Toward Routine Automatic Pathway Discovery from On-line Scientific Text Abstracts

See-Kiong Ng [1] **Marie Wong** [2]
skng@krdl.org.sg marie@bic.nus.edu.sg

[1] Kent Ridge Digital Labs, 21 Heng Mui Keng Terrace, Singapore 119613
[2] NUS Bioinformatics Centre, National University of Singapore, Singapore 119260

Abstract

We are entering a new era of research where the latest scientific discoveries are often first reported online and are readily accessible by scientists worldwide. This rapid electronic dissemination of research breakthroughs has greatly accelerated the current pace in genomics and proteomics research. The race to the discovery of a gene or a drug has now become increasingly dependent on how quickly a scientist can scan through voluminous amount of information available online to construct the relevant picture (such as protein-protein interaction pathways) as it takes shape amongst the rapidly expanding pool of globally accessible biological data (e.g. GENBANK) and scientific literature (e.g. MEDLINE).

We describe a prototype system for automatic pathway discovery from on-line text abstracts, combining technologies that (1) retrieve research abstracts from online sources, (2) extract relevant information from the free texts, and (3) present the extracted information graphically and intuitively. Our work demonstrates that this framework allows us to routinely scan online scientific literature for automatic discovery of knowledge, giving modern scientists the necessary competitive edge in managing the information explosion in this electronic age.

1 Introduction

Current progress in genomics and proteomics projects worldwide has generated an increasing number of putative new proteins — the biochemical functional characterization of these proteins are continuously being discovered and reported. At the same time, progress in information technology has resulted in an increasing number of useful databases for biological data (e.g. GENBANK) and scientific literature (e.g. MEDLINE) on the Internet, as well as an increasing number of traditionally paper-based research journals coming online. The traditional trip to the library will soon be replaced with a simple search on the world wide web from the computer keyboard.

The ease of such electronic dissemination of biological data and scientific discoveries on the Internet has effectively globalized and accelerated modern research. The race to a new gene or drug is now increasingly dependent on how quickly a scientist can keep track of the voluminous information online to capture the relevant picture (such as protein-protein interaction pathways) hidden within the latest research articles that are continuously coming online from all over the world. To cope with this information explosion in the electronic age, scientists need a tool that will help them automatically scan the internet for research literature, extract the relevant knowledge from the (possibly multiligual) free text sources, and then present the information in an intuitive and readable form. In this paper, we describe our ongoing attempt to address this need for timely and routine extraction of knowledge from online texts for scientific research. Our initial project focuses on the processing of online abstracts for discovering specific protein-protein interactions to automatically construct the underlying pathway maps.

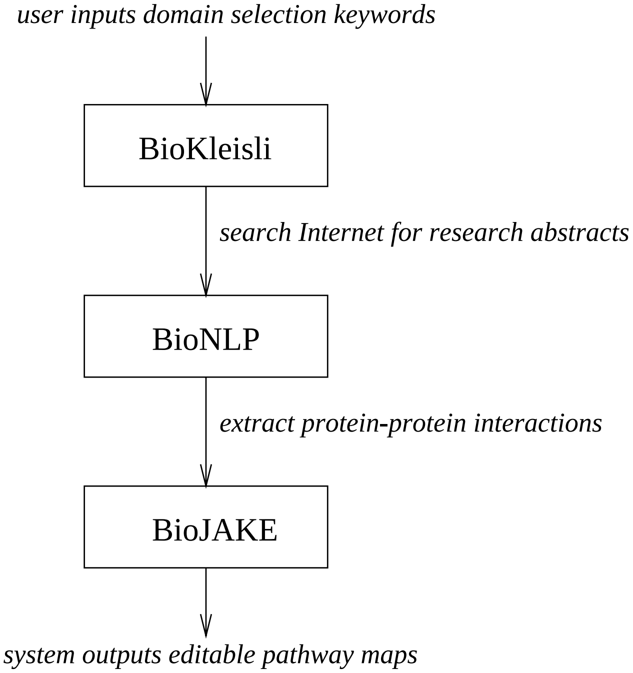

user inputs domain selection keywords
BioKleisli
search Internet for research abstracts
BioNLP
extract protein-protein interactions
BioJAKE
system outputs editable pathway maps

Figure 1: System architecture.

2 System

We have enlisted three key software engines in our system (Fig. 1):

1. *BioKleisli*[1]. In coping with the dispersed and heterogeneous nature of scientific databases on the Internet, we use the database-independent BioKleisli system [7, 15] as our query engine for retrieving scientific abstracts and literature from (possibly multiple) online bibliographic resources on the Internet [17]. BioKleisli is a well-tested query system that has been built on the formal foundations of modern query languages [4, 3, 16] and specially designed for integrating data from disparate sources across many geographies and systems;

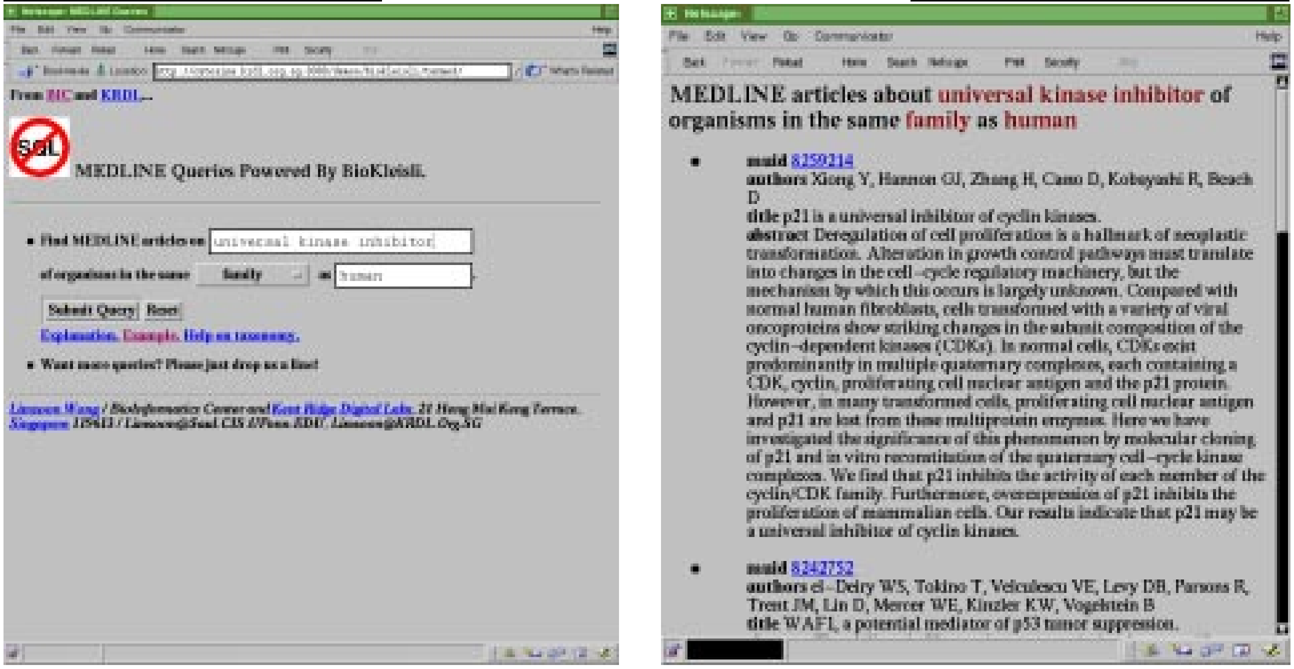
2. *BioNLP*. We have developed the BioNLP module to process the free texts in the scientific abstracts retrieved by BioKleisli. BioNLP identifies protein names mentioned in the free texts and performs function word pattern matching to discover protein-protein relation expressed in the abstracts;

3. *BioJAKE*. We have adapted the BioJAKE visualization engine for organizing the information extracted by BioNLP, graphically presenting and managing the automatically constructed pathway maps in an intuitive manner to the user. BioJAKE is a software tool that has been designed specifically for the creation, manipulation, and visualization of metabolic pathway diagrams [12].

3 Method

Integration of BioKleisli, BioNLP, and BioJAKE has resulted in a useful tool for timely and routine discovery of knowledge from online scientific literature. In Fig. 2, we show a web interface to the BioKleisli query engine in which a user can easily query scientific databases on the Internet for protein-protein interaction pathways by entering specific keywords to specify the domain of interest (e.g.

[1] BioKleisli is now marketed by Kris Technology Inc., 713 Santa Cruz Avenue, Suite 2, Menlo Park, CA 94025. Email: info@kris-inc.com.


File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Stop
Bookmarks Location What's Related

From BIC and KRDL...

MEDLINE Queries Powered By BioKleisli.

- Find MEDLINE articles on universal kinase inhibitor
  of organisms in the same family as human.

  Submit Query Reset

  Explanation. Example. Help on taxonomy.

- Want more queries? Please just drop us a line!

Limsoon Wong / Bioinformatics Center and Kent Ridge Digital Labs, 21 Heng Mui Keng Terrace, Singapore 119613 / Limsoon@Saul.CIS.UPenn.EDU, Limsoon@KRDL.Org.SG

MEDLINE articles about universal kinase inhibitor of organisms in the same family as human

- **muid** 8259214
  authors Xiong Y, Hannon GJ, Zhang H, Casso D, Kobayashi R, Beach D
  title p21 is a universal inhibitor of cyclin kinases.
  abstract Deregulation of cell proliferation is a hallmark of neoplastic transformation. Alteration in growth control pathways must translate into changes in the cell–cycle regulatory machinery, but the mechanism by which this occurs is largely unknown. Compared with normal human fibroblasts, cells transformed with a variety of viral oncoproteins show striking changes in the subunit composition of the cyclin–dependent kinases (CDKs). In normal cells, CDKs exist predominantly in multiple quaternary complexes, each containing a CDK, cyclin, proliferating cell nuclear antigen and the p21 protein. However, in many transformed cells, proliferating cell nuclear antigen and p21 are lost from these multiprotein enzymes. Here we have investigated the significance of this phenomenon by molecular cloning of p21 and in vitro reconstitution of the quaternary cell–cycle kinase complexes. We find that p21 inhibits the activity of each member of the cyclin/CDK family. Furthermore, overexpression of p21 inhibits the proliferation of mammalian cells. Our results indicate that p21 may be a universal inhibitor of cyclin kinases.

- **muid** 8242752
  authors el–Deiry WS, Tokino T, Velculescu VE, Levy DB, Parsons R, Trent JM, Lin D, Mercer WE, Kinzler KW, Vogelstein B
  title WAF1, a potential mediator of p53 tumor suppression.

Figure 2: A web interface for user query in BioKleisli.

"universal kinase inhibitors). From then on, BioKleisli will automatically perform the necessary database-specific queries to retrieve the latest set of research abstracts from the online databases. In this way, the busy scientist no longer needs to learn about the idiosyncrasies of the various Internet databases.

The research abstracts thus collected by BioKleisli are then passed on to the BioNLP module, which automatically processes the free text abstracts and extracts protein-protein interaction relations that are of interest to the scientist. For example, the scientist may specify to look only at protein-protein relations such as *"activate"* and *"inhibit"*.

The extracted knowledge of protein-protein interactions are then passed on to BioJAKE, which organizes the discovered relations into pathway maps and graphically displays them as a manageable *hyperlinked* and *editable* network. Fig. 3 shows the output to the user's *"universal kinase inhibitors"* query.

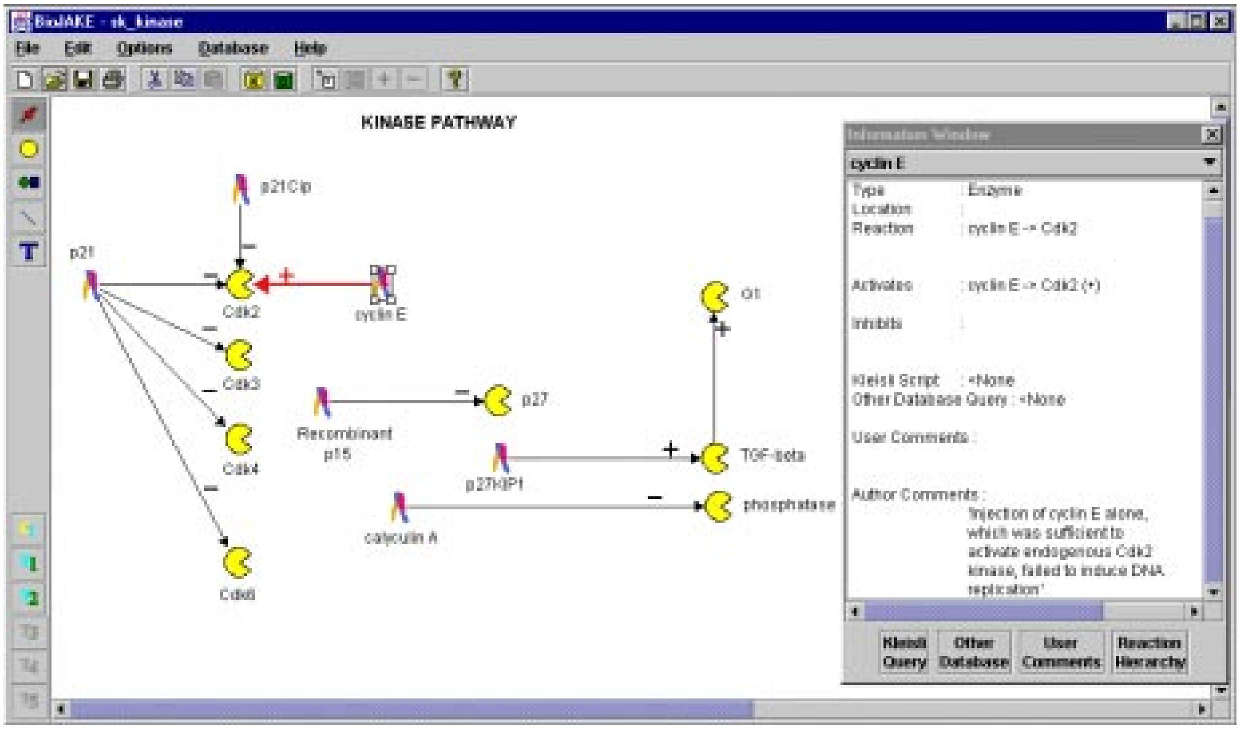
Using this integrated tool, the scientist can easily automate the routine discovery of pathway maps from the latest information on the Internet. Such timely discovery can provide an invaluable edge in the modern race for a new gene or drug.

In the next three sections, we describe the three modules in more details. Other than BioNLP, both BioKleisli and BioJAKE have been previously developed for other applications [7, 12]; we will therefore provide only a brief description for these two modules.

3.1 BioKleisli

The BioKleisli system [6, 7] is an advanced broad-scale integration technology that has proved useful in the bioinformatics arena [2, 1, 5, 9]. Many bioinformatics problems require access to data sources that are high in volume, highly heterogeneous and complex, constantly evolving, and geographically dispersed. Solutions to these problems usually involve multiple carefully sequenced steps and require information to be passed smoothly between the steps. BioKleisli is designed to handle these requirements directly by providing a high-level query language, CPL, [3], that can be used to express complicated transformation across multiple data sources in a clear and simple way.

BioKleisli is extensible in many ways. It can be used to support many other high-level query



cyclin E

Type : Enzyme
Location :
Reaction : cyclin E -> Cdk2

Activates : cyclin E -> Cdk2 (+)

Inhibits :

Kleisli Script : <None
Other Database Query : <None

User Comments :

Author Comments :
'Injection of cyclin E alone, which was sufficient to activate endogenous Cdk2 kinase, failed to induce DNA replication'

Figure 3: A protein-protein interaction pathway map automatically created from a user query of ”universal kinase inhibitor”. Information about a selected interaction (shown here as a dotted arrow between `cdk2` and `cyclin E`) is shown in an information window on the right.

languages by replacing the CPL module. BioKleisli can also be used to support many different types of external data sources by adding new drivers, which forward BioKleisli’s requests to these sources and translate their replies into BioKleisli’s exchange format. The standard installation of BioKleisli contains over sixty drivers for all popular bioinformatics systems.

3.2 BioNLP

The BioNLP module is a rule-based system written in Perl to perform simple natural language processing (NLP) on the extracted scientific abstracts using pattern matching. There are two major tasks when extracting protein-protein interaction information from scientific abstracts:

1. *Protein name identification.* Straightforward use of a dictionary of protein names is inadequate in this domain because new names are continuously being invented and quoted in medical and biological papers. The names of the new proteins must therefore be identified by linguistic means;

2. *Information extraction.* Co-occurrence of protein names in an article abstract, a sentence, or a phrase generally implies that the proteins are related in some way. Such co-occurrence is a useful heuristic for extracting information about protein-protein interactions.

There are two corresponding sets of rules in BioNLP specifying the patterns for identifying protein names and for extracting specific protein-protein interactions from the free texts.

3.2.1 Identifying protein names

Identifying protein names can be challenging because the standard nomenclature is often only loosely followed by authors naming new proteins [8, 11]. Even under the standard nomenclature, protein names can still be difficult to identify, as some of the protein names are long compound words or have multiple variants. To tackle this task, Fukuda *et al.* [8] have devised a set of rules to identify protein names based on lexical considerations such as the presence of upper cases and of special characters. We have incorporated their lexical rules in BioNLP, and augmented it with the following strategies:

1. *Exclusion by standard dictionaries.* We filter out most of the non-proper nouns in the abstracts by looking up the words in a classical dictionary. In the current version of BioNLP, we simply do a *'grep'* on */usr/dict* on UNIX;

2. *Inclusion with semantic clues.* Proper nouns that are not recognized, but that are linked together by protein-protein interaction function words (e.g. "activate" or "inhibit"), are classified as potential new protein names;

3. *Inclusion with protein dictionaries.* We maintain a local protein dictionary for the rapid identification of common protein names. This dictionary also allows the re-inclusion of protein names that are made up of the nouns excluded by Step 1. The dictionary may be manually edited by the user, or automatically learned from the protein-protein relations subsequently extracted from the abstracts. If necessary, new protein names (e.g. unclassified proper nouns learned from the semantic clues described above) can be automatically verified against protein databases using BioKleisli.

3.2.2 Extracting protein-protein interactions

Papers written to present specific results often contain information on subjects which are secondary to the main topic, but which may be quite useful for researchers working on new areas where a complete picture is unavailable and there are plenty of missing links to be filled. In a new field such as protein-protein interaction discovery, it is therefore important to extract as many protein-protein relationships as possible, including those that are only mentioned in the literature in a cursory manner, to build a reasonably *complete* interrelated network for the proteins of interest. The BioNLP rules are therefore designed to capture as many protein-protein relationships from the literature as possible, whether or not they are the main topic in the papers which they are mentioned.

In BioNLP, the user can specify the type of protein-protein interactions (e.g. "inhibit-activate") to extract from the abstracts. BioNLP maintains a set of function words for each supported interaction types. These function words may be edited by the user. Their roles are as keys into the literature for seeking out sentences that may contain related protein-protein information. For example, some of the key function words for the *inhibit-activate* relationship are:

inhibitor: {`inhibit, suppress, negatively regulate`}

activator: {`activate, transactivate, induce, upregulate, positively regulate`}

BioNLP seeks out sentences containing any of the function words and then searches for any protein names mentioned. These protein names are then associated with the function words using BioNLP's suite of pattern matching rules to determine their actor-patient roles. Some examples of the pattern matching rules in BioNLP are shown below. In these examples, both <*A*> and <*B*> can denote individual or a conjunction of protein names, while <*fn*> denotes a matched function word:

1. *<A> ... <fn> ... <B>:* This rule models the basic sentence pattern such as *"A inhibits B, C, and D"*;

2. *<A> ... <fn> of ... <B>:* This rule models sentences such as *"A, an activator of B, is found to be lacking in the patient population".*;

3. *<A> ... <fn> by ... <B>:* This models sentences in passive voice, such as *"A is inhibited by the activities of B."*;

4. *<A> ..., which ... <fn> ... <B>, ...:* This models sentences such as *"A, which inhibits the activities of B, is found to be lacking in the patient population"*;

5. *<fn> of <A> is ... <B>:* This template models sentences such as *"Induction of A is caused by B."*.

These rules can be straightforwardly implemented in a pattern-matching language such as Perl. The rules are currently hand-coded in BioNLP; however, the simplicity of the pattern rules, as well as the regularity of the sentence structures in the corpus, indicate that it should be possible for BioNLP to automate the learning of the pattern matching rules in future.

3.3 BioJAKE

The extracted protein relations are passed on to BioJAKE, with the actual sentences on which BioNLP's pattern matching rules apply retained as pertinent textual annotations. In BioJAKE, these extracted relations are displayed collectively as an editable, hyperlinked pathway network. The Java-based BioJAKE program [12] was originally created for the visualization, creation, and manipulation of metabolic pathways. It has been designed to provide a familiar and easy-to-use interface while still allowing for the input and manipulation of complex and detailed metabolic data. In coping with the detailed and diverse sources of data available across the Internet, BioJAKE also provides a mechanism by which remote database queries can be stored and performed with respect to individual molecules within a pathway. This remote database access functionality is offered in addition to comprehensive local database creation, management, and querying capability. BioJAKE is therefore a particularly useful tool for visualizing and maintaining the information for the purpose of routine pathway discovery from online databases.

4 Example

We look at an example in which the user's query is for the *"inhibit-activate"* interactions in proteins that are *"effective human cyclin inhibitors"*. When the user enters the search phrase into the web query form, BioKleisli automatically connects to the relevant scientific literature databases to download the related article abstracts. In this example, 26 articles in MEDLINE found to be related to *"effective human cyclin inhibitors"* were automatically downloaded from Internet.

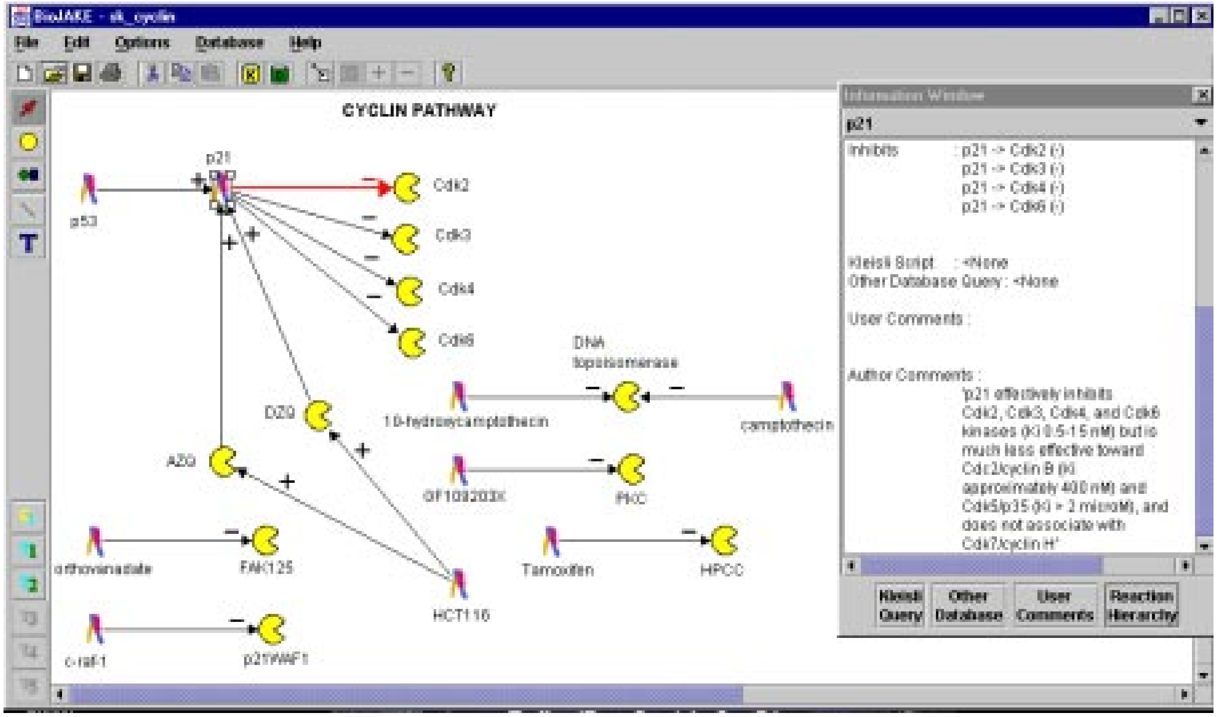
BioNLP then proceed to process the downloaded text abstracts. It identifies protein names from the free texts, and then extracts sentences containing meaningful relations between the protein names and the function words related to the protein-protein interaction that the user is interested in. Here is an example of an extracted sentence (Example sentence 1):

```
[Article 7626805] p21 effectively inhibits Cdk2, Cdk3, Cdk4, and Cdk6 kinases
(Ki 0.5-15 nM) but is much less effective toward Cdc2/cyclin B (Ki approximately
400 nM) and Cdk5/p35 (Ki > 2 microM), and does not associate with Cdk7/cyclin
H.
```

The above sentence matches the pattern template *<A> ... <fn> ... <B>*; the matched function key phrase is:

```
‘‘ effectively inhibits’’
```

The following protein-protein relations are then extracted from the sentence:



```
Inhibits        : p21 -> Cdk2 (-)
                  p21 -> Cdk3 (-)
                  p21 -> Cdk4 (-)
                  p21 -> Cdk6 (-)

Kleisli Script    : <None
Other Database Query : <None

User Comments :

Author Comments :
                  'p21 effectively inhibits
                  Cdk2, Cdk3, Cdk4, and Cdk6
                  kinases (Ki 0.5-15 nM) but is
                  much less effective toward
                  Cdc2/cyclin B (Ki
                  approximately 400 nM) and
                  Cdk5/p35 (Ki > 2 microM), and
                  does not associate with
                  Cdk7/cyclin H'
```

Figure 4: A protein-protein interaction pathway map automatically constructed from a user query of "effective human cyclin inhibitor".

```
p21 --inhibit--> Cdk2
p21 --inhibit--> Cdk3
p21 --inhibit--> Cdk4
p21 --inhibit--> Cdk6
```

As another example, the next sentence (Example sentence 2) matches the pattern template *<fn> of <A> is ... <B>*:

```
[Aricle 9685373] Induction of p21 is p53-dependent; it does not occur in
cells with mutant p53 or in cells expressing human papillomavirus E6.
```

The matched function key phrase is:

```
''Induction of''
```

The protein-protein interaction extracted is:

```
p53 --activate--> p21
```

In this exercise, 11 sentences were extracted from the 26 MEDLINE abstracts, resulting in 16 unique protein-protein interactions. Fig. 4 shows the the pathway network output by BioJAKE.

5 Further work

Our ultimate goal is to be able to automatically organize the scientific knowledge extracted from online sources into networks of interactions which can be tracked, visualized, edited, analyzed, and eventually simulated over time. The system that we have described here is our first step toward this goal. Further work include:

- *Textbook pathway libraries incorporation.* Existing textbook pathways are unlikely to be mentioned in the research abstracts. It is therefore necessary to maintain libraries of textbook pathways so that newly discovered interactions can be incorporated with these known textbook pathways to form complete pathway maps;

- *Routine pathway discovery tracking.* In order to track the development of protein-protein interactions, the encoding of the extracted interaction relations should be stored such that the pathway maps can be *incrementally* constructed over time. The system can then automatically flag the scientist when a new link in a pathway map of interest is discovered and reported;

- *More sophisticated NLP techniques.* Pattern matching is inherently limiting and can only detect simple syntatic relations like the ones mentioned in this paper. For example, the relationships between `p21` and `Cdk7/cyclin H` in Example sentence 1 cannot be detected with simple pattern matching and requires more sophisticated NLP techniques.

6 Conclusion

Automatic retrieval of electronic scientific literature, integrated with rule-based information extraction and natural language processing, and graphical and hyperlinked visualization of extracted knowledge, forms a useful framework facilitating routine automated scientific knowledge discovery from online texts. Our prototype system implements such a framework, extracting protein-protein interaction information from scientific text abstracts automatically downloaded from online bibliographic sources, and graphically displaying the discovered relations as an editable and manageable pathway networks with hyperlinked nodes. Such a system can be a useful tool for the routine and timely discovery of knowledge from online scientific literature [10, 13], providing up-to-date accumulation of knowledge that is useful for, say, the screening of novel therapeuticals. By further integrating the system with a pathway simulation module [14], we can even hope that researchers may someday be able to understand and experiment with the pathways *in silico* as they are being discovered.

Acknowledgments

We are grateful to Dr. Limsoon Wong for his invaluable feedback and advice on this project.

References

[1] Baker, P. G. and Brass, A., Recent development in biological sequence databases, *Current Opinion in Biotechnology*, 9:54–58, 1998.

[2] Benton, D., Bioinformatics—principles and potential of a new multidisciplinary tool, *Trends in Biotechnology*, 14:261–272, 1996.

[3] Buneman, P., Libkin, L., Suciu, D., Tannen, V., and Wong, L. Comprehension syntax, *SIGMOD Record*, 23:87–96, 1994.

[4] Buneman, P., Naqvi, S., Tannen, V., and Wong, L., Principles of programming with complex objects and collection types, *Theoretical Computer Science*, 149:3–48, 1995.

[5] Chen, J., Strauss, D., and Wong, L. Using Kleisli to bring out features in BLASTP results, *Genome Informatics*, 9:102–111, 1998.

[6] Chung, S.-Y. and Wong, L., Kleisli, a new tool for data integration in biology, *Trends in Biotechnology*, 1999 (to appear).

[7] Davidson, S., Overton, C., Tannen, V., and Wong, L., BioKleisli: a digital library for biomedical researchers, *International Journal of Digital Libraries*, 1:36–53, 1997.

[8] Fukuda, K. , Tamura, A., Tsunoda, T., and Tagagi, T., Toward information extraction: identifying protein names from biological papers, *Proc. the Pacific Symposium on Biocomputing '98*, 707–718, 1998.

[9] Karp, P.D. Database links are a foundation for interoperability, *Trends in Biotechnology*, 14:273–279, 1996.

[10] Ohta, Y., Yamamoto, Y., Okazaki, T., Uchiyama, I., and Takagi, T., Automatic construction of knowledge base from biological papers, *Proc. the Fifth International Conference on Intelligent Systems for Molecular Biology*, 218–225, 1997.

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

[12] Salamonsen, W., Mok, K.Y.-C., Kolatkar, P., and Subbiah, S., BioJAKE: a tool for the creation, visualization and manipulation of metabolic pathways, *Proc. the Pacific Symposium on Biocomputing '99*, 392–400, 1999.

[13] 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*, 9;62–71, 1998.

[14] Tomita, M., Hashimoto, K., Takahashi, K., Matsuzaki, Y., Matsushima, R., Saito, K., Yugi, K., Miyoshi, F., Nakano, H., Tanida, S., and Shimizu, T., E-CELL project overview: towards integrative simulation of cellular processes, *Genome Informatics*, 9:242–243, 1998.

[15] Wong, L., Kleisli, a functional query system, *Journal of Functional Programming*, 1999 (to appear).

[16] Wong, L., Normal forms and conservative extension properties for query languages over collection types, *Journal of Computer and System Sciences*, 52:495–505, 1995.

[17] Wong, L., Some MEDLINE queries powered by Kleisli, *ACCESS*, 25:8–9, 1998.