# Kinase Pathway Database: An Integrated Protein-Kinase and NLP-Based Protein-Interaction Resource

Asako Koike, Yoshiyuki Kobayashi and Toshihisa Takagi

*Genome Res.* 2003 13: 1231-1243
Access the most recent version at doi:10.1101/gr.835903

Resource

# Kinase Pathway Database: An Integrated Protein-Kinase and NLP-Based Protein-Interaction Resource

Asako Koike,[1,2,4,5] Yoshiyuki Kobayashi,[1,3,4] and Toshihisa Takagi[1]

*[1]Human Genome Center, Institute of Medical Science, University of Tokyo, Shirokane-dai, Minato-Ku, Tokyo 108-8639, Japan; [2]Central Research Laboratory, Hitachi, Ltd., Higashi-Koigakubo, Kokubunji-shi, Tokyo 185-8601, Japan; [3]Life Science Group, Hitachi, Ltd., Minamidai, Kawagoe-shi, Saitama, 350-1165, Japan*

**Protein kinases play a crucial role in the regulation of cellular functions. Various kinds of information about these molecules are important for understanding signaling pathways and organism characteristics. We have developed the Kinase Pathway Database, an integrated database involving major completely sequenced eukaryotes. It contains the classification of protein kinases and their functional conservation, ortholog tables among species, protein–protein, protein–gene, and protein–compound interaction data, domain information, and structural information. It also provides an automatic pathway graphic image interface. The protein, gene, and compound interactions are automatically extracted from abstracts for all genes and proteins by natural-language processing (NLP). The method of automatic extraction uses phrase patterns and the GENA protein, gene, and compound name dictionary, which was developed by our group. With this database, pathways are easily compared among species using data with more than 47,000 protein interactions and protein kinase ortholog tables. The database is available for querying and browsing at http://kinasedb.ontology.ims.u-tokyo.ac.jp/.**

Cellular signaling in eukaryotes plays a key role in the processes of tissue growth, cell differentiation, and rapid response to environmental changes. Defects in signaling components have been found to cause diseases such as cancer (Hunter 2000). The phosphorylation and dephosphorylation of proteins through the mediation of protein kinases and phosphatases is one of the most frequently used signaling techniques. Some signaling pathways that include phosphorylation and dephosphorylation have been conserved quite well in evolutionarily distant organisms. Knowing the orthologous protein kinases and the pathways that include the protein kinases among major eukaryotes can aid our understanding of all signaling pathways, their evolution, and organism characteristics. It will also allow us to predict unknown protein functions.

The Protein Kinase Resource (Smith 1997), a well known and useful database related to protein kinases, provides various types of information such as protein conformations, all members of the protein kinases, and mutations. However, it does not provide pathways that include protein kinases and ortholog tables among species. Some databases such as Transpath (Schacherer et al. 2001), KEGG (Kanehisa et al. 2002), and DIP (Xenarios et al. 2002) provide biological pathways and/or protein–protein interaction information, which can be summarized quite well manually and is thus useful in many respects. However, manual information extraction is very time-consuming and costly. Consequently, recent information that is relevant is not likely to be accumulated in such databases due to the rapid progress in molecular biology.

On the other hand, automatic extraction of information from articles has been studied, and some systems for extracting protein–protein interactions have been developed (Sekimizu et al. 1998; SUISEKI, by Blaschke and Valencia 2001; GeneWays, by Krauthammer et al. 2002). However, most of these systems are not open to the public and are used only locally. Furthermore, many of them use original-term tagging components to identify gene and protein names by using rules, external knowledge, or both, and the problem of synonyms seems to be unresolved. That is, even if the same protein interaction information is extracted, it is not recognized as the same interaction when written as synonyms.

Toward a resolution of these problems, we developed the Kinase Pathway Database using sequence analysis and natural-language processing (NLP) techniques. The database provides ortholog tables for the protein kinases of the major eukaryotes *Drosophila melanogaster*, *Mus musculus*, *Rattus norvegicus*, and *Homo sapiens*. It also provides manually commented functional conservation information on protein kinases in these species, domain information, structural information, gene/protein/compound interactions, and a Java-based graphic viewer. The interaction data were automatically extracted from MEDLINE abstracts using natural-language processing for all genes, not only protein kinases. This system offers the following main features. (1) A protein, gene, and compound name dictionary called GENA (http://gena.ontology.ims.u-tokyo.ac.jp/search/servlet/gena) is used to identify as many synonyms as possible. GENA specifies the relationship between gene and protein names extracted from text and the locus (sequence) or external database. (2) By combing automatically extracted information and ortholog information, users can easily get interaction information on the ortholog genes or proteins of the target genes or proteins and can easily compare the pathways of different species on the graphic viewer.

Here we present an overview of the Kinase Pathway Da-

**[4]These authors contributed equally to this work.**
**[5]Corresponding author.**
**E-MAIL akoike@ims.u-tokyo.ac.jp; FAX 81-3-5449-5434.**
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.835903.

tabase and discuss the recall and precision of extracted protein, gene, and compound interaction information. In the main text, we briefly summarize the performance of automatic information extraction. However, this subject requires an extended discussion for the examination of its validity. Accordingly, our analysis of extraction errors and a comparison with other natural-language processing methods are added in the Appendix.

# RESULTS AND DISCUSSION

## Database Contents

### *Protein Kinases*

The predicted numbers of protein kinases are summarized in Table 1. Their family classification and ortholog tables are contained in the current version (March 2003) of the Protein Kinase Database. The numbers of predicted protein kinases are slightly different from those in previous reports (Hunter and Plowman 1997; Plowman et al. 1999; Morrison et al. 2000; Manning et al. 2002) This is mainly due to the following reasons: (1) Our data do not include Celera human data (www.celera.com), EST, or pseudogene data; (2) the public data have been refined better compared to that used in previous studies; (3) a difference in fragment threshold; (4) a different assignment of the eukaryotic protein kinase group name (when the sequence is similar to multiple groups). Because the difference due to Celera and EST data will be decreased with increases in the precision of public databases, we did not use those data.

### *Automatically Extracted Protein–Interaction*

The total numbers of abstracts used, the protein, gene, and compound interactions extracted, and the protein types extracted are summarized in Table 2. They are contained in our database currently and are updated monthly. To evaluate our method, the precision and recall are calculated. The precision of extraction (=true positive/[true positive+false positive]) is summarized in Table 3. "True positive" indicates the number of correctly extracted interactions, and "false positive" indicates the number of wrongly extracted interactions. Here, we do not evaluate the condition of protein, gene, or compound interactions, such as direct or indirect interaction, and active, inhibited, regulated or undirected interactions, because they are too complex and semantic analysis may be required to raise the precision of extraction. Therefore, we evaluate only protein name and direction of interaction (which protein is the signal donor or acceptor in the signal cascade). Furthermore, "mutant", "dominant negative", "gene promoter", etc. are not distinguished from gene or protein wild type in this version. To distinguish parsing or phrase-pattern errors and errors caused by GENA (e.g., errors due to names that are the same as a synonym of another gene or the same as a general noun or adjective; see Appendix), two types of precision are used. In Table 3, the numbers in front of the parentheses indicate the precision without GENA error, and those in parentheses indicate the precision with GENA error. The precision was measured by manually checking randomly extracted interactions for each species (the numbers of samples are given in Table 3).

For the calculation of recall, 500 abstracts (200 for *S. cerevisiae* and 300 for *H. sapiens*) were checked manually, and the results were compared with automatically extracted results. The recall rate (=true positive/[true positive + false negative]) values of *S. cerevisiae* and *H. sapiens* are 12(automatic extraction)/46(manual extraction) = 26% (12/46 = 26%) and 38/153 = 25% (27/120 = 23%), respectively. "False negative" indicates the number of wrongly unextracted interactions. The numbers in parentheses are the results without compounds. Further discussion of extraction errors and the comparison with other methods are given in the Appendix.

This program leaves room for further analysis of coordinate clauses and anaphora of pronouns, and improvement in these areas will increase the recall and precision. We are working on those improvements now. About half of the molecular names are not written as gene names; rather they are written as family names (including family-like names) or as unspecified gene names (see Appendix). To extract interaction information between families and between families and proteins, we are preparing an ontology of family names that contains enough synonyms.

**Table 1. The Number of Each Group or Family of Protein Kinases in Each Organism**

| Protein kinase family name | *S. cerevisiae* | *C. elegans* | *D. melanogaster* | *H. sapiens* |
|---|---|---|---|---|
| Protein kinase | | | | |
| Eukaryotic protein kinase | | | | |
| AGC group | 17 | 33 | 29 | 64 |
| CaMK group | 17 | 30 | 26 | 47 |
| CMGC group | 24 | 44 | 29 | 54 |
| STE group | 18 | 25 | 21 | 53 |
| CK group | 8 | 75 | 13 | 12 |
| PTK group | 0 | 79 | 29 | 90 |
| OTHER group | 31 | 38 | 54 | 66 |
| Histidine kinase | 2 | 1 | 1 | 5 |
| Protein kinase-like | | | | |
| Guanylyl cyclase kinase | 1 | 27 | 12 | 9 |
| Choline/ethanolamine kinase | 2 | 9 | 2 | 3 |
| Inositol kinase | 10 | 10 | 14 | 21 |
| Diacylglycerol kinase | 2 | 8 | 8 | 10 |
| Adenylate kinase | 0 | 10 | 12 | 6 |

## Comparison With Other Databases

We compared our data with other manually constructed databases to evaluate our database contents. MIPS contained 7351 nonredundant physical interactions and 1368 nonredundant genetic interactions as of October 2002. They consist of 4479 protein types, 4416 of which are registered in GENA. GENA-IDs are assigned for 7256 physical interactions and 1356 genetic interactions. Only 399 physical interactions and 149 genetic interactions of these are common with our extracted data (Kinase Pathway Database). Of the MIPS physical interactions (7351), 6356 interactions are yeast two-hybrid results. They may not be written in the text, but are summarized in tables.

In the DIP database, there is a total of 16,581 nonredundant interactions regarding our target organisms. They consist of 5720 proteins (*H. sapiens*, 687; *M. musculus*, 177; *R. norvegicus*, 82; *D. melanogaster*,

**Table 2.** Numbers of Abstracts, Proteins, and Protein Interactions

| Species | Number of abstracts | Number of extracted protein/compound types | Total number of protein interactions (Numbers in parentheses are nonredundant types) |
|---|---|---|---|
| *S. cerevisiae* | 37,921 | 2,144 | 3,854 (2,529) |
| *C. elegans* | 4,109 | 211 | 232 (157) |
| *D. melanogaster* | 10,918 | 1,067 | 1,967 (1,470) |
| *M. musculus* | 163,886 | 5,276 | 17,493 (11,633) |
| *R. norvegicus* | 179,930 | 2,452 | 8,129 (4,999) |
| *H. sapiens* | 480,074 | 7,752 | 46,354 (26,543) |

45; *C. elegans*, 5; *S. cerevisiae*, 4724). Of these, 5327 proteins (*H. sapiens*, 585; *M. musculus*, 116; *R. norvegicus*, 56; *D. melanogaster*, 37; *C. elegans*, 4; *S. cerevisiae*, 4529) are registered in GENA. Therefore, GENA-IDs are assigned for 14,888 interactions (*H. sapiens*, 567; *M. musculus*, 47; *R. norvegicus*, 17; *D. melanogaster*, 38; *C. elegans*, 2; *S. cerevisiae*, 14,217). Of the *S. cerevisiae* interactions (14,217), 6256 interactions are yeast two-hybrid results. Only 706 interactions (*H. sapiens*, 144; *M. musculus*, 16; *D. melanogaster*, 1; *C. elegans*, 2; *S. cerevisiae*, 543) of those are found in our extracted data. In DIP, some interactions are extracted from tables and figures in review articles. Many interactions are not written in abstracts. In TRANSPATH, 3021 molecules are connected by 3225 reactions from 977 extracted papers (almost all reviews). Unfortunately, those data could not be downloaded, and we could not compare it with our results.

The overlap between manually gathered data and our data is not so large. Uncollected data will be probably be complemented by the analysis of review articles.

## The Web Interface of the Kinase Pathway Database

The architecture and contents of Kinase Pathway Database are summarized in Figure 2, cited in the Methods section. The homepage of the Kinase Pathway Database (http://kinasedb.ontology.ims.u-tokyo.ac.jp/) allows the contents of the database to be accessed by the following search methods: (1) protein kinase classification, (2) pathway search, (3) protein interaction search, (4) phylogenetic tree search, (5) orthologous protein search, (6) multiple-domain search, and (7) hierarchical-structural information search. In the classification of protein kinase families, the group → family → subfamily → members can be followed hierarchically. The searches by methods 2–5 are done by gene name queries. Because the retrieval system uses the GENA gene name dictionary, synonyms are also accepted in queries. The pathway search involves conditions such as the specification of starting point proteins or starting and ending proteins, and whether or not the direction of the pathway is to be considered. By combining these options, this database provides the answers for various questions, for example, whether the two proteins genetically interact, which molecules interact with this protein, and which molecules interact with an ortholog of this protein in other organisms.

Figure 1 shows the results of searching with the Web interface. The manually developed comment file for each family (search method 1) is shown in Figure 1A, and Figure 1B is an ortholog table (search method 5). To express gene duplication, "level" has been used. The "III" indicates genes that have diverged from "II". The "II" indicates those that have diverged from "I". The phylogenetic tree for kinases is shown in Figure 1F (search method 4), and the interaction data automatically extracted from abstracts are shown in Figure 1D (search method 3). As evidence of the extracted interaction information, the corresponding sentence from which the interaction data is extracted and the author information are provided as shown in Figure 1E. Using the information in Figure 1B and D, the pathways for one or two organisms can be drawn side by side, as shown in Figure 1C (search method 2). In this figure, the orthologous proteins have been specified as the starting proteins of the pathways, and the interacting proteins within two steps from the starting proteins are drawn. In the pathway viewer, orthologous proteins are selected according to the ortholog table level. For example, when *M. musculus* PAK1 is selected, *S. cerevisiae* STE20, CLA4, *D. melanogaster* PAK, *R. norvegicus* PAK1, and *H. sapiens* PAK1 are shown as orthologous proteins with interaction information in the intermediate Web interface, and one protein can be selected (STE20 in this example; Fig. 1C). However, other PAK subfamily members such as *H. sapiens* PAK2–7 are not shown, because they are set as paralogs in the ortholog table. By using the pathway viewer, the known interlog (both proteins are ortholog) pathways such as STE11–→PBS2–→HOG1 in *S. cerevisiae* and MAP3K1–→MAP2K4–→MAPK8 in *H. sapiens* are easily found. It is expected that the unknown interlog pathways will also be found.

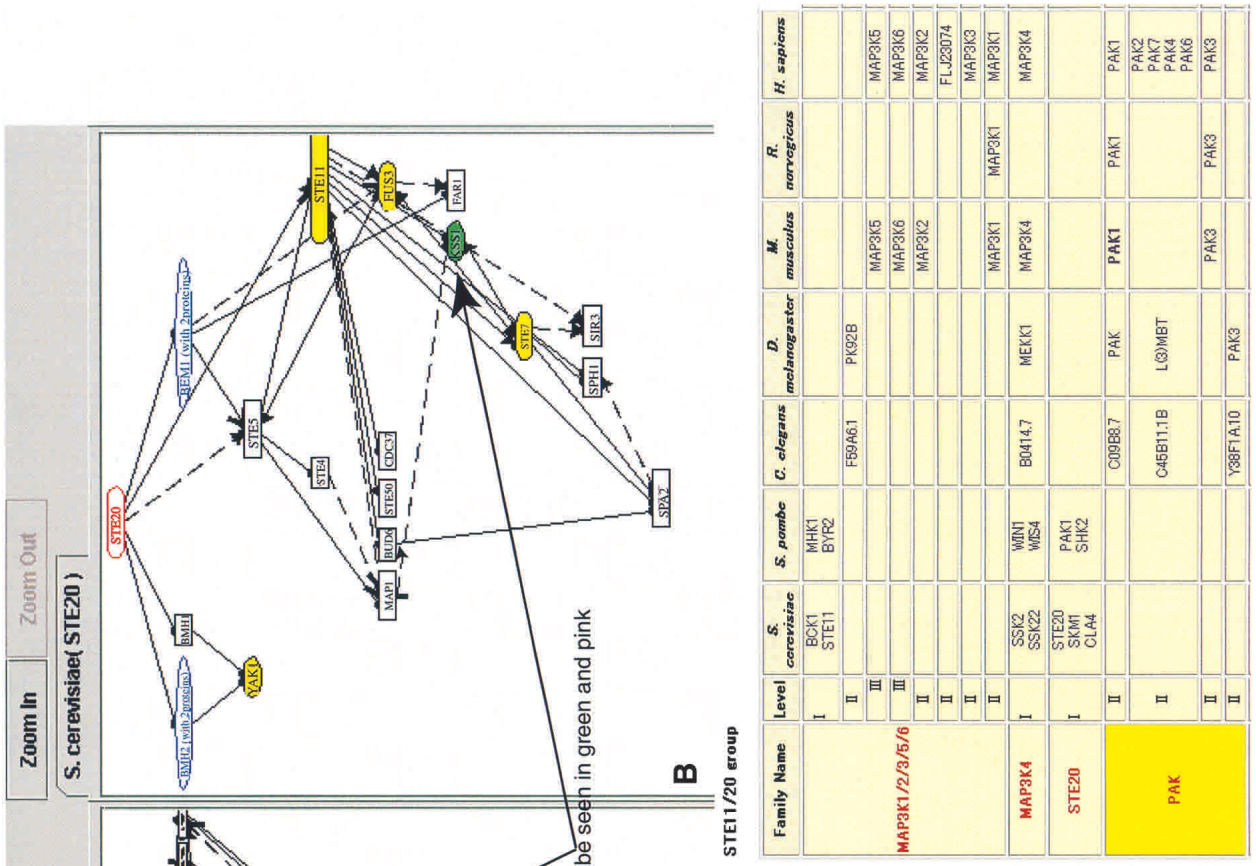
Further information related to the use of this database, statistical information regarding its contents, and the structure of the relational database can be accessed from the database Web page.

**Table 3.** Summary of Precision

| Species | Precision (including error caused by GENA)/ number of interaction samples |
|---|---|
| *S. cerevisiae* | 92% (90%)/196 |
| *C. elegans* | 97% (95%)/171 |
| *D. melanogaster* | 96% (80%)/200 |
| *M. musculus* | 95% (85%)/291 |
| *R. norvegicus* | 94% (87%)/281 |
| *H. sapiens* | 94% (83%)/263 |

## Application

The main purpose of the Kinase Pathway Database is to answer questions the user has about data, as shown in the previous section. The ultimate goals are to extract general interaction rules and predict unknown pathways by analyzing these interactions, and to understand the evolution of signaling transduction and/or protein interactions. The conservation rate of protein–protein interactions between organisms, the structural slant of protein interaction pairs, and the evolutionary meaning of protein–interaction domain pairs are given as example subjects. It is expected that these numerous interaction data will be useful for gaining new insights into protein interactions and signal transduction.




| Family Name | Level | S. cerevisiae | S. pombe | C. elegans | D. melanogaster | M. musculus | R. norvegicus | H. sapiens |
|---|---|---|---|---|---|---|---|---|
| MAP3K1/2/3/5/6 | I | BCK1<br>STE11 | MHK1<br>BYR2 | | | | | |
| | II | | | F59A6.1 | PK92B | | | |
| | III | | | | | MAP3K5 | | MAP3K5 |
| | III | | | | | MAP3K6 | | MAP3K6 |
| | II | | | | | MAP3K2 | | MAP3K2 |
| | II | | | | | | | FLJ23074 |
| | II | | | | | | | MAP3K3 |
| | II | | | | | MAP3K1 | MAP3K1 | MAP3K1 |
| MAP3K4 | I | SSK2<br>SSK22 | WIN1<br>WIS4 | B0414.7 | MEKK1 | MAP3K4 | | MAP3K4 |
| STE20 | I | STE20<br>SKM1<br>CLA4 | PAK1<br>SHK2 | | | | | |
| PAK | II | | | C09B8.7 | PAK | PAK1 | PAK1 | PAK1 |
| | II | | | C45B11.1B | LQ(3)MBT | | | PAK2<br>PAK7<br>PAK4<br>PAK6 |
| | II | | | | | PAK3 | PAK3 | PAK3 |
| | II | | | Y38F1A.10 | PAK3 | | | |

**Figure 1** (Continued on facing page)

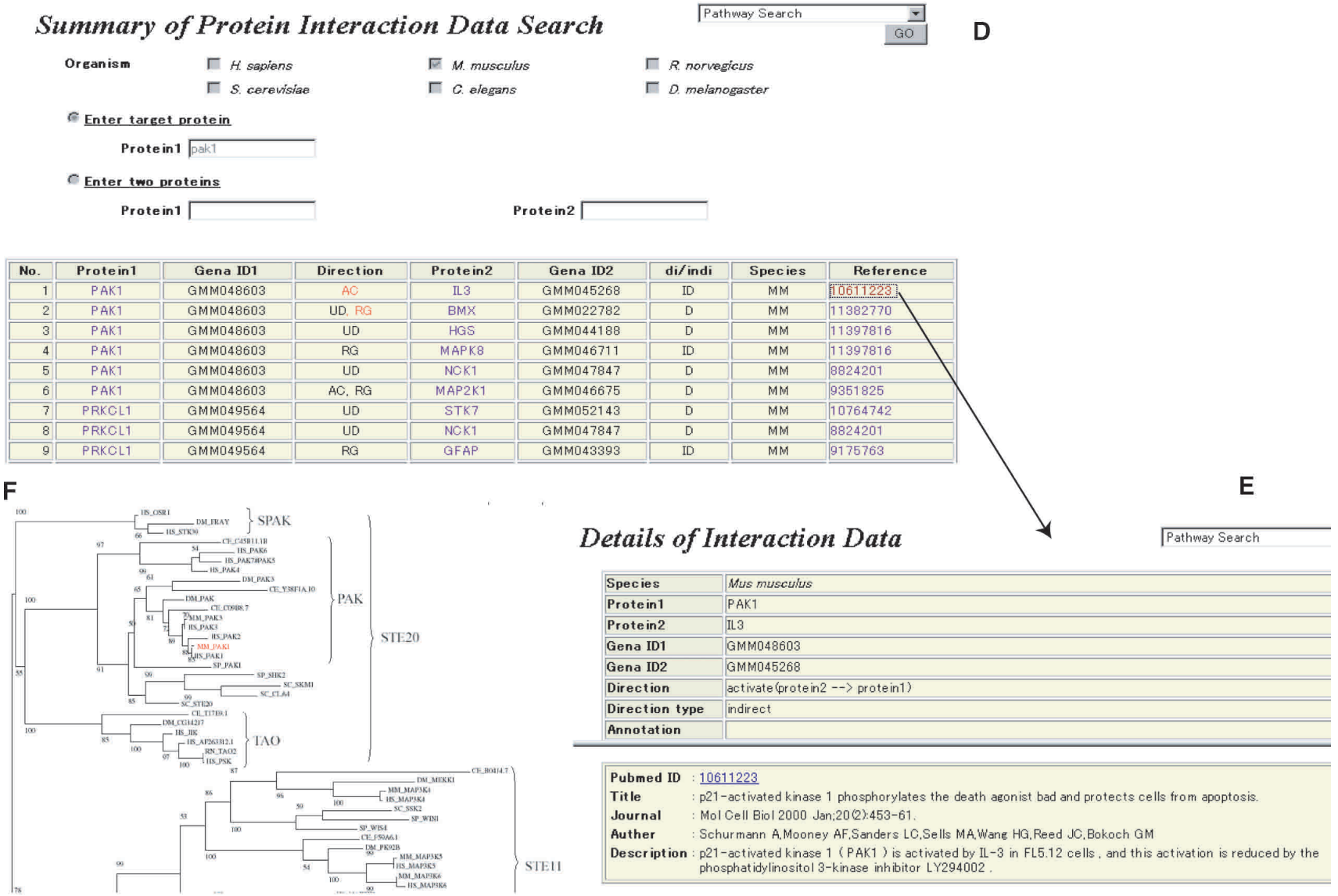



**Figure 1** Interface for Kinase Pathway Database. (*A*) Comments about functional conservation in each family. (*B*) Orthologous table. (*C*) Pathways viewer. Direct interaction and indirect interaction are respectively indicated by solid lines and dotted lines. "Activate", "inhibit", and "regulate" are distinguished by the arrow shape. The blue nodes represent the clustered proteins, which have only single-step interactions. The yellow nodes represent kinases, which have ortholog tables. Moving the cursor, orthologous proteins are depicted in red (corresponding orthologous proteins) and green (by pointing with the cursor). (*D*) Protein interaction data. (*E*) References for protein interaction data (*F*) Phylogenetic tree. The red name is the query protein name.

## Summary

We developed the Kinase Pathway Database, which contains the classification of protein kinases and their functional conservation and ortholog tables among species, protein, gene, and compound interaction data, domain information, structural information, and an automatic pathway graph image interface. The protein, gene, and compound interactions are automatically extracted from abstracts for all genes and proteins by natural-language processing using phrases. The main feature of the automated extraction of protein, gene, and compound interactions is the use of the GENA dictionary to identify as many synonyms as possible. The identification of corresponding official gene symbols and sequences or loci with extracted gene or protein names facilitates the accumulation and retrieval of interaction data and various statistical analyses using the interaction data, even when a large amount of text is processed. Furthermore, by the combination of ortholog table and interaction information, pathways can be compared between species. The precision of automated extracted protein, gene, and compound interactions is 80%–97%, which may not yet be adequate, but this system will help users to find protein, gene, and compound interactions as a method of highly functional text retrieval. Our program is still in development. The complexity of the sentences in abstracts and repeated descriptions are considered, and the totally extracted information coverage will be improved to about 70% by further refinement of our program through the use of simple phrase patterns, a gene-name dictionary method, and large text. Many protein interactions are not written with gene names, and are written with unspecified names such as family or superfamily. We are now constructing a hierarchical family-name dictionary to extract that information. The use of review articles and a new dictionary, along with improvement of our programs, will allow much more worthwhile knowledge to be gathered in the near future. The enormous extracted interaction data will be useful to determine the general aspects of signal transductions and/or protein interactions and study their evolution. The automatic extraction of relationships between two terms by full or partial parsing plus phrase and word patterns is also applicable for gathering more extended information such as disorder and protein function.


# METHODS

## Architectural Design

An overview of the Kinase Pathway Database is shown in Figure 2. The Kinase Pathway Database and GENA are implemented using PostgreSQL. In the following section, the construction of each component is discussed in detail. The table layout is written on the Web (http://kinasedb.ontology.ims.u-tokyo.ac.jp/comment/table.files/slide0001.htm).

## The Kinase Classification and Ortholog Table

The analysis sequences were collected as follows. Amino acid sequences for *S. pombe*, *S. cerevisiae*, *C. elegans*, and *D. melanogaster* were obtained as complete sequence sets from the NCBI (http://www.ncbi.nlm.nih.gov/). Those for *C.elegans* and *H. sapiens* were downloaded from WormBase (http://www.wormbase.org/) and NCBI RefSeq (http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html), respectively. Sequences from a nonredundant amino-acid database (nr-aa) that are not contained in RefSeq, and Ensemble (http://www.ensembl.org/) sequences that are not contained in either RefSeq or nr-aa were also added. Any redundancy was removed by BLAST (Altschul et al. 1997), and finally checked manually. Amino acid sequences for *M. musculus* and *R. norvegicus* were downloaded from NCBI RefSeq. It is difficult to discriminate between a receptor protein-tyrosine kinase and a nonreceptor protein-tyrosine kinase by sequence similarities alone, because these sequence homologies are high. Therefore, unknown function sequences can be distinguished by whether or not the sequence has a transmembrane part by

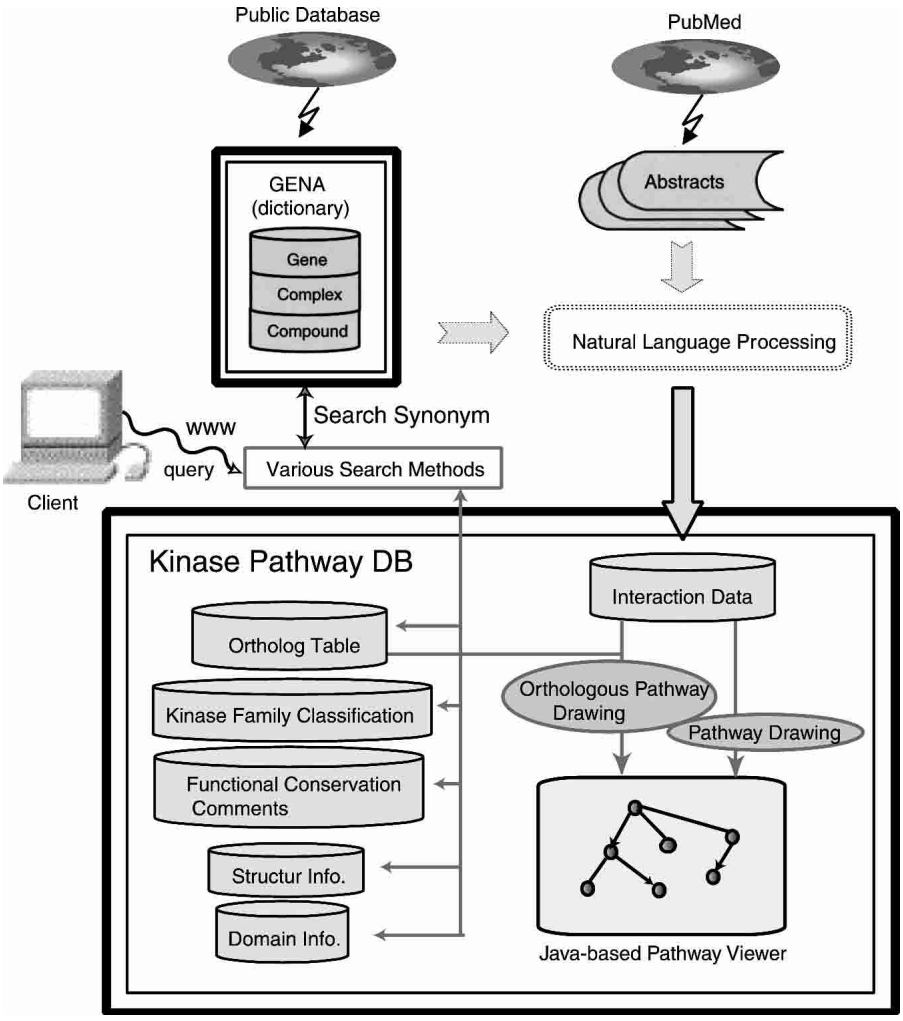



**Figure 2** Overview of kinase pathway database.

using HMMTOP (Tusnady and Simon 2001) and SOSUI (http://sosui.proteome.bio.tuat.ac.jp/sosuiframe0.html).

Pfam (Bateman et al. 2000) was used to extract the various types of kinase domains. In terms of large superfamilies such as eukaryotic protein kinases, the core group/family/subfamily classification was determined from the *S. cerevisiae* literature (i.e., the Hanks classification in Hanks and Quinn 1991). Under conditions where these sequences were subjects and other species sequences were set as queries, a BLAST search was employed, and the group/family/subfamily classification was based on the results of those similarities. CLUSTALW 1.7 (Thompson et al. 1994) or parallel PRRN (Gotoh 1996) was applied in multiple alignments. When the aligned sequences were close, a phylogenetic tree was drawn using the njplot of CLUSTALX ver. 1.81 (Thompson et al. 1997); otherwise it was constructed using the Molphy maximum likelihood method program (Adachi and Hasegawa 2000). Phylogenetic trees drawn by treeview (Page 1996) are accessible in this database.

The ortholog tables including paralog were produced from these phylogenetic trees. "Level," to express gene duplication in ortholog tables, was also determined from these phylogenetic trees. These tables can be used to search for the orthologous pathways, using the graphic pathway viewer, as is discussed in the Results and Discussion section. Further, family names and functional conservation for eukaryotic protein kinases have hierarchical definitions and can be summarized manually.

## Domains and Structural Information

Signaling pathways are expected to produce complicated networks using a variety of domains and their combinations. The domains of proteins with protein interaction information were identified by InterPro (Apweiler et al. 2000) and are stored in this database. The combinations of domains are searchable (proteins that have specified domain combinations can be retrieved).

On the other hand, protein structures that are expected to be biased in function and the distribution of structural information on pathways are attracting attention (Hegyi and Gerstein 1999). Structural information was extracted as follows. PSI-BLAST was applied for each sequence under conditions where the expected value was 0.001. The threshold expected value that was included to construct Position Specific Scoring Matrix (PSSM) on the next iteration was 0.001. The maximum number of iterations was 10. The nr-aa+pdb sequences were used as the subject database. The hit regions for PDB in all iterations were extracted, and a corresponding SCOP classification ID (Lo Conte et al. 2002) was assigned.

## Graphic Pathway Viewer

The kinase pathway database has a Java-based graphic viewer. Pathways are drawn by connecting protein interactions instead of using predefined pathways. The query protein is set in the start position, and the interacting proteins are set in the lower position in turn. When the edge (representing interaction) passes over the node (protein name), the node positions are shifted on the right or left and the node coordinates are optimized. The minimum paths between two nodes are calculated using the Program Evaluation and Review Technique (Yasuzawa and Nomura 1994).

## Protein Interaction Information: Automatic Extraction

The protein, gene, and compound interactions are extracted by natural-language processing. The subject is not only kinases, but all genes, proteins, and compounds. The methods are similar to those used in GeneWays (Friedman et al. 2001), SUISEKI (Blaschke and Valencia 2001), and the Ono program (Ono et al. 2001), but we extended their methods. Here, the definitions of protein–protein, protein–gene, and protein–compound interactions are broader than those used in the previous methods. The targets are all of the relationships between two terms, and they are not restricted to actions such as bind, activate, associate, or complex. As a result, genetic interactions were extracted as well as physical interactions. Further, we attempted to resolve the problem of synonyms by using the gene-name dictionary GENA as far as possible. GENA gathers gene and compound names from various database sites, and the corresponding sequences could be determined simultaneously, even for *H. sapiens*.

The general procedure is shown in Figure 3. The details are described in following subsections.

### *GENA: The Gene, Protein, and Compound Name Dictionary*

The identification of protein, gene, and compound names was done using GENA, which was developed in our group. Although the identification of protein and gene names from characteristic spelling is also possible, the recognition is not very precise, and the problem of synonyms remains. In addition, the identification of each corresponding protein or gene of a sequence becomes a problem. In recent years, gene names and their synonyms in public databases have been supplemented considerably. For major species, most gene and protein names, except for small variations of spelling, appear to be contained in the public databases. Because the public databases identify the correspondence between gene names and loci or sequences, various analyses of automated extracted information will be facilitated by using such sequence data.

GENA automatically and periodically gathers full official gene names, official gene symbols, gene synonyms, gene products, and family names from PomBase (http://www.sanger.ac.uk/Projects/S_pombe/), SGD (http://genome-www.stanford.edu/Saccharomyces/), MIPS (http://mips.gsf.de/projects/fungi/yeast.html), WormBase (http://www.wormbase.org/), FlyBase (http://flybase.bio.indiana.edu/), MGI (http://www.informatics.jax.org/), RGD (http://rgd.mcw.edu/), HUGO(http://

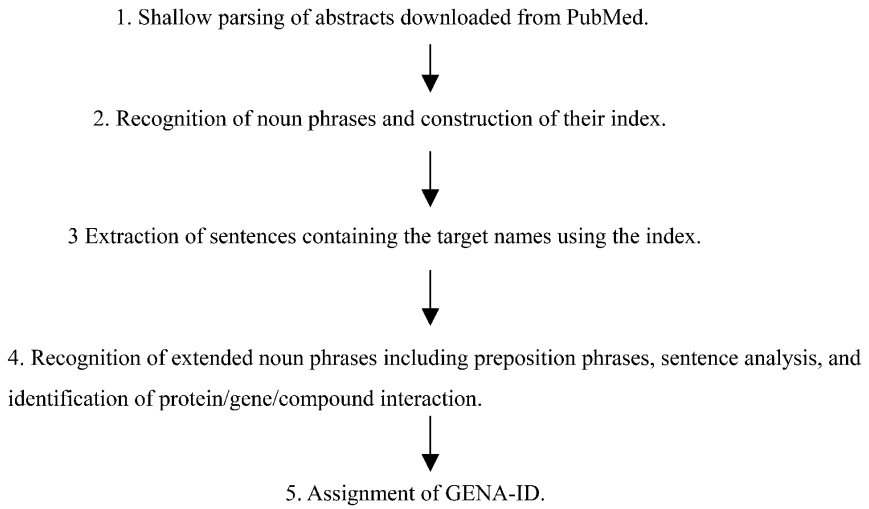



**Figure 3** Overview of the automatic interaction extraction process.

www.gene.ucl.ac.uk/hugo/), GDB (http://gdbwww.gdb.org/), GenAtlas (http://www.dsi.univ-paris5.fr/genatlas/), OMIM (http://www.ncbi.nlm.nih.gov/omim/), LocusLink (http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html), SWISS-PROT (Bairoch and Apweiler 2000), TrEMBL (http://www.expasy.ch/sprot/), and PIR (McGarvey et al. 2000). GENA also gathers various proper nouns other than gene, protein, family, and superfamily names that affect genes or proteins as compounds. For example, x-ray is not a chemical compound, but it is included in the compounds. The recognition of common genes is done using the consistency of official gene symbol and link information in the various databases. The numbers of gene, protein, and compound names used are summarized in Table 4. Because cDNA and ORF entries are included, the number of genes is greater than the number of genes with official gene names (especially in mouse). Most of these systematic names are not used for the interaction extraction. The standard official gene symbol is set to be that of the main public database for each species (e.g., HUGO for *H. sapiens*). The retrieval is done for each species. In the Kinase Pathway Database, all genes are expressed as official gene symbols. In the GENA system, new synonyms can be added and mistakes in the data gathered from outside databases have been corrected as far as possible. Both Web-based and command-line searches are available.

### *Shallow Parsing*

In the present study, abstracts were downloaded from NCBI for the period 1985–2002 using the MeSH term of each species, "saccharomyces cerevisiae," "caenorhabditis elegans," "drosophila melanogaster," "mice," "rats," and "human". For the human, rat, and mouse categories, a limit by journal name (about 200 names) was added, because there are too many abstracts on them. The MeSH term is used only for the recognition of each species.

The shallow parser FDG-Lite by Conexor (http://www.conexoroy.com/products.htm) assigns word form, base form, part-of-speech, and light syntactic representation to all targeted abstracts. FDG-Lite was originally developed by A. Voutilainen, P. Tapanainen, and T. Järvinen at the University of Helsinki. The parsed example can be seen on the Web (http://www.conexoroy.com/lite.htm).

Although most protein, gene and compound names are not stored in the FDG-Lite dictionary, correct syntactic tags and part-of speech tags are assigned for them in almost all cases. Stemming and abbreviation errors caused by protein, gene, or compound names are observed. For example, "c-fos" is wrongly recognized as "plural of c-fo". A simple program corrected these errors.

### *Recognition of Noun Phrases*

Noun phrases are extracted from information provided by the syntactic tags and part-of-speech tags with the following regular expression and are indexed. The correspondence between each tag (the meaning of tags can been seen on the Web http://www.wjh.harvard.edu/soc_help/cnxfdgen.pdf) and the following terms is easily anticipated, and thus the explicit relationship between them is omitted here. When multiple tags are assigned for a word after shallow parsing (i.e., FDG-Lite could not decide which tag is appropriate), if either tag is matched to the noun-phrase regular expression, it is recognized as a noun phrase.

**Table 4.** Summary of Dictionary Data

| Species | Number of genes | Average number of synonyms |
|---|---|---|
| *S. cerevisiae* | 7,434 | 8.6 |
| *C. elegans* | 22,718 | 4.1 |
| *D. melanogaster* | 29,891 | 4.7 |
| *M. musculus* | 80,174 | 2.2 |
| *R. norvegicus* | 5,705 | 7.5 |
| *H. sapiens* | 22,762 | 8.4 |
| Compound | 7,290 | 1.6 |

### *Regular Expression of Noun Phrases*

noun_phrase = (adjective_modifier_phrase*noun_modifier+)?[left_parenthesis (adjective_modifier_phrase*noun_modifier+)*head_noun right_parenthesis]?(head_noun)[left_parenthesis (adjective_modifier_phrase*noun_modifier+)*head_noun right_parenthesis]?

Here, special characters are used similarly to perl. In this expression, each term corresponds to the syntactic tag and part-of-speech tag pairs. For instance, when the syntactic tag is "premodifer" and the part-of speech tag is "noun", "noun_modifier" is assigned for the corresponding word.

The following noun phrases enclosed in angle brackets (< >) are recognized, and their base forms and positions in the sentence are indexed.

"*Using < leptomycin B > <we> demonstrate that <transport> of the <FXR proteins> out of the <nucleus> is mediated by the < export receptor exportin1>*".

This step is not necessary if abstracts are used only once for information extraction. The indexes are useful when gene/protein/compound names are added to GENA, because the search of sentences including gene/protein/compound names becomes a CPU time-consuming task with an increase in their names. The indexes are also used for the extraction of molecular function under development.

When prepositions are included in the target protein or compound name, they are not extracted correctly. However, the percentage of protein and compound names that include a preposition is less than 0.5% for all of those used in this study. In addition, if the other interacting protein name does not include a preposition, the protein with the preposition will be extracted as an interacting partner protein. Therefore, we ignored them in this version.

### *Identification of Protein Interactions*

Each protein, gene, and compound name registered in GENA is searched for in the index produced in step (2) of Figure 3. If the name is in the index, the corresponding parsed sentence is extracted, and the extended noun phrase and verb phrase are identified. The definition of 'extended noun phrase' is broader than is used in Figure 3 step (2); 'extended noun phrase' includes prepositional phrases and/or coordinate conjunctions + noun phrases. Further, coordinate clauses, parenthetical expressions, and relative clauses are analyzed. Matching of the template phrase patterns is investigated, and the corresponding noun phrases (expected to include protein, gene, or compound names) are extracted.

The template phrase patterns were mainly divided into two types: the noun-phrase type (e.g., the interaction between proteins A and B; protein A induced activation of protein B) and the predicate verb type (e.g., protein A interacts with protein B; protein A induces the expression of protein B). Actually, instead of only protein, gene, or compound names, extended noun phrases (noun phrases or noun phrase + prepositional phrase, or noun phrase + prepositional phrase + coordinate clause) are used in these template phrase patterns.

For example:

1. *Using < leptomycin B > <we> demonstrate that <transport of the FXR proteins out of the nucleus> can* **be mediated by** *the <export receptor exportin1>.*
2. *Our results show that <Sp1>* **plays an essential role in** *the <regulation of the differential gene expression of the HCK gene>.*
3. *The <***interaction between** *ILK* **and** *PINCH>* *has been*

*consistently observed under a <variety of experimental conditions>*.

The articles and demonstrative pronouns are ignored in the pattern matching procedure. The auxiliary verbs and verb modifier phrases are recognized as the verb phrase and are ignored in the pattern matching procedure. The verb phrases and preposition patterns and the noun phrase in bold typeface are incorporated in the template phrase patterns. The angle brackets (< >) indicate extended noun phrases, and the underlined angle brackets (< >) show noun phrases extracted by phrase pattern matching.

The detailed pattern types and sentence examples are summarized in Table 5. Actually, some restriction is imposed on noun phrases including protein names in every pattern. For example, most noun-phrase 2 (NP2s) including a protein name in the extended noun phrase (surrounded by < >) in Table 5 must be the noun phrase without prepositions or protein name + special expression (such as "activation of protein name"). More than 600 complete template phrase patterns, including differences in prepositions, are provided. Patterns are made for every similar verb/noun. To increase pattern variety, we use ambiguous pattern expressions. For example, instead of "play an essential role in" we use "play <noun phrase ended with 'role'> in", because there are various expressions such as "play an important role" and "play a crucial role". Template phrase patterns are extracted by manually checking reviews and abstracts (more than 1000 in all). Whether proteins directly or indirectly interact with other proteins, the direction of interaction (the specifications for the signal donor and acceptor proteins in the signaling cascade), and the type of interaction, such as activation, inhibition, or unknown (but signal direction is known, which indicates regulation) are also recognized. When the direction of the signal cascade is not known, "undirected" is assigned.

**Table 5. Summary of the Template Phrase Pattern Types**

| Pattern type/pattern example | Example of sentence |
|---|---|
| **Predicate verb** | |
| NP1 activate NP2 | It has recently been reported that <**rifampicin**> activates the <**glucocorticoid receptor**>. |
| NP1 bind NP2 | We were able to confirm that <**Tap**> can indeed bind <**p15** specifically both in_vivo and in_vitro>. |
| NP1 induce NP2 | These results demonstrate that <**TNF-alpha** and **IFN-gamma**> induce <expression of functional **FasL** in adenocarcinoma cels>. |
| NP1 regulate NP2 | <Transfection-mediated expression of wild-type **p53**> has been shown to negatively regulate <basal promoter activity of **MGMT** in_vitro>. |
| NP interact | Further analysis confirmed that <**Vav** and **Ly-GDI**> interact both in_vitro and in_vivo assays. |
| etc. | |
| **Predicate verb-preposition** | |
| NP1 interact with NP2 | <**Gadd45**> was also able to physically interact with <**Cdc2**>. |
| NP1 bind to NP2 | Here we report that <**PKR**> binds to <**p53**>. |
| etc. | |
| **Predicate verb-phrase** | |
| NP1 form <[a-z]* complex with NP2> | We demonstrate that <**Pap**> forms <a stable complex with **Pyk2**>. |
| NP1 involve in <regulation of NP2> | Recent animal studies indicate that <**leptin**> is involved in the <regulation of blood pressure through the **leptin receptor**>. |
| NP1 be <[a-z]* substrate of NP2> | <**HIP-55**> were <direct substrates of **caspase 3**>. |
| NP1 play <[a-z]* role in NP2> | <**Hsp90**> plays an <important role in the rapid, estrogen receptor-mediated modulation of **eNOS**>. |
| NP1 regulate <[a-z]* through degradation of NP2> | Here we report that <**Ubr1p,** the main recognition component of this pathway>, regulates <peptide import in the yeast Saccharomyces cerevisiae through degradation of **Cup9p,** a 35 kDa homeodomain protein>. |
| NP1 serve as <inhibitor of NP2> | **CHOP** serves as an <inhibitor of the activity of **C/EBP** proteins>. |
| etc. | |
| **Noun-phase** | |
| Induction of NP1 by NP2 | It was concluded that the <induction of **MCP-3** by **IFN**> is regulated differently in fibroblasts and PBMC. |
| Activation of NP1 by either NP2 or NP2′ | <Activation of **p38** by either **opsonized zymosan** or IgG-coated **SRBC**> was similar in wild-type and rac2 (−/−) cells. |
| Effect of NP1 on NP2 | The <effect of **c-myc** on down-regulation of **Fas** expression> was found. |
| Interaction between NP1 and NP2 | The <interaction between **ILK** and **PINCH**> has been consistently observed under a variety of experimental conditions. |
| Phosphorylation of NP1 through NP2 | These data suggest that SLF enhances <integrin-fibronectin-dependent tyrosine phosphorylation of **pp125FAK** through activation of **integrin**>. |
| NP1 mediated NP2 | Here we demonstrate that an enhancer element can suppress the <**P-Ph**-mediated inhibition of **yellow transcription**>. |
| Binding of NP1 to NP2 | Phosphorylation of MAPK of the critical serine residue (Ser 127) of the Drosophila transcription factor Yan depends on Mae, and is mediated by the <binding of **Yan** to **Mae** through their Pointed domain>. |
| etc. | |

NP1 and NP2 are noun-phrase including protein/gene/compound names.
< > shows the extended noun-phrase concerning information extraction.
Bold letters indicate protein/gene/compound names, and the underlined parts indicate template phrase patterns.
The articles and demonstrative pronouns are ignored in the pattern matching.

These types of interactions are determined by matching template phrase patterns. For example, "complex," "interact," "associate," "bind," and "verb of chemical modifier" are recognized as "direct interaction." "Activate" and "inhibit" are recognized as "indirect interaction," because they are not always used to indicate direct interaction. In only special cases, the order of pattern matching is fixed for the correct extraction. For example, protein A ↔ protein B and protein A ↔ protein C are correctly extracted from "interaction of protein A with protein B and protein C" using "interaction NP1 with NP2" (NP2 can include "and") instead of "interaction of NP1 and NP2". When the previous pattern is matched, the latter pattern is not tried. Blaschke and Valencia (2002) also use frames and summarized them into several types. Although they do not recognize the subject and object by noun-phrase bracketing and prepositions, they count the number of the words between the verb and protein name (example frame: [protein-name] (0–5 words) [verb] (0–5 words) [protein name]). They evaluate the extraction precision with the number of the interval words and provide the probability score for each frame. Our phrase patterns are more detailed than their frames, in order to recognize the preposition phrase and noun phrase. Because the causes of extraction errors are common to all phrase patterns, as discussed in the Appendix, the probability score is not used in our system.

In the present study, anaphora (the definition of anaphora is coreference of an expression with its antecedent) of pronouns was not resolved. The antecedent of an interrogative is always considered to be the previous noun phrase. For example, "Recently we cloned the <cDNA encoding human **Fe65L2**, which> interacts with <**Alzheimer's beta-amyloid precursor protein (APP)**>." Because < > are recognized as the subject and object of "interact", "which" is considered to point to "cDNA encoding human Fe65L2". When negative words such as "neither," "not," "never, " or "fail" are included in a verb phrase, or other species names are included in the sentence, the sentence is discarded. When "investigate," "examine," "design" and so on are included in the sentence, the sentence does not denote a fact or experimental result, so we discarded such sentences. NCBI taxonomy (http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/) is used as the list of species names. When other species names are included in the abstract, a "problematic" mark was assigned to avoid wrong recognition of other species interactions. Data marked "problematic" are searchable in protein interaction data but are not shown in the pathway viewer.

### *Assignment of GENA–ID*

From the extracted noun phrases (e.g., the underlined noun phrase enclosed in angle brackets (< >) in the previous sample sentence), protein, gene, and compound names are recognized using GENA. They are then converted to official gene symbol names, and GENA-IDs are assigned to them.

For example:

FXR proteins → official symbol:NR1H4, Gena-ID:GHS139996
exportin1 → official symbol:XPO1, Gena-ID:GHS020866

The names in the Kinase Pathway Database are unified by this official symbol and Gena-ID. In principle, uppercase and lowercase characters are not distinguished. However, for some gene names that have the same spelling as common nouns or prepositions such as "yellow" and "on", only all-uppercase letter names, names that include uppercase letters, or exact case matches are used to avoid the extraction of unrelated information and reduction of precision. (This is not applied when the gene name has all lowercase characters in GENA). Concerning some words such as 'cAMP', exact matching, including the case pattern, is used to distinguish gene names from words that have other meanings (in this case 'cAMP' must be distinguished from 'CAMP [cathelicidin antimicrobial peptide]' in human). Because the synonyms of GENA are insufficient to allow use by only complete matching, special treatment of some special marks/letters, for example, "-" (hyphen), "." (period), Greek letters, etc. and spaces was devised. Because the characteristics of gene name spelling vary considerably among species, different treatments are required for different species. When extracted and assigned gene names were abbreviated and their full names were written in the abstracts (full names and synonyms were written preceding parentheses and abbreviations were written within parentheses), we checked the consistency of GENA-IDs to identify them. When the full names or synonyms and the abbreviations in parentheses were not registered with the same GENA-ID, those interaction data were not saved in the database.

When the extracted gene name was assigned to multiple genes and the actual gene name was not distinguishable, a "problematic" mark was added and the interactions of all candidate gene patterns were stored. "Problematic" data are searchable, but are not shown in the pathway viewer. Biological experts check gene-like and protein-like noun phrases that do not completely match the GENA entries but match GENA entries through heuristics such as the treatment of hyphens, spaces, and Greek letters, and those were then registered in GENA as synonyms. Because mammalian gene names are similar to one another and there are not as many synonyms for mice and rats as there are for humans, we used dictionary information for other species. That is, gene-like or protein-like noun phrases (e.g., '-ase') that are not registered as GENA entries for targeted species but are registered for other species, are also checked by biological experts and registered in GENA as synonyms if proven true.

## ACKNOWLEDGMENTS

We thank Drs. T. Takai and K. Nakai for their helpful discussions, Mrs. K. Kodama and S. Asahi at Hitachi ULSI Systems for helping us by programming the database, and Ms. A. Nakata, Ms. Y. Shidahara, and Mr. K. Yamada for their careful reading of a large number of abstracts. This work was supported in part by a Grant-in-Aid for Scientific Research on Priority Areas (C) Genome Information Science from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

## APPENDIX

In this Appendix, we discuss the cause of errors in the automatic protein interaction extraction and compare our method with other natural-language extraction methods.

### The Precision of Automatically Searched Abstracts

In Table 2 in the main text, the precision with GENA errors is lower for *H. sapiens* and *D. melanogaster* than for other species. This is because there are some names that are indistinguishable from common nouns, adjectives, and prepositions. In addition, some abbreviations for gene or protein names, compound names, reagent names, and equipment names share the same spelling. Except for recognizing them by using uppercase and lowercase letters or full names, they can hardly be distinguished in the sentence without prior background knowledge.

The following errors (false positive) were detected in extracted gene, protein, or compound interactions.

*1. Failure in Sentence Parsing*

1.1 Failure of syntactic and morphological analysis by FDG-Lite (e.g., the past participle is wrongly recognized as "verb" instead of "adjective", and vice versa.)

1.2 Stemming error by FDG-Lite (e.g., the final "s" of a gene name is wrongly removed as that of a plural form.)

1.3 Failure of sentence analysis: wrong recognition of subject, object, or verb phrase, including wrong recognition of noun phrase, coordinate conjunction, parenthetical expression, and relative clause. For example, in '*<Protein A, a target hormone for Protein B>, has been shown to regulate <Protein C>.*', the angle brackets (< >) represent an extracted noun phrase. In this program, the interactions Protein A → Protein C and Protein B → Protein C are extracted. The latter is wrong. The noun phrases surrounded by commas cannot be distinguished as to whether they constitute a parenthetical expression or an expression in apposition.

*2. Failure in Recognizing Gene Name (Errors Caused by GENA)*

2.1 A gene/compound name that is the same as a common noun, such as "Set" in mouse.

2.2 A gene or compound name that is the same as another gene or compound name.

*3. Failure in Phrase Pattern Recognition*

3.1 Shortage in variety of phrase patterns. For example, "*the interaction between protein A and protein B is enhanced by protein C*", shows the relations among more than two items, but we do not provide multiterm interaction phrase patterns for more than two items in this version.

3.2 The phrase pattern is provided, but is used for another meaning. For example, in "*The observation that FUR4 was closely linked to GAL10, one of the three genes forming the gal cluster, we could determine precisely the position of the gene on chromosome II.*", 'link to' is used for the description of protein, gene, or compound relations, but it also has other meanings. Here, it describes the chromosome position, so the extraction of interaction between FUR4 and GAL10 is wrong.

*4. Others*

4.1 The name of another organism is extracted.

4.2 Failure in the recognition of a negative word.

The occurrence of extraction errors caused by #1.1 is quite low, probably less than 1%. Except for the errors caused by GENA, more than half of the errors are parsing errors [#1.3, mainly failures in coordinate clause analysis and noun-phrase (object-phrase) recognition]. Most of the noun-phrase (object-phrase) recognition errors described in #1.3 are avoidable if analysis of the constituent words of noun phrases (especially prepositional phrases) is added. For example, consider the sentence, "<p38> was also activated by <cisplatin with similar kinetics as JNK>." With our program, the wrong interaction "JNK → p38" and the correct interaction "cisplatin → p38" are extracted. However, the classification conditions by the analysis of constituent words of noun phrase would be numerous, so we did not check them in this version. Only frequently occurring errors were finally removed by a simple program.

Blaschke and Valencia (2002) reported precision which differs with each phrase-pattern type. They use frames and the distance between protein-name and verb (see Methods). With our results, a statistically significant difference in the precision with phrase-pattern types (Table 5 in Methods) is not observed, because the problem of parsing and noun-phrase recognition is common to all phrase patterns. The precision difference in the point of #3.2 is occasionally observed among instances of "verb" or "noun" instead of "phrase-pattern". For example, "link" is used for the description of "chromosome position" in five sentences out of 47 sentences which are used for interaction extractions. Unfortunately, we do not have enough extracted data of each phrase-pattern for further discussion, because they do not appear frequently in abstracts.

## The Recall of Automatically Searched Abstracts

Calculation for recall is a laborious task, because protein–protein interaction is not frequently described in abstracts. Five hundred abstracts (200 for *S. cerevisiae* and 300 for *H. sapiens*) were checked manually, and the results were compared with automatically extracted results as described in the main text.

For *S. cerevisiae*, 93 (90) interactions were manually extracted, and among these, 47 (44) included names that were unregistered or unrecognizable in GENA (such as family names and the existence or nonexistence of a number of the proteins such as 'PYK' and 'PYK1'). The above numbers in parentheses are the results without compounds. Some interactions with names that are unregistered or unrecognizable in GENA are also extracted, but we did not evaluate them in this study. To separate the problem of name detection caused by GENA and parsing/phrase-pattern error, we evaluated 46 (46) interactions whose names are identified by GENA out of 93 (90) interactions. As a result, 12 (12) were extracted by our program. The recall rate (=true positive/[true positive + false negative]) is 26% (26%). About 40% of the false negatives are due to the complexity of the sentences (e.g., the problems of anaphora by relative pronouns, coordinate clauses, and parenthetical expressions). About 10% is due to insufficient sentence analysis, which will be resolved by improvement of our program. The rest of the false negatives are mainly due to the lack of template phrase patterns. Anaphora covering multiple sentences seems to be a trivial problem, because only two interactions are extracted as descriptions that were covered in two sentences.

For *H. sapiens*, 349 (244) interactions were extracted manually, and among them 196 (124) included names that are unregistered, unrecognizable names in GENA. Accordingly, the names of 153 (120) interactions are registered in GENA. Although some of the interactions of these unregistered or unrecognizable names were extracted, we did not evaluate them here. As a result, 38 (27) out of 153 (120) were extracted by our programs. The recall rate is 25% (23%). About 30% of the unregistered names are compound names, about 10% are gene names, and about 60% are family-level or unspecified gene names. About 6% of the gene names that were recognized as genes registered in GENA manually were not found automatically by simple heuristics. For example, if only 'c-kit' is registered in GENA, it is difficult to recognize 'c-kit receptor' as a 'c-kit' synonym, because similar name patterns such as 'insulin receptor' and 'insulin' are not correct. About 40% of the false negatives are due to sentence complexity. About 10% of the false negatives are caused by insufficient analysis of coordinate clauses and parenthetical expressions; the rest are mainly due to the lack of template phrase patterns. Interestingly, although the signaling cascade of *H. sapiens* is more complicated than that of *S. cerevisiae*, the complexity of sentences seems to be on the same level.

## Comparison With Other NLP Methods

The appropriate extraction standard of protein–protein interaction depends on each purpose. For example, the extraction of proteinA–proteinB interaction from the sentence "We investigated the interaction of protein A and protein B" is wrong, because the sentence does not indicate the experimental result. However, if a user wants to search abstracts related to proteinA–proteinB, the extraction may be useful. Handling the negative phrase and mutant information also depends on their purpose. There are many detailed standards, and they are not clearly described in most NLP-related papers. As imposed conditions increase, the recall and precision will be

decreased. Therefore, the simple comparison with other NLP systems based on the precision/recall described in papers may be inappropriate.

There are mainly two kinds of protein, gene, or compound relation extraction. One is the co-occurrence of two proteins, genes, or compounds in text; the other is the use of phrase patterns, as was done in this study. The former seems to be useful for well known interactions that co-occur in the text (Stapley and Benoit 2000; Jenssen et al. 2001), but more detailed information such as the direction of interaction cannot be extracted. Ono's program (Ono et al. 2001), GeneWays (Friedman et al. 2001), and SUISEKI (Blaschke and Valencia 2002) use phrase patterns. Our program is similar to them. Our program differs from Ono's program mainly in sentence analysis and the variety of phrase patterns. The Ono method basically uses only the relative position of protein names and the pattern verb. Although that may not create a problem when only extracting patterns related to direct interaction, such as "bind," "interact," "associate," or "complex," its application to the extended phrase patterns that include genetic interaction descriptions may decrease precision. Accordingly, we analyze the sentence further to recognize the subject and object. For example, consider the sentence, "*Protein A induced the intracellular translocation of protein B whose function is associated with that of protein C*." With the Ono method, the wrong interaction "protein A and protein C" is extracted by the relative positions of the underlined term and the protein names. With our program, the correct interaction "protein A → Protein B" is extracted, because protein A is recognized as the subject of "induce." The SUISEKI system uses phrase patterns with probability score and distance (number of words) between protein names and verb. Although high precision and recall are obtained for short distances, this method may not be appropriate for complex sentences that contain parenthetical expressions. GeneWays uses phrase patterns and a full and/or partial parser, which extract interactions and relations between process names as well as between protein, gene, and compound names. Interaction is extracted with additional information such as semantic action category (activate, attach, break bond, etc.). The recognition of gene and protein names uses both rules and external knowledge sources, but gene synonyms and the relationship between gene names and the corresponding sequences and synonyms seem not to be considered. Unfortunately, GeneWays is not open to the public, so we cannot compare it with our system. Some other systems that use full and/or partial parsing and phrase or word patterns have been developed, for example, EDGAR (Rindflesch et al. 2000) and the method described by Sekimizu et al. (1998). EDGAR uses contextually identified gene and cell names. The Sekimizu system, on the other hand, does not offer gene and protein names, but extracts the noun phrases of the subjects and objects of selected verbs that are expected to represent protein interaction (such as interact, bind, etc.). Many of those systems are under development and have not been evaluated sufficiently, or they are not open to the public.

## REFERENCES


Adachi, J. and Hasegawa, M. 2000. MOLPHY Version 2.3: Programs for phylogenetics, ver. 2.3. Institute of Statistical Mathematics, Tokyo.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M.D., et al. 2000. InterPro—An integrated documentation resource for protein families, domains, and functional sites. *Bioinformatics* **16:** 1145–1150.

Bairoch, A. and Apweiler, R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28:** 45–48.

Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L., and Sonnhammer, E.L.L. 2000. The Pfam protein families databases. *Nucleic Acids Res.* **28:** 263–266.

Blaschke, C. and Valencia, A. 2001. The potential use of SUISEKI as a protein interaction discovery tool. *Genome Inform. Ser. Workshop Genome Inform.* **12:** 123–134.

Blaschke, C. and Valencia, A. 2002. The frame-based module of the SUISEKI information extraction system. *IEEE Intelligent Systems II* 2–8.

Friedman, C., Kra, P. Yu, H., Krauthammer, M., and Rzhetsky, A. 2001. GENIES: A natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics* **17:** 74–82.

Gotoh, O. 1996. Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. *J. Mol. Biol.* **264:** 823–838.

Hanks, S.K. and Quinn, A.M. 1991. Protein kinase catalytic domain sequence database, identification of conserved features of primary structure and classification of family members. *Methods Enzymol.* **200:** 38–62.

Hegyi, H. and Gerstein, M. 1999. The relationship between protein structure and function: A comprehensive survey with application to the yeast genome. *J. Mol. Biol.* **288:** 147–164.

Hunter, T. 2000. Signaling–2000 and beyond. *Cell* **100:** 113–127.

Hunter, T. and Plowman, G.D. 1997. The protein kinases of budding yeast: Six score and more. *Trends Biochem Sci.* **22:** 18–22.

Jenssen, T.K., Laegreid, A., Komorowski, J., and Hovig, E. 2001. A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet.* **28:** 21–28.

Kanehisa, M., Goto, S., Kawashima, S., and Nakaya, A. 2002. The KEGG databases at GenomeNet. *Nucleic Acids Res.* **30:** 42–46.

Krauthammer, M., Kra, P., Iossifov, I., Gomez, S.M., Hripcsak, G., Hatzivassiloglou, V., Friedman, C., and Rzhetsky, A. 2002. Of truth and pathways: Chasing bits of information through myriads of articles. *Bioinformatics* (Suppl.) **17:** 249–257.

Lo Conte, L., Brenner, S.E., Hubbard, T.J., Chothia, C., and Murzin, A.G. 2002. SCOP database in 2002: Refinements accommodate structural genomics. *Nucleic Acids Res.* **30:** 264–267.

Manning, G., Plowman, G.D., Hunter, T., and Sudarsanam, S. 2002. Evolution of protein kinase signaling from yeast to man. *Trends Biochem Sci.* **27:** 514–520.

McGarvey, P.B., Huang, H., Barker, W.C., Orcutt, B.C., Garavelli, J.S., Srinivasarao, G.Y., Yeh, L.S., Xiao, C., and Wu, C.H. 2000. PIR: A new resource for bioinformatics. *Bioinformatics* **16:** 290–291.

Morrison, D.K., Murakami, M.S., and Cleghon V. 2000. Protein kinases and phosphatases in the *Drosophila* genome. *J. Cell Biol.* **150:** 57–62.

Ono, T., Hishigaki, H., Tanigami, A., and Takagi, T. 2001. Automated extraction of information on protein–protein interactions from the biological literature. *Bioinformatics* **17:** 155–161.

Page, R.D. 1996. TreeView: An application to display phylogenetic trees on personal computers. *Comput. Appl. Biosci.* **12:** 357–358.

Plowman, G.D., Sudarsanam, S., Bingham, J., Whyte, D., Hunter, T. 1999. The protein kinases of *Caenorhabditis elegans:* A model for signal transduction in multicellular organisms. *Proc. Natl. Acad. Sci. USA* **96:** 13603–13610.

Rindflesch, T.C., Tanabe L., Weinstein J.N., and Hunter L. 2000. EDGAR: Extraction of drugs, genes, and relations from the biomedical literature. *Pac. Symp. Biocomput.* 517–528.

Schacherer, F., Choi, C., Gotze, U., Krull, M., Pistor, S., and Wingender, E. 2001. The TRANSPATH signal transduction database: A knowledge base on signal transduction networks. *Bioinformatics* **17:** 1053–1057.

Sekimizu, T., Park, H.S., and Tsujii, J. 1998. Identifying the interaction between genes and gene products based on frequently seen verbs in Medline abstracts. *Genome Inform. Ser. Workshop Genome. Inform.* **9:** 62–71.

Smith, C.M. 1997. The protein kinase resource and other bioinformation resources. *Prog. Biophys. Mol. Biol.* **71:** 525–533.

Stapley, B.J and Benoit, G. 2000. Biobibliometrics: Information retrieval and visualization from co-occurrences of gene names in Medline abstracts. *Pac. Symp. Biocomput.* 529–540.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. ClustalW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., and Higgins, D.G. 1997. The CLUSTAL_X windows interface: Flexible

strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25:** 4876–4882.
Tusnady, G.E. and Simon, I. 2001. The HMMTOP transmembrane topology prediction server. *Bioinformatics* **17:** 849–850.
Xenarios, I., Salwinski, L., Duan, X.J., Higney, P., Kim, S.M., and Eisenberg, D. 2002. DIP, the Database of Interacting Proteins: A research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* **30:** 303–305.
Yasuzawa, Y. and Nomura, K. 1994. *Operation Research: Its Technique and Application.* pp. 137–54. Corona Publishing Co., Tokyo, Japan.

## WEB SITE REFERENCES

http://www.wjh.harvard.edu/soc_help/cnxfdgen.pdf; Conexor tag information Web site.
http://www.conexoroy.com/products.htm; Conexor Web site.
http://www.ensembl.org/; ENSEMBL Web site.
http://flybase.bio.indiana.edu/; FlyBase Web site.
http://gdbwww.gdb.org/; GDB Web site.
http://gena.ontology.ims.u-tokyo.ac.jp/search/servlet/gena; GENA Web site. A. Koike, T. Takai, and T. Takagi; a partial database is open to the public.
http://www.dsi.univ-paris5.fr/genatlas/; GENATLAS Web site.
http://www.gene.ucl.ac.uk/hugo/; HUGO Web site.
http://mips.gsf.de/projects/fungi/yeast.html; MIPS Web site.
http://www.ncbi.nlm.nih.gov/Entrez/; NCBI Entrez Web site.
http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html; NCBI locus link Web site.
http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/; NCBI taxonomy Web site.
http:// www.ncbi.nlm.nih.gov/; NCBI Web site.
http://www.ncbi.nlm.nih.gov/omim/; OMIM Web site.
http://www.sanger.ac.uk/Projects/S_pombe/; PomBase Web site.
http://genome-www.stanford.edu/Saccharomyces/; SGD Web site.
http://sosui.proteome.bio.tuat.ac.jp/sosuiframe0.html; SOSUI Web site.
http://www.expasy.ch/sprot/; SWISS-PROT and TrEMBL Web site.
http://www.wormbase.org/; WormBase Web site.
www.celera.com; Celera Web site.
http://kinasedb.ontology.ims.u-tokyo.ac.jp; Kinase Pathway Database Web site.
http://www.informatics.jax.org; MGI Web site.
http://rgd.mcw.edu; RGD Web site.
http://www.conexoroy.com/lite.htm; Conexor FDG Lite parser information Web site.

*Received September 24, 2002; accepted in revised form March 26, 2003.*