The Potential Use of SUISEKI as a Protein Interaction Discovery Tool

Christian Blaschke `blaschke@cnb.uam.es`
Alfonso Valencia `valencia@cnb.uam.es`

Protein Design Group, CNB/CSIC, Campus Universidad Autónoma, 28049 Madrid, Spain

Abstract

Relevant information about protein interactions is stored in textual sources. This sources are commonly used not only as archives of what is already known but also as information for generating new knowledge, particularly to pose hypothesis about new possible interactions that can be inferred from the existing ones. This task is the more creative part of scientific work in experimental systems. We present a large-scale analysis for the prediction of new interactions based on the interaction network for the ones already known and detected automatically in the literature.

During the last few years it has became clear that part of the information about protein interactions could be extracted with automatic tools, even if these tools are still far from perfect and key problems such as detection of protein names are not completely solved. We have developed a integrated automatic approach, called SUISEKI (System for Information Extraction on Interactions), able to extract protein interactions from collections of Medline abstracts.

Previous experiments with the system have shown that it is able to extract almost 70% of the interactions present in relatively large text corpus, with an accuracy of approximately 80% (for the best defined interactions) that makes the system usable in real scenarios, both at the level of extraction of protein names and at the level of extracting interaction between them.

With the analysis of the interaction map of Saccharomyces cerevisiae we show that interactions published in the years 2000/2001 frequently correspond to proteins or genes that were already very close in the interaction network deduced from the literature published before these years and that they are often connected to the same proteins. That is, discoveries are commonly done among highly connected entities. Some biologically relevant examples illustrate how interactions described in the year 2000 could have been proposed as reasonable working hypothesis with the information previously available in the automatically extracted network of interactions.

Keywords: SUISEKI, information extraction, protein-protein interactions, frame-based systems, discovery of interactions

1 Introduction

Scientists in areas such molecular biology and biochemistry aim to discover new biological entities and their functions. Typical cases could be the discovery of the implication of new proteins and genes in an already known process, or the implication of proteins with previously characterised functions in a separated process. This is for example the case of the discovery of new signalling pathways linking previously known ones.

The use of the available information (published papers, talks in conferences, etc.) are a key step for the discovery process, since in many cases week or indirect evidences about possible relations hidden in the literature are used to substantiate working hypothesis that are experimentally explored.

A similar situation is found after the application of massive proteomic methods that enable the discovery of hundred of protein interactions, i.e. yeast two hybrid, mass spectrometry applied to the

resolution of protein complexes. In this case it is interesting to realise quickly the presence of new interactions between already known proteins that might constitute a relevant new discovery.

In practice both, the generation of new hypothesis about interactions and highlighting the new discoveries, require a painful step of analysis of the available information in databases and literature sources. The automatic systems for information extraction of interactions have the capacity of facilitating this task.

1.1 The SUISEKI System

We are developing a System for Information Extraction on Interactions (SUISEKI) able to automatically identify protein or gene names and their biological interactions. On the one hand our system takes advantage of the analysis of the syntactical structure of phrases and other developments in computational linguistics. And at the same time it counts on the statistics and number of occurrences of indications of interactions in the framework of what are generally known as pattern-matching approaches. The SUISEKI approach can be considered a hybrid between the purely statistical methods [15, 9, 1, 2, 4] and the more computer linguistically oriented approaches such as [13, 16, 18].

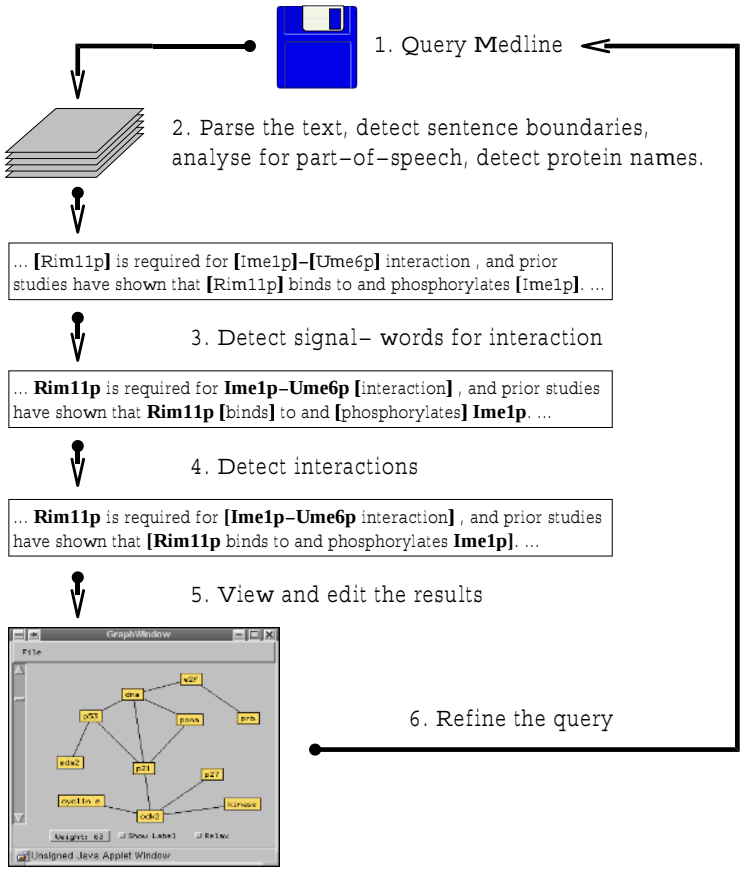

Figure 1: Overview of the SUISEKI system.

Different components of the SUISEKI system have been previously described in [3] and [5, 6]. The basic steps during the analysis of protein interactions are summarised in Fig. 1. In the first step the user has to perform a query on MEDLINE to extract the entries that form the text corpus (step 1 in the figure). Then the text is separated in sentences and parsed by a part-of-speech tagger. This information is used to detect the protein names in combination with a set of rules and dictionaries (2). Sentences containing at least two protein names are matched against a collection of predefined frames that contain the basic forms of expressing protein interactions in scientific text (3). The results are kept in a protein-protein interaction database that allows the analysis with an interactive query interface (4), the interface allows additional manipulations of the information by human experts. Additional modules are applied to extract more specific information on proteins like possible synonyms and functional descriptions (5).

With the practical limitations of this methodology in mind the possibility of obtaining massive amount of data has prompted us to ask some basic questions about the structure and organisation of the interaction networks, with the aim of using this information to discover new interactions. This allows us for the first time to go beyond the facts that are contained in the text and extracted by

an IE system and to predict totally new facts based on this information. These proposals could be helpful in the interpretation of proteomic data or being followed by detailed experimental approaches.

2 Methods

The core of the system is the definition of the rules that capture different language constructions that are commonly used to express interactions. These rules are implemented in the system as so-called *frames*. The frames with the highest recall are those of the form "[protein/gene] binds/associates/... [protein/gene]". They incorporate a distance measure (number of words) between particles (names and specific verbs indicating an interaction). Short distances (between zero to five words) represent more clearly relations and are associated to higher probability scores, in this case our precision estimation is of 68% (percentage of frames of this type with a distance of maximal five words between the particles that detect correctly an interaction). Frames with distances larger than 10 words between the protein/gene names and the action-keyword have associated a lower score and correspond to a precision of 21%. A number of other frames are created with different combinations of distances. These frames also detect negations to reduce the number of false positives.

Most of the remaining frames are dedicated to specific relations, such as "[noun indicating interaction] of [protein/gene] with [protein/gene]", these frames are very precise, for example the precision of this particular frame is better than 95%, even if they cover a relatively small number of cases.

The final result is a list of interactions each of them associated with a score (the sum of the probability scores according to the frames that matched) that is related with the reliability of this information.

The use of the frames relies heavily on the previous identification of protein and gene names. The intrinsically complex nature of this problem makes the application of different heuristics necessary. Names can be composed of English words of general use (e.g. *checkpoint protein 1*), they can be part of other names (e.g. *Cdc7* and *Cdc7 protein kinase* are two different proteins) and non-protein names may form part of protein names (e.g. *RNA* is not a protein name but *RNA polymerase II*). Additional complications are words like *alpha*, *multisubunit* or *promotor* that may form part of the names or may better be excluded. Finally, the commonly present abbreviations, substance names and experimental techniques require specific treatment to differentiate them from protein and gene names.

Furthermore in the class of protein or gene names there exist semantic problems because names can refer to protein families like for example Fus3p and Kss1p that are MAP kinases or CLN1, CLN2 and CLB5 that are all G1/S cyclins. In this case it is possible that in the future the use of annotated classifications of proteins and functions (ontologies) may provide viable solutions.

SUISEKI uses a combination of different indications for this task. First the syntactical information provided by a part-of-speech tagger to detect the noun-phrases that contain the names. Then these phrases are checked for lexical criteria (by the use of a slightly adapted English dictionary) and for morphological criteria (if it looks like a protein or gene name like for example p53 or Arf1 or words ending with *-ase*). Finally the terms are accepted as protein or gene names if they fit in the context (this step is still very limited and it checks if the terms refer to cell or phage names).

2.1 Text Corpora

The first text corpus analyzed consists of the 43417 Medline abstracts [19] that contain "Saccharomyces cerevisiae" in MeSH terms. The second corpus is a subset of 5283 that additionally contain the words "cell cycle". The detailed analysis of a biological scenario presented here is based on the cell-cycle corpus. The text corpora for the cell cycle and yeast have sizes of 12Mb and 95Mb approximately.

2.2 Application of SUISEKI

The analysis carried out with SUISEKI requires a considerable preparation time (access to the text corpus, indexing and parsing), while the statistical parts of the implementation are much faster. The most time consuming step is the part-of-speech tagging of the text what can take some hours for the approximately 40,000 abstracts of the entire yeast corpus. The following steps will take about 2-3 hours on a common workstation.

The interactions are represented as a graph in an interactive window. The interface allows adjustment of a threshold (score of the interactions) to restrict the displayed interactions to a higher or lower level of accuracy. The interface has basic search capacities implemented and a basic layout algorithm to "unfold" the graph avoiding the collapse of the entities in the graph. The viewer has been implemented in Java as a stand-alone piece of software with an applet that enables the access of the data via the internet. The graph can be edited, wrong interactions can be deleted, and nodes can be fused (necessary for synonyms). The data on the interactions (which proteins interact, the scores, types of interactions, links to the literature, etc.) are stored in a simple relational database schema that facilitates the access to the information. The results of the analysis are written to HTML-pages and linked to the graph-viewer. This allows the direct access to all the underlying information (the text or individual sentences) used by the system.

3 Results

3.1 Performance of the SUISEKI System

We have previously analysed large data sets to establish the performances of the system in real biological scenarios [5, 6]. The efficiency of SUISEKI in directly related with the frequency of the interactions, since the accuracy of the extracted interactions increases with their relative abundance. That is, the most commonly found interactions have higher chances of corresponding to true interactions. The system found a total of 4657 interactions in the 5283 abstracts comprising the cell cycle corpus.

As mentioned before each of the detected interactions gets a score that is related to the reliability of this information. Interactions in the first quarter of this list (ordered by the scores) have a reliability of 80% (20% false extractions) which decreases when we move down in the list (69%, 63% and 42% in the second, third and forth quarter). This allows the application of thresholds to adapt the precision/recall level.

The second important figure to consider here is the capacity of the system for retrieving a significant number of the interactions contained in the text. The recall (percentage of detected interactions in relation to the number of interactions detected by manual inspection) is low when small samples are considered (recall of 38%), but increases considerably when large text collections are analysed, since then the system benefits from the repetition of information in different parts of the text. The recall raised to 72% when the system had to recover 154 interactions selected from a sample in the cell cycle corpus. We would expect an increase in the system recall in larger text corpora (e.g. the full yeast corpus or the full text of the articles instead of the abstracts), since the yeast cell cycle represents a relatively small set of abstracts.

3.1.1 Sources of Incorrectly Detected Interactions

Our previous analysis and the recently published evaluation of the capacity of the SUISEKI system for retrieving the protein-protein interactions contained in DIP database [17] only from MEDLINE abstracts [7] has made us aware of the importance of the several areas in which the information extraction applications require improvement. They can be summarised in three main topics, treated in order of importance:

1. **Erroneous detection of protein names.** A systematic nomenclature for gene and protein names does not exist for most organisms. And in addition to many possible writing variants for the same name synonyms can be associated with the genes or proteins what makes the detection and classification of the terms very difficult. Our analysis of the DIP database [7] shows that 50% of the names given in the database were impossible to find in the corresponding abstracts.

2. **Insufficient information in the text sources.**

   (a) Medline abstracts, commonly used do not contain all the information about interactions. The above mentioned experiment [7] has shown that for 30% of the cases interactions that were described in the database were not found in the abstracts but in the full text of the publication. The analysis of the full text of the articles remains a challenge at the moment.

   (b) Indirect references and anaphoric expression. This is a well-known problem in computational linguistics (e.g. information contained in previous sentences is referred implicitly in the following text), and this is a key question for the analysis of Medline abstracts where protein names can be given in the title or initial sentences and later treated with forms such as "the protein" or mentioned as a general class of proteins like "the kinase".

3. **Incorrect detection of the interactions due to deficiencies in the information extraction technology.**

   (a) Incorrect parsing of sentences, mainly due to the limitations imposed by the parsers and their lack of adaptation to this particular domain.

   (b) Limitations of the current set of frames that SUISEKI uses for describing interactions. In the analysis of the DIP database mentioned above, we discovered that in 20% of the cases in which the information was contained in the abstracts and the names were correctly identified our system failed to detect the corresponding interactions because the structure of the sentences was too complicated for the available frames. This problem has to be seen in context, since even if some of the sentences are not captured by the frames the same information would be most likely repeated in other sentences that would conform to the standard structures expected by the frames.

3.2 The Interaction Network of the Cell Cycle

The analysis of the interaction network formed around the cell cycle corpus provides a good practical example of the type of information provided by SUISEKI and the possibilities open to the analysis by human experts. The overview provided in Fig. 2, shows how the main protein components implicated in this biological system were detected automatically. In the upper left corner we see Cdc28, a key cyclin dependent kinase (CDK) involved in cell cycle control. The activity of Cdc28 is controlled by the G1 cyclins (cln1, 2, 3) and the G2 cylins (clb1, 2, 3 and 4). The interaction between Cdc28 and the cyclins controls the activity of the transcription factors Sbf and Mbf.

The graph shows also the relation between the Rad proteins and DNA. They are generally implicated in DNA repair and recombination. The group of genes close to them (Rme1p, Ime1, etc.) are related with the developmental control of meiosis. Ste20 is a protein kinase required for pheromone signal transduction, pseudohyphal and invasive growth, as well as mating. It binds to and is activated by Cdc42, a small GTPase of the Rho family, which has a role in regulating actin organisation. Cdc24 is the guanine-nucleotide exchange factor for Cdc42 and together they are required for the establishment of cell polarity and for bud formation.

The overview provided in Fig. 3 hides some of the complexities of the interaction network to focus on Cdc28, the cyclins, and the major transcription factors. The Medline sentences matched by the SUISEKI frames are organized in web pages and linked to the graphical web interface. The information

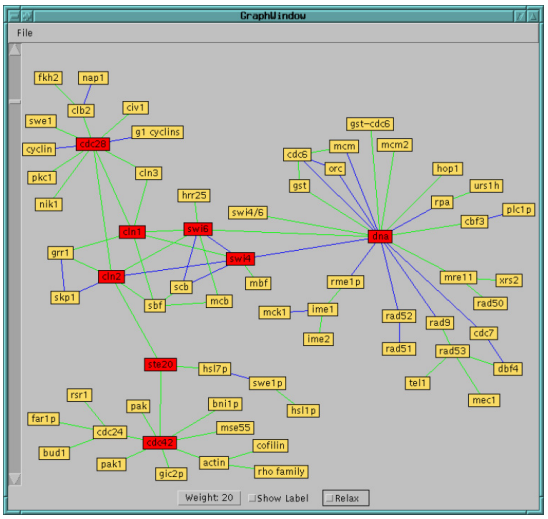

Figure 2: A part of the interaction network for the cell cycle corpus.

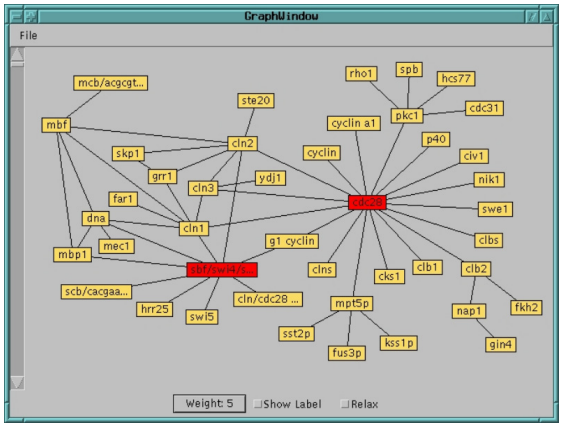

Figure 3: Focus on the cyclins that regulate the cell cycle, their effector the cyclin dependent kinase Cdc28 and the major transcription factors implicated in this process.

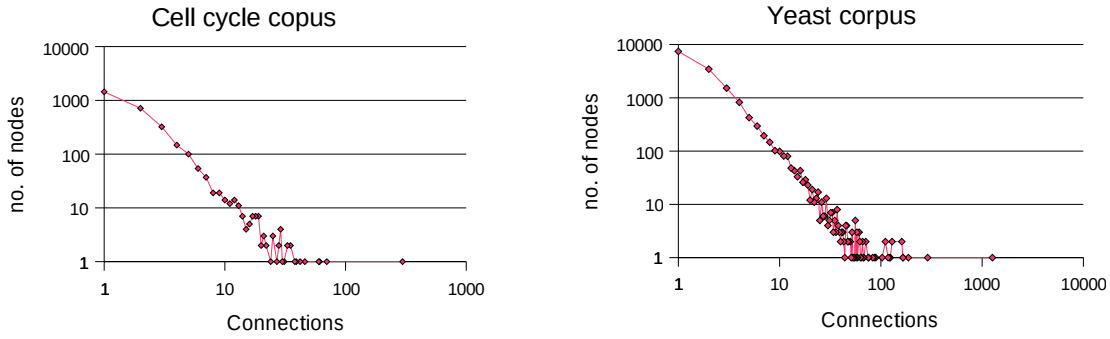

Figure 4: Distribution of the number of connections per node (entity) in the yeast and the cell cycle corpus (The values are given in logarithmic scale in both axis).

about interactions is separated from the information about protein functions that is also presented to the user. For example, we used this information to understand that Sbf is a dimer of Swi4 and Swi6, while Mbf is a dimer of Mbp1 and Swi6. A sentence where the interaction between Swi4 and Swi6 was detected clarifies that Swi4 is the DNA binding part, and that Swi6 binds to Swi4. From other sentences it became clear that Mbp1 also binds to the DNA, and that Swi6 is the regulatory part of these two transcription factors (Sbf and Mbf). It was also easy to detect that MCB and SCB are specific DNA sequences where these transcription factors bind.

3.3 The Connectivity in the Network of Interactions

For the two corpora the relation between the number of connections and nodes in the interaction graph (the proteins or genes) is represented in Fig. 4. In both cases most of the nodes are connected by less than 10 connections, while only a few of them are the centre of a remarkable high number of connections.

The most connected node is DNA, followed by key chemical compounds such as ATP, GTP (in the

general yeast corpora) and proteins such as actin and microtubules in both corpora (Table 1). This correlates well with their central activity in the functioning of the cell, and in the case of actin and microtubules to the role of the cytoeskeleton in the cell cycle.

In the current implementation entities such as metabolites (ATP, GTP, GDP) and general classes of proteins such as DNA and microtubules are included, because even if they are not protein or gene names they have a considerable nucleation power and help to understand some of the relations. This is also the case for some of the general protein names, e.g. cyclin, Cdk or DNA polymerase that cover a class of interactions for that protein families. Besides this nucleation centres it is clear in the figures that some particular proteins are very well represented in terms of connections in direct correlation with their role as key players in the corresponding biological systems.

Remarkable examples in the yeast cell cycle are for example Cdc28 that is an essential protein controlled both by G1 and G2 cyclins, and in interaction with them it controls the transcription factors SBF and MBF integrating cell-cycle signals and controlling the response at the level of transcription activation.

A second example could be Cdc42, a small GTPase of the Rho family (a member of the Ras superfamily). Cdc42 is controlled by its exchange factor, Cdc24 and itself controls the actin polymerisation during the bud formation. This interaction is key for the establishment of the cell polarity. Ste20 is a protein kinase required for pheromone signal transduction as part of the MAP kinase pathway where it is activated by Cdc42. At the same time Ste20 is controlled via phosphorylation by the cyclin Cln2, more precisely it is phosphorylated by the complex Cln2p-Cdc28 kinase (see graph). Therefore Cdc42 provides a link between the cell structure (cytoskeleton) and the regulation of transcription (via Ste20).

Table 1: Highly connected entities in the two text corpora.

Cell cycle corpus		Yeast corpus	
Number of connections	Entities	Number of connections	Entities
297	DNA	1264	DNA
70	actin	291	ATP
62	Cdc28	187	Gal4
59	cyclin	166	Tbp
46	Cln2	164	actin
42	bud	161	ATPase, GTP
38	Cdc42, Swi4	129	DNA-binding, RNA polymerase II
35	Cdc6, Cdk	124	Ras
33	Cln1, Pho85	122	GTPase
31	microtubule	117	Gcn4
30	DNA polymerase	112	phosphatase
29	Ras, Sbf	111	Uas
28	GTPase, Ime1, Swi6	103	Gal1 promoter
27	GTP	89	polypeptide
26	Cdc25	86	Map
25	Apc, Uas	83	Rap1
24	Cdc14, Ste20	76	histone
22	Cdc45, Map	75	polymerase
21	cyclin-dependent, Ste11	71	GDP, TATA
20	Cdc24, Far1, Mapk	67	Cdc28

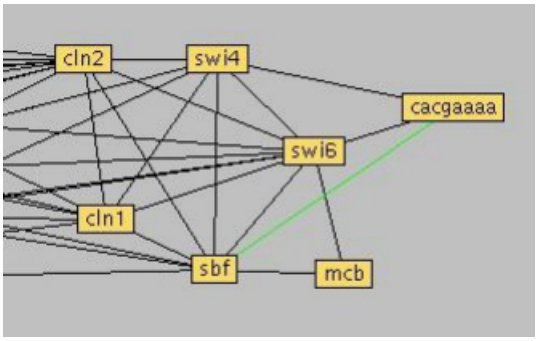

Figure 5: Subgraph showing the relation between the promotor element CACGAAAA and cyclins Cln1 and Cln2.

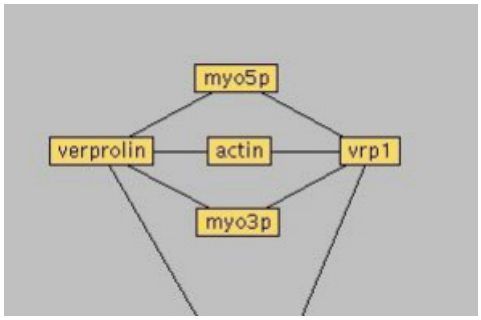

Figure 6: The subgraph for Verprolin and Vrp1 which are linked to the same nodes and turned out to be synonyms for the same protein name.

3.4 Can the SUISEKI System Be Used to Produce New Discoveries?

3.4.1 Closeness in the Interaction Map Indicates Functional Relationships

In the literature the interactions are listed one by one, but by recovering a significant part of them one can add a new dimension to this information. The interaction map puts them in a context and represents a knowledge that is normally restricted to a specialist working with a particular group of proteins. During the deeper examination of protein networks we realised that nodes that are close but not directly connected are in most cases functionally related or form part of the same macromolecular structure.

An example can illustrate this point. The subgraph in Fig. 5 shows the relation between the promotor element CACGAAA, the proteins Swi4 and Swi6 that form part of the transcription factor Sbf and the G1 cyclins Cln1 and Cln2. There exists no direct relation between the promotor and the cyclins but the sentence "SWI4 and SWI6 are components of a factor (SBF) that binds the CACGAAAA (SCB) promoter elements responsible for activation in late G1 of the HO endonuclease, CLN1 and CLN2 genes [MED 1386897]" demonstrates their close functional relationship.

Another example is the relation between Ntc20 and Ntc30. Both of them are, among others, connected to the same nodes but not directly to each other. But they form part of the same functional complex as shown in the following sentences: "Like other identified components of the complex, both Ntc30p and Ntc20p are associated with the spliceosome in the same manner as Prp19p immediately after or concurrently with dissociation of U4, indicating that the entire complex may bind to the spliceosome as an intact form" and "Neither Ntc30p nor Ntc20p directly interacts with Prp19p, but both interact with another component of the complex, Ntc85p [MED 11018040]".

Not all the proteins connected by the same nodes represent related proteins and in some cases we discovered synonyms of the same protein name that appear connected to the same network of proteins. Two of such cases are Verprolin and Vrp1p that have very similar connections (see Fig. 6) but are synonyms of the same protein name ("The proline-rich protein verprolin (Vrp1p) binds to the SH3 domain of Myo3p or Myo5p in two-hybrid tests, coimmunoprecipitates with Myo5p, and colocalizes with Myo5p [MED 9628892]") and Ime2 and Sme1 ("These results suggest that IME1 product stimulates meiosis by activating transcription of SME1 (IME2) and that protein phosphorylation is required for initiation of meiosis [MED 219643]").

But the limit between a functional relationship and a direct interactions between two proteins is not that clear. Therefor me performed an experiment and divided our cell cycle text corpus in two

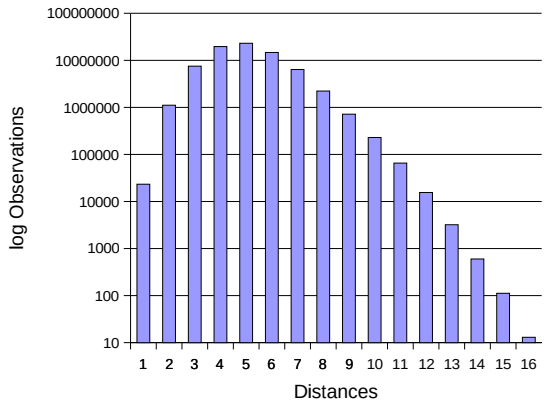

Figure 7: Minimal distances between all nodes in number of edges in the before2000 corpus. Distances of 1 represent a direct connection between two entities (some nodes are not connected at all what is not represented in this graph).

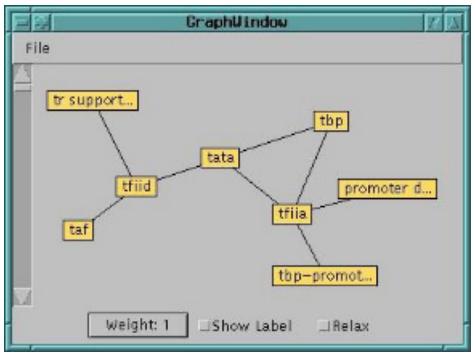

Figure 8: The interactions of transcription factors IIA and IID later shown to form part of a macromolecular complex.

parts, one containing abstracts published till the year 1999 (corpus before2000) and the rest comprising the years 2000 and 2001 (till April 2001; corpus after1999). The corpus before2000 was used as the reference set and compared to the "new" interactions detected in the corpus after1999.

We first analysed the graph generated for the before2000 corpus. The graph obtained is quite inhomogeneous, not all the nodes are interconnected, several regions are separated, and other regions are very densely populated. The distribution of the minimal distances of all nodes (all against all) in the interaction map is shown in Fig. 7. It shows a mean distance of 5 with most of the values spread between 3 and 7 connections.

The analysis of the after1999 corpus revealed a total of 855 new interactions of which 149 corresponded to connections between protein (or gene) pairs that were already part of known interactions (in the before2000 corpus) but with no direct connection. For these new interactions the mean distance between the nodes was 3.2. This means that on average there were 3 edges or 2 nodes between the new interactions (the minimum would be 2 because 1 means a direct interaction). Only for 24 of them there was no connection at all (regardless of the distance) in the literature collected before the year 2000.

These numbers can be compared with a random experiment where 149 nodes were randomly chosen from the interactions before2000 corpus. This experiment was repeated 25 times to calculate standard errors. The mean distance in this simulation was 5.0 ± 0.3 (a number that was expected because this was the mean value of the distribution represented in Fig. 7) and 106.8 ± 5.7 nodes had no connection in the graph (more than 2/3 of the selected nodes). The average distance of the simulated experiment was 5 edges or 4 nodes.

Therefore the new relations discovered after1999 corresponded in most cases to entities that were already connected by shorter distances in the graph before2000 (shorter than average and shorter than expected from a random distribution of interactions). In these cases it is tempting to think that researches in the field were closely following proteins and genes that were becoming suspiciously closer by having similar interactions with other proteins or genes, and finally proved that they interact directly with each other.

3.4.2 Deducing New Interactions

One example of these "new" interactions is illustrated in Fig. 8, where the connection between the transcription factors IIA and IID and TATA box element is visible as an indirect relation mediated by the common binding to the TATA box in the before2000 corpus. It could have been concluded from this analysis that a relation between the IIA and IID transcription factors was to be expected. Indeed, in the after2000 corpus this relation was demonstrated ("The general transcription factor IIA (TFIIA) forms a complex with TFIID at the TATA promoter element, and it inhibits the function of several negative regulators of the TATA-binding protein (TBP) subunit of TFIID [MED 10581267]"). This example could be seen as reminiscent of the mental process of the researches in this field that knowing that various transcription factors activated the same genes decided to investigate their possible participation in molecular complex that would cooperate during the regulation of transcription as a single unit.

4 Discussion

A number of publications have addressed the problem of detecting protein and other molecular interactions from the literature. Large-scale applications of these systems have been limited to co-occurrence studies of gene symbols in MEDLINE abstracts for yeast [15] and human genes [9] that perform very well but are not able to specify the type of relation that is detected between the genes. This is a serious limitation if one is interested in the analysis of interaction networks.

Other, more computer linguationally orientated systems can give more detailed information about the interactions that are reported but they are still limited to small text corpora of a few hundred sentences and have not provided a biological analysis of the problem (e.g. [13, 10, 12, 14, 16]). Furthermore none of these systems explicitly deals with the detection of protein names even if this has been discussed in specific publications [8, 11].

SUISEKI uses concepts from both of these extremes to select only the proteins that interact with each other (and eliminate other types of relations) and to detect the protein names directly from the text and not with a fixed list of names. But it was not overloaded by using methodologies that would have prevented its application to big text corpora. Grammar analysis was substituted by specialised frames that catch frequently used language constructions that are used to express protein interactions. The probability scores associated with the frames allowed a classification of the detected interactions according to their reliability. The precisions range from 80% for the high scoring interactions to less than 50% for low scoring ones, the recall was with more than 70% considerably high.

These values of precision can not be considered as existing but they were good enough for our studies and are the price that has to be paid for the large-scale application of IE techniques.

4.1 The Network of Known Interactions

We have analysed some of the basic properties of the interaction networks for both text corpora, the yeast cell cycle and the full Medline collection about yeast. In both cases the graphs obtained showed interesting properties. First they are very inhomogeneous, with very populated (connected) areas and very sparse ones, including the presence of different disconnected sets of interactions.

Second, they show the capacity of some key biological entities, such as DNA, central metabolites or major macromolecular complexes (cytoskeleton) for nucleating interactions.

Third, they make obvious the presence of central proteins and genes that concentrate different signalling pathways and connect full systems, for example cytoskeleton and transcription regulation. The interest in these proteins and their presence in different areas of research is well correlated with their key role in the interaction graphs. Different examples are clearly in line with these interpretations.

4.2 The Network as a Potential Source of New Discoveries

We propose here the use of the network of interactions extracted from the literature as a potential source of information for the generation of hypothesis about new interactions. These hypotheses could be seen as suggestions for further directed experimental work or as guides to set the priorities in the interpretation of systematic proteomic data.

We use the example of the yeast cell cycle to compare the connectivity of the network of interactions documented before and after the year 2000. In this case it was possible to see how the more recently discovered new interactions between previously known proteins were frequently established between entities (proteins, genes or macromolecular complexes) that were very close in the previously established graph (closer than average and closer than the random distributions). Therefore, it seems feasible to propose a searching strategy based on the study of new interactions among the highly connected nodes that share many similar connections.

Following this strategy we have analysed the relations between transcription factors IIA and IID. Before the year 2000, they belong to a well-connected set of interactions in which they have in common some key connections. This information could have helped to propose a direct interaction between them before it was experimentally discovered in the following year.

It is unavoidable to think that other discoveries of higher biological value than this simple example are hidden in the interaction graphs. Bringing them to the scientific community requires deep knowledge of the research area and orderly access to massive amount of experimental information. The knowledge has to be provided by human experts, while the SUISEKI system can facilitate the access to the information and the formulation of hypothesis. We are confident that the experts will appreciate the possibilities offered by the system at the same time that they are able to avoid the pitfalls introduced by erroneous name detections and incorrectly assigned interactions.

Acknowledgments

This work is support in part by a EC TMR grant (Genequiz). We thank our colleges of the protein design group for interesting discussions and continuous support.

References

[1] Andrade, M.A. and Valencia, A., Automatic annotation for biological sequences by extraction of keywords from MEDLINE abstracts, *ISMB97*, 5:25–32, 1997.

[2] Andrade, M.A. and Valencia, A., Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families, *Bioinformatics*, 14:600–607, 1998.

[3] Blaschke, C., Andrade, M.A., Ouzounis, C., and Valencia, A., Automatic extraction of biological information from scientific text: protein–protein interactions, *ISMB99*, 7:60–67, 1999.

[4] Blaschke, C., Oliveros, J.C., and Valencia, A., Mining functional information associated to expression arrays, *Funct. Integr. Genomics*, 1:256–268, 2001.

[5] Blaschke, C. and Valencia, A., The frame–based module of the Suiseki information extraction system, *IEEE Intelligent Systems in Biology*, in print.

[6] Blaschke, C. and Valencia, A., The SUISEKI information extraction system, In revision.

[7] Blaschke, C. and Valencia, A., Can bibliographic pointers for known biological data be found automatically? Protein interactions as a case study, *Comp. Funct. Genom*, 2:196–206, 2001.

[8] Fukuda, K., Tsunoda, T., Tamura, A., and Takagi, T., Information extraction: identifying protein names from biological papers, *Pacif. Symp. Biocomp. '98*, 707–718, 1998.

[9] Jenssen, T.K., Lægreid, A., Komorowski, J., and Hovig, E., A literature network of human genes for high–throughput analysis of gene expression, *Nature Genetics*, 28:21–28, 2001.

[10] Park, J.C., Kim, H.S. and Kim, J.J., Bidirectional incremental parsing for automatic pathway identification with combinatorial categorial grammar, *Pacif. Symp. Biocomp. '01*, 396–407, 2001.

[11] Proux, D., Rechenmann, F., Julliard, L., Pillet, V., and Jacq, B., Detecting gene symbols and names in biological texts: a first step toward pertinent information extraction, *Genome Informatics*, 9:72–80, 1998.

[12] Proux, D., Rechenmann, F., and Julliard, L., A pragmatic information extraction strategy for gathering data on genetic interactions, *ISMB2000*, 8:279–285, 2000.

[13] Rindflesch, T.C., Tanabe, L., Weinstein, J.N., and Hunter, L., EDGAR: extraction of drugs, genes and relations from the biomedical literature, *Pacif. Symp. Biocomp. '00*, 515–524, 2000.

[14] Sekimizu, T., Park, H.S., and Tsujii, J., Identifying the interaction between genes and gene products based on frequently seen verbs in medline abstracts, *Genome Informatics*, 8:62–71, 1998.

[15] Stapley, B.J. and Benoit, G., Biobibliometrics: information retrieval and visualization from co–occurrences of gene names in Medline abstracts, *Pacif. Symp. Biocomp. '00*, 529–540, 2000.

[16] Thomas, J., Milward, D., Ouzounis, C., Pulman, S., and Carrol, M., Automatic extraction of protein interactions from scientific abstracts, *Pacif. Symp. Biocomp. '00*, 541–552, 2000.

[17] Xenarios, I., Rice, D.W., Salwinski, L., Baron, M.K., Marcotte, E.M., and Eisenberg, D., DIP: the Database of Interacting Proteins, *Nuc. Acids Res.*, 28:289–291, 2000.

[18] Yakushiji, A., Tateisi, Y., Miyao, Y., and Tsujii, J., Event extraction from biomedical papers using a full parser, *Pacif. Symp. Biocomp. '01*, 408–419, 2001.

[19] http://www.ncbi.nlm.nih.gov/pubmed/ or http://www.nlm.nih.gov/Entrez/medline.html