1A2-1

NTG を利用した薬物分子グラフマイニングのための知識ベースの構築と活性推定への応用

Construction of Knowledge-base for Graph Mining of Drug Molecules Using NTG and Application to Activity Prediction

栗林 滝　　高橋 由雅
Ryo Kuribayashi　　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 proposed a basic skeletal feature representation of molecules with a graph called Non-Terminal vertex Graph (NTG) that has no vertices of degree of 1. A large number of NTGs were extracted from a structure database of investigative new drugs MDDR. In this paper, we have constructed a knowledge base of drug molecules that is based on the NTGs. And Active class prediction of drugs has also investigated using the knowledgebase. The details of the approach will be discussed with an illustrative example.

1. はじめに

先に当研究室では，大野[ohno 02]，青木[aoki 04]らによって，NTG（Non-Terminal vertex Graph）を利用した分子グラフマイニングの研究が行われてきた．それは，薬物構造データベース中の分子グラフから NTG を取り出し，表現レベルごとに NTG データベースを作成して，ある薬物活性に特徴的な NTG の解析を行うというものであった．

本研究では，先行研究で作成された NTG データベースを利用して知識ベースを構築し，これを利用して，活性未知化合物が持つ薬物活性クラスの推定を試みた．

2. Non-Terminal vertex Graph(NTG)

NTG とは頂点次数が 1 となるような頂点および孤立頂点を持たないグラフとして定義される．従って，化学構造のグラフ表現（分子グラフ）から抽出される NTG は環を中心とした基本骨格を表すものと見なすことができる．また，NTG には原子や結合タイプの区別の有無に応じて様々な表現が可能であり，異なるグラフ表現に対応した 5 つの表現レベルが定義されている．これらをまとめて図１に示す．

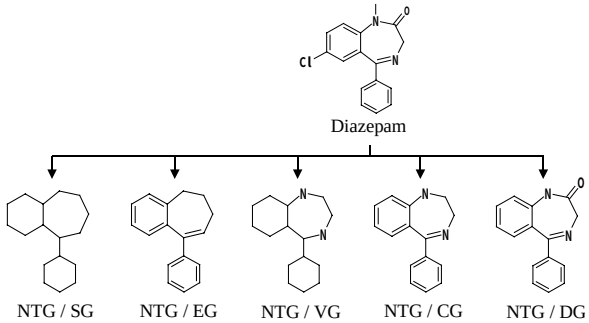

図 1　Diazepam から抽出した各表現レベルでの NTG

連絡先:高橋由雅，〒441-8580 愛知県豊橋市天伯町雲雀々丘 1-1 豊橋技術科学大学 知識情報工学系，Tel: 0532-44-4878, taka@mis.tutkie.tut.ac.jp

3. NTG を利用した活性推定の流れ

活性未知化合物（以降，クエリと呼ぶ）が持つ活性クラスを推定する流れを図 2 に示す．まず，先行研究で作成された NTG データベース群を利用して，NTG 関係データベースを構築する．一方，先行研究で開発された NTG 抽出ツールを用いて，クエリの NTG を抽出しておく．次に，NTG 関係データベースを参照し，クエリから抽出した NTG/SG と同じ構造を持つ NTG/SG を検索する．これが見つかれば，その NTG/SG を親とする NTG を NTG 関係データベース中から全て取り出し，知識ベースを作成する．この知識ベース中に記述された活性情報を解析し，クエリが持つ可能性の最も高い活性クラスを推定する．

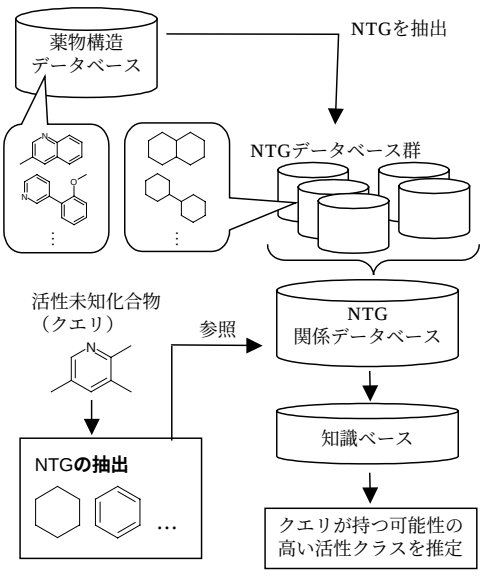

図 2　活性クラス推定の流れ

3.1 NTG 関係データベースの構築

本研究では，先行研究で作成された NTG データベース群を利用して NTG 関係データベースを構築し，Morgan 符号列を用いた構造情報だけでなく NTG 間の包含関係まで記述するこ

ととした. NTG 関係データベースは, 1 つのユニークな NTG につき 1 セットで情報を記述している. その仕様を図 3 に示す.

NTGID	NTG レベル	親 NTGID	由来構造数	活性情報
構成原子数				
Morgan 符号列（From Attachment，親リスト）				
Morgan 符号列（Ring Closure，閉環リスト）				
Morgan 符号列（Node Values，原子型リスト）				
Morgan 符号列（Line Values，結合型リスト）				

図 3　NTG 関係データベース（1 セット）の仕様

3.2　知識ベースの構築

クエリから抽出した NTG/SG と同じ構造の NTG/SG が NTG 関係データベースから見つかれば，その NTG/SG を親とする NTG を NTG 関係データベース中から全て取り出して，知識ベースを構築する．知識ベースの 1 セットの仕様を図 4 に示す．atom difference，bond difference とは，その NTG の構造と，クエリの NTG の構造を同じレベルで比較して，原子の種類，結合の多重度がどれだけ異なっているかを示す指標である．

NTGID	NTG レベル	活性情報
atom difference	bond difference	

図 4　知識ベース（1 セット）の仕様

4. 実データによる活性クラス推定

本研究では, MDDR データベース[MDL 02]から収載件数の多い上位 20 種類の活性クラスに属する治験薬 66808 化合物を抽出し, 原データとして用いた. これらの薬物構造データからすべての NTG を抽出し, 前述の方法に従って NTG 関係データベースを作成し, 実験を試みた. 本実験で用いた活性未知化合物と見立てたクエリ構造と単純グラフに対応する表現レベルで得られた NTG（NTG/SG）を図 5 に示す.

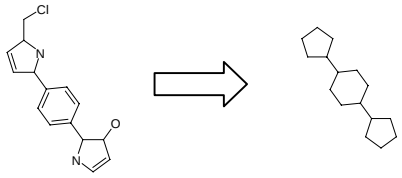

図 5　活性推定実験に用いた候補薬物の構造と抽出された NTG/SG

ここでのクエリ構造に対する活性クラスの推定は以下の手順で行った.

まず, 構築した NTG 関係データベースから, 図 5 の NTG/SG と一致する NTG/SG を持つ NTG を全て取り出し, 知識ベースに格納する. 知識ベース中の各 NTG はそれぞれ由来構造を持ち, それぞれの由来構造は様々な活性を持っているので, 一つの活性につき, いくつの由来構造がその活性を持っているかということを計算し, 推定に利用する. このとき, 単純に度数をカウントしていくだけでなく, NTG の表現レベルが詳細になるにつれて, 大きな値が得られるように重みづけを行う. これは, より詳細な表現レベルの NTG ほど, もとの化学構造の特徴をよく表しているとの考えに基づくものである. また, atom difference, bond difference も考慮し, これらの値が大きくなるほど, つまり, クエリとの構造的な差異が大きいほど, スコアが小さくなるように重みを設定することがより合理的であると考えられる. 本実験で設定した重み値 w(NTG)は, w(SG)=1, w(EG)=2, w(VG)=2, w(CG)=3, w(DG)=4 を用いた. また, atom difference および bond difference に対する重み値にはそれぞれ 0.9 を用いた. これらの値はすべて経験的に定めた.

例えば, 知識ベース中で atom difference = 3, bond difference = 0 の値を有するある NTG/CG に対し, A という活性を持つ由来構造が 10 件あるとする. このとき, atom difference の重みが 0.9 の場合, A 活性におけるその NTG/CG の由来構造数と重みの積（スコアと呼ぶ）は, 10 に 0.9 を 3 回掛けることになる. また, CG レベルの重みが 3 であるので, さらにこれに 3 を掛ける. つまり, この NTG/CG の A 活性におけるスコアは, $10 \times 0.9^3 \times 3 = 21.87$ となる. 一つの活性について知識ベース中の全ての NTG のスコアを求め, これらの合計をその活性の合計スコアとし, この合計スコアが最も高い活性を, クエリが持つ可能性の最も高い活性とする. ある活性の合計スコア S は, (1)式で求められる.

$$S = \sum_{l=SG}^{DG} \sum_{i=0}^{m_l} n_i w_l (w_a)^{d_a} (w_b)^{d_b} \qquad (1)$$

ここで, n は一つの NTG の中で現在注目している活性を持つ由来構造数, w は重み値, d は difference 値を表す. また, ℓ は NTG の表現レベル, i は レベルℓでの個々の NTG の識別子, a は atom difference, b は bond difference を表している.

図 5 の構造をクエリとして, MDDR 収載件数上位 20 件の NTG 関係データベースから知識ベースを構築し, 上述の方法で活性クラスを推定した結果, 表 1 のような結果を得た.

表 1　活性クラス推定の結果

活性名	合計スコア	割合（%）
Antibacterial	261.51	48.76
Oxazolidinone	223.41	41.66
Antimycobacterial	42.84	7.99
Platelet Antiaggregatory	4.25	0.79
gpIIb/IIIa Receptor Antagonist	4.25	0.79

5. まとめ

表 1 の結果から, 図 5 の仮想クエリは Antibacterial 活性を持つ可能性が最も高いことが分かった. しかし, Oxazolidinone 活性を持つ可能性もかなり高くなっており, これらは同時に持つ活性であるのか, 単独で持っている活性であるのかは, ここからは読み取れない. 今後は, この問題について検討するとともに, 別の活性推定アルゴリズムについても検討していきたい.

参考文献

[ohno 02]　大野貴生: Non-Terminal vertex Graph(NTG)を利用した薬物の構造特徴解析, 第 30 回構造活性相関シンポジウム講演要旨集, 41-42, 2002

[aoki 04]　青木寛人: Non-Terminal vertex Graph(NTG)を利用した薬物構造データマイニングツールの開発, 第 32 回構造活性相関シンポジウム講演要旨集, 131-132, 2004

[MDL 02]　MDL Drug Data Report, http://www.mdl.com/