

Excel/VBA を利用した分子類似性指数計算のための 教材プログラム開発とダイオキシン異性体への応用

大前 貴之

山口県立大学 生活科学部 生活環境学科, 〒 753-8502 山口市桜島 3 - 2 - 1

e-mail: ohmae@yamaguchi-pu.ac.jp

(Received: June 29, 2005; Accepted for publication: September 1, 2005; Published on Web: November 21, 2005)

化学を専門としない学生が分子構造の類似性に関する概念を容易に学ぶことができるようにするために、Excel/VBA を利用したプログラムを作成した。この教材用プログラムによって、分子構造の類似性を数値化して表現できる Pearson 係数と Tanimoto 係数の計算が可能になった。作成したプログラムを用いて、2,3,7,8-ポリクロロジベンゾ-*p*-ジオキシン（以下、2,3,7,8-PCDD のように略記する。）に対するダイオキシン異性体の計算を行った。計算された類似性指数と TEF の間に良い相関のあることが示唆される結果が得られた。

キーワード：類似性指数, Pearson 係数, Tanimoto 係数, 分解能, ダイオキシン

1 はじめに

酵素反応の反応機構に見られるように、化学反応において分子の幾何学的な形が重要な役割を果たす場合のあることが、広く知られている [1]。このことから、化学をひとつの手段として生命に関わる現象を解明しようとする栄養学や環境学などを専攻する人々にとって、分子の幾何学的な形の類似性に関する概念を修得することは必要不可欠であると考えられる。しかしながら筆者の知る限り、分子の幾何学的な形の類似性に関する章立てをした化学の教科書は見あたらず、この概念は経験を通じて体得するものであるかのような扱いが従来なされてきた。

この状況は、日常的に分子構造をながめる時間を持つことのできる化学を専門とする者にとっては、さほど痛痒を感じるものではない。実際、例えば Figure 1 を化学専攻の人々に示して I の構造に対する a, b, c 各構造の類似性の高低を問えば、その豊富な経験からほとんどの人々が躊躇なく $a > b > c$ と答えるものと考えられる。しかしながら、限られた時間の中で手段である化学を学ぶ人々に同じ質問をすると、その回答

は可能な 6 種類の順列にほぼ均等に分布してしまうのが現実なのである。

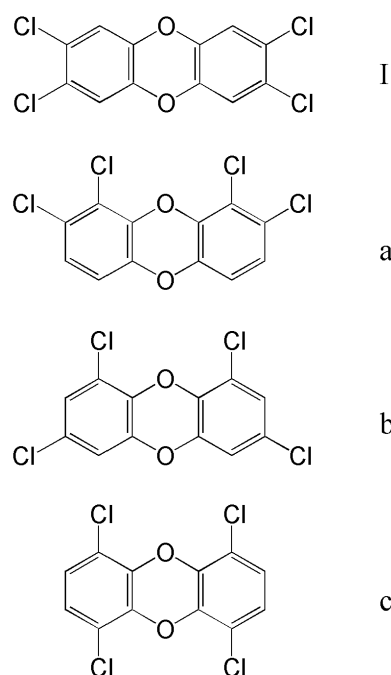


Figure 1. Molecular Framework of Dioxin Isomers.

筆者はこの状況を改善する必要に迫られ, Excel/VBA を利用した自作プログラムを数年前から授業と学生の自習のために用いて, 分子構造の類似性に関する概念を環境学専攻の学生に修得させる試みを実践してきた。幸い学生の多くが所有するパーソナル・コンピュータには標準的に Excel が搭載されていたこともあり, この試みは短期間で上述の回答の分布を $a > b > c$ へ収斂させる効果があった。そこで筆者と同様に, 手段としての化学を学ぶ学生に分子構造の類似性に関する概念を修得してほしいと考える方々の参考に供するために, 本報で上述のプログラムとダイオキシン異性体の計算例を紹介する。

2 類似性指数計算プログラム

定量的構造活性相関の研究に利用する目的で, さまざまな分子の類似性を数値化する方法が提案されているが [2, 3], 本報では距離対の比較に基づいて定義される次の 2 つの係数に注目した [4, 5]。これは分子構造の幾何学的な形の類似性を考察する上で, 距離に着目することが非化学専攻の学生にとって直感的に最も理解しやすいと考えたためである。

Pearson Correlation Coefficient :

$$r_P \equiv \frac{\sum(d_{A_{ij}} * d_{B_{ij}})}{\sqrt{\sum(d_{A_{ij}})^2 * \sum(d_{B_{ij}})^2}}, \quad (1)$$

Tanimoto Coefficient :

$$r_T \equiv \frac{N_C}{N_A + N_B + N_C}, \quad (2)$$

ただし, 式 (1) で $d_{A_{ij}}$ は分子 A の原子 ij 間の距離を意味する。また式 (2) で N_A と N_B はそれぞれ分子 A と分子 B の距離対の数を意味し, N_C は次式で定義される d の大きさが基準値を超えない距離対の数を意味する。なお, 次式で計算される距離の差の基準値としては, $1 \times 10^{-11} \text{m}$ から $1 \times 10^{-10} \text{m}$ の間の適当な値が用いられる。

$$\Delta d = d_{A_{ij}} - d_{B_{ij}}. \quad (3)$$

式 (1), (2) から明らかなように, r_P と r_T は両者ともに, その値が 0 の時は比較している分子間に類似性がまったく存在しないことを意味し, その値が 1 の場合には比較している分子間の構造が完全に一致していることを意味する。

2.1 r_P 計算プログラム

教材として提供する際の学生の利便性を考慮して, データの入出力を Excel のワークシートから行えるよう, プログラムを作成した [6]。これは, 日常のレポート作成などを通じて学生が Excel の操作法に習熟していると考えられることと, 新規な操作画面を用いさせることによって生じると予測される学習意欲の低下とその操作法に習熟するまでの時間の無駄を省くためである。

Figure 2 に作成したデータ入力部分のリストを示した。ただし, Figure 2 で RX, RY, RZ はそれぞれ基準分子の各原子の 3 次元直交座標を示す。また, RD は原子 i, j 間の距離を表す。

本プログラムでは, Figure 2 に示したプログラムによって読み込んだ基準分子と比較分子の 3 次元直交座標から原子間距離が計算され, その後, 式 (1) に従って r_P の値が計算されるという手順を採った。なお, 筆者の担当する授業では本プログラムを使用する前に行う MOPAC の演習で得られた 3 次元直交座標を csv データとして Excel の入力シートに読み込んで利用してい

```

1600 Dim N As Integer, I As Integer, J As Integer
1700 '基準分子構造の計算
1800 Dim RX() As Double, RY() As Double, RZ() As
      Double, RD() As Double
1900 Sheets("基準分子").Select
2000 Range("A1").Select
2100 Range("G1:Z100").ClearContents
2200 Cells(1, 8) = "RD"
2300 N = Cells(2, 2)
2400 ReDim RX(N - 1), RY(N - 1), RZ(N - 1),
      RD(N - 1, N - 1)
2500 For I = 0 To N - 1
2600   RX(I) = Cells(4 + I, 3)
2700   RY(I) = Cells(4 + I, 4)
2800   RZ(I) = Cells(4 + I, 5)
2900 Next I
3000 For I = 0 To N - 2
3100   For J = I + 1 To N - 1
3200     RD(I, J) = (((RX(J) - RX(I)) ^ 2)
      + ((RY(J) - RY(I)) ^ 2)
      + ((RZ(J) - RZ(I)) ^ 2)) ^ 0.5
3300     Cells(1, 10 + I) = I + 1
3400     Cells(2 + I, 9) = I + 2
3500     Cells(1 + J, 10 + I) = RD(I, J)
3600   Next J
3700 Next I

```

Figure 2. List of data input part in this program.

るが，MOPAC の演習と切り離して本プログラムを利用する場合も想定して，現在，3 次元直交座標の簡便な生成プログラムの作成を検討中である．

上述のリストで示したプログラムに続けて，式 (1) に従って r_P の値を計算する部分のリストを Figure 3 に示した．ただし，Figure 3 で比較分子の原子 i, j 間の距離を表す D を計算する部分は，Figure 2 に示したリストの各変数名から R を取り除くだけで作成できるので，ここではリストを示さなかった．また，Figure 3 の $D2, RD2, DRD$ はそれぞれ式 (1) の $d_{A_{ij}}^2, d_{B_{ij}}^2, d_{A_{ij}} * d_{B_{ij}}$ を示す．

```

5800 'rP の計算
5900 Dim D2 As Double, RD2 As Double,
      DRD As Double, INDEX As Double
6000 Sheets("比較分子").Select
6100 Range("A1").Select
6200 Cells(1, 4).ClearContents
6300 Cells(1, 5).ClearContents
6400 Cells(1, 4) = "INDEX : "
6500 D2 = 0
6600 RD2 = 0
6700 For I = 0 To N - 2
6800     For J = I + 1 To N - 1
6900         D2 = D2 + D(I, J) * D(I, J)
7000         RD2 = RD2 + RD(I, J) * RD(I, J)
7100         DRD = DRD + D(I, J) * RD(I, J)
7200     Next J
7300 Next I
7400 INDEX = DRD / (((D2 * RD2)) ^ 0.5)
7500 Cells(1, 5) = INDEX

```

Figure 3. List of part calculating Pearson correlation coefficient in this program.

	A	B	C	D	E
1	分子名:	1469-PCDD		INDEX:	0.9968
2	原子数:	22			
3	No.	Atom	x	y	z
4	1	C	0.0000	0.0000	0.0000
5	2	O	1.3862	0.0000	0.0000
6			1.9504	1.2663	0.0000
7			1.1850	2.4451	-0.0009
8			-0.2013	2.4451	-0.0020

Figure 4. An output page of this program.

```

5800 'rT の計算
5900 Dim NA As Integer, NB As Integer
6000 Dim SVL As Single,
      SVU As Single, SVP As Single
6100 Dim NC(10) As Integer
6200 Dim INDEX(10) As Double
6300 Dim RNAME As String, NAME As String
6400 Sheets("基準分子").Select
6500 Range("A1").Select
6600 RNAME = Cells(1, 2)
6700 SVL = Cells(1, 4)
6800 SVU = Cells(1, 6)
6900 SVP = (SVU - SVL) / 10
7000 For II = 0 To 10
7100 NC(II) = 0
7200 INDEX(II) = 0
7300 For I = 0 To N - 2
7400     For J = I + 1 To N - 1
7500         If Abs(D(I, J) - RD(I, J)) <= (SVL + SVP * II)
7600             Then GoTo 7600 Else GoTo 7700
7700             NC(II) = NC(II) + 1: GoTo 7800
7800         Next J
7900 Next I
8000 NA = N * (N - 1) / 2
8100 NB = NA
8200 INDEX(II) = NC(II) / (NA + NB - NC(II))
8300 Next II
8400 Sheets("比較分子").Select
8500 Range("A1").Select
8600 NAME = Cells(1, 2)
8700 Sheets("分析").Select
8800 Range("A1").Select
8900 Range("A1:Z100").ClearContents
9000 Cells(1, 1) = "基準分子 : "
9100 Cells(1, 2) = RNAME
9200 Cells(1, 3) = "比較分子 : "
9300 Cells(1, 4) = NAME
9400 Cells(2, 1) = "基準値"
9500 Cells(2, 2) = "INDEX"
9600 Cells(1, 8) = "D-RD"
9700 For II = 0 To 10
9800     Cells(II + 3, 1) = SVL + SVP * II
9900     Cells(II + 3, 2) = INDEX(II)
10000 Next II
10100 For I = 0 To N - 2
10200     For J = I + 1 To N - 1
10300         Cells(1, 10 + I) = I + 1
10400         Cells(2 + I, 9) = I + 2
10500         Cells(1 + J, 10 + I) = D(I, J) - RD(I, J)
10600     Next J
10700 Next I

```

Figure 5. List of part calculating Tanimoto coefficient in this program.

さらに、以上のプログラムを用いて、2,3,7,8-PCDD に対する 1,4,6,9-PCDD の r_P 値を求めた際の出力画面を Figure 4 に示した。今回紹介する試みにおいては、複数の比較分子の r_P を同時に計算するようにプログラムを作成していないので、Figure 4 に示したように r_P の計算値は比較分子の座標を入力したワークシートの右上隅に出力させた。

2.2 r_T 計算プログラム

r_T の計算プログラムにおいても、3次元直交座標から原子間距離を計算する部分として、Figure 2 に示したプログラムを用いた。この部分に続けて、式 (2) から r_T の値を計算するために、Figure 5 に示した r_T 計算部のプログラムを作成した。ただし、Figure 5 で SVL 、 SVU 、 SVP はそれぞれ基準値の下限、上限、刻み幅を表す。また、 NA 、 NB 、 NC は式 (2) の N_A 、 N_B 、 N_C を表す。

a. Input page

	A	B	C	D	E	F
1	分子名:	2378-PCDD	SVL	0.0000	SVU	1.0000
2	原子数:	22				
3	No.	Atom	x	y	z	
4	1	C	0.0000	0.0000	0.0000	
5	2	C	0.0000	0.0000	0.0000	
6	3	C	1.2667	0.0000	0.0000	
7	4	C	2.4473	-0.0024	0.0000	
8	5	O	-0.1998	2.4473	-0.0050	

b. Output page

	A	B	C	D	E	F	G
1	基準分子:	2378-PCDD	比較分子:	1469-PCDD			
2	基準値	INDEX					
3	0.0000	0.0000					
4	0.1000	0.2798					
5	0.2000	0.2941					
6	0.3000	0.3391					
7	0.4000	0.4043					
8	0.5000	0.4760					
9	0.6000	0.8331					
10	0.7000	0.8331					
11	0.8000	0.8331					
12	0.9000	0.8331					
13	1.0000	0.8660					

Figure 6. An input page and an output page of this program.

ここでは後に触れるように類似性における分解能類似概念の導入のため、Figure 6a に示した入力画面に距離対の類似性を判定する基準値の上限と下限を入力しておくことで、この区間を 10 等分したそれぞれの基準値で r_T の値が計算されるようにした。これによって、 r_T の出力範囲をあらかじめ適当なグラフのデータに指定しておけば、Figure 6b に示した出力画面のように基準値の変化による類似性判定の変化を観察することができる。

また、このグラフを示した上で、Figure 7 の様な図を例示して多角形と円の類似性を考察させることで、類似性判定における分解能類似概念を学生に触発することができた。

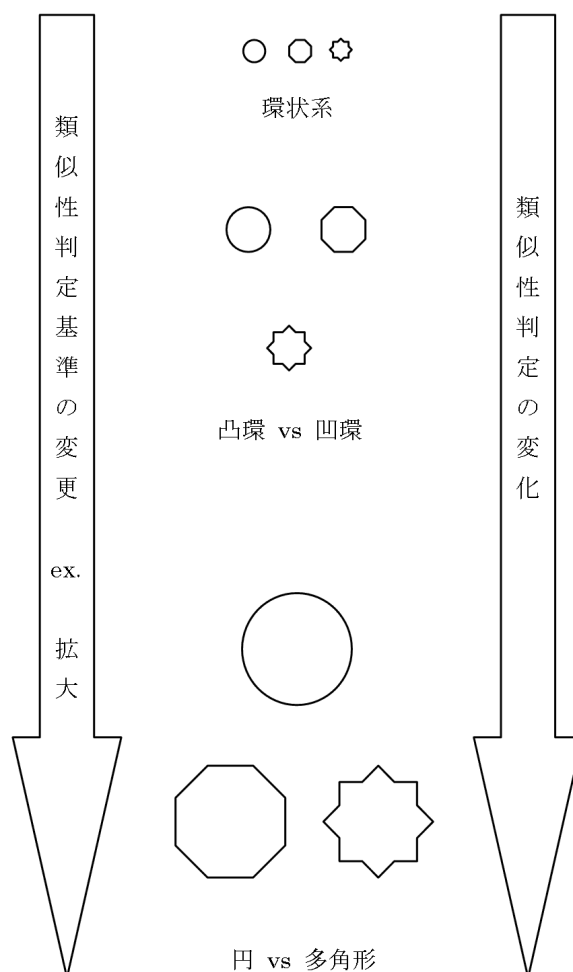


Figure 7. Diagram to remind students of concept of resolving power.

Table 1. Calculated values of r_P and r_T of some Dioxin isomers.

Chlorination	TEF	r_P	$r_T (t=0.5)$
2,3,7,8-	1	1.0000	1.0000
1,2,3,7,8-	1	0.9997	0.9412
1,2,3,4,7,8-	0.1	0.9994	0.9012
1,2,3,6,7,8-	0.1	0.9994	0.8780
1,2,3,7,8,9-	0.1	0.9994	0.8857
1,2,3,4,6,7,8-	0.01	0.9991	0.8333
1,2,3,4,6,7,8,9-	0.0001	0.9989	0.7838

t : Reference value, $\times 10^{-10}$ m.

3 ダイオキシン異性体への適用例

前節までに紹介したプログラムを用いて, 2,3,7,8-PCDD に対するダイオキシン異性体の r_P と r_T の値を計算した. 計算したすべての異性体の中から, TEF の値 [7] が既知のもの r_P と r_T の値を Table 1 に示した. TEF の相対的な順位と r_P および r_T の値の間に良い相関のあることは Table 1 から明らかである.

ただし, r_P の値の降順にすべての異性体を配列すると, 例えば TEF の値が 0.01 と 0.001 の間には TEF の値が未知の異性体が 6 種類入ることからも明らかのように, ひとつの類似性指数だけで化合物群の活性を説明することは困難であり, 小林ら [8] はダイオキシン類の電子構造に基づく考察からハードネスの概念が特に重要であることを指摘している.

また, r_T の値は Figure 6 にも示したように基準値によって変化し, 場合によっては値の大小関係が逆転することもあるため, 活性相関の検討に用いる際には注意を要すると考えられる. Table 1 に示したダイオキシン異性体の r_T の値が, 基準値の選び方によって変化する様子を Table 2 に示した.

なお, Table 2 のような基準値に依存する類似性指数の値の変化と塩素置換数の等しい異性体の構造式, 例えば 6 置換体の構造式を比較することで, 分子構造の類似性が一意的なものではなく区別する基準によって変化するものであることを啓発することができる.

4 おわりに

類似性指数を計算する実用的プログラムという観点から見ると, 本報で紹介したプログラムにはいくつかの改善すべき点が存在する. 例えば, 本文中でも触れ

Table 2. A change of calculated value of r_T with a change of reference value.

Chlorination	$r_T(t=0.1)$	$r_T(t=0.3)$	$r_T(t=0.5)$
2,3,7,8-	1.0000	1.0000	1.0000
1,2,3,7,8-	0.8406	0.8704	0.9412
1,2,3,4,7,8-	0.6739	0.7769	0.9012
1,2,3,6,7,8-	0.7111	0.7634	0.8780
1,2,3,7,8,9-	0.7048	0.7567	0.8857
1,2,3,4,6,7,8-	0.5986	0.6739	0.8333
1,2,3,4,6,7,8,9-	0.5049	0.6098	0.7838

t : Reference value, $\times 10^{-10}$ m.

た複数の比較分子の類似性指数が同時に計算できないことなどはその最たるもののひとつであろう. しかしながら, この短所を教材プログラムという観点から見ると, ひとつひとつの比較分子の類似性指数を計算する度に, 自分が今なにをしているのかを確認することができるという意味で, むしろ短所は長所になり得るものと考えられる. プログラムの入力と出力の部分に数行を書き加えれば容易に改善できることばかりではあるが, 上述の観点からいくつかの短所は気づいていながらそのままにしておいた.

また, 本文中で述べた類似性における分解能の概念については, 今後さらに多くの類似性指数に関する計算を行った後に, 考察を深める予定である.

なお, 本研究の一部は山口県立大学研究創作活動助成事業によるものであることを, ここに記して感謝する.

参考文献

- [1] 例えば, 上代淑人監訳, イラストレイテッドハーパー・生化学, 丸善 (2003), 51-62, など.
- [2] R. C.-Dorca, X. Girones, P. G. Mezey, *Fundamentals of Molecular Similarity*, Kluwer Academic / Plenum Publishers, New York (2001).
- [3] P. M. Dean, *Molecular Similarity in Drug Design*, Blackie Academic & Professional, London (1995).
- [4] P. Willett, V. Winterman, *Quant. Struct. -Act. Relat.*, **5**, 18 (1986).
- [5] 中馬 寛, 唐沢真美, 佐々木幹夫, 長嶋雲兵, *J. Chem. Software*, **4**, 143 (1998).

- [6] 吉村忠与志, 化学とソフトウェア, **23**, 11 (2001).
[7] 相澤寛史, ダイオキシン類の基礎知識, 東京教育
情報センター (2001), 151.

- [8] 小林茂樹, 田中 彰, 鮫島圭一郎, 化学と工業, **52**,
42 (1999).

Program Development to Calculate Similarity Index of Molecules for Excel/VBA and Application to Dioxin Isomers

Takayuki OHMAE

Department of Environmental Science, Yamaguchi Prefectural University
3-2-1 Sakurabatake, Yamaguchi, Yamaguchi 753-8502, Japan
e-mail: ohmae@yamaguchi-pu.ac.jp

A computer program which used Excel/VBA was made so that the student who did not major in chemistry could learn concepts about similarity of molecular structure easily. By this program, it is possible to calculate Pearson correlation coefficient (r_P) and Tanimoto coefficient (r_T). From the point of view of education, both programs were done in a convenient way which plural comparison molecule cannot handle. We kept the function which changed reference value into the program which calculated r_T . The possibility that there was a concept of resolving power in similarity was suggested by the results of calculation with this function. We calculated r_P and r_T of Dioxin isomers for 2,3,7,8-PCDD with this program. It was suggested that there was good correlation between calculated indexes and TEF.

Keywords: Similarity index, Pearson coefficient, Tanimoto coefficient, Resolving power, Dioxin isomer