ELSEVIER

Journal of Molecular Graphics and Modelling 21 (2003) 333–339

Journal of Molecular Graphics and Modelling

www.elsevier.com/locate/JMGM

MolSpace: a computer desktop tool for visualization of massive molecular data

Yoshimasa Takahashi*, Mitsuru Konji, Satoshi Fujishima

Department of Knowledge-based Information Engineering, Toyohashi University of Technology, 1-1 Hibarigaoka, Tempaku-cho, Toyohashi 441-8580, Japan

Received 3 July 2002

Abstract

The authors have developed a software tool, MolSpace, to visualize massive molecular datasets. MolSpace can project a set of massive multivariate data onto a visual space (two- or three-dimensional space) by means of principal component analysis. MolSpace allows users not only to draw a scatter diagram of the data but also to display their two- or three-dimensional molecular structures as the objects in that space. With a probe (a molecular object) the user can navigate vast data spaces, thus facilitating understanding of the data structure. In addition, partial space searching is also available that is based on similarity searching techniques. It is possible to interrogate a three-dimensional structure of a chemical compound that corresponds to each object on the space in real time. The detail of the system is discussed with an illustrative example.

Keywords: Molecular similarity; Structural similarity; Data visualization; Similarity searching; Topological fragment spectra; PCA; Chemical data space

1. Introduction

The visualization of massive molecular datasets is important in the area of rational drug design, especially for understanding the latent data structure of chemical space and for finding new lead candidates. This need has become stronger with the development of combinatorial chemistry using automated synthesis and high-throughput screening techniques. Typical applications for such visualization methods include the identification of compounds with desired activity by database searching, feature analysis and modeling of the activity for data mining, selection of representative subsets from large chemical libraries or virtual libraries [1,2]. The basic idea behind these applications is that similar compounds in a molecular space are likely to possess similar chemical properties and similar biological activities [3]. Our research interests are in the area of similarity estimation for exploring leads, and in methods for selecting a representative set of chemicals to be considered as lead candidates.

There are two different aspects to similarity analysis (or diversity analysis). One is focused on the chemical space based on a variety of molecular properties. The other is focused on structural feature space in which one may find common or similar substructures among the molecules [4]. In the former case, various physical and chemical properties that can be calculated by theoretical or empirical models are often used in describing the data space [5,6]. In the latter, many kinds of molecular descriptors can be used to represent structural features [7–9]. Graph theoretical indices also can be used as numerical descriptors for the ordering of chemical structures [10–12]. In either case we often have to handle massive datasets that derive from the modern methods of combinatorial drug design and development.

In the present work, we have developed a software tool, MolSpace, to visualize multivariate molecular data space. MolSpace can project a set of massive multivariate data onto a visual space (two- or three-dimensional space) by means of principal component linear mapping method. MolSpace allows us not only to draw a scattering diagram of the data, but also to display their three-dimensional molecular structures as the objects on the space. With a probe (object) the user can explore the data space, facilitating understanding of the data structure. In this paper, the basic concept of the system and the implementation are described with the illustrative examples.

2. Methods

2.1. Molecular descriptors and the feature space

Generally, understanding of the structural features of molecules which are relevant to biological activity is extremely

* Corresponding author.

1093-3263/02/$ – see front matter © 2002 Elsevier Science Inc. All rights reserved.
PII: S1093-3263(02)00180-8

important in a wide range of research. Structural feature analysis can be done by manual approaches for a small set of molecules. However, such methods become quite tedious and time consuming for a large set of molecules. Many other computational approaches use a set of predefined substructure fragments such as substructure search screens of chemical information systems. The set of substructure fragments can be used as descriptors for depicting molecular data space. For example, *d* selected descriptors define *d*-dimensional data space for the molecules, and each molecule under investigation is assigned as coordinates in this space based on the values of the descriptors. Such molecular data space is often used to evaluate intermolecular similarity or diversity [7]. In general, two-dimensional substructure descriptors are extremely useful for large-scale data because most chemical databases contain only connection tables without three-dimensional information [13]. However, quantification of such structural similarity still has problems of the dependency on the chosen set of substructures defined as the descriptors in advance. To avoid the problem, MolSpace employs the topological fragment spectra (TFS) method as the following discussion shows.

2.2. Topological fragment spectrum (TFS)

The authors proposed topological fragment spectra (TFS) method [14] as a tool for describing the topological structure profile of a molecule. The approach is based on enumeration of all possible substructures from a chemical structure and numerical characterization of them. The procedure involves two main steps: (1) enumeration of all possible substructures from each chemical structure, (2) numerical characterization of the substructures.

Firstly, for a given structure represented as a chemical graph (hydrogen suppressed graph), all the possible substructures embedded in it are enumerated. Here, all the substructures of the parent structure (the original chemical graph) are taken into account in the following processes. Subsequently, all the individual substructures are characterized numerically. To perform this characterization we have used a method which computes the overall sum of the mass numbers of the atoms (atomic groups) corresponding to the nodes of the chemical subgraph enumerated. For the method, attached hydrogen atoms are taken into account as augmented atoms and are represented by weighting correspondingly their respective nodes in the chemical graph. The TFS is defined with the histogram obtained by displaying the frequency distribution of a set of individually characterized substructures (structural fragments) according to the value of their characterization index. An illustrative scheme of the procedure is shown in Fig. 1.

Fig. 1 shows an example for enumerating all the possible substructures with 2-methylbutane and generating the TFS. Here, every substructure is characterized by the sum of atomic mass numbers of constituent atoms that involve the nodes of subgraph obtained from the chemical graph of 2-methylbutane. Fig. 1(b) shows the histogram of the characterized substructures. The histogram is referred as a TFS.

This approach was fully computerized and used in the following structural similarity analysis and similar structure searching in a chemical database. The fragment spectrum generated according to this manner is a representation of the topological structural profile of a molecule. It should be noticed that the system does not require any predefined list of substructures to be considered. The fragment spectrum of promazine that is characterized by the method mentioned above is shown in Fig. 2.

Generally, the computational time required for the exhaustive enumeration of all possible substructures from a chemical structure is often very large especially for molecules which contain highly fused rings. In addition to this, a large difference in the dimensionality between the fragment spectra to be compared may lead to the unexpected result. To avoid the problems, we employed a subspectral approach, in which every TFS are described with structural fragments that have the specified size or less. The size is the number of bonds of the fragment or edges of the subgraph to be generated. In this work, all the fragments with the size of 5 or less were used for generating TFS.

2.3. Quantitative evaluation of structural similarity based on the TFS

Obviously, the fragment spectrum obtained by the method mentioned above can be described as a kind of multidimensional pattern vector. Consequently, using the pattern representation of a spectrum it is possible to apply various quantitative measures for the evaluation of similarity. In the present system, Euclidean distance, the cosine and Tanimoto coefficients [7] can be used for evaluating the similarity or the dissimilarity. The similarity and dissimilarity between TFS pattern vectors X_i and X_j is calculated with the following equations:

$$D(X_i, X_j) = \sqrt{\sum (x_{ik} - x_{jk})^2} \quad \text{(Euclidean distance)} \qquad (1)$$

$$C(X_i, X_j) = \frac{\sum (x_{ik} x_{jk})}{\sqrt{\sum (x_{ik})^2 \sum (x_{jk})^2}} \quad \text{(Cosine coefficient)} \qquad (2)$$

$$T(X_i, X_j) = \frac{\sum (x_{ik} x_{jk})}{\sum x_{ik}^2 + \sum x_{jk}^2 - \sum (x_{ik} x_{jk})} \quad \text{(Tanimoto coefficient)} \qquad (3)$$

where x_{ik} and x_{jk} are pattern vectors which represent the frequency value of peak *k* of fragment spectra of *i*th molecule and *j*th molecule, respectively. The different dimensionalities of the spectra to be compared are adjusted

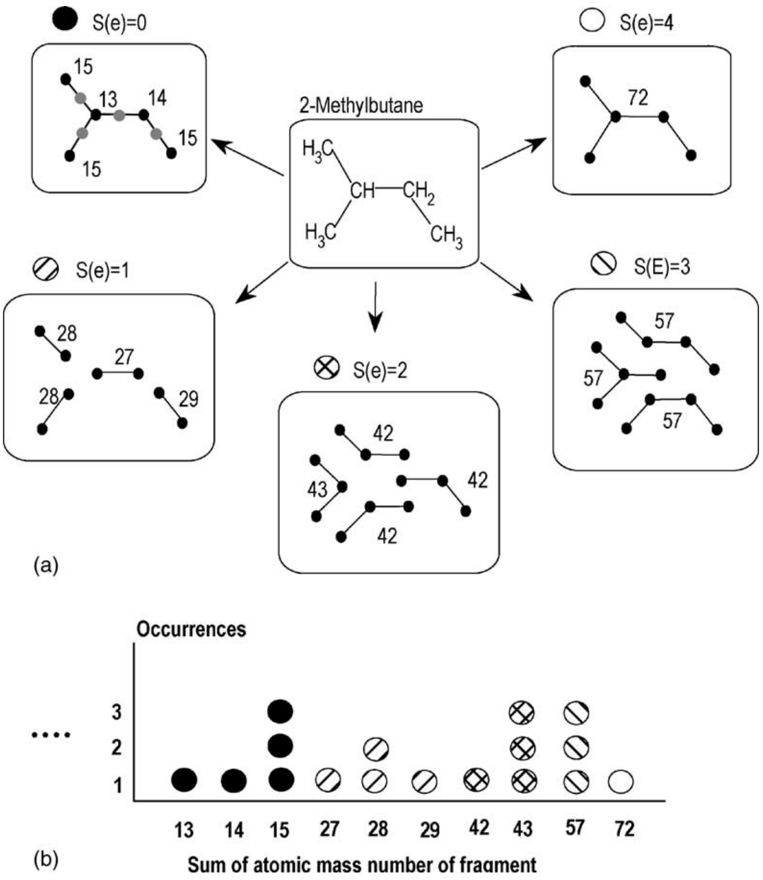


Fig. 1. An illustrative scheme of the procedure of TFS generation. (a) Enumeration and characterization: every substructure is characterized by the sum of atomic mass numbers of constituent atoms that involve the nodes of subgraph obtained from the chemical graph of 2-methylbutane. $S(e)$ is the size(number of edges) of subgraph to be enumerated. (b) The histogram of the characterized substructures. The histogram is referred as a TFS.

as follows:

if $X_i = (x_{i1}, x_{i2}, \ldots, x_{iq})$ and

$$X_j = (x_{j1}, x_{j2}, \ldots, x_{jq}, x_{j(q+1)}, \ldots, x_{jp}) \quad (q < p),$$

then X_i is redefined as

$$X_i = (x_{i1}, x_{i2}, \ldots, x_{iq}, x_{i(q+1)}, \ldots, x_{ip}) \tag{4}$$

where

$$x_{i(q+1)} = x_{i(q+2)} = \cdots = x_{ip} = 0.$$

The searching process was also computerized and used in the following structural similarity analysis and similar structure searching in a chemical database.

2.4. Visualization of the TFS data space

To visualize the data structure in the high-dimensional TFS space, it must be transformed into two- or three-dimensional space in which the original data structure should be retained as closely as possible. To achieve this dimensional reduction, a technique of principal component analysis (PCA) [15] based the eigenvalue analysis of the data

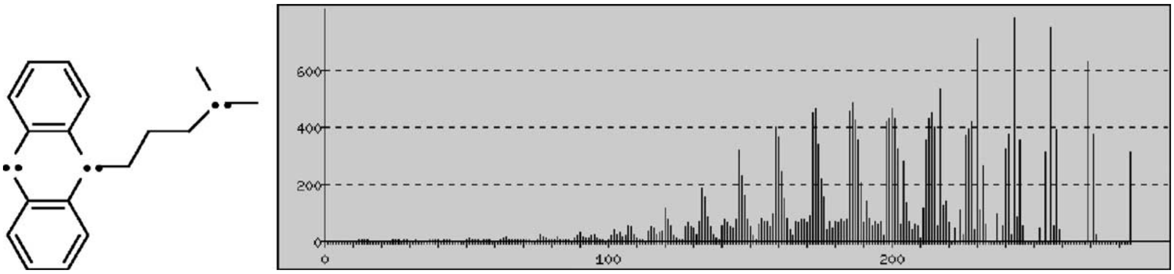


Fig. 2. TFS of promazine that was characterized by the sum of atomic mass numbers.

matrix is used in the present system. Let a data matrix $\boldsymbol{X}$ be a set of descriptors expressing n samples by d descriptors. In the principal component analysis, d eigenvalues λ_β ($\beta = 1, 2, \ldots, d$) and the eigenvector matrix $\boldsymbol{T}$ are obtained by computing the covariance matrix $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ of the data matrix $\boldsymbol{X}$ and diagonalyzing it:

$$\boldsymbol{T}^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}) = \Lambda = \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_d \end{bmatrix} \tag{5}$$

where $(\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d)$ and

$$\boldsymbol{T} = (\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_d) \tag{6}$$

$\boldsymbol{X}^{\mathrm{T}}$ denotes the transpose matrix of $\boldsymbol{X}$. The principal component matrix $\boldsymbol{Z}$ is determined by the eigenvector as follows:

$$\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{T} \tag{7}$$

Then, considering r principal components with large eigenvalues, let us reduce the dimensionality of the data matrix. For more details of PCA see [15]. Each principal component is a linear combination of original descriptors. When the data has a large number of interrelated variables, PCA is a very powerful method for analyzing the intrinsic data structure and reducing the dimensionality of the data space. MolSpace is an enhanced PCA tool for doing this.

3. Results and discussion

3.1. Implementation of the system

MolSpace provides us a desktop tool for multivariate chemical data analysis. The system consists of five basic modules; file handling, data reduction, scattering diagram visualization, partial space searching and molecular viewer. These functions offer the useful visual information to users. For example, the data reduction module makes it possible to visualize the higher-dimensional data space of molecular objects. The scattering diagram viewer can display not only the scattering map of the data by the point object representation but can also map two- or three-dimensional chemical structures when they are available. MolSpace allows users to navigate the data space on the screen. To investigate individual chemical structures, the molecular viewer allows display of three-dimensional molecular models in several ways; simple wire frame, stick model, ball and stick model, etc. The user can select any molecule in the data space by means of a mouse click operation on the PC screen. The partial space searching module allows users to clip out a part of data space and view the subspace. An outline of the data visualization process in the case of using TFS module that is offered in the system is shown in Fig. 3. The computer programs were written in Microsoft Visual C/C++ (Version 6.0) and OpenGL [16] under the Windows 98 operating system. Compaq Visual Fortran (Version 6.0) was also used for some of the function programs.

3.2. Example of data visualization by MolSpace

Examples of massive dataset visualization by MolSpace are shown in Fig. 4. A subset of World Drug Index (WDI) database [17] was used for a trial to illustrate the ability of MolSpace. The trial database consists of 3600 drugs randomly selected from the WDI database. The TFS of all the molecules were generated and they were prepared as a TFS database by using an inherent function of MolSpace. Fig. 4(a) shows a scattering map of the 3600 drug molecules in the TFS space. Here, all the TFS data were reduced into a three-dimensional space by PCA technique.

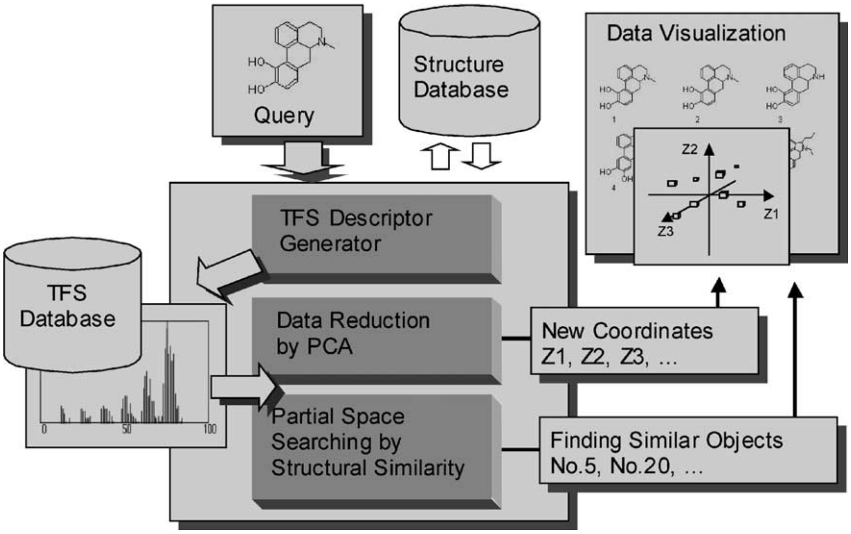
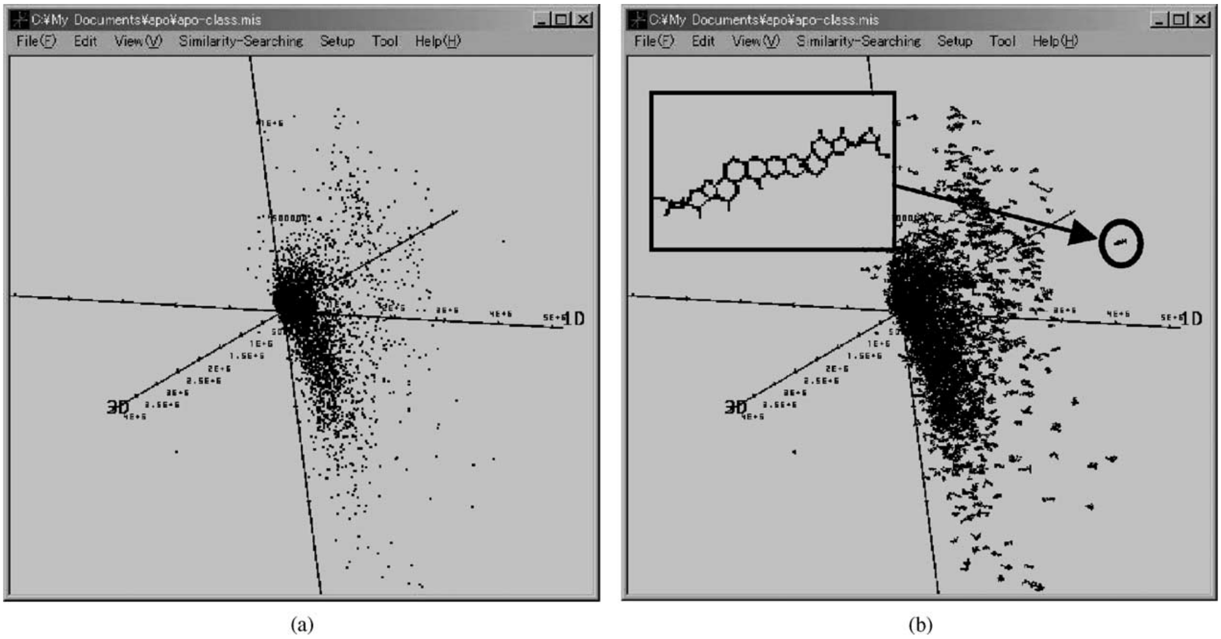

Query
Structure Database
Data Visualization
TFS Descriptor Generator
TFS Database
Data Reduction by PCA
Partial Space Searching by Structural Similarity
New Coordinates Z1, Z2, Z3, ...
Finding Similar Objects No.5, No.20, ...
Z1
Z2
Z3
0
50
100

Fig. 3. A schematic diagram of visualization of TFS data space by MolSpace.


C:¥My Documents¥apo¥apo-class.mis
File(F) Edit View(V) Similarity-Searching Setup Tool Help(H)
1D
3D
C:¥My Documents¥apo¥apo-class.mis
File(F) Edit View(V) Similarity-Searching Setup Tool Help(H)
1D
3D

(a)

(b)

Fig. 4. Examples of massive data visualization by the MolSpace. (a) Scattering map on the three-dimensional PCA space for the TFS data of 3600 drugs with the point object representation. (b) Scattering map on the same space with the structure object representation.

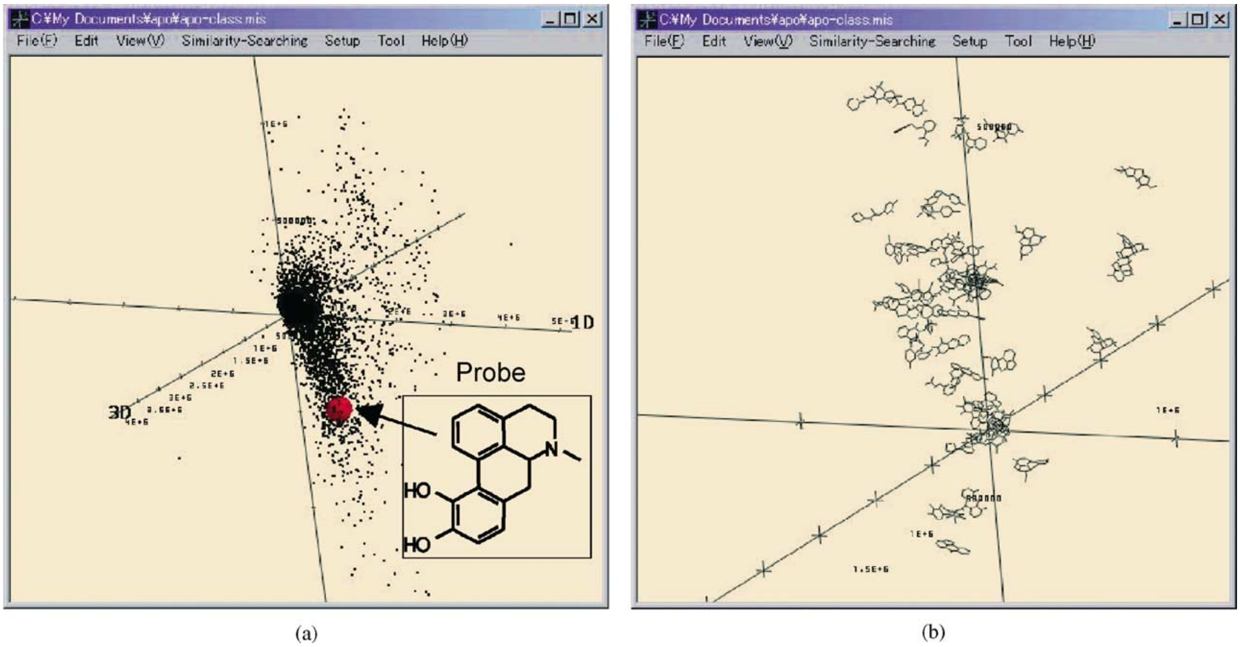
The system can also display a scattering map of the data in structure object representation when the structure data are available for the system (Fig. 4(b)). But they are dispensable. In the present work, the three-dimensional coordinates of all the molecules were calculated by the CONCORD [18] in advance, and stored with the connection tables. For the present system our own format is required in preparing the structure data database. However, it is possible to convert the data of MDL format (molfile) into that of the MolSpace by a conversion program. The details of the file specification will be available with the system from the authors.

The TFS data space is a kind of structural feature space. For the reason, structurally similar molecules would be located on the data space near each other. Thus, the user can find similar objects in the subspace of interest. However, it is still hard to locate the subspace where the particular molecule of the interest is located in the entire data space. To overcome this problem, MolSpace allows us to use a probe molecule. The probe molecule can be mapped approximately onto the data space using the mapping vectors (i.e. eigenvectors) obtained by PCA for the current trial data. The result shows users the region of feature space where the desired structures may be located. Fig. 5(a) shows an example of a probe point mapped onto the TFS data space.

3.3. Partial space searching

The MolSpace also allow us to explore the particular subspace with a probe object (molecule) and clip it out. Then the users can easily go into the space on the screen of MolSpace. Exploration is carried out using similarity searching. Here, the partial space searching was employed for understanding the structure of a data subspace that involves the molecular objects similar to the probe of apomorphine. The trial was carried out for the TFS space of 3600 drug molecules mentioned above. Fig. 5 shows the probe point of apomorphine mapped onto the TFS data space and an example of the partial space searching with the probe.

The system shows the location of the probe with a colored marker (Fig. 5(a)). On the basis of TFS, similar searching was carried out. At searching the users can specify the number of molecules to be searched. After searching the partial space that involves the hit molecules is automatically clipped from the TFS space, where the structurally similar objects are located in near each other. For the case, the origin of the data space is moved to the probe point. In MolSpace, the similarity searching is always carried out in the original data space but not in the reduced space obtained by PCA. For the searching, MolSpace generates the TFS of the probe molecule and evaluates the similarity between the TFS and those of database molecules with a specified measurement function. For the present trial, Euclidean distance measure was used for evaluating the similarity. The clipping of the partial space is to visualize only the data extracted from the entire TFS space. Fig. 5(b) shows the result of a partial space view of 50 most similar molecules found and clipped by the similar structure searching. This function allows us identify artifacts mapped into the same space due to PCA mapping errors. MolSpace also offers us an additional tool of molecular viewer, called MolView, which can be used for displaying and rotating a three-dimensional molecular model in several ways, e.g. stick, ball & stick and space filling models. Fig. 6 shows an example for using the viewer. Fig. 6(a) is a screen snapshot of MolSpace when we went


C:¥My Documents¥apo¥apo-class.mis
File(F) Edit View(V) Similarity-Searching Setup Tool Help(H)
1D
3D
Probe
N
HO
HO
C:¥My Documents¥apo¥apo-class.mis
File(F) Edit View(V) Similarity-Searching Setup Tool Help(H)

(a)

(b)

Fig. 5. Partial space searching with the probe of apomorphine. (a) A probe point mapped onto the TFS data space. (b) Partial space that involves first 50 structurally similar molecules. The origin of the new space is located on the probe of apomorphine. The similar structure searching was carried out in the original data space but not in the reduced data space obtained by PCA.

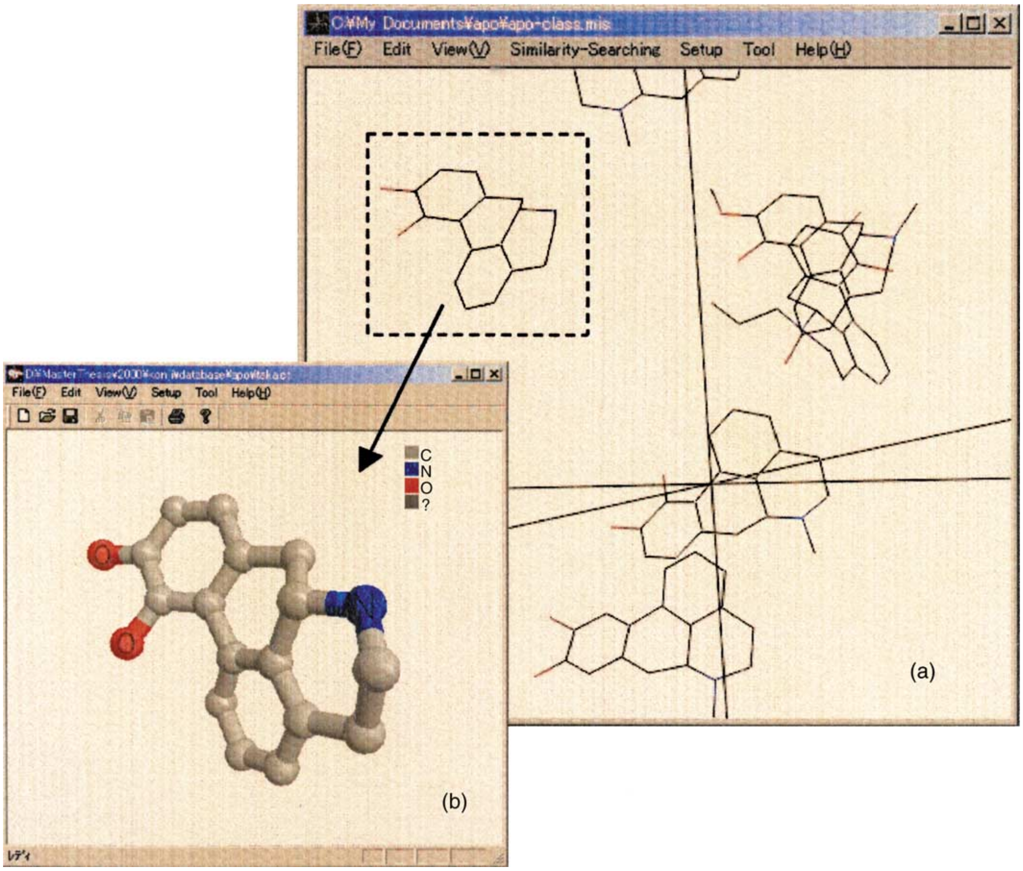

C:¥My Documents¥apo¥apo-class.mis
File(F) Edit View(V) Similarity-Searching Setup Tool Help(H)
File(F) Edit View(V) Setup Tool Help(H)
C
N
O
?
(a)
(b)
レディ

Fig. 6. Partial space view and three-dimensional molecular viewer. (a) Magnified partial space view around the probe of apomorphine of Fig. 5(b). (b) A three-dimensional molecular model of the object specified on the screen.

into the subspace of Fig. 5(b). Fig. 6(b) shows the result for displaying with the ball and stick model of a molecule that is located near to the probe of apomorphine.

4. Conclusion

The software tool, MolSpace, has been developed for visualizing massive multivariate molecular data space. MolSpace allows us to easily navigate in huge data spaces. It provides partial space searching with a probe molecule. It is possible to locate a two- or three-dimensional structure (if available) of a molecule that corresponds to each object in the reduced data space in real time. This feature makes it easy to compare an object molecule with neighbors in the same region of data space. It also should be noted that the system can be used not only in TFS space but also in any other multivariate data space when you prepare the data file in the specified format elsewhere. Most of functions of the MolSpace can be used for visualizing a large set of multivariate chemical data without three-dimensional structure data. We believe that MolSpace provides us a useful tool for multivariate chemical data analysis.

Acknowledgements

Part of this work was supported by Grant-in-Aid for Scientific Research on Priority Areas (B) 13131210 and Grant-in-Aid for Scientific Research on Priority Areas (B) 12480088.

References

[1] J. Bajorath, Selected concepts and investigations in compounds classification, molecular descriptor analysis, and virtual screening, J. Chem. Inf. Comput. Sci. 41 (2001) 233–245.

[2] M. Rarey, M. Stahl, Similarity searching in large combinatorial chemistry spaces, J. Comput.-Aided Mol. Des. 15 (2001) 497–520.

[3] M.A. Johnson, G.M. Maggiora (Eds.), Concepts and Applications of Molecular Similarity, Wiley, New York, 1990.

[4] Y. Takahashi, Identification of structural similarity of organic molecules, Top. Curr. Chem. 174 (1995) 105–133.

[5] S.K. Kearsley, S. Sallamack, E.M. Fluder, J.D. Andose, R.T. Mosley, R.P. Sheridan, Chemical similarity using physicochemical property descriptors, J. Chem. Inf. Comput. Sci. 36 (1996) 118–127.

[6] D.J. Livingstone, The characterization of chemical structures using molecular properties, a survey, J. Chem. Inf. Comput. Sci. 40 (2000) 195–209.

[7] P. Willett, J.M. Barnard, G.M. Downs, Chemical similarity searching, J. Chem. Inf. Comput. Sci. 38 (1998) 983–996.

[8] R. Benigni, G. Gallo, F. Giorgi, A. Giuliani, On the equivalence between different descriptions of molecules, J. Chem. Inf. Comput. Sci. 39 (1999) 575–578.

[9] L. Xue, J. Godden, J. Bajorath, Evaluation of descriptors and minifingerprints for the identification of molecules with similar activity, J. Chem. Inf. Comput. Sci. 40 (2000) 1227–1234.

[10] M. Randic, On characterization of molecular branching, J. Am. Chem. Soc. 97 (1975) 6609–6614.

[11] L.B. Kier, L.H. Hall, Molecular Connectivity in Chemistry and Drug Research, Academic Press, New York, 1976.

[12] L.H. Hall, L.B. Kier, Molecular similarity based on novel atom type electro topological state indices, J. Chem. Inf. Comput. Sci. 35 (1995) 1074–1080.

[13] R.D. Brown, Y.C. Martin, Use of structure–activity data to compare structure-based clustering methods and descriptors for use in compounds selection, J. Chem. Inf. Comput. Sci. 36 (1996) 572–584.

[14] Y. Takahashi, H. Ohoka, Y. Ishiyama, Structural similarity analysis based on topological fragment spectra, in: R. Carbo, P. Mezey (Eds.), Advances in Molecular Similarity, vol. 2, JAI Press, Stamford, CT, 1998, pp. 93–104.

[15] S. Wold, K. Esmensen, P. Geladi, Principal component analysis, Chemom. Intell. Lab. Syst. 2 (1987) 37–52.

[16] Clayton Walnum_Win32 Programming with OpenGL Toppan Tokyo, 1996.

[17] World Drug Index Darwent Inc. Release, 1999.

[18] CONCORD, A program for the rapid generation of high quality approximate 3-dimensional molecular structures, The University of Texas at Austin and Tripos Associates, St. Louis, MO.