2F2-03

TFS を用いた化学構造データマイニング

Chemical Structure Data Mining using Topological Fragment Spectra (TFS)

藤島 悟志　　高橋 由雅
Satoshi Fujishima　　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

We have developed a peak identification system for topological fragment spectra (TFS). The TFS is a kind of quantitative representation of a chemical structure, which is based on enumeration and numerical characterization of all the possible substructures from a chemical structure. We have investigated the applicability of the TFS approach to chemical data mining with practical chemical data. The system can pursuit the meaning of each peak of TFS and identify the substructures behind the peaks. Usefulness of the system has been validated in Quantitative Structure-Toxicity Analysis with the TFS descriptors.

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.

現在, 筆者らは先に提案した化学構造情報の数値的記述表現のための手法の一つである Topological Fragment Spectra (TFS)法 [Takahashi 98] の化学データマイニングへの応用について様々な検討を進めている. これまで, この TFS を構造類似性評価や構造活性相関研究に利用してきたが, 得られた結果に対する解釈に際して, 個々の TFS ピークの意味を追求できなかった. このことから, 別途, TFS の個々ピークと生成フラグメント(部分構造)との関係を自動的に解析提示することを目的とした TFS ピーク同定システムを開発した[藤島 01].

本研究では, TFS を利用した構造活性相関解析と化学構造データマイニングにおける同システムの有用性について, 実データを用いて例示する.

2. 方法

2.1 TFS による構造特徴の数値的記述表現

TFS とは, 図 1 に示すように, 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, 構造情報を多次元数値ベクトルとして取り扱うことができる. また, TFS は部分構造の特徴付け(次数和, 質量数和, etc)を工夫することにより, 様々な特性スペクトルを生成することができる.

2.2 TFS ピーク同定システム

TFS の各ピークに対応する部分構造の同定を行うことができれば, その構造活性相関解析等への応用に際し, 部分構造情報にもとづくより詳細な考察も可能となると同時に構造データマイニングへの有力なツールとしても期待できる. このことから, 先に筆者らは TFS ピークの自動同定システムの開発を行った. システムの処理の流れを図 2 に示す.

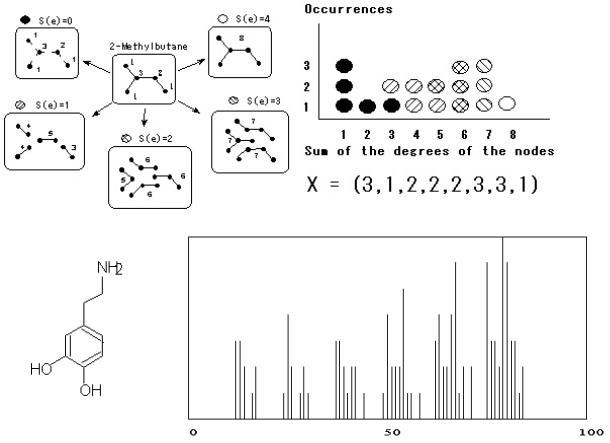

図 1　TFS の生成手順と生成例

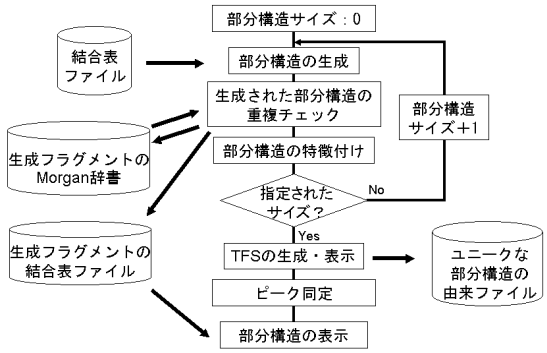

図 2　TFS ピーク同定システムの流れ

対象とする化学構造に対して, 部分構造を順に生成していく. ピーク同定の際にはユニークな部分構造を取り扱うため, 重複した部分構造を排除する必要がある. この重複チェックに, Morgan の規範化アルゴリズム [Morgan 65] を使用した. 生成された部分構造の規範化によって得られた Morgan 名を辞書に登録する. その後生成された部分構造に対して辞書内の Morgan 名との重複チェックを行う. 重複していなければ Morgan 辞書に登録する. 指定された結合サイズまでの部分構造を順次

連絡先：高橋由雅, 〒441-8580 愛知県豊橋市天伯町雲雀ヶ丘 1-1 豊橋技術科学大学 知識情報工学系, Tel: 0532-44-6878, taka@mis.tutkie.tut.ac.jp

生成しながらこれらの操作を行う. このようにして作成された部分構造の Moragan 辞書をもとに, システムは, 生成された TFS の各ピークに対して対応する生成フラグメント(部分構造)の同定を行う. すなわち, 任意のピークに対して, それに含まれている部分構造を表示する.

2.3 構造活性相関への適用

ここでは TFS ピーク同定システムの有用性を実証するため, 実データによる定量的構造活性相関 QSAR(Quantitative Structure-Activity Relationships) への適用を試みた.

田村ら[田村 00]は, 遺伝的アルゴリズムによる関数近似 GFA(Genetic Function Approximation)の構造活性相関への応用の試みの中で, 種々の化学物質の構造と魚毒性に関する構造毒性相関のモデル化に対し, TFS を構造変数パラメータとして用いた良好なモデルを報告している. ここでは, この田村らのモデルに含まれる構造変数パラメータに対して, ピーク同定を行うことにより, 特定の活性を持つ構造群に対して, TFS からどのような情報が得られるかを検討した.

構造データには上記の GFA 解析で使用された, 魚毒性データが既知の 463 件を用いた. これらの化合物データセットはその化学性にもとづいて 3 つのクラス(非極性, 極性脂肪族, 極性芳香族)に分割することができる. ピーク同定のためのモデルは, 各クラスにおいて作成された適用度最良モデルを使用した.

3. 結果と考察

例として, クラス 1 の非極性化合物に対する解析結果を示す. クラス 1 の訓練集合 168 化合物に対する魚毒性データをもとに, TFS を利用した GFA 解析より得られた適用度最良モデルを図 3 に示す. ここで, y は毒性値を表し, [ ]とともに示される数値が構造パラメータとして使用された TFS ピーク変数である. また, n はサンプル数, LOF は適応度, r は相関係数, s は標準誤差を表す.

```
NO.1   y= + 0.727      + 0.831 * LogP -  0.411 * [36]   + 0.577 * [49]
           - 0.171 * [62] + 0.036 * [68]  + 0.027 * [76]   + 0.238 * [118]
n = 168  bfunc_num = 7  LOF = 0.535  r = 0.822  s = 0.670
```

図 3 TFS 記述子を用いた化学物質の構造毒性相関モデル

筆者らの開発した TFS ピーク同定システムを利用し, クラス 1 の構造(訓練集合 168 件)全てに対して TFS を生成し、データベースを作成するとともに(図 4), クラス1対する適応度最良モデルに使用された 6 種の TFS パラメータのピーク同定を行った. ピーク変数[118]に対するピーク同定によって得られた部分構造の一覧を図 5 に示す. 図中, 点線で表示されている結合は, 芳香族結合の一部分であることを示している. ここでは, 化学的な議論の詳細は避けるが, このモデルに大きな寄与を示す TFS ピークの意味を具体的な部構造情報として考察することができることは極めて大きな意義を持つ. また GFA による毒性近似モデルからはこれらの部分構造が毒性の発現に正(+)の寄与を有することが示唆される.

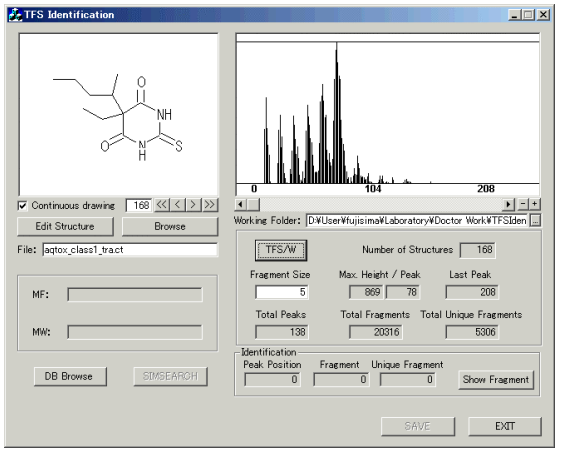

図 4 クラス 1 の訓練集合に対する TFS の生成

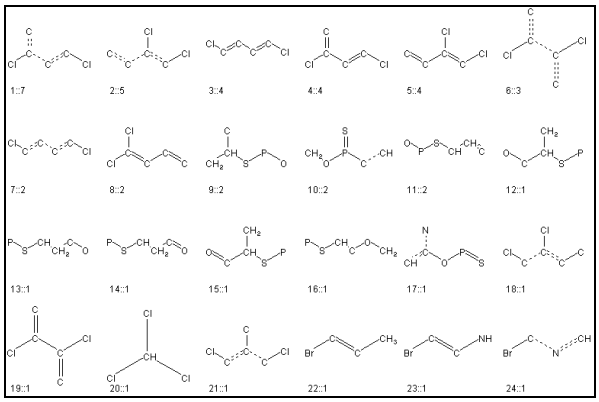

図 5 ピーク［118］に含まれるユニークな部分構造

一方, ピーク同定システムの機能の一つである, 構造データベース参照機能を利用することにより, ピーク内の指定した部分構造を含む全ての由来構造を検索することも可能である. これにより, 重要と思われるピーク内に複数の部分構造が存在する場合にもその由来構造を比較することにより詳細な知見を得ることができる. 発表に際してはこれらについても合わせて議論したい.

4. まとめ

ピーク同定システムを構造活性相関解析に適用することにより, 活性に重要と思われる具体的な部分構造を抽出することが可能となった. またそれら部分構造の由来構造を参照することにより, 活性に関与している構造特徴の詳細な比較が容易に行うことができる.

今後は, 活性クラスの異なる 2 つの化合物群の TFS の差などに注目し, このことによって得られる情報についても合わせて検討していきたい.

参考文献

[藤島 01] 藤島悟志, 高橋由雅: 第 24 回情報化学討論会講演要旨集, pp.57-58 (2001).

[Morgan 65] Morgan H.L.: The Generation of a Unique Machine Description for Chemical Structures, *J. Chem. Doc.* 5, pp.107-113 (1965).

[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).

[田村 00] 田村広志, 高橋由雅: 第 28 回構造活性相関シンポジウム講演要旨集, pp.242-245 (2000).