知能と複雑系　132－28
(2003. 3. 15)

薬物分子の三次元構造類似性にもとづくデータマイニング

加藤 博明　高橋 由雅　阿部 英次

豊橋技術科学大学知識情報工学系 〒441-8580 豊橋市天伯町雲雀ヶ丘 1-1

E-mail: hiro@cilab.tutkie.tut.ac.jp, taka@mis.tutkie.tut.ac.jp, abe@cilab.tutkie.tut.ac.jp

あらまし　化学物質の種々の性質はその化学構造と密接に関連していることは良く知られている事実である．筆者らはこれまでに，グラフ論的なアプローチにもとづく三次元共通構造特徴の自動認識システム COMPASS (COMmon geometric PAttern Search System)の開発を進めてきた．本研究ではこれを基礎とし，与えられたデータセットに対する三次元構造類似性検索への応用を試みた．ここでは，クエリー分子とデータセット中の各分子の三次元構造を COMPASS を用いて比較し，探索された共通部分構造のサイズ（ここでは構成原子(団)の数）をこれらの間の類似性の尺度と定義した．治験医薬品データベース MDDR-3D から抽出したテストデータベースを対象に，ドーパミン D2 アンタゴニスト活性を持つ分子をクエリーとした類似構造検索の結果，同様の活性を持つ化合物が上位にランクされ，本法の有用性を強く示唆する結果を得た．

キーワード　構造類似性，COMPASS，化学グラフ，三次元構造特徴，データマイニング

Data Mining Based on 3D Structural Similarity of Drug Molecules

Hiroaki KATO　Yoshimasa TAKAHASHI　Hidetsugu ABE

Department of Knowledge-based Information Engineering, Toyohashi University of Technology
1-1 Hibarigaoka, Tempaku-cho, Toyohashi, 441-8580 Japan

E-mail: hiro@cilab.tutkie.tut.ac.jp, taka@mis.tutkie.tut.ac.jp, abe@cilab.tutkie.tut.ac.jp

Abstract　It is well known the structure of an organic molecule has rich information related to various physicochemical properties and biological activities of it. In the preceding works, we have developed a computer program, named COMPASS (COMmon geometric PAttern Search System), for automated identification of 3D common structural features among molecular structures by a graph theoretical approach. In the present work, we have applied it to the task of similar 3D structure searching on a given data set. Here, a query molecule and every molecule in the data set are compared using COMPASS. A similarity criterion for a pair of structures is defined as the size, i.e. the number of coincident atoms (or atomic groups), between them. The search experiment was made with a dopamine D2 antagonist molecule as a query and a test data set extracted from MDDR-3D. The result strongly suggested that the present approach can be applicable to evaluate 3D structural similarity of drug molecules.

Keyword　Structural Similarity, COMPASS, Chemical Graph, 3D Structural Feature, Data Mining

1. はじめに

化学物質の種々の性質はその化学構造と密接に関連していることは良く知られている事実であり，合成技術等の進歩に伴い蓄積された大量の構造データからその構造-活性（物性）相関に関する有用な知識を効率よく抽出するための新たなマイニング技術の確立が切望されている．特に，新規有用物質の候補構造探索やリスク評価における特性予測問題では，トポロジカル（二次元的）な構造情報だけでなく，その立体構造を考慮したより詳細な構造特徴解析もまた極めて重要な意味を持つと考えられる[1]．筆者らはこれまでに，グラフ論的なアプローチにもとづく三次元共通構造特徴の自動認識，並びに三次元部分構造検索のためのシステムの開発を進めてきた[2,3]．本研究では，これらの成果を基礎とし三次元構造類似性検索のためのアプローチの検討と，治験医薬品データベース MDDR-3D [4]を対象とした三次元構造類似性にもとづくデータマイニングへの応用を試みた．

2. 方法

2.1. COMPASS アルゴリズムの概要

三次元共通構造特徴の自動認識システム COMPASS (COMmon geometric PAttern Search System) [2]では，与えられる化合物分子の三次元構造は rigid なものと仮

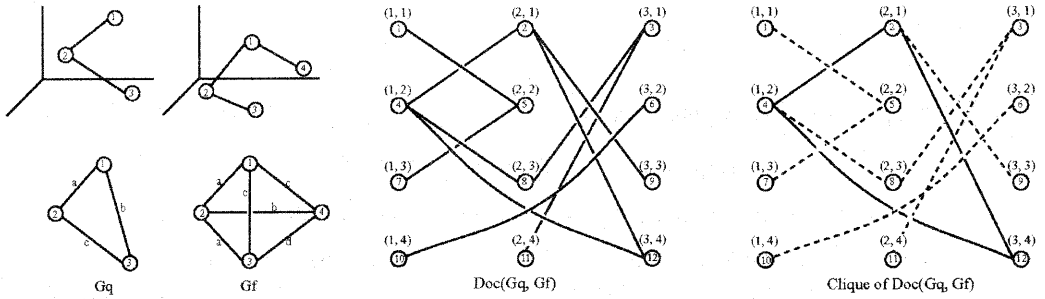

図1 二つのグラフ G_q, G_f から得られたドッキンググラフとクリークの例.

定する．今，全ての原子を等価とみなせば，与えられた各化合物の構造は，三次元空間上の各構成原子に対応する点の集合として取り扱うことができる．また，これらの点の集合を原子間の（ユークリッド）距離行列として記述し，三次元幾何情報を含めた化合物分子の構造をエッジ重み付き完全グラフとして表現する．これにより，化合物分子の三次元部分構造検索の問題を，部分グラフのマッチング問題として取り扱うことができる．二つの化合物構造 q と f に対応する分子グラフ G_q, G_f から次式のように定義されるドッキンググラフ $Doc(G_q, G_f)$ を生成する．

$$Doc(G_q, G_f) = < \mathbf{V}, \mathbf{E} >$$

$$\mathbf{V} = < (\sigma, \mu) \mid \sigma \in G_q, \mu \in G_f >$$

$$\mathbf{E} = < [(\sigma_i, \mu_k), (\sigma_j, \mu_l)] \mid |w_q(i, j) - w_f(k, l)| \leq \delta >$$

ここで，$\mathbf{V}$ と $\mathbf{E}$ はそれぞれドッキンググラフのノード集合並びにエッジ集合を，σとμはそれぞれグラフ G_q, G_f 中の構成ノード（すなわち，化合物構造 q と f の構成原子）を表わす．また，$w_q(i, j)$ はグラフ G_q 中のノード i とノード j との間のエッジの重み（化合物 q の原子 i と原子 j との間の距離），$w_f(k, l)$ はグラフ G_f 中のノード k と l との間のエッジ重みを表わす．なお，δは両者のエッジを等価とみなすための，エッジ重み（原子間距離）の許容度である．ドッキンググラフ $Doc(G_q, G_f)$ の例を図1に示す．以上のようにして生成したドッキンググラフからクリーク（すなわち最大完全部分グラフ）を探索することは，元の二つのグラフの最大共通部分グラフを探索することと等価であり，最終的に得られたクリークを構成するノードの集合はここで求める最大共通幾何パターンの構成原子の集合に対応付けられる．なお，各構成原子の種類やその周辺環境の情報を対応するノードに重み付けることにより，これらの違いを考慮した化学的により詳細な構造特徴を探索することも可能である．

2.2. 三次元構造類似性検索

COMPASS では，いわゆる「最大公約数」的な要素を探索するので，探索対象とするグループ（例えば共通の活性を持つ化合物群）の中に一つでも例外的な構造がある場合，あるいはそれらが本質的に複数の構造クラスから構成されるなどの場合，結果がそれに引っ張られ，意味のある特徴抽出ができないことがある．一方，部分構造検索[3]は，クエリーとして指定した構造（部分構造）を完全に内包するものだけを検索・出力するものであり，また注目する三次元部分構造の情報（例えば活性部位）をあらかじめ定義できなければ利用できないという問題がある．

そのため本研究では，クエリー（プローブ）として指定した化合物構造と，探索対象データベース中の各構造との間で COMPASS による共通構造特徴探索を行ない，探索された最大共通部分構造のサイズ（ここでは構成原子(団)の数）をこれら二つの構造間の類似性の尺度と定義する．以上の処理を，データベース中の

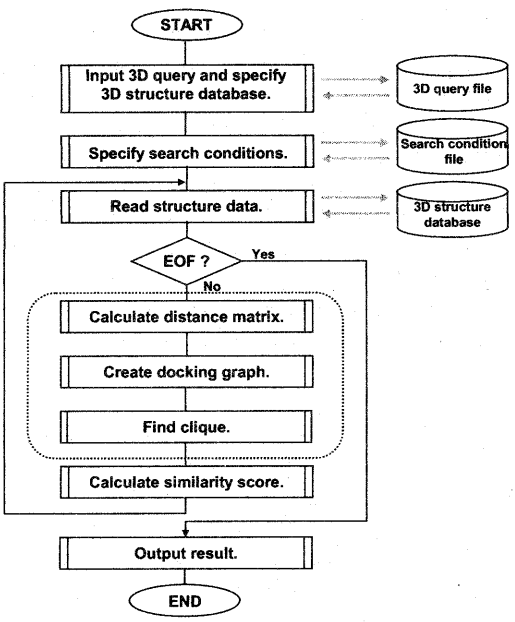

図2 三次元類似性検索の処理の流れ.

それぞれの構造に対して行ない，類似度の高い順に，指定された候補数だけ出力することで構造類似性検索を実現した．全体の処理の流れを図２に示す．

3. 結果と考察

3.1. テストデータセットによる評価実験

CSD (Cambridge Structural Database)から抽出した９種の化合物（エステル類）からなる小規模なテストデータセット（図３）を利用して性能評価のための検索実験を試みた．なお，本研究では水素原子は省略して表現し，また分子の三次元構造は rigid なものと考え，配座等は考慮しないこととする．図３の(2)の構造をクエリーとし，検索条件：距離の許容度 0.6Å・原子の種類を区別する，のもとでの三次元類似性検索の結果を表１に示す．クエリー自身を除いて共通部分構造サイズ(MCS)の最も大きいもの（最も類似しているもの）は(7)，逆に最も類似していないものは(5)と(8)となり，視覚的にも妥当な結果であることが確認できた．

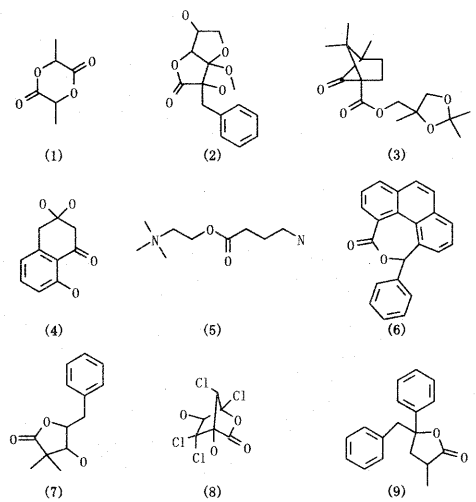

図３ CSD から抽出したテストデータセット.

表１ (2)をクエリーとした構造類似性検索結果.

	Size	MCS	R_DB	R_Min
(1)	10	8	80	80
(2)	20	20	100	100
(3)	23	10	43	50
(4)	14	10	71	71
(5)	13	7	53	53
(6)	24	11	45	55
(7)	16	12	75	75
(8)	14	7	50	50
(9)	20	11	55	55

ところで，共通構造サイズを尺度とした比較では，データベース中の各化合物構造のサイズ(構成原子数)に大きなバラつきがある場合，あるいはクエリーに完全に内包される部分構造が存在し，それらの特徴を強調したい場合には，十分であるとはいえない．そのような場合，比較対象構造のサイズ（あるいは，クエリーと比較してより小さい方）に対する共通部分構造サイズの割合(%)を求め，この相対的な値を尺度として評価した方がより適切であると考えられる．このような相対的な評価尺度(R_DB または R_Min)に注目して前述のデータセットを評価したところ (1), (4), (7)の類似度が高く，先のものとは異なる視点での類似性検索結果を得ることができた．

3.2. 薬物データベースに対する類似性解析

治験医薬品データベース MDDR-3D（図４）には，ID 番号(EXTREG)，CAS 登録番号，化合物名などの基本情報と，生物活性情報(登録活性記述子 約 800 種)，構造情報（二次元構造式，並びに CORINA によりモデリングした三次元座標情報）などの情報が収められており，Ver. 2001.1 には，約 12 万件のデータが登録されている [4]．その活性情報の分布を調査したところ，出現頻度が最も多いものは抗腫瘍活性(Antineoplastic)の 12,847 であり，逆に，少ない方ではゼロ（現在の DB では記述のない）の活性が 98 種，１回しか出現しない（対応する化合物がただ一つの）ものが 39 種であった．化合物データの側から見ると，各化合物は平均 2 種類の活性を持つことが分かった（活性情報を持たないデータも 28 件存在した）．

神経伝達物質であるドーパミン(Dopamine) に注目すると，そのアンタゴニスト活性を有する化合物は計 1,364 件登録されていた．ここでは，このうち約 10%（144 件）を任意に選択したものに，MDDR 全データセットから約 1%（1,154 化合物）抽出したものをノイズとして加えた，計 1,298 件からなるテストデータベースを作成した．

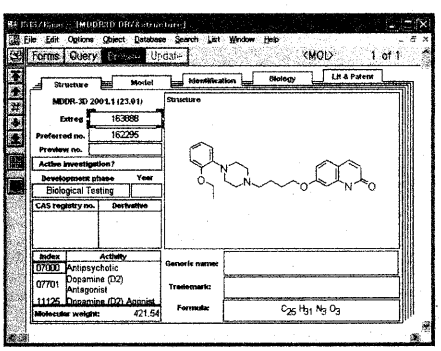

図４ MDDR-3D データベースの実行画面例.

クエリーとしてドーパミン D2 アンタゴニスト活性を持つ化合物（EXTREG: 163888，図４）を設定し，先と同様の検索条件のもとで行なった類似構造検索の結果の一例をその図５に示す．ここでは，各化合物分子の ID 番号(EXTREG)，三次元構造式，活性情報を示した．その結果，最も類似しているものはクエリー自身（自明）で，２番目，３番目に類似している構造として，それぞれ D3，D2 受容体に対するアンタゴニスト活性を有する化合物が検索された．上位 20 件について調べると，うち５件（クエリー自身を含む）が同じ D2 受容体に対する活性を持ち，また D3, D4 タイプのものがそれぞれ３件ずつ検索された．また，これらの活性はよりジェネリックな活性クラスとして抗精神病薬（Antipsychotic）の活性情報を持つが，上位 20 件中 13 件が同活性クラスに属していることが示された．

4. まとめと今後の課題

本研究では，二つの化合物間の共通部分構造サイズに注目した三次元構造類似性の尺度を定義し，構造類似性にもとづくデータマイニングの可能性について検討を試みた．MDDR-3D を対象にドーパミンアンタゴニスト活性を持つ化合物をクエリーとした類似構造検索の結果，同様の活性を持つ化合物が上位にランクされ，本法の有用性を強く示唆する結果を得ることができた．今後は，別途検討を進めている TFS 法[5]との併用による，より詳細な構造類似性解析に対する効果を調査するとともに，ベンゼン環の縮約など化学的により合理的な構造表現の導入による改良と，得られた類似構造の活性との関係について引き続き検討を行なう．

謝　辞

本研究は文部科学省科学研究費補助金・特定領域研究(B)「アクティブマイニング」のもとに行われたものであることを明記して謝意を表する．

文　献

[1] P.Willett, Chemical similarity searching, *J. Chem. Inf. Comput. Sci.*, **38**, pp.983-996 (1998).

[2] Y.Takahashi, S.Maeda, and S.Sasaki, Automated Recognition of common geometrical patterns among a variety of three-dimensional molecular structures, *Anal. Chim. Acta*, **200**, pp.363-377 (1987).

[3] H.Kato, and Y.Takahashi, Development of a three-dimensional substructure search program for organic molecules, *Bull. Chem. Soc. Jpn.*, **70**, pp.123-127 (1997).

[4] MDL Information Systems, Inc., *MDL Drug Data Report -3D*, Ver. 2001.1 (2001).

[5] Y.Takahashi, H.Ohoka, and Y.Ishiyama, Structural similarity analysis based on topological fragment spectra, *Advances in Molecular Similarity*, **2**, (Eds. R.Carbo & P.Mezey), JAI Press, Greenwich, CT, pp.93-104 (1998).

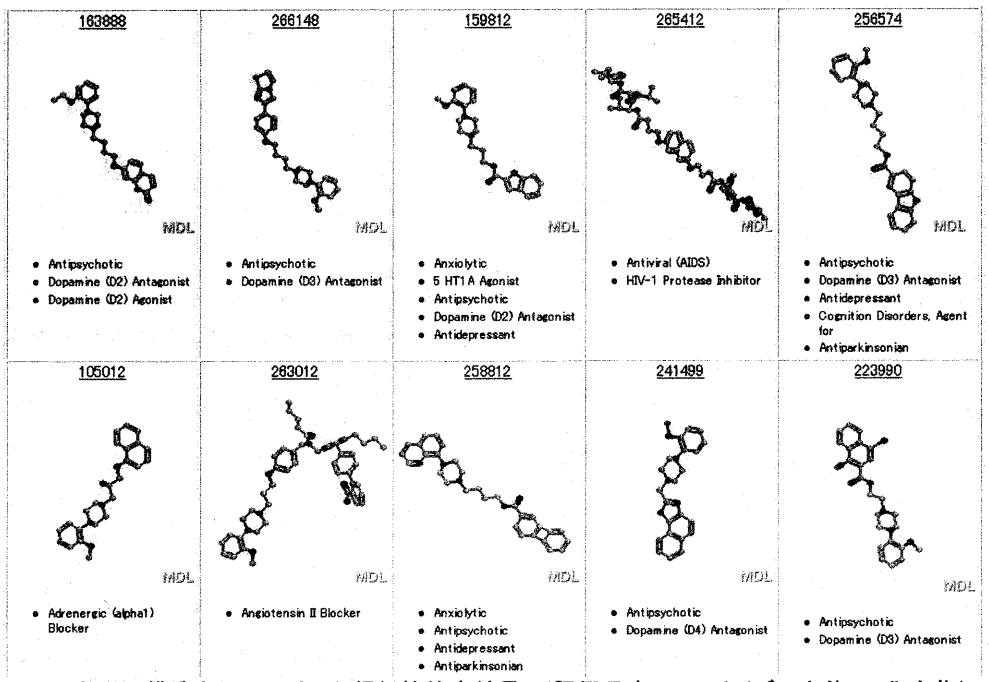

図5　左上の構造をクエリーとした類似性検索結果．（評価尺度 MCS サイズの上位 10 化合物）