DEVELOPMENT OF A COMPUTER SYSTEM FOR THE AUTOMATED RECONSTRUCTION OF MOLECULAR-GENETIC INTERACTION NETWORKS

Aman E.E.1, Demenkov P.S.2, Pintus S.S.3, Nemiatov A.I.1, Apasieva N.V.1, Dubovenko E.A.1, Ignatieva E.V.1, Podkolodny N.L.1, Ivanisenko V.A.*1
1 Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia; 2 Institute of Mathematics, SB RAS, Novosibirsk, 630090, Russia; 3 Novosibirsk State University, Novosibirsk, 630090, Russia
* Corresponding author: e-mail: salix@bionet.nsc.ru

Key words: information extraction, text mining, molecular-genetic interaction networks

SUMMARY

Motivation: The body of data related to molecular-genetic objects reported in the literature and stored in the computer databases keeps growing substantially over the years. The benefit of using computer assisted tools to overview and analyze bulky information is that it allows for the integration of information extracted from the databases with that retrieved from the automated processing of published reports.

Result: We are creating a system enabling us to detect interactions involving genes, proteins, low molecular weight substances, and human diseases. The broad basis is the information we retrieved from the databases about the interactions of these objects, protein and gene functional characteristics, gene expression profiles and also automated analysis of PubMed abstracts that relies on statistical approach. By using a combination of text- and data-mining, robust results can be achieved for the detection and systematization of information pertaining to molecular-genetic object interactions.

INTRODUCTION

Molecular biologists owe credit to the current achievements of computer-assisted reconstruction of gene networks. The respectable precision and higher recall of methods is, indeed impressive. The software GeneScene (Leroy, Chen, 2002) and MedScan (Daraselia et al., 2004) were developed to predict protein-protein and protein-gene interactions using text-mining. The software implements algorithms on the basis of an in-depth linguistic analysis of published texts. The authors of the systems have reported high precision, over 90 %, and recall of about 20 %. The systems based on linguistic text analysis yield very accurate estimates for the interaction detected in the text under consideration. However, their accuracy is restricted by the credibility of the analyzed paper. Low efficiency is another drawback.

Cooper have generated a simple system that predicts protein-protein interactions by using text-mining methods based on the search of particular words that describe the interactions, of protein name synonyms, and simple rules for their occurrence in the retrieved article. The system is admirably efficient, its accuracy is about 60 % (Cooper, Kershenbaum, 2005).
With this in mind, it appeared expedient to develop a combined approach that would enable us to use the shallow parsing of the PubMed abstracts. The envisaged combination will also include data on protein function, their intercellular location, gene expression profiles, among others, which can determine object interaction occurrence. The novelty of our approach is to use neural network algorithm for combining data obtained from abstracts with functional-structural data, which allow us to improve the prediction exactness. The accuracy of the automated reconstructed gene networks will be increased through a set of experimentally supported interactions retrieved from the GENE, KEGG and other publicly available databases. The system will give the user the opportunities to analyze the networks for interactions associated with biological processes and diseases of interest.

METHODS AND ALGORITHMS

We have developed a complex algorithm to detect real interactions between molecular-genetic objects. The algorithm was grounded on analysis of the PubMed abstracts and data on the functional-structural characteristics of the objects. The algorithm is schematically represented in Fig. 1.

![Figure 1. Schematic representation of the molecular-genetic interaction recognition algorithm.](image)

Thesauruses (dictionaries of synonyms) were generated for databases of genes (NCBI GENE), proteins (SwissProt), substances (ChEBI), diseases (PharmGKB), and species (NCBI Taxonomy). Each entry in the dictionary contained a list of synonyms for each and every object, also references to the entries in the databases from which the information was retrieved.

To increase parsing accuracy of the abstracts, we listed the linking words whose occurrence in the sentence, along with the names of two objects, were evidences of interaction between them.

The collected dictionaries were used for parsing the PubMed abstracts. Four matrices for the names of objects and linking words occurrence in the texts were derived from the texts. Matrix building obeyed one of following rules:

- Two object names must occur in the same article’s name.
- Two object names must occur in the same abstract.
Two object names must occur in the same sentence.
Two object names must occur in the same sentence and be joined by linking words.

A back propagation network (BPN) was used to predict object interactions. To train the neural network, experimentally supported data were retrieved from the DIP, Gene, IntAct, MIPS, MINT, DrugBank, GRID, KEGG and other relevant databases.

Besides the object name occurrence matrices, we exploited additional information about the objects. The information included expression profiles, gene coregulation, object location in the cell, protein functions, and certain biological processes. The idea of training of the neural network was to make it capable of answering the question: “Can two molecular-genetic objects really interact?”

Taken together, the results provided by the BPN and the experimentally supported data were used for building the matrices for the pairwise interactions. A fragment of the matrix derived interaction network can be reconstructed depending on the parameters chosen by the user. These parameters may include concrete objects, network size, network degree of relationship among others.

RESULTS

Dictionaries of object name synonyms. Dictionaries of object name synonyms were compiled from the SwissProt database. To list the synonyms, card fields ID, Synonyms were utilized. Synonym lists from two cards were joined when the synonym sets in two cards overlapped by more than 50%. To compile the dictionary of gene name synonyms, we took advantage of the ID, Gene name, Gene description, and Gene aliases fields, also of the NCBI GENE database. The compilation of the dictionary of substance name synonyms was based on the ChEBI database data and its Name, IUPAC Name, Synonym fields. Information about substance hierarchy was also included in the dictionary. The dictionary of disease name synonyms was compiled from the information in the PharmGKB database using its Name, Alternate names, Related gene, Related drug fields. The two latter fields contain information about genes and substances (drugs), associated with diseases. To compile the dictionary of words that are synonyms of species names, we used the NCBI Taxonomy database, fields Name, GeneBank Common name, and Common name.

All the synonyms shorter than 3 characters and those overlapped with the authorized English dictionary (20,000 words) were omitted.

<table>
<thead>
<tr>
<th>Table 1. The statistics for the dictionaries of synonyms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proteins</td>
</tr>
<tr>
<td>----------</td>
</tr>
<tr>
<td>Number of entries</td>
</tr>
<tr>
<td>Terms (synonyms) per entry</td>
</tr>
<tr>
<td>Average length per term (words)</td>
</tr>
<tr>
<td>Ambiguous terms</td>
</tr>
<tr>
<td>Terms overlapping with English</td>
</tr>
</tbody>
</table>

The dictionary for the gene names contains the maximum entry names (Table 1, row 1). Diseases dictionary is the richest containing more than 9 terms per entry (Table 1, row 2). The compactness of Substances and Genes dictionaries is the highest, 1.35 and 1.56 words per term respectively; the least is Proteins dictionary with more than 3 words per term (Table 1, row 3). On the one hand long terms can be recognized more precisely due to lesser ambiguity, but on the other hand longer terms tend to vary significantly.
texts. The number of ambiguous terms, i.e. the number of those associated with multiple object IDs, is very large for Substances, it is smaller for Proteins and Genes, and still smaller for Diseases (Table 1, row 4). The percent of term overlapping with English dictionary in Diseases is maximum among all the dictionaries (exceeding 0.5 %), its minimum estimate is for Species (less than 0.03) (Table 1, row 5). The term ambiguity is a source of many recognition errors. To disambiguate terms in the same dictionary the text can be analyzed in order to find keywords corresponding to a definite object. To solve ambiguity of terms in different dictionaries we can search nouns like “gene” or “protein” and verbs “express” or “catalyze” determining object type close to term mention.

**Linking word dictionary.** This dictionary was compiled manually using expertise analysis of the PubMed abstracts. The dictionary contains more than 100 words and word-forms occurring in sentences that describe interactions between objects, for example “regulate expression” or “interact with”.

The next steps of the algorithm are now under development. This process is ongoing.

**ACKNOWLEDGEMENTS**

Work was supported in part by Russian Foundation for Basic Research (Nos 04-01-00458, 05-04-49283, 06-04-49556), the CRDF Rup2-2629-NO-04, the State contract No. 02.434.11.3004 of 01.04.2005 and No. 02.467.11.1005 of 30.09.2005 with the Federal Agency for science and innovation “Identification of promising targets for the action of new drugs on the basis of gene network reconstruction” federal goal-oriented technical program “Study and development of priority directions in science and technique, 2002-2006, Interdisciplinary integrative project for basic research of the SB RAS No. 115 “Development of intellectual and informational technologies for the generation and analysis of knowledge to support basic research in the area of natural science”.

**REFERENCES**