

# 化学物質の構造類似性にもとづくデータマイニング

高橋 由雅\*, 藤島 悟志, 加藤 博明

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

\*e-mail: taka@mis.tutkie.tut.ac.jp

(Received: April 17, 2003; Accepted for publication: June 30, 2003; Published on Web: August 29, 2003)

本研究では筆者らが先に提案したトポロジカルフラグメントスペクトル ( Topological Fragment Spectra, TFS ) による新たな構造情報の記述手法をもとに, 特定の部分構造の有無のみならず化学構造全体の漠然とした類似性をも考慮したよりやわらかな構造情報の取り扱いとその化学データマイニングへの応用について検討した. 構造類似性評価における TFS 法の有用性と薬物構造データベースを対象としたデータマイニングへの応用の可能性を検証するため, World Drug Index ( WDI ) より抽出した 3,637 件の薬物構造データベースを対象に, TFS データベースを作成した. これを利用し, 神経伝達物質として知られるドーパミンを Query とした類似構造検索を試みたところ, 化学者の直感に矛盾しない良好な結果を得ることができた. また, 上記の検索結果をもとに得られた構造類似化合物についてのドーパミン活性の有無を調べてみたところ, 複数の化合物が同活性を有するものであることが明らかとなった. このことは構造類似性にもとづく候補構造発見のためのデータマイニングの有用性を示唆するものであると考える.

キーワード: データマイニング, 構造類似性, ドラッグデザイン, TFS, 化学データベース

## 1 はじめに

「AはBに似ている」,あるいは「CはDと××が似ている」といったいわゆる“類似性”の概念は科学における様々な問題解決の場で利用される極めて重要な概念の一つである. このことは化学の分野においても例外ではなく, その対象となる分子の間の類似性がしばしば言及される [1, 2]. 特に, これまでの明示的な部分構造マッチングとは別に, “似た構造”あるいは“似た反応”といったいわゆる化学的な“類似性”の概念を如何に取り扱うかは, 関連分野におけるコンピュータのより高度な利用を図る上で極めて重要な問題の一つであり, こうした分子の類似性の概念に基礎を置く, より柔らかな構造情報処理に向けた新たな技術の確立が望まれている.

ところで, “分子の類似性”と言っても, その対象は様々である. 例えば化学構造の類似性のほかに, 融点・沸点などの物理的な特性に注目した類似性, さ

らには生物学的な作用に関する類似性など, その取り扱い問題によって以下のように様々な視点がある.

- (1) 生物学的活性に関する類似性.
- (2) 物性に関する類似性.
- (3) 化学反応に関する類似性.
- (4) 分光スペクトルの類似性.

また, 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明らかであり, その関係は次のように表すことができる.

$$\text{Molecular Property} = f(\text{chemical structure}) \quad (1)$$

このことは, 化学構造が変われば物質の性質もこれに依じて変わるとの考えを示したものにほかならない. 言い替えれば, 化学物質の種々の性質についてその類似性を比較・検討することは, その起因となる化学構造の類似性を解析することと等価な問題と見なすことができる.

化学物質の構造類似性を評価するにはまず初めに、個々の分子の構造特徴を調べる必要がある。いわゆる構造特徴解析である。その代表的なものには部分構造検索の技法 [3] を基礎とした官能基解析 [4] や部分構造解析 [5]、構造中の ring system に注目した環解析 [6] などがある。これらは特定の官能基や環構造の有無あるいは数を調べ、個々の分子の構造特徴を記述するのに用いられる。そしてこれらの部分構造特徴のうち比較対象とする化合物間に共通に含まれるものを調べ、その構造間の類似性を比較・評価するための様々な試みが報告されている [7, 8]。一方、これとは別に、解析の対象とする化合物間の最も大きな部分構造を探索する MCS (Maximum Common Substructure) の問題もまた構造類似性解析と密接に結び付けられる [9]。現在、主にドラッグデザインの分野を中心に、そのシステム化に向けた様々な試みの中でこうした共通構造特徴解析の問題はトポロジカルあるいは 2 次元構造式レベルのみならず 3 次元構造レベルでの解析も行われるようになってきた [10, 11]。そしてこれらの明示的な共通構造特徴解析の考え方はそこでの要素技術である 2 次元ならびに 3 次元部分構造検索と共に、いわゆる“形状の類似性”などに注目したより柔らかな構造情報処理の問題へと引き継がれている。

本研究は化学構造全体から見た漠然とした“類似性”も含め、これらの定量的な評価の方法並びにそのシステム化に必要な要素技術の確立と化学データマイニングへの積極的な活用を目指すものである。分子構造の類似性評価に関する研究は新薬開発やこれに関連した類似化学物質の検索あるいは分子の特性予測問題に関連して現在活発に研究が進められている分野である。先に述べたように、これまでこれらの研究では、トポロジカル、2 次元、3 次元を問わず、化合物間の構造類似性を評価するための構造情報記述子としては予め注目される特定の部分構造特徴(官能基)が用いられてきた。しかしながら、こうしたアプローチではその類似性評価の結果は構造特徴を記述するために事前に定義された部分構造集合に大きく依存することが避けられない。そこで、本研究では筆者らが先に提案したトポロジカルフラグメントスペクトル (Topological Fragment Spectra, TFS) [12] による新たな構造情報の記述手法を利用した構造類似性の定量的評価のための方法をもとに、特定の部分構造の有無のみならず化学構造全体の漠然とした類似性をも考慮したより柔ら

な構造情報の取り扱いとその化学データマイニングへの応用に向け、その有用性を実用規模のデータベースを用いて検討した。

## 2 方法

### 2.1 TFS による構造特徴の定量的記述表現

TFS と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多次元数値ベクトルとして表現しようとするものである。その生成手順は、(1) 構造情報の記述表現に際しては化学構造式の原子を点(頂点)、結合を辺と見なし、原子や結合の種類の違いを区別する重み付きグラフ(化学グラフ)として取り扱う。ただし、水素原子はすべて省略する。(2) 与えられた構造式に対応する化学グラフから可能なすべての部分グラフを列挙する。ここでは、親グラフ(もとの化学グラフ)中の異なる要素からなる部分グラフについては同型のものも全て考慮した。(3) 次に、得られた個々の部分グラフの定量的特徴づけを行う。これらの特徴づけには様々な方法が考えられる。たとえば、与えられた親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えば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す特性スペクトルを得ることができる。また、生成された部分グラフを各部分グラフ中の頂点に対応する原子の質量数の総和(フラグメント重量)によって特徴づければ、原子の種類を考慮した構造フラグメントに関する特性スペクトルを得ることも可能である。

本研究では生成フラグメントの特徴づけに対しては後者のフラグメント重量を用いた。結合水素原子を有するものについてはそれを含めた拡張原子として各頂点の重みづけを行った。このようにして特徴づけられた個々の部分グラフ(構造フラグメント)集合をもとに、その特徴指数に従って度数を調べ、その結果をヒストグラム表示したものがここでの TFS となる。これら、TFS 生成手順の概要を Figure 1 に示す。TFS は一種のデジタルスペクトルと見なすことができる。これはまた多次元数値パターンベクトルとして取り扱うことができ、そのパターン間の類似性評価に対しては、種々の類似度関数の適用が可能となる。

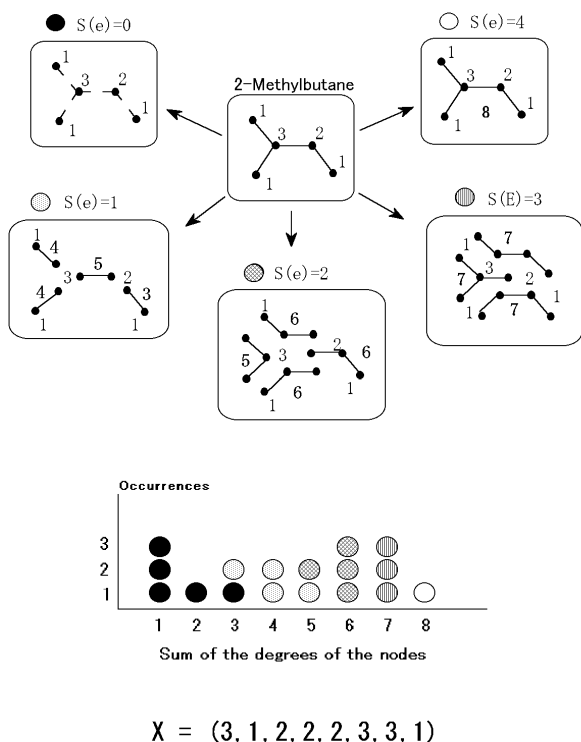


Figure 1. An illustrative scheme of the procedure of TFS generation.

## 2.2 類似度評価関数

化学物質の構造類似性の評価に際しては、「構造情報をどのように表現するか」がきわめて重要となることは言うまでもない。この他、もう一つ、その評価結果を大きく左右するものに評価関数を挙げることができる。すなわち、定量的に記述表現された構造情報の数値記述子を「どのような関数を用いて評価すればより適切な結果を得ることができるのか?」、という問題である。

ここでは、このような視点から 5 種の異なる類似度 (あるいは相違度) 関数を対象とし、ここでの TFS 法を利用した構造類似性評価における各々の関数の適否並びに相互の結果の比較検討を行った。類似度の評価には様々な評価尺度の利用が考えられる。本研究では、評価尺度としてユークリッド距離 ( $S_{ED}$ )、Tanimoto 係数 (実数データ) ( $T_C$ )、Tanimoto 係数 (バイナリデータ) ( $T_B$ )、Cosine 係数 ( $S_C$ )、ピアソンの積率相関係数 ( $S_P$ ) の五つの相違度または類似度関数を利用可能とした。ここで  $x_{ik}$ ,  $x_{jk}$  はそれぞれ化合物  $i$  およ

び化合物  $j$  についての  $k$  番目の記述子の値を表す。

(a) ユークリッド距離 ( $S_{ED}$ ):

$$S_{ED} = \sqrt{\sum (x_{ik} - x_{jk})^2} \quad (2)$$

(b) Tanimoto 係数 ( $T_C$ ):

$$T_C = \frac{\sum (x_{ik}x_{jk})}{\sum x_{ik}^2 + \sum x_{jk}^2 - \sum (x_{ik}x_{jk})} \quad (3)$$

(c) Tanimoto 係数 (二値データ) ( $T_B$ ):

$$T_B = \frac{C}{Q + D - C} \quad (4)$$

(d) Cosine 係数 ( $S_C$ ):

$$S_C = \frac{\sum (x_{ik}x_{jk})}{\sqrt{\sum (x_{ik})^2} \sqrt{\sum (x_{jk})^2}} \quad (5)$$

(e) Pearson's 相関係数 ( $S_P$ ):

$$S_P = \frac{\sum (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum (x_{ik} - \bar{x}_i)^2} \sqrt{\sum (x_{jk} - \bar{x}_j)^2}} \quad (6)$$

尚、式 (4) においては各 TFS をピークの有無にしたがって '1' または '0' の値をとる 2 値スペクトルに変換している。ここでは、 $Q$  は Query 構造の TFS 中で値が '1' の要素の数、 $D$  はデータベース中の比較する構造の TFS 中に含まれる値が '1' の TFS 要素の数、 $C$  は Query 構造およびデータベース構造の両者ともに値 '1' をもつ要素の数を表す。

## 2.3 データセット

本研究では Darwent 社より市販されている World Drug Index (WDI) より、ドーパミン活性を有する化合物 117 件を抽出し、その他ランダムに抽出した化合物を加えることによって 3,637 件からなる薬物構造データベースを作成し、テストデータベースとして用いた。

## 3 結果及び考察

上記の考えをもとに、化学物質の構造類似性評価における TFS 法の有用性と薬物構造データベースを対象としたデータマイニングへの応用の可能性を検証するため、上述の World Drug Index より抽出した 3,637 件の薬物構造データベースを対象に、これら全ての化合物に対する TFS を生成、データベースを作成し計算機実験を行った。著者らは先に、TFS を利用した類似

構造検索において、生成フラグメントの完全列挙により得られる全スペクトルと生成するフラグメントのサイズを制限することによって得られる部分スペクトルとの比較の中で、両者が良好な相関を有することを示した。このことから、本実験では TFS の生成に際してはサイズ（生成部分構造に含まれる結合の数）が 5 までのフラグメントにもとづく TFS を生成、利用した。ここで用いたサイズ '5' は、ベンゼン環上の *para* 位に位置する原子間のパスが直接記述可能であることを意図したものである。

作成した TFS データベースをもとに、ドーパミンを

Query とした類似性検索を試みた。ドーパミンの化学構造とそのサイズ 5 までの生成フラグメントにもとづく TFS Query パターンを Figure 2 に示す。はじめに、類似性の評価関数として単純ユークリッド距離を用いて検索を試みた。本研究では、後述の種々の評価関数による検索結果の比較を容易にするため、全ての検索実験において、類似度の高い（距離の小さい）ものから上位 20 件を検索した。ユークリッド距離を用いて検索されたこれら 20 化合物の構造式一覧を Figure 3 に示す。

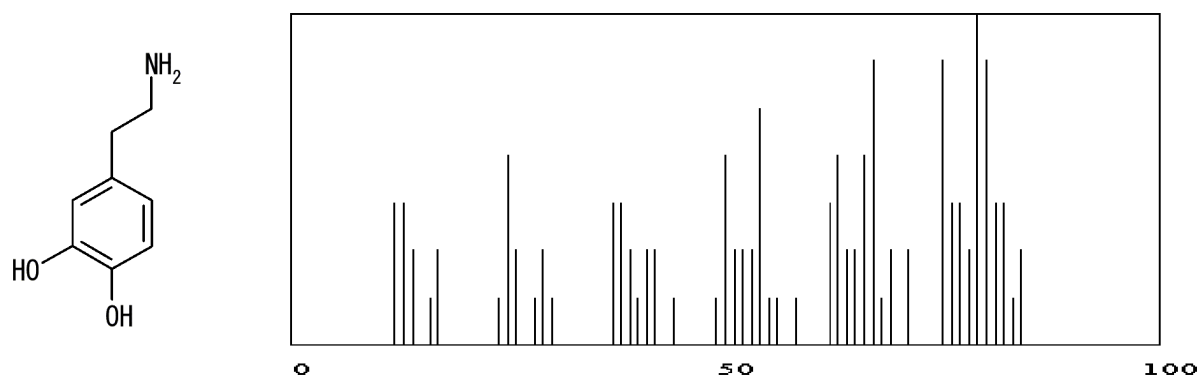


Figure 2. TFS of dopamine that was characterized by the sum of atomic mass numbers.

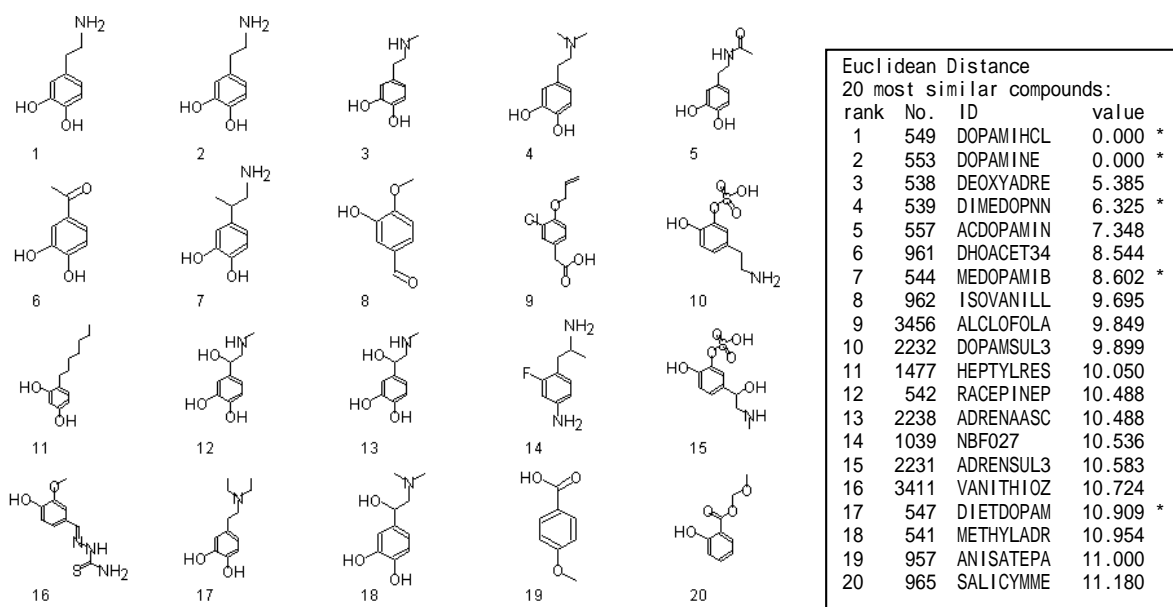


Figure 3. Result of structural similarity searching for a query of dopamine by TFS method. (Twenty most similar compounds found by Euclidean distance). The asterisk in the table shows that the compound is dopaminergic active.

Table 1. Results of similar structure search for dopamine by different similarity functions.

Tanimoto Coefficient				Tanimoto Coefficient (Binary)				Cosine Coefficient				Correlation Coefficient			
20 most similar compounds:				20 most similar compounds:				20 most similar compounds:				20 most similar compounds:			
rank	No.	ID	value	rank	No.	ID	value	rank	No.	ID	value	rank	No.	ID	value
1	549	DOPAMHCL	1.000 *	1	549	DOPAMHCL	1.000 *	1	549	DOPAMHCL	1.000 *	1	549	DOPAMHCL	1.000 *
2	553	DOPAMINE	1.000 *	2	553	DOPAMINE	1.000 *	2	553	DOPAMINE	1.000 *	2	553	DOPAMINE	1.000 *
3	538	DEOXYADRE	0.934	3	554	OXIDOPAMI	0.958	3	539	DIMEDOPNN	0.968 *	3	539	DIMEDOPNN	0.946 *
4	539	DIMEDOPNN	0.919 *	4	2074	HOTESTO6B	0.900	4	538	DEOXYADRE	0.967	4	538	DEOXYADRE	0.944
5	557	ACDOPAMIN	0.890	5	148	NAIN	0.885	5	544	MEDOPAMIB	0.962 *	5	544	MEDOPAMIB	0.942 *
6	544	MEDOPAMIB	0.872 *	6	893	STRONGYL2	0.885	6	2787	SKF89615	0.951 *	6	2787	SKF89615	0.920 *
7	961	DHOACET34	0.855	7	1316	DOMOPREDN	0.885	7	557	ACDOPAMIN	0.951	7	542	RACEPINEP	0.919
8	542	RACEPINEP	0.826	8	3264	NO500	0.885 *	8	542	RACEPINEP	0.948	8	2238	ADRENAASC	0.919
9	2238	ADRENAASC	0.826	9	3265	NAXAGOHCL	0.885 *	9	2238	ADRENAASC	0.948	9	557	ACDOPAMIN	0.916
10	962	ISOVANILL	0.818	10	91	PD129289	0.882	10	541	METHYLADR	0.944	10	541	METHYLADR	0.906
11	1477	HEPTYLRES	0.818	11	1319	RIMEXOLON	0.882	11	396	FEPENTOLA	0.939	11	2788	SKF89626	0.900 *
12	2231	ADRENSUL3	0.818	12	1660	BREFONALO	0.882	12	2231	ADRENSUL3	0.936	12	396	FEPENTOLA	0.899
13	3456	ALCLOFOLA	0.815	13	2982	RS82856	0.882	13	2788	SKF89626	0.936 *	13	2231	ADRENSUL3	0.891
14	541	METHYLADR	0.814	14	538	DEOXYADRE	0.878	14	961	DHOACET34	0.931	14	961	DHOACET34	0.889
15	3411	VANITHIOZ	0.804	15	3478	PIRIBEMES	0.878	15	3536	E4101	0.921 *	15	3536	E4101	0.878 *
16	2232	DOPAMSUL3	0.802	16	301	TINCTORAM	0.868	16	148	NAIN	0.920	16	2825	VESTITOL	0.875
17	547	DIETDOPAM	0.799 *	17	1310	BETULINAT	0.868	17	2825	VESTITOL	0.920	17	944	BUCAFFEAT	0.869
18	2915	HOMEMEXIL	0.773	18	1315	DEPRONDP	0.868	18	1477	HEPTYLRES	0.919	18	943	ETCAFFEAT	0.867
19	1039	NBFO27	0.769	19	1509	KATONATE	0.868	19	3411	VANITHIOZ	0.919	19	17	ERBSTATIN	0.865
20	728	LEVODOPA	0.760 *	20	1994	MORDANBR1	0.868	20	944	BUCAFFEAT	0.916	20	1477	HEPTYLRES	0.863

\* Asterisk shows that the compound is dopaminergic active.

Table 2. The rate of structures shared by the searching results with different measurement functions.

	$S_{ED}$	$T_B$	$T_C$	$S_C$	$S_P$
$S_{ED}$	1.00	0.15	0.90	0.65	0.60
$T_B$		1.00	0.15	0.20	0.15
$T_C$			1.00	0.65	0.55
$S_C$				1.00	0.80
$S_P$					1.00

Figure 3 から明らかのように構造的によく似ているものが上位にランク付けされている。また、ここでは Query であるドーパミン自身がデータベース中に異なる商品名で 2 件含まれており、これらの二つが距離ゼロで最初に検索されていることがわかる。

次に、TFS 法における類似性評価に際しての種々の類似度関数の適否について検討を行った。評価尺度としてユークリッド距離 ( $S_{ED}$ ) のほか、方法の部で述べた Tanimoto 係数 (実数データ) ( $T_C$ )、Tanimoto 係数 (バイナリ・データ) ( $T_B$ )、Cosine 係数 ( $S_C$ )、ピアソンの積率相関係数 ( $S_P$ ) について検討を行った。これらの検索実験の結果をまとめて Table 1 に示す。ここで、Table 1 の Rank は各評価尺度での類似度の順位を示しており、value はそれぞれの評価関数の値を示す。類似性評価尺度としての評価関数の値を相互に直

接比較することはできないが、各検索結果の 20 位に位置する構造の類似度を比較すると、用いた 4 つの尺度のうちでは Tanimoto 係数 ( $T_C$ ) が最も厳しい評価値を与えていることがわかる。

次に、これらの結果をもとに、評価関数の違いによる検索結果の相違を共通にヒットした化合物数の割合 (重なり率) によって比較した。これらの結果をまとめて Table 2 に示す。

Table 2 より、ユークリッド距離と Tanimoto 係数 ( $T_C$ ) を用いた場合の結果に注目すると、重なり率は 0.9 となり上位 20 化合物のうち 18 化合物が両者に共通であることが分かる。また、Figure 3 と Table 1 の検索結果の順位リストをみると、 $T_C$  の類似度が 0.8 以上の類似性の高い構造 16 件についてはその全てがユークリッド距離による検索の上位 20 件の中にヒットしていることがわかる。このことは TFS を基礎とした構造類似性検索においてはこれらの評価関数はほぼ同様な結果を与えることを示唆している。比較のため、Figure 4 に Tanimoto 係数 ( $T_C$ ) による検索結果の構造式一覧を示す。また、これとは別に、cosine 係数とピアソンの相関係数の間においても重なり 20 件中 16 件 (重なり率 0.8) の構造が共通にヒットしており、その順位リスト (Table 1) をからも相互に高い相関を有することがわかる。それぞれの評価関数によって検索された構造式一覧を Figure 5, Figure 6 に示す。

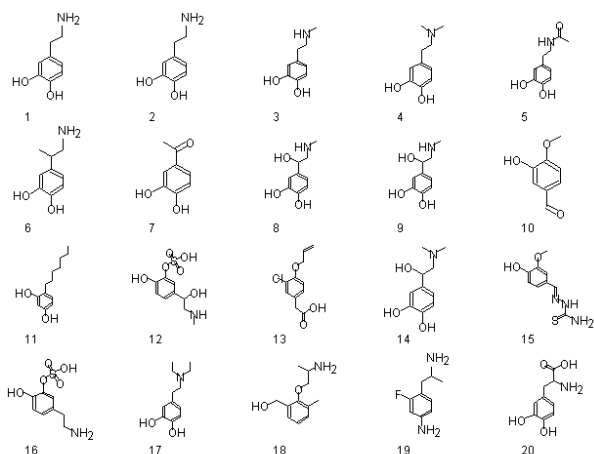


Figure 4. Twenty most similar structures obtained by Tanimoto coefficient ( $T_C$ ).

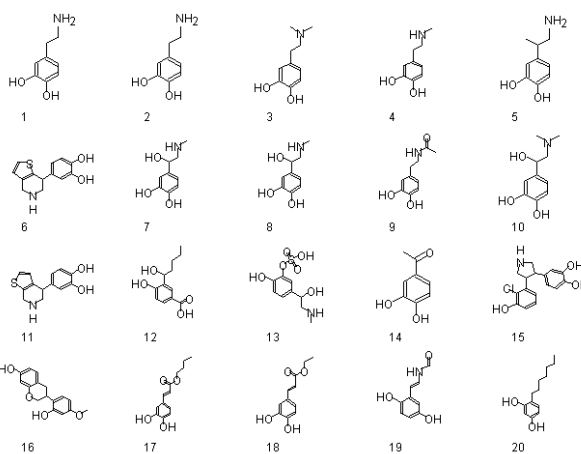


Figure 6. Twenty most similar structures obtained by correlation coefficient (binary,  $S_P$ ).

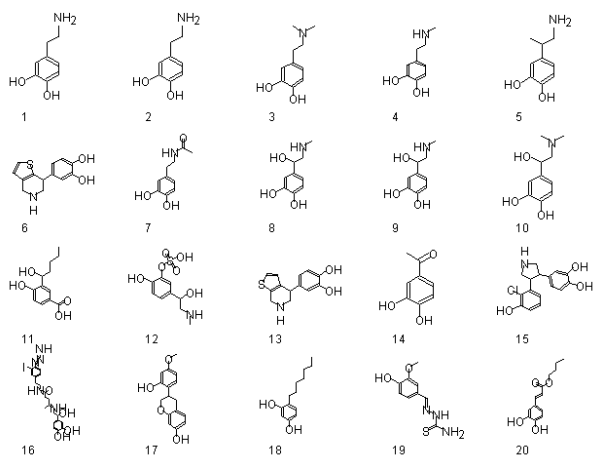


Figure 5. Twenty most similar structures obtained by cosine coefficient ( $S_C$ ).

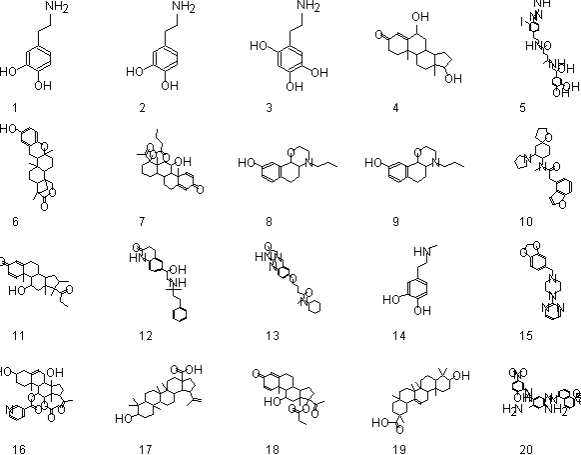


Figure 7. Twenty most similar structures obtained by Tanimoto coefficient (binary,  $T_B$ ).

一方、TFSの2値表現を基礎としたTanimoto係数( $T_B$ )は他の類似度関数を用いた場合に比べ、大きく異なる検索結果を与えていることが分かる(Table 2)。このことはFigure 7に示した検索結果の構造式一覧からも明らかである。すなわち、上位20件での検索の結果を比較すると、ユークリッド距離、Tanimoto係数、cosine係数、ピアソンの相関係数に対し、それぞれヒットした化合物の重なり率は、0.15、0.15、0.20、0.15であり、検索結果の重なりは極めて小さな値を示している。これは、データのバイナリ化を基礎としたTanimoto係数( $T_B$ )が、TFSの類似度評価に際してスペクトルの強度(頻度)情報を利用していないことに

よるものと考えられる。

以上のことから、本研究で使用した上記TFSにもとづく構造類似性評価のための5種類の評価関数は(1)ユークリッド距離、Tanimoto係数、(2)cosine係数、Pearsonの相関係数、(3)Tanimoto係数(2値データ)の3つの異なるグループに大別できる。ここで、(1)はTFSのピーク強度を直接考慮した方法であり、他の方法に比べて分子のサイズに関する情報が反映されたものとみなすことができる。これに対して(2)はTFSのピーク強度そのものよりはスペクトルパターンの相対的な変化を考慮したものであり、検索実験の結果からも(1)に比べてサイズ依存性が低いことが分かる。ま

た，(3)はその定義から明らかなようにピークの強度情報は利用していないことから，特に本実験で使用した部分スペクトルによる類似性検索においてはサイズ依存性のない方法とみなすことができる．

これらのことは類似性検索の結果にもよく反映されている．ここでの構造類似性検索実験に Query として用いたドーパミンは神経伝達物質としての作用を示すことが知られている．上記の検索結果をもとに得られた構造類似化合物についてのドーパミン様活性の有無を調べてみたところ，それぞれ Query のドーパミン以外に複数の化合物が同活性を有するものであることが明らかとなった．これらの構造については Figure 3 の順位リストおよび Table 1 に \* 印を付して示した．本研究で使用したデータセット ( 3,637 件 ) 中には重複も含め 117 件のドーパミン様活性化合物が含まれている．しかしながら，この中には構造的に大きく異なるアポモルフィンの誘導体や類縁体が多数含まれており，ドーパミンの直接的な誘導体と考えられる単環系の活性化化合物は全部で 10 件 ( 9 種 ) であった．このうち，(1)の二つの方法 ( ユークリッド距離，Tanimoto 係数 ) ではそれぞれ 5 件 ( 4 種 ) ，6 件 ( 5 種 ) の構造をヒットしている．一方，(2)，(3)の方法では互いに (1)とは異なる活性化化合物の構造もヒットしている．また，(3)の Tanimoto 係数 ( 2 値データ ) においては，サイズ的にも多様な構造をヒットしていることが分かり興味深い．これらの結果は TFS を利用した構造類似性にもとづくデータマイニングの有用性を示唆するとともに，ここでの類似度評価のための評価関数はその特質によって 3 つの異なるグループに分けられ，これらを併用することによって視点の異なる構造データマイニングが可能であることを示している．

## 4 おわりに

以上，本研究では実用規模の薬物データベースを用いた類似構造検索の実験を通じて，TFS 表現にもとづく構造類似性評価を基礎とした化学データマイニングの有用性を示した．また，あわせて筆者らの提案する TFS 法を用いた構造類似性評価における類似度関数相互の関係を解析するとともに，その適否を検討した．しかしながら，TFS を利用した構造類似性解析においてはどの類似度関数が最適であるかを結論づけることは困難であり，特に構造データマイニングの視点からは特質の異なる複数の類似度 ( 相違度 ) 関数を目的に

応じて利用することが重要と思われる．

尚，本研究の一部は文部科学省科学研究費補助金 ( 特定領域研究 B 2 ) ( 課題番号 : 13131210 ) 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．

## 参考文献

- [1] Johnson M. A., Maggiora G. M., *Concepts and Applications of Molecular Similarity*, Wiley, New York (1990).
- [2] Carbo R., *Molecular Similarity and Reactivity*, Kluwer Academic Publishers, Boston (1995).
- [3] Sussenguth, Jr. E. H., A Graph Theoretic Algorithm for Matching Chemical Structures, *J. Chem. Doc.*, **5**, 36-43 (1965).
- [4] Cramer III, R.D., Redl G., Berkoff C.E., Substructural Analysis: Novel Approach to the Problem of Drug Design, *J. Med. Chem.*, **17**, 533 (1974).
- [5] Adamson G.W., Bush J.A., Evaluation of an Empirical Structure- Activity Relationship for Property Prediction in a Structurally Diverse Group of Local Anaesthetics, *J. Chem. Soc. Perkin Trans. I*, 168-172 (1976).
- [6] Downs G.M., Gillet V.J., Holiday J.D., Lynch M.F., Review of Ring Perception Algorithms for Chemical Graphs, *J. Chem. Inf. Comput. Sci.*, **29**, 172 (1989).
- [7] Stanton D.T., Morris T.W., Roychoudhury S., Parker C.N., Application of Nearest-Neighbor and Cluster Analysis in Pharmaceutical Lead Discovery, *J. Chem. Inf. Comput. Sci.*, **39**, 21-27 (1999).
- [8] Bajorath J., Selected concepts and Investigations in compounds classification, molecular descriptor analysis, and virtual screening, *J. Chem. Inf. Comput. Sci.*, **41**, 233-245 (2001).
- [9] Takahashi Y., Identification of structural similarity of organic molecules, *Topics Curr. Chem.*, **174**, 105-133 (1995).
- [10] Thorner D.A., Wild D.J., Willett P., Wright P.M., Similarity Searching in Files of Three-Dimensional

- Chemical Structures: Flexible Field-Based Searching of Molecular Electrostatic Potentials, *J. Chem. Inf. Comput. Sci.*, **36**, 900 (1996).
- [11] Rarey M., Stahl M., Similarity searching in large combinatorial chemistry spaces, *J. Computer-Aided Mol. Des.*, **15**, 497-520 (2001).
- [12] Takahashi Y., Ohoka H., Ishiyama Y., Structural Similarity Analysis Based on Topological Fragment Spectra, *Advances in Molecular Similarity*, **2**, 93-104 (1998).

## Chemical Data Mining Based on Structural Similarity

Yoshimasa TAKAHASHI\*, Satoshi FUJISHIMA and Hiroaki KATO

Dept. of Knowledge-based Information Engineering, Toyohashi University of Technology  
1-1 Hibarigaoka, Tenpaku-cho, Toyohashi, Aichi 441-8580, Japan  
*\*e-mail: taka@mis.tutkie.tut.ac.jp*

This paper describes an approach to chemical data mining based on the quantitative evaluation of structural similarity. The topological fragment spectrum (TFS) method reported by the authors was used for describing a chemical structure by numerical representation. The TFS is based on enumeration and numerical characterization of all possible substructures derived from the chemical structures. The TFS was applied to similar structure searching with over 3,600 drugs extracted from the World Drug Index. All the spectra were characterized for fragments having five or less bonds. Five different similarity (or dissimilarity) functions were investigated for their suitability for similarity searching with the TFS. Computational trial of similar structure searching on the database suggested that the present approach is successfully applicable to chemical and pharmaceutical data mining based on the evaluation of structural similarity of drug molecules.

**Keywords:** Data mining, Structural similarity, Drug design, TFS, Chemical database