様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２４年　６月１４日現在

機関番号：５３３０２
研究種目：若手研究（B）
研究期間：2009　～　2011
課題番号：２１７００１５９
研究課題名（和文）　パスフラグメントによる分子グラフのクイックマイニング

研究課題名（英文）　Quick Mining of Molecular Graph Using Path Fragment

研究代表者
藤島　悟志（FUJISHIMA SATOSHI）
金沢工業高等専門学校・電気電子工学科・准教授
研究者番号：10411787

研究成果の概要（和文）：構造特徴表現として考案したパスフラグメントを使用した，頻出パスフラグメントの統計的解析の手順を検討し，化学構造に特徴的なパスフラグメントの獲得を試みた．解析結果から，原子ラベルは同じでも，その間の距離によって「ドーパミン活性の発現が異なる」という，先行研究では見出せなかった知見を獲得することができた.また,構造記述子としての薬理活性クラス分類への適用可能性についても検証を行い，良好な結果が得られた.

研究成果の概要（英文）：A quick approach to chemical structure data mining using path fragments ware proposed. The molecular graph of each drug is featured by the path fragments and the frequent path fragments that are specific to individual drug actions were analyzed. We could obtain characteristic fragments for each activity. The applicability of the SVM based on the path fragments for classification of pharmacological activities of drugs was validated.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００９年度	500,000	150,000	650,000
２０１０年度	500,000	150,000	650,000
２０１１年度	400,000	120,000	520,000
年度			
年度			
総　計	1400,000	420,000	1820,000

研究分野：分子情報工学
科研費の分科・細目：情報学・知能情報学
キーワード：パスフラグメント，分子グラフ，薬物構造データマイニング，薬物活性クラス

１．研究開始当初の背景

(1)　医薬や農薬などの開発研究の場では既存薬物のデータを背景に，薬化合物の化学構造と生理活性などの種々の作用（特性）との関係を見出し，これらを活用して新規有用化学物質の合理的な開発を進める様々な試みが行われている．これらの基礎となる，知識の獲得や種々のデータ予測の実現には化学物質の構造特徴の理解が不可欠となる．

(2) 一方で，情報化学の分野において，集積された大量のデータを系統的，機械的に解析・処理しながら，知識獲得やチャンス発見を目的としたデータマイニングの技法が注目を集めている．しかし，マイニングによって得られた情報は，豊富ではあるが，その豊富さ故に解釈が困難になるなど，そのままでは知識として不十分であり，有用な知識獲得のためには，その情報の精錬作業が必要となる．未だ化学者に既知の骨格を提示するに留まっており，新規の発見には至っていない．

(3) こうした化学構造情報を基礎としたデータマイニングにおいては，その要素技術である，構造特徴解析の新たな技法の開発やシステム化が極めて重要な課題となっている．

２．研究の目的

(1) 化学構造データマイニングの一連の作業の中でも重要な位置を占める，化学構造の構造特徴表現に注目し，直感的に理解容易な構造記述表現として「パスフラグメント」を考案する．化学構造の各構成原子とそれらの間の最短パス長を利用した直観的な表現を用いる．

(2) パスフラグメントを利用した新たなマイニング法も併せて提案する．様々な活性クラス群の構造特徴差異の抽出を行うことで，新薬開発のスピードアップに繋がる化学構造データマイニング手法の確立を目指すとともに，活性クラス分類や薬物分子の毒性予測など構造活性相関モデルへの応用を試みる．

３．研究の方法

(1) パスフラグメントの定義と拡張表現

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．

より特徴的なフラグメントの抽出のために，化学的な見地も取り入れた拡張表現の提案を行う．

①カルボニル（O=C）のような酸素の 2 重結合などは重要であるため，ヘテロ原子の結合多重度を考慮する．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，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，これまで不可能だった両者の区別が可能となる．

②パスの途中に存在するヘテロ原子も考慮する（図 2 の(c)，(d)）．

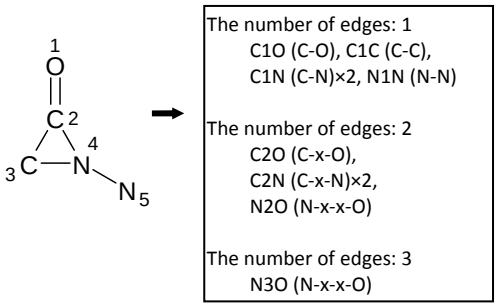


図 1 分子グラフ G のパスフラグメント

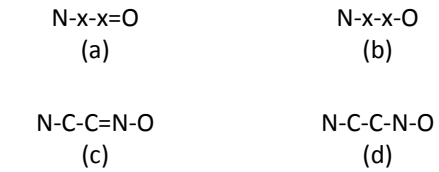


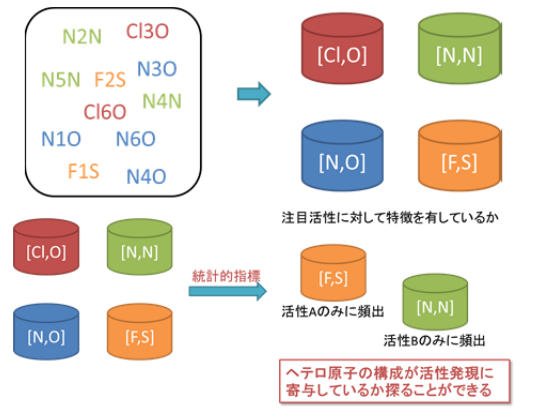
図 2 パスフラグメントの拡張表現例（結合多重度を考慮(a),(b)，パス中のヘテロ原子を考慮(c),(d)）

(2) パスフラグメントによる化学構造データマイニング手法

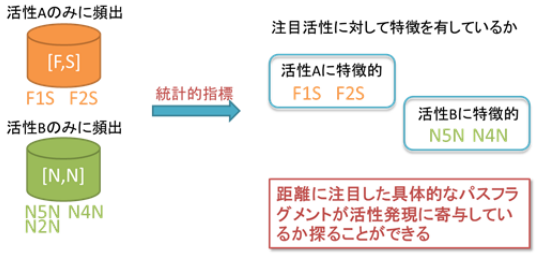
化学構造群から得られたパスフラグメントに対して，頻出パスフラグメントの統計的解析に基づき，化学構造に特徴的なパスフラグメントの獲得について検討を行う．

生成されたパスフラグメントを系統的にまとめることで，構造特徴解析を段階的に行うことを試みる．解析手順の第 1 ステップは生成されたパスフラグメントを系統的にまとめる作業であり，パスフラグメントを両端の原子ラベルのみに注目してグループ化し，その頻度分布をみる．この作業を行うことで，原子ラベルのみで薬理活性クラスの特徴解析が行える（図 3(a)）．

次に第 2 ステップとして，特定のグループに対して，距離情報を横軸とした頻度分布を生成する．この頻度分布からは薬理活性クラスに特徴的な，フラグメントの距離情報などが得られる可能性がある（図 3(b)）．



(a) パスフラグメントグループに分ける



(b) グループ内のフラグメントに注目

図 3 パスフラグメントによる化学構造データマイニングの概念図

このように段階的に解析を行うことで，原子ラベルは同じでも，その間の距離によって「薬理活性の発現が異なる」という，先行研究では見出せなかった知見が得られる．

各段階において特徴的なパスフラグメントを抽出する指標として，統計量 Internal Sensitivity (IS) と Inherency (H) を定義する．内部感度 IS は，注目クラス内の相対頻度を表す（式 1）．ここで，$n_{i,j}$ はクラス i における特徴（パスフラグメント）j を持つサンプルの数を，N_i はクラス i のサンプル数をそれぞれ示す．

注目クラスにおけるある特徴の H（固有度）は，注目クラス内のその特徴の IS と他の全クラスにおけるその特徴の IS の平均の比によって定義される（式 2）．p はクラス数，そして，$H(i,j)$はクラス i における特徴 j の固有度を表す．

$$IS(i,j) = \frac{n_{i,j}}{N_i} \tag{1}$$

$$H(i,j) = \frac{IS(i,j)}{IS(i,j) + \frac{1}{p-1}\sum_{k=1(k\neq i)}^{p} IS(i,j)} \tag{2}$$

H の値は $0 \leq H \leq 1$ の範囲をとり，$H(i,j) \leq 0.5$ は特徴 j がクラス i において特徴的な意味を持たないことを，$H(i,j) > 0.5$ は特徴 j がクラス i に特徴的であることをそれぞれ表す．つまり，ある特徴の $H(H > 0.5)$ がより大きければ，その特徴はそのクラスにより特徴的であることを意味する．本研究では，上記 2 つの統計指標を使ってパスフラグメントの評価を行った．

(3) 構造特徴を用いた活性クラス分類や物性（毒性）予測モデルなどの構造活性相関モデルへの応用

パスフラグメントを構造特徴記述子とした活性クラス分類への応用が考えられる．上記データマイニング手法によって得られた各クラスの構造特徴を入力記述子とした，活性クラス分類実験について検討する．活性クラス分類には，申請者の先行研究で行ってきた，人工ニューラルネットワークやサポートベクタマシンが利用できる．

4．研究成果

(1) 薬物構造データベース MDDR 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．

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パスフラグメント	出現頻度 >0.1
従来法	190	31
本手法	775	19

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 4 本手法により得られた多様なパスフラグメント（一部）

両方の手法による実験結果を表 1 に示す．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 4 に示す．

(2) 頻出パスフラグメントの統計的解析の手順から，化学構造に特徴的なパスフラグメントの獲得を試みた．ドーパミン受容体アゴニスト 3 種（370 化合物，D1：63，D2：143，Autoreceptor：180）を用いて，本手法の妥当性について検討を行った．構造特徴の抽出には，上記の統計指標を使い，閾値の設定を行いながら解析結果の評価を行った．

まず，データセットから両端にヘテロ原子を持つパスフラグメントを生成した．次に，各パスフラグメントを，両端原子ラベルをもとに距離情報を考慮しないパスフラグメントグループへ割り当てた．その結果，4,663 パスフラグメントが生成され，19 パスフラグメントグループに割り当てられた．パスフラグメントグループごとの各クラスにおける IS 値のグラフを図 5 に示す．

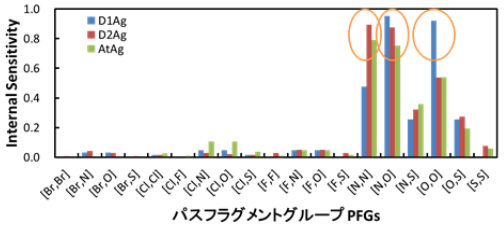


図 5 パスフラグメントグループごとの Internal Sensitivity（D1Ag，D2Ag および AtAg はそれぞれ Dopamine D1 Agonist，D2 Agonist および Autoreceptor Agonist を表す）．

[N,O] フラグメント群は D1 アゴニストで最も頻度が高く，その IS は 0.952 であった．それに続いて，D2 アゴニストは 0.874，Autoreceptors アゴニストは 0.754 と順に減少している．[N,N] については D2 アゴニストの IS が最も高く 0.895 であった．また，[O,O]フラグメント群では D1 アゴニストの IS が 0.921 と最も高かった．このように，パスフラグメントグループでクラス寄与がみられる結果となった．

次に，第 2 段階のマイニングとして，距離情報を考慮した個々のパスフラグメントに注目した．ここでは，パスフラグメントの出現頻度が高い，[N,O]の結果を示す．図 6 に [N,O] のパスフラグメント群における，距離情報ごとの IS 値のグラフを示す．パスフラグメント N6O と N7O において D1 アゴニストが最も頻度が高いことが分かる．N2O と N5O では D2 アゴニストが他のクラスより頻度が高い．一方で，Autoreceptors については，N3O と N4O が特徴的なフラグメントであると考えられる．これらの結果から，各クラスにおいて，両端原子間の適切な距離が存在していることが分かる．

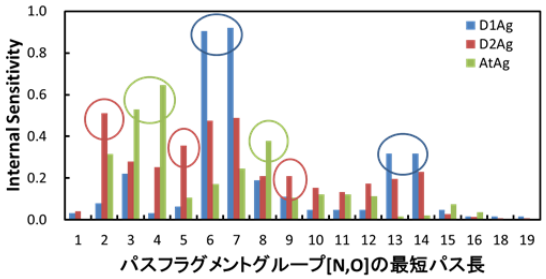


図 6 距離情報を考慮した[N,O]パスフラグメントグループの活性寄与

表 2 ドーパミンアゴニストに対する縮約パスフラグメントにおけるマイニング結果

	[N,N]	IS	H	[N,O]	IS	H	[O,O]	IS	H
D1	-	-	-	N6O	0.905	0.737	O3O	0.905	0.790
				N7O	0.921	0.715			
				N13O	0.317	0.750			
				N14O	0.317	0.716			
D2	N2N	0.371	0.671	N2O	0.510	0.721	-	-	-
	N5N	0.441	0.757	N5O	0.357	0.807			
	N7N	0.483	0.606	N9O	0.210	0.658			
Auto	N2N	0.348	0.643	N3O	0.529	0.678	O4O	0.289	0.771
	N3N	0.246	0.637	N4O	0.647	0.820	O5O	0.417	0.626

パスフラグメントグループ[N,O]，[N,N] および [O,O]に対する詳細解析（第 2 段階解析）の結果を表 2 に示す．表中の特徴的パスフラグメントは，IS≥0.2 かつ H≥0.6 の条件で抽出した．ここでも，各パスフラグメントグループにおいて，クラスごとに特徴的な距離が存在していることが分かる．フラグメント N6O と N7O は D1 アゴニストにおいて 91% と 92% の確率で出現していることが分かり，かつ両フラグメントの Inherency は他のクラスと比べて極めて高い値を持つことが分かった．また，O3O も D1 アゴニストにおいて IS と H が極めて高いことから，これらのフラグメントは D1 アゴニストに特徴的であると考えることができる．

得られたパスフラグメントによって示される“特徴”をいくつかのドーパミン D1 アゴニストの構造上に実際に示す（図 7）．図 7 からパスフラグメント N6O, N7O と O3O が構造的に多様な分子グラフ中で確認できることが分かる．これらは，本アプローチが化学構造データマイニングに適用可能であることを示唆すると考える．

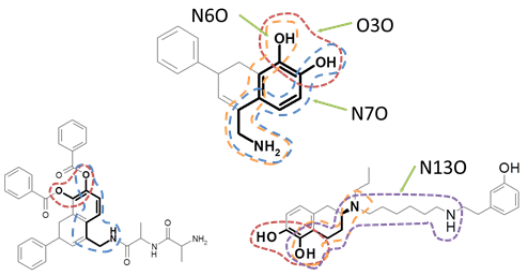


図 7 N6O，N7O および O3O を持つドーパミン D1 アゴニストの構造例.

表 3 SVM による Dopamine 予測結果（10 回平均）

	全体	Agonists	Antagonists
学習率	97.5%	96.0%	98.9%
予測率	96.4%	92.6%	100.0%

以上の結果から，本研究で提案する化学構造データマイニング手法が有効であることが分かった．特に，ヘテロ原子に注目したパスフラグメントを使用することで，直観的に理解可能な特徴記述表現となり，マイニングも高速に行える．また得られる結果（ルール）も複雑にならないため，化学者にとって理解が容易になると言える．しかし，構造特徴の抽出には，頻度分布から得られる統計量の定義や閾値の設定などが必要であり，現段階では未だ決定的なものが得られていない．今後は，自然言語処理分野で使用されている TF/IDF 法の適用も含め引き続き統計量の検討を行っていく．

(3) パスフラグメントを構造特徴記述子とした活性クラス分類への適用実験を試みた．

PubMed から抽出した 4 種（ドーパミン拮抗薬／作動薬，変異原性，発癌性）のデータに対して薬理活性クラス分類実験を行った．ドーパミン 2 種を対象とした 2 クラス分類や 4 種すべてを使用した 4 クラス分類などの実験を行ったところ，いずれも良好な精度でクラス分類を行うことができた．ここではドーパミン 2 種（作動薬（Agonits）104 件，拮抗薬（Antagonists）115 件）のクラス分類結果を示す．

ドーパミンデータの 90%を訓練データ，残り 10%を予測データとして分ける．訓練データを使用して SVM による学習を行い，モデルを構築する．予測データをモデルに入力し，その予測率からモデルの予測能力を検証する．10-fold cross validation 法によって本手法の汎用性を検証した．結果を表 3 に示す．全体として 96.4%と高い予測率が得られた．特に拮抗薬の予測率が 100%と全て正しく予測することができた．

このことから，SVM による活性クラス分類についてパスフラグメントの応用可能性も期待できる．今後は，創薬研究の初期段階における，副作用などを避けたリード構造の提案や，既存薬に対する未知の副作用などに関するリスク推定問題への応用可能性について引き続き検討したい．

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 1 件）

① Takashi Okada, Masumi Yamakawa, Norihito Ohmori, Sachio Mori, Hiroshi Horikawa, Taketo Hayashi, Satoshi Fujishima, The Development of a Knowledge Base for Basic Active Structures: An Example Case of Dopamine Agonists, Chemistry Central Journal, 査読有, Vol.4, 2010, DOI:10.1186/1752-153X-4-1

〔学会発表〕（計 3 件）

① 南成光，藤島悟志，パスフラグメントを用いた分子構造の構造類似性評価，平成 21 年度 北陸地区学生による研究発表会，2010/03/06，（金沢工業大学，石川）
② 徳野聖児，福野達郎，藤島悟志，薬理活性クラス分類におけるパスフラグメント表現の有用性，平成 21 年度 北陸地区学生による研究発表会，2010/03/06，（金沢工業大学，石川）
③ 加納正章，藤島悟志，高橋由雅，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，第 23 回人工知能学会全国大会，2009/06/19，（サンポートホール高松，香川）

６．研究組織

(1)研究代表者

藤島 悟志（FUJISHIMA SATOSHI）
金沢工業高等専門学校・電気電子工学科・准教授
研究者番号：10411787