Visualization and Level-of-Detail Control for Multi-Dimensional Bioactive Chemical Data

Maiko Yamazawa*, Takayuki Itoh*, Fumiyoshi Yamashita**
*Ochanomizu University, **Kyoto University
{maiko, itot}@itolab.is.ocha.ac.jp

Abstract

We previously applied our own hierarchical data visualization technique for structure-activity relationship (SAR) analyses of biochemical data. The study applied a recursive partitioning to store the drugs as hierarchical data, based on their chemical structures, and visualized the hierarchy of drugs. Though the activity data of drugs is usually multi-dimensional, our previous work did not represent the multi-dimensional values onto one display space. This paper presents a technique for visualizing hierarchical multi-dimensional data, and its level-of-detail (LOD) control, for visualization of multi-dimensional bioactive chemical data. The technique is an extension of our hierarchical data visualization technique, where the extended technique places leaf-nodes as well as the original hierarchical data visualization technique, and represents multi-dimensional values by dividing the icons of the leaf nodes.

1 Introduction

Drug discovery and development is costly, time consuming, and high risk activity. The process starts with the discovery of chemicals or clusters of chemicals with particular biological activity. Information visualization techniques should be useful to discover characteristics from such multi-dimensional and large-scale data. We are especially interested in the visualization of collection of multi-dimensional and large-scale chemical compound data for drug discovery.

Our previous paper presented the visualization of bioactive chemicals [7] by applying our own hierarchical data visualization technique. The previous work visualized correlation between chemical structures and receptivity to Cytochrome P450 (CYP) enzymes, which are a super-family of drug metabolizing enzymes that extensively affect the elimination of drugs from the body. The work visualized the structure-activity relationship (SAR) analyses of CYP-related metabolism. Here, the previous work visually analyzed metabolic susceptibility of 161 drugs to major five CYP isoforms (i.e., 1A2, 2C9, 2C19, 2D6, and 3A4), because only the five CYP isoforms account for 95% of hepatic drug metabolism though more than 40 CYPs encoded in the human genome [9]. Design of molecules or their libraries becomes more effective, by understanding what molecular structural attributes relate to substrate specificity of each CYP isoform. Our work applied a recursive partitioning method to find the relationship between metabolic susceptibility profile and chemical structure, and it finally stored the drugs as hierarchical data. Our previous paper [7] discussed about structure-activity relationship the drugs from our visualization results. Though the activity data of drugs is usually multi-dimensional, our previous work did not attempt to visualize the multi-dimensional values onto one display space. We need to develop a hierarchical multi-dimensional data visualization technique for bioactive chemical data, to visually understand the structure-activity relationship from the single visualization result.

The paper presents a visualization technique for hierarchical multi-dimensional data of bioactive chemical data. The technique is an extension of our hierarchical data visualization technique [6, 7], where the extended technique places leaf-nodes of tree structure as well as our own hierarchical data visualization technique, and then represents multi-dimensional values by dividing the icons of the leaf nodes. The technique assigns independent hue to the subregions of the icons, so that each hue denotes each dimension of the multi-dimensional values. It then calculates the saturation and intensity of the subregions from each of multi-dimensional values, so that saturations and intensities denote the values. We think this representation is intuitive to visually understand the correlation between chemical structures and experimental values of drugs: if correlation between chemical structures and experimental values of drugs in a cluster is high, the technique visualizes the drugs as icons with uniform color patterns.

The paper also presents the level-of-detail (LOD) control technique, which unifies multiple icons in a lower-level cluster into a representative icon. The technique divides the subregions of icons into several triangles based on the histogram of the experimental values, and the triangles represent variation of the values. This LOD control technique enables visualization of variation of experimental values in high-level clusters of drugs, as well as experimental values of independent drugs.

2 Related Work

2.1 Multi-dimensional Data Visualization Technique

There have been a lot of multi-dimensional data visualization techniques, including Parallel Coordinates [5], Worlds within Worlds [3], scatter plots techniques, dimension-reduction-based techniques such as Design Galleries [8], and several glyph-based techniques [4].

The technique presented in this paper is aimed to be used as a user interface, where data elements are displayed as clickable icons. The former four techniques do not always avoid overlaps among data elements on the display, however, it is better to avoid them if we would like to display the data as clickable icons. Our technique therefore places the icons by applying our own hierarchical data visualization technique, and represents the multidimensional data by a glyph-like approach.

2.2 Hierarchical Data Visualization Technique

There have been a lot of hierarchical data visualization techniques as well as multi-dimensional data visualization techniques, which are categorized as space-filling and tree-based approaches.

Our space-filling hierarchical data visualization technique [6, 7] represents hierarchical data as small icons and nested rectangles. Quantum Treemap [1] is very analogous to our technique, because both techniques subdivide display spaces into rectangular areas, and represents leaf nodes of tree structure as non-overlapped equally-shaped icons. Actually Quantum Treemap can be an alternative method to our technique for the purpose of this paper. Experiments described in [6] discusses trade-offs between Quantum Treemap and our technique: ours had better numerical results in aspect ratios of rectangular subregions, and similarity of display results among similar data. One more feature of our technique is that it can display independent images as equally-sized thumbnails; even the depth of hierarchy is deep or inhomogeneous. It is unclear if Quantum Treemap can display every leaf-node as equally-sized icons if the depth of hierarchy is inhomogeneous.

3 Hierarchical Multi-dimensional Data Visualization Technique

3.1 Requirements

Following are requirements we believe are important for the visualization of multi-dimensional hierarchical bioactive chemical data.

1) We would like to equally visualize each of drugs; it is therefore preferable that all drugs are represented as equally-shaped and equally-sized icons, and they never overlap each other on a display space.

2) We would like to equally visualize each dimension of values; it is therefore preferable that all dimensions of values are represented as equally-shaped and equally-sized metaphors.

3) We would like to visualize distribution of experimental values at multiple levels; it is therefore preferable that the experimental values can be represented either drug-by-drug or cluster-by-cluster. The cluster-by-cluster representation is also useful, when the data is very large-scale and it is difficult to display all the icons of drugs in one display.

4) We would like to satisfy the above requirements even if the depth of hierarchy is deep or inhomogeneous.

For the first requirement, we apply our own hierarchical data visualization technique. For the second requirement, we present an extension of our hierarchical data visualization technique to represent multi-dimensional values. For the third requirement, we present a LOD control technique to visualize the data at multiple levels.

3.2 Multi-dimensional Value Representation

This section presents our hierarchical multi-dimensional data visualization technique, which is an extension of our hierarchical data visualization technique [6, 7]. The presented technique represents the hierarchy of the data as well as our technique, and then subdivides the icons of leaf-nodes into n subregions if the data has n-dimensional values. It then assigns independent hue to each subregion, and represents each of the n-dimensional values by saturations and intensities of the subregions. This section denotes the i-th value of a leaf-node as t_i ($0 \leq i < n$).

The technique first subdivides square icons representing leaf-nodes as $l \times m$ grid subregions. Our implementation calculates l and m, as $l = [\sqrt{n}] + 1$, and $m = [n/l] + 1$. Here, $[t]$ denotes an integer value that does not exceed t, and products of l and m are always equal to n or more than n. The technique assigns each of n-dimensional values to each subregion. It is possible that the product of l and m is larger than n, but in this case our implementation lets odd subregions as blank. This representation is quite simple, but we think it is reasonable. The presented technique aims to display hundreds or thousands of leaf-nodes without overlapping in one display, and therefore areas of the leaf-nodes would be small. At the same time, it is better to subdivide display spaces into squares rather than thin shapes, especially when they are small on the display [2]. Based on this discussion, the technique simply divides the icons into square-like subregions.

The technique then calculates the colors of the subregions. It uses HSI color system, where this section denotes hue as H, saturation as S, and intensity as I. It first selects n subregions, and independently assigns hues to each n subregion. Our implementation simply calcu-

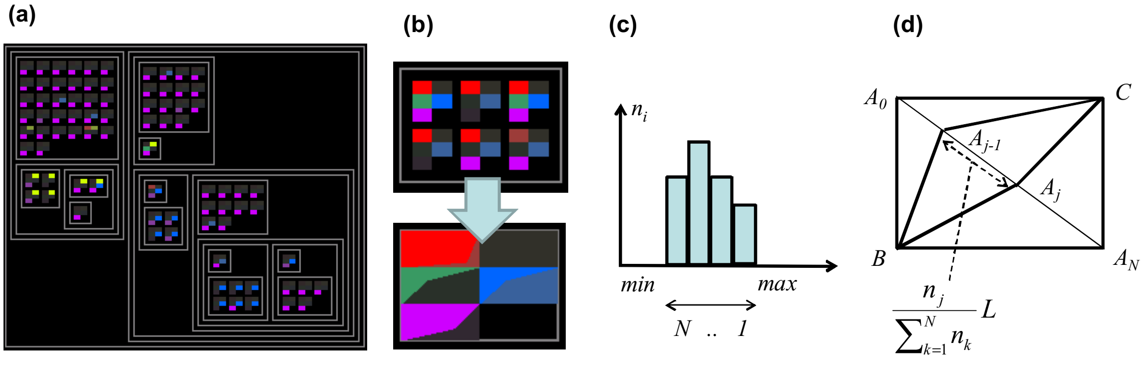

Figure 1: Unified representation of values of lower-level nodes into one higher-level representative node.

lates H ($0 \leq H < 2\pi$) as $H = 2\pi i/n$. It then calculates S ($0 \leq S \leq 1$) and I ($0 \leq I \leq 1$) from the i-th value t_i ($0 \leq i < n$), where we assume t_i is normalized as $0 \leq t_i \leq 1$. Our implementation simply calculates S and I as $S = I = 0.2 + 0.8t_i$. Consequently, the technique assigns different hue for each subregion to represent different dimensions, and represents the values by saturation and intensity. This implementation is suitable while the data has a property that importance of a node increases when its values get higher. Otherwise, we need to reconsider the equation to calculate S and I.

3.3 Level-of-Detail Control

Figure 1(a) shows an example of visualizing hierarchical multi-dimensional data by our technique, and its zoom-in view of a part of the data. This example shows that we need zoom-in operations to visually recognize each value of leaf-nodes of large-scale data. In other words, it may be difficult to visually recognize the values of every leaf-node of the large-scale data in one display, since the displayed leaf-nodes very small. To solve the problem, the technique provides a level-of-detail (LOD) control technique that adjusts the number and sizes of icons on the display, by unifying lower-level nodes as a representative higher-level node. Figure 1(b) shows an example of five icons of leaf-nodes which have 5-dimensional values. Our technique unifies the five icons as a representative icon.

The technique forms a histogram of values of lower-level nodes by dividing their range into N intervals, where the first interval is the maximum, and the N-th interval is the minimum, as shown in Figure 1(c). It uses the histogram to represent the variation of the values by the representative node. Our implementation fixes N as 3.

Let upper-left, upper-right, lower-left, and lower-right corners of the subregion as A_0, B, C, and A_N, as shown in Figure 1(d). The technique draws a diagonal line between A_0 and A_N, and divides the line into N segments, while generating vertices A_1 to A_{N-1} between A_0 and A_N. It calculates d_j, the distance between A_{j-1} and A_j ($j = 1..N$), by the following equation:

$$d_j = \frac{n_j}{\sum_{k=1}^{N} n_k} L \quad (1)$$

where L is the length between A_0 and A_N, n_j is the number of nodes categorized in the j-th interval of the histogram. Finally, the technique paints two triangles, $A_{j-1}A_jB$ and $A_{j-1}A_jC$, to represent the j-th interval of the histogram. Here, it calculates S and I, where t_i of j-th interval is calculated by the following equation:

$$t_i = \frac{(j - 0.5)min_i + (N + 0.5 - j)max_i}{N} \quad (2)$$

where min_i is the minimum value, and max_i is the maximum value of the i-th dimension, respectively. This representation makes easier to visually distinguish between leaf and non-leaf nodes, because diagonal intensity borders only appear in the representative nodes of clusters.

In addition to the above representation, our implementation automatically controls the LOD interlocking to the zooming operation of a user. The technique unifies lower-level icons into less number of representative higher-level icons according to the zoom-out operation. It also inversely replaces representative higher-level icons by larger number of lower-level icons according to the zoom-in operation.

Figure 2 shows an example of LOD control interlocking to the zooming operation. The zoom-out display of Figure 2 represents representative nodes of higher-level clusters. While the zoom-in operation, the representative icon in the red box is replaced by four representative icons, and finally replaced by icons corresponding to leaf-nodes of the tree structure. Users can explore the interesting lower-level hierarchy by the zooming operation.

Our implementation simply draws rectangular areas of branch-nodes as representative icons, because we would

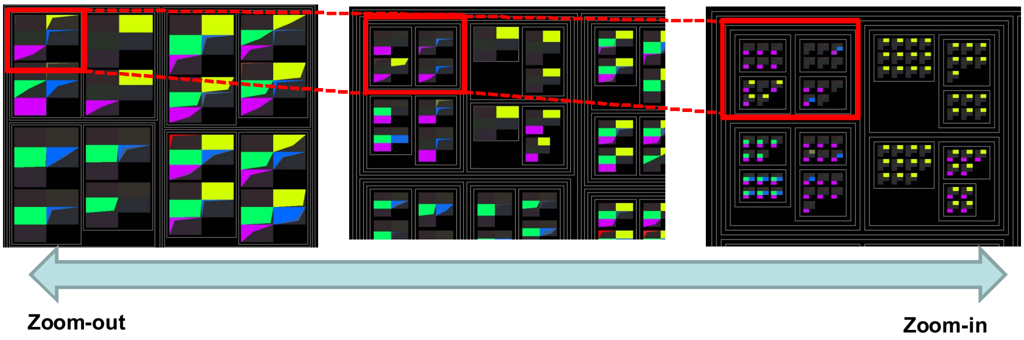

Figure 2: Level-of-detail control (LOD) interlocking to the zoom operation.

like to represent them as large as possible without overlapping each other. Here, it is usually better to draw the representative icons as squares rather than thin rectangles. Since our previous paper experimentally proved that our hierarchical data visualization technique was good at aspect ratio of rectangular subregions for representing clusters [6], it appears that our data visualization technique is preferable for this purpose.

4 Visualization of Bioactive Chemical Data

4.1 Construction of Hierarchical Multi-dimensional Data

This section describes the experiments of visualization of multi-dimensional data of bioactive chemicals. In this experiment, we used the metabolism data of 161 drugs against five CYPs (CYP1A2, CYP2C9, CYP2C19, CYP2D6, and CYP3A4). Here, CYP is an abbreviation for cytochrome P450, which is a collective term of proteins work as isozymes. The above five CYPs mainly work in livers. In this experiment we analyzed metabolic susceptibilities of drugs with the five CYPs. Here some drugs have high susceptibility values with many of CYPs, and some others have none or only one of CYP. Our interest is to discover correlations between chemical structures and susceptibility values as many as possible; we expect that such discover can contribute to predict the experimental values of newly designed drugs.

We first gathered five values of metabolic susceptibilities for each of 161 drugs, and then normalized the values. We then constructed hierarchical five-dimensional data by recursively dividing the drugs according to their structural features. Each of the division process formed two subsets of the drugs to maximally increase the information gain, which is defined as the reduction of information entropy.

In our study, the information entropy h was defined as:

$$h = \sum_{i=1} -P(s_i)logP(s_i), \qquad P(s_i) = \frac{n_i}{N} \qquad (3)$$

where n_i is numbers of drugs in the i-th cluster, and N is the total number of the drugs, respectively.

Let the information entropy of a drug group as h_0, and the information entropy of the two subsets as h_1 and h_2. We applied various molecular constitutional descriptors as a trial, to divide the drugs into two subsets, and calculated the information gain $G = h_0 - (h_1 + h_2)$. We took on the descriptor which brought the maximum G value. Recursively repeating this division, we constructed a binary classification tree, and treated the tree structure as hierarchical data. We used Dragon 5.2 (Talete srl, Italy) [10] as the molecular constitutional descriptors derived from chemical structure.

4.2 Visualization Results

Figure 3(Left) shows an example of the visualization of hierarchical multi-dimensional data constructed by aforementioned procedure. Here, metabolic susceptibility of the five CYPs is represented as the following five colors: CYP1A2 as red, CYP2C9 as yellow, CYP2C19 as green, CYP2D6 as blue, and CYP3A4 as magenta.

In the recursive partitioning analysis, the primary description raised for classification was whether sum of atomic Sanderson electronegativity (Se) would be less than 44.89 or not. In Figure 3(Left), left cluster (pointed as (A)) contains drugs whose Se values are less than 44.89, and in the cluster there are no icons with red and green subregions that are bright. The visualization result proves that the primary description is very correlative with CYP1A2 and CYP2C19.

Figure 3(Upper-center) and (Upper-right) is two typical parts of the visualization result shown in Figure 3(Left). Here, it is often caution needed if a drug is susceptible with only one isozyme, because the dynamics of the drug strongly depends on the isozyme, and therefore risk of drug interaction may get higher. In this case, only one color of icons becomes bright.

Figure 3(Upper-center) shows caution needed clusters of drugs, since only one subregion is bright in many of

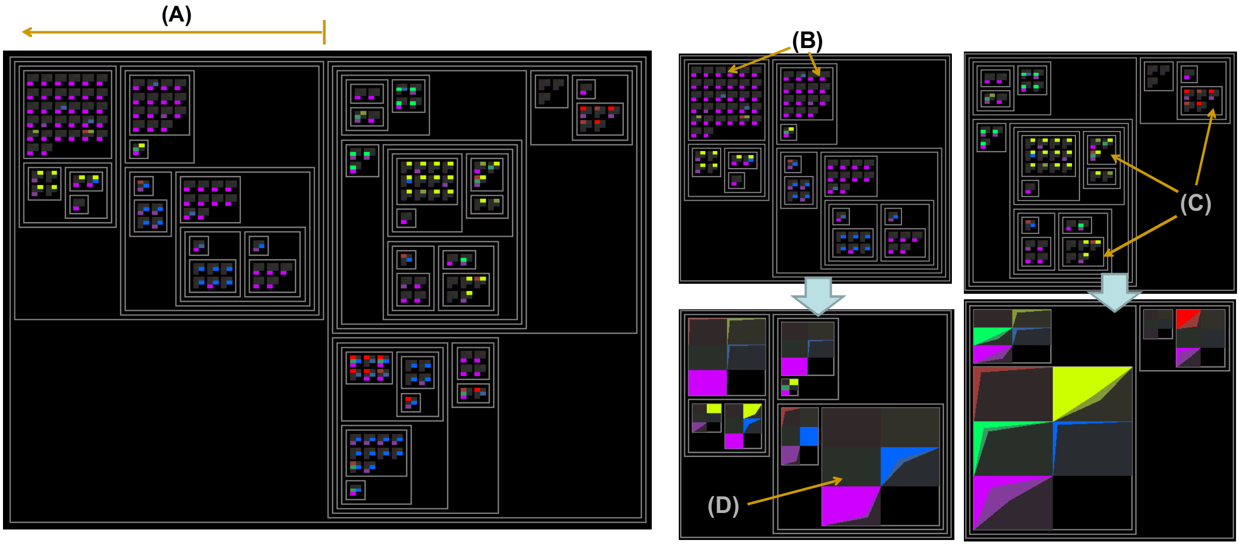

Figure 3: Example of visualization result.

icons, and same colors are bright in the most of icons of the clusters (pointed as (B)). If untested drugs have same chemical structures with the clusters shown in Figure 3(Upper-center), users can expect that the untested drugs have experimental values similar to the drugs in the clusters, and they are caution-needed.

On the other hand, Figure 3(Upper-right) shows many preferable drugs, since multiple subregions are bright in many of icons. In this case, the drugs are susceptible with multiple isozymes, and therefore risk of drug interaction may get lower. However, different colors of icons are bright in some of the clusters (pointed as (C)) in Figure 3(Upper-right). If untested drugs have same chemical structures with the clusters shown in Figure 3(Upper-right), it is difficult to predict the experimental values of the untested drugs.

Figure 3(Lower-center) and (Lower-right) represents the same two parts as representative icons of mid-level clusters, also shown in Figure 3(Upper-center) and (Upper-right). Figure 3(Lower-center) shows that only one or two colors are bright in most of representative nodes, and most of their subregions look almost rectangular, except a lower-right large representative node (pointed as (D)) have non-rectangular bright blue and magenta subregions. We can easily observe that Figure 3(Lower-center) contains many clusters which contribute to the prediction of experimental values, thanks to the LOD control. On the other hand, Figure 3(Lower-right) shows that all three representative nodes have subregions which do not look rectangular. Again, we can easily observe that Figure 3(Lower-right) contains clusters which do not contribute to the prediction of experimental values, comparing with the clusters shown in Figure 3(Lower-center).

One of our motivations for the visualization of bioactive chemical data is that we would like to visualize a variation (maximum and minimum values) of experimental values in high-level clusters of drugs, as well as experimental values of each drug. We think such kind of LOD control is a good approach to interactively get and return high- and low-level information of bioactive chemical data.

As mentioned above, we think that this kind of bioactive chemical data visualization technique will be useful for prediction of functionality or experimental values (e.g. metabolic susceptibility) of untested new drugs. If chemical structures of the new drugs are already known, and the correlation between their chemical structures and experimental values are high, we expect that we can predict their experimental values before experiments, and the estimation can contribute to reduction of experiment costs. Our visualization technique really has been already used in the research and development divisions of a pharmaceutical company, and contributes to screening of test drugs during their development.

4.3 User Experiments and Discussion

We had user experiments of the presented technique, and especially discussed about effectiveness of the LOD control technique. We asked 11 examinees to use and discover our user-interface for several minutes, and evaluate it. All the examinees were grad or undergraduate students belonging to computer science division.

First, we prepared a hierarchical 5-dimensional data, and asked examinees to discover specific lowest-level clusters within 30 seconds. The specific clusters were that "the clusters which all dimensions of experimental values of every node are constant". Actually, the data contained 16 clusters satisfying the condition. We provided two implementations of the presented technique: one supported the LOD control, and the other did not support it. This test proved that the LOD control helped the users in discovering more specific clusters [1] .

Second, we prepared various hierarchical 5-dimensional data, and asked examinees to look for the lowest-level cluster which have a specific feature, by operating our technique. We measured the time taken by the examinees to discover the specific lowest-level cluster. In this experiment we specified the following feature: Two specific dimensions vary, and three other dimensions are constant, among the values of the leaf-nodes in a cluster. We prepared four hierarchical data (called "Data 1" to "Data 4"), where the numbers of leaf-nodes are 9140, 9216, 727, and 729. Sizes and depths of clusters are not uniform in Data 1 and 3, but they are uniform in Data 2 and 4. We asked examinees to use two versions of the presented technique: one version supported the LOD control, and the other did not implement it. This test proved that LOD control usually works well to assist the quick discover of specific clusters. It also proved that it is effective if sizes and depths of clusters are uniform.

5 Conclusion

This paper presented hierarchical multi-dimensional data visualization and its LOD control technique, for visualization of bioactive chemical data. The technique is an extension of our own hierarchical data visualization technique, which represents the hierarchy as nested rectangular borders. The technique applies grid-like subdivision of the icons corresponding to leaf-nodes of the hierarchical data, and represented each dimension of the data as hue of icons, and the variance of the multi-dimensional values as saturation and intensity. The LOD control technique unifies icons in lower-level clusters as a larger representative icon of a higher-level cluster, interlocking to the zoom-out operations. The paper also presented several examples and user experiments which indicated the effectiveness of the presented technique, and discusses the contribution of the technique for the visualization of bioactive chemical data.

The technique has a limitation on scalability. We experimentally evaluate that it is not always well-readable if the dimension is more than 20. Also, we experimentally evaluate that it is not always clickable if the number of icons are more than 5000. We need further evaluation on scalability, and development of improved techniques.

Unfortunately we do not have larger publishable data, since we usually test the technique with confidential data of companies. As a future work, we would like to prepare the larger publishable data and introduce more experiments.

The presented technique is not essentially limited to the visualization of bioactive chemical data, and therefore another future work is exploring the usefulness of the technique in other fields.

References

[1] Bederson B., Schneiderman B., Ordered and Quantum Treemaps: Making Effective Use of 2D Space to Display Hierarchies, *ACM Transactions on Graphics*, 21, 4, 833-854, 2002.

[2] Bruls D.M., Huizing K., Wijk J. J., Squarified Treemaps, *Data Visualization 2000 (joint Eurographics and IEEE TCVG Symposium on Visualization)*, 33-42, 2000.

[3] Feiner S., Beshers C., Worlds within Worlds: Metaphors for Exploring n-Dimensional Virtual Worlds, *ACM Symposium on User Interface Software and Technology (UIST'90)*, 76-83, 1990.

[4] Forsell C., Seipel S., Lind M., Simple 3D Glyphs for Spatial Multivariate Data, *IEEE Information Visualization 2005*, 119-124, 2005.

[5] Inselberg A., Dimsdale B., Parallel Coordinates: A Tool For Visualizing Multidimensional Geometry, *IEEE Visualization '90*, 35-38, 1990.

[6] Itoh T., Yamaguchi Y., Ikehata Y., Kajinaga Y., Hierarchical Data Visualization Using a Fast Rectangle-Packing Algorithm, *IEEE Transactions on Visualization and Computer Graphics*, 10, 3, 302-313, 2004.

[7] Itoh T., Yamashita F., Visualization of Multi-dimensional Data of Bioactive Chemicals Using a Hierarchical Data Visualization Technique "Heianky-oView", Asia Pacific Symposium on Infomation Visualization (APVIS) 2006, 23-29, 2006.

[8] Marks J., et al., Design Galleries: A General Approach to Setting Parameters for Computer Graphics and Animation, *ACM SIGGRAPH '97*, 389-400, 1997.

[9] Rendic, S., Di Carlo, F. J., Human Cytochrome P450 Enzymes: A Status Report Summarizing Their Reactions, Substrates, Inducers, and Inhibitors, *Drug Metabolism Reviews*, 29, 413-580, 1997.

[10] Todeschini R., Consonni V., Handbook of Molecular Descriptors, Wiley-VCH, Weinheim, 2000.

[1] Detailed statistics of the evaluation is uploaded at http://itolab.is.ocha.ac.jp/~maiko/Junihitoe-evaluations.pdf.