	–	–	(4)	(4)	(5)	11
Tong *et al*.	2	4	4	6	(1)	(2)	–	16
Ursic *et al*.	19	9	4	4	(3)	(6)	(6)	36
Yatherajam *et al*.	17	1	1	–	(2)	(1)	(2)	19
Totals:	154	73	54	53	(31)	(40)	(26)	334

Table 1. Enumeration of various types of lexical chains in sample corpus of protein-interaction articles (Numbers in parentheses are not included in totals)

interaction screens reveal functional networks involving Sm-like proteins, *Yeast*, **17**, 95–110.

Halliday, M.A.K., and Hasan, R. (1976) *Cohesion in English*, Longman.

Hirst, G., and St.-Onge, D. (1997) Lexical chains as representation of context for the detection and correction of malapropisms, In: Fellbaum *et al*. 1998, 305–332.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S-L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., et al. (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry, *Nature*, **415**, 180–183.

Hsu, C-L., Chen, Y-S., Tsai, S-Y., Tu, P-J., Wang, M-J., and Lin, J-J. (2004) Interaction of *Saccharomyces* Cdc13p with Pol1p, Imp4p, Sir4p, and Zds2p is involved in telomere replication, telomere maintenance and cell growth control, *Nucleic Acids Research*, **32**(2), 511–521.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakari, Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome, *PNAS*, **98**(8), 4569–4574.

Jiang, J., and Conrath, D. (1997) Semantic similarity based on corpus statistics and lexical taxonomy, *Proceedings, International Conference on Research in Computational Linguistics*, Taiwan.

Lu, L., Arakaki, A.K., Lu, H., and Skolnick, J. (2003) Multimeric threading-based prediction of protein-protein interactions on a genomic scale: Application to the *Saccharomyces cerevisiae* proteome, *Genome Research*, **13**, 1146–1154.

Marcotte, E.M., Xenarios, I., and Eisenberg, D. (2001) Mining literature for protein-protein interactions, *Bioinformatics*, **17**(4), 359–363.

Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. (1990) Introduction to WordNet: An on-line lexical database, *International Journal of Lexicography*, **3**(4), 235–244.

Morris, J., and Hirst, G. (1991) Lexical cohesion, the thesaurus, and the structure of text, *Computational Linguistics*, **17**(1), 21–48.

Printen, J.A., and Sprague, G.F. Jr. (1994) Protein-protein interactions in the yeast pheromone response pathway: Ste5p interacts with all members of the MAP kinase cascade, *Genetics*, **138**, 609–619.

Pustejovsky, J., Castaño, J., Zhang, J., Cochran, B., and Kotecki, M. (2002) Robust relational parsing over biomedical literature: Extracting inhibit relations, *Pacific Symposium on Biocomputing*.

Schwikowski, B., Uetz, P., and Fields, S. (2000) A network of protein-protein interactions in yeast, *Nature Biotechnology*, **18**, 1257–1261.

Silber, H.G., and McCoy, K.F. (2002) Efficiently computed lexical chains as an intermediate representation for automatic text summarization, *Computational Linguistics*, **28**(4), 487–496.

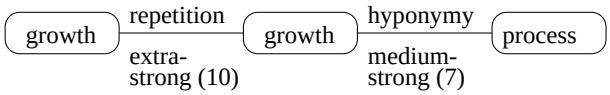
Thomas, J., Milward, D., Ouzounis, C., Pulman, S., and Carroll, C. (2000) Automatic extraction of protein interactions from scientific abstracts, *Pacific Symposium on Biocomputing*.

Tong, A.H.Y., Evangelista, M., Parsons, A.B., Xu, H., Bader, G.D., Page, N., Robinson, M., Raghibizadeh, S., Hogue,

(4) No lexical chains found.
Score = 0.

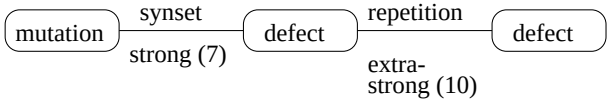
(5) No lexical chains found.
Score = 0.

(6)



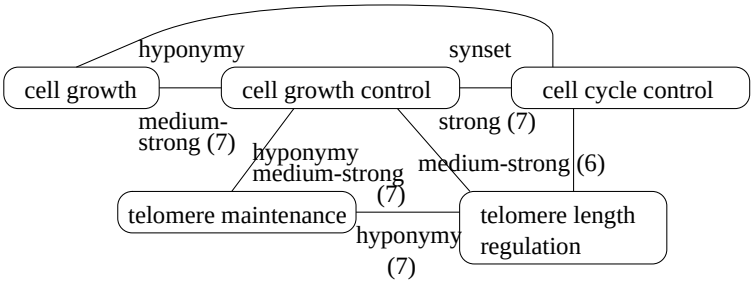
Lexical chain: {growth, growth, processes}
Score = 10 + 7 = 17

(7)



Lexical chain: {mutations, defects, defects}

Score = 7 + 10 = 17



Lexical chain: {cell growth, cell cycle control, telomere length regulation,
telomere maintenance, telomere length regulation, cell growth control}

Score = 7 + 6 + 7 + 7 + 6 = 33 Total score = 17 + 33 = 50

Fig. 3. Sample rankings of lexical chains for protein interactions in examples (4) to (7)

C.W.V., Bussey, H., et al. (2001) Systematic genetic analysis with ordered arrays of yeast deletion mutants, *Science*, **294**, 2364–2368.

Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., et al. (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*, *Nature*, **403**, 623–631.

Ursic, D., Chinchilla, K., Finkel, J.S., and Culbertson, M.R. (2004) Multiple protein/protein and protein/RNA interactions suggest roles for yeast DNA/RNA helicase Sen1p in transcription, transcription-coupled DNA repair and RNA processing, *Nucleic Acids Research*, **32**(8), 2441–2452.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G. (2002) Comparative assessment of large-scale data sets of protein-protein interactions, *Nature*, **417**, 399–403.

Yatherajam, G., Zhang, L., Kraemer, S.M., and Stargell, L.A. (2003) Protein-protein interaction map for yeast TFIID, *Nucleic Acids Research*, **31**(4), 1252–1260.

Zheng, Y., Bender, A., and Cerione, R.A. (1995) Interactions among proteins involved in bud-site selection and bud-site assembly in *Saccharomycese cerevisiae*, *Journal of Biological Chemistry*, **270**(2), 626–630.