BirdsAnts: Bringing Informative Rules from a Database System, Aimed at Novel Targets Search

Motoi Tobita tobita-m@reprori.jp
Ken Horiuchi horiuchi-k@reprori.jp
Kenji Araki araki-k@reprori.jp
Masashi Nemoto nemoto-m@reprori.jp
Tetsuo Nishikawa nisikawa-t@reprori.jp

Reverse Proteomics Research Institute Co., Ltd., 2-6-7 Kazusa-Kamatari, Kisaradu-si, Chiba 292-0818, Japan

Keywords: data visualization, correlation data, proteins-drugs interaction, clustering, data mining

1 Introduction

A large portion of time in the drug discovery process is consumed in an analysis of complex mass data which often include correlation data such as proteins-drugs binding data and mRNAs-'expression libraries' expression-strength data. Due to their size and a complicated network between those correlation data, it is a formidable task to discover meaningful information from mass data and to acquire novel knowledge of interests hidden in them. We have developed a web-based software, "BirdsAnts", in order to help researchers analyzing complex mass data. The software demonstration is given using proteins-drugs complex data obtained from protein data bank(PDB) [3], augmented by annotations obtained either by calculations or from public sources, to exhibit how the software aid in finding target proteins and corresponding drug candidates, while in this abstract, key technologies behind "BirdsAnts" is described.

2 Functionalities

"BirdsAnts" is designed to provide information packed into easily human-understandable amount, regardless of the number of data. In other words, when the number of data is small, detailed information for each datum is shown, while the information contents get packed as the number of data increases, thus keeping the amount of information displayed relatively constant, that helps users easily understand and handle information. The above feature is realized by a combination of three technologies; 1: Data packing technology based on clustering, 2: Data abstraction technology, 3: Automatic selection technology of the order of data packing and the order of data abstraction. The former two of those technologies are briefly explained in Section 2.1 and 2.2.

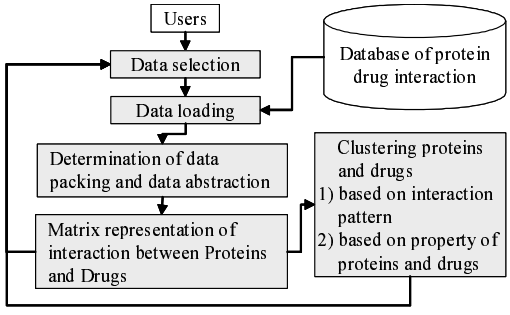

Figure 1: An overview of the systems.

Also, "BirdsAnts" features data clustering based both on correlation data and on property data. As an example, the correlation data can be proteins-drugs binding affinities, and property data can be the molecular weight of drug compounds and gene ontology of proteins. This feature is best presented in our software demonstration. The capability of clustering, based both on correlation data and on property data, makes it possible to discover properties that best explain correlation data. This technology lays the foundation of a data mining engine.

Figure 1 shows the overview of the system. Selected data by users are loaded from the database of proteins-drugs interactions [1]. Data are displayed in a matrix representation with the order of data packing and of data abstraction determined depending on the number of data. The clustered data can also be shown in the matrix representation window.

Current version of "BirdsAnts" runs on a web browser which supports a macromedia FLASH player 6.0 or higher [2]. Data are locally stored in XML and CSV files written in a specified format.

2.1 Data Packing Technology

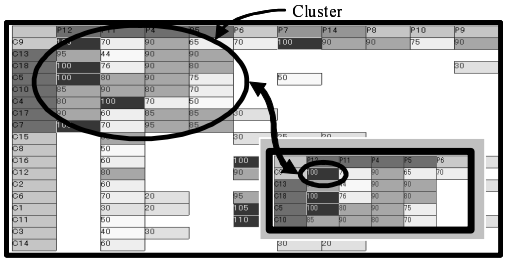

Figure 2: Two forms of data packing.

When correlation data, such as proteins-drugs interactions, are displayed in a table form, as the number of proteins and drugs increases, so does the table size. When there are more than hundreds of proteins or drugs, it is almost impossible to understand those data. In such cases, many cells of similar values can be represented by a cell, resulting in reduced size of the table. The discovery of similar cells is realized by clustering calculations. Figure 2 shows an example of data visualization in unpacked and packed form. A cell in the larger table represents a correlation data, whereas a cell in the smaller table (in a box) represents the average value of a cluster, which includes one or more correlation data as components. These two forms of data visualization can be switched by a single mouse click.

2.2 Data Abstraction Technology

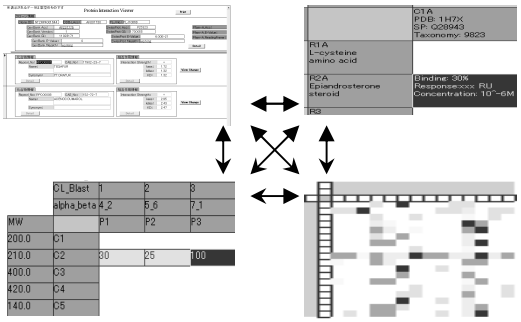
Figure 3: Four forms of data abstraction and their transition.

"BirdsAnts" is also featuring in showing information in a cell with different orders of data abstraction. When detailed information about a particular correlation datum is of interest, those can be viewed in a free format as in Figure 3, upper-left [1]. Meanwhile, when we would like to see several correlation data at a time, "BirdsAnts" selects only important information and pack them within a cell as in the upper-right of the figure. As the number of data increases, we strengthen the order of abstraction as in the lower-left and lower-right of the figure. In the lower-left of the figure, only the small amount of character type data is shown as text information, and other type of data is represented by a color of the cell. The lower-right of the figure contains information represented only by colors, while this form of data visualization is capable of displaying large amount of data. Data can be classified by a color pattern, and parts of data which share a color can be visualized in a less abstract format.

3 Future Developments

We plan to integrate the capability of "BirdsAnts" as a part of client-server system to enhance capabilities to handle larger amount of data. Also, a function to retrieve connections among different types of correlation data will be implemented. As an application of "BirdsAnts", it is our goal to develop a database system which contains proteins-drugs binding affinities measured in-house, and related annotation data. Using these data, we aim at finding new drug-target proteins. This work was supported by a grant from NEDO Project of the Ministry of Economy, Trade and Industry of Japan.

References

[1] Nemoto, M., Araki, K., Horiuchi, K., Tobita, M., and Nishikawa, T., An integrated database of interaction between human proteins and commonly used drugs, *Genome Informatics*, 14:599–600, 2003.

[2] `http://www.macromedia.com/`

[3] `http://www.pdb.org/`