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

Maiko Yamazawa*
Graduate School of Humanitics and Sciences, Ochanomizu University

Takayuki Itoh†
Graduate School of Humanitics and Sciences, Ochanomizu University

Fumiyoshi Yamashita‡
Graduate School of Pharmaceutical Sciences, Kyoto University

ABSTRACT

We previously presented a technique for structure-activity relationship (SAR) analyses of biochemical data [1]. The study applied a recursive partitioning to store the drugs as hierarchical data, based on their chemical structures. It then visualized the data by our own hierarchical data visualization technique. 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 while using the hierarchical data visualization technique. This poster presents a technique for visualization of hierarchical multi-dimensional data, and its level-of-detail (LOD) control, for visualization of multi-dimensional bioactive chemical data.

1 HIERARCHICAL MULTI-DIMENSIONAL DATA VISUALIZATION TECHNIQUE

1.1 Requirements

Following are requirements we suppose for visualization for multi-dimensional hierarchical bioactive chemical data. Firstly, 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 on a display space. Secondly, 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. Thirdly, 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. Finally, 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 technique to visualize the data at multiple levels.

1.2 Multi-dimensional Value Representation

This section presents our hierarchical multi-dimensional data visualization technique, which is an extension of our hierarchical data visualization technique [1]. The presented technique represents the hierarchy of the data as well as our previous 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 of the subregions, 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 $(0 \leq i < n)$.

*e-mail: maiko@itolab.is.ocha.ac.jp
†e-mail: itot@is.ocha.ac.jp
‡e-mail: yama@pharm.kyoto-u.ac.jp

The technique first subdivides square icons representing leaf-nodes as $l \times m$ grid subregions. Our implementation calculates l as $l = [\sqrt{n}] + 1$, and m as $m = [n/l] + 1$. 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 of the subregions. 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.

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. The technique first selects n subregions, and independently assigns hues to each of the n subregions. Our implementation simply calculates H, as $H = 2\pi i/n$, where $0 \leq H < 2\pi$. The technique 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 $S = I = 0.2 + 0.8t_i$.

1.3 Level-of-Detail Control

Figure 1 shows an example of visualization result of 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 leaf-nodes are displayed 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: (Upper) Example of hierarchical multi-dimensional data visualization. (Lower) Zoom-in view of a part of the data.

Figure 2(Upper-left) shows an example of five icons of leaf-nodes which have 5-dimensional values. Our technique unifies the icons as one representative icon, as shown in Figure 2(Upper-right). The technique forms a histogram of values of lower-level nodes by dividing their range into N intervals, where the first interval is the

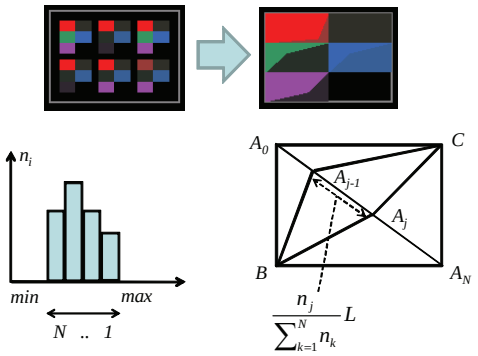

Figure 2: Unified representation of values of lower-level nodes into one higher-level representative node.

maximum, and the N-th interval is the minimum, as shown in Figure 2 (Lower-left). It uses the histogram to represent the variation of the values by the representative node. 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 2(Lower-right). 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:

$$d_j = \frac{n_j}{\sum_{k=1}^{N} n_k} L \tag{1}$$

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. The technique calculates the saturation and intensity of the triangles, where t_i of j-th interval is calculated by the following equation, where min_i and max_i are the minimum and maximum values of the i-th dimension:

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

In addition to the above representation, our implementation automatically controls the LOD interlocking to the zooming operation of a user. It 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.

2 Visualization of Bioactive Chemical Data

2.1 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).

We constructed hierarchical multi-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 $h = \sum_{i=1} -P(s_i)logP(s_i)$, where $P(s_i) = n_i/N$, n_i is numbers of drugs in the i-th cluster, and N is the total number of the drugs in the (sub)data set, 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 as $G = h_0 - (h_1 + h_2)$. We took on the descriptor which brought the maximum G value. Recursively repeating this division, we constructed a binary classification tree, and treated as hierarchical data.

2.2 Visualization Results

Figure 3(Upper) 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 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's electronegativity (Se) be less than 44.89 or not. In Figure 3(Upper), left cluster contains drugs whose Se values are less than 44.89, and in the cluster there are no icons that red and green subregions are bright. The visualization result proofs that the primary description is very correlative with CYP1A2 and CYP2C19.

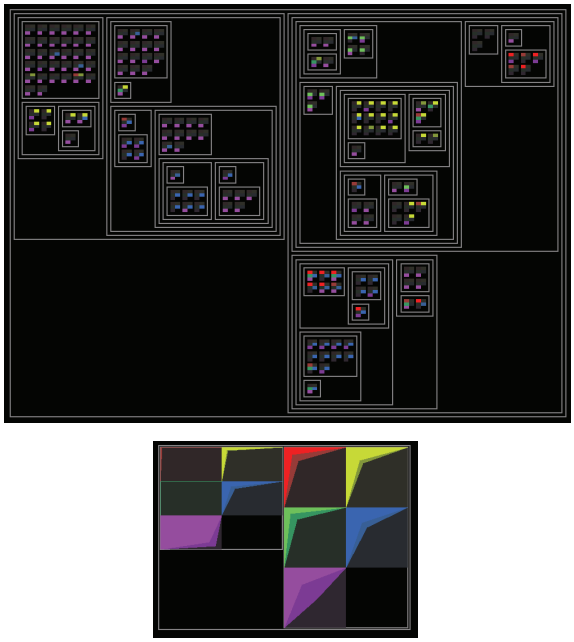

Figure 3: (Upper) Example of visualization result. (Lower) Example of LOD control.

Figure 3(Lower) represents the same data as two representative nodes by applying the LOD control. We can observe that the left representative node does not contain bright red and green parts, but the right representative node does. It is easier to find that the primary description is very correlative with CYP1A2 and CYP2C19 by looking Figure 3(Lower), rather than looking Figure 3(Upper). Also, it is easily found that most of drugs in the left cluster are susceptive to CYP3A4 because bright magenta part looks almost rectangular. On the other hand, about a half of drugs in the right cluster are susceptive to CYP3A4 because bright magenta part looks almost triangular. Again, such feature is found easier by looking Figure 3(Lower), rather than looking all icons in Figure 3(Upper). The result proofs that the LOD control well-summarizes the distribution of values, and assists users to easily discover the features.

References

[1] Itoh T., Yamashita F., Visualization of Multi-dimensional Data of Bioactive Chemicals Using a Hierarchical Data Visualization Technique "HeiankyoView", Asia Pacific Symposium on Information Visualization (APVIS) 2006, pp. 23-29, 2006.