Chem-Bio Informatics Journal, Vol. 6, No. 1, pp.17-28(2006)

BirdsAnts: A protein-small molecule interaction viewer

Motoi Tobita[1, 2*], Ken Horiuchi[1], Kenji Araki[1], Masashi Nemoto[1], Hiroyasu Shimada[1], and Tetsuo Nishikawa[1*]

[1]*Informatics Department*, *Reverse Proteomics Research Institute Co., Ltd.*, *2-6-7 Kazusa-Kamatari, Kisarazu-shi, Chiba 292-0818, Japan*
[2]*Current Address: Hitachi, Ltd., Advanced Research Laboratory, 1-280, Higashi-Koigakubo Kokubunji-shi, Tokyo 185-8601, Japan*

**E-mail: 1. nisikawa-t@reprori.jp, 2. toby@rd.hitachi.co.jp*

(Received November 22, 2005; accepted March 6, 2006; published online March 27, 2006)

Abstract

A data-matrix viewer, BirdsAnts, is developed for the effective visualization of large scale protein-small molecule interaction data. One of two design concepts is the simultaneous visualization of three kinds of data: data pertaining to proteins, those to small-molecules, and those to interactions between them. The other is switching of global and local views of data back and forth. To evaluate BirdsAnts, a data-matrix was generated by processing a protein-small molecule interaction database. Analyzing the matrix, two known facts were rediscovered. First rediscovery was the knowledge that each function of vitamin B12 is exercised by interaction to a different protein. Second was that cyclooxygenase inhibitors were correctly classified into known classes by two-step clustering calculations. The results of the evaluation show that BirdsAnts has future potential for the effective visualization of protein-small molecule interaction data.

Key Words: visualization, viewer, matrix, interaction, protein, small molecules, drug discovery, clustering, vitamin B12, COX

Area of Interest: Genome Wide Experimental Data Analyses

Chem-Bio Informatics Journal, Vol. 6, No. 1, pp.17-28(2006)

1. Introduction

Various kinds of biological data that are related to the understanding of the biology of human being are emerging[1]. Among them, microarray data provide some insight into the expression of genes[2]. Protein network data relates to the understanding of functions of organisms at the molecular level[3]. Protein-small molecule interaction data may be linked to the discovery of a novel drug because the pharmacological actions of most small molecules are the results of their interactions with proteins. Here, the term “interaction” means binding of a small molecule to a protein that causes change in protein function. Aspirin (a small molecule) binds to cyclooxygenase (a protein) to cause relief of headache is an example of protein-small molecule interaction and its outcome.

It is mandatory to use data-analysis and -visualization tools to understand the biology of human being due to the massiveness of the generated data. Such tools[4] have been developed with great focus on the visualization of microarray data and protein network data. So-called "heat-map" is typically used for visualizing microarray data. In the heat-map, two orthogonal axes are genes and experimental conditions. Experimental results are visualized in the form of a colored-matrix. Protein network is often visualized by shapes and lines, where a protein is represented as a shape and an interaction between proteins is expressed as a line connecting the two corresponding shapes.

There have been growing attempts to exhaustively obtain protein-small molecule interaction data. A work by Blueprint known as small molecule interaction database (SMID) has been open to public since 2004[5]. National Institute of Health in the United States announced its plan to collect 500,000 small molecules, so-called “Molecular Libraries”, to perform chemical genomics research, an attempt to analyze the biological system with small molecular probe. In Reverse Proteomics Research Institute, interactions between human proteins and commonly used drugs have been studied since 2001. However, there have been no tools to best suit to the visualization of such exhaustive protein-small molecule interaction data. Current viewers that come with most of existing protein-small molecule interaction databases[6][7][8][9] displays data related to a single protein of interest, or those related to a single small molecule of interest as a list. It is difficult to understand the exhaustive data as a whole by viewing many lists separately. Heat-map is more suitable for this purpose. However, access to detailed information that is related to each protein-small molecule interaction is limited with the heat-map.

We extended functionality of a heat-map viewer and developed a protein-small molecule interaction viewer, named BirdsAnts, based on two concepts[10]. First is the simultaneous visualization of three kinds of data; data pertaining to proteins, those to small molecules, and those to interactions between proteins and small molecules. Second is switching of global and local views of data back and forth. The second concept closely relates to the name of the tool, BirdsAnts. The "Birds" part of BirdsAnts means the capability of observing whole data globally by bird's-eye view. The "Ants" part means the capability of displaying local and detailed information as if an "ant" observes the data.

BirdsAnts are applied to the analysis of a data matrix that is derived from SMID[5]. SMID is a database of small molecule-domain interactions determined from molecular modeling database (MMDB) records. MMDB itself is originated from the Protein Data Bank (PDB). Each domain of any given protein found to interact with a small molecule is then PSI-BLASTed against NCBI's conserved domain database (rpsBLAST). Any result with an E-value less than 0.003 (a measure of similarity of a given domain to conserved domain in rpsBLAST) are stored in SMID. Setting the E-value threshold as 0.003 would extract most of possibly existing interactions, while part of those can be false positives. The purpose of data analysis is to extract knowledge that is useful for drug discovery from massive small molecule-protein interaction data. As a result of analyzing the

data-matrix, two known facts are rediscovered, which serves as a proof of the design concepts of BirdsAnts.

2. Method

2.1 New feature of BirdsAnts

The simultaneous visualization of data pertaining to protein, those to small molecules, and those to protein-small molecules interaction is achieved using three overlapped tables in BirdsAnts. The three tables showing interaction between general class *A* and *B* are shown in figure 1 as areas enclosed by broken line. In particular case, class *A* is "small molecule" and class *B* is "protein". The first area labeled as "Attributes of entity A" is a table of data pertaining to class *A*. The table contains one or more attributes, as shown by labels "Attribute A-1" to "Attribute A-4". "Attribute A-1" is mandatory as opposed to the other attributes. It represents an identification of each item in class *A*. If *A* is small molecule, "Attribute A-1" can be a name of each small molecule. The other attributes can be Log P, molar refractivity and so on. Further, if *A* is drug, therapeutic effect and its target gene are possible entries for the attributes. The second area labeled as "Attributes of entity *B*"

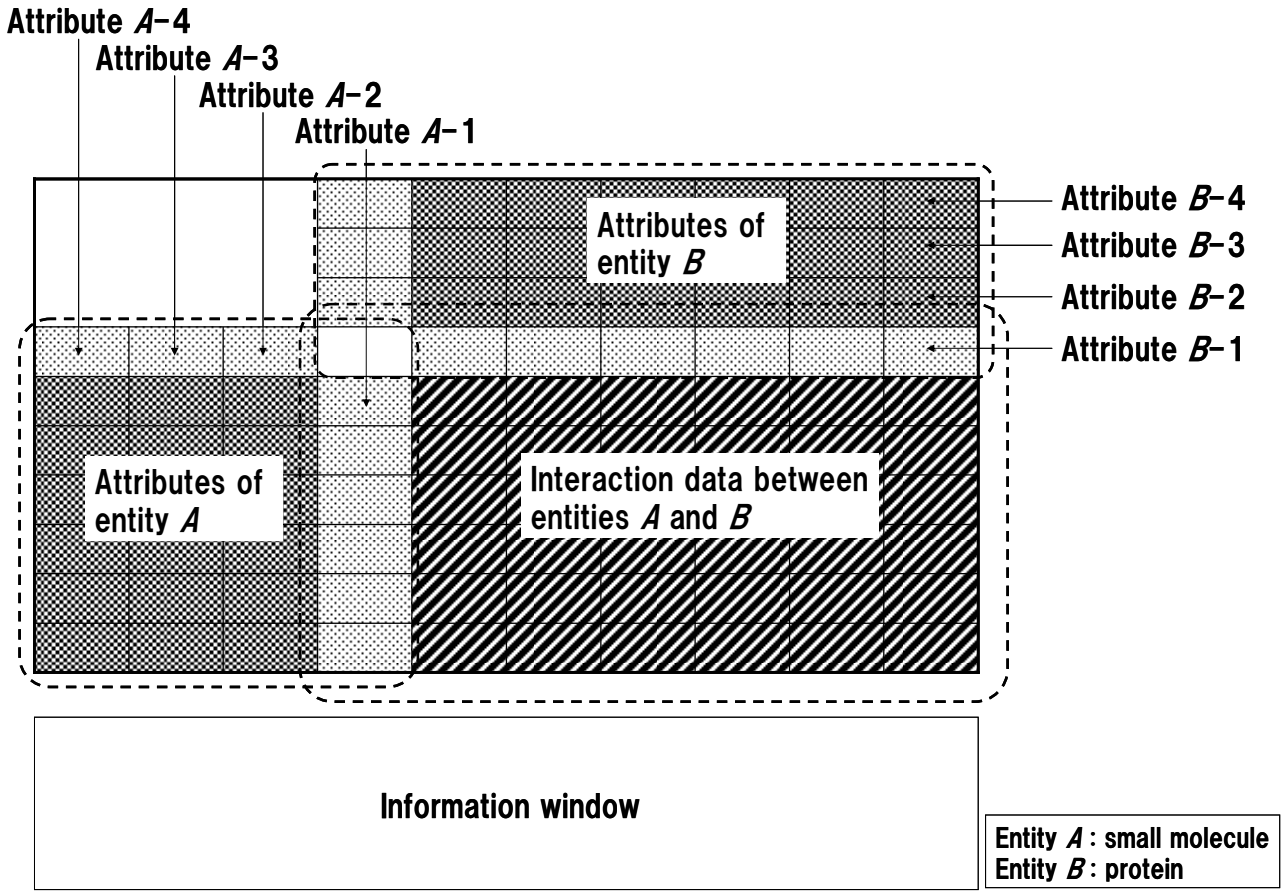

Figure 1. Data visualization scheme by BirdsAnts

holds the same structure as the area for entity *A*. Attributes for proteins are often called "annotations" which contain functional and structural information. The third area labeled as "Interaction data between A and B" is a matrix table containing data related to both classes *A* and *B*. More specifically, they can be experimentally measurable variables such as K_d, K_i, and IC_{50}. We call a rectangular area composing the tables "cell", and each cell in the tables can contain numerical values, text information, and background color as sources of information. More detailed information related to the cell can be displayed at an information window shown at the bottom of figure 1.

In order to access to global and local information back and forth, data can be visualized in three different levels of abstraction. The three views introduced are named "heat-map view", "summary view", and "in-depth view". As the view goes from the heat-map view toward the in-depth view, the amount of displayed information for each cell increases, while the number of simultaneously displayed cell decreases. The heat-map view shows representative values by colors of each cell. The color of a cell is determined according to the strength of an interaction. The summary view provides short text information as well as colored information. Examples of the short text information are names of proteins and molecular weight of small molecules. The summary view is designed so that relationships among interaction data, attributes of small molecules, and attributes of proteins are easily observed. The in-depth view provides long text information for each cell. For example, the title of protein-small molecule complex can be displayed in a cell. Also, structural and functional information of a protein and a name of a small molecule are shown. A cell which contains underlined text information in it has an external URL link to connect to a source of information, such as external database or literature abstract.

2.2 Materials and procedure for data analyses

We have picked up 178 drug-like small molecules, majority of them are drugs approved for therapeutic use in the United States or in Japan, from SMID records. The number of protein domains that interact with at least one of 178 small molecules resulted in 606. A matrix of 178 small molecules and 606 protein domains is shown in Figure 2(a). Each cell in the figure is colored according to the reliability of corresponding interaction between a small molecule and a protein using the E-value. Coloring scheme is as follows: red, 0; dark orange, $<10^{-50}$; light orange, $<10^{-20}$, yellow, $<10^{-10}$; green, $<10^{-5}$; and sky-blue, $<10^{-3}$. Note that as the color goes from red to sky-blue, possibility of an interaction being false-positive increases. In addition, each cell contains relevant information regarding the small molecule, protein domain or their interaction in question.

The matrix is first clustered row-wise and column-wise separately. In this study, a combination of maximum linkage agglomerative clustering scheme and Pearson correlation distance function is used. The chosen combination of clustering scheme and distance function gave the smallest number of resulting clusters among the tested combinations between four variants of the agglomerative hierarchical clustering scheme and eight different distance functions implemented in the C clustering library [11]. The number of clusters is defined as the number of separated groups of cells in the clustered matrix. Here, a separated group is defined as a set of colored cells which are adjacent to at least one colored cell.

We propose a 7-step procedure for analyzing protein-small molecule interaction data using BirdsAnts. This procedure guides in determining when and how we should change the view of a data matrix.

1. Given matrix-data, rows and columns are re-ordered so that proteins of similar interaction profile and small molecules of similar interaction profile are displayed adjacent to each other by means of a clustering technique. Alternatively, a data-matrix file, which is already clustered elsewhere,

can be imported to BirdsAnts.
2. By visually inspecting cell color, we identify an interesting cluster, or region in the matrix. An interesting cluster is a group of cells which are adjacent horizontally and/or vertically on the clustered matrix.
3. The region of interest is visualized in the summary view.
4. By visually inspecting relationships among the value of interaction data, attributes of small molecules, and those of proteins, we identify an interesting set of interaction data.
5. The data of interest is visualized in the in-depth view.
6. Detailed information about the interaction of interest and its components (*i.e.*, small molecule and protein) are obtained. If a URL link is provided, information from external sources is also accessible.
7. A hypothesis is constructed to connect measured or calculated interaction strength between a protein and a small molecule and known information.

If there are more than one cluster of interest, steps 2 through 7 may be repeated. Views can be changed backwards by reversibly following the above steps. Also, if several different clustering results should be compared, step 1 and 2 can be repeated. The above procedure is based on a simple idea that a data matrix is better understood by re-ordering rows and columns of the matrix. The re-ordering makes similar rows and similar columns displayed adjacent to each other, respectively. The re-ordering can be done by using build-in clustering algorithms as well as by manual operations.

BirdsAnts is available at http://www.reprori.jp/BirdsAnts/. Flash™ player version 6.0 or higher (version 7.0 is strongly recommended) needs to be installed.

3. Results

The matrix of 178 drugs by 606 protein domains is shown in Figure 2(a) and its clustering result is shown in Figure 2(b). A visual inspection of the clustered matrix Figure 2(b) shows that, near the left-side and at the right-bottom of the matrix, there are several clusters consisting of a small number of domains and many small molecules as indicated by vertically spread color-patterns. Near the bottom of the matrix, there are several clusters consisting of a small molecule and a relatively large number of domains seen as horizontally spread color-patterns. There are other clusters of various sizes, and each of these may be worth further investigation. We choose to further investigate two clusters; one surrounded by a blue oval and the other surrounded by a red oval. Each of these clusters consists of a small number of both small molecules and domains and because each of those clusters includes data of low E-values as indicated by red colored cells.

3.1 Analysis of Vitamin B12 related data

The summary view of the region surrounded by the blue oval in Figure 2 (b) is shown in Figure 3. Five small molecules respectively labeled COB, TAR, B12, CNC and COY by so-called “HET” code are pre-clustered with respect to structural similarity on the basis of the Tanimoto index[12] The four molecules labeled COB, B12, CNC, and COY are green in color and indicate that they belong to the same cluster with respect to structure. TAR in brown, on the other hand, forms a singleton. Domains of proteins are clustered with respect to enzymatic activity (EC) and structure (CATH) [13][14], separately. Formed clusters were identical in both enzymatic activity-based or structure-based clustering, indicating that each structural domain corresponds to a single type of enzymatic activity. The first digits of the EC numbers are 2, 4, and 5, indicating the possibility that

B12 influences different types of enzymatic reactions. At the left most of the matrix data, there are two interaction data labeled “1G64_A”. This means that chain A of the PDB entry 1G64 has sequence level homology to two pre-defined domains labeled “CobA Cot.” and “RecA-like.”. The homology is higher for one domain (the orange cell) than the other (the yellow cell). Near the center of the matrix, there are three interaction data labeled “1CB7_A” or “1CB7_B”. TAR relates only to

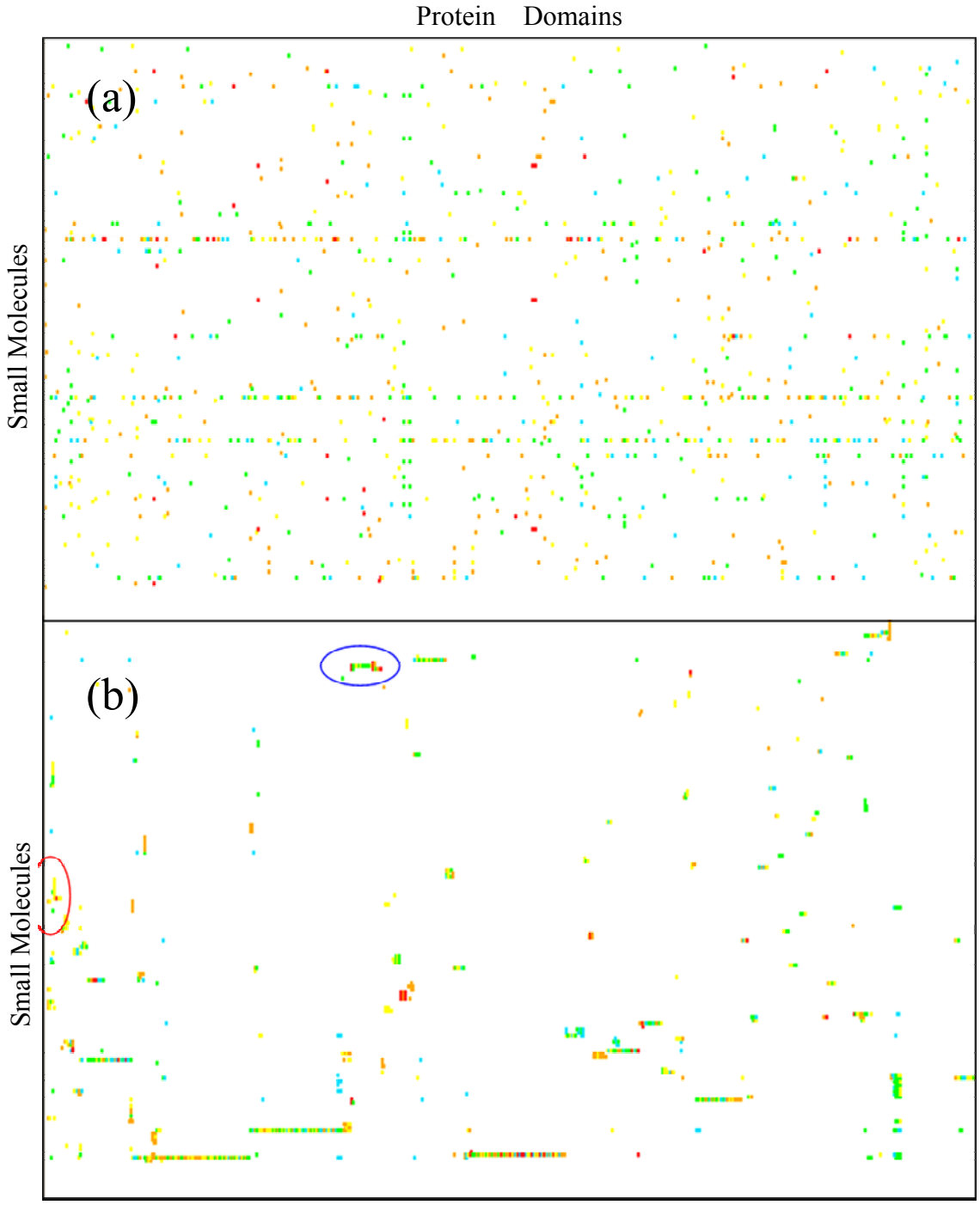

Figure 2. Data matrix between small molecules and domains of proteins in the heat-map view. (a) Before clustering (upper), (b) after clustering (lower)

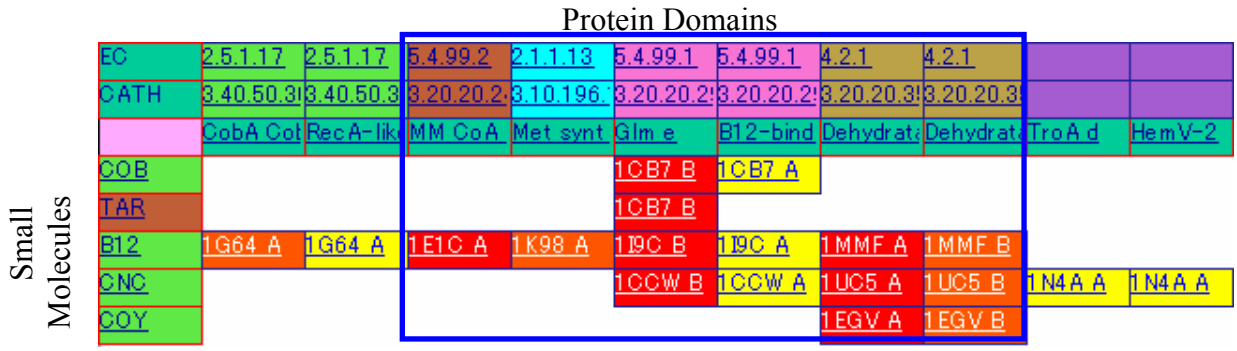

Figure 3. Data matrix of a vitamin B12 related cluster in the summary view.

"1CB7_B" and is thought to bind to an area present in chain B. On the contrary, COB relates both to "1CB7_A" and "1CB7_B", implying that COB contacts both chain A and chain B.
The in-depth view of the region surrounded by the blue rectangle in Figure 3 is shown in Figure4. The selected region contains many red cells indicative of high domain similarity to conserved domains. The four small molecules in green are vitamin B12 and its related molecules. Vitamin B12 is known to be involved in several biological reactions and be related to certain diseases.

Two of the reactions and their relevance to biological processes including a disease are explained by the information contained in the matrix. First, the sky-blue EC information in Figure 4indicates from what enzyme this domain is derived and states that it transfers a methyl group to synthesize methionine. Clicking the URL link on the cell shows that this reaction converts homocysteine to methionine. Failure of methionine synthesis results in accumulation of homocysteine in the blood which is known to trigger atherosclerosis [15]. Second, inspection of the left most column of the matrix in Figure 4 reveals that this protein is methylmalonyl-CoA mutase. The reaction involving methylmalonyl-CoA mutase is located at an upstream of tri-carboxylic acid (TCA) cycle. VitaminB12 acts as the coenzyme to methylmalonyl-CoA mutase, which is related to carbohydrate and lipid metabolism.

3.2 Analysis of COX inhibitors

The summary view of the region surrounded by the red oval in Figure 2(b) is shown in Figure 5. Seven small molecules are shown with two attributes. First attribute is labeled as "Effect" that means therapeutic effect of a small molecule. Five of the seven small molecules in brown have an anti-inflammatory effect. Another attribute is molecular weight (MW) and the formula of small molecules. Colors of cells in the MW column are determined based on the similarity of molecular weight. Colors of cells showing labels of small molecules are determined according to structural similarity as done in the clustering of vitamin B12s. Comparing those three different clustering structures, *i.e.* by effect, by molecular weight, and by structure, there is little consensus among them except that small molecule IBP and 34C lie in the same cluster in all the three different view-points. Thus, those 7 small molecules are considered to be diverse in chemical space. Interestingly, such diverse molecules bind to a common domain of a protein labeled "An perox..".

The in-depth view of the region surrounded by the blue rectangle in Figure 5 is shown in Figure 6. Also, corresponding molecular structures are presented in the figure. Inspection of the orange cells showing title of complex reveals that the protein is cyclooxygenase (COX). Three of the seven

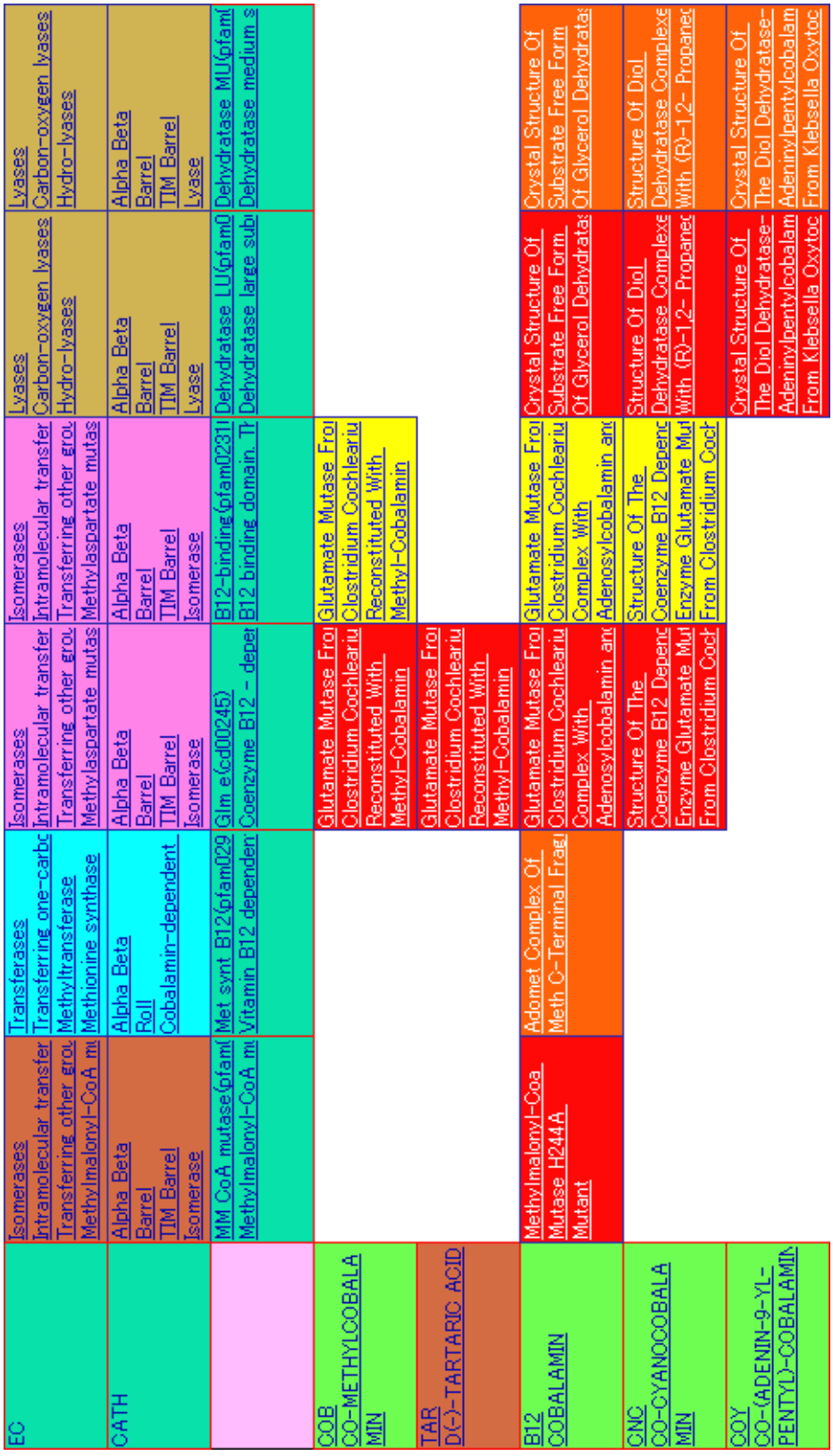

Figure 4. Data matrix of a vitamin B12 related cluster in the in-depth view.

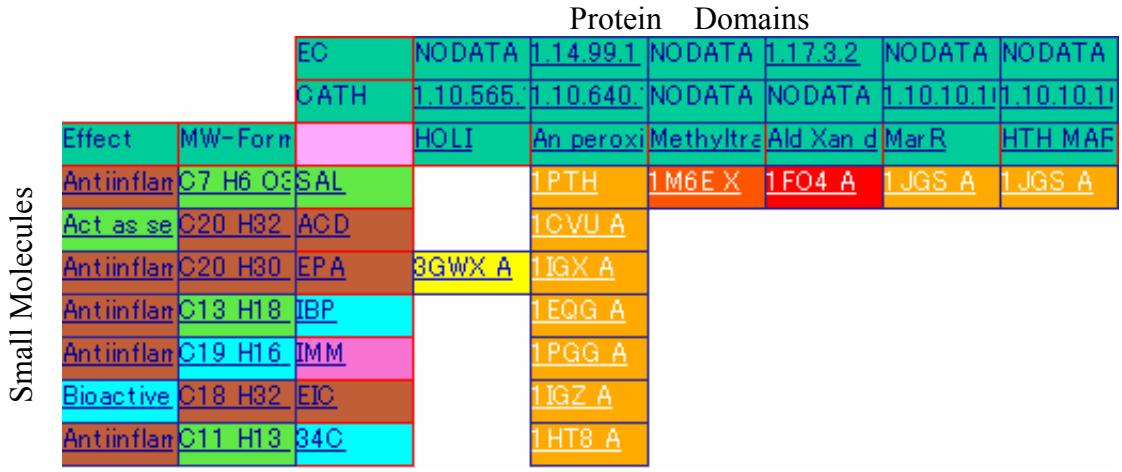

Figure 5. Data matrix of a COX inhibitors-related cluster in the summary view.

small molecules, arachidonic acid, eicosapentaenoic acid, and linoleic acid are native ligands to COX. They are similar in structure and in chain-length. Thus the molecules are within the identical cluster with respect to similarity in molecular weight and structure. The other four molecules are COX inhibitors. Small molecules clustered based on similarity in structure correspond to a known classes of COX inhibitors. According to a classification of COX inhibitors by Kurumbail and co-workers [16], there are four types of inhibitors. The first type they defined is aspirin. Aspirin expresses its anti-inflammatory effect by acetylating serine residues of the COX active site and is metabolized to form salicylic acid, that is molecule 1 in Figure 6. The second type includes ibuprofen (molecule 4) which acts as reversible, competitive inhibitors of COX-1 and COX-2. The third type includes indomethacin and (3-choloro-4-propoxyphenyl)-acetic acid (molecule 5 and 7) which causes a slow, time-dependent inhibition of COX-1 and COX-2. The latter is alclofenac analogue, where allyl-group is substituted by pentyl-group. The fourth type, COX-2 selective inhibitors, is not included in Figure 6.

4. Discussion

The goal of our research was to develop an alternative visualization scheme for protein-small molecule interaction data. Achievements and remaining problems are discussed below.

4.1 Current development status of BirdsAnts

A key factor for the effective use of BirdsAnts is the preparation of data, when data size is large. A data matrix used in the previous section contained data between 178 drugs and 606 protein domains. In addition, known information about the drugs, proteins, and drug-protein interactions are included. Such data are compiled as an XML file, which is then read into BirdsAnts. While using XML file as input data makes BirdsAnts operate as a stand-alone program, the reading of an XML file became to take more than a few seconds when the file exceeds about 10 thousand lines using an 800MHz Pentium III laptop with 512Mbytes of memory. In order to overcome this problem, a direct connection between BirdsAnts and a database has been considered. A database containing

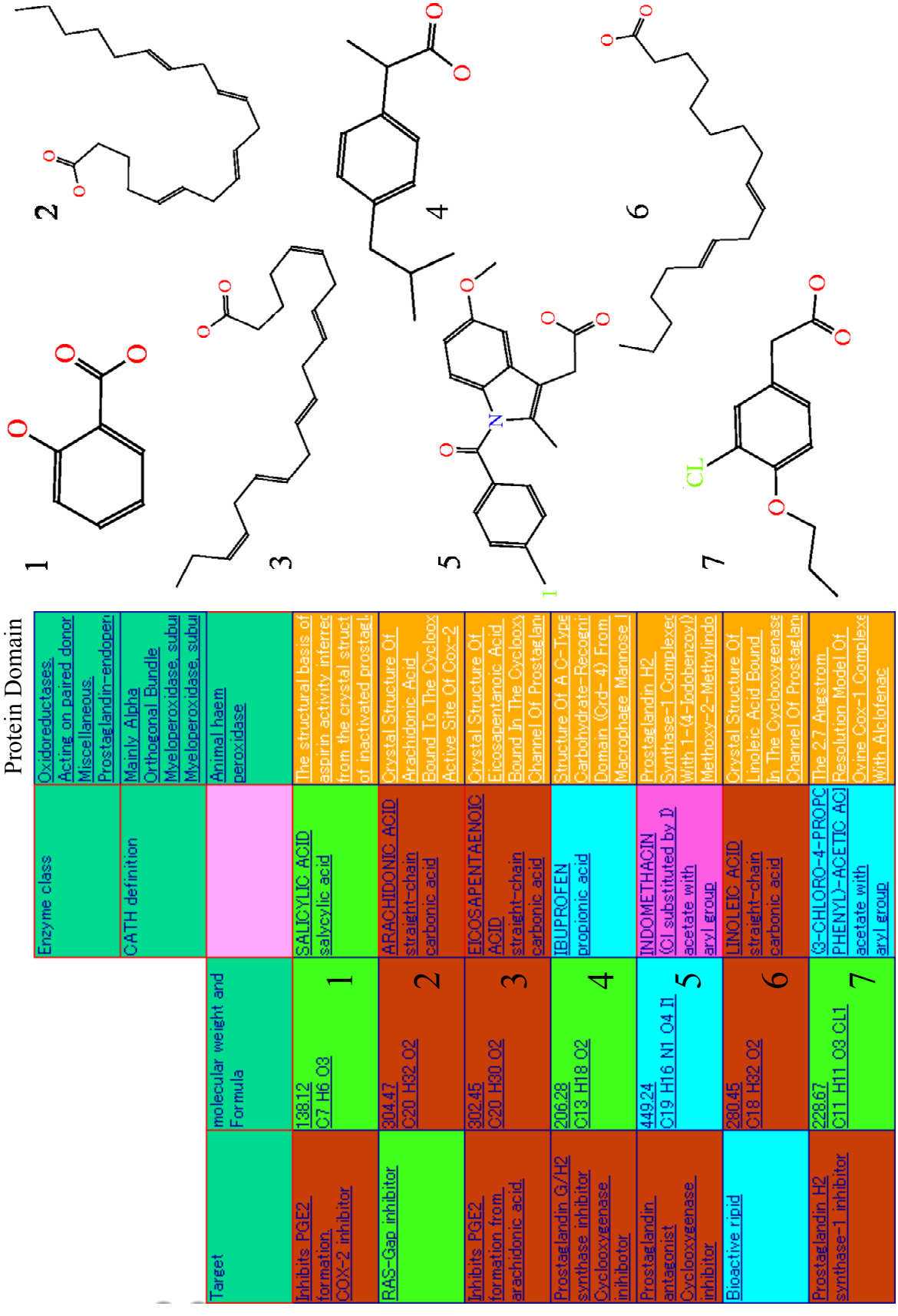

Figure 6. Data matrix of a COX inhibitors-related cluster in the in-depth view.

protein-small molecule interaction data was briefly described elsewhere [17].

Another important factor to accelerate the data analysis is availability of various computational tools which are hidden behind visualization. Although various visualization effects can be realized with less effort in Flash™, a clustering tool that is mandatory for the data analysis cannot conveniently be installed in it. Time-consuming numerical calculations are not best handled in the Flash™ environment, which is a current limitation. Therefore, we found it more effective to rather introduce a clustering program externally albeit having an internally implemented K-means clustering routine.

The maximum number of attributes is not technically limited for the current version. However, when the number of attributes is large, the table showing interaction data between entity A and B shifts outside the computer screen. Thus, showing 5 or less attributes is recommended.

4.2 Choice of clustering scheme and distance measure

The choice of clustering scheme and distance measure depends on types of dataset. From our limited experience, choosing maximum-linkage or average-linkage clustering scheme tends to give the smaller number of resulting clusters as compared to single-linkage or centroid-linkage clustering scheme. As for the distance measure, the eight distance measures are classified into Euclid distance-type measures and Pearson correlation-type measures. When a data-matrix on which we perform a clustering calculation is sparse, the Pearson correlation-type distance measures are more preferable.

5. Conclusion

The purpose of our research was to develop a scheme to effectively visualize protein-small molecule interaction data. We have implemented a data-matrix viewer, BirdsAnts which possesses two features. One is the simultaneous visualization of three kinds of data, *i.e.*, data pertaining to proteins, to small-molecules, and to interactions between them. The other is switching of global and local views of data back and forth.

A BirdsAnts analysis of a matrix derived from SMID successfully rediscovered known facts. The rediscovery was demonstrated in vitamin B12 and COX inhibitors examples. These examples clearly showed that BirdsAnts has future potential to effectively analyze of protein-small molecule interaction data.

BirdsAnts is being applied to analyze internally measured data. We expect that new knowledge which relates to the drug discovery will be extracted.

References

[1] Searls D. B., “Data Integration: Challenges for Drug Discovery”, *Nat. Rev. Drug Discov.,* **4**, 45-58 (2005).

[2] Stoll D., Bachmann J., Templin M. F., and Joos T. O., “Microarray Technology: An Increasing Variety of Screening Tools for Proteomic Research”, *Drug Discov. Today : TARGETS,* **3**, 24-31 (2004).

[3] Nikolsky Y., Nikolskaya T., and Bugrim A., “Biological Networks and Analysis of Experimental Data in Drug Discovery”, *Drug Discov. Today,* **10**, 653-662 (2004).

[4] Tao Y., Lin Y., Friedman C., and Lussier Y. A., "Information Visualization Techniques in Bioinformatics During the Postgenomic Era", *Drug Discov. Today : BIOSILICO*, **2**, 237-245 (2004).
[5] http://smid.blueprint.org/index2.php
[6] Puvanendrampillai D., and Mitchell J. B. O., "Protein Ligand Database (PLD): Additional Understanding of the Nature and Specificity of Protein-Ligand Complexes", *Bioinformatics*, **19**, 1856-1857 (2003).
[7] Zhang J., Aizawa M., Amari S., Iwasawa Y., Nakano T., and Nakata K., "Development of *Ki*Bank, a Database Supporting Structure-Based Drug Design", *Comp. Biol. Chem.*, **28**, 401-407 (2004).
[8] Yamaguchi A., Iida K., Matsui N., Tomoda S., Yura K., and Go M., "Het-PDB Navi.: a Database for Protein-Small Molecule Interactions", *J. BioChem. (Tokyo)*, **135**, 79-84 (2004).
[9] Alfarano C., Andrade C. E., Anthony K. *et. al.,* "The Biomolecular Interaction Network Database and Related Tools 2005 Update"*, Nuc. Acids. Res.*, **33**, D418-D424 (2005).
[10] Tobita M., Horiuchi K., Araki K., Nemoto M., and Nishikawa T., "BirdsAnts: Bringing Informative Rules from a Database System, Aimed at Novel Targets Search", *Genome Informatics*, **14**, 286-287 (2003).
[11] De Hoon M. J. L., Imoto S., Nolan J., and Miyano S., "Open Source Clustering Software", *Bioinformatics*, **20**, 1453-1454 (2004).
[12] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, U.S.A: Academic Press, (1999).
[13] Orengo C. A., Michie A. D., Jones S., Jones D. T., Swindells M. B., and Thornton J. M., "CATH--a Hierarchic Classification of Protein Domain Structures", *Structure*, **5**, 1093-1108 (1997).
[14] Pearl F. M. G., Lee D., Bray J. E., Sillitoe I., Todd A. E., Harrison A. P., Thornton J. M., and Orengo C. A., "Assigning Genomic Sequences to CATH", *Nucleic Acids Res.*, **28**, 277-282 (2000).
[15] Sumner A. E., Chin M. M. , Abrahm J. L. , Berry G. T. , Gracely E. J., Allen R. H., and Stabler S. P., "Elevated Methylmalonic Acid and Total Homocysteine Levels Show High Prevalence of Vitamin B12 Deficiency After Gastric Surgery", *Annals of Internal Medicine*, **124**, 469-76 (1996).
[16] Kurumbail R. G., Stevens A. M., Gierse J. K., McDonald J. J., Stegeman R. A., Pak J. Y., Gildehaus D., Miyashiro J. M., Penning T. D., Seibert K., Isakson P. C., and Stallings W. C., "Structural Basis for Selective Inhibition of Cyclooxygenase-2 by Anti-Inflammatory Agents", *Nature*, **384**, 644-648 (1996).
[17] 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).