3A3-01

化合物の線形フラグメントに基づいた特徴的リード構造抽出システムの開発

Development of Characteristics Lead Structure Extraction System based on Liner Fragments of Compounds.

藤島悟志
Satoshi Fujishima

岡田孝
Takashi Okada

関西学院大学 理工学部 情報科学科
Department of Informatics, School of Science and Technology, Kwansei Gakuin University

Rules derived by the cascade model could give useful insights for characteristic features affecting the pharmaceutical activity. But the extraction and the synthesis of characteristic feature were performed by the hand of an expert and it took too long to reach the final characteristic structure. The characteristics lead structure extraction system based on liner fragments of compounds has been developed to solve this problem. The system has shown its usefulness in the refinement from the seeds found in the rule to the expert level knowledge. The details of the system will be discussed with illustrative examples.

1. はじめに

当研究室ではデータマイニング技法の一つとして，相関ルールを拡張したカスケードモデルを提案している[Okada 99]. これまで，線形フラグメント属性を利用した化学構造からのマイニングを行ってきた. 結果として，注目する薬理活性に特徴的なリード構造を抽出し，カスケードモデルを使用したルール群の表現が，知識発見に有効であることを示してきた[Okada 02a, Okada 02b]. しかし，最終的なリード構造の抽出には，出力されたルール表現を利用した専門家の手作業による解析作業が必要であり，その解析に多大な時間と労力を要している.

そこで本研究では，より効率的な構造特徴の抽出および知識発見を目的とし，ルール群に記述されている線形フラグメントを種として，対象化合物群により特徴的な構造に洗練していくシステムの開発を行った. また，実データを用いたシステムの有用性について検討を行ったので報告する.

2. 方法

2.1 線形フラグメント

カスケードモデルによる解析では，属性として線形フラグメントを用いる. これは解析対象の化合物群から生成されたものであり，解析データでは，各化合物中での有無を y/n で表記している.

線形フラグメント生成法は，下記の方法によっている. (1) 指定した種類の元素および結合両端の原子を起点として，最短 path 長が max-length 以内のすべての原子との間で線形フラグメントを取り出す. (2) このフラグメントを構成元素と結合の種類，各原子の配位数と付随水素原子の有無，および分岐構造で枝上に現れる最初の元素により特徴づける. (3) 利用者の指定した詳細度に従い，これらの線形フラグメントを記述する.

これまでの研究[Okada 05]から，線形フラグメントは両端から各 2 個の原子のみに配位数と付随水素原子の有無を記載し，フラグメント長は原子数が 10 以内に限っている.

図 1 左側の構造式で OH から NH2 に至る部分の線形フラグメント表記を，同図の右側に示す. ここで，O2H は 2 配位の酸素原子に少なくとも 1 個の水素が付随しており，一重結合で繋がった c は芳香族の 3 配位炭素原子であること，次の::− −は 2 つの芳香族結合に続き 2 つの一重結合が存在することを表す. この間の元素記号は省略されている. 最後の C4H-N3H は少なくとも 1 個の水素が付随した 4 配位炭素および 3 配位窒素を示す.

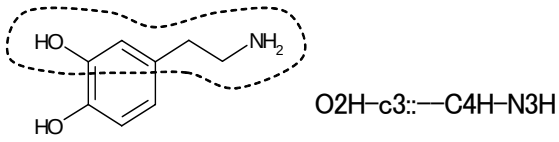

図 1 線形フラグメントの例

2.2 線形フラグメントの洗練

カスケードモデルから得られるルールは豊富であるが，知識としては不十分である. そこで，豊富にあるルール（情報）を洗練することにより，注目薬理活性に特徴的な部分構造が知識として取り出せる. 先行研究では，専門家の手作業によって洗練作業を行っていたが，その解析に多大な時間と労力を要していた. 本研究では，より効率的な構造特徴の抽出および知識発見を目的とし，洗練作業の自動化を試みた.

線形フラグメントのシステムへの入力方法として，SMILSE/SMARTS 表記[Weininger 88]を採用した. SMARTS 表記を用いることで，原子をパターン記述で表すことが可能となる. 図 1 右の線形フラグメントの SMARTS 表記例を図 2 に示す.

[O;H1]c:*:*−*−C[N;!H0]

図 2 SMARTS 表記の例

元のフラグメントの O2H は SMARTS では[O;H1]となる. これは，酸素かつ，1 つの水素が結合していることを意味する. また，元のフラグメントで省略されている原子は“*”または“A”で表す. 省略原子が芳香族原子である場合は，“a”とする. 右端の記述[N;!H0]は元のフラグメントの N3H に対応しており，窒素かつ，1 つ以上の水素が結合していることを意味する. この例は一例であり，厳密に線形フラグメントを SMARTS 表記に変換する必要

連絡先：岡田孝，〒669−1337 兵庫県三田市学園 1−2，関西学院大学 理工学部 情報科学科，Tel：079−565−7709，okada−office@ksc.kwansei.ac.jp

はない．より一般性を持たせた表記にするなど，ユーザ（専門家）の知識を活かせることを考慮している．

SMARTS 表記による線形フラグメントの洗練には Greedy Search を用いて，以下のように行う．(1) 洗練時に用いる原子リストおよび結合リストを用意する．(2) 原子リストおよび結合リストの全組み合わせを，種線形フラグメントの全原子間に順に挿入し，その都度，新フラグメントの BSS 値を求める．(3) ここまでを 1 ステップとし，その中で BSS 値が最大となる新フラグメントをそのステップの代表として，次ステップの種フラグメントとする．(4) 前ステップの BSS 値を下回るまで，上記の操作を繰り返し，最終的な洗練構造を決定する．ここで，BSS（Between groups Sum of Squares）値（式 1）は，カスケードモデルで用いられている，ルール強度を示す指標である[Okada 99].

$$\mathrm{BSS}^g = \frac{n^g}{2}\sum_{\alpha}(p^g(\alpha) - p(\alpha))^2 \qquad (1)$$

3. 実データによる特徴構造の抽出

上記の考えをシステムに実装し，実データを用いて先行研究との比較検討を行った．データセットは，米国 MDL 社の開発医薬品データベース MDDR（MDL Drug Data Report）[MDL 01] に収載されている，3 種の異なる受容体（D1, D2, Autoreceptor）に作用するドーパミンアゴニスト（370 件）を用いた．先行研究において同一データを用いて解析を行ったカスケードモデルの出力ルールを利用し，ルールに示されている線形フラグメントを種として解析を行い，結果の比較を試みた．ここでは D1 アゴニスト（以下 D1Ag）に注目した解析結果を示す．

3.1 主ルール：Rule1

Rule1（最も強力なルール）の主条件として示されている線形フラグメント（O2H-c3:c3-O2H）を種フラグメントとし，OccO として解析を行った．解析結果を表 1 に示す．

表 1 Rule1 の主条件フラグメントを用いた解析結果

	種フラグメント	洗練フラグメント
SMARTS	OccO	Oc(:c(:c(:c(:c))(-C(-C(-N))))(-C))cO
BSS	12.175	22.262
支持事例数	115: (Y = 57, N = 58)	37: (Y = 35, N = 2)

Y：D1Ag 活性を持つ事例数，N：D1Ag 活性を持たない事例数

種フラグメントでは，D1Ag 活性を持つ事例と持たない事例数がほぼ同じであり，種フラグメント自体では D1Ag に特徴的とは言えない．システムによって得られた洗練フラグメントの支持事例数は 37 件であり，その内の 35 件が D1Ag 活性を持っている．このことから洗練フラグメントは，D1Ag に特徴的な構造であると考えられる．システムに実装されている，支持事例構造群表示機能を利用することで，実際に構造を確認することができる．洗練フラグメントの支持事例構造を図 3 に示す．洗練フラグメントに対応する部分の色を変えることにより視覚的にも理解しやすい．また，この洗練フラグメントは，先行研究で導かれた D1Ag に特徴的な構造と一致することも確認できた．先行研究では，Rule1 に記述されている全てのフラグメントと対象構造群を総合的に解釈し，特徴的なフラグメントを抽出してきたことと比べると，主条件のみの入力で同一の特徴的構造を抽出できることは，本システムの有用性を示唆するものと考える．

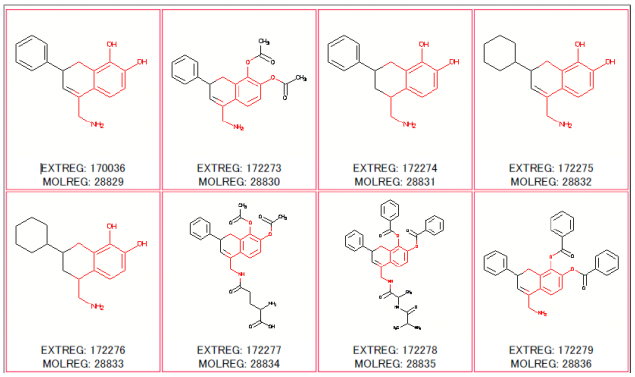

図 3 洗練フラグメントの支持事例構造（一例）

種フラグメントの，異なる SMARTS 表記（[O;H1]cc[O;H1]）を用いた解析結果を表 2 に示す．表から分かるように，種フラグメントの支持事例 52 件のうち D1Ag 活性を持つ構造は 50 件であり，洗練フラグメントでもその内訳はほぼ同じである．また，表 1 の結果で得られた D1Ag 活性の構造 35 件も含まれていることが分かった．このように，種フラグメントの記述によって，より特徴的な洗練フラグメントの取得が可能となることが分かった．さらに，種フラグメントの記述に専門家の知識を導入することで，より効率的な洗練が行えることを示唆する結果となった．

表 2 種フラグメントの異なる記述による解析結果

	種フラグメント	洗練フラグメント
SMARTS	[O;H1]cc[O;H1]	[O;H1]c(:c(:c(:c(:c))(-C(-C(-N)))))c[O;H1]
BSS	32.557	33.471
支持事例数	52: (Y = 50, N = 2)	51: (Y = 50, N = 1)

4. まとめ

化合物の特徴的リード構造洗練による知識抽出システムの開発を行った．洗練作業の自動化により，効率的な構造特徴の抽出が可能となった．また，専門知識を反映させることで，フラグメントをより効果的に洗練できることを示した．これらの結果は，カスケードモデルを用いた，活性化合物に関する効果的な知識獲得に大きく寄与するものと考える．

参考文献

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

[Okada 99] Okada, T.: Rule Induction in Cascade Model Based on Sum of Squares Decomposition., in *PKDD*, pp.468–474 (1999).

[Okada 02a] Okada, T.: Datascape Survey Using the CascadeModel., in *Discovery Science*, pp. 233–246 (2002).

[Okada 02b] Okada, T.: Discovery of Structure Activity Relationships using the Cascade Model: The Mutagenicity of Aromatic Nitro Compounds, *J. Comput. Aided Chem.*, Vol. 2, pp. 79–86 (2002).

[Okada 05] Okada, T., Yamakawa, M., and Niitsuma, H.: Spiral Mining using Attributes from 3D Molecular Structures, to be published in Active Mining (2005).

[Weininger 88] Weininger, D.: SMILES, a chemical language and information system. 1. Introduction and Encoding Rules, *J.Chem. Inf. Comput. Sci.*, **28**, 31-36 (1988).