

報 文

スペクトル位置の重心及び標準偏差値による NMR スペクトルのファイル検索

佐々木 慎一* , 四居 靖彦** , 落合 周吉***

(1974 年 7 月 15 日受理)

有機化合物構造同定の一手段として NMR スペクトルファイル検索システムを試作した。スペクトルの規格化は 1 枚のスペクトルを各シグナルにプロトンの重みを持たせたときのスペクトル全体の重心 (G) とその点からの各シグナルへの分布状態を表す標準偏差 (S) で代表させることによって行った。なるべく広範囲に集めた 232 化合物のスペクトルの G, S 値を格納してファイルを作った後、それら格納物を一つ一つ未知試料に見立ててファイル検索を行った結果、検体の (72~89) % が正解のみ、残りは正解 (+1~3) 個の不正解を出力したので、この方法が構造同定に十分な検索精度を有することが分かった。

1 緒 言

有機化合物構造同定の一つの手段として検体スペクトルを既測定スペクトルと照合する方法があげられるが、照合すべきスペクトルの数は近時、ますます膨大なものとなつてきているため、検索能率向上の点からスペクトル類による構造同定を計算機で行う方法が講ぜられつつあることは周知のとおりである。この際スペクトル全体を忠実にしまい込むより、むしろ構造同定であれ何であれその利用目的を損なわぬ範囲でスペクトルを規格化してファイル化するほうが一般的である。スペクトルの規格化とファイル検索は赤外スペクトル¹⁾、質量スペクトル²⁾についてはよく研究されており、両者とも構造同定のための情報検索として実用化の段階に入っている。著者らは有機化合物構造決定の自動化研究の一部として、NMR スペクトルのファイル検索システムを試作した。スペクトルの規格化にあたってはファイル検索の結果正解を落としてはならぬことを重要事とするが、その際に副生する正答以外の答 (ノイズ) の量をなるべく少なくし、実用上十分な検索精度を保たせること及びシステムの的になるべく簡潔なものにすることを目標として実

験を行った。以下にその大要を記す。

2 NMR スペクトルの規格化

2.1 規格化の方法

著者らによる NMR スペクトルの規格化は当初全領域を一定の基準を設けて切断し、各部分につき化学シフト、強度、水素数などを記録することによって行われてきた³⁾が、若干の難点もあるので今回は

(1) スペクトルの特徴をよりよく反映し、そのスペクトルに関する固有値もしくはそれに近い値を引き出すこと。

(2) 1 スペクトルあたりの記憶量をできるだけ少なくすること。

(3) 測定条件 (機種、分解能など) の影響をなるべく受けぬようにすること。

などを目標にして規格化を工夫した。その結果、1 枚のスペクトルを各シグナルにプロトン数の重みを持たせたときのスペクトル全体の重心 (G) と、その点からの各シグナルへの分布状態を表す標準偏差 (S) の 2 種類の値をもって代表させることにした。本報では以下 G, S という略号を用いるが、それらはそれぞれ (1), (2) で求められる。

$$G = \frac{\sum_{i=1}^n (\omega_i A_i)}{\sum_{i=1}^n A_i} \quad \dots\dots\dots (1)$$

$$S = \left(\frac{\sum_{i=1}^n (\omega_i - G)^2 A_i}{\sum_{i=1}^n A_i} \right)^{1/2} \quad \dots\dots\dots (2)$$

* 宮城教育大学：宮城県仙台市荒巻字青葉

** 現在 第一製菓株式会社中央研究所：東京都江戸川区南船堀町 2810

*** 日本電子株式会社：東京都昭島市中神町 1418

ここに A_i : i 番目のシグナルのプロトンに関する重み

ω_i : i 番目のシグナルの位置 (化学シフト TMS 基準)

G と S の組み合わせは原スペクトル固有の値と違って差し支えないこと、G, S はともに小数点以下 2 けたをとれば検索精度は十分であり、これに要する記憶容量は (1~2) words で済むこと、更に G はスペクトルの分解能が漸次小さくなり最終的に 1 本のシグナルにまとまってしまったとしたときのそのシグナルの化学シフトと考えることができ、測定機器の分解能には関係しないといえることなどから、この節のはじめに述べたファイルシステム作成の目標をほぼ満たすものと考えられる。

2.2 規格化の実際

データの規格化は NMR スペクトルのデータ集を材料にして行われる。このとき各シグナルの高さは直ちにプロトン数に比例するものとして扱い、またあるシグナル群にプロトンの数を明らかに割り当てることができるスペクトルについては、まずシグナル群ごとに割り当てた上で、更にそれを構成するシグナルに対してそれぞれの高さに応じてプロトンを配分した。プロトン割り当ての困難なシグナル群に対しては全シグナルの高さの総和 ($\sum_{i=1}^n A_i$) に対する各シグナルの高さ (A_i) の比で割り当てを行った。式 (1) から分かるように、プロトンの総和の分かっていること (換言すれば分子式があらかじめ分かっていること) や、シグナル群にプロトン数を正しく割り当てることが可能か否かなどは、いずれもこの検索システムのためのデータ規格化において必ずの条件とはならない。

データ集からスペクトルを読み取るときの基準は次のように定めた。

(1) 他から独立しているシグナル群があって、それに対して明確にプロトンを割り当てうるときはそれを行う。次にこのシグナル群を構成する各シグナルの高さの総和 ($\sum_{i=1}^k a_i$) に対する各シグナルの高さ (a_i) の割合をもとに各シグナルに対してプロトンの重みの配分を行う。独立したシグナル群を見いだすことのできないとき、もしくはできてもそれに明らかな数のプロトンを割り当てられぬときは、スペクトル全体を 1 群とみなす。群に分ける分けないにかかわらずプロトンの絶対数が分かっている必要はない。

(2) 1 本のシグナルについては (3) の場合を除きピーク頂点の位置と高さを読み取る。

(3) ≤ 3 Hz の範囲で近接し、しかも高さに著しい違いのない (2~3) 本のシグナルについては、それらの中心に 1 本のシグナルがあるものとみなして位置を決める。高さはそれらの総和とする。ただし、先端が極めてわずかに割れているようなシグナルは 1 本とみなす。

(4) 肩は原則として読まない。

(5) 広い吸収幅をもつシグナルの場合はそのシグナルの頭にシグナルがあるものとみなす。また ≤ 3 Hz の範囲で多数のシグナルが接近しているときは、中央のシグナルの位置から 5 Hz ごとに高さを読み、それぞれその位置に独立したシグナルがあるものとみなす。

サンプリングはこのような基準に基づいて行われるが、エチルベンゼン (1) の場合を例示すれば、そのスペクトルを読み取ってそれを位置、高さ、シグナル数、プロトン数などの項目別に表示すると Table 1 のようになる。すなわち Table 1 には 8 本のシグナルの位置 (TMS を基準とした Hz)、強度 (高さ) 及び各シグナルもしくは、シグナル群に割り当てられたプロトン数、1 群を構成するシグナルの本数が示されている。この数値からまず各シグナルの持つプロトンの重み (A_i) が計算され、ついで式 (1), (2) に従って $G=4.43$, $S=2.70$ がそれぞれ求められ、最終的には ppm 単位に変換されて出力される (Fig. 1)。

Table 1 Nmr spectrum of ethylbenzene (1) (digitalized)

Signal No.	Chem. shift (Hz)	Height	No. of proton	No. of signal
1	66	86	3	3
2	72	219		
3	81	125		
4	145	36	2	4
5	152	86		
6	160	73		
7	167	22	5	1
8	425	522		

$G=4.43$, $S=2.70$

2.3 検索の基準

スペクトル検索において検索の精度に影響を及ぼす要因として次の 2 点が考えられる。

(i) 測定条件の違いにより同一化合物が異なるスペクトルを示す。

(ii) スペクトルから規格化データ (本法では G, S 値) を求めるときの読み取り個人差。

(i) については、更に測定機器の違いに起因するものとその他の測定条件 (測定温度、溶媒) の違いに起因するものの二つに区別されよう。前者は理想的には消去可能な要因であり、また本法はこの影響を受けにくいも

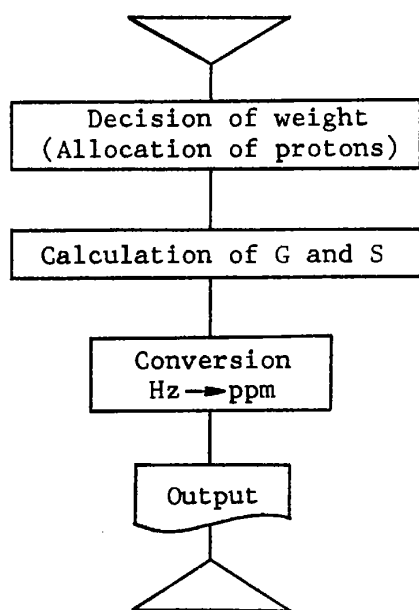


Fig. 1 Flowdiagram of calculation of G and S values

のであることはすでに記した。しかし後者, すなわち測定温度, 溶媒などの差は同一化合物に対して異なるスペクトルの異同を検索するものであるという本法の基本に照らして同一化合物でも測定条件の違いにより, 著しく異なるスペクトルを与える場合にはファイル中にも複数の規格化データを格納しておくという立場をとっている。このことは各種スペクトルデータ集をデータ源として採用するときにも同じことがいえる。

(ii) の規格化データ作成についてはスペクトル読取装置から on line で G, S 値の得られることが望ましいが, 現段階では人力にたよっている。従ってこの段階で発生する G, S 値の読み取りのばらつきが本検索法の精度に大きく影響するものと思われる。そこでスペクトルの形が複雑であるなどの理由で規格化データが求めにくく, かつそれに相当して個人差の発生しやすいと予想されるスペクトルについて同一スペクトルから複数の人間が G, S 値を求めたときのばらつき程度を調べその値をスペクトル検索の際の有意差判定の基準として用いることとした。これを検討するために 52 の化合物のスペクトルを選び各々 2 回ずつ規格化してみたところ, 1 回目と 2 回目の差の最も大きかったものは 2-methyl-3-heptanol (2) と 3-methyl-4-heptanol (3) であった。そこでこれら 2 化合物を, 規格化データを求めにくい例に見立てて更に詳しく検討するため, (2), (3) のスペクトルについて各々複数の人間が 10 回ずつ G, S を求めたところ, Table 2 に示す結果が得られた。例え

Table 2 Calculation of G and S values of 2-methyl-3-heptanol (2) and 3-methyl-4-heptanol (3)

	Obsd. value	Average	Range	Standard deviation (σ)
2 {G	1.26, 1.27×7, 1.28×2	1.271	0.02	0.005
{S	0.56, 0.57×5, 0.58×4	0.573	0.02	0.006
3 {G	1.27×2, 1.28×8	1.278	0.01	0.004
{S	0.59×2, 0.60×7, 0.61×1	0.599	0.02	0.005

ば, 化合物 (2) のスペクトルは G が 1.26(ppm) は 1 回, 1.27 が 7 回, 1.28 が 2 回と計算されたのでその平均値は 1.271, 分布の範囲 (R) は 0.02, 標準偏差 (σ) は 0.005 となる。また (2) の S については $\sigma=0.006$ である。(3) に関しても Table 2 に示すとおりである。そこで一般に標準偏差は G については $\sigma_G=0.005$, S については $\sigma_S=0.006$ を採用することにした。今 99.7% 確率の範囲を求めるとすると 3σ を有意差の基準とすれば十分に思えるが, 現実には規格化は (1~2) 回のサンプリング回数で行われるのが普通であるから, その分の危険性を考慮に入れて 6σ とればよろしかろうと判断した。すなわち今化合物 X と Y の G, S をそれぞれ G_x , S_x ; G_y , S_y としたとき

$$|G_y - G_x| \leq 0.03$$

$$|S_y - S_x| \leq 0.04$$

を同時に満足させるとき X と Y は同一物の可能性があるものと判定することとした。

3 実験と結果

3.1 スペクトルの収集

化合物の種類, 化合物型になるべく変化を持たせるよう配慮しつつ 232 のスペクトルを集めた (Table 3)。

Table 3 Scope of samples taken

Alkanes	5
Alkyls	41
Alkenes	20
Acetylenes	2
Total	68
Benzenes	
(mono-substd.)	23
(di-substd.)	53
(tri-substd.)	34
Total	110
Heteroaromatics	18
Steroids	22
Others	14
Sum total	232

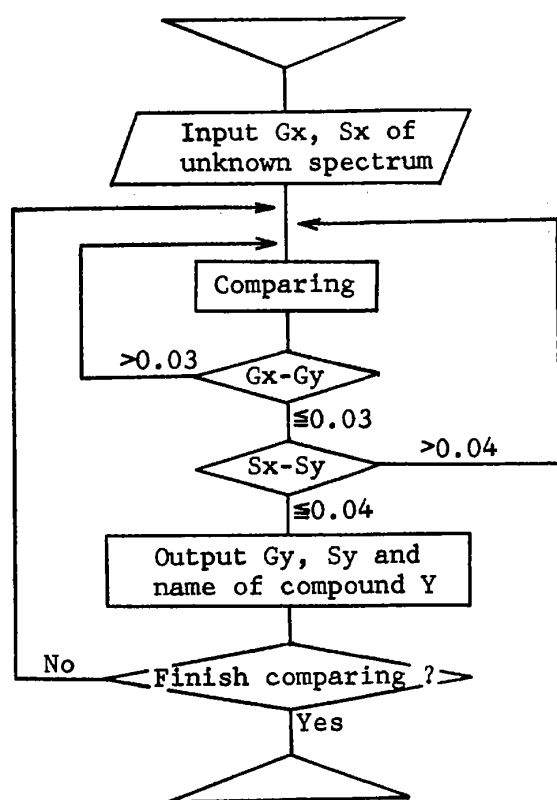


Fig. 2 Flowdiagram of structure identification by comparison of G and S values

3.2 検索の方法と結果

検索の方法は簡単で Fig. 2 のフローチャートに沿って進む。まず被検スペクトル (X) の G_x , S_x が入力されると、ファイル中に格納された化合物 G_y , S_y と比較され、2・3 に記した基準に合うか否かが検討される。1 回の照合に要する時間は (100~150) μ s 程度である。前記のように現在、ファイルには 232 件の格納があるので、その中から一つをとって順次被検試料として検索を行ったところ、Table 4 に示す結果を得た。Table 4 は分子式を未知としたときの検索結果で、判定基準 (0.03, 0.04) の場合とそれをややゆるめて (0.04, 0.05) とした場合の結果が表示されている。前者の基準を用いる限り被検試料の 72% (166 検体) が正解のみをもたらし、最も不良の場合でも正解以外に 3 個の不正解 (ノイズ) を発生するにすぎないことが分かる。3 個のノイズを伴う 3 検体とは 2-methyl-3-heptanol, 3-methyl-4-octanol 及び 2,4-xyleneol である。今、メチルヘプタノールとその関連化合物の G, S 値を列挙してみると Table 5 のとおりであって、2-methyl-3-heptanol (A) は 3-methyl-4-heptanol (B), 2-methyl-3-octanol (H), 3-methyl-4-octanol (I) と区別されえぬことが明らかである。同様に (I) は (A), (B), (H) と区別できない。

Table 4 Computed results by G, S-file searching (Molecular formula : unknown)

Noise	$\Delta G=0.03$ $\Delta S=0.04$	$\Delta G=0.04$ $\Delta S=0.05$
0	166 (72)	147 (63)
1	53 (23)	56 (24)
2	10 (4)	18 (8)
3	3 (1)	8 (3)
4	0	3 (1)
5	0	0
6	0	0
7	0	0
8	0	0
9	0	0

Table 5 G and S values of methylheptanols and methyloctanols

		G	S	Noise
A	2-Methyl-3-heptanol	1.27	0.58	B, H, I
B	3- " -4- "	1.30	0.60	A, I
C	6- " -2- "	1.25	0.66	
D	5- " -3- "	1.31	0.66	
E	3- " -2- "	1.37	0.67	
F	4- " -4- "	1.14	0.17	
G	3- " -3- "	1.16	0.22	J
H	2-Methyl-3-octanol	1.30	0.56	A, I
I	3- " -4- "	1.30	0.59	A, B, H
J	4- " -4- "	1.18	0.20	G

ただし分子式が分かっているものとすれば (A) のノイズは (B) のみ、(I) のノイズは (H) のみとなり検索の精度が向上することは論をまたない。またキシレノール類もノイズを発生しやすいが、そのなかで 2,4-xyleneol (O) が 3 個のノイズ (M, P, Q) を伴う事情は Table 6 から理解されよう。なお判定基準を (0.04, 0.05) としても相当よい検索精度が期待できる (Table 4)。最後に分子式が既知のときにいかにノイズの発生が少なくなるかを Table 7 で示す。例えば共通の分子式 $C_8H_{18}O$ を持つ検体 17 を検索したところ 15 検体が正解のみ 1 個 (ノイズ 0)、2 検体が正解とノイズ 1 個を出力した。そのほかについてもノイズを 2 個以上出力するのは例外ともいえることが Table 7 から分かる。

Table 6 G and S values of xyleneols

		G	S	Noise
L	3,5-Xyleneol	3.79	1.98	
M	2,6 " "	3.77	2.05	N
N	3,4 " "	3.79	2.07	M, O
O	2,4 " "	3.82	2.08	M, P, Q
P	2,3 " "	3.83	2.07	O, Q
Q	2,5 " "	3.83	2.07	O, P

Table 7 Computed results by G, S-file searching
(Molecular formula : known)

	(Total)	Noise			
		0	1	2	3
C ₈ H ₁₀ O	18	13	1	3	1
C ₈ H ₁₁ N	22	22	0	0	0
C ₈ H ₈ O ₃	17	17	0	0	0
C ₈ H ₈ O ₂	16	12	4	0	0
C ₉ H ₁₀ O ₃	16	16	0	0	0
C ₈ H ₉ NO ₂	15	15	0	0	0
C ₈ H ₁₄ O	17	15	2	0	0
C ₈ H ₁₆ O	16	14	2	0	0
C ₉ H ₂₀ O	11	9	2	0	0
Others	84	74	10	0	0
	232	207	21	3	1
		89%	9%	1%	0.4%

4 結 語

構造同定を NMR スペクトルのファイル検索で行う方法についての試案を述べた。実験の結果から「正解を落とさない」という配慮をしても実用上十分な検索精度を保ちうる事が分かった。必要とする記憶容量も 1 スペクトルにつき (1~2) words で済み、演算の過程も簡潔であるから検索上の要求はすべて満たされているものと思える。本法は NMR スペクトルによる構造同定法として独立に用いられる程度の検索精度を有するものではあるが、著者らの計画では他スペクトル (赤外, 質量) と 1 組にして格納し, 完全な検索を行う機能を持たせた上でシステム CHEMICS-F⁴⁾ の一部に取り入れる予定になっている。

研究費は特定研究 (広域大量情報の高次処理) 補助金によってまかなわれた。付記して謝意を表する。

文 献

- 1) 田辺和俊, 佐伯慎之助, 田村禎夫: 本誌, **23**, 626 (1974).

- 2) "Mass Spectral Search System", May 1974, U. S. Environmental Protection Agency Office of Planning and Management.
- 3) 阿部英次, 佐々木慎一, 斉藤啓治, 石田嘉明: 日本化学会第 29 秋季年会講演予稿集, I, p. 72 (1973).
- 4) 佐々木慎一: 化学の領域, **27**, 1117 (1973).
工藤喜弘, 落合周吉, 佐々木慎一: 化学と工業, **27**, 395 (1974).

☆

File searching of organic compounds with the use of the center of gravity and the standard deviation of NMR spectral line. Shin-ichi SASAKI, Yasuhiko YOTSUI and Shukichi OCHIAI (Miyagi University of Education, Aoba, Sendai-shi, Miyagi)

The nmr spectrum was coded numerically into G and S values so as to prepare the file searching system for the structure identification of organic compounds. Here, G stands for the center of gravity of the whole spectrum and S for the standard deviation of each signal to the center (Equations 1 and 2, respectively). The two values of each of 232 nmr spectra of various compounds were stored in the computer. The file searching for the unknown compound was carried out by comparing the spectrum of the compound with the store. The result of the search proved fairly good usefulness of this method. Even in case that the molecular formula was uncertain, proper structure was reached in all cases — 72% of the tests gave unique answer of correct structure, 1% the correct structure plus three noises, the other 27% the correct structure plus one or two noises (Table IV). In the case when the molecular formula was known, the probability of finding unique answer of correct structure by this method was found as 89% (Table VII).

(Received July 15, 1974)

Keywords

File search

Nmr spectra

Structure identification, organic compounds