知能と複雑系　132－18
(2003. 3. 14)

ＡＧＭによる３次元構造と生理活性の相関解析

西村 芳男† 鷲尾 隆† 吉田 哲也† 元田 浩† 猪口 明博‡ 岡田 孝※

†大阪大学産業科学研究所 〒567-0047 大阪府茨木市美穂ヶ丘 8-1
‡日本アイ・ビー・エム(株)東京基礎研究所 〒242-8502 神奈川県大和市下鶴間 1623 番 14
※関西学院大学情報メディア教育センター 〒662-8501 兵庫県西宮市上ヶ原 1-1-155
E-mail: †{nishimura, washio, yoshida, motoda }@ar.sanken.osaka-u.ac.jp ,
‡ inokuchi@jp.ibm.com , ※okada@kwansei.ac.jp

あらまし ＡＧＭアルゴリズムは，グラフ構造データベース中のグラフデータに共通して多く現れる部分構造パターンを完全探索で抽出するアルゴリズムである．ＡＧＭアルゴリズムは頂点と辺にラベルを持つ一般のグラフデータを扱うことができ，グラフのトポロジーを解析できる．本稿では，頂点が３次元座標で表されたグラフデータに対し，各頂点間の距離を計算して辺ラベルに情報を追加する前処理を行い，ＡＧＭアルゴリズムで３次元構造の解析を行う手法を提案する．また，実データによって３次元構造解析を行い，生理活性の相関を調査する．

キーワード グラフデータ，アプリオリアルゴリズム，立体構造，生理活性，化学

Correlation Analysis of 3-dimensional Chemical Structure and its Activity by AGM

Yoshio NISHIMURA† Takashi WASHIO† Tetsuya YOSHIDA†
Hiroshi MOTODA† Akihiro INOKUCHI‡ and Takashi OKADA※

† Institute of Scientific and Industrial Research, Osaka University, 8-1, Mihogaoka, Ibaraki, Osaka, 567-0047, Japan
‡ Tokyo Research Laboratory, IBM Japan, 1623-14, Shimotsuruma, Yamato, Kanagawa, 242-8502, Japan
※Kwansei Gakuin University, Center for Information & Media Studies,
1-1-155 Uegahara, Nishinomiya, Hyogo, 662-8501 Japan
E-mail: †{nishimura, washio, yoshida, motoda }@ar.sanken.osaka-u.ac.jp ,
‡ inokuchi@jp.ibm.com , ※okada@kwansei.ac.jp

Abstract Apriori-based Graph Mining (AGM) algorithm efficiently extracts all the subgraph patterns which frequently appear in graph structured data. The algorithms can deal with general graph structured data with multiple labels of vartices and edges, and is capable of analyzing the connective structure of graphs. We have proposed a faster algorithm of AGM by adding an extra constraint to reduce the number of generated candidates for seeking frequent subgraphs. In this paper, we propose a new method to analyze graph structured data which are represented with a 3-dimensional coordinate by AGM. In this method the distance between each vertex of a graph is calculated and added to the edge label in pre-processing so that AGM can handle 3-dimensional graph structured data. Chemical compounds with dopamine antagonist in MDDR database were analyzed by AGM to analyze their 3-dimensional chemical structure and correlation with physiological activity.

Keyword Graph Structured Data, Apriori Algorithm, 3-Dimensional Structure, Physiological Activity, Chemistry

1. はじめに

AGM (Apriori-based Graph Mining)アルゴリズム[1]は，グラフ構造データベース中のグラフデータに共通して多く現れる部分構造パターンを完全探索ですべて抽出するように，Apriori アルゴリズム[2]を拡張したものである．この手法は頂点と辺にラベルを持つ一般のグラフデータを扱うことができ，グラフの結合構造を解析できる．我々はこの手法の部分構造パターン抽出を高速化するアルゴリズムを提案している[3] [4] [5].

本稿では，頂点が 3 次元座標で表された立体構造を持つグラフデータに対し，各頂点間の距離を計算して辺ラベルに情報を追加する前処理をし，AGM アルゴリ

ズムで立体構造の解析をする手法を提案する．また，MDDR データベースのドーパミン活性を持つ化合物の立体構造を解析し，生理活性との相関を調査する．

2. AGM アルゴリズム

2.1. 概略

AGM アルゴリズム[2]は，グラフ構造データベース GD が与えられたとき，ユーザが指定した最小支持度 (minsup) と呼ばれる閾値を使用して，GD の中に最小支持度を上回る頻度で誘導部分グラフとして含まれるグラフ構造のみを効率よく抽出するアルゴリズムである．グラフ G_s の支持度 $sup(G_s)$は，GD の全データのうち G_s を誘導部分グラフとして含むグラフの割合で定義される．グラフ G_s の支持度 $sup(G_s)$が最小支持度を上回る場合，グラフ G_s を多頻度グラフと呼ぶ．AGM アルゴリズムは，頂点数が 1 の多頻度グラフをから逐次的に頂点数が多い多頻度グラフをレベルワイズに抽出する．図 1 に AGM アルゴリズムの概略を示す．始めに，頂点数が 1 の多頻度グラフをデータベースより抽出し，それを F_1 に代入する．次に，関数 apriori-gen-join では，頂点数が k の多頻度グラフから頂点数 k+1 の多頻度グラフの候補を生成し，それを $\hat{C}_{k+1}$ に代入する．次に，関数 apriori-gen-prune では $\hat{C}_{k+1}$ に格納されている多頻度グラフの各候補について，多頻度グラフであるための必要条件を調べる．この条件を調べることで多頻度グラフの候補の数を絞りこむ．絞りこみで残った多頻度グラフの候補のみを C_{k+1} に格納する．次に，関数 count ではグラフ構造データベース GD にアクセスして，C_{k+1} の各要素の支持度を求める．C_{k+1} の各要素の支持度が最小支持度を上回る場合は，そのグラフを多頻度グラフとし，それを F_{k+1} に格納する．以上の操作を F_k が空集合になるまで繰り返し，グラフ構造データベース GD に含まれる多頻度グラフをすべて抽出する．

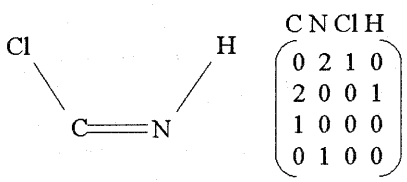

図 2 ラベルつきグラフと隣接行列

```
// GD :グラフ構造データベース
// F_k :頂点数 k の多頻度グラフの集合
// Ĉ_{k+1} :頂点数 k の多頻度グラフを合成したものの集合
// C_k :頂点数 k の多頻度グラフの候補の集合
// minsup :最小支持度(閾値)
1) F_1 = {Frequent subgraph of size=1};
2) for (k =1; F_k ≠ φ ; k++) do begin
3)   Ĉ_{k+1} = apriori-gen-join (F_k);
4)   C_{k+1} = apriori-gen-prune (Ĉ_{k+1});
5)   count (GD;C_{k+1});
6)   F_{k+1} = {c_{k+1} ∈ C_{k+1} | sup(G(c_{k+1})) ≧ minsup};
7) end
8) Answer= ∪_k F_k ;
```

図1 AGMアルゴリズム

2.2. 定義

AGM アルゴリズムで扱うグラフは頂点，辺にラベルを持ち，以下のように定義される．

- ラベル付きグラフ

頂点の集合 V(G)，辺の集合 E(G)，頂点のラベル集合 $L_V(V(G))$，辺のラベル集合 $L_E(E(G))$が

$$V(G) = \{v_1, v_2, \ldots, v_k\},$$
$$E(G) = \{e_h = (v_i, v_j) \mid v_i, v_j \in V(G), i \neq j\},$$
$$L_V(V(G)) = \{lb(v_i) \mid v_i \in V(G)\},$$
$$L_E(E(G)) = \{lb(e_h) \mid e_h \in E(G)\}$$

と与えられたとき，グラフ G は

$$G = (V(G), E(G), L_V(V(G)), L_E(E(G)))$$

と表現される．ここで，頂点の数 $|V(G)| = k$ をグラフ G の大きさとする．$lb(v_i)$および $lb(e_h)$はそれぞれ頂点 v_i のラベル，辺 e_h のラベルである．例えば，図 2 の左側に示すグラフの場合，C, N, Cl, H などの原子の種類が頂点のラベル，単結合,二重結合などの結合の種類が辺のラベルとなる．

- 隣接行列

大きさ k のグラフ $G=(V(G),E(G),L_V(V(G)),L_E(E(G)))$ が与えられたとき，隣接行列 X_k の(i,j)要素 $x_{i,j}$ は，

$$x_{i,j} = \begin{cases} num(lb(e_h)) & \text{if } e_h = (v_i, v_j) \in E(G) \\ 0 & \text{if } e_h = (v_i, v_j) \notin E(G) \end{cases}$$

で与えられる．ここで，$i, j \in \{1, \ldots, k\}$であり，

$num(lb(v_i))$および $num(lb(e_h))$は頂点ラベル $lb(v_i)$，辺ラベル $lb(e_h)$に割り当てられた整数である．図 2 の右側に示す隣接行列は図 2 の左側のグラフに対応し，単結合に 1，二重結合に 2，結合なしに 0 の整が割り当て

られている．

3. 立体構造の解析手法(AGM-3D)

AGM アルゴリズムは頂点と辺にラベルを持つグラフの結合構造を解析することはできるが，立体構造を解析することはできない．

本稿で提案する立体構造の解析手法(AGM-3D)では，グラフの各頂点が図 3 に示すような 3 次元座標で表される立体構造が与えられたときに，それを解析するための手法を提案する．提案手法は AGM で解析を行う前に，距離行列の計算と離散化の 2 つの前処理を行い，AGM で解析可能なグラフデータに変換を行う．

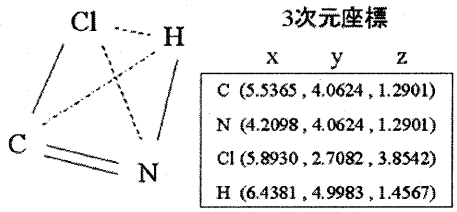

図3　3次元座標で表現される立体構造

3.1. 距離行列の計算

図 3 に示すような 3 次元座標で表される立体構造が与えられたとき，各頂点間の距離を計算し，図 4 の左側に示す距離行列[6]で表現する．距離を用いて 3 次元構造を解析する手法を用いる利点は，距離以外の 3 次元構造データを使用しないため，3 次元モデルの回転や移動など座標系の影響を受けないこと，同一の立体構造を判定する際は距離行列の各要素を比較すればいいため，判定が容易にできることである．

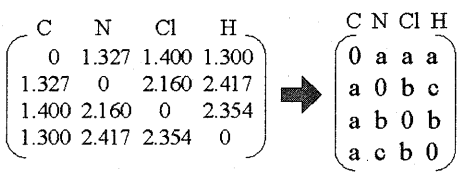

図 4　距離行列と離散化

3.2. 離散化

AGM アルゴリズムで扱えるグラフのラベルは，C, H などの頂点ラベルや単結合,二重結合などの辺ラベルのように離散値を持つラベルであるため，距離値のように連続値からなるラベルは扱えない．そこで，本稿では連続値からなる距離値をある閾値で区切り，離散値に変換する．例えば，図 4 左側の距離行列の各要素 dist を以下の閾値によって離散化した場合，図 4 右側となる．

If　(1.2 ≤ dist < 1.8) then 離散値= a
If　(1.8 ≤ dist < 2.4) then 離散値= b
If　(2.4 ≤ dist < 3.0) then 離散値= c

ここで離散化した距離値は図 5 に示すように，隣接行列の辺ラベルの情報に追加し，AGM で解析可能なグラフデータに変換する．

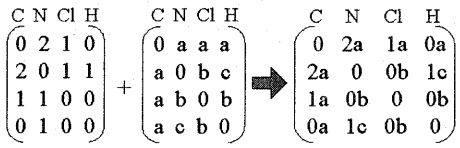

図5　隣接行列に距離の情報を追加

4. 応用

本稿で提案した立体構造を解析する手法の応用例として，MDDR データベース(Trademark of MDL Information Systems Inc.[7])からドーパミン活性を持つ化合物群を抽出し，その化合物の結合構造と 3 次元座標，原子団寄与法で計算された LogP 値を用いて解析を行った．

ドーパミン活性 Dopamine (Di) Antagonist は Di が D1, D2, D3,D4 の 4 種類があり，その活性をもつ化合物はそれぞれ 173 個,430 個,254 個,574 個であった．全データの 10%にあたる 144 個の化合物群をアクティブマイニングのための新規化合物群に設定し，残りの 90%のデータを学習用データとして表 1 のようにデータを分割した．

表1 新規化合物群と学習用データの化合物数

	D1	D2	D3	D4	合計
新規化合物群	20	40	24	60	144
学習用データ	153	390	230	514	1287
合計	173	430	254	574	1431

4.1. 実験 1

学習用データに設定した 1287 個の化合物を対象として，その化合物の結合構造と 3 次元座標データを用いて部分構造パターンの分析を行った．

まず最初に，図 6 の左側で表されるような分子構造データに仮想リンクを追加した．仮想リンクは実際には結合が存在しない辺に，仮想的な辺のラベルをつけたものである．例えば，図 6 の場合は C と H の原子間に結合はないが，2 つの結合を通じてつながっているため，Path2 という仮想的な辺ラベルを追加する．同

様に結合のない他の原子間にも仮想リンクを追加した．

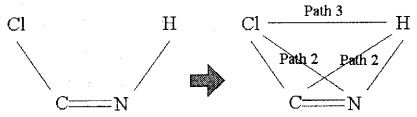

図6　仮想リンクの追加

次に化合物の3次元座標データからAGM-3Dで距離行列の計算をし，離散化した距離の情報を隣接行列の辺ラベルに追加した．離散化では，横軸に距離，縦軸にデータ数を取ったヒストグラムで最大値となる点（モード）から，±0.4Å の点に最初の閾値を設定し，残りの閾値は一定間隔 0.8Å ごとに設定した．こうするのは，ヒストグラムの分布に偏りがあるためであり，辺ラベルごとに異なった閾値を使用した．こうして得られたグラフは，原子の種類を表す頂点のラベル数が12，結合の種類を表す辺のラベル数が108であった．

AGM-3Dで前処理を行ったデータに対し，最小支持度である30%以上の個数の化合物に含まれる部分構造パターンを抽出した．部分構造パターンの抽出に使用した計算機は，Pentium4 2.0GHz と 1.5GB のメモリ、Windows2000 を搭載し，計算時間は3986秒かかった．本実験では多数の部分構造パターンが抽出されたが，以下で説明するχ^2検定[8]によって，抽出したパターンを統計的に評価する．

表2は抽出されたある部分構造パターンに対して，学習用データを活性の種類と部分構造パターンを含むか含まないかによって分割し，分割された各ブロックのデータ数を数えるための表である．Nは学習用データの総数，C_1は部分構造パターンを含む学習用データ数，C_2は部分構造パターンを含まない学習用データ数，O_j(j=1,2,3,4)は活性が Dj を持つ学習用データ数，$C_{1,j}$ (j=1,2,3,4)は活性が Dj を持つデータの中で、部分構造パターンを含む化合物数，$C_{2,j}$ は活性が Dj を持つデータの中で，部分構造パターンを含まない化合物数である．

表2　χ^2-検定の分割表

	D1	D2	D3	D4	合計
部分構造を含む	C_{11}	C_{12}	C_{13}	C_{14}	C_1
部分構造を含まない	C_{21}	C_{22}	C_{23}	C_{24}	C_2
合計	O_1	O_2	O_3	O_4	N

このとき，その部分構造パターンを含む学習用データの活性分布とすべての学習用データの活性分布をχ^2検定で統計的に適合するか調べる式が以下の式(1)である．

$$\chi^2 = \sum_{i,j} \frac{\left(C_{i,j} - E_{i,j}\right)^2}{E_{i,j}}, \quad (1)$$

$$E_{i,j} = N \times \frac{C_i}{N} \times \frac{O_j}{N}$$

図 7,8,9 に実験で抽出された部分構造パターンのなかでχ^2値が大きなパターンを例に示す．各図は左側に部分構造パターンの3次元モデルを，右側に結合構造を示す．図7の部分構造パターン(1)はχ^2値が82.0，図8のパターン(2)はχ^2値が49.7である．2つのパターンはベンゼン環に接続している原子が1つ異なるだけであるが，パターン(1)の分子は活性がD4の化合物に含まれる数が少なく，パターン(2)の分子はD3，D4の化合物に多く含まれる．また，図9のパターン(3)はχ^2値が31.3である．D1の化合物に含まれる数が少なく，D2の化合物に多く含まれる．

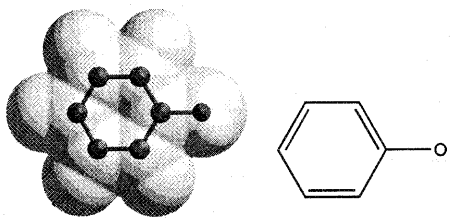

図7　抽出された部分構造パターン(1)

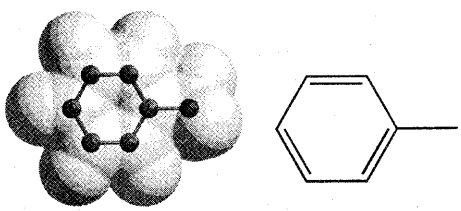

図8　抽出された部分構造パターン(2)

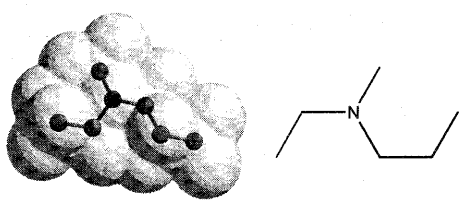

図9　抽出された部分構造パターン(3)

4.2. 実験 2

分類規則学習法では標準的な手法である C4.5[9]を使用して，学習用データからドーパミン活性の分類規則を作成し，新規化合物群を想定したテスト用データ

でこの分類誤差の測定を行った.

各化合物のデータを各種属性項目を要素とするベクトルで表現する. 各要素の値は数値のみではなく記号であってもよい. C4.5 は各化合物のベクトル情報から目的とするクラス属性（本実験ではドーパミン活性レベル）を分類する手法である. 化合物のデータについて, 実験 1 の AGM-3D で抽出された 111 個の各パターンを部分構造に持つかどうか判定し, それぞれに「ある」、「なし」の記号を割り当て, 各化合物をベクトルで表す. このベクトル形式のデータに, LogP 値とドーパミン活性レベルの情報を付け加え, C4.5 の入力とした. この場合, ドーパミン活性レベルの分類誤差は 23.6 %となった. C4.5 によるデータの分類分布を表 6 に示す. 横行は実際の活性レベル, 縦列は C.4.5 による分類結果を表し, 対角部分の個数が多いほうが精度が高いことを表す.

表6 C4.5によるデータ分類の分布(AGM-3D)

C4.5 による分類→	D1	D2	D3	D4
D1	11	7		2
D2	1	29	5	5
D3		2	18	4
D4		7	1	52

表7 C4.5によるデータ分類の分布(AGM)

C4.5 による分類→	D1	D2	D3	D4
D1	12	8		
D2	1	33	2	4
D3		1	21	2
D4	1	5	3	51

また, 立体構造を示す 3 次元座標データを使わずに AGM アルゴリズムで最小支持度である 30%以上の個数の化合物に含まれる部分構造パターン 275 個を抽出し, 同様の実験を行った. この場合の分類誤差は 18.8 %となった. C4.5 によるデータの分類分布を表 7 に示す.

AGM と AGM-3D を用いて部分構造の属性項目を生成し C4.5 で分類規則を得る手法とも, ほとんど部分構造パターンの属性によって生理活性の分類を行う規則を生成した. また, 両手法を比較した場合, AGM-3D の結果が良くないが, これはもともとは連続値である距離を離散化して取り扱っているためと思われる. AGM-3D では距離が近く実際の立体構造も似ている化合物が, 離散化によって異なる物質と判断される可能性がある. そのため, これを回避するための手法を実装することが今後の課題である.

5. 結論

AGM アルゴリズムを利用して立体構造の解析を行う AGM-3D を提案し, ドーパミンデータで特徴的な部分構造パターンを抽出し, 部分構造パターンと生理活性の相関を解析し, 相関の高いことを確認した.

文　献

[1] A. Inokuchi, T. Washio, and H. Motoda ,"An Apriori-Based Algorithm for Mining Frequent Substructures from Graph Data," Proc. of the 4th European Conference on Principles and Practice of Knowledge Discovery in Databases, pp.13-23 ,2000.

[2] R. Agrawal and R. Srikant ,"Fast Algorithms for Mining Association Rules in Large Databases," Proc. of the 20th VLDB, pp.487-499 ,1994.

[3] 猪口明博, 鷲尾隆, 元田浩, "Apriori-Based Graph Mining アルゴリズムの効率化," 第 15 回人工知能学会全国大会 ,2001.

[4] 猪口明博, 鷲尾隆, 西村芳男, 元田浩 ,"グラフ構造データからの連結多頻度グラフ抽出手法," 第 16 回人工知能学会全国大会 ,2002.

[5] 西村芳男, 鷲尾隆, 吉田哲也, 元田浩, 猪口明博 , "Apriori-based Graph Mining アルゴリズムの高速化," 第 128 回情報処理学会知能と複雑系研究会 ,2002.

[6] H. Kato and Y. Takahashi ,"SS3D-P2:a three-dimensional substructure search program for protein motifs based on secondary structure elements," 13, pp.593-600 ,1997.

[7] http://www.mdli.com/.

[8] S. Brin, R. Motwani and C. Silverstein, "Beyond market baskets: Generalizing association rules to correlations," Proc. of the 1997 SIGMOD, pp.265-276 ,1997.

[9] J. R. Quinlan, "C4.5:Programs For Machine Learning," Morgan Kaufmann Publishers ,1993.