Automatic extraction of biological information from scientific text: protein–protein interactions

Christian Blaschke[1], Miguel A. Andrade[2], Christos Ouzounis[3] and Alfonso Valencia[1*]

[1]Protein Design Group
CNB–CSIC E–28049 Madrid. Spain
[2] European Molecular Biology Laboratory
D–69012 Heidelberg. Germany
[3] Computational Genomics Group
The European Bioinformatics Institute–EMBL
Cambridge CB10 1SD. UK

[*] corresponding author
Protein Design Group. CNB–CSIC
Cantoblanco, Madrid 28049. Spain
Tfn. +34–91–585 45 70 Fax. +34–91–585 45 06
email: valencia@cnb.uam.es

Abstract

We describe the basic design of a system for automatic detection of protein–protein interactions extracted from scientific abstracts. By restricting the problem domain and imposing a number of strong assumptions which include pre–specified protein names and a limited set of verbs that represent actions, we show that it is possible to perform accurate information extraction. The performance of the system is evaluated with different cases of real–world interaction networks, including the *Drosophila* cell cycle control. The results obtained computationally are in good agreement with current biological knowledge and demonstrate the feasibility of developing a fully automated system able to describe networks of protein interactions with sufficient accuracy.

Introduction

Despite the widespread use of computers in biological research, the end result of all scientific experiments is a publication, in the form of text and figures. Even if standards are developed in the future for the deposition of some of this valuable information into some computer–readable form, the problem of retrieving all past knowledge of molecular biology research is staggering.

The success of bioinformatics partly originated from the fact that some data (those for sequences and structures) had already been captured in a computer–readable form in various molecular biology databases, assisting and guiding further experimentation (Andrade and Sander 1997). Yet, the bottleneck remains to effectively associate macromolecular data and other, less well–specified, experimental information. To this end, there is a great need to expand computational representations for other subjects, such as metabolic networks (Karp et al. 1996; Karp and Riley 1998; KEGG) or human genetic disease (OMIM).

The necessity for automatic methods to assist human operators in collecting, curating and maintaining various databases is becoming even more critical, with the exponential increase of the sheer size and complexity of biological information (Kyrpides; Gaasterland et al.).

Here, we present a simple system for the information extraction of protein–protein interactions from scientific journal abstracts available in Medline (NLM). The system is based on our previous experience in the detection of significant, characteristic keywords in sets of Medline abstracts referring to protein families (Andrade and Valencia 1997; Andrade and Valencia 1998). In that case, the use of statistical methods was sufficient to generate meaningful results without the further need of implementing syntax analysis.

The problem here was chosen in such a way that, apart from its inherent scientific interest, it would be amenable to simple approaches in information extraction. In addition, the selection of examples was such that a large corpus of abstracts was available, in order to support statistical analysis of case studies.

A number of possible applications for the system are presented and discussed, including its use in the process of genome annotation or database curation or the

Copyright ©1999, American Association for Artificial Intelligence (www.aaai.org). All rights reserved.

automatic generation of networks of macromolecular interactions.

System Design

We present here the design and first implementation of a system for the identification of protein–protein interactions directly from textual information (**Fig. 1**). The system is based on the counting of the number of sentences containing protein names separated by interaction verbs. The principal issues that we have addressed in the development of this application are the following:

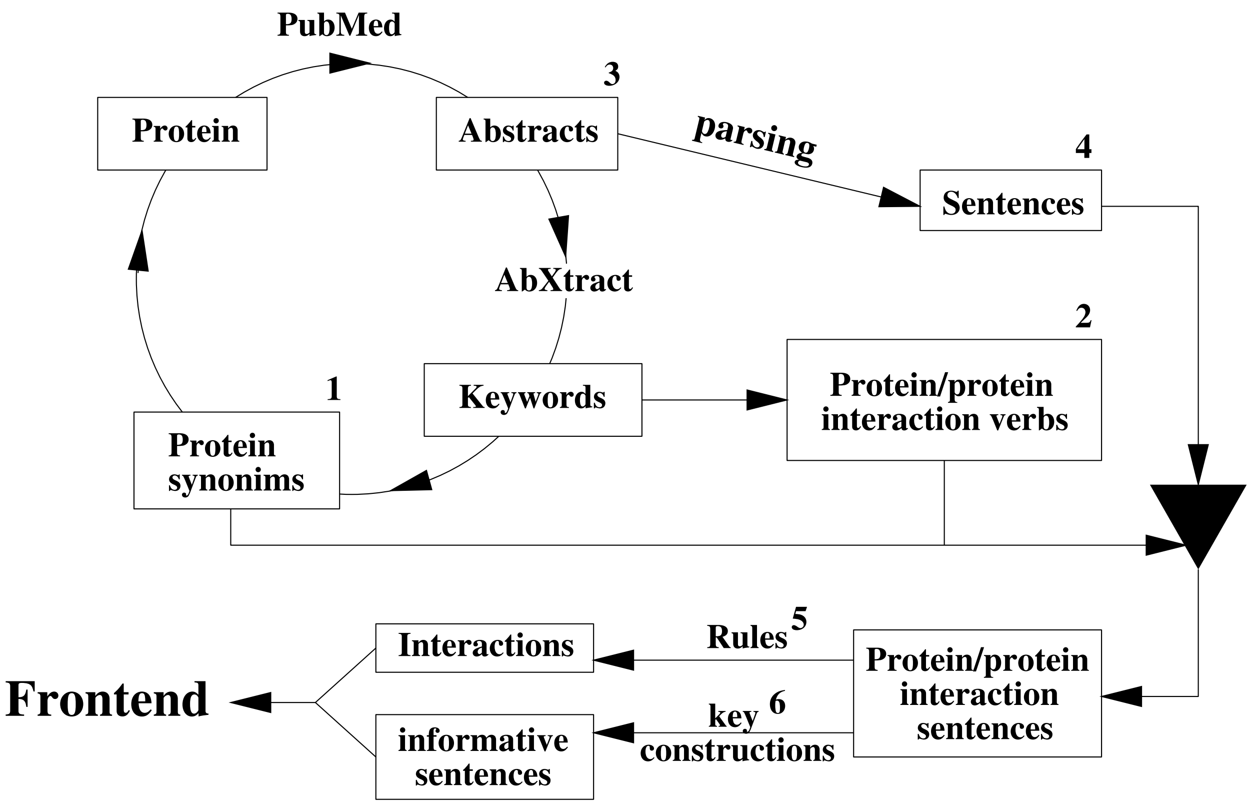

Figure 1: **Description of the system for the extraction of protein interactions from scientific text.** Upper–left corner, a protein is chosen. Abstracts related to this protein are selected using a public Medline server (NLM). These abstracts are analyzed using the AbXtract system (Andrade and Valencia 1998). Synonyms of the target protein are manually selected among the relevant keywords. A new search may be done in Medline with the new set of synonymous protein names. The final set of abstracts is parsed for sentences. Sentences containing any of the synonymous and verbs related to protein interaction are selected as prone to contain information about the interaction with other proteins. The sentences are analyzed with a set of rules and translated to a schema of interactions and dissected in the search for some key contractions containing protein function description. All this information is provided to the human user. The different steps (1.- Protein names, 2. Action verbs, 3. Collection of the corpus of text, 4. parsing of sentences, 5. Set of rules and 6. Extraction of key constructions) are described in detail in the text.

1. Identification of protein names. The issue of name and synonym identification remains as one of the main problems in this type of applications. Here, we assume that protein names are specified by the user. In the future, it may be possible to adopt more specific approaches, like the one developed by (Fukuda et al. 1998), based on morphological criteria.

2. Collection of a corpus of data. Collections of abstracts were obtained for each of the examples described below by accessing all Medline abstracts corresponding to the protein names. These sets are further expanded by searching for related abstracts with the use of the Neighbors utility (Wilbur and Coffee 1994). We have collected five different sets of abstracts:

a) 1435 Medline abstracts directly referenced from each of the *Drosophila* Swissprot entries

b) 4109 Medline abstracts referenced directly from FlyBase

c) 111747 abstracts retrieved obtained extending set b, with the Neighbors utility

d) 518 Medline abstracts containing any of the protein names (related with cell cycle control) and *Drosophila* in the MESH list of terms and,

e) 6278 Medline abstracts by expanding set (d) using Neighbors to identify all related abstracts

3. Identification of actions. In the current implementation, we used a set of 14 pre–specified words indicating actions related to protein interactions. Namely:

- acetylat–e (–ed, –es, –ion)
- activat–e (–ed, –es, –ion)
- associated with
- bind (–ing, –s, –s to, /bound)
- destabiliz–e (–ed, –es, –ation)
- inhibit (–ed, –es, –ion)
- interact (–ed, –ing, –s, –ion)
- is conjugated to
- modulat–e (–ed, –es, –ion)
- phosphorylat–e (–ed, –es, –ion)
- regulat–e (–ed, –es, –ion)
- stabiliz–e (–ed, –es, –ation)
- suppress (–ed, –es, –ion)
- target

In future extensions, it would be desirable to include other action definitions derived from specific keywords extracted with AbXtract (Andrade and Valencia 1998), from dictionaries of common actions in Molecular Biology (Julian) or from selected subsets of public dictionaries of English verbs (WordNet 1998).

4. Identification of passages containing names and actions. The original text was parsed with simple rules, that appear to be very effective in this particular domain, as previously described (Andrade and Valencia 1998). Obvious extensions here would involve more sophisticated systems (Reynar and Ratnaparkhi 1997).

In the current implementation, we have not addressed two hard problems. First, negative sentences that constitute a well–known problem in language understanding (Salton 1989) were omitted. Second, the recognition of implicit information in a number of sentences and/or titles of papers may contain key information, e.g. protein names, that later are used implicitly in other sentences.

5. Rules. The original text is parsed into fragments preceding grammatical separators (”.”/”,”/”;”). We deal separately with each of these text fragments. In order to recognize one interaction in a text fragment, we have elaborated a series of simple rules based on protein/verb arrangement and proximity. First, we select text fragments containing at least two protein names and an action verb. The order of proteins and verbs is used to deduce the interaction. The construction easiest to interpret is ”protein A – action – protein B” (indicating that protein A interacts with protein B and the action verb specifies the type of interaction). Two other variants in which the action is not found between the two protein names, i.e. ”action – protein A – protein B” and ”protein A – protein B – action”, are more difficult to interpret and will be included in a future extension.

6. Extraction of key constructions. See ”future directions”.

Analysis of real–world case studies

The problem of protein–protein interactions is of fundamental importance in biology. In addition, there are virtually no computational methods that allow the identification of interacting partners in a set of protein sequences. The examples presented below form an illustration of the complex set of decisions that have to be made in the analysis of scientific literature.

Reconstruction of the network of interactions in the *Drosophila* Pelle system

A set of nine known interactions were identified by manual inspection of the relevant literature ((Protein Design Group) for a full description of the data generation steps). The results of the automatic analysis are represented in **Fig. 2**, including the correct identification of eight out of nine interactions. The interaction between **toll** and **pelle** is apparently wrong and is produced by sentences such as: ”**tube** and **pelle** then transduce the signal from activated **toll** to a complex of **dorsal** and **cactus**” (Medline 98274203). In this case ”activated” refers to the state of **toll** and not to the action of **pelle** on **toll**, as wrongly interpreted by the system. But at least there exists a weak functional interaction between these proteins.

The corpus of text analyzed was composed of 6728 abstracts (see description in System design).
The proteins represented in the graph are:

pelle: a serine threonine protein kinase

dorsal: transcription factor

toll: a transmembrane receptor

tube: unknown function associated to the membrane

spatzle: an extracellular ligand for toll

cactus: inhibitor of dorsal

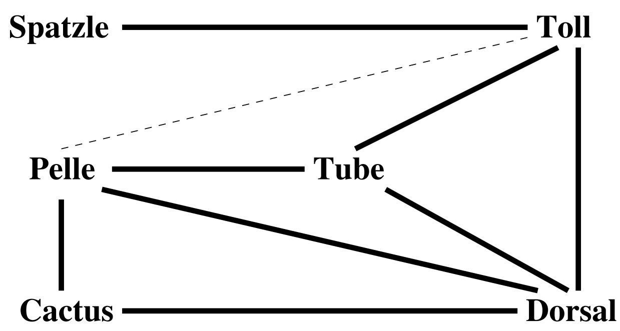

Figure 2: **Graph representing the relations between proteins in the *Drosophila* Pelle system and the comparison with the relations extracted automatically.**

The interactions between these proteins that are represented in a condensed form in the graph are the following:

1. **dorsal** and **cactus** form a cytoplasmatic complex
2. **spatzle** arrives to the cell surface where it binds to a receptor called **toll**
3. **pelle** arrives to the cell surface where it binds/phosphorylates to **tube** which is associated to the membrane
4. The activated **toll** induces a localized recruitment of **tube** and **pelle** to the plasma membrane
5. **pelle/tube** bind dorsal (bound to **cactus**) forming a complex of four components
6. **pelle** phosphorylates **cactus**
7. phosphorylation of **cactus** produces the release and degradation of **cactus** and then **dorsal** translocates into the nucleus and directs gene expression (establishing the **dorsal** ventral polarity of early drosophila embryo)

The interactions detected automatically and the number of occurrences are:

1. dorsal → binds → tube (5)
2. dorsal → regulate → toll (5)
3. pelle → activate(d)→ dorsal (10)
4. pelle → activated → toll (5)
5. pelle → interact → dorsal (6)
6. pelle → regulates → dorsal (5)
7. pelle → regulates → tube (5)
8. spatzle → activate → toll (8)
9. spatzle → activates → toll (6)
10. toll → phosphorylates → tube (5)
11. toll → regulated → dorsal (6)
12. tube → activate → dorsal (5)
13. tube → activates → pelle (4)
14. tube → interact → dorsal (10)
15. tube → interact → pelle (4)

The sentences that were used by the system to deduce the relation between **spatzle** and **toll** are given in **Fig. 3a**. "Activation" has a clear meaning and the relation between proteins corresponds to a common syntax involving the two protein names separated by the action verb. It is interesting to consider some of the sentences that could have been selected by more complex rules (see System Design) in which the action does not lay between the protein names (**Fig. 3b**). The first of them contains the same information but it is more difficult to follow. The second one corresponds to more specific actions like "inhibition" and "regulates" and contain more than two protein names. They indicate a relation between **spatzle** and **toll** with **cactus**, but this relation is described in a particular experimental system, and cannot be easily generalized. The last sentence in **Fig. 3b** implicitly proposes a relation between **spatzle** and toll with **dorsal**, but it would be a too strong simplification to conclude that **spatzle** and **toll** are regulating **dorsal**.

From the analysis of this example, it seems that the clear interactions are readily detected and their significance is enhanced by their frequent occurrence in different abstracts, while a number of other, more complex, relations are seldom repeated. Therefore, the analysis should rely on those simple characteristics that are consistently repeated in the reference data set.

3a. Sentences that correspond to the construction proteinA–action–proteinB and were utilized by the system for building the network of interactions.

1.- MED 97070052: actions: **activate**; Proteins: **spatzle**; **toll**;
this process is thought to restrict the action of three follicle cell gene functions, encoded by windbeutel, nudal, and, pipe, to the ventral follicle cells, where they lead to the localized activation of a serine protease cascade required to produce the active **spatzle** ligand to **activate** the **toll** receptor
2.- MED 94170368: actions: **activates**; Proteins: **spatzle**; **toll**;
spatzle acts immediately upstream of the membrane protein toll in the genetic pathway, suggesting that **spatzle** could encode the ventrally localized ligand that **activates** the receptor activity of **toll**
3.- MED 98362749: actions: **activate**; **cleaved**; Proteins: **spatzle**; **toll**;
proteolytically **cleaved spatzle** could therefore dimerize and **activate** the **toll** receptor by inducing receptor dimerization.

3b. Example of sentences that correspond to construction different from proteinA–action–proteinB and were NOT utilized by the system for building the network of interactions.

1.- MED 96033803: actions: **activated**; Proteins: **spatzle**; **toll**;
the ligand for the **toll** receptor is thought to be **spatzle**, a secreted protein that is **activated** by proteolytic cleavage
2.- MED 98175880: actions: **inhibited**; Proteins: **spatzle**; **toll**; *cactus*;
here we demonstrate a dorsalizing activity for the heterologous easter, **spatzle** and **toll** proteins in uv ventralized xenopus embryos, which is **inhibited** by a co injected dominant *cactus* variant
3.- MED 96422863: actions: **regulates**; Proteins: **spatzle**; **toll**; *dorsal*;
after fertilization, the initial asymmetry of the egg chamber is used by the **spatzle toll** pathway to generate within the embryo a nuclear gradient of the transcription factor *dorsal*, which **regulates** the regional expression of a set of zygotic genes

Figure 3: **Sentences analyzed by the system for extracting the relation between spatzle and toll.** The keywords (protein names and action verbs) are highlighted. The identifiers of the Medline abstracts are indicated.

Analysis of a complex system: the cell cycle control in *Drosophila*.

Cell cycle control is a much more complex example than the one already mentioned. The network of protein interactions identified in this process is shown in **Fig. 4**. For simplicity, the types of the particular interactions are not represented in the figure. The first observation concerns the coverage, stability and accuracy of the system. In this case, the coverage is relatively small, since a significant number of interactions were identified only for 33 of the 91 protein names used in the initial bibliographic screening. This is a underestimate, since the list of 91 names includes many synonyms (for example: **Cdk7**, **Cyclin–dependent kinase 7** and **DmCdk7**) that were considered as a single entity during the construction of the interaction network. This low coverage is reasonable, since at this stage it is difficult to cover the most subtle instances.

The protein names included in the initial literature screening were: anachronism, Arrowhead, bHLH, c Myb, cak, cdc kinase, cdc2, cdc21 homolog, cdc25 homolog, cdc2c, CDC2Dm, CDC5 homolog, cdc6, Cdi3, Cdk activating kinase, Cdk1, cdk2, Cdk7, cdks, Chk1 homolog–kinase, Chk1 homolog kinase, crm, CycA, CycE, cyclin A, cyclin B, cyclin B2, cyclin D, Cyclin dependent kinase 7, cyclin dependent kinase inhibitor, cyclin E, cyclin H, cyclin J, dacapo, DEAD box RNA helicase, disc proliferation abnormal, Dm Myb, Dm-cdc2c, DmCdk7, dmcycd, DMcycE, dmcyce, DmMO15, dmyc, DP, dpp, E2F, escargot, fizzy, fleabag, G1 Cyclin, G2 M Cyclin, gadd45, grapes, Licensing factor, LIM homeodomain transcription factor, m cdk, MCM4 homolog, medullaless, mpf, mus209, Mutagen sensitive 209, Myb, myc, p21, p34cdc2, PCNA, peanut, pitchoune, polo, Polymerase delta/epsilon processivity factor, Proliferating cell nuclear antigen, Protein phosphatase 2A, RBF, rca1, Retinoblastoma family protein, rfc, roughex, rux, secreted glial glycoprotein, string, terribly reduced optic lobes, trol, twine, twins, twist, ubiquitin, wee1, wee1 kinase, zen, zerknuellt.

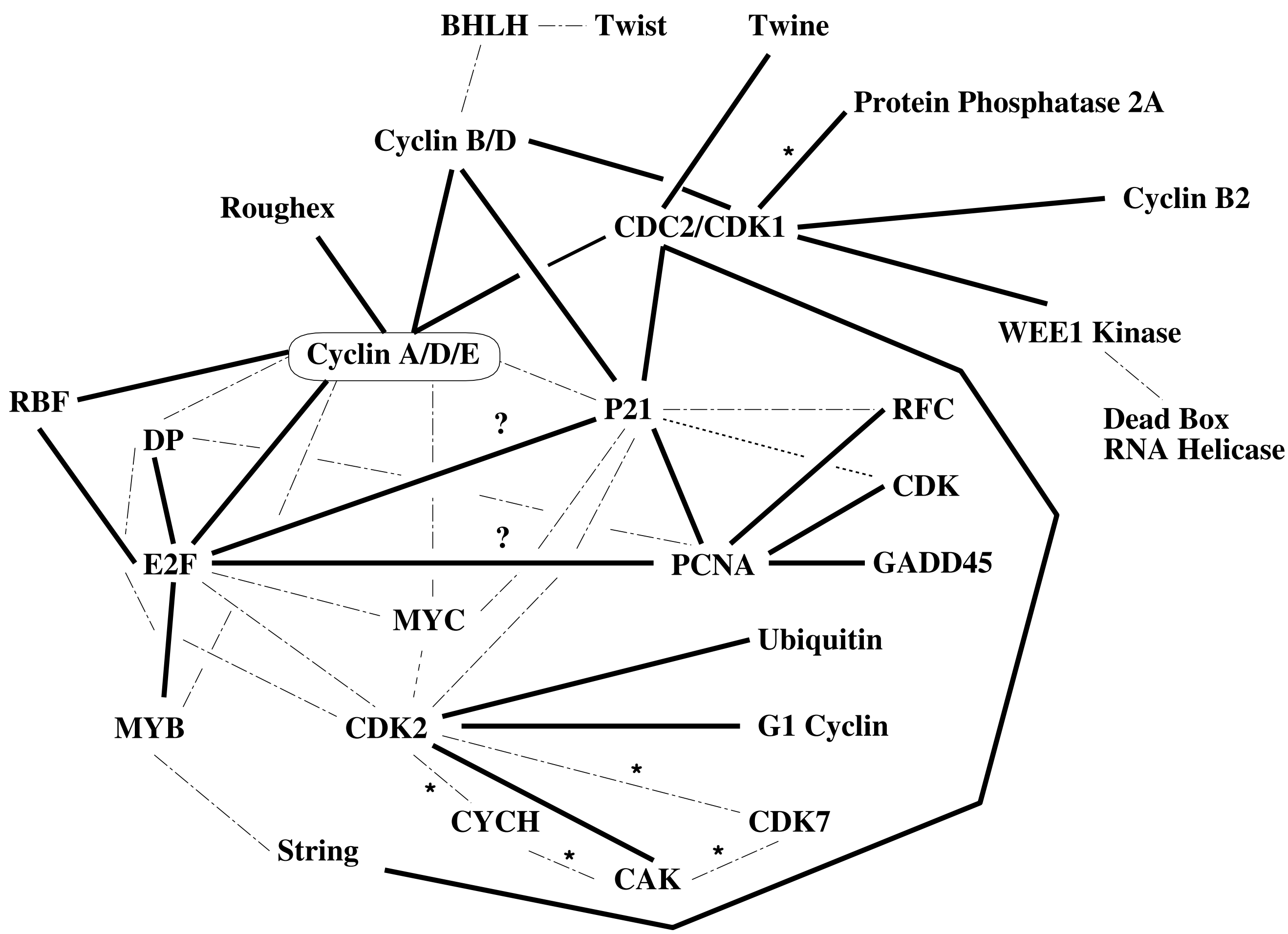

Figure 4: **Drosophila cell cycle interactions automatically detected by the system.** The interactions between these proteins are represented in a condensed form in the graph. For simplicity the type of interaction are not shown, but they can be retrieved from (Protein Design Group). The interactions that correspond to well known cases in the literature are shown in tight lines (—), dubious interactions are marked with interrogations, other possible interactions that were extracted by the system but could not be confirmed by human experts are shown with $(- \cdot - \cdot -)$, finally the interaction between cdk and p21 that was not detected by the system $(\cdots)$. (for * see comments in the text)

From a biological standpoint, performance is impressive since the system can identify all the key interactions, protein–protein associations of various types around the central protein p21, the kinase–cyclin complexes and the set of proteins controlling the kinase action.

One of the interesting parts of the output is the relation between **Cdk7**, **Cdk2**, **Cak** and **CycH**. The system detected the following relations: **Cak** inhibits/phosphorylates **Cdk7**, **Cak** activates/phosphorylates **Cdk2**, **Cdk7** phosphorylates **Cdk2**, **CycH** phosphorylates **Cak** and **CycH** phosphorylates **Cdk2**. Although these interactions are all described in early literature, the current knowledge records that **Cak** is a complex formed by **Cdk7** and **CycH**, and the **Cak** complex regulates **Cdk2**. There is an element of generalization implicit in later publications, embodying previous, more dispersed findings. A clear improvement here would be the generation of associated weights for texts according to their level of generality.

The analysis of one of the sentences is instructive (the full text cases are given in (Protein Design Group)). The sentence "while **cdk7 cyclin h** is sufficient for **phosphorylation** of **cdk2**, we show that p36mat1 is required for efficient **phosphorylation** of p53 by **cdk7 cyclin h**, suggesting that p36mat1 can act as a substrate specificity determining factor for **Cdk7 cy-**

clin H" was extracted from Medline abstract 98038791, and contains information for the action, **phosphorylation** and for the proteins: **cyclin H**, **Cdk7**, and **Cdk2**. This sentence represents a partial view of the system; **Cdk7** and **cyclin H** are already presented as a complex (implicit information) and the complex is proposed to phosphorylate Cdk2.

Missing information can also lead to somewhat misleading conclusions. For example, in the case of **Myc**, this protein works in a complex with another protein called **Max**. This second protein was not included in our search and is not reported in the final graph, leading to the impression that **Myc** is the only effector of different interactions.

Another source of potential problems is coming from the impossibility of distinguishing between biologically significant interactions and other results obtained in particular experimental conditions. One such case is the interaction between **cyclins A**, **B**, **D** or **E** with **p21**. Only interactions with **cyclins A** and **B** seem to be biologically relevant while reports of interactions with **cyclins D** and **E** have only been demonstrated in vitro and are probably irrelevant in *Drosophila*. The same is true for the interactions between **Cdc2/Cdk1** and **protein phosphatase 2A**, since this phosphatase has a very broad specificity and is not the specific phosphatase that activates **Cdc2/Cdk1**. Therefore, the current analysis is extracting true information but it cannot distinguish the quality of this information or its relevance for a given system.

One final example can illustrate the gain in the robustness of the conclusions produced by the accumulation of relations. The protein name **p21** corresponds to two different proteins: an inhibitor of **Cdk**, related with the cell cycle, and an oncogene, involved in intracellular signaling. Even if the abstracts corresponding to both proteins were included in the initial selection, the interactions corresponding to the **p21**–oncogene were finally not included in the network of interactions since they are not connected with the other cell cycle proteins.

Conclusions

We describe the design of a system for the automatic detection of protein–protein interactions from scientific text. The basic idea is that sentences derived from sets of abstracts will contain a significant number of protein names connected by verbs that indicate the type of relation between them. By pre–specifying a limited number of possible verbs, we avoid the complexity of semantic analysis.

It is important to notice that the current design relies on the peculiarities of this knowledge domain. On the one hand, we are dealing with a very specialized type of information, including a very restricted and direct use of English, short sentences and a great abundance of highly specialized terms in Molecular Biology. On the other hand, the number of abstracts is adequate for a quantitative approach, where the abundance of occurrences of different events is more significant than the single occurrence of a useful sentence.

Beyond the possibilities of this approach, illustrated here with the analysis of two real systems, we are well aware of the limitations of simple statistical algorithms for detecting the structure of written sentences. The following sentence contains three names and two actions (one positive and one negative): "the **toll** signaling pathway, which controls dorsal ventral patterning during drosophila embryogenesis, regulates the nuclear import of **dorsal** in the immune response, but here we show that the **toll** pathway is <u>not</u> required for nuclear import of **dif**" (MED 98169073). It would be very difficult to use the interesting information contained in this sentence with the help of simple methods: the assumption is that if this fact is common, it will be present in the same collection of abstracts in shorter and easier sentences.

In this sense, it is relevant to compare the number of interactions discovered by the system with five different corpus of data analyzed (see System design). The simple corpus of text (set **a**), containing mainly information about protein sequencing renders only two interactions while a more sophisticated corpus (set **b**) that includes text that human experts selected as relevant for the function of the proteins discovers 117 interactions. The results are comparable (28 interactions) to the ones obtained with a set of Medline abstracts taken directly by querying with the relevant protein names (set **d**). Interestingly, this result can be improved (285 interactions) by including in the analysis those Medline abstracts detected as related by the Neighbor facility (set **e**). This is the set that has been used for the analysis reported here. The same extension of the network of interactions obtained by the expansion of the number of abstracts can be seen in the case of set **c**, obtained by the expansion with Neighbors of set b. In this case the number of interactions for the proteins corresponding to the cell cycle control example is of 554.

This comparison points to two important conclusions, first the system gains tremendously when a large corpus of text is analyzed, like the collections of 6728 or 111747 abstracts, and second, the results can be improved by creating data sets that are directly relevant for a particular subject, such as the set of abstracts selected from FlyBase to describe the function of each one of the proteins.

To summarize, the three main results of this work are: first, restricting the number of relevant names and verbs virtually eliminates the problem of text understanding; second, specifying a particular subproblem enhances quality and accuracy of the results obtained from free text (here we dealt with small problems of protein–protein interactions in specific *Drosophila* systems) and finally the repeated occurrence of certain facts can enhance the quality of the discovery and strengthen the identification of particular relationships.

Future directions

Front end for human expert analysis. The most immediate use of the current system is as an interface to facilitate the access to bibliographic information for human experts working in database curation or in the annotation of large quantity of genomic data. It can also serve as a front end for other systems like Medline or as part of genome analysis tools like Genequiz (Andrade et al. 1999) or Magpie (Gaasterland and Sensen 1996).

Description of protein function with rules about the structure of the sentences. As part of the future front end facility we have implemented a first set of 13 rules to cover the most obvious descriptions of protein functions, like "protein A is a .." or "protein B is a new member of a family" ((Protein Design Group) for the description of the rules and their application) . The initial analysis shows that it is possible to detect sentences containing a short and clear description of the function of some proteins, for example: "**cak** is a cdk cyclin complex implicated in the control of multiple cell cycle transitions" or "**cak** is a trimeric enzyme containing **cdk7**, **cyclin h**" that would have solved the conflicting situation described above during the analysis of the relation between **cdk7** and **cak**.

Reconstruction of large network of interactions. The system itself can be used for building extended networks of possible protein interactions. We are currently analyzing 111747 abstracts for building a first networks connecting 4851 protein names. This type of applications can be seen as analogous to experimental methods like two hybrid systems, where many true interactions can be detected with the cost of identifying many irrelevant relations.

Acknowledgments

We are indebted to Manuel Serrano, CNB–CSIC for the expert analysis of the cell cycle interaction network.

References

Andrade, M.A., Brown, N.P., Leroy, C., Hoersch, S., Daruvar, A.D., Reich, C., Franchini, A., Tamames, J., Valencia, A., Ouzounis, C. and Sander, C. (1999) *Automated genome sequence analysis and comparison.* Bioinformatics in press.

Andrade, M.A. and Sander, C. (1997). *Bioinformatics: from genome data to biological knowledge.* Current Opinion in Biotechnology. 8, 675-683

Andrade, M.A. and Valencia, A. (1997). *Automatic annotation for biological sequences by extraction of keywords from MEDLINE abstracts. Development of a prototype system.* ISMB 97. 5, 25-32

Andrade, M.A. and Valencia, A. (1998). *Automatic extraction of keywords from scientific text: Application to the knowledge domain of protein families.* Bioinformatics. 14, 600-607
http://www.cnb.uam.es/cgi-bin/blaschke/abxs

Fukuda, K., Tsunoda, T., Tamura, A. and Takagi, T. (1998) *Toward information extraction: identifying protein names from biological papers.* Pacific Symp. Biocomputing 3: 705-716
www.hgc.ims.u–tokyo.jp/∼ichiro/

Gaasterland, T., Andersson, A. and Sensen, C. *Magpie genome sequencing project list.*
http://www–c.mcs.anl.gov/home/gaasterl/genomes.html

Gaasterland, T. and Sensen, C.W. (1996) *Fully automated genome analysis that reflects user needs and preferences – a detailed introduction to the MAGPIE system architecture.* Biochimie 78, 302-310

Julian
www.mblab.gla.ac.uk/∼julian/

Karp, P.D., Ouzounis, C. and Paley, S.M. (1996). *HinCyc: A knowledge base of the complete genome and metabolic pathways of H. influenzae.* Intelligent Systems for Molecular Biology 1996 (ISMB96), 116-124. St Louis MO– AAAI Press.

Karp, P.D. and Riley, M. (1998). *EcoCyc: Encyclopedia of E. coli Genes and Metabolism.*
http://ecocyc.PangeaSystems.com/ecocyc/ecocyc.html

KEGG: Kyoto Encyclopedia of Genes and Genomes. http://www.genome.ad.jp/kegg/kegg2.html

Kyrpides, N. *Completed and Ongoing Genome Projects.* http://geta.life.uiuc.edu/∼nikos/genomes.html

NLM – National Library of Medicine. MEDLINE.
http://www.nlm.nih.gov/ and
BioMedNet. MEDLINE service.
http://biomednet.com/db/medline

OMIM – Online Mendelian Inheritance in Man.
http://www3.ncbi.nlm.nih.gov/omim/

Protein Design Group CNB–CSIC, Madrid
www.gredos.cnb.uam.es/medline_interactions/
contains additional information about **a,** manual derivation of the Pelle system. **b,** full description of interactions in the Drosophila cell cycle control example. **c,** sentences used for the analysis of the Drosophila cell cycle control network. **d,** sentences extracted in the Drosophila cell cycle control example with a simple set of 13 rules about protein function description.

Reynar, J.C. and Ratnaparkhi, A. (1997). *A maximum entropy approach to identifying sentence boundaries.* In Proceedings of the 5th Conf. on Applications of Natural Language Processing, 16-19.

Salton, G. (1989). *Automatic Text Processing.* Addison–Wesley (Addison–Wesley series in Computer Science), Reading, Massachusetts, USA

The interactive Fly
http://sdb.bio.purdue.edu/fly/aimain/1aahome.htm

Wilbur, W.J. and Coffee, L. (1994) *The effectiveness of document neighboring in search enhancement.* Inf. Process. Manage. 30, 253-266

WordNet – a Lexical Database for English. 1998.
http://www.cogsci.princeton.edu/∼wn/