FILLING PREPOSITION-BASED TEMPLATES TO CAPTURE INFORMATION FROM MEDICAL ABSTRACTS

G. LEROY, H. CHEN
Department of Management Information Systems, University of Arizona, 1030 E. Helen St, Tucson, AZ 85721, USA

Due to the recent explosion of information in the biomedical field, it is hard for a single researcher to review the complex network involving genes, proteins, and interactions. We are currently building GeneScene, a toolkit that will assist researchers in reviewing existing literature, and report on the first phase in our development effort: extracting the relevant information from medical abstracts. We are developing a medical parser that extracts information, fills basic prepositional-based templates, and combines the templates to capture the underlying sentence logic. We tested our parser on 50 unseen abstracts and found that it extracted 246 templates with a precision of 70%. In comparison with many other techniques, more information was extracted without sacrificing precision. Future improvement in precision will be achieved by correcting three categories of errors.

1 Introduction

The explosion of information in the biomedical field provides researcher with great opportunities to study cell growth, differentiation and death, and the associated regulating processes. The biochemical pathways seem to be interconnected and consequently form a complex network involving numerous genes and proteins. The enormous amount of information available on individual pathways and their potential connections makes it hard for a single researcher to investigate and formulate relationships, especially in a new or unfamiliar domain. We believe researchers would benefit from a toolkit to assist them in summarizing and reviewing the existing literature.

We are currently building such a toolkit for the biomedical field, called GeneScene. GeneScene will derive information from the relevant journals and assist in reviewing existing literature, identifying gaps in existing knowledge, and as such help lead the way to new and interesting hypotheses and field research. The complete toolkit will contain four components: 1. the extracted, stored, and integrated gene pathway analysis data from abstracts from several journals, 2. a visualization component that will allow researchers to browse and search for information, get an overview of the collection, retrieve particular abstracts, and modify the representational map, 3. personalization and collaboration options for the researchers, and 4. the possibility to map microarray data onto the literature-based data. GeneScene will be developed in three consecutive phases (see Figure 1). Initially, information will be extracted from individual sentences and put into preposition-based templates. Then, the sentence-based information will be

combined with information from existing knowledge sources, allowing additional checking. At this point, meta-information such as the publication date will also be extracted. Finally, all information will be made available to researchers in a software toolkit allowing revision, modification, and information sharing.

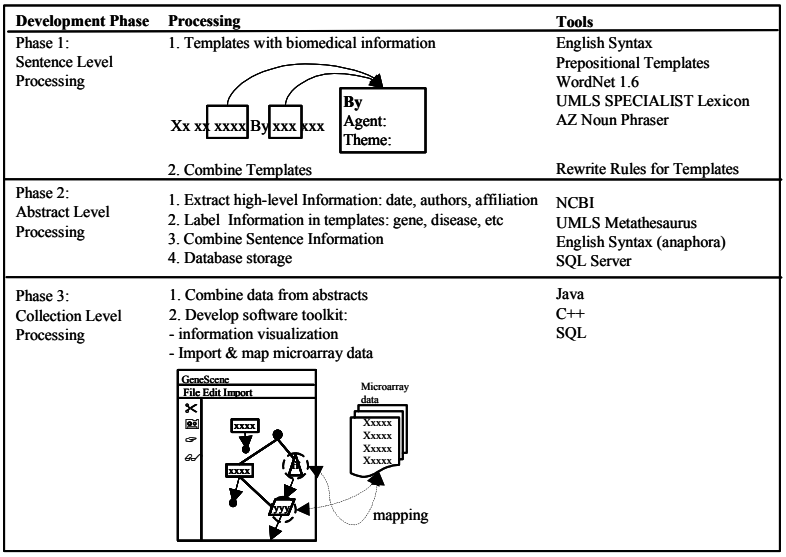

Figure 1: GeneScene Development Overview.

This paper discusses our approach to and initial results for the first phase of the project: extracting the relevant information from individual sentences in medical abstracts. Careful review of the literature and our own strengths and weaknesses led us to a new approach to this problem: a preposition-based medical parser. Our approach is new since we do not focus on pre-specified genes and interactions; additionally, we do not try to parse the complete sentence structure. Instead, we use basic templates as building blocks. These templates are based on English closed word classes, such as prepositions and conjunctions. We use rewrite rules to combine the basic templates and rewrite them into more complex patterns that reflect the underlying sentence logic, which is necessary to correctly represent the information. In the following sections we describe previous research, followed by our own approach and evaluation, and a discussion of future work.

2 Background

The approaches currently described in the literature range from general-purpose parsers to pre-specified extraction of particular information. The general-purpose parsers are based on sound linguistic principles and aim to detect the complete structure of a sentence. The complexity of the medical language used instigates parsing errors and problems with overall processing speed. Yakushiji et al.[1] built a full parser and increased its speed with two preprocessors to reduce the workload of the full parser. The first preprocessor recognizes noun chunks; the second reduces parts-of-speech ambiguity. The authors discovered that medical abstracts use more complicated sentence structures than the ones their parser was based on. They reported that 53% of their test bed's structures was not extracted. Park et al.[2] used a slightly more specific approach with a bi-directional incremental parser based on a combinatory categorical grammar. With this grammar, verbs are expected to be surrounded by a particular sentence structure. For example, "inhibits" expects a noun phrase to its left and to its right. The authors focused their parser on a few verbs of interest. Their approach resulted in high precision (80%) and somewhat lower recall (48%) of protein-protein relations. We believe that a perfect medical parser would be invaluable; however, it would still need an additional logic module since, as Rindflesh et al.[3] point out, a linguistic analysis does not provide a semantic interpretation.

Several approaches focus on extracting specific gene, protein and interaction information from abstracts. Sekimizu et al.[4] collected the most frequently used verbs in their collection of abstracts. They used partial and shallow parsing techniques to extract noun phrases from sentences and developed rules to find the subject and object of the high-frequency verbs. They estimated their precision at 73%. Thomas et al.[5] used a statistical parser to fill templates with information on proteins and their interactions. They concentrated on three verbs (interact with, associate with, bind to) for which they developed templates. They calculated recall and precision in four different manners for three samples of abstracts. Recall ranged from 24% to 63%, and precision from 60% to 81%. BioNLP[6] uses three components, two of which are BioKleisli[7] to query multiple medical databases and BioJAKE[8] to visualize and manipulate metabolic pathways. The third component is of interest here; it extracts gene names and their relations from free text based on an existing thesaurus, together with additional rules to identify existing and new gene names. The relations are limited to a predefined set of verbs. Once the genes are found, the sentences are matched against predefined syntactic structures and the verb thesaurus to identify the nature of the relation between the genes. Unfortunately, there was no evaluation data and the authors indicated that their pattern matching was not sophisticated enough to handle all sentences. The rules

used by BioNLP are based on work by Fukuda et al.,[9] who achieved very high precision extracting proteins (95% to 98%).

Other specific approaches extract information about a subset of genes and interactions. The PIES project,[10] requires users to submit key terms, such as "calyculin," and searches Medline for abstracts containing these terms. From the matching abstracts, "inhibit" and "activate" interactions are considered. The authors use BioNLP to extract the relevant information from the sentences, and the Graphviz software package (available online at http://www.research.att.com/) to visually display the results. An interesting addition to their system is that users can save and update the retrieved information. Unfortunately, no evaluation was provided. Blaschke et al.[11] used a comparable approach and asked users to provide the protein names to retrieve abstracts. They focused on the sentences containing the protein names and one of 14 pre-defined words representing actions. No systematic evaluation was reported. Stephens et al.[12] started from thesauri containing gene names and possible relations. They represented documents as vectors with a dimension equal to the size of the thesaurus and calculated the association between the genes based on the similarity of the vectors. When related genes were found, they retrieved the verb in that sentence. If it was found in their relation thesaurus, they accepted it as the relation between the two genes. The information is represented in a representational graph where distance represents similarity.

3 GeneScene

3.1 Selecting Abstracts and Sentences

GeneScene will ultimately integrate gene pathway information from thousands of abstracts. We will not require researchers to pre-specify genes or interactions. Instead, to extract the relevant information with sufficiently high precision, we plan to filter at three levels: the journal, the abstract, and the sentence level. Filtering at the "journal level" will be straightforward: we will initially concentrate on journals with a high impact factor, as defined by the Institute for Scientific Information (ISI, http://www.isinet.com/isi/index.html), that are also indicated as top journals by the biomedical researchers advising us in this project. The journal impact factor measures the frequency with which the "average article" in a journal has been cited in a particular year. It indicates a journal's relative importance in the field. At the "abstract level" we plan to focus on general abstracts. For abstracts describing clinical studies we plan to extract information only from the conclusion and

discussion sections. Finally, at the "sentence level," we will evaluate individual sentences based on WordNet information, to ensure that actual information and not e.g. the hypothesis is extracted. WordNet is a general English ontology (http://www.cogsci.princeton.edu/~wn/). We are currently building a WordNet-based thesaurus of catch phrases that will help us classify sentences. Sentences containing phrases such as "we show," "we demonstrate," "we established," "we hypothesized," "we expect" can be mapped to WordNet and its verb hierarchies. For example, "hypothesize" and "speculate" are both more specific ways (hypernyms) of "expect" and as such belong to the "expect" hierarchy. We will identify hierarchies containing phrases that indicate sentences to be included and other hierarchies that indicate sentences to be excluded. The classification system's main contribution will be to exclude sentences discussing expectations and hypotheses instead of results from GeneScene.

3.2 Preposition-based Parsing

There are two major phases our parser works through when processing a sentence. During the first phase, the extraction phase, the basic templates are identified. Prepositions form the entry point in a sentence. We then retrieve the main verb, adverbs, negation, and noun phrases around the preposition to fill the templates. Classification of words into word classes is currently based on WordNet 1.6. However, we noticed that the WordNet vocabulary will be insufficient to process a large collection of medical abstracts. Terms that necessarily need to be recognized as verbs, adjectives or adverbs, are not always found in WordNet 1.6. Fortunately, they are part of the SPECIALIST Lexicon, a component of the Unified Medical Language System (UMLS). For example, the verbs "phosphorylate," "overexpress," and "dysregulate," and the adjectives "oncogenic," "mitogenic," and "transcriptional" are not part of WordNet 1.6 but can be found in the SPECIALIST Lexicon. Lack of time prevented us from integrating the SPECIALIST Lexicon as a component in our parser for this evaluation. Instead, a small lexicon was added for the terms we discovered so far that are not found in WordNet.

We believe that by building templates around prepositions, we are able to capture more information than when looking for particular genes. We capture genes and proteins, but also e.g. diseases, cell phases, gene locations. In addition, we believe that precision will be high because, while we cover all possible sentence structures, we only extract the information that fits our templates. Although we intend to cover most prepositions, we report here on initial results of the templates developed for the two prepositions: "by" and "of." These were chosen because they frequently appear in medical abstracts and are representative of the complexity involved in processing medical abstracts. Additionally, our choice of prepositions allows us to demonstrate the second phase of parsing, the recombination phase,

where we rewrite the basic templates into combined templates that capture the underlying logic of the abstract. In the following, we discuss both phases in detail.

During the extraction phase, we focus on filling basic templates with phrases surrounding the preposition. We first retrieve the main verb close to the preposition. Then, we search for noun phrases to the left and right of the verb and preposition. Noun phrase detection is currently based on a variant of stop word phrasing: punctuation, auxiliaries, verbs, and closed class words are used as indicators of the start and end of phrases. For example, in the sentence "Remarkably, despite the inhibition **of** cell proliferation and apoptosis, the degeneration **of** lens fibers and aberrant expression **of** filensin were only …" we can extract three templates (described later) surrounding "of." For example, for the second template the closest boundaries are a comma on the left and a conjunction on the right. We also use a stop word list to cleanse the strings. For example, auxiliaries should not be part of an agent or theme. We keep track of begin and end indices of the template in the sentence. This information will be necessary to take overlapping arguments into account when combining templates. We employ additional selectional restriction to limit the phrases that can be agents or themes. A determiner, adjective, adverb, closed class word, a number, or a phrase containing a percentage cannot be the agent or theme. For example, in the sentence "…, JNK activity was increased by 150%," the "150%" is not the agent of the activity. It is restricted from this function.

In the following, we provide an overview of the templates currently being tested. A first template is built around the preposition "by." For this template we capture two main sentence structures (Structure 1 and 2) used to fill the by-template:

Structure 1: String1 – [modifier | negation] – main verb – **by** – String2
Structure 2: String1 – [modifier | negation] – nominalized verb – **by** – String2
Rule: agent = String2, action = verb, theme = String1
By-template: agent – [modifier | negation] action - theme

Both structures are used to fill the template as follows: *action* is the main verb or the underlying verb form of the nominalized verb and can be modified by a negation or a modifier. String2 is the *agent* of the action and String1 is the *theme* of the action. Auxiliaries can appear in the structure but they will not be part of the final template. If no verb is found, then only an agent is searched for; otherwise, both agent and theme are searched for. A modifier can be an adjective, an adverb, or a verb in the past tense. For example, the sentence "Apoptosis induced **by** the p53 tumor suppressor can attenuate cancer growth in preclinical animal models," results in the following template: (p53 tumor suppressor – induce – apoptosis).

A second template is built around the preposition "of." We capture two similar structures as with the by-template. However, with a nominalized verb, an agent is not searched for and "null" is inserted instead. The theme is found after the preposition in the sentence. For example, the sentence "This effect was accompanied by an increased expression **of** the cyclin-dependent kinase inhibitor p21(WAF1/CIP1) and a decreased expression **of** cyclin A," results in the following of-templates: (null – [increased] express – cyclin-dependent kinase inhibitor p21(WAF1/CIP1)) and (null – [decreased] express – cyclin A). The nulls in templates are important for the rewrite rules of the second recombination phase. Negation is also captured, for example the sentence "However, E2F is not a general regulator **of** oxidative phosphorylation genes since …," results in the following template: (E2F – [not][general] regulate – oxidative phosphorylation genes).

We do not only capture genes and proteins, but all information. For example the sentence "This arrest response appeared independent **of** p53/p21cip1/waf-1 function," results in the following template: (arrest response – [independent] appear – p53/p21cip1/waf-1 function). Other approaches miss this information. Labeling the content of the templates, e.g. "gene" or "bacteria," will follow in a later phase by mapping to data from the UMLS and the National Center for Biotechnology Information (NCBI) database.

During the recombination phase, templates are combined and rewritten. A first set of rewrite rules looks at specific prepositional combinations. In the following, we describe the individual templates that need to be extracted from a sentence, as described above, and the resulting combined template. We use the "*" notation to indicate a pointer to another template.

Prepositional Combination 1:

Of-template:	null – [modifier| negation] action1 – theme1
By-template:	agent2 – null – null
Rule:	no other by- or of-template can be found in between
Combined:	agent2 – [modifier| negation] action1 – theme1

For example, "Inactivation **of** the pRb proteins in mouse brain epithelium **by** the T121 oncogene induces aberrant proliferation and …," resulted in the following combined template (T121 oncogene – inactivate – pRb proteins).

Prepositional Combination 2:

Of-template:	null – [modifier| negation] action1 – theme1
By-template:	agent2 – [modifier| negation] action2 – theme2
Rule:	theme1 = theme2
Combined:	agent2 – action2 – *of-template

For example, the sentence "… suggests the existence **of** cell type-specific inhibitory pathways induced **by** these signals," results in the combined template (signals – induce – (NULL – exist – cell type-specific inhibitory pathways))

Prepositional Combination 3:

Of-template1:	null – [modifier| negation] action1 – theme1
Of-template2:	null – [modifier| negation] action2 – theme2
Rule:	action2 = verb form of theme1
Combined:	null – [modifier| negation] action1 – *of-template2

An example of this third combination is the following: "…distribution through the modulation **of** the expression **of** cell cycle-related genes …" which results in the template (null – modulate – (null – express – cell cycle-related genes)).

Prepositional Combination 4:

By-template:	agent1 - action1 - null
Of-template:	null – [modifier] action2 - theme2
Rule:	[modifier] + verb form of agent 1 = [modifier] action2
Combined:	*of-template – action1 – null

An example of this combination is the sentence "…that are activated **by** severe depletion **of** cell energy stores." The by-template (severe depletion - activate – null) and the of-template (null – [severe] deplete – cell energy stores) are combined into ((null – [severe] deplete – cell energy stores) – activate – null).

A second set of rewrite rules focuses on conjunctions. Two non-overlapping templates based on the same preposition and connected by "and" are combined. The missing element in the second template (following the "and") is copied from the first template. Currently, we only test for missing themes.

Conjunctional Combination:

X-template1:	agent1 – [modifier| negation] action1 – theme1
X-template2:	agent2 – [modifier| negation] action2 – null
Rule:	conjunction "and," no overlap between templates, prepositions in both templates have to be identical
X-template1:	agent1 – [modifier| negation] action1 – theme1
X-template2:	agent2 – [modifier| negation] action2 – theme1

For example, from the sentence "Given that E2F1 activity is stimulated **by** p300/CBP acetylase and repressed **by** an RB-associated deacetylase, we …," the following templates are extracted: (p300/CBP acetylase – stimulate – E2F1 activity)

and (RB-associated deacetylase – repress – null). These are connected by "and," and the rewrite rule changes the second template to (RB-associated deacetylase – repress – E2F1 activity).

3.3 Evaluation

Following a tuning-phase, we used the keyword "E2F1" to retrieve 50 new abstracts. Both titles and the actual abstracts were processed, resulting in a total of 474 sentences and 246 templates. Table 1 provides an overview of the results. We only consider templates that contained at least two non-null elements. For example, when an agent name is captured, but no other information, the resulting template (e.g. pRb – null – null) is currently not considered for evaluation. A template was scored as correct when all noun phrases were complete, when no modifier or negation was missing, and when the template correctly represented that subpart of the sentence.

To calculate recall, we counted the instances where templates could have been built. For the of-template this meant all occurrences of the preposition except when it was used in expressions such as "some of which," numeric expressions such as "5 of 7," or noun phrases without action words such as "B-subunits of replicative DNA polymerases." For the by-template this meant all occurrences of the preposition except when it was used in expressions such as "by which," or as the first word in a sentence. For the combination templates, there were no exceptions. Precision, recall, and F-measure were calculated according to the following formulas:

Precision = total correct templates / total extracted templates

Recall = total correct templates / total possible templates

F-measure = (2 * recall * precision) / (recall + precision).

Table1: Performance Analysis

	Total	Average Per abstract	Average Precision (%)	Average Recall (%)	F-measure
General Analysis:					
Abstracts:	50	-	-	-	-
Sentences:	474	9.5	-	-	-
Templates built:	246	4.9	70	47	56
Template Specific Analysis:					
Of-Templates built:	189	3.8	74	52	61
By-Templates built:	58	1.2	72	43	54
Combo-Templates built:	22	0.5	45	38	42

The average precision was 70% for all templates combined. It was slightly higher for of-templates (74%) and by-templates (72%) separately. Since combined templates can only be correct if the two underlying templates are correct, this

precision is lower (45%). Recall was 47% in general, 52% for of-templates, and 43% for by-templates. As with precision, recall of combined templates depends on the other two templates being recalled and, as such, was lower (38%). If we had taken a less general approach and concentrated on only those relations that contain the term "E2F1," then we would have extracted a maximum of 110 templates. Many approaches take an even more specific approach and require not one, but two genes to be present in a sentence. In that case, fewer relations would have been extracted. Although it is possible to test all possible combinations of known genes, our approach does not depend on any pre-specified name list. Additionally, we also extract information elements that are not genes or proteins.

In Table2, we provide an overview of the distribution of errors that shows there are major categories of errors that can be systematically addressed.

Table 2: General Error Analysis

Error Type:	Fraction (%)
Template not yet developed:	24
Agent/Theme overextension:	28
Modifier incomplete:	9
Agent/Theme incomplete:	4
Agent/Theme contains rubbish terms:	15
Error in Combinations:	4
Error due to WordNet limitation:	1
Other:	14

A closer look reveals that almost 70% of the errors belong to just three categories. The first category accounts for 24% of the errors. These were incorrect because combinational templates not yet designed were not incorporated, resulting in a misrepresentation of the information. For example, the sentence "… for the induction **of** the p21 promoter **by** activated Ras, …" resulted in the templates (NULL – induct – p21 promoter) and (activated Ras – promote – p21). Since the "activated Ras" does not promote "p21" but the "induction of the p21 promoter," this is a missed "of-by" combination resulting in an erroneous second template. These errors will be corrected with additional combination rules. Although it is a challenge to add more template combinations without introducing new errors, correcting this category of errors would increase precision significantly.

The errors due to overextension of the agent and theme phrases form a second main error category, representing 28% of the total errors. In almost all cases, these errors were due to a word not being recognized as a conjugated verb. For example, in the sentence "We show that the E2Fs control the expression of several genes that are involved in cell proliferation," the word "control" was not recognized as the conjugated verb, resulting in an erroneous agent "E2Fs control." To address this second category of errors, we will try and implement proven noun phrasing techniques based on our experience with the Arizona Noun Phraser.[13]

A final major error category contains the agents or themes with rubbish terms. For example, from the sentence "Increased expression of neutrophins (e.g. NGF, BDNF) and …," the "(e.g." became part of the theme. We expect improvements by processing more abstracts since that will make our stop word list, which is used to filter and cleanse this irrelevant information from the templates, more complete.

We want to remark on our decision to convert nominalized verbs to their base verb form. This was done to increase the compilation powers of GeneScene when we combine all information. In some cases, the transformation of nominalized verbs to their base verb form might seem unsuitable. However, by transforming e.g. "the expression of CDK4" and "CDK4 is expressed" to the same form "null – express – CDK4" the relation is strengthened. This will provide researchers with important clues since a frequently found relation often indicates consistent findings. A very rarely found relation can be an erroneous finding stated by an author, an error in the processing of the abstracts, or a very interesting and rare finding. Furthermore, this process will allow us to represent more information visually in the same manner, making the overall picture less demanding to understand. For example, name labels in "green" ink to indicate "expression," or a colored arrow from the agent to the theme indicating that the agent is responsible for the expression of the theme.

4 Conclusion

We feel that our approach has a lot of potential for different reasons. First of all, we achieved an average precision of 70% without focusing on a subset of the available information. We expect to improve this precision by correcting the main error categories discussed earlier. Most approaches to automated extraction of biomedical information report precision between 60% and 80%,[2,3,5] depending on the different definitions of precision used and also on the diversity of the extracted information. It can be expected that systems focusing on a very specific subset of the information will be more precise than general system. However, we do not focus on certain types of information. The agent and themes do not have to be proteins or genes; the action does not need to belong to a pre-specified set of interaction verbs. We also use a liberal definition of modifiers for verbs, allowing us to capture details about the relation. Furthermore, by focusing on the prepositions and their particular combinations, we are able to capture the underlying sentence logic. The combination of e.g. a by-template followed by an of-template is different from an of-template followed by a by-template. Finally, we want to note that the development of our parser is a continuing effort. We expect to improve its precision, and to process larger sets of abstracts in the near future.

Acknowledgments

We would like to thank Ann Lally and Jesse Martinez for their inspiring comments and suggestions. We feel grateful to the National Library of Medicine and Princeton University for making the UMLS and WordNet available to researchers. This research was sponsored in part by the following grants:

NSF Digital Library Initiative-2, "High-performance Digital Library Systems: From Information Retrieval to Knowledge Management," April 1999-March 2002.

National Institute of Health and National Library of Medicine, "UMLS Enhanced Dynamic Agents to Manage Medical Knowledge," February 2001-February 2004.

References

1 A. Yakushiji, Y. Tateisi, Y. Miyao, and J. Tsujii, *Pacific Symposium on Biocomputing*, 408 (2001)
2 J.C. Park, H.S. Kim, and J.J. Kim, *Pacific Symposium on Biocomputing*, 369 (2001)
3 T.C. Rindflesch, L. Hunter, and A.R. Aronson, *Proc AMIA Symp***:** 127 (1999)
4 T. Sekimisu, H.S. Park, and J.i. Tsujii, *Genome Informatics***:** 62 (1998)
5 J. Thomas, D. Milward, C. Ouzounis, S. Pulman, and M. Carroll, *Pacific Symposium on Biocomputing* **5:** 538 (2000)
6 S.-K. Ng, and M. Wong, *Genome Informatics* **10:** 104 (1999)
7 S.B. Davidson, C. Overton, V. Tannen, and L. Wong, *International Journal on Digital Libraries***:** (1996)
8 W. Salamonsen, K.Y.C. Mok, and P. Kolatkar, *International Journal of Digital Libraries* **1**(1)**:** 36 (1997)
9 K. Fukuda, T. Tsunoda, A. Tamura, and T. Takagi, *Pacific Symposium on Biocomputing*, 705 (1998)
10 L. Wong, PIES, *Pacific Symposium on Biocomputing*, 520 (2001)
11 C. Blaschke, M.A. Andrade, C. Ouzounis, and A. Valencia, *ISMB***:** 60 (1999)
12 M. Stephens, M. Palakal, S. Mukhopadhyay, R. Raje, and J. Mostafa, *Pacific Symposium on Biocomputing*, 483 (2001)
13 K.M. Tolle, and H. Chen, *Journal of the American Society of Information Systems* **51**(4)**:** 352 (2000)