The Frame-Based Module of the SUISEKI Information Extraction System

Christian Blaschke and Alfonso Valencia, *Protein Design Group*

SUISEKI, an information extraction system, uses morphological, syntactical, and contextual information to detect gene and protein names and interactions in scientific texts. This article describes the system's rules (called frames) used to detect and analyze interaction networks described in the molecular biology literature.

Scientific discovery hinges on the ability to build on existing knowledge. The fields of molecular biology and biomedicine, however, have recorded substantial amounts of knowledge in a free-text form that is difficult to use directly as a source of information. To facilitate retrieval and analysis of the huge amounts of data contained in texts on experimental approaches to genomics and proteomics, researchers are now developing dedicated information extraction systems.

The first applications of information extraction techniques, although still very experimental, address such problems as the extraction of protein interactions, the functions common to protein families and gene groups, the interaction between proteins and chemical compounds, and the description of cellular compartments. More computationally oriented approaches that include text parsing, part-of-speech tagging, disambiguation, or grammars[1–3] work well on limited text corpora but have not been proven equally effective on a text corpora of biologically interesting size.

The simple co-occurrence of two gene or protein names in the same abstract already indicates a relation between them. Because of its simplicity, we can apply this approach to large amounts of text and establish interaction networks for *Saccharomyces cerevisiae* and humans.[4,5] The limiting factor is that we can't know the type of relation between the genes and proteins this way. To address this problem, we developed SUISEKI, an information extraction system that takes an intermediate view of the problem by requiring the two names to be in a frame that indicates a direct or indirect interaction between them.

The SUISEKI system

On the one hand, SUISEKI uses statistics and frequency of occurrence, while on the other, it uses analysis of the syntactical structure of phrases and other developments in computational linguistics. This simplified view of the possible complications in the context of text analysis finds its justification in the field of natural language understanding.

Both the grammar and pattern-matching approaches offer advantages. Generally, the less syntax used, the more domain-specific the system will be. This permits constructing a robust system relatively quickly, but also risks losing many subtleties in the sentence's interpretation. In some applications, however, the domain-dependent pattern-matching approach might be the only way to attain reasonable performance in the foreseeable future.[6]

Figure 1 shows how SUISEKI works. The system core defines the frames that capture the various language constructions used to express protein interactions. We developed them manually by filtering large amounts of text to find the most frequent constructions that implicate two protein names and express a direct or indirect interaction. Our first version of the system used only one frame and a fixed list of protein names; we discussed the possibilities of this approach and initial results for interaction networks in *Drosophila*.[7] The current version implements the protein name detection module and the frames, takes into account negations, and introduces probability scores for the frames.

Natural language has almost unrestricted potential for expressing the same fact in different ways, but in practice it often uses a limited number of constructions, which facilitates detection using a set of patterns. In particular, the most obvious pattern ("protein A binds/interacts/ ... (with) protein B") covers many of the interactions.

1094-7167/02/$17.00 © 2002 IEEE

The system requires minimal user intervention and can easily be applied to large text collections. Let's examine in detail the frames used for the system, their matching frequency in a large text corpus, and the accuracy with which each pattern detects interactions.

Methods for detecting interactions

SUISEKI simultaneously detects protein names and signal words that indicate interactions in predefined frames. Here we describe the frames-based module; we describe the other modules elsewhere.[8]

The frame-based module

The analysis splits the text into sentences that the system then treats separately. It analyzes sentences that contain two protein (or gene) names and one of the interaction keywords by comparing them with a predefined list of frames. No phrase separators (“,” or “;”) are allowed in the string that matches a given frame.

We have assigned a probability score to each frame depending on its reliability, and use these values to score the interactions. The results depend not only on the quality of the matched frames but also on repetition of the same interactions within different text fragments. The interaction scores permit viewing and manipulating the interactions at different reliability levels.

The frames also account for negations and the distance between names and action keywords. Negations have an associated score of zero to prevent them from contributing to the establishment of associations between the corresponding names. Increased distance between particles (names or keywords) decreases a frame's score. More specific frames have higher scores than less specific ones.

SUISEKI calculates the interaction's final score as the sum of scores for all the frames it matches. Protein pairs that frequently match high-scoring patterns tend to represent true interactions.

Interaction types

Interactions between proteins can be quite indirect, such as when a protein influences another's expression or when two interact with a third protein (such as in a complex). For example, the sentence, “The expressed p53 protein showed nuclear localization and its expression was associated with an induction of p21 and bax expression,” relates p53 with p21 and bax but does not imply a physical interaction between them. Some action keywords such as “activate” tend to be more strongly related to such indirect interaction.

SUISEKI's current implementation does not explicitly address the different qualities of interactions and treats all extracted relations in the same way. But when the results are visualized, the user can select the interaction keywords (such as “bind” and “phosphorylate”) to analyze different types of interactions.

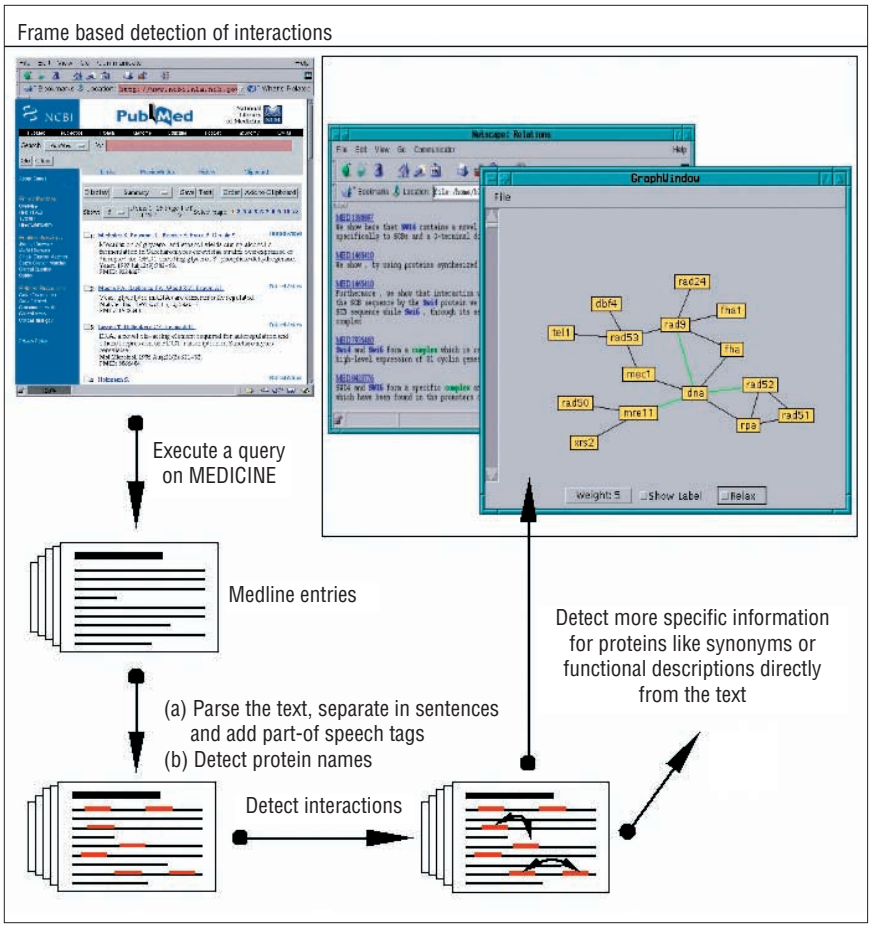

Figure 1. Overview of SUISEKI and its components. The user first performs a Medline query to extract entries that are subsequently analyzed (1). A part-of-speech tagger then separates the text into sentences and parses them, using this information to detect protein names (2). The system uses frames to detect the interactions described within the sentences (3), and stores the results in a protein-protein interaction database that permits analysis using an interactive query interface (4). It then applies additional modules to extract more specific information on proteins, such as possible synonyms and functional descriptions (5).

The text corpus

We have two text corpora of different sizes. The first one consists of Medline abstracts that contain *“Saccharomyces cerevisiae”* in the MeSH terms. (MeSH terms are a curated list of medically relevant terms appended to the abstract that normally contain information about the organism used in the experiments.)

The second corpus, a restricted version, contains only the abstracts including the words “cell cycle.” The cell-cycle corpus contains 5,283 abstracts, whereas the *saccharomyces corpus* has 43,417 abstracts. We've conducted most of the detailed analysis with the cell-cycle corpus, whereas the larger corpus complements the analysis for the frames that appear infrequently in the smaller corpus.

Evaluating system accuracy

We consider interactions to be correctly detected if the names are sufficiently well identified and they form part of a true interaction, as assessed manually. The only interactions labeled as fully correct are those in which both names are completely correct and the interaction detected is a true interaction. This is obviously the most stringent possible evaluation of the interaction quality detected.

The evaluation is done in terms of

- "recall," calculated as the number of interactions detected compared to the number of interactions detected by manual inspection
- "precision," the proportion of real interactions among all interactions the system has identified

Analysis results

We detected a total of 6,778 interaction instances in the 5,283 cell-cycle abstracts, resulting in 4,657 distinct interactions. In the 43,417 *saccharomyces* abstracts, we detected 39,126 instances, corresponding to 25,988 distinct interactions. We first analyzed the interaction extraction results for each frame individually, then analyzed the precision of interactions according to their score and system recall.

Individual frame precision

For each frame, we analyzed a maximum of 100 sentences to calculate the precision (see Table 1). We considered only sentences where both names were detected correctly because we wanted to separate name detection problems from the different frames' efficiency. Brackets indicate the syntactical class, for example, [proteins] or [verbs]; and parentheses indicate how many words can appear in that position, for example, (0-5) for zero to five words. The probability score (0 = lowest) expresses the likelihood that a sentence matched by a frame represents a true interaction. "NA" means a field does not apply or that precision was not calculated.

The most frequent pattern and the first implemented in the system,[7] protein-verb-protein accounts for almost 96 percent of the matches. The efficiency of the protein-verb-protein pattern decreases with the distance between the names, but even when the protein names are far away, the sentence can express a relation between them. For example, the sentence, "The CPC2 gene of the budding yeast *Saccharomyces cerevisiae* encodes a G beta-like WD protein, which is involved in regulating the activity of the general control activator Gcn4p," matches the frame [CPC2]...[regulating]...[Gcn4p] and states an interaction between CPC2 and Gcn4p.

Negation patterns also occur frequently, which supports their inclusion in the system for reducing the number of false-positive identifications. Other patterns of the form verb/noun-protein-protein, and those describing protein complexes, account for about 3 percent and 1 percent of the matches, respectively. Patterns such as verb-of-protein-to-protein occurred more frequently than verb-of-protein-by-protein.

Patterns such as verb/noun-protein-protein offer improved precision due to stringent rules that allow few intervening words. The verbs-of-protein-by-proteins frame provides an exception; this construction refers in many cases to a different event than that identified by the patterns. For example, the sentence, "Most Cdks require binding of a cyclin and phosphorylation by a Cdk-activating kinase (CAK) to be active," incorrectly matches the rule [binding]...[cyclin]...[Cdk-activating kinase], when the real meaning is that Cdks bind to cyclins and are phosphorylated by Cdk-activating kinases.

Even the most accurate frames can mistakenly interpret interactions because of the text's structure. In a typical case, "HDA1 and HDA3 are components of a yeast histone deacetylase (HDA) complex" matches the frame "Name (other name for A) complex" and misleads the system by inferring an interaction between histone deacetylase and HDA for the formation of a complex. A more detailed treatment of the parentheses could have prevented this type of error.

Relation between precision and the interaction score

From the 5,283 abstracts of the cell-cycle corpus, the system extracted 4,657 different interactions described in 1,471 abstracts. How efficiently did the system detect interactions in the full text corpus?

To analyze the quality of these interactions, we first classified the interactions into four groups of scores and then manually analyzed 100 abstracts from each group. Table 2 shows that the extracted interactions' accuracy correlates with their scores, which helps us associate error estimators to the various scores.

In this case, we did not evaluate interactions at the level of individual sentences but instead checked whether the interaction was clearly stated in at least one of the sentences extracted for an interaction. Therefore, if a sentence read, "These results demonstrate that Sst2 and Gpa1 interact physically and suggest that Sst2 is a direct negative regulator of Gpa1" (it matched the frame "Gpa1 interact ... Sst2") where none of the frames matched correctly, we still considered the result to be correct because the system extracted a biologically correct fact and provided a correct sentence. Even if linguistically not correct, this result will satisfy a user interested in the biological facts.

SUISEKI recall and information repetition

SUISEKI's frames cover only part of the many possible ways to express gene or protein interactions. Fortunately, information is repeated, and if a frame does not match in one sentence, it can still match elsewhere and extract the fact of interest. We therefore analyzed how many individual fragments referring to interactions SUISEKI can detect and how many interactions it can detect in a given text corpus when the only requirement is to detect at least one fragment referring to the interaction.

SUISEKI recall

We randomly chose 100 abstracts from the cell-cycle corpus and compared the results to all individual text fragments referring to interactions. Table 3 shows that the system correctly detected about 40 percent of the individual interaction instances with a precision of about 45 percent.

SUISEKI recall with repetition

We expect that most interactions not detected in a given sentence will be detected elsewhere in the same or a different abstract. For example, the interaction between PTIP and Pax2 appears repeatedly in the following text:

> PTIP, a novel BRCT domain-containing protein, interacts with Pax2 and is associated with active chromatin. In this report, we describe the isolation and characterization of a novel gene and its encoded protein, PTIP, which binds to the activation domain of Pax2 and other Pax proteins?PTIP binds to Pax2 in vitro, in the yeast two-hybrid assay and in tissue culture cells. The binding of PTIP to Pax2 is inhibited by the octapeptide repression domain.

The first two sentences' structure doesn't permit the system to detect that PTIP and Pax2 interact, but two sentences later in the same abstract state the interaction in terms that match the system's patterns. Therefore, even if the system does not detect the first description of the interaction, the subsequent instances will contribute to its correct detection.

To evaluate this assumption more thoroughly, we analyzed three samples of 100 abstracts chosen randomly from the cell-cycle corpus. For these samples we evaluated how many interactions given in the text the system detected in the corresponding abstracts (at least once in an abstract if an interaction is stated in more than one sentence) and how many it detected when the system had the

Table 1. Frame representation and accuracy for 100 randomly selected cases.

Frame	Probability score	Number of hits in cell-cycle corpus	Number of hits in saccharomyces corpus	Precision, saccharomyces corpus (percentage)
Type I				
[syntactical class = proteins] (0-5 words) [verbs] (0-5) [proteins]	4	2628	13667	68
[proteins] (0-5) [verbs] (6-10) [proteins]	3	969	5380	50
[proteins] (6-10) [verbs] (0-5) [proteins]	3	892	5090	54
[proteins] (0-10) [verbs] (0-10) [proteins]	2	278	1672	33
[proteins] (*) [verbs] (*) [proteins]	1	1632	11080	21
protein verbs protein	NA	6399	36889	NA
[proteins] (*) [verbs] (0-3) but not (0-3) [proteins]	0	26	64	NA
[proteins] (*) cannot (0-3) [verbs] (*) [proteins]	0	7	24	NA
[proteins] (*) does not (0-3) [verbs] (*) [proteins]	0	38	235	NA
[proteins] (*) did not (0-3) [verbs] (*) [proteins]	0	34	218	NA
[proteins] (*) was not (0-3) [verbs] (*) [proteins]	0	12	77	NA
[proteins] (*) not (0-3) [verbs] (*) by (*) [proteins]	0	6	101	NA
[proteins] (*) not required for (0-3) [verbs] (*) [proteins]	0	4	10	NA
[proteins] (*) failed to (0-3) [verbs] (*) [proteins]	0	2	67	NA
Negations	NA	129	796	NA
Type II				
[verbs] of (0-3) [proteins] (0-3) by (0-3) [proteins]	5	1	17	40 (*)
[verbs] of (0-3) [proteins] (0-3) to (0-3) [proteins]	5	29	294	97
[nouns] of (0-3) [proteins] (0-3) by (0-3) [proteins]	5	93	400	91
[nouns] of (0-3) [proteins] (0-3) with (0-3) [proteins]	5	66	386	95
[nouns] between (0-3) [proteins] (0-3) and (0-3) [proteins]	5	83	437	94
Verb/noun protein protein	NA	242	1223	NA
Type III				
[proteins] (0-2) [proteins] (0-2) complex	5	43	239	68
Complex containing (0-3) [proteins] (0-2) and (0-2) [proteins]	5	7	21	100
Complexes containing (0-3) [proteins] (0-2) and (0-2) [proteins]	5	1	7	100
Complex formed between (0-3) [proteins] (0-2) and (0-2) [proteins]	5	0	1	- (*)
Complex of (0-3) [proteins] (0-2) and (0-2) [proteins]	5	3	31	100
Complexes of (0-3) [proteins] (0-2) and (0-2) [proteins]	5	1	20	89 (*)
Formation of a complex between (0-3) [proteins] (0-2) and (0-2) [proteins]	5	0	1	- (*)
Formation of complexes between (0-3) [proteins] (0-2) and (0-2) [proteins]	5	0	1	- (*)
[proteins] (0-2) form a complex with (0-2) [proteins]	5	5	13	100 (*)
[proteins] (0-2) [proteins] (0-2) complexes	5	11	67	55 (*)
[proteins] (0-2) [proteins] (0-2) dimer	5	0	7	- (*)
[proteins] (0-2) [proteins] (0-2) heterodimer	5	2	16	64 (*)
[proteins] (0-2) [proteins] (0-2) homodimer	5	0	3	- (*)
Complexes	NA	73	430	NA

(*) fewer than 10 sentences were available for analysis

entire cell-cycle corpus available (5,283 abstracts). Table 4 presents the results.

Table 4 lists the total number of interactions detected in sample 1 (see Table 3), the number of unique instances, the number of these instances detected in the samples, and the interactions detected by extending the search to the whole cell-cycle corpus.

The recall for different interactions in the sample compares to individual interactions and is relatively low. Extending the search to a larger text corpus greatly enhances the recall, to around 70 percent. Thus, the more text available, the higher the recall of interactions expressed at different positions.

Table 2. Relation between scores of the extracted interactions and their precision.

	Correctly identified interactions (percentage)	Mean score of the interactions
First quarter	80	8.5
Second quarter	69	4.0
Third quarter	63	3.2
Fourth quarter	42	1.5

Table 3. Detection of interactions among 100 abstracts from the cell-cycle corpus (*).

	Number of interactions	Recall (percent)	Precision (percent)
Identified manually	297	NA	NA
Identified by SUISEKI	263	NA	NA
Correctly identified	118	39.7 (118 of 297)	44.9 (118 of 263)

(*) All occurrences are counted individually even if they refer to the same proteins.

Main error sources

Incorrect detection can result from erroneous detection of protein names, incorrect parsing of sentences separated by commas, indirect references from previous sentences, and current frame set limitations.

Name detection

Several problems occur in this step when gene or protein name detection errors are translated directly in incorrect interaction identification. Incorrect detection can result from insufficient distinction between words that form part of a protein name because they are English words also used outside the domain of molecular biology. For example, SUISEKI incorrectly removes "alpha" from some protein names and does not remove other particles such as "multisubunit" or "promotor" because they do not appear in the standard dictionaries. Also, SUISEKI can confuse abbreviations, substance names, and experimental techniques with protein names because formulating heuristics for these cases proves difficult, and a dictionary does not cover them adequately.

Protein names being intrinsically complex, we must apply different heuristics to catch all of them in the text. In detecting checkpoint protein 1, for example, difficulties arise in separating out from the surrounding text names that are constructed from normally used English words. Also, protein names can be part of other names; for example, Cdc7 and Cdc7 protein kinase are two different proteins. And non-protein names can form part of protein names; for example, RNA is not a protein name but RNA polymerase II is.

Semantically, names can refer to protein classes, not just individual proteins. For example, Fus3p and Kss1p are MAP kinases, CLN1, CLN2, and CLB5 are all G1/S cyclins. To model this, we'd have to apply an ontology of protein names.

For more discussion of protein name detection, please refer to the sidebar, "Detecting protein interactions in text."

Difficult sentences and indirect referents

Our current parsing techniques separate sentences following commas, making detection of the relationship between corresponding objects impossible. Examples of difficult sentences include

- "HCS26 does not associate with CDC28, but instead associates with PHO85, a closely related protein kinase."
- "PTIP, a novel BRCT domain-containing protein interacts with Pax2 and is associated with active chromatin."

Indirect referents between sentences also prove difficult, as in the following:

- "The CLN1, CLN2 and CLN3 family of cyclin homologues is required for cells to pass START. They probably act by activating the CDC28 protein kinase."
- "These results indicate that cyclophilin A and Ess1 function in parallel pathways and act on common targets by a mechanism that requires prolyl isomerization. Using genetic and biochemical approaches, we found that one of these targets is the Sin3-Rpd3 histone deacetylase complex, and that cyclophilin A increases and Ess1 decreases disruption of gene silencing by this complex."

Limitations of the current frames

In the following example, a relationship between two proteins goes undetected because the action keyword ("phosphorylation") is a noun and SUISEKI did not contain a suitable protein-noun-protein frame:

> LCD1 is also required for efficient DNA damage-induced phosphorylation of Rad9p and for the association of Rad9p with the FHA2 domain of Rad53p after DNA damage.

The relationship between two other proteins also went undetected because the distance between words exceeds that allowed by the corresponding frame (association [1-word] Rad9p with [4-words] Rad53p).

Detecting protein interactions: SUISEKI performance

To test SUISEKI's accuracy, we first ana-

Table 4. SUISEKI recall when repeated information is taken into account.

	Sample 1		Sample 2		Sample 3	
Identified interactions	Number of interactions	Recall (percent)	Number of interactions	Recall (percent)	Number of interactions	Recall (percent)
Manually: total	297	NA	NA	NA	NA	NA
Manually: unique	154	NA	118	NA	115	NA
By SUISEKI in the sample	58	37.6	59	50.0	65	56.5
By SUISEKI in the cell-cycle corpus	111	72.1	86	72.9	79	68.7

Detecting protein interactions in text

Several publications address the problem of detecting protein and other molecular interactions from the literature, but many problems common to the field remain unresolved. Our own analysis shows that the main problem remains protein name detection.[1] As far as we know, no published systematic approach directly addresses the problem of detecting protein names. (The EDGAR system addresses it partially by using the UMLS metathesaurus to detect gene and cell names,2 although the problem itself has been discussed in several specific publications.[3,4])

Most research has analyzed the detection of protein interactions using rather over-optimistic scenarios, such as selected text pieces and small sets of hand-tagged text, without addressing the name problem.[5,6–8] Numerous systems use the simple approach of counting co-occurrence of names within the same text (abstract),[9,10] in most cases with explicit lists of protein names. For example, a recent analysis[10] of genetic networks based on the simple co-occurrence of names in a large publication corpus on human genes used a predefined set of names.

Comparing different methods in this field is difficult because they employ different assumptions about what an error is, how the protein names are treated, and how the text is selected for the evaluation. This situation can only change by calling for competitions like in other fields (information extraction systems are compared at the Message Understanding Conferences (MUC) or protein structure predictions are compared at the Critical Assessments of Structure Prediction (CASP)). This could be done with a standardized text or by the use of experimental data that has to be recovered from the text (see also our earlier discussion of this problem).[1]

The results of application of the full system are available at www.pdg.cnb.uam.es/suiseki for such biological systems as cell cycle, DNA replication, cytoskeleton, or nuclear proteins. How well the integrated approach performs depends directly on name and keyword detection module results and on how efficiently the system describes the interactions in the frames. We discuss this in a separate publication.[11]

References

1. C. Blaschke and A. Valencia, "Can Bibliographic Pointers for Known Biological Data Be Found Automatically? Protein Interactions as a Case Study," *Comparative and Functional Genomics*, vol. 2, no. 4, 2001, pp. 196–206.
2. T.C. Rindflesch et al., "EDGAR: Extraction of Drugs, Genes and Relations from the Biomedical Literature," *Proc. Pacific Symp. Biocomputing*, 2000, pp. 515–524.
3. K. Fukuda et al., "Information Extraction: Identifying Protein Names from Biological Papers," *Proc. Pacific Symp. Biocomputing*, World Scientific Publishing, River Edge, N.J., 1998, pp. 707–718.
4. D. Proux et al., "Detecting Gene Symbols and Names in Biological Texts: A First Step toward Pertinent Information Extraction," *Proc. 9th Workshop Genome Informatics*, Universal Academy Press, Tokyo, 1998, pp. 72–80.
5. J. Thomas et al., "Automatic Extraction of Protein Interactions from Scientific Abstracts," *Proc. Pacific Symp. Biocomputing*, World Scientific Publishing, River Edge, N.J., 2000, pp. 384–395.
6. J.C. Park, H.S. Kim, and J.J. Kim, "Bidirectional Incremental Parsing for Automatic Pathway Identification with Combinatory Categorial Grammar," *Proc. Pacific Symp. Biocomputing*, World Scientific Publishing, River Edge, N.J., 2001, pp. 396–407.
7. D. Proux, F. Rechenmann, and L. Julliard, "A Pragmatic Information Extraction Strategy for gathering Data on Genetic Interactions," *Proc. Int'l Conf. Intelligent Systems for Molecular Biology*, AAAI Press, Menlo Park, Calif., 2000, pp. 279–285.
8. T. Sekimizu, H.S. Park, and J. Tsujii, "Identifying the Interaction between Genes and Gene Products Based on Frequently Seen Verbs in Medline Abstracts." *Proc. 9th Workshop Genome Informatics*, Universal Academy Press, Tokyo, 1998, pp. 62–71.
9. B.J. Stapley and G. Benoit, "Biobibliometrics: Information Retrieval and Visualization from Co-occurrences of Gene Names in Medline Abstracts," *Proc. Pacific Symp. Biocomputing*, World Scientific Publishing, River Edge, N.J., 2000, pp. 529–540.
10. T.K. Jenssen et al., "A Literature Network of Human Genes for High-Throughput Analysis of Gene Expression," *Nature Genetics*, vol. 28, May 2001, pp. 21–28.
11. C. Blaschke and A. Valencia, "The SUISEKI Information Extraction System," in preparation.

lyzed each frame individually on 100 sentences (or fewer, if the frame did not match a minimum of 100 times) and calculated the precision. Then we evaluated the entire system's precision by analyzing 4×100 interactions with different scores that the system extracted. Because these evaluations reveal nothing about the system's recall (how much information it missed in the text), we used 100 abstracts to assess how many fragments expressing interactions the system can detect correctly and 3 ¥ 100 abstracts to estimate how many real interactions (independent of how many fragments they appear in) it can recover from a bigger text corpus.

To further biological knowledge, we would ideally evaluate system performance using databases of well-characterized interactions as a reference. Direct exploration of one of the first protein interaction databases (the DIP database9) reveals that we must still address key issues related to protein name identification and information sources before we can consider the database a valid reference.[10] The challenge therefore remains to obtain large reference sets for evaluating automatic information extraction systems.

Individual frame accuracy

In the current implementation we have defined a list of 31 frames, including eight that define negations. We have evaluated different frames' detection accuracy for those cases in which both names were correctly detected. Some frames are clearly very constrained and more accurate than others; for example, the frame protein-5-verb-5-protein (see Table 1) has almost 70 percent accuracy and is matched by 13,667 sentences (35 percent of all the possible hits).

The most predominant frame type has the general form of "protein A ... interacts/binds/ ... ... protein B" and offers, on average, 48 percent accuracy. Interestingly, its accuracy decreases with the distance between the words, but even with 20 words between the names we find a significant number of positive matches. We could probably improve these results by including more information in the frames. Preliminary unpublished results (Blaschke, unpublished) suggest that we could use the part-of-speech tagger to more precisely identify the relationship between the actions and the corresponding protein names.

The Authors

Christian Blaschke is a postdoctoral fellow in the Protein Design Group at the Centro Nacional de Biotecnologia, Universidad Autonoma de Madrid. His research focuses on the extraction of biologically relevant information from scientific literature using statistical and linguistic methods and the application of machine learning techniques to reduce the cost of adapting these systems to new domains. He has an MSc in plant physiology from the University of Salzburg, Austria, and a PhD in molecular biology from the Universidad Autónoma de Madrid.

Alfonso Valencia is Group Leader of the Protein Design Group at the Centro Nacional de Biotecnologia, Universidad Autonoma de Madrid. His research interests include the use of the genomic and proteomic information for the study of molecular evolution and the development of new biotechnological resources; the development of bioinformatics and computational biology methods; and the analysis and comparison of genomes, prediction of protein structure and function, analysis of protein interactions, and extraction of information from scientific text. He earned an MSc in genetics from the Universidad Complutense, Madrid, and a PhD in biochemistry from the Universidad Autonoma, Madrid. He is also a senior scientist of the Spanish research council (CSIC), coordinator of the Spanish Network of Bioinformatics, member of the editorial board of Bioinformatics, and vice president of the International Association for Computational Biology (ISCB).

Evaluating detected interactions

Analyzing groups of detected interactions shows how those with higher scores (those that match more significant frames in more cases) clearly provide better information than interactions with lower scores. Sorting interactions by score offers a simple mechanism for representing confidence in them and an ideal parameter for representing the interaction system. Our current evaluation shows how the most frequent interactions can be detected with error rates of less than 20 percent.

System recall and repeated information

The frames' limitations and the lack of preparsing steps to resolve coordination and anaphora (information distributed over more than one sentence and referenced by pronouns such as "it" and "them") mean SUISEKI can correctly detect only about 40 percent of text fragments that indicate interactions. Improvements in these points will obviously elevate this baseline performance, although, as mentioned, the same or different abstracts generally repeat information several times. One of the highest scoring interactions in the cell-cycle corpus between Swi4 and Swi6 was detected 19 times in nine different abstracts (plus an unknown number of repetitions that the system did not detect). Studying 300 abstracts from the cell-cycle corpus showed us that the system could detect 70 percent of the interactions mentioned in these samples when it uses the entire corpus (about 5,000 abstracts).

Limits of the frame-based approach

Our analysis shows that a single frame type, protein-verb-protein, dominates the system's coverage and that more specific frames are more reliable but match with a much lower frequency. (We must relativize this because in absolute numbers the type II frame, verb/noun-protein-protein, hits more than 1,200 times in the *saccharomyces* corpus, and type III hits more than 400 times, affecting the results considerably). Obviously, we're missing frames for some constructions (such as protein-protein-noun in "the A and B interaction" or protein-noun-protein in "A is required for the phosphorylation of B") that we're currently evaluating to verify their reliability.

We must go beyond the problem of coordination and commas, however, to reduce sentence complexity in a parsing step before applying the current frames. For example, in "A, a novel XYZ protein, interacts with B," the correct detection and deletion of the intervening phrase would make the fact detectable by the current frames; the same would be true for the transformation of "A binds to B, C and D" into "A binds to B," "A binds to C," and "A binds to D." These extensions will likely improve system performance without making it too inflexible and overloaded to apply to large text collections. We rarely find anaphora in abstracts where the language used is more precise and to the point, but to extend the frame-based method to full publication texts, we might have to address this problem as well.

Obviously, we must accelerate the time-consuming manual collection and engineering of the rule set to make the system more flexible and extensible. We might achieve this with systems that can learn information extraction rules from domain-specific text and support the knowledge engineer in evaluating the proposed rules. Such techniques will let us extend SUISEKI to the detection of facts such as protein–drug interactions, relations between genes and diseases, and other useful information. ■

Acknowledgments

We acknowledge the discussions held with Biopath consortium members, particularly Dietrich Schuhman and Harald Kirsch. Juan C. Oliveros and R. Hoffmann have contributed to SUISEKI-related information extraction work. We are grateful for the continuous support from other members of the Protein Design Group. This work was supported in part by a Training and Mobility of Researchers (TMR) grant from the European Commission.

References

1. T.C. Rindflesch et al., "EDGAR: Extraction of Drugs, Genes and Relations from the Biomedical Literature," *Proc. Pacific Symp. Biocomputing,* 2000, pp. 515–524.
2. J. Thomas et al., "Automatic Extraction of Protein Interactions from Scientific Abstracts," *Proc. Pacific Symp. Biocomputing,* 2000, pp. 384–395.
3. A. Yakushiji et al., "Event Extraction from Biomedical Papers Using a Full Parser," *Proc. Pacific Symp. Biocomputing,* 2001, pp. 408–419.
4. B.J. Stapley and G. Benoit, "Biobibliometrics: Information Retrieval and Visualization from Co-occurrences of Gene Names in Medline Abstracts," *Proc. Pacific Symp. Biocomputing,* World Scientific Publishing, River Edge, N.J., 2000, pp. 529–540.
5. T.K. Jenssen et al., "A Literature Network of Human Genes for High-Throughput Analysis of Gene Expression," *Nature Genetics*, vol. 28, May 2001, pp. 21–28.
6. J. Allen, *Natural Language Understanding,* Benjamin/Cummings Publishing Co., Redwood City, Calif., 1995.
7. C. Blaschke et al., "Automatic Extraction of Biological Information from Scientific Text: Protein-Protein Interactions," *Proc. Int'l Conf. Intelligent Systems for Molecular Biology,* AAAI Press, Menlo Park, Calif., 1999, pp. 60–67.
8. C. Blaschke and A. Valencia, "The SUISEKI Information Extraction System," in preparation.
9. I. Xenarios et al., "DIP: The Database of Interacting Proteins," *Nucleic Acids Research*, vol. 28, no. 1, 2000, pp. 289–291.
10. C. Blaschke and A. Valencia, "Can Bibliographic Pointers for Known Biological Data Be Found Automatically? Protein Interactions as a Case Study," *Comparative and Functional Genomics,* vol. 2, no. 4, 2001, pp. 196–206.