APPLICATION OF THE STEPWISE CLUSTERING METHOD FOR EFFICIENT DRAWING OF BIOPOLYMER STRUCTURES

Mayumi OYAMA
Information Processing Research Center, Kwansei Gakuin University, 1-1-155 Uegahara, Nishinomiya, 662 Japan

Shin-ichi SASAKI
Department of Knowledge-based Engineering and Science, Toyohashi University of Technology, 1-1 Hibarigaoka, Tenpaku, Toyohashi, 440 Japan

and

Mototsugu YOSHIDA
Takatsuki Research Laboratory, Sumitomo Chemical Co., Ltd., 2-10-1 Tsukahara, Takatsuki, 569 Japan

Abstract

The reduction of drawing time is desirable in writing a complex molecular structure by use of a plotter. The stepwise clustering method is applied to find efficient drawing paths for six structures of protein and DNA molecules. All the optimization ratios of path lengths exceed 50%, while the necessary CPU times are of the order of 0.1 second. These results show the effectiveness of the method for molecular graphics. A summary of the algorithm and other possible applications are also discussed.

1. Introduction

In chemical research and education, the preparation of high quality hard copies illustrating molecular structures is an important task. The drawing of molecular structures by a plotter requires much time, since it is constituted of relatively slow mechanical pen movements. However, the plotter often takes the pen-up state in order to change its pen position from one line segment to another. So, the necessary drawing time would be reduced if we could find a shorter path which connects all the given line segments. This subject is called the problem of efficient drawing path, which is defined as the determination of a short path passing through all the line segments at least once. This problem becomes very important when we want to draw complex chemical structures such as those found in biopolymers.

The efficient drawing path problem is a kind of combinatorial optimization problem. In fact, if every line segment shrinks and changes to a point, this problem is simplified to a traveling salesman problem on a perfect graph. The number of possible combinations which can form a path from the given line segments increases

exponentially as the number of segments increases. Therefore, an exhaustive search of the shortest path requires an extremely large amount of time, even if a high-speed computer is used.

Various heuristic methods were proposed to find an approximate solution for the above-mentioned problem [1,2]. However, there are few efficient methods which can deal with the optimization in connecting the line segments scattered in a three-dimensional space.

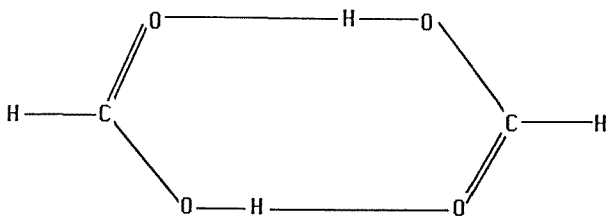
One of the authors of this paper formulated the stepwise clustering method to solve this problem [3,4]. It can effectively determine an efficient solution for the drawing of maps, Chinese characters, and randomly generated line segments. The application of the stepwise clustering method to the drawing of biopolymer structures is interesting in the following two respects. First, we can check the adequacy of the clustering method for the line segments in a three-dimensional space by using the geometries determined by X-ray crystallography. Secondly, it is significant in relation to the character of the molecular structure data. That is, in contrast to the other data so far examined, chemical bonds have nearly uniform lengths and most of them are connected to each other. Thus, if the stepwise clustering method is proved to be useful in molecular structure drawing, the effectiveness is not limited to the practical availability in drawings, but we can add evidence to show the generality of the heuristics employed in this method.

In the next section of this paper, we first briefly introduce the stepwise clustering method. Then, its applications to the drawing of biopolymer structures are shown in the following section. Lastly, we discuss the effectiveness of the method in various fields of science and technology. The possibility of relating the clustered substructures to chemically meaningful entities is also examined.

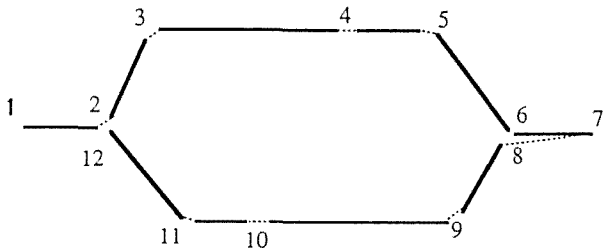
2. Stepwise clustering method

Let us think of the task of drawing the structure of a formic acid dimer depicted in fig. 1(a). Figures 1(b) and (c) show two examples of drawing paths, both of which pass through all the bonds denoted by the solid lines once in their respective ways. The total path length is definitely shorter in the case of (b), since the length of the dotted lines denoting the unnecessary pen-up-state movements in (b) is shorter than that in (c). The efficient drawing path problem is to find a near optimum short path like (b) within a practical computation time.

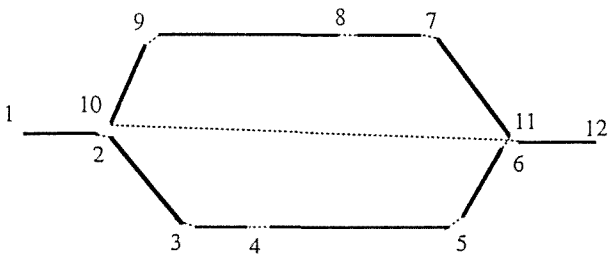
We construct the path (c) by starting at the extreme left node, and by tracing the line segments which share terminals. The longer segment is traced when there are two candidates, as at the 2nd and the 6th nodes. On the other hand, path (b) is devised by the stepwise clustering method, and it is one of the shortest paths to draw this diagram. Figure 2 illustrates the procedure to construct the shortest path (b) in fig. 1 following the stepwise clustering method on a two-dimensional plane. In this example, the process is postulated as follows.



(a)



(b)



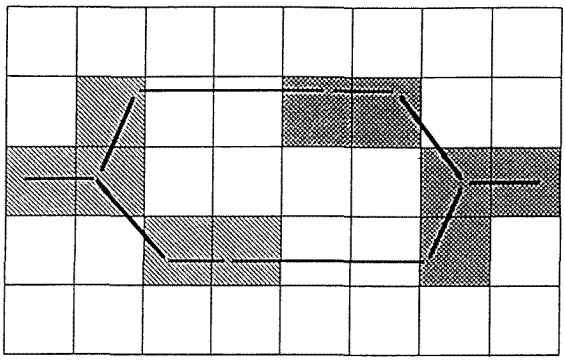
(c)

Fig. 1. Two sample paths to draw the structure of a formic acid dimer. (a) Hydrogen-bonded formic acid dimer; (b) sample of the shortest path; (c) sample of a longer path. In (b) and (c), the numbers show the sequence of pen movements. The solid and dotted line segments indicate the pen movements with pen-down and pen-up states, respectively.

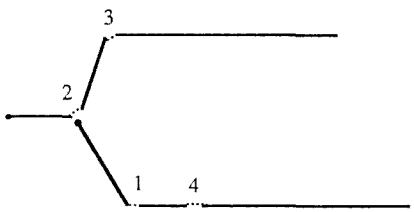
(1) The drawing area is divided into cells, as shown in fig. 2(a). The cell size h is determined by the following formula:

$$h = ((\Delta x \cdot \Delta y)/n \cdot \lambda)^{1/2},$$

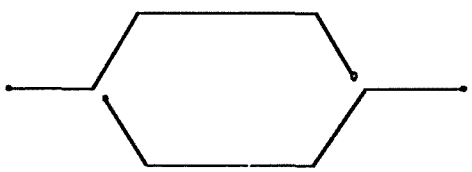
where Δ's are the differences between the maximum and the minimum coordinates of the terminals of the line segments and n is the number of terminals. λ is an empirical parameter to provide an adequate cell size.



(a)



(b)



(c)

Fig. 2. Sample procedure to generate a drawing path by the stepwise clustering method. (a) Grid formation and recognition of clusters. The cells hatched by ☐ and ☐ show the two clustered regions, respectively. (b) Order of intracluster connections. Dotted lines show the connections between line segments, • indicates a terminal of the segment after connections. (c) Two new segments.

(2) Clustering regions are generated by aggregating the cells which contain segment terminals and share a corner or a side with each other. Terminals in a clustering region make up a cluster, and they are to be connected to form an efficient drawing path. Two clusters appear, in fig. 2(a), each of which contains ten terminals of six line segments.

(3) A path in the clustering region is given by connecting segment terminals one by one. We use the following criteria to select the terminals to be connected:

(a) Calculate the distances between all the terminal pairs, and connect two terminals with the shortest distance.

(b) When there are plural terminal pairs which share the shortest distance, select a pair in which the shortest segment participates.

(c) If there still exists more than one candidate connection, compare the length of the other segment in the connections and choose the one with the shortest length. However, the connection is forbidden in the following two cases. Firstly, it is not permitted when the terminals at both ends of a segment are not connected. Secondly, a connection is not allowed if both segments span to other cluster regions. Here, we note that the maximum value norm is employed in the comparison of terminal distances and segment lengths to save computation time. The intracluster connection procedure does not execute an exhaustive search, but employs the nearest neighbor search. This choice is reasonable since the loss cannot be large in the intracluster region, even if the nearest neighbor connection does not lead to the best path.

The numbers in fig. 2(b) show the order of intracluster connections derived by the above criteria. The connnected segments constitute a new segment to be used as part of the final drawing path. Figure 2(c) illustrates two new segments derived from the connections in (b) after the execution of this step.

(4) Steps (1)–(3) are repeated using the new segments generated. The decrease in the number of segments expands the cell size in the new iteration. Then, some line segments previously spanning two clustering regions are now transformed into intracluster segments. They can be connected with other segments to make up a larger segment in the new iteration.

When the number of segments becomes sufficiently small, the iteration stops and the whole drawing area is considered to be one clustering region. Then, the intracluster connections give us only one segment, which is the final drawing path. The four terminals of the two segments in fig. 2(c) are regarded as belonging to a cluster, and they are connected to give the path in fig. 1(b). For the effective execution of the stepwise clustering method, the key factor is to use an adequate cell size. If the cell size is too large, there appears to be only one cluster. On the contrary, a too small cell size gives too many clusters, most of which contain only one segment terminal. Effective and efficient intracluster connections of segments cannot be expected in these cases. Therefore, the actual program changes the parameter λ dynamically so that the cell size is adjusted to a reasonable value range.

Lastly, we examine the necessary number of path length comparisons in the following three methods, and we show the efficiency of the current method.

I. Exhaustive search method.

II. Stepwise clustering method with exhaustive search for intracluster connections.

III. Stepwise clustering method with nearest neighbor search for intracluster connections (current method).

Assuming that there are n line segments, the number of paths generated by the exhaustive search is

$$2^{n-1}n!. \tag{1}$$

On the other hand, the nearest neighbor search necessitates length comparisons of $4n(n-1)/2$ segments to find a pair of terminals to be connected. Thus, the total number of segments to be compared is not more than

$$\sum_{i=2}^{n} 2i(i-1) = 2n(n+1)(n-1)/3. \tag{2}$$

We assume that q steps of clustering take place in the stepwise clustering process, and that there exist k line segments in every cluster. That is, there are no segments spanning to two clusters, and k segments in a cluster are always reduced to one segment by intracluster connections. The initial number of segments n is given by k^{q+1}. Then the number of paths to be compared is

$$2^{k-1}k!(k^q + k^{q-1} + \ldots + k + 1) \tag{3}$$

in method II, while method III gives

$$2k(k+1)(k-1)(k^q + k^{q-1} + \ldots + k + 1)/3 = 2k(k+1)(k^{q+1} - 1)/3. \tag{4}$$

Table 1 illustrates common logarithms of the necessary number of comparisons for $k = 5$, 10 and $q = 0-3$ using the above equations. We can see that the number of comparisons increases linearly with n in the stepwise clustering method, and that the current method is efficient even if the number of segments in a cluster becomes large.

Table 1

Number of length comparisons in the three methods

Clustering steps	$k = 5$				$k = 10$			
	No. of segments	log no. of comparisons[a]			No. of segments	log no. of comparisons[a]		
		I[b]	II	III		I[b]	II	III
0	5	3.3	3.3	1.9	10	9.3	9.3	2.3
1	25	32.4	4.1	3.4	100	181.0	10.3	3.3
2	125	247.0	4.8	4.1	1000	2800.0	11.3	4.3
3	625	1670.0	5.5	4.8	10000	38000.0	12.3	5.3

[a] Method I : Exhaustive search method.
Method II : Stepwise clustering method with exhaustive search for intracluster connections.
Method III : Stepwise clustering method with nearest neighbor search for intracluster connections (current method).

[b] Stirling's approximation is used to compute the factorial for $q = 1-3$.

3. Application to the drawing of biopolymer structures

We have applied the stepwise clustering method to draw the structures of proteins and DNAs listed in table 2. The atomic coordinates and the connectivities of these molecules are taken from the Protein Data Bank of the Brookhaven National Laboratory.

Table 2

The results of stepwise clustering

Molecule no.[a)]	No. of line segments	Parameter	No. of iterations	CPU time [msec][b)]	Optimization ratio [%][c)]
1	940	20	13	145.1	67
2	543	19	5	116.9	68
3	529	20	8	129.5	66
4	544	18	7	39.4	55
5	791	16	10	152.6	68
6	550	20	17	53.7	56

[a)]The molecules are

1: Cytochrome C (rice),
2: Insulin-like growth factor I (Somatomedin),
3: Insulin-like growth factor II (Somatomedin),
4: B-form DNA (crystal).
5: DES-*PHE B1 insulin,
6: DNA¥(Z-II,5*-¥D(P*CP*GP*CP*GP*CP*GP*CP*GP*CP*GP*CP*G)-3*).

[b)]CPU time shows the execution time of the program on a HITAC M-680D mainframe computer.

[c)]The optimization ratio is computed by the formula $100(Xi - X0)/Xi$; where Xi and $X0$ are path lengths before and after the optimization, respectively. The path for Xi follows the sequence of bonds in the Protein Data Bank.

The procedure described in the previous section is expanded to handle the line segments distributed in a three-dimensional space. An application of this procedure to the geometry of a molecule is expected to give a short path that can trace all the bonds in the three-dimensional space. Once we obtain this path, the structure projected onto a plane from any direction can be drawn effectively following the path.

Table 2 shows the ratio of optimization obtained by the stepwise clustering method, as well as the used CPU time and the number of iterations.

Figure 3 shows the planar projections of the following data for three molecules.

(a) Molecular structure to be drawn.

(b) Drawing path with pen-up state before optimization.

(c) Drawing path with pen-up state after optimization.

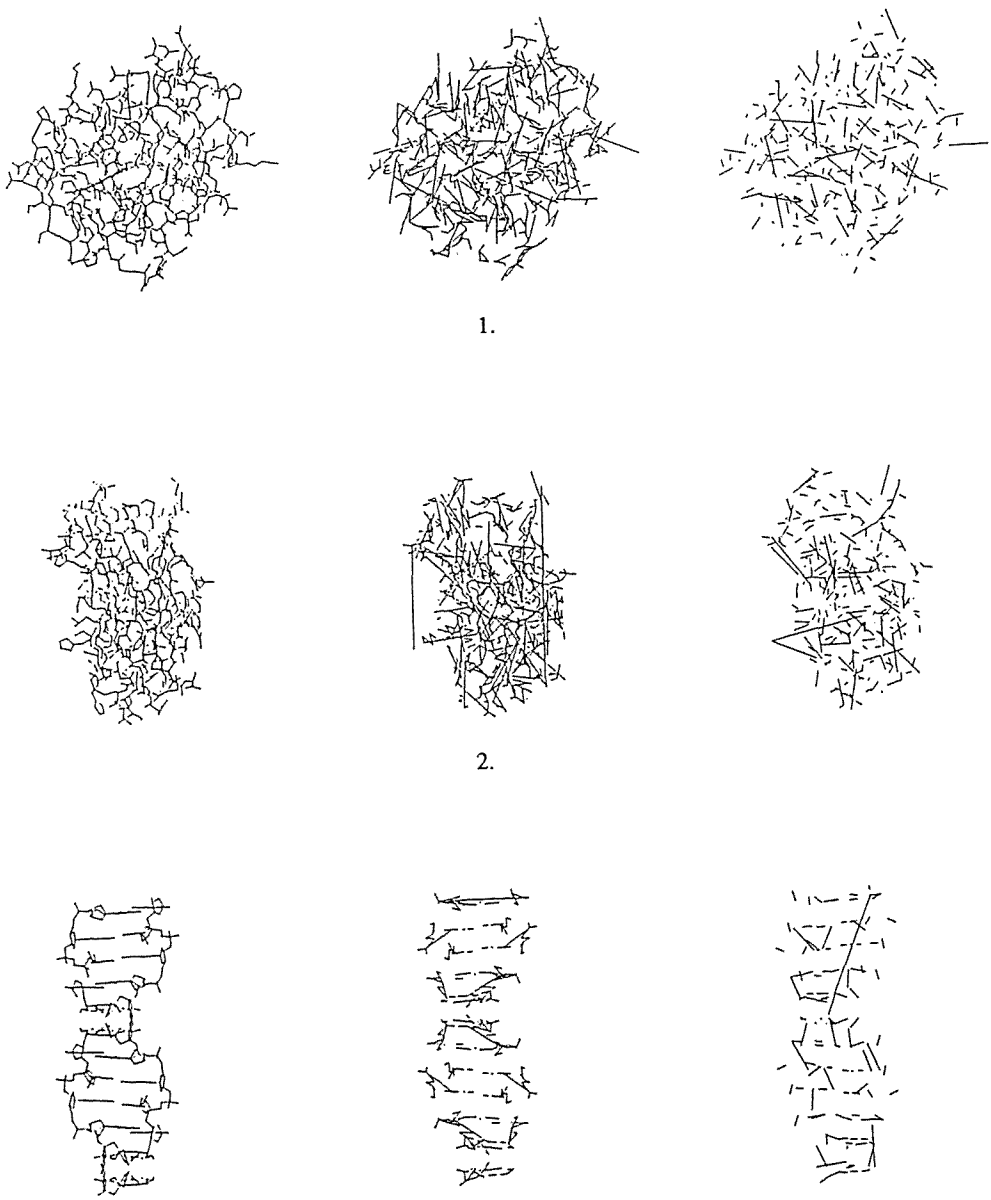


Fig. 3. Optimization of drawing path by the stepwise clustering method. 1. Cytochrome C (rice). 2. DES-* B1 insulin. 3. DNA¥(Z-II,5*-¥D(P*CP*GP*CP*GP*CP*GP*CP*GP*CP*GP*CP*G)-3*): (a) Molecular structure; (b) path with pen-up state before optimization; (c) path with pen-up state after optimization.

Since the effective drawing path problem is to minimize the pen-up state movements, the difference between the path lengths of (b) and (c) in this figure is a measure of the effectiveness of the stepwise clustering method. When we use a Hitachi plotter (H-8292-2) with 100 m/sec pen movement to draw Cytochrome C of fig. 3(a) on a 30 cm × 45 cm area, it takes 265 sec and 158 sec before and after optimization, respectively.

The optimization ratio and the CPU time in table 2 indicate that the current method can provide fairly good drawing paths within a reasonable CPU time. Therefore, we can state that the stepwise clustering method can be applied effectively to draw a diagram consisting of many line segments, even if they have nearly uniform lengths and they share the terminals frequently.

Here, we have to mention that the value of the parameter λ necessary to determine the cell size is set to fairly large values, between 16 and 20, as shown in table 2. That is, we begin the calculation employing a relatively small cell size. If we employ a larger cell size, the two terminal atoms of a bond are likely to be placed in the neighboring cells and then all terminal atoms in a molecule tend to form a single cluster. Even if the initial value of the parameter λ is not adequate, we anticipate that the automatic justification of the cell size leads to the formation of an efficient drawing path, though the necessary CPU time becomes greater.

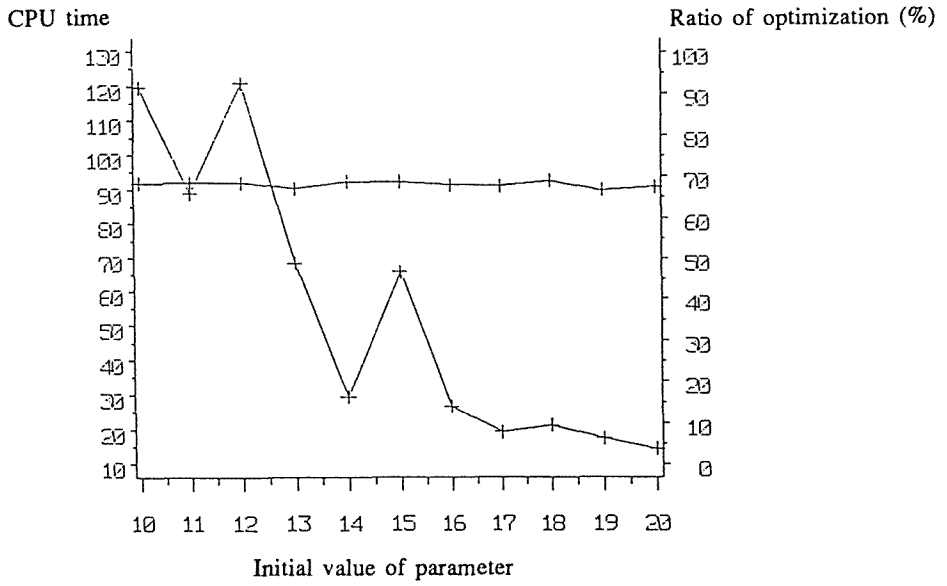


Fig. 4. Effects of the change in the initial value of the parameter λ to the optimization ratio and the CPU time in the drawing of Cytochrome C. Bars show the ratio of optimization. Crosses show the used CPU time.

Figure 4 shows the effects of a variation in the parameter value λ in relation to the efficiency of the resulting path and to the computation time for the drawing

of the Cytochrome C structure. The parameter λ is fixed to the cited value in all the iterations. As expected from the characteristics of the molecular structure data, a longer CPU time is necessary if λ becomes less than 16, while the ratio of optimization is almost constant.

4. Conclusions

The stepwise clustering method is proved to be very effective in the drawing of complex molecular structures. Hence, we anticipate that this method can be well applied to a wide range of drawing problems when we take into account the successful applications to the drawings of Chinese characters and maps so far examined.

The stepwise clustering method can be developed in two ways. First, it can be extended to effective robot arm movements, which is of principal importance in factory automation. In this way of development, it is necessary to incorporate various forms of constraints to the path optimization, and a study in this direction has been fruitful [5].

The other way of development is to try to give some chemical meaning to the clusters generated during the path formation. Then it will be a means of molecular modeling. Figure 5 illustrates the schematic expression of an imaginary proton transport

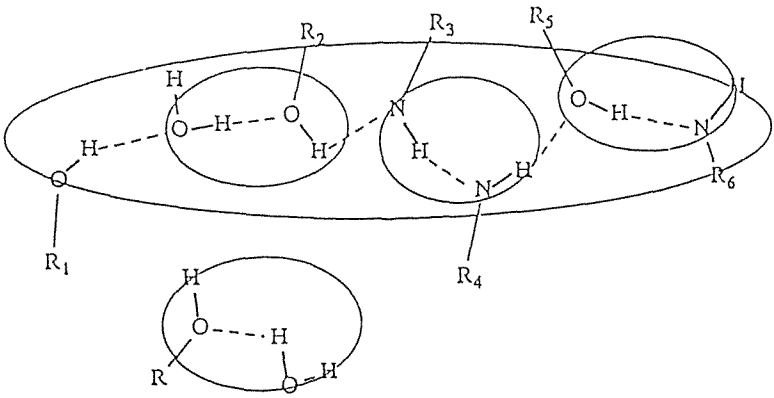


Fig. 5. Searching for a proton transport path in proteins.

path in a protein molecule which could be found by the stepwise clustering method. That is, if we employ components with hydrogen bonding capability like –OH and $-NH_2$ as line segments to be drawn, and if we stop the execution of the clustering at some intermediary step, then it will be possible to find a proton transport path as a cluster which traverses a molecule passing through hydrogen bonds. Since it is very difficult for a human being to find such a path in a complex structure, the clustering

method may give us a new viewpoint to inspect the molecular structure. We look foward to studies in this direction.

Acknowledgement

The authors would like to thank Professor Takashi Okada for his helpful suggestions.

References

[1] M. Iri, K. Murota and S. Matsui, *10th IFIP Conf. on System Modeling and Optimization* (1981), p. 572.
[2] D. Avis, Technical Report No. SOCS-82.4, McGill University (1982).
[3] M. Oyama, Trans. Inf. Proc. Soc. Japan 28(1987)1135 (in Japanese).
[4] M. Oyama, Trans. Inf. Proc. Soc. Japan 29(1988)1091 (in Japanese).
[5] M. Oyama and K. Abe, Trans. Inf. Proc. Soc. Japan 31(1990)1221 (in Japanese).