Visualization of Large-Scale Aqueous Solubility Data Using a Novel Hierarchical Data Visualization Technique

Fumiyoshi Yamashita,*,† Takayuki Itoh,‡ Hideto Hara,† and Mitsuru Hashida†

Department of Drug Delivery Research, Graduate School of Pharmaceutical Sciences, Kyoto University, 46-29 Yoshidashimoadachi-cho, Sakyo-ku, Kyoto 606-8501, Japan, and Department of Information Sciences, Faculty of Science, Ochanomizu University, 2-1-1 Otsuka, Bunkyo-ku, Tokyo 112-8610, Japan

Received October 31, 2005

It is a difficult task to recognize the trends in molecular physical properties relevant to a specific chemical class and find a way to optimize potential compounds. We present here a novel hierarchical data visualization technique, named "HeiankyoView", to visualize large-scale multidimensional chemical information. HeiankyoView represents hierarchically organized data objects by mapping leaf nodes as colored square icons and nonleaf nodes as rectangular borders. In this way, data objects can be expressed as equishaped icons without overlapping one another in the two-dimensional display space. HeiankyoView has been applied to visualize aqueous solubility data for 908 compounds collected from the published literature. When the results of a recursive partitioning analysis and hierarchical clustering analysis were visualized, the trends hidden in the solubility data could be effectively displayed as intuitively understandable visual images. Most interestingly, the data visualization technique, without any statistical computations, was able to assist us in extracting from such large-scale data meaningful information establishing that ClogP and the molecular weight are critical factors in determining aqueous solubility. Thus, HeiankyoView is a powerful tool to help us understand structure−activity relationships intuitively from a large-scale data set.

INTRODUCTION

It is a difficult task to recognize the trends in molecular physical properties relevant to a specific chemical class and find a way to optimize potential compounds. Medicinal chemists, who are in charge of drug design and discovery, want interactive tools to assist them to understand structure−activity relationships. Increasing attention has been paid to large-scale data visualization techniques, which are expected to provide intuitively understandable images of the structure−activity relationships.

An essential challenge associated with data visualization techniques is how to visualize data objects containing multidimensional chemical information. Data visualization techniques which have been used in this field can be basically classified into two categories. One is the mapping of multidimensional data onto a two-dimensional (2D) or three-dimensional (3D) space. Principal component analysis is a common method of reducing the dimensionality of a data set, where linear combinations of variables are sequentially explored to explain the largest amount of variation of the original data set. Multidimensional scaling[1] is a nonlinear mapping method that arranges data objects on a 2D or 3D conceptual map so as to reproduce the distances between the data. The self-organizing map invented by Kohonen[2] is also a method of creating an array of nodes onto which multidimensional data are mapped. These mapping methods generally aim to reduce the dimensionality of chemical structural information to 2D or 3D. Data of interest (e.g., biological activity) might be represented by the color, size, or shape of icons on the plot, to investigate structure−activity relationships. Although the structural similarity of molecules can be assessed by measuring the distances on conceptual plots with the eye, it is essentially difficult to understand what structural factors determine the location of compounds. Another approach to visualizing multidimensional data is to graphically represent the results of cluster analyses. Hierarchically organized data are often represented as dendrograms (structural similarity) associated with color mosaics (biological activity). However, dendrographic representations are difficult to navigate, especially when applied to large-scale data. Schneiderman[3] developed Treemap as an alternative for visualizing hierarchical structures. Treemap is a 2D space-filling approach in which each leaf node is a rectangle the size and color of which correspond to specific attributes. Molecular Property eXplorer, reported very recently,[4] is structure−activity relationship visualization software implementing the Treemap algorithm.

This article presents another hierarchical data visualization technique called "HeiankyoView"[5] for the visualization of multidimensional chemical information. HeiankyoView visualizes the hierarchical data by mapping leaf nodes as colored square icons and nonleaf nodes as rectangular borders (see Figure 1). The technique aims to represent all leaf nodes of large-scale hierarchical data in one display space without any focus-and-context techniques. The technique satisfies four requirements for large-scale data visualization: (1) no overlaps between leaf nodes, (2) efficient use of display space, (3) squarish images of subspaces, and (4) flexible placement of arbitrarily shaped nodes. These features of

* Corresponding author phone: 81-75-753-4535; fax: 81-75-753-4575; e-mail: yama@pharm.kyoto-u.ac.jp.
† Kyoto University.
‡ Ochanomizu University.

10.1021/ci0504770 CCC: $33.50
Published on Web 03/28/2006

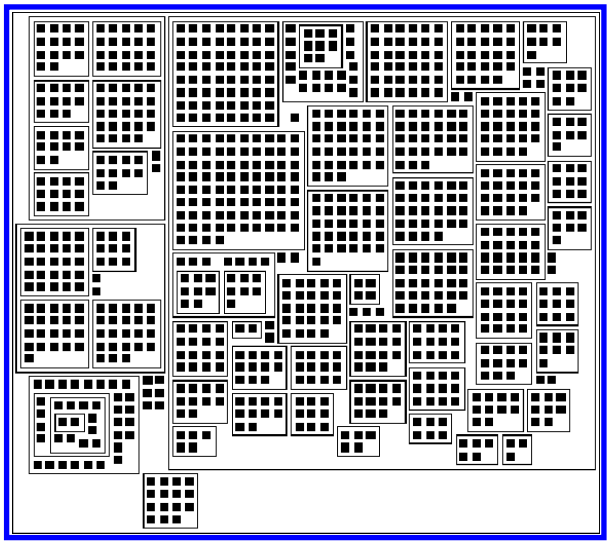

Figure 1. Example of hierarchical data visualization by HeiankyoView.

HeiankyoView are the same as those of the rectangle packing technique developed earlier by Itoh et al.,[6] while HeiankyoView is superior in terms of the computation time (unpublished result). These techniques have been applied to visualize various data, including the access trend of Web sites,[7] jobs in distributed computing environments,[8] and the distribution of network intrusion detection data.[5] These investigations demonstrated that these rectangle packing visualization techniques are useful not only for the overview of large-scale hierarchical data but also for the discovery of interesting but minor local characteristics and the exploration of detailed local information.

In this article, HeiankyoView was applied to the analysis of the structure−activity relationship regarding aqueous solubility. Aqueous solubility is a crucial physical property that plays a significant role in various physical and biological processes. The prediction of aqueous solubility has been extensively studied in pharmaceutical and environmental research fields. The methods proposed so far include Yalkowsky's general solubility equation,[9−11] an amended solvation energy relationship,[12] and other quantitative structure−activity relationship models.[13−16] In general, a quantitative structure−activity relationship model is a mathematical relationship between a property of interest and the structural features of compounds, where the structural features of a particular compound are quantified as a series of molecular descriptors that encode the topological, geometric, and electronic information about the molecule. Besides quantitative structure−property relationship studies using mathematical equations, it would be of great interest to discover if the structure−property relationships present in aqueous solubility data can be intuitively understood. First, we developed a data set consisting of the aqueous solubility of 908 organic compounds available in the published literature[10,11] and their molecular descriptors. Next, we applied two different methods, that is, recursive partitioning and hierarchical clustering analyses, to construct the hierarchically organized aqueous solubility data of the compounds. Finally, we visualized the solubility data using HeiankyoView, demonstrating the usefulness of the data visualization technique in the study of structure−activity relationships.

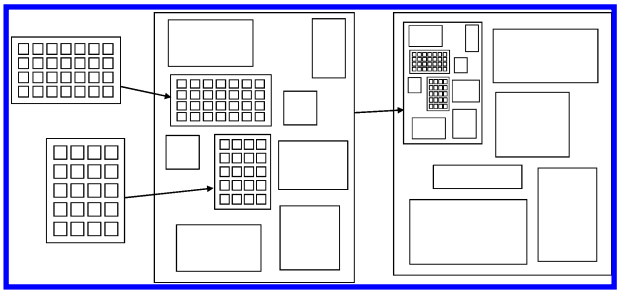

Figure 2. Processing flow of HeiankyoView.

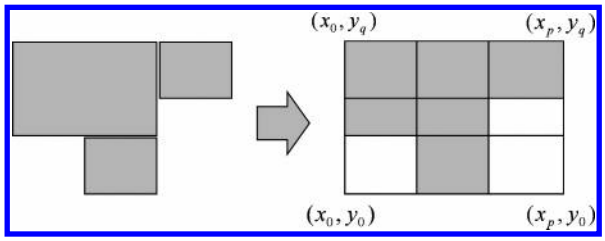

Figure 3. Division process of the display space into rectangular subdivisions for the exploration of candidate positions.

THEORY

A hierarchical data visualization technique called "HeiankyoView" represents leaf nodes of hierarchical data as square icons and nonleaf nodes as nested rectangular borders. Figure 2 shows the processing order of the display layout of the nodes. The technique first places the leaf nodes at the lowest level of the hierarchy onto a display space and represents a nonleaf node by enclosing the leaf nodes. The technique then places leaf nodes and nonleaf nodes at a higher level and again encloses them by another rectangle. The technique represents all of the hierarchical data by repeating the process until it reaches the top of the hierarchy.

The technique solves the rectangle packing problem of finding the optimal display layout of leaf and nonleaf nodes. It also attempts to satisfy the following conditions: (condition 1) Rectangles never overlap one another. (condition 2) The area of the rectangular region enclosing the placed rectangles is to be minimized. (condition 3) The aspect ratio of the rectangular region enclosing the placed rectangles is to be optimized.

The technique places the rectangles onto the display one-by-one, selecting each position of the rectangles from multiple candidate positions. The basic processing order was as follows:

1. Divide the display space into rectangular subdivisions using extension lines of edges of the already placed rectangles (Figure 3).

2. Calculate multiple candidate positions for a target rectangle *R*.

3. Repeat processes a−e for each candidate position.

(a) Place *R* on the candidate position.

(b) If the candidate does not satisfy condition 1, then select another candidate.

(c) Calculate the penalty value reflecting the extension of the display area and the deviation of the aspect ratio from unity (i.e., square) due to the placement of *R*.

(d) If the penalty value is zero, fix *R* to the position.

(e) If the penalty value is not zero but smaller than any of the penalty values of the previously processed candidates, record the candidate position and its penalty value.

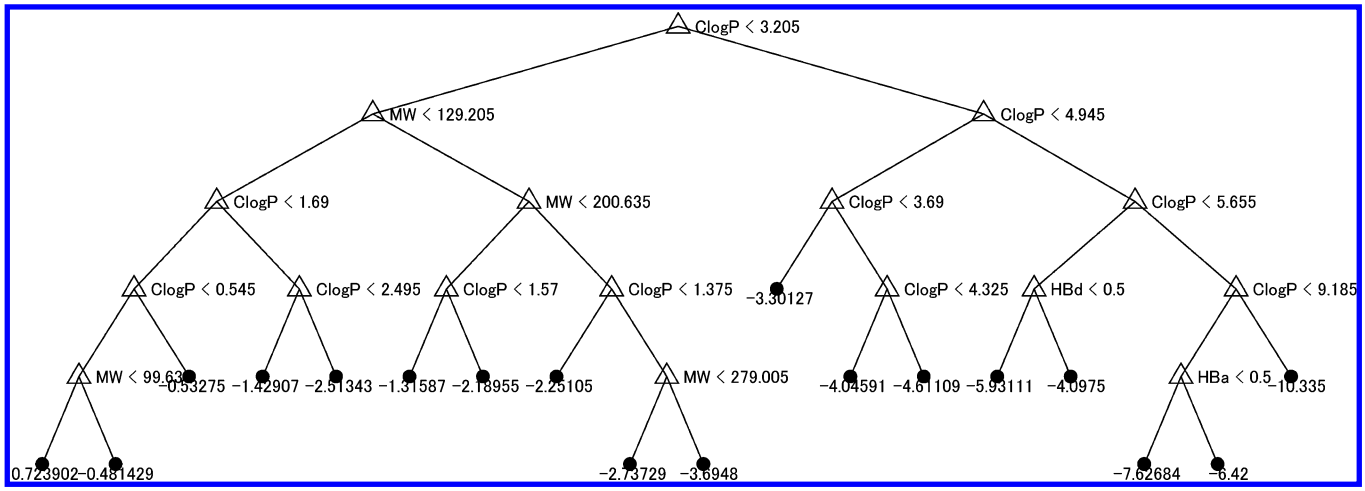

Figure 4. Decision tree model for predicting aqueous solubility from molecular properties. The model was developed by a recursive partitioning analysis.

4. Fix R to the recorded candidate, and update the data structure for the placement of rectangles.

Because of such processing, data objects are laid out in a grid pattern like the old capital city of Kyoto (so-called Heiankyo). The details of the rectangle placement algorithm are described elsewhere.[5]

METHODS

Data Set. Aqueous solubility data for 908 compounds were collected from the published literature,[10,11] which ranged from −10.41 to +2.06 in terms of the common logarithm (log S). The molecular descriptors used were ClogP, molecular weight (MW), counts of hydrogen bonding donors (HBd), and that of hydrogen bonding acceptors (HBa), all of which were descriptors for Lipinski's five rules.[17] The ClogP values were obtained from the published literature,[10,11] while the others were computed using ADMET Predictor (Simulations Plus, Inc., Lancaster, CA).

Cluster Analysis. A cluster analysis was carried out using the Statistics Toolbox 5.0 implemented in the MatLab 7.0 platform (MathWorks, Inc., Natick, MA). The similarity between compounds was evaluated by means of the normalized Euclidean distance in a chemical space consisting of the four attributes (molecular descriptors). Ward's method [i.e., *linkage*('ward')] was adopted to compute a hierarchical cluster tree. The distance cutoff value was set to 2.0.

Recursive Partitioning Analysis. A recursive partitioning analysis was also performed using the Statistics Toolbox 5.0 implemented in the MatLab 7.0 platform. A decision tree was created for predicting aqueous solubility as a function of the four molecular descriptors. The *treefit* function was used for generating a large decision tree. After the best level of pruning was estimated by a cross-validation procedure [i.e., *treetest*('crossvalidate')], the *treeprune* function was applied to obtain the final decision tree.

RESULTS AND DISCUSSION

Recursive Partitioning Analysis and Data Visualization. Recursive partitioning, also known as classification and regression tree analysis, is an exploratory technique for uncovering hidden patterns within data. The technique successively splits a data set into increasingly homogeneous subsets. As a result, the technique gives intuitive and interpretable models in the forms of if−then logics. Recursive partitioning is available, regardless of the type of dependent variables, that is, categorical or quantitative. If the variable is categorical, the misclassification ratio is evaluated at each partitioning step. On the other hand, if quantitative, the intragroup variances are evaluated. When the minimal value of these measures is given, the criterion of an explanatory variable is regarded as optimal. The technique has been widely used to solve various classification and prediction problems including medical diagnosis, phylogenic classification, psychological assessment, and information retrieval. The structure−activity relationship research field is not exceptional; for example, tree models for the prediction of biological activity,[18,19] toxicology,[20,21] physical properties,[22,23] drug absorption,[24,25] and metabolism[26] have been proposed.

In the present study, the structure−activity relationship of aqueous solubility for 908 compounds was analyzed by recursive partitioning. Using ClogP, MW, HBa, and HBd as molecular descriptors, a regression tree was generated (Figure 4). It should be noted that, when the regression tree model was obtained, pruning was carried out to avoid overfitting to the data. The model had a correlation coefficient (R^2) of 0.890. For comparison, the same data set was analyzed by the multiple linear regression method. Multiple linear regression analysis gave the following equation:

$$\log S = 0.781(\pm 0.072) - 0.887(\pm 0.028)\ \mathrm{ClogP} - 5.43(\pm 0.59) \times 10^{-3}\ \mathrm{MW} - 0.0944(\pm 0.0448)\ \mathrm{HBd} - 0.0838(\pm 0.0263)\ \mathrm{HBa}$$
$$(R^2 = 0.810,\ s = 0.886,\ n = 908) \qquad (1)$$

When the regression coefficients were compared, it was obvious that the regression tree model provides a better accuracy than the multiple linear regression one. In addition, the ability of the regression tree model to predict unknown data was investigated. The entire data set was randomly split into training (720 compounds) and testing (188 compounds) data sets. The regression tree model developed from the subdata set provided an R^2 of 0.907 (n = 720) and a predictive R^2 of 0.824 (n = 188). The validation study indicated that recursive partitioning was effective in developing a structure−activity relationship model for the prediction

of aqueous solubility. Xia et al.[22] previously constructed and tested a recursive partitioning model for predicting the solubility of vendor compounds. They classified the compounds into two categories (soluble and insoluble) and analyzed the semiquantitative data by a classification tree method. They demonstrated that the classification tree model improved the success rate of selecting "acceptably" soluble compounds as compared with the use of a simple ClogP cutoff. Although the model of Xia et al. and our own differ in terms of the type of dependent variables (categorical vs quantitative), both investigations suggest the potential benefit of recursive partitioning in the structure−activity relationship analysis of aqueous solubility data.

As shown in Figure 4, the dendrogram or decision tree is a popular representation of the results of a recursive partitioning analysis. It simply represents the structure of a classification model. It is like a mathematical equation which represents a regression model. Therefore, from the dendrogram, it is difficult to grasp how the model was constructed in relation to the data or how reliable the model is. In Figure 4, each leaf node represents an average of the data belonging to it. Although all data objects instead of the averages may theoretically be represented as leaf nodes, it is not practical to represent such a large-scale data set. Because the dendrographic method is essentially a single-dimensional visualization one, there is a limitation in representing large-scale data. Therefore, a conceptually different visualization technique is needed. HeiankyoView visualizes hierarchical data by mapping leaf nodes as painted square icons and nonleaf nodes as rectangular borders. The technique utilizes the 2D display space much more efficiently than dendrographic methods. Figure 5 shows a HeiankyoView representation of the solubility data organized hierarchically by recursive partitioning. The aqueous solubility of each compound is expressed by the "hot-to-cold" color of the square icon. Because HeiankyoView provides an overview of the whole data set, the trends in hierarchical data can be evaluated fully. The graphical image in Figure 5A shows that the regression tree constructed by recursive partitioning is reasonably accurate. In addition, as shown in Figure 5B, one can request annotations for leaf and nonleaf nodes of interest through a graphical user interface. This function assists us to understand the structure of data in more detail.

Cluster Analysis and Visual Data Mining. Cluster analysis is another technique for exploring the structure of data. The technique groups data objects into clusters, such that objects from the same cluster are similar and objects from different clusters are dissimilar. Clustering is referred to as unsupervised classification, which does not use a priori knowledge for the classification of data objects as recursive partitioning does. Therefore, cluster analysis classifies data objects into "natural" groups solely based on their similarity or dissimilarity; on the other hand, recursive partitioning has a greater risk of misclassification by splitting data objects with a single variable. Cluster analysis does not directly aim to construct the model of predicting the target variable from explanatory variables. Rather, the technique is useful for developing some hypotheses concerning the cause-and-result relationship.

We consider that effective visualization of the cluster analysis enhances the discovery of the trends hidden in large-scale data. In this context, the solubility data were analyzed by the hierarchical clustering method, and then, the data organized were visualized using HeiankyoView. Figure 6 shows the dendrographic and HeiankyoView representations of solubility data organized by hierarchical clustering. In the HeiankyoView representation, it was found that a group of the compounds located at the bottom left were mostly highly soluble and another group of the compounds located at the top middle were mostly insoluble in water. Thus, the HeiankyoView representation clearly indicates that structurally similar compounds tend to have a similar aqueous solubility.

When the base plane of HeiankyoView is tilted, another chemical attribute can be represented using the bar height. The bar heights in Figure 7A,B indicate ClogP and MW, respectively. These figures clearly demonstrate that, as the compound is more hydrophilic or has a lower MW, it is more

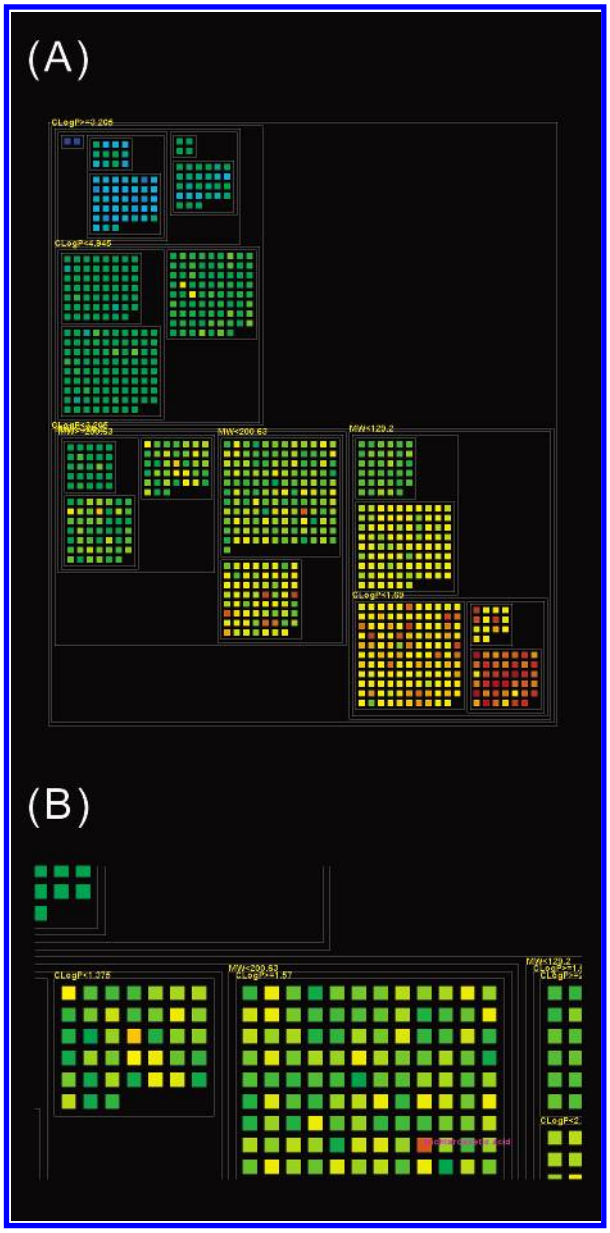

Figure 5. Visualization of solubility data objects split by recursive partitioning using HeiankyoView. Solubility data objects are represented as color icons, while nonleaf nodes are represented as rectangular borders. The upper figure (A) covers all of the data objects, while the lower figure (B) is a magnification of a part of the figure (A).

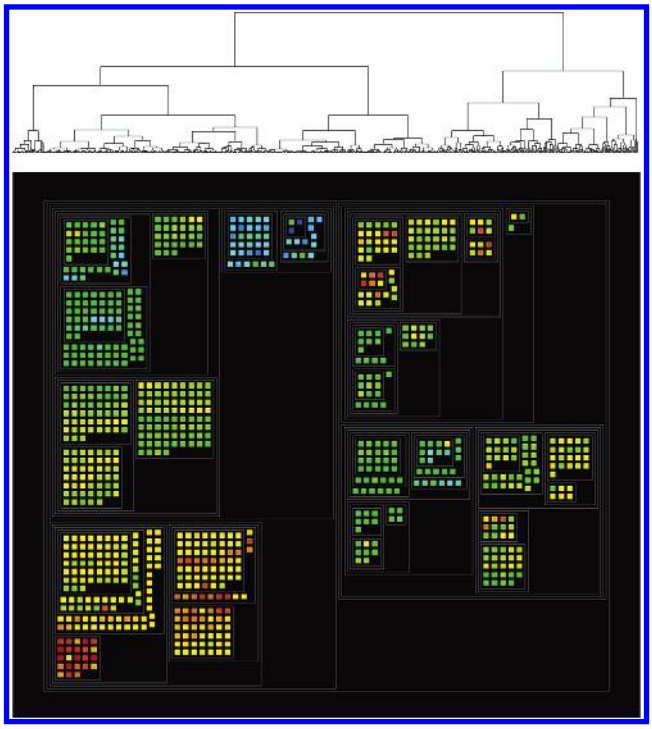

Figure 6. Dendrographic and HeiankyoView representations of solubility data organized by hierarchical clustering.

soluble in water. In fact, when the simple correlation coefficient (r^2) between log *S* and each predictor variable was calculated, ClogP and MW had significant r^2 values (0.719 and 0.355, respectively). On the other hand, no significant relationship was observed with HBd and HBa by tilted HeiankyoView representations (data not shown). An important point to recognize here is that these trends in the structure−activity relationship could be extracted visually from the large-scale raw data.

Concluding Remarks. This study presents a visualization technique, that is, HeiankyoView, for large-scale hierarchical chemical data, demonstrating that this technique can be effectively used to visualize results of data mining or to extract meaningful information from visualized images. As mentioned in the Introduction, a 2D space-filling visualization technique called Treemap has been introduced into structure−activity relationship analysis very recently.[4] The technique represents a hierarchical data structure by splitting a rectangular region along alternating vertical and horizontal lines. A major advantage of HeiankyoView over Treemap is that it is able to represent all data objects as equishaped icons. Another advantage is that HeiankyoView comprises squarish images of subspaces that are low in aspect ratio. While Treemap guarantees that the area of each generated rectangle is proportional to an input number, it does not make any promise about the aspect ratio of the rectangles. Long skinny rectangles can be hard to see, select, compare in size, and label.[27] Thus, HeiankyoView can provide a more acceptable image than Treemap. Although they have never been applied in the structure−activity relationship field, ordered and Quantum Treemaps[27] were also presented, which could overcome the problem associated with Treemap. On the other hand, Itoh et al.[6] developed a rectangle packing visualization technique before HeiankyoView and compared it with Quantum Treemaps, demonstrating that their technique exhibited a better aspect ratio of the layout and a better stability of the layout against slight modification of the data structure but a slower computation time compared with Quantum Treemaps. HeiankyoView is an improved version of rectangle packing algorithms in terms of computation time (unpublished result). An essentially indispensable issue in large-scale data visualization is the user-friendliness of the graphical representation. Further application studies need to be carried out to confirm the usability of this technique.

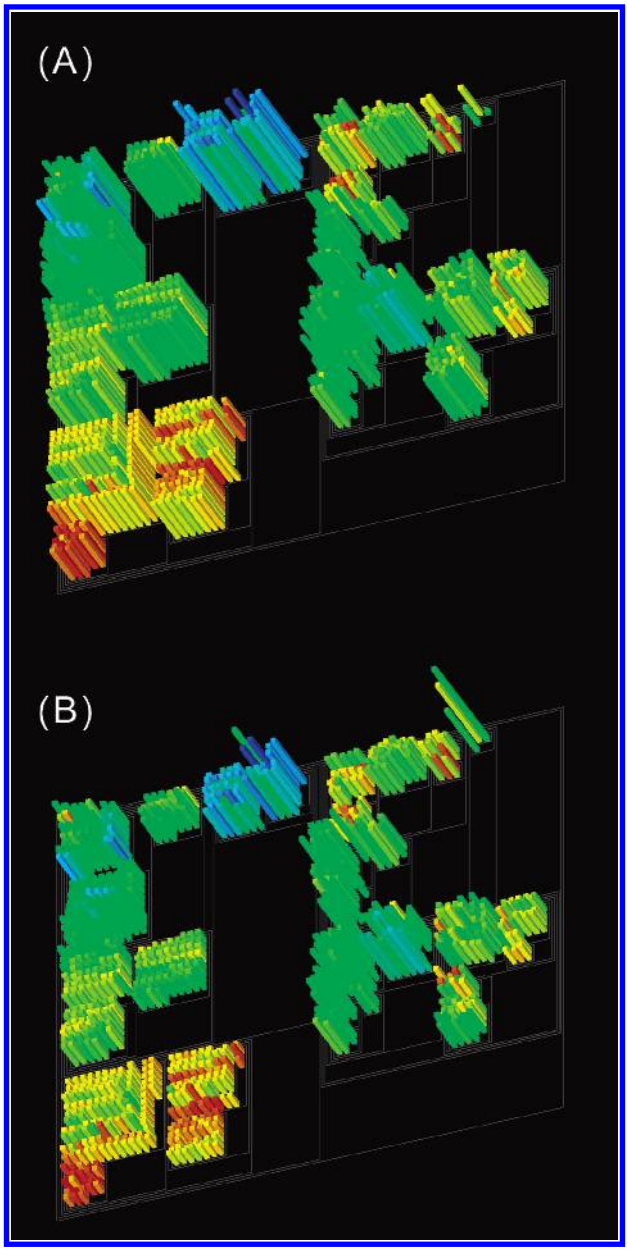

Figure 7. Tilted HeiankyoView images of solubility data arranged by hierarchical clustering. The bar heights represent ClogP (A) and molecular weight (B), while the bar color represents aqueous solubility.

ACKNOWLEDGMENT

This research was supported in part by a Grant-in-Aid for a Leading Project for Biosimulation from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

REFERENCES AND NOTES

(1) Kruskal, J. B.; Wish, M. *Multidimensional Scaling*; Sage Publications: Beverly Hills, CA, 1977.

(2) Kohonen, T. *Self-Organizing Maps*; Springer: Berlin, 1995.
(3) Schneiderman, B. Tree visualization with tree-maps: 2-d space-filling approach. *ACM Trans. Graph.* **1992**, *11*, 92−99.
(4) Kibbey, C.; Calvet, A. Molecular Property eXplorer: a novel approach to visualizing SAR using tree-maps and heatmaps. *J. Chem. Inf. Model.* **2005**, *45*, 523−532.
(5) Itoh, T.; Takakura, H.; Sawada, A.; Koyamada, K. Hierarchical visualization of network intrusion detection data in the IP address space. *IEEE Comput. Graph. Appl*. **2006**, in press.
(6) Itoh, T.; Yamaguchi, Y.; Ikehata, Y.; Kajinaga, Y. Hierarchical data visualization using a fast rectangle-packing algorithm. *IEEE Trans. Vis. Comput. Graph.* **2004**, *10*, 302−313.
(7) Yamaguchi, Y.; Itoh, Y.; Ikehata, Y.; Kajinaga, Y. Interactive poster: web site visualization using a hierarchical rectangle packing technique. *Proc. IEEE Symp. Inf. Vis. 2002*; PUBLISHER: PLACE OF PUBLICATION, 2002.
(8) Yamaguchi, Y.; Itoh, T. Visualization of distributed processes using "Data Jewelry Box" algorithm. *Proc. Comput. Graph. Int. 2003*; PUBLISHER: PLACE OF PUBLICATION, 2003; pp 162−169.
(9) Jain, N.; Yalkowsky, S. H. Estimation of the aqueous solubility I: application to organic nonelectrolytes. *J. Pharm. Sci*. **2001**, *90*, 234−252.
(10) Yang, G.; Ran, Y.; Yalkowsky, S. H. Prediction of the aqueous solubility: comparison of the general solubility equation and the method using an amended solvation energy relationship. *J. Pharm. Sci.* **2002**, *91*, 517−533.
(11) Ran, Y.; Jain, N.; Yalkowsky, S. H. Prediction of aqueous solubility of organic compounds by the general solubility equation (GSE). *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1208−1217.
(12) Abraham, M. H.; Le, J. The correlation and prediction of the solubility of compounds in water using an amended solvation energy relationship. *J. Pharm. Sci.* **1999**, *88*, 868−880.
(13) Huuskonen, J. Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 773−777.
(14) Wanchana, S.; Yamashita, F.; Hashida, M. Quantitative structure/property relationship analysis on aqueous solubility using genetic algorithm-combined partial least squares method. *Pharmazie* **2002**, *57*, 127−129.
(15) Jorgensen, W. L.; Duffy, E. M. Prediction of drug solubility from structure. *Adv. Drug Delivery Rev.* **2002**, *54*, 355−366.
(16) Butina, D.; Gola, J. M. Modeling aqueous solubility. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 837−841.
(17) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.
(18) Giordanetto, F.; Fossa, P.; Menozzi, G.; Mosti, L. In silico rationalization of the structural and physicochemical requirements for photobiological activity in angelicine derivatives and their heteroanalogues. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 53−64.
(19) Sutherland, J. J.; Weaver, D. F. Development of quantitative structure−activity relationships and classification models for anticonvulsant activity of hydantoin analogues. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1028−1036.
(20) Bahler, D.; Bristol, D. W. The induction of rules for predicting chemical carcinogenesis in rodents. *Proc. Int. Conf. Intell. Syst. Mol. Biol*. **1993**, *1*, 29−37.
(21) Phillips, J. C.; Purchase, R.; Watts, P.; Gangolli, S. D. An evaluation of the decision tree approach for assessing priorities for safety testing of food additives. *Food Addit. Contam.* **1987**, *4*, 109−123.
(22) Xia, X.; Maliski, E.; Cheetham, J.; Poppe, L. Solubility prediction by recursive partitioning. *Pharm. Res.* **2003**, *20*, 1634−1640.
(23) Put, R.; Perrin, C.; Questier, F.; Coomans, D.; Massart, D. L.; Vander Heyden, Y. Classification and regression tree analysis for molecular descriptor selection and retention prediction in chromatographic quantitative structure-retention relationship studies. *J. Chromatogr., A.* **2003**, *988*, 261−276.
(24) Bai, J. P.; Utis, A.; Crippen, G.; He, H. D.; Fischer, V.; Tullman, R.; Yin, H. Q.; Hsu, C. P.; Jiang, L.; Hwang, K. K. Use of classification regression tree in predicting oral absorption in humans. *J. Chem. Inf. Comput. Sci*. **2004**, *44*, 2061−2069.
(25) Deconinck, E.; Hancock, T.; Coomans, D.; Massart, D. L.; Heyden, Y. V. Classification of drugs in absorption classes using the classification and regression trees (CART) methodology. *J. Pharm. Biomed. Anal.* **2005**, *39*, 91−103.
(26) Ekins, S.; Berbaum, J.; Harrison, R. K. Generation and validation of rapid computational filters for cyp2d6 and cyp3a4. *Drug Metab. Dispos.* **2003**, *31*, 1077−1080.
(27) Bederson, B.; Schneiderman, B. Ordered and quantum treemaps: making effective use of 2D space to display hierarchies. *ACM Trans. Graph.* **2002**, *21*, 833−854.

CI0504770