分子グラフの簡易マイニングにおけるパスフラグメント表現の拡張

Extended Path Fragment Representation in Simple Mining of Molecular Graph

加納 正章 Masaaki kanou
藤島 悟志 Satoshi Fujishima
高橋 由雅 Yoshimasa Takahashi

豊橋技術科学大学 工学部 知識情報工学系
Department of Knowledge-based Information Engineering, Toyohashi University of Technology

In the preceding work, we investigated quick approach to structure data mining of chemicals using path fragments that both of the terminal atoms are hetero atoms. In this work, we defined an extended path fragment. The new path fragment is characterized by multiple bonds and hetero-atoms at intermediate part of the path but not only the terminal hetero-atoms. The extended path fragment makes us possible to identify different fragments that are regarded as the same path fragment in the former definition. The paper describes the extended definition of a path fragment and a quick mining algorithm for finding them. To test the usage of our approach, the dopamine agonists of 370 chemicals were used for a computational trial. The result suggests the present method would be useful for quick mining to get some more information on the path fragments.

1. はじめに

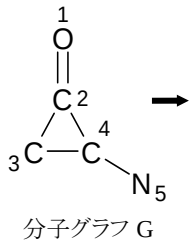
構造活性相関の研究において，活性予測のためのモデル構築手法の研究が多く行われている．しかし，そのデータ表現に使用する構造記述子について議論されることは少ない．そこで，本研究室ではヘテロ原子に注目したパスフラグメントによる分子グラフの簡易マイニング手法を提案し，その有用性を示してきた[栗林 04, 藤島 07].

本研究では，両端の頂点ラベルがヘテロ原子のみで構成される従来のパスフラグメントの拡張として，パスの途中に存在するヘテロ原子やヘテロ原子に接合する多重結合を考慮することで，先に提案したパスフラグメントに比べ，より詳細な構造情報を有する拡張パスフラグメントの提案とその生成法について検討を行った．

2. パスフラグメント

与えられた分子グラフ中の，任意の 2 つの原子と，それらの間の最短パスの長さ(エッジ数)で表わされる部分グラフをパスフラグメントと定義する．また，その表記は極めて簡便であり，かつ解釈も大変容易であるなどの利点を有する．例えば，部分グラフ N-x-x-O(x は任意の原子)は前述の定義に基づくパスフラグメントであることは明らかであり，その表記は N3O として表すことができる．図 1は分子グラフ G から取り出せるパスフラグメントを最短パスの長さごとに列挙したものである．

グラフ G の全パスフラグメント



分子グラフ G

エッジ数：1
C1O (C-O), C1C (C-C)×3,
C1N (C-N)

エッジ数：2
C2O (C-x-O)×2, C2N (C-x-N)×2,
C2C(C-x-C)×3

エッジ数：3
N3O (N-x-x-O)

図 1 分子グラフ G のパスフラグメント

連絡先：高橋由雅，〒441-8580 愛知県豊橋市天伯町雲雀ヶ丘 1-1 豊橋技術科学大学 知識情報工学系，Tel: 0532-44-6878, taka@mis.tutkie.tut.ac.jp

3. 拡張パスフラグメント

(1) ヘテロ原子に接合する多重結合を考慮する

図 2 の(a)と(b)のグラフ(x は任意の原子)は，従来のパスフラグメント定義ではいずれも N3O で表され，同一のパスフラグメントみなされる．しかしながら，=O と-OH の酸素原子の化学的な性質は明らかに異なる．従って，たとえ簡易マイニングを主眼とする方法であっても，これらの化学種を考慮することは極めて重要である．

このことから，ここでは，従来のパスフラグメントの拡張定義として，「両端ヘテロ原子に接合する結合多重度を考慮」することとした．この場合の拡張パスフラグメントの表記は「原子+<結合多重度>+距離+<結合多重度>+原子」とする．ただし，単結合の場合は結合多重度の情報は明記しない．これにより，図 2 の(a),(b)はそれぞれ N3<2>O および N3O と表記され，両者の区別が可能となる．

N－x－x＝O　　N－x－x－O

(a)　　(b)

図 2 結合多重度を考慮した場合の比較

(2) パスフラグメント内のすべてのヘテロ原子を考慮する

次に図 3の例について考える．この場合も(a), (b)の二つのグラフは従来のパスフラグメント表記では同一のフラグメント(N4O)として表記される．

N－C－C＝N－O　　N－C－C－N－O

(a)　　(b)

図 3 パス中のヘテロ原子を考慮した場合の比較

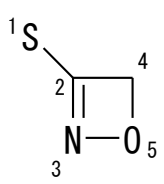
しかし，上記(1)，(2)の定義拡張により，両端のヘテロ原子以外にもパス中のすべてのヘテロ原子とこれらに接合する多重結合度の相違を明確に区別することができる．この場合の拡張パスフラグメントの表記は「原子[+<結合多重度>+距離+<結合多重度>+原子]」とする．ただし，[ ]内は 1 回以上の繰り返しである．これにより，図 3 の(a), (b)はそれぞれ N3<2>N1O および N3N1O と表記され，両者の区別が可能となる．

4. 拡張パスフラグメントの生成

本手法では新たにパス中のヘテロ原子や結合多重度の情報が必要になるため，Schultz の提案する距離行列の生成アルゴリズム[Schultz 00]を修正し，パス中の情報を拡張結合行列として得るアルゴリズムを提案する．

4.1 拡張結合行列

拡張結合行列において対角要素は「元素記号」，それ以外の要素は「結合多重度[+元素記号+結合多重度]([ ]内は任意の回数繰り返し)」で表現されるパスの情報である．図 4(a)の分子グラフの結合行列を図 4(b)に示す．生成される拡張距離行列を図 4(c)に示す．拡張距離行列で「/」を含む要素は複数のパスが存在することを表す．



$$\begin{pmatrix} S & 1 & 0 & 0 & 0 \\ 1 & C & 2 & 1 & 0 \\ 0 & 2 & N & 0 & 1 \\ 0 & 1 & 0 & C & 1 \\ 0 & 0 & 1 & 1 & O \end{pmatrix}$$

(a) 分子グラフ　　(b) 結合行列

$$\begin{pmatrix} S & 1 & 1C2 & 1C1 & 1C2N1/1C1C1 \\ 1 & C & 2 & 1 & 2N1/1C1 \\ 2C1 & 2 & N & 2C1/1O1 & 1 \\ 1C1 & 1 & 1C2/1O1 & C & 1 \\ 1N2C1/1C1C1 & 1N2/1C1 & 1 & 1 & O \end{pmatrix}$$

(c) 拡張距離行列

図 4 拡張距離行列の生成

4.2 拡張距離行列の生成アルゴリズム

アルゴリズムの流れを以下に示す．ここで，Step1～Step6 は Schultz の距離行列生成手順を示したものであり，Step1.1 と Step5.1 は拡張結合行列を生成するために追加した手順である．

Step1　対角要素を元素記号，隣接している要素を「1」，その他の要素「0」と表記した隣接行列を作成

Step1.1　対角要素を元素記号，隣接している要素を結合多重度，その他の要素を「0」と表記した結合行列を作成

Step2　行の左側から要素を見ていき，距離が未知の要素(要素が 0 のもの)を探す
※この行を r1 とする

Step3　距離が未知の要素から対角要素まで下に行き，横方向へ要素が「1」の位置を探す
※未知の要素があった列を c1 とする

Step4　「1」が見つかった場合，r1 行に注目する
※「1」が見つかった位置の行を r2，列を c2 とする
※「1」が複数見つかる場合もある

Step5　r1 行 c2 列の要素と r2 行 c2 列の要素の和を求め距離が未知の要素(r1 行 c1 列)へ代入し，対称な位置(c1 列 r1 行)にも代入する
※Step4 で「1」が複数見つかった場合は和が最小の値を代入する

Step5.1　r1 行 c1 列に「r1 行 c2 列の要素」「c2 行 c2 列の要素」「r2 行 c2 列の要素」の順に並べたものを代入し，対称な位置には逆の並びにしたものを代入する
※和が最小となる位置が複数ある場合はすべてのパターンを記録する

Step6　全ての行から未知の要素がなくなるまで Step2~5 を繰り返す

5. 実データを用いた計算機実験

薬物構造データベース MDDR[MDL 01]]に登録されているドーパミンアゴニスト 370 件のデータを使用して，計算機実験を行った．実験では，従来の定義に基づくパスフラグメントと本手法について，抽出されたパスフラグメントの総数および，データセット中 10%以上の出現頻度を持つフラグメントについて検討を行った．両方の手法による実験結果を表 1に示す．

表 1 従来法と本手法によるパスフラグメント抽出結果

	抽出されたユニークな全パスフラグメント	出現頻度 ＞0.1
従来手法	190	31
本手法	775	19

従来法では総数 190 種類のユニークなパスフラグメントが抽出されたのに対し，本手法では約 4 倍の 775 種類のユニークなパスフラグメントが抽出された．このことは，本提案手法によるパスフラグメントの方がより詳細な構造情報を記述できることから，従来法に比べ，多様なフラグメントが抽出されたことを示している．一方，出現頻度が 0.1 以上のフラグメントのみを抽出した場合には，それぞれ 31 種類，19 種類が抽出された．本手法により得られた多様なパスフラグメントの一部を図 5 に示す．

N2<2>N4N	N4N	N2<2>O
N2S	O<2>1<2>S	N<8>6<2>O
N<2>2S	O1S	N6<8>N
N5S	O1S<2>1<2>O	N3N4O4N
N2<2>N	O<2>1<2>S<2>1<2>O	N3N4O

図 5 本手法により得られた多様なパスフラグメント(一部)

6. おわりに

ここでは，パスの途中に存在するヘテロ原子やヘテロ原子に接合する多重結合を考慮することで，先に提案したパスフラグメントに比べ，より詳細な構造情報を有する拡張パスフラグメントの抽出について述べた．今後の課題としては，特定の薬物群に対する相関ルールのマイニングを具体的に進めていくと同時に，引き続き，より柔軟なパスフラグメント表現を工夫していきたい．

参考文献

[栗林 04] 距離行列に基づくパスフラグメントの生成とデータマイニング: 平成 16 年度豊橋技術科学大学卒業論文, 2004.

[藤島 07] 藤島悟志, 高橋由雅, 岡田孝: ヘテロ原子に注目したパスフラグメントによる化学構造データマイニング， 2007 年度人工知能学会全国大会(第 21 回)論文集， 3F7-4, 2007.

[Schultz 00] Harry P. Schultz: Topological Organic Chemistry. 13. Transformation of Graph Adjacency Matrixes to Distance Matrixes, J. Chem. Inf. Comput. Sci. 40 pp.1158-1159 , 2000.

[MDL 01] MDL: MDL Drug Data Report, 2001.1, 2001.

[Nijssen 04] S. Nijssen, J. N. Kok: Frequent Graph Mining and its Application to Molecular Databases, Proceedings of the 2004 IEEE Conference on Systems, Man & Cybernetics (SMC2004), 2004.