知能と複雑系　132－27
(2003. 3. 15)

TFSを利用した薬物活性クラス分類とリスクレポート

高橋 由雅　　藤島 悟志　　横江 恭子

豊橋技術科学大学工学部 〒441-8580 豊橋市天伯町雲雀ケ丘 1-1

E-mail: taka,@mis.tutkie.tut.ac.jp　fujisima@mis.tutkie.tut.ac.jp　yokoe@mis.tutkie.tut.ac.jp

あらまし　本研究では先に提案したTopological Fragment Spectra(TFS)法のリスクレポートにおける有用性を検証するため，4種のドーパミン受容体(D1, D2, D3, D4)ごとのアンタゴニスト活性について，MDDRデータベース中の治験薬1227件からTFS人工ニューラルネットによる学習モデルを作成し，別途用意した137化合物について受容体を予測したところ，その予測精度は81%に上り，本方法がこれまでにない有効性を示すことを明らかにできた．また，TFS法による構造類似性検索の実験の結果，異なる活性クラスに属するにもかかわらず構造類似性の極めて高い例外分子が見出されることを示し，リスクレポートへの有効性を実証した．一方，化合物データベースをTFS仮想空間上に配置し，化合物の直接探索が可能なデータ可視化ツールMolSpaceを開発した．これにより，例外分子のリスクレポートからの専門家による考察の支援が可能となった．

キーワード　構造類似性，TFS，クラス分類，データマイニング，構造特徴解析，リスク評価

Risk Report Based on Structural Similarity of Chemicals

Yoshimasa TAKAHASHI　Satoshi FUJISIMA　and　Kyoko YOKOE

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

E-mail: taka,@mis.tutkie.tut.ac.jp　fujisima@mis.tutkie.tut.ac.jp　yokoe@mis.tutkie.tut.ac.jp

Abstract　The applicability of the Topological Fragment Spectra (TFS) method, which was reported in our preceding work, was validated in discriminating active classes of pharmaceutical drugs. Dopamine antagonists of 1,227 that interact with different type of receptors (D1, D2, D3 and D4) were used for training an artificial neural network(ANN) with their TFS to classify the type of action. The ANN classified 88% of the drugs into their own classes correctly. Then, the trained ANN model was used for predicting class unknown compounds. For other 137 compounds the active classes of 81% of all the compounds were correctly predicted. Beside, to validate an instance-based chemical risk report approach based on structural similarity, TFS-based similar structure searching was employed for identification of active molecular analogues with different activities. The TFS successfully identified structurally similar molecular analogues of our interest. In addition, a desktop software tool, called MolSpace, was also developed for visualizing massive molecular data space or TFS space. It makes us easy to compare an object molecule with neighbors in the same region of data space.

Keyword　Structural Similarity, TFS, Pattern Classification, Data mining, Structural Feature Analysis, Risk Assessment

1. はじめに

今日、計算機処理能力の飛躍的な向上と大容量記憶媒体の低廉化に伴い、膨大な情報の収集が容易に行えるようになった。その一方、その情報を解析、理解し、有効に活用する技術が追いついていないとの現状がある。このことから最近では様々な分野においてデータマイニングあるいはチャンス発見と呼ばれる大量のデータから有用な知識を発掘するための新たな技術の確立に多くの期待が寄せられている。

そこで本研究では化学物質の構造データマイニングのための基盤技術の一つとして、構造類似性を基礎とした化学データマイニング並びにチャンス発見のための効果的な手法の確立をめざすとともに，新規有用化学物質の候補構造の探索やリスク評価における特性予測問題への応用の可能性について実データを用いて検討を行った．

Topological fragment spectra (TFS)法は化学物質の構造式から可能な部分構造を列挙し，その数値的な特徴づけにもとづいて化学物質のトポロジカルな構造プロフィールをデジタルスペクトル（あるいは多次元数値ベクトル）として表す．化学グラフの各頂点原子をその隣接原子の数でラベルづけし，生成された部分グラフをこれらの総和によって特徴づけを行なうことによって化学構造を単純グラフと見なした場合の骨格のトポロジーを表わす特性スペクトルを得るこ

とができる．また，生成部分グラフを頂点に対応する原子の質量数の総和（フラグメント重量）によって特徴づけることにより，質量スペクトルに類似の構造フラグメントに関する特性スペクトルが得られる．このことから TFS 法を基礎とし，定義部分構造を必要としない構造全体の漠然とした類似性を考慮したより柔らかな構造情報の取り扱いが可能となる．

本研究では先に提案した Topological Fragment Spectra(TFS)法 [1]のリスクレポートにおける有用性を検証するため，市販の治験薬構造データベースをもとに，実データを用いて活性クラス識別問題における構造記述子としての TFS 法の有効性について評価するとともに、併せて構造類似性に基づくリスクレポートへの応用の可能性を検証した．

2. 方法

化学構造式の TFS 表現： 筆者らは構造特徴の定量的表現の一つとして，Topological Fragment Spectra (TFS)法を提案している．この手法は，部分構造の定義ファイルを必要とせず，与えられた構造から TFS を生成して，構造全体の漠然とした類似性評価を行うことができる．TFS は，対象とする化学構造の可能な部分構造をすべて列挙し，列挙したそれぞれの部分構造に対して数値的な特徴付けを行う．その特徴付けの値と出現頻度のヒストグラムを生成する．このヒストグラムが TFS であり，これを多次元パターンベクトルとして用いることで，化学構造を定量的に扱うことができる．また，TFS は部分構造の特徴付け(次数和，重量和，etc)を変えることにより，様々なスペクトルを生成することができる．

データセット： ここでは，米国 MDL 社の治験薬構造データベース MDDR（MDL Drug Data Report） [2]より抽出した 4 種の異なる受容体（D1, D2, D3, D4）に作用するドーパミンアンタゴニスト 1,364 種を対象に検討を行った．解析に際してはこれら全ての化学構造をもとに TFS を生成し，データベース化を行った．ここでの TFS 表現には結合数が 5 までのフラグメントを生成し，その質量数で特徴づけを行った．また，このようにして用意されたデータはそこから約 1 割にあたる 137 件をランダムに抽出し，予測集合として別途確保し，残り 9 割にあたる 1227 件を訓練集合として用いた．

人工ニューラルネットワーク： これら薬物の活性クラス識別のモデル化には人工ニューラルネットワークを用いた．本研究では通常の完全結合型 3 層ネットワークモデルを用いた．入力シグナルには上述の TFS を利用し，ネットワークの学習には誤差逆伝播法を用いた．これらの処理には筆者らが別途作成した NNQSAR [3]を用いた．

3. 結果及び考察

まず初めに，MDDR より抽出した 4 種の異なる受容体（D1, D2, D3, D4）に作用するドーパミンアンタゴニスト 1,364 種を対象に，活性クラス識別問題における構造記述子としての TFS 法の有効性を検証した．分類機には誤差逆伝播法にもとづく人工ニューラルネットワーク（ANN）を用いた．実験に際しては対象データを事前に訓練集合（1,227 化合物）と予測集合（137 化合物）に分割し，ネットワーク学習には訓練集合 1,227 化合物を用い，これらのすべてについてエッジサイズ（生成フラグメントの結合数） 5 以内のフラグメントにもとづく TFS を生成し，入力シグナルとした．学習には入力層，中間層（1 層），出力層からなる 3 層ニューラルネットワークを用いた．入力層，中間層，出力層のユニット数はそれぞれ 165, 3, 4 とした．学習の結果，訓練集合 1,227 化合物中 1,087 化合物（88.6%）の活性クラスを正しく学習・認識することができた(表 1)．一方，学習済み ANN モデルをもとに，別途用意した予測集合 137 化合物（D1：18 化合物，D2：39 化合物，D3：24 化合物，D4：56 化合物）の活性予測を試みた．その結果，各々のクラスごとに 61%, 69%, 95%, 89%の予測率を得た．また，全体で 111／137 化合物(81%)についてその活性クラスを正しく予測することができた．このことは化学構造からの活性クラス識別における構造特徴記述子としての TFS の有効性を強く示すものである．

表 1. ニューラルネットワークによるドーパミン受容体アンタゴニストの活性クラス分類

Class	Training		Prediction	
	No. of samples	Correct (%)	No. of samples	Correct (%)
All	1227	1087 (88.6)	137	111 (81.0)
D1	155	112 (72.3)	18	11 (61.1)
D2	356	312 (87.6)	39	27 69.2)
D3	216	193 (89.4)	24	23 (95.8)
D4	500	470 (94.0)	56	50 (89.3)

また，構造類似性にもとづくリスク評価の観点から，上記のドーパミンアンタゴニストデータセットに対し，別途 MDDR よりランダムに抽出した 10,000 件（ドーパミンアンタゴニストは除く）をノイズデータとして加え，類似構造検索の実験を試みた．これらの全てについて TFS を生成するとともに，構造(１)を query と

した探索結果を図1に示す．これは11,227件を対象にqueryに対して最もよく似ているもの上位10件の化学構造を示したものである．図からも分かるように，これらの化学構造は相互に極めて類似しているものであることは明らかである．また，ここに示された10化合物は全てqueryと同じD1アンタゴニスト活性を有するものであった．

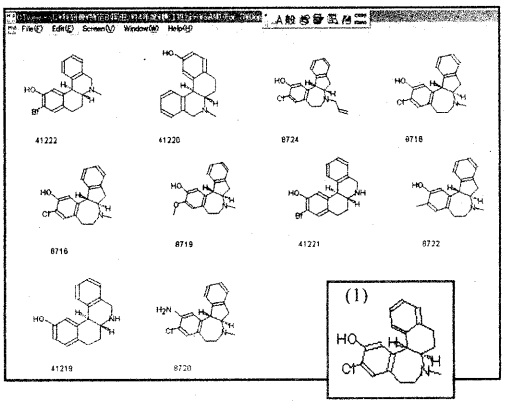


図1. 類似構造検索（queryに対しての上位10構造）．クエリーは(D1)アンタゴニストであり，探索された10件も全て同活性を持つものであった．

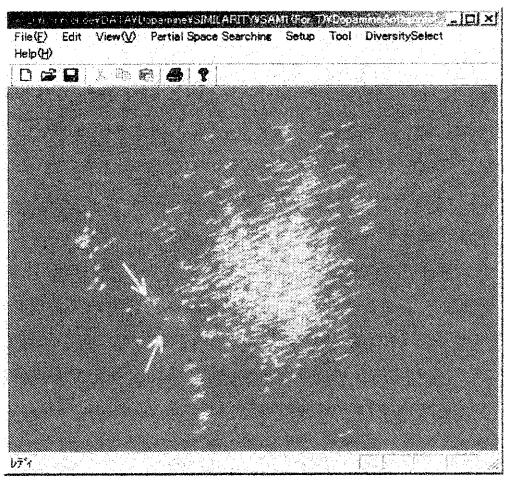


図2. MolSpaceによるデータ空間の可視化：1,227件のドーパミンアンタゴニストに10,000件のノイズ化合物を加えたTFS仮想空間．

また，本研究を通じて別途開発した多変量化学データ可視化ツールMolSpace [4]を用いてこれらのTFSデータの次元縮約による構造類似性空間の可視化を試みた．図2にその結果を示す．ここでは，全化合物構造のTFSデータを主成分分析による3次元縮約空間に写像し，各サンプルを構造オブジェクトとして直接表示している．図中，矢印で示した部分が上で述べた類似性検索での結果に対応する構造群を表す．このように，多次元数値ベクトルとして記述されるTFSデータ空間の可視化は，後述の例外分子に対するリスクレポートにおける考察支援にも極めて有効であると考える．

次に，このTFSを利用した類似性検索によるリスクレポートへの応用について検討を行った．上述のドーパミン受容体アンタゴニストのTFSデータベースを用い，D4受容体アンタゴニスト活性を持つ化合物を新規候補活性分子と想定して，前述の例の場合と同様に訓練集合（1227件）に1万件のノイズ化合物を加えたテストデータセットに対して，類似構造検索にもとづくリスク推定を行った．実験に用いたクエリー構造を図3に，TFSを利用した構造検索の結果を表2に示す．

ここでは類似性が高いものから10件を検索した．検索の結果，これら10化合物のうち，1～9位までの化合物が，すべて統合神経失調症の薬であり，また，そのうちの7位までの化合物がD4受容体に対するアンタゴニスト活性を持つものであることが判明した．このことから，この候補活性分子がD4受容体に対するドーパミンアンタゴニスト活性を持つ可能性が高いことが推察できる．8位の化合物についてはデータベース中に詳細な記述はなく，D4受容体アンタゴニストであるか否かは不明であるが，同じ統合神経失調症の薬であること，また類似性が高いとしてD4 antagonistの活性を持つ化合物と一緒に上位にランクされていることを考えると，この化合物もD4受容体アンタゴニストである可能性が高いと考えることができる．

一方，9位の化合物は，D2アンタゴニスト活性を有することが示されている（表2）．最近の遺伝子研究からは，D2受容体とD4受容体は同一タンパクファミリーであることが明らかにされていることを考えると，このD2アンタゴニストとの構造類似性も興味深い結果といえる．

最後に，表2の10位にランクされた化合物に注目すると，他の化合物とはことなり，抗高血圧薬活性を有することが分かる．他の9件の化合物（抗精神薬）と異なる活性にも関わらず，10位に検出されることを考えると，この化合物には抗精神薬関連の副作用の可

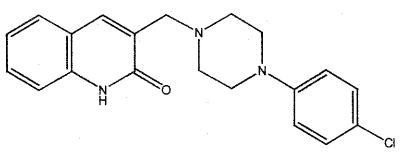
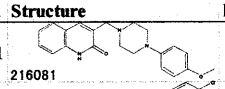
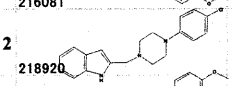
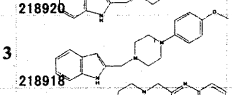
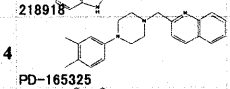
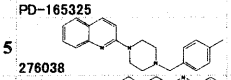
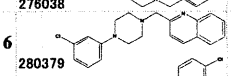
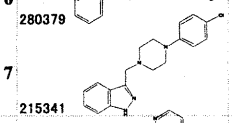
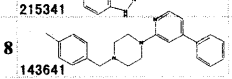
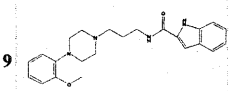
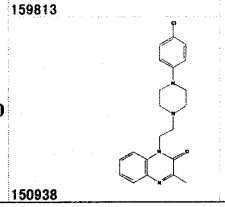


図3. リスクレポートの実験に用いたクエリー構造

表 2. TFS による類似構造検索の結果*

No.	Structure	Distance	Activity
1	 216081	10.954	D4 Antagonist; Antipsychotic
2	 218920	14.353	D4 Antagonist; Antipsychotic
3	 218918	15.875	D4 Antagonist; Antipsychotic
4	 PD-165325	16.912	D4 Antagonist; Antipsychotic; 5 HT2A Antagonist
5	 276038	17.146	D4 Antagonist; Antipsychotic
6	 280379	18.493	D4 Antagonist; Antipsychotic
7	 215341	18.947	D4 Antagonist; Antipsychotic
8	 143641	19.157	Antipsychotic
9	 159813	19.468	Dopamine (D2) Antagonist; Antidepressant; Anxiolytic; 5 HT1A Agonist; Antipsychotic
10	 150938	20.025	Antihypertensive; Adrenergic (alpha) Blocker

*訓練集合 1227 件の TFS データベースに対する類似度（ユークリッド距離）上位の 10 件の構造と薬理活性情報を示している.

能性があると推測することができる.

以上のことから，構造類似性にもとづく薬物候補の副作用など，広く化学物質の薬物のリスク推定，リスクレポートへの応用も十分期待できると考える.

4. まとめと今後の課題

本研究では先に提案した Topological Fragment Spectra(TFS)法のリスクレポートにおける有用性を検証するため，4 種のドーパミン受容体ごとの活性について，MDDR データベース中の治験薬 1227 件から TFS 人工ニューラルネットによる学習モデルを作成した. 別途用意した 137 化合物について受容体を予測したところ 81%の化合物について正しく予測することができ，本方法がこれまでにない有効性を示すことを明らかにした. また，TFS 法による構造類似性検索の実験の結果，異なる活性クラスに属するにもかかわらず構造類似性の極めて高い例外分子が見出されることを示し，リスクレポート問題への有効性を実証した. さらに，化合物データベースを TFS 仮想空間上に配置し，化合物の直接探索が可能なデータ可視化ツール MolSpace を開発した. これにより，例外分子のリスクレポートからの専門家による考察の支援が可能となった.

リスク評価レポートには様々な参照事例が不可欠である. 今後は，試行対象を拡大するとともにより大規模な実データを用いながら検討を進める必要がある. このことから，引き続き MDDR をデータソースとし，他の GPCR 関連活性群についても同様な検討を進め，その性能を検証する. また，TFS ピークからリスク推定に寄与する構造特徴を解析し，ユーザに分かり易い形で提供するためのツールについても併せて工夫する必要がある.

文　献

[1] 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).

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

[3] 安藤秀一，高橋由雅，構造活性相関研究のためのニューラルネットワークツール NNQSAR の開発，第 24 回情報化学討論会講演要旨集，pp.117-118 (2001).

[4] Y. Takahashi, M. Konji, and S. Fujishima, MolSpace: A computer desktop tool for visualization of massive molecular data, *J. Mol. Graph. Model.*, **21**, pp.333-339 (2003).