1F1-03

# 化学構造の TFS 表現に対する特徴抽出と薬物活性クラス分類

## Feature Selection for TFS Representation of Chemical Structure and Classification of Pharmacological Activity of Drugs

藤島悟志[*1]
Satoshi Fujishima

高橋由雅[*2]
Yoshimasa Takahashi

[*1] 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

[*2] 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

This paper describes a feature selection technique for TFS-based pattern classification of chemicals that have a particular drug activity. In the present method, the frequencies of all the TFS peaks are analyzed to every activity class. The results are used for finding predominant peaks of individual classes. The reduced TFS patterns that consist of the predominant peaks were applied to TFS-based support vector machine, and tested their utility for classification of drug activities. Computational trial using 1354 dopamine antagonists gave us successful results. And several characteristic structural features were also identified on the predominant peaks by means of the TFS peak identification.

## 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. 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.

筆者らは, 薬物構造データマイニング手法の確立を目的に, 先に構造類似性を基礎とした事例ベースの活用による薬理活性の推定やリスクレポートの可能性について検討するとともに, 構造類似性評価を基礎としたデータマイニングにもとづく知識発見への応用の可能性を検討した[Takahashi 02, 高橋 03]. そこでは, 事前の部分構造知識を必要としない構造特徴のプロファイリング手法である Topological fragment spectra (TFS) 法[Takahashi 98]を用い, 化学構造に対応する固有の TFS を生成し, TFS 空間における類似性をもとに, 類似構造薬物の検索を実現した. 同様な観点から, 薬物活性クラスの分類・識別問題における TFS 法の応用を試み, TFS を入力シグナルとした人工ニューラルネットワーク (Artificial Neural Network, ANN) やサポートベクターマシン (Support Vector Machine, SVM) の有用性を明らかにした[Fujishima 04 , 藤島 04b, 錦織 03]. これらの研究から, TFS を基礎とした構造類似性検索や活性クラス分類が有効であることは明らかであり, TFS が化合物の構造特徴を良く表現していると考えることができる.

また, 筆者らは, TFS から化合物の構造特徴の抽出について検討を行い, TFS の差スペクトルの概念を導入することにより, 注目する活性に特徴的なピークとそれに含まれる特徴的なフラグメントを抽出できることを示した[藤島 04c].

本研究では, TFS 差スペクトルにおける活性クラスごとの厳密なピークの有無に注目した特徴抽出法とは別に, 統計的な“頻度”を考慮した TFS ピークの特徴抽出を試みるとともに, 薬物活性クラス分類における本法の有用性について検討を行った.

連絡先：藤島悟志, 〒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：079-565-8149, fujishima@ksc.kwansei.ac.jp

## 2. 方法

### 2.1 TFS とは

TFS とは, 化学構造の定量的な類似性評価を目的に, 高橋らによって考案された構造情報の数値的な記述手法の一つである. その生成手順は, (1) 対象とする化学構造式から可能なフラグメントをすべて列挙し, (2) 列挙したそれぞれのフラグメントに対して数値的な特徴付けを行う. そして, (3) 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, 化学物質の構造特徴を数値的・定量的に表すことができる (図 1). この手法は, フラグメントの定義ファイルを必要とせず, また, 生成されたフラグメントの特徴付けの方法を工夫することによって, 様々な特性スペクトルを生成することができる[Takahashi 03].

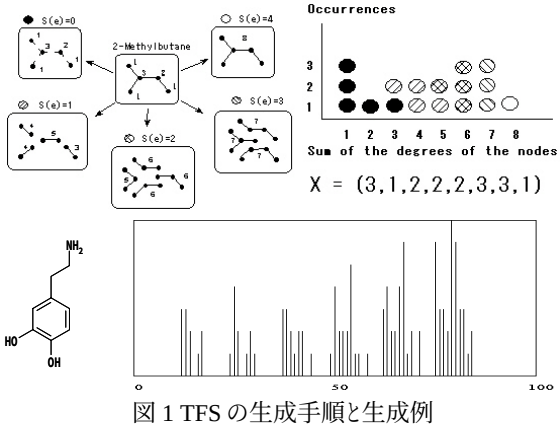


図 1 TFS の生成手順と生成例

### 2.2 TFS ピーク出現頻度とクラス識別力

本研究では, TFS ピークの特徴抽出として, ピーク出現頻度の利用を検討した. クラス毎のピーク出現頻度をもとに, クラス間でのピーク出現率の差をとることにより, 個々のピークの各クラスへの寄与度を算出する.

ここでは例として，2 つのクラス a, b を考える．まず各サンプルの TFS ピークを，以下に示す関数（$bin(\mathbf{X})$）によって 2 値化を行う（ピークの有無を表す）．

$$bin(X) = (x'_1, x'_2, ..., x'_d)$$
$$x'_j = \begin{cases} 1: \text{注目位置にピークが存在} \\ 0: \text{注目位置にピークが存在しない} \end{cases} \quad (1)$$

$k$ 番目の特徴に対するクラス毎の出現率とクラス識別力 $Da_k$ を以下に定義する．

$$Da_k = \left( \frac{\sum_i^{n_a} x'^{(a)}_{ik}}{n_a} - \frac{\sum_i^{n_b} x'^{(b)}_{ik}}{n_b} \right)^2 \quad (2)$$

ここで，$n_a$, $n_b$ はそれぞれクラス a, b のサンプル数を表す．$k$ 番目の特徴に対するクラス毎の出現率を求め，クラス間での差をとることによって，その注目するピークのクラス依存度（クラス識別力）を計算する．多クラスの場合は，各クラス間での $Da_k$ を求め，その平均値を算出することによって，そのピークが持つクラス識別力とする．どのクラスにも頻繁に現れるピーク $k$ の $Da_k$ は 0 に近づき，特定クラスにのみ多く表れるピーク $k$ の $Da_k$ は 1 に近づく．

特徴抽出は，クラス識別力の最も高い特徴（ピーク）から順に，指定されたピーク数分を抽出することによって行われる．生成された合成スペクトルから，これら一連の処理を行うシステムの開発を行った．また，特徴抽出後の TFS を入力とした，SVM による活性クラス分類の精度について検証を行った．

### 2.3 データセット

本研究での計算機実験には，米国 MDL 社の治験薬構造データベース MDDR（MDL Drug Data Report）[MDL 01]に収載されている，4 種の異なる受容体（D1～D4）に作用するドーパミン(D*n*)アンタゴニスト（1,354 化合物）を用いた．

### 2.4 SVM

SVM は近年様々な応用領域において，優れた性能を示している学習アルゴリズムである．SVM ではクラスラベルとして $y_i \in \{-1,+1\}$ を有する学習サンプル $\mathbf{x}_i \in R^d$ を，$f(\mathbf{x}_i) = \mathrm{sgn}(g(\mathbf{x}_i))$ で分類するような識別関数 $g(\mathbf{x})$ を学習する．SVM の発展的な特徴の一つとして，カーネルトリックを用いた高次元写像による非線形分離性の向上がある．データ空間で定義され，Mercer の条件を満たすカーネル関数 $K(\mathbf{x}, \mathbf{x}')$ を導入することにより，写像空間での複雑な計算を避けて，元の空間で直接解くことができる．そのカーネル関数の一つとして，次式で定義される Gaussian Kernel があり，本研究ではこれを用いている．

$$K(\mathbf{x}, \mathbf{x}') = \exp\left( -\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{\sigma^2} \right) \quad (3)$$

学習の問題は凸二次計画問題に帰着でき，最適化問題を解いて得られる識別関数は，最終的に次式で表すことができる．

$$g(\mathbf{x}) = \sum_{i=1}^{l} \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + b \quad (4)$$

SVM は基本的には 2 クラス分類モデルであり，複数クラスに関する分類のためには SVM を組み合わせる必要がある．ここでは，$k$ クラスの分類問題を解くための一般的な組み合わせ法である，one-against-the-rest を利用した．実験に際しては，Dong らの提案した SMO(Sequential Minimal Optimization)改良アルゴリズム[Dong 02]をもとに当研究室で別途作成した SVM 学習ツールを用いた．

## 3. 結果と考察

### 3.1 特徴抽出と活性クラス分類

特徴抽出によって得られた TFS が，どの程度活性クラスを正しく識別･予測できるか検証するために，SVM による活性クラス分類を行った．ドーパミンアンタゴニストのデータから生成した TFS（164 次元）に対して特徴抽出を行い，5 種類の TFS（それぞれ，100, 80, 60, 40, 20 次元）を生成した．例として元の TFS と 20 次元の TFS を図 2に示す．

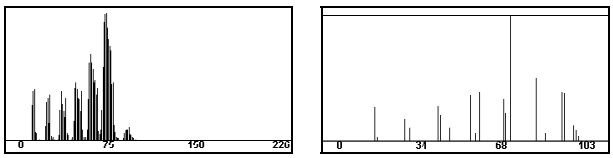


図 2 元の TFS（164 次元）と特徴抽出によって得られた TFS（20 次元）

特徴抽出前の元の TFS と特徴抽出後の次元数の異なる 5 種類の TFS を使用して SVM による活性クラスの識別実験をそれぞれ行った．表 1にその結果を示す．学習および予測共に次元数が減少するに従ってその精度は低下している．しかし，その低下は決して急激なものではなく，いずれも高精度で学習および予測が行えることがわかった．特に 20 次元の TFS では，学習においては 10%，予測おいては 4%程度低いだけであり，1/8 程度の次元数であるにもかかわらず，高い精度を保っていることは注目に値する．これらの結果から，出現頻度を利用した TFS の特徴抽出はクラスに依存した TFS ピークを抽出することができると考える．

表 1 特徴抽出によって得られた各次元数の TFS を使用した SVM による活性クラス分類の学習･予測結果

| 次元数 | | 164 | 100 | 80 | 60 | 40 | 20 |
|---|---|---|---|---|---|---|---|
| 学習 | ALL | **100** | **100** | **98.7** | **98.7** | **98.4** | **89.9** |
| | D1An | 100 | 100 | 99.3 | 98.5 | 96.3 | 92.6 |
| | D2An | 100 | 100 | 97.6 | 98.4 | 93.0 | 86.4 |
| | D3An | 100 | 100 | 98.1 | 98.1 | 94.8 | 86.9 |
| | D4An | 100 | 100 | 99.6 | 99.2 | 97.8 | 93.0 |
| 予測 | ALL | **94.1** | **93.4** | **94.1** | **91.9** | **90.4** | **90.4** |
| | D1An | 86.7 | 86.7 | 86.7 | 86.7 | 86.7 | 86.7 |
| | D2An | 95.3 | 93.0 | 93.0 | 88.4 | 90.7 | 88.4 |
| | D3An | 84.2 | 84.2 | 89.5 | 89.5 | 89.5 | 78.9 |
| | D4An | 98.3 | 98.3 | 98.3 | 96.6 | 93.2 | 96.6 |

### 3.2 ピーク同定による構造特徴解析

特徴抽出によって得られた TFS を用いたクラス分類は，元の TFS でのそれと同程度の精度を維持できることが分かった．これ

は, 注目する活性に特徴的なピークが効果的に抽出されていると考えることができる. そこで本研究では次に, 特徴的なピークが具体的にどのようなフラグメントを含んでいるのか解析を行った. 先に報告した TFS ピーク同定システム[藤島 04a]を用いてピーク同定を行うことで, 活性に特徴的なピークに含まれるフラグメントの解析を行った. 各ピークに含まれるフラグメントが活性クラスに寄与していることが分かれば, クラス識別力を考慮した特徴抽出の効果を示すことができる.

ここでは, 特徴抽出によって得られた 20 次元の TFS に対してピーク同定を行った結果を示す. 図 3に質量数 29 のピーク同定によって得られたフラグメントを示す.

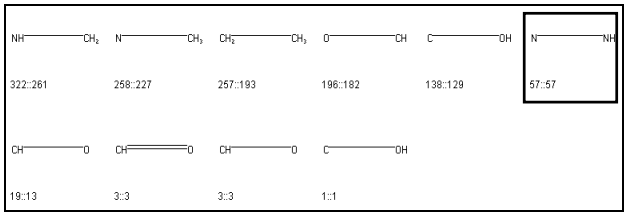


図 3 特徴抽出によって選ばれた質量数 29 のピークに含まれるフラグメント

このフラグメントのうち, 四角で囲んだフラグメントに注目する. 左下の数値 (57::57) は, このフラグメントが 57 回出現 (生成) し, 57 件の由来 (親) 構造を持つことを意味する. そこで, その 57 件の由来構造の出力を行ったところ, 図 4に示す構造が得られた. 視覚的にも類似した構造が表示されていることが分かる. これら 57 件の由来構造の活性クラスを調べたところ, 55 件の構造が D4 アンタゴニスト活性を持つことが分かった. すなわち, 図 1 の 6 番のフラグメントだけで 55 件の D4 アンタゴニストを説明することができ, クラス識別力を考慮した特徴抽出の効果が表れている一例であると考える.

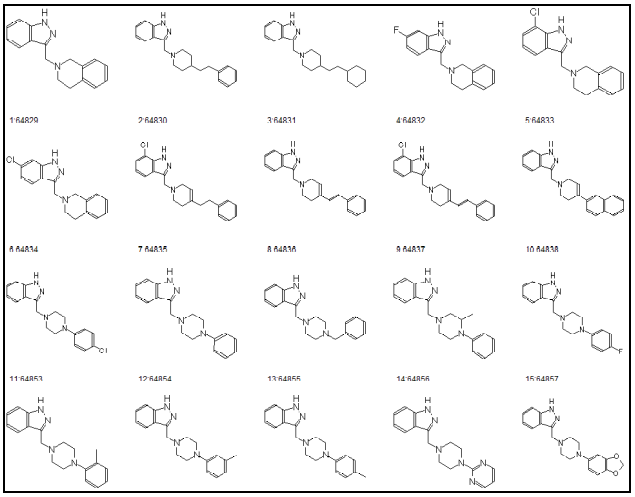

図 4 質量数 29 のピークに含まれる 6 番目フラグメントの由来構造 (一部)

また, 由来構造表示における, D4 アンタゴニスト活性を持つ 55 件の構造に注目すると, どれも類似した構造であり, 図 5に示すような 3 つの共通部分構造を見つけることができる. これらの構造は, 山川らによって報告されているドーパミンアンタゴニストの特徴的構造[Yamakawa 05]における, D4 アンタゴニストの特徴的な構造 (図 6) と一致し, 図 6の Ar に Indazole が結合していることと同じになることが分かる.

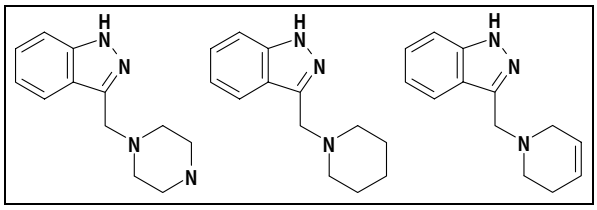

図 5 由来構造の共通フラグメント

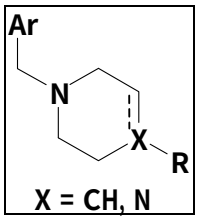


図 6 山川ら[Yamakawa 05]によって報告された D4 アンタゴニスト活性に特徴的なフラグメントの一つ

このように, 特徴抽出によって選択されたピークをもとに, 活性に特徴的なフラグメントおよび由来構造を具体的に解析することができた. 化学構造データファイルから生成される TFS ピークに含まれるフラグメントには, 活性が異なる数百件の由来構造が存在するものもある. そのようなフラグメントが多い中, TFS ピークの特徴抽出によって得られたピークの, 具体的意味として, 同活性を持つ数十件の構造を説明できる, 1 つのフラグメントを同定することができた. また, ピーク同定によって得られたフラグメントを一つのきっかけに, その由来構造から特徴的な共通フラグメントを見つけることができたことも注目に値すると考える.

## 4. まとめ

TFS ピークの特徴抽出においては, ピーク出現頻度を利用し, クラス毎のピーク出現頻度から, 各クラスへの寄与度が高いピークの抽出を試みた. 得られた TFS を使用した SVM による活性クラス分類においては, 特徴抽出前の TFS を使用した結果と比較しても, ほぼ同等の学習･予測能力を示すことができ, ピーク出現頻度での特徴抽出の有用性を示した.

また, 特徴抽出において得られた TFS ピークに対して, ピーク同定による構造特徴解析を行った. その結果, 得られた TFS ピーク内のフラグメントにおいて, 同活性の構造のみ (または大半) が由来構造であるフラグメントをいくつか発見することができた. これらの結果は, TFS ピークの特徴抽出が効果的に行われたことを表し, ピーク同定システムの併用によって, 薬物構造データマイニングにおける知識発見に, 本手法が有効であることを示すものと考える.

## 参考文献

[Dong 02] J. X. Dong, A. Krzyzak, and C. Y. Suen, A Fast SVM Training Algorithm, *International workshop on Pattern Recognition with Support Vector Machines. S.-W. Lee and A. Verri (Eds.)*: Springer Lecture Notes in Computer Science LNCS 2388, pp.53-67, Niagara Falls, Canada, Au-gust 10 (2002).

[Fujishima 04] S. Fujishima, Y. Takahashi : Classification of Pharmacological Activity of Drugs Using TFS-Based Neural Network, *J. Chem. Inf. Comput. Sci.*, **44**, 1006-1009 (2004).

[藤島 03] 藤島悟志, 高橋由雅 : “TFS を用いた化学構造データマイニング”, 第 17 回人工知能学会全国大会論文集, 2F2-03 (2003).

[藤島 04a] 藤島悟志, 高橋由雅 : “化学構造データマイニングのための TFS ピーク同定システムの開発”, *J. Comput. Chem. Jpn.*, **3**, 49-58 (2004).

[藤島 04b] 藤島悟志, 錦織克美, 加藤博明, 岡田孝, 高橋由雅 : “ノイズデータを含むドーパミン受容体アゴニスト/アンタゴニストの活性クラス分類”, 人工知能学会 第 64 回知識ベースシステム研究会, pp.125-128 (2004).

[藤島 04c] 藤島悟志, 高橋由雅 : “TFS 差スペクトルによる薬物構造データマイニング”, 第 18 回人工知能学会全国大会論文集, 1F2-03 (2004).

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

[錦織 03] 錦織克己, 高橋由雅 : “薬物活性クラス分類へのサポートベクターマシン（SVM）の応用”, 第 17 回人工知能学会全国大会論文集 (2003).

[Takahashi 98] Y. Takahashi, H. Ohoka, and Y. Ishiyama: Structural Similarity Analysis Based on Topological Fragment Spectra, *In: R. Carbo and P. Mezey (Eds), Advances in Molecular Similarity* 2, pp.93-104, JAI Press, Stamford CT, (1998).

[Takahashi 02] Y. Takahashi, S. Fujishima and K. Yokoe: Chemical Data Mining Based on Structural Similarity, Proceedings of International Workshop on Active Mining, IEEE ICDM 2002, 132-135 (2002).

[Takahashi 03] Y. Takahashi, S. Fujishima, H. Kato, Chemical Data Mining Based on Structural Similarity, *J. Comput. Chem. Jpn.*, **2**, 119-126 (2003).

[高橋 03] 高橋由雅, 藤島悟志, 横江恭子 : “TFS を利用した薬物活性クラス分類とリスクレポート”, 電子情報通信学会「人工知能と知識処理」研究会・情報処理学会「知能と複雑系」研究会・人工知能学会「人工知能基礎論」研究会・人工知能学会「知識ベースシステム」研究会, 「アクティブマイニング合同研究会」, (2003).

[Yamakawa 05] 山川眞透, 岡田孝 : “ドーパミン・アンタゴニストの特徴的構造について”, 宝塚ワークショップ：アクティブマイニングによる化学構造からの知識発見, 26-32 (2005).