

コンピュータによる赤外スペクトルからの構造推定

田辺 和俊^{a*}, 松本 高利^b, 伊藤 祥司^c, 上坂 博亨^d,
都築 誠二^e, 田村 禎夫^e, 佐伯 慎之助^e, 小野 修一郎^a

^a 千葉工業大学社会システム科学部, 〒 275-0016 習志野市津田沼 2-17-1

^b 東北大学多元物質科学研究所, 〒 980-8577 仙台市青葉区片平 2-1-1

^c 筑波大学工学研究科, 〒 305-0006 つくば市天王台 1-1-1

^d 富山国際大学地域学部, 〒 930-1292 富山県上新川郡大山町東黒牧 65-1

^e 産業技術総合研究所計算科学研究部門, 〒 305-8568 つくば市梅園 1-1-1

*e-mail: tanabe@pf.it-chiba.ac.jp

(Received: May 12, 2004; Accepted for publication: May 31, 2004; Published on Web: December 28, 2004)

コンピュータを用いて赤外スペクトルから化学構造を推定するには、データベースを利用する帰納的方法、知識ベースを利用する経験的方法、計算化学を利用する理論的方法、など様々な方法がある。赤外スペクトルからの構造推定能力の向上をめざして、コンピュータを用いて化学構造を推定するこれら様々な手法を総合的に検討した。まず、赤外スペクトルデータベースを利用する帰納的方法を検討し、問題点を考察した。次に、線形学習機械やニューラルネットワークを用いて、赤外スペクトルから官能基の有無を推定する経験的手法を検討し、問題点を考察した。さらに、計算化学を利用する理論的方法の能力を検証するために、非経験的分子軌道法により赤外スペクトルの振動数と強度を計算し、実測値をどの程度再現できるかを検討した。しかし、どの方法を用いても、現状のコンピュータの能力では赤外スペクトルから構造を推定するには多くの問題があることが分かった。

キーワード: 赤外スペクトル, 構造推定, データベース, ニューラルネットワーク, 非経験的分子軌道計算

1 緒言

赤外スペクトルは他のスペクトル分析法と比較して、気体、液体、固体のどのような化合物でも測定できること、1 mg 程度の微量の試料でも測定できること、迅速な測定も可能であり化学反応の制御にも利用できること、さらに、歴史的に最も古くから利用されているためスペクトルが豊富に蓄積されていること、などの特徴がある。そのため、材料、生体、環境などの分野での学術的な研究目的だけでなく、産業界などでの実用的な目的でも非常に幅広く利用されている。

しかし、赤外スペクトルの感度は質量スペクトルに比べると劣るため、微量分析では質量スペクトルの方がよく利用される。さらに、構造解析の手段としては NMR スペクトルに遅れをとっている。以前は未知化

合物の構造決定ではまず赤外スペクトルを測定していたが、最近では NMR のみで構造決定が行われることが多い。この理由としては、NMR の解析からは確実な構造情報が得られるのに対し、赤外スペクトルの解釈では 50 年以上前に作られた Colthup の相関表 [1] などが現在でも用いられているように未だ経験的な手法に止まっており、NMR ほど確実な構造推定ができないためであると考えられる。

しかし、近年著しく進展したコンピュータの情報処理および計算処理能力を駆使して赤外スペクトルの豊富なデータを解析することにより、赤外スペクトルからの構造推定の威力が格段に向上することが期待される。

そこで、我々は赤外スペクトルからの構造推定能力の向上をめざして、コンピュータを用いて赤外スペク

トルから化学構造を推定する様々な手法を長年にわたって総合的に検討してきた。本報ではそれらの研究成果を総括し、様々な手法の問題点を明らかにするとともに、将来の方向を展望する。

2 帰納的方法 - スペクトルデータベース

コンピュータを用いて赤外スペクトルから化学構造を推定する第1の方法は、多数の既知の化合物についてスペクトルを測定し、それらの標準スペクトルを集積したデータベースを構築しておき、未知化合物の実測スペクトルとデータベース中の標準スペクトルとを逐一照合するという帰納的な方法である。この目的でいろいろなスペクトルについてデータベースが世界各国で構築され、利用されている。

我々の研究所でも20年以上前から赤外スペクトルを含む各種のスペクトルのデータを集積した統合的なスペクトルデータベースシステム SDBS [2] の構築、および赤外スペクトルデータの検索アルゴリズムの研究 [3-18] を行ってきた。SDBS は同一化合物について各種のスペクトルを自ら測定し集積している統合的な

スペクトルデータベースであること、またきわめて精度の高いデータがデータベースに集積されていることが他に類をみない特徴であり、海外から高い評価を受けている。現時点で約3万の化合物について総計約5万件の赤外スペクトルが集積されているが、スペクトルの測定とデータの集積は現在も継続されている。

しかし、この帰納的方法にも幾つかの問題点がある。情報検索の一般の問題として、正答を確実に与え、雑音が少なく、しかも処理時間が短いといった相矛盾する条件を満たす最適の検索アルゴリズムをみつけないのは困難な問題である。我々は様々な検索アルゴリズムについて SDBS のデータを用いて検索実験を行い、Table 1 に示す結果から最適の検索アルゴリズムを見出した [13]。しかし、最適の検索アルゴリズムはデータベースの規模、インデックスデータの作成方法、用いるコンピュータの性能 (処理速度) などに依存するので、時代とともに変化する可能性がある。

また、最近では化学反応プロセスのモニタリングが関心を集めており [19]、そのためには反応途中の不安定化学種を同定する必要がある。しかし、スペクトルデータベースは元来、安定化合物について構築されることが多いので、不安定化学種のような未知化合物の関与する問題には無力である。

Table 1. Average ranks of correct answers

	Scoring equation	Average rank
A	$\Sigma I_h / \Sigma I_s \times 100$	27.1
B	$[\Sigma I_h - \Sigma I_n] / \Sigma I_s \times 100$	10.4
C	$[\Sigma I_h - \Sigma I_{n'}] / \Sigma I_s \times 100$	6.3
D	$[\Sigma I_h - \Sigma I_s - I_h] / \Sigma I_s \times 100$	17.4
E	$[\Sigma I_h - \Sigma I_n - \Sigma I_s - I_h] / \Sigma I_s \times 100$	7.5
F	$[\Sigma I_h - \Sigma I_{n'} - \Sigma I_s - I_h] / \Sigma I_s \times 100$	5.6
G	$[\Sigma I_s - \Sigma I_n] / \Sigma I_s \times 100$	21.5
H	$[\Sigma I_s - \Sigma I_{n'}] / \Sigma I_s \times 100$	4.5
I	$[\Sigma I_s - \Sigma I_s - I_h] / \Sigma I_s \times 100$	14.0
J	$[\Sigma I_s - \Sigma I_n - \Sigma I_s - I_h] / \Sigma I_s \times 100$	9.8
K	$[\Sigma I_s - \Sigma I_{n'} - \Sigma I_s - I_h] / \Sigma I_s \times 100$	6.0
L	$[N_s \Sigma I_s I_h - (\Sigma I_s)(\Sigma I_h)] / [N_s \Sigma I_s^2 - (\Sigma I_s)^2]^{1/2} / [N_s \Sigma I_h^2 - (\Sigma I_h)^2]^{1/2} \times 100$	37.9
M	$[N_f \Sigma I_s I_h - (\Sigma I_s)(\Sigma I_f)] / [N_f \Sigma I_s^2 - (\Sigma I_s)^2]^{1/2} / [N_f \Sigma I_f^2 - (\Sigma I_f)^2]^{1/2} \times 100$	1.7

I: peak intensity (absorbance), N: number of peaks, s: sample (input), f: file, h: hit peaks among file peaks, n: non-hit peaks among file peaks, n': non-hit peaks stronger than the weakest sample peak.

しかし、最大の問題点はデータベースの規模の問題である。データベース中の標準スペクトルと照合して正答が出るためにはその化合物のスペクトルがデータベースに集積されていることが不可欠である。データベースに集積されていない化合物はいかにしても正答が得られないからである。したがって、正答率を上げるためには可能な限りの化合物のスペクトルを集積してデータベースの規模をできるだけ大きくすることが必要となる。

ところが、これまでに化学者が論文で扱ってきた化合物の総数は 2,000 万種類以上にも達しており、その中で現実社会で流通している化合物だけでも数十万種類以上になるといわれている。それに対して、これまでに構築されているスペクトルデータベースは最大規模のものでも数万化合物程度であり、正答率の点からは十分な規模ではない。我々の SDBS も現在約 3 万化合物にとどまっており、大規模なスペクトルデータベースの構築はきわめて困難な問題である。

3 経験的手法

3.1 相関表

第 2 の方法は、多数の化合物についての化学構造とスペクトルの特徴との間の関係についての経験的な知識を利用して未知化合物の実測スペクトルからその構造を推定するという経験的な方法である。赤外スペクトルは化合物を構成する分子の振動に起因するものであり、スペクトルと構造との間には何らかの関係が存在する。事実、赤外スペクトルに関してはスペクトルの特徴と構造との間にはこれまで非常に多くの経験的な知識が特性吸収帯という名前で集積されている。

それらの経験的な知識はいろいろな形に整理されているが、代表的なものとして 1950 年に Colthup が発表した相関表 [1] がある。これは代表的な官能基についてスペクトル中に現れるピークの振動数範囲と大凡の強度を棒グラフの形式で示したものである。赤外スペクトルからの未知化合物の構造推定が現在でもこのような相関表にもとづいて行われ、また、そのような相関表を利用して構造推定するソフトウェアが市販の赤外スペクトル測定装置に搭載されている。

しかし、Colthup の相関表やソフトウェアを用いて未知化合物の構造推定を行おうとしても一般にはあま

りうまくゆかない。それは赤外スペクトルは図形であり、その特徴をこれらの相関表のように棒グラフで表してしまうと多くの特徴が失われるからである。

すなわち、赤外スペクトルには多数のピークが現れるが、ピークには振動数、強度、線幅の 3 要素があり、振動数と強度だけでなくピークが鋭いか幅広いかということも赤外スペクトルでは重要な特徴であり、スペクトル解析の専門家はピーク幅の情報も考慮して構造推定を行っているからである。しかし、Colthup の相関表などではこのようなピークの幅に関する情報が欠落している。

3.2 パターン認識

Colthup の相関表などに代表される経験的な知識は赤外スペクトルの特徴と構造の関係を簡潔に整理したものであるが、スペクトルの情報をこのような棒グラフで表すのではなく、図形そのまま扱えば構造推定の正答率も向上すると期待される。このような図形解析手法としてパターン認識があり、赤外スペクトルの問題にも多く適用されている。

我々もパターン認識の 1 手法である線形学習機械を用いて赤外スペクトルからの構造推定を検討した [20-23]。SDBS に集積されている赤外スペクトルデータを用いて 6 種類の官能基の有無を判定する識別関数を作成し、分類を行った結果、Table 2 に示すように 75 ~ 98% 程度の正答率で分類できることを見出した。

Table 2. Number of compounds which contain the functional group (Nc), number of compounds which contain it but are estimated not to contain (Ncn), number of compounds which do not contain it (Nn), number of compounds which do not contain it but are estimated to contain (Nnc), and error rate % (E)

Functional group	Nc	Ncn	Nn	Nnc	E
C=O	4570	342	5046	389	7.6
C=C	2135	967	7481	1364	24.2
C≡N	223	68	9393	150	2.3
C-O-C	1264	459	8352	752	12.6
Benzene ring	4572	698	5044	705	14.6
NO ₂	485	109	9131	212	3.3

しかし、この正答率は教師データに対する学習分類率であり、未知データに対する予測率はこの数値よりかなり低下すると思われる。なぜならば、赤外スペクトルパターンと構造との関係は本来、線形ではなく非線形であり、線形学習機械で取り扱うのは限界があると考えられるからである。

3.3 ニューラルネットワーク

スペクトルのような図形情報と化学構造との非線形関係の解析に有効な武器として期待されるのがニューラルネットワーク [24, 25] である。このニューラルネットワークは多くの利点をもっている。第 1 は相関関係を予め仮定する必要がなく、非線形な相関関係の解析が行えること、第 2 はスペクトルパターンのような図形情報がそのまま扱えること、第 3 は構造とスペクトルの特徴との関係についての知識が自動的に獲得できること、第 4 は並列化により高速な処理が行えること、第 5 は学習により高度な解析が行えることである。

このような利点のためにニューラルネットワークは化学の分野でもさまざまな問題に適用されており、中でもスペクトル、特に赤外スペクトルへの応用については多数の論文がある。我々も赤外スペクトルへの適用について幾つか論文を発表している [26–35] が、我々も含めて大多数の論文は赤外スペクトルからの化合物の同定に関するものである。それに対して、赤外スペクトルと構造との相関の解析、すなわち赤外スペクトルからの構造推定への応用に関する論文は少数である [29, 35–40]。

そこで、我々はニューラルネットワークを利用して赤外スペクトルから構造を推定する実用的なシステムを開発することを目的として、SDBS に集積されている大量のスペクトルデータを活用することにより、赤外スペクトルから多種類の官能基を推定するためのニューラルネットワークの最適化について研究を行った。

3.3.1 入力データ

本研究でニューラルネットワークの学習およびテストに用いた赤外スペクトルデータは SDBS の原データから作成した。SDBS では各化合物についての基本データを集積した化合物辞書があり、化合物名、分子式、構造式、CAS 登録番号などの他に、官能基を示

す 15 種類のコードが集積されている [2]。しかし、赤外スペクトルから構造を推定するシステムとしての実用性を考えると、推定する官能基がこの 15 種類だけでは不十分である。そこで、この 15 種類の官能基を主官能基とし、それぞれについて数百化合物以上の構造式を調べ、その主官能基の隣接の構造によって細かく分類して副官能基とした。このようにして分類された官能基と各官能基を有する化合物数 (スペクトル数) を Table 3 に示す。

赤外スペクトルの波数範囲は通常 $4,000 \sim 400 \text{ cm}^{-1}$ であり、SDBS の赤外スペクトルもこの波数範囲のデータを集積している。一方、スペクトル中のピーク幅 (半値幅) はピークによって様々であるが、ニトリル基のように鋭いピークでは $2 \sim 3 \text{ cm}^{-1}$ 程度のものもある。したがって、このような鋭いピークをもつ赤外スペクトルパターンを忠実に再現するためには 1 cm^{-1} 程度の波数間隔が必要となり、ニューラルネットワークに入力する波数点は 3,600 にもなる。しかし、本研究で用いたニューラルネットワークのソフトウェアはユニット数の上限が 255 であり、全波数範囲のスペクトルパターンを 1 cm^{-1} の波数間隔で入力することは不可能である。

この問題が赤外スペクトルなどのスペクトルパターンをニューラルネットワークで解析する場合の最大の問題であり、これまでの論文でも入力点を減らすためのスペクトルの特徴抽出の前処理方法として高速フーリエ変換 (FFT)、wavelet 変換、Hadamard 変換など色々な方法が試みられている。しかし、我々はスペクトル解析の専門家が構造推定を行う際に果たして FFT や wavelet 変換に類する特徴抽出操作を行っているのかどうかという点に疑問を抱き、むしろ専門家は実測の赤外スペクトルそのものに対してパターン認識を行っているのではないかと考えた。

そこで、本研究では専門家のそのような思考過程にできるだけ近い構造推定システムの開発を目指して、ニューラルネットワークに入力するデータとしてはこのような特徴抽出の変換操作をせず、実測のスペクトルパターンを入力することにした。したがって、ニューラルネットワークに入力する波数点が 255 以下になるように波数範囲と波数間隔を設定する必要がある。むしろ、入力点が 255 一杯になるよう波数範囲を広くとるよりも狭い波数範囲の方が正答率が高くなる可能性もある。

Table 3. Main functional groups, sub-functional groups, structures, numbers of substances that contain those groups (N), and identification accuracies (%) for present (P) and absent (A) groups

Main functional group	Sub-functional group	Structure	N	P	A
AMIDE			1024	83.0	80.7
	AMIDE1	CO-NH ₂	136	91.2	85.3
	AMIDE2	CO-NHR	608	81.9	82.9
	AMIDE3	CO-NR ₂	214	72.9	75.7
	AMIDE4	beta-lactam	12	83.3	33.3
	AMIDE5	gamma-lactam	10	80.0	80.0
	AMIDE6	delta-lactam	2	100.0	100.0
	AMIDE7	R-O-CO-NHR	4	100.0	0.0
	AMIDE8	NHR-CO-NHR	132	89.4	72.7
AMINO			682	80.6	81.2
	AMINO1	R-NH ₂	232	84.5	86.2
	AMINO2	Ar-NH ₂	188	88.3	81.9
	AMINO3	R-NH-R	82	75.6	75.6
	AMINO4	R-NH-Ar	50	80.0	76.0
	AMINO5	Ar-NH-Ar	16	100.0	62.5
	AMINO6	NR ₃	152	61.8	63.2
BENZ			886	72.0	75.6
	BENZ1	mono-substituted benzene ring	262	87.0	87.8
	BENZ2	o-di-substituted benzene ring	256	70.3	82.8
	BENZ3	m-di-substituted benzene ring	72	86.1	66.7
	BENZ4	p-di-substituted benzene ring	188	78.7	87.2
	BENZ5	1,2,4-tri-substituted benzene ring	124	66.1	75.8
	BENZ6	1,2,3-tri-substituted benzene ring	76	76.3	68.4
	BENZ7	1,3,5-tri-substituted benzene ring	10	20.0	60.0
CDC			1304	63.8	66.6
	CDC1	CH ₂ =CHR	186	79.6	86.0
	CDC2	CH ₂ =CR ₂	80	77.5	75.0
	CDC3	CHR=CHR(cis)	462	66.2	64.5
	CDC4	CHR=CHR(trans)	84	83.3	73.8
	CDC5	CHR=CR ₂	528	60.6	65.9
	CDC6	CR ₂ =CR ₂	148	44.6	74.3
	CDC7	C=C=C	2	100.0	100.0
	CDC8	C≡C	86	90.7	76.7
CHO			154	72.7	80.5
	CHO1	R-CHO	34	88.2	70.6
	CHO2	C=C-CHO	30	100.0	86.7
	CHO3	Ar-CHO	90	66.7	77.8
COOH			742	91.1	86.5
	COOH1	R-COOH	302	74.2	82.8
	COOH2	CX-COOH	24	91.7	83.3
	COOH3	C=C-COOH	62	90.3	71.0

	COOH4	Ar-COOH	174	88.5	81.6
	COOH5	amino acid	182	96.7	95.6
CTN			670	74.3	98.2
	CTN1	R-C≡N	292	63.0	95.9
	CTN2	C=C-C≡N	162	86.4	93.8
	CTN3	Ar-C≡N	220	63.6	96.4
ESTER			518	89.2	89.2
	ESTER1	R-COO-R	304	86.8	95.4
	ESTER2	C=C-COO-R	76	94.7	63.2
	ESTER3	R-COO-C=C	12	83.3	100.0
	ESTER4	C=C-COO-C=C	4	100.0	50.0
	ESTER5	Ar-COO-R	84	90.5	78.6
	ESTER6	R-COO-Ar	24	91.7	100.0
	ESTER7	Ar-COO-Ar	6	66.7	100.0
	ESTER8	C=C-COO-Ar	6	100.0	0.0
ETHER			486	68.3	77.8
	ETHER1	R-O-R	236	89.8	83.9
	ETHER2	R-O-C=C	14	85.7	85.7
	ETHER3	R-O-Ar	194	74.2	91.8
	ETHER4	C=C-O-C=C	28	64.3	78.6
	ETHER5	Ar-O-Ar	12	33.3	100.0
	ETHER6	C=C-O-Ar	12	83.3	66.7
HALO			906	57.2	68.0
	HALO1	C-F	96	52.1	77.1
	HALO2	C-Cl	548	55.5	65.0
	HALO3	C-Br	252	43.7	66.7
	HALO4	C-I	64	37.5	59.4
KETONE			632	71.2	76.9
	KETONE1	R-CO-R	154	72.7	92.2
	KETONE2	R-CO-C=C	136	64.7	75.0
	KETONE3	R-CO-CX	16	87.5	37.5
	KETONE4	R-CO-CO-R	12	83.3	66.7
	KETONE5	5-membered ring ketone	38	89.5	89.5
	KETONE6	R-CO-Ar	158	62.0	84.8
	KETONE7	Ar-CO-Ar	42	85.7	95.2
	KETONE8	p-quinone	78	79.5	79.5
NITRO			1028	90.7	94.4
	NITRO1	R-NO ₂	34	94.1	100.0
	NITRO2	C=C-NO ₂	36	44.4	77.8
	NITRO3	Ar-NO ₂	948	87.8	92.4
	NITRO4	>N-NO ₂	10	40.0	80.0
	NITRO5	O-NO ₂	2	100.0	0.0
OH			772	76.4	85.0
	OH1	CH ₂ R-OH	248	83.1	87.9

	OH2	CHR ₂ -OH	248	73.4	90.3
	OH3	CR ₃ -OH	48	79.2	62.5
	OH4	Ar-OH	322	68.3	84.5
PHOS			482	75.5	80.5
	PHOS1	H-P(=O)(OR) ₂	8	75.0	50.0
	PHOS2	R-P(=O)(OR) ₂	226	82.3	85.0
	PHOS3	P(OR) ₃	4	100.0	100.0
	PHOS4	others	244	68.9	81.1
SULF			926	68.3	57.2
	SULF1	SO ₃ H	86	81.4	76.7
	SULF2	SO ₃ R	34	76.5	88.2
	SULF3	SO ₂ -O-	6	33.3	100.0
	SULF4	SO ₃	136	83.8	94.1
	SULF5	SH	108	66.7	46.3
	SULF6	-S-	326	58.9	59.5
	SULF7	SO ₂ -	102	76.5	88.2
	SULF8	-S-S-	44	45.5	45.5
	SULF9	>C=S	108	55.6	61.1
	SULF10	-SO-O-	6	66.7	66.7
	SULF11	-SO-	4	100.0	100.0
Overall			5944	74.5	80.3

R: alkyl group , Ar: aromatic group

すなわち、Colthupの相関表などに示されているように各官能基のピークが現れる波数範囲はほぼ一定であり、入力データをそれらの波数範囲に限定することも考えられる。なぜならば、スペクトル解析の専門家が構造推定を行う場合、全波数範囲のスペクトルパターンに着目しているのではなく、各官能基ごとに特定の波数範囲のパターンに注目しているのではないかと考えられるからである。したがって、ニューラルネットワークに入力する場合も必要最小限度の波数範囲のスペクトルデータを入力する方が有効である可能性もある。

そこで、15種類の各官能基について、SDBSに集積されている赤外スペクトルデータからその官能基を含む化合物と含まない化合物をそれぞれ数百以上抽出した。次に、それぞれについて平均スペクトルP、Aを計算し、両者の差スペクトルDを計算した。この方法で得られた各官能基ごとのスペクトルP、A、DをFigure 1に示す。

この差スペクトルDは各官能基に特有のスペクトルパターンであり、Colthupの相関表をコンピュータを用いてより定量的に表現したものと考えることができる。その吸光度が正の波数領域はその官能基の存在を、また吸光度が負の領域はその官能基の不在を示す特徴的な波数範囲であると考えられる。したがって、この差スペクトルの吸光度の絶対値が高い波数範囲をニューラルネットワークに入力すべき波数範囲と考えることができる。しかし、上に記したように、波数範囲は広い方が必ずしも適切とは言えない可能性がある。

そこで、この差スペクトルのデータを用いて波数範囲の影響を調べるために、官能基BENZについて波数範囲を4種類設定し、正答率への影響を調べた。その結果、波数範囲が最も狭い場合に正答率が最大になり、波数範囲はできるだけ狭い方が適切であることが分かった。そこで、この方針に基づき、Figure 1の差スペクトルのデータを参考にして各官能基の波数範囲と波数間隔を決定した。

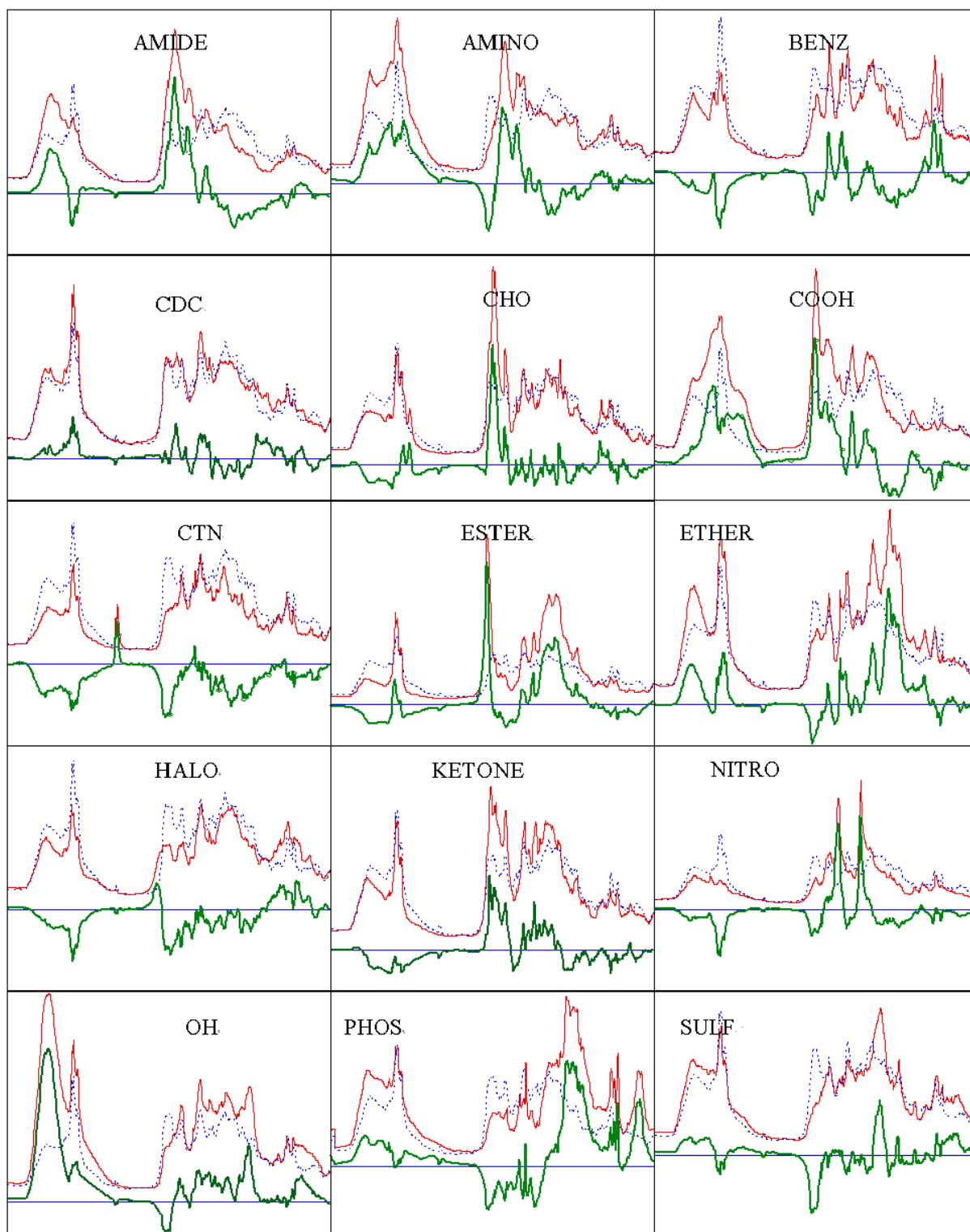
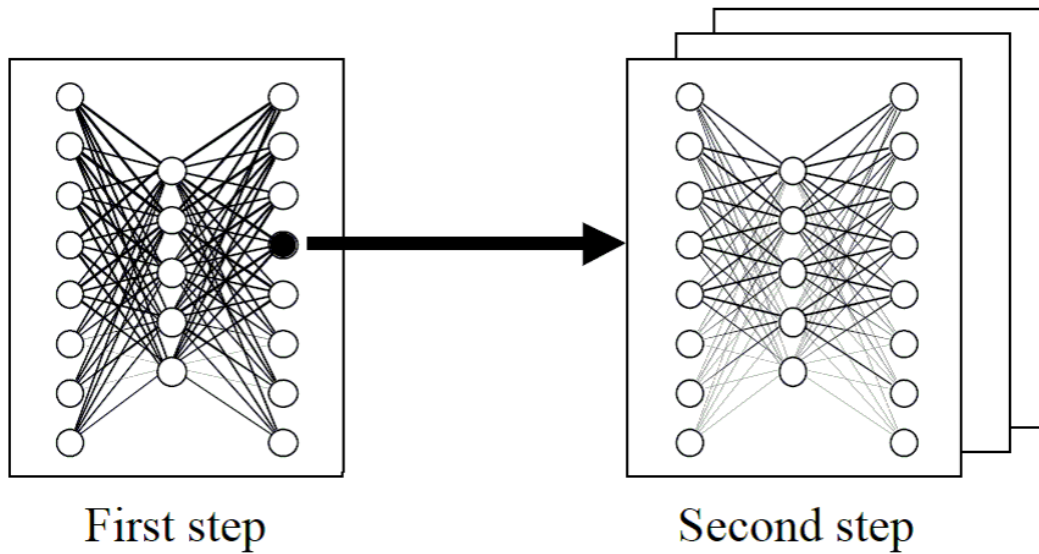


Figure 1. Averaged spectra for substances that contain (thin) and that do not contain (dotted) functional groups and difference spectra (thick).

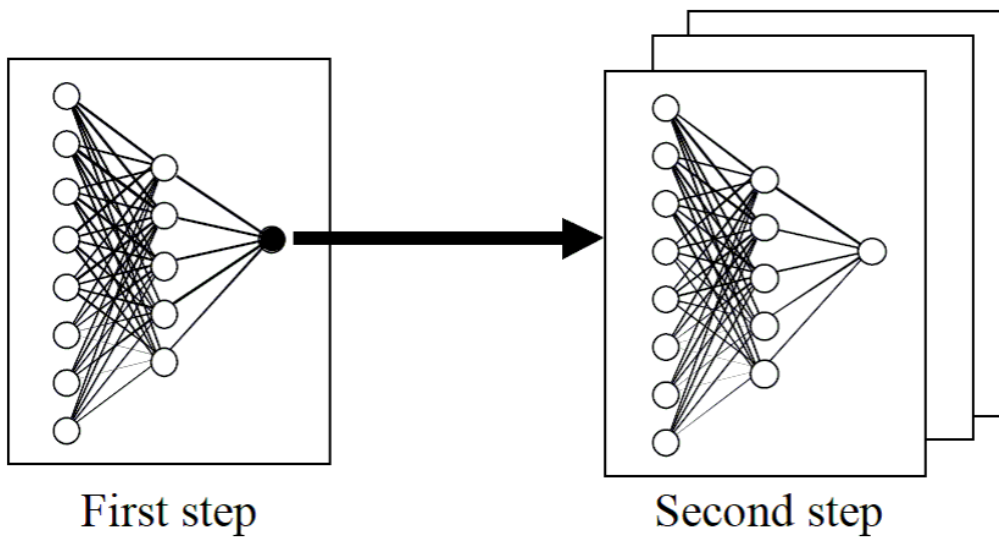
3.3.2 ニューラルネットワークの構造

ニューラルネットワークには幾つかの型があるが、赤外スペクトルから構造を推定する場合には、Figure 2 に示す入力層、中間層、および出力層から成る階層型のニューラルネットワークが有効である。スペクトルパターンデータを入力層に入力し、出力層には構造を教師データとして学習を行うものである。本研究では

ニューラルネットワークのソフトウェアには富士通の NEUROSIM/L を使い、Windows PC 上で学習とテストを行った。この NEUROSIM/L は階層型のニューラルネットワークシミュレータであり、学習にはエラーバックプロパゲーション方式が用いられている。NEUROSIM/L では中間層は 1 層または 2 層の設定が可能であるが、本研究では入力層、中間層、出力層各 1 層から成る 3 層構造のニューラルネットワークを用いた。



(a) For all main functional groups



(b) For each main functional group

Figure 2. Two types of neural network structures

このニューラルネットワークを用いて Table 3 に示した合計 90 種類の官能基を推定するために本研究では Figure 2 に示す 2 段構造のニューラルネットワークシステムを用いた。すなわち、初段のニューラルネットワークでは 15 種類の主官能基の判定を行い、初段のニューラルネットワークの出力層に結合している次段のニューラルネットワークでは副官能基の判定を行った。

赤外スペクトルを入力して 15 種類の官能基の有無を推定するニューラルネットワークを構築する場合、Figure 2 に示す (a)、(b) の 2 種類の構造のニューラルネットワークが考えられる。構造 (a) は、初段では 1 個のニューラルネットワークを構築し、出力層のユニット数を 15 にしてそれぞれに主官能基を割り当て、全てのデータを一括して用いて学習を行う方法である。構造 (b) は、初段の 15 種類の主官能基ごとに 1 個のニューラルネットワークを構築し、出力層のユニット数は 1 にして、それぞれの主官能基ごとに学習を行う方法である。

両方の方法でニューラルネットワークを構築、学習し、テストを行った。その他の条件は以下の検討における最適条件を用いたが、この点は以下の検討においても同様である。幾つかの官能基について検討した結果、(b) の場合に正答率が高いことが分かった。そこで、以下の検討においてはこの構造を採用した。

3.3.3 ニューラルネットワークの学習とテスト

本研究では実用的な構造推定システムを開発することを目的としてニューラルネットワークの最適条件の検討を行った。そのために、Table 3 に示した各官能基について SDBS のスペクトルデータを用いてニューラルネットワークの学習とテストを行った。その際、各官能基に対してその官能基を含む化合物と含まない化合物のスペクトルデータを同数作成して学習とテストを行った。官能基を含む化合物のデータは Table 3 に示した各官能基を含む化合物について作成し、含まない化合物のデータはそれ以外の官能基を含む化合物群からアトランダムに同数抽出して作成した。

ニューラルネットワークの教師データは官能基を含む場合を 1、含まない場合を 0 として学習を行った。スペクトルデータは学習用とテスト用に分割し、ニューラルネットワークの構造や学習法を種々変えながら正答率が最高になる最適条件を探索した。テストの正答率の算出には、その官能基を含む時に出力値が 0.5 以

上の場合、およびその官能基を含まない時に出力値が 0.5 以下の場合に正解とした。

ニューラルネットワークによる構造推定システムを開発するためには、システムの性能を求めるために予測テストを行う必要がある。そのためにはデータ全体を学習用とテスト用に分割し、前者を用いて学習したニューラルネットワークに後者を入力して正答率を求める必要がある。その際の学習用およびテスト用のデータを作成する方法には、データ全体を適当な割合で学習用とテスト用に分割する方法や、leave-one-out 法など色々な方法が考えられるが、本研究では前者の方法を採用した。

データ全体を学習用とテスト用に分割する際の比率 50% : 50% および 90% : 10% について、および学習用とテスト用データの作成位置前半/後半について、その影響を調べた。幾つかの官能基について検討した結果、データ作成の比率や位置にはほとんど依存しないことが分かった。学習用とテスト用データの比率が 50% : 50% および 90% : 10% の両者の場合で差がなかったことは、今回検討した官能基を含む化合物数が十分な大きさであったためであり、次段の副官能基のように化合物数が少ない場合には学習用とテスト用データの比率は影響を与える可能性がある。

また、本研究でニューラルネットワークの学習とテスト用に作成したデータは SDBS に集積されている化合物の中から該当する官能基を含む化合物を登録番号順に抽出したが、正答率が学習用、テスト用データの作成位置に依存しなかった結果は、SDBS に登録されている化合物が系統的な偏りが少なく、アトランダムに集積されていることを示していると考えられる。そこで、以下の検討においては、学習用データは前半の 50%、テスト用データは後半の 50% で作成することにした。

次段の副官能基の学習用およびテスト用データを作成する場合に、その副官能基を含む化合物のデータの作成法は問題ないが、その副官能基を含まない化合物のデータをどのように作成するかについては種々の方法が考えられる。本研究では次の 2 通りの方法を検討した。その 1 は、その副官能基を含む全ての主官能基からデータをアトランダムに抽出する方法であり、その 2 は、その副官能基が含まれる主官能基の内の他の副官能基のみからデータを抽出する方法である。以上の 2 つの方法について幾つかの副官能基について正答率を調べた。その結果、前者の作成法の方が正答率が高くなった。したがって、次段データの作成法は前者

の方法を採用して以下の検討を行った。

3.3.4 ニューラルネットワークによるテスト結果

以上のような各種の検討を行って最適条件を決定し、その最適条件においてニューラルネットワークを構築し、学習とテストを行った。15種類の主官能基および90種類の副官能基について得られた正答率を Table 3 に示す。

まず、主官能基の正答率について考察すると、15種類の主官能基は CTN を除いていずれも十分な数のデータでテストが行われた。したがって CTN 以外の 14種類の主官能基の正答率は十分信頼できるものと考えられる。それらの内で正答率の高いものは NITRO (88.9%)、ESTER (87.5%)、COOH (84.8%)、ETHER (83.7%) などであり、これらの官能基は赤外スペクトルからその存在が推定しやすいといわれているものである。一方、正答率が低かった主官能基は HALO (58.5%)、PHOS (67.9%)、CDC (68.5%) などであり、これらの官能基も赤外スペクトルからは存在が推定しにくいとされているものである。したがって、各主官能基についての正答率はこれまで経験的に定性的にいわれてきた赤外スペクトルからの官能基の推定の確度を数値化したものと考えられる。

次に、副官能基についての正答率を考察すると、上記

のように、データ数が少ない場合には正答率の信頼性が低いので、データ数が比較的多い副官能基の内では正答率の高いものをあげると ESTER1 (91.1%)、NITRO3 (90.1%)、AMIDE2 (82.4%) などであり、これらの官能基は赤外スペクトルからその存在が推定しやすいといわれているものである。逆に正答率の低いものは SULF6 (59.2%)、HALO2 (60.2%)、CDC5 (63.3%) などである。BENZ7 (40.0%)、SULF8 (45.5%)、HALO4 (48.4%) などの副官能基はデータ数が少なく信頼性が低い、正答率がきわめて低い点で注目される。

3.4 経験的方法の問題点

このように正答率がきわめて低い官能基が存在することは赤外スペクトルの本質によるものと解釈される。そのような赤外スペクトルの 1 例を Figure 3 に示す。この化合物は分子内に塩素原子を含んでいるが、Figure 3 の赤外スペクトルをニューラルネットワークに入力すると C-Cl 結合は存在しないという推定結果を与える。これは分子内に塩素原子が存在するにも拘わらず、C-Cl 結合に起因する特性吸収帯 (Colthup の表によれば $780 \sim 700 \text{ cm}^{-1}$ に強いピーク) が Figure 3 のスペクトルには見られないので、ニューラルネットワークは C-Cl 結合は存在しないという推定結果を与える。

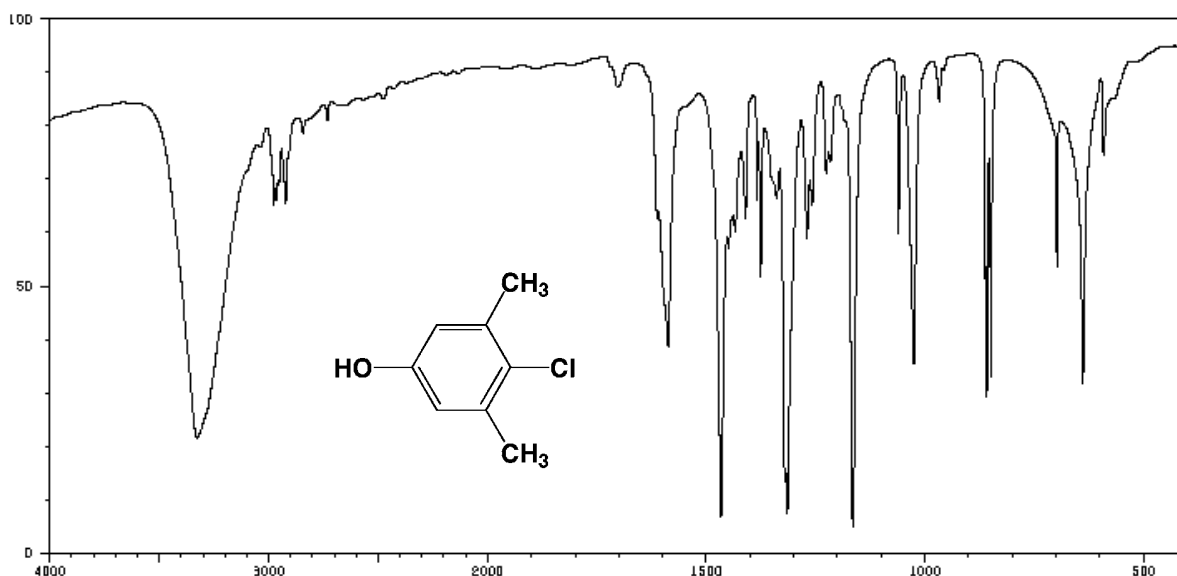


Figure 3. Infrared spectrum of 4-chloro-3,5-dimethylphenol

このように、ある官能基が分子内に存在するにも拘わらず、その特性吸収帯が赤外スペクトルに現れないために、ニューラルネットワークによる官能基推定の正答率が平均として 77.2 % という低い数値になったと考えられる。このような現象を与える原因としては

1. 分子の対称性によりその官能基を含む振動が赤外不活性、または振動双極子能率の値が小さくなり、赤外スペクトルにほとんど現れない。
2. その官能基による特性吸収帯が現れる波数領域で他の非常に強いピークと偶然重なったために、その官能基の存在が推定できない。
3. 赤外スペクトルは一般にスペクトルの中で最も強いピークの透過率が 5% 程度になるように試料セルの厚さや溶液濃度を調整して測定するので、ある官能基の特性吸収帯がある程度の強度で現れたとしても、他の領域で極めて強いピークが存在すると、そのピークの強度で全体を規格化するために、他のピークの強度は低くなり、最初の官能基が存在しないという推定になる。

など様々な原因が考えられる。上記の Figure 3 の場合に C-Cl に帰属されるべきピークが現れていない原因は不明である。

このように考察すると、ニューラルネットワークにより赤外スペクトルから官能基を推定する際の正答率として本研究で得られた平均 77.2% という正答率は NMR スペクトルから構造推定する場合の数値よりは低いと考えられる。しかし、この数値は赤外スペクトルから化学構造を推定する場合の限界を示していると考えられ、赤外スペクトルの本質に起因している以上はこれ以上の改善は不可能であろう。したがって、スペクトルパターンの特徴と化学構造との関係を利用する経験的方法にはここに最大の問題点が存在する。

4 理論的方法

第 3 の方法は、想定した構造の分子について分子軌道法などにより赤外スペクトルを計算し、その結果から実測の赤外スペクトルを与えるような構造を推定するという理論的方法である。近年のコンピュータの演算性能の向上により、かなり大きな分子でも赤外スペクトルの理論計算が可能になり、また計算法の進展により、計算されたスペクトルも実測スペクトルにかなり近いものが得られるようになった。中でも分子軌道

法による振動スペクトルの計算については多くの研究が行われている。

しかし、これまでの理論計算では専ら振動数の再現に重点が置かれており、振動スペクトルの 3 大要素である振動数、強度、線幅を総合的に再現することを目的とした理論的研究はほとんどない。その中で Schlegel らは振動数と赤外強度の非経験的分子軌道計算を行っている [41, 42] が、電子相関を取り入れた計算は不十分である。これは当時のコンピュータ能力の限界のためであるが、現在ではコンピュータの性能向上により電子相関をさらに十分に取り入れた計算が可能である。

また、彼らは計算値との比較対照として、振動数と赤外強度の実測値に振動の非調和性を補正した推定値を用いている。これは通常分子軌道計算のプログラムでは振動数の計算値は調和振動の近似のもとに算出されているが、実測の振動数には振動の非調和性が影響しているためである。しかし、分析対象の未知化合物についてこのような非調和性の補正は不可能であり、未知化合物の構造決定のための振動スペクトルの解析手法という観点からは、計算値の比較対照は実測値そのものに対して行われるべきである。

本研究では、振動スペクトルの理論的解析法として非経験的分子軌道法がどの程度実測スペクトルを再現できるかを検証するために、メタンから四塩化炭素に至る 13 種の塩素化メタン分子について GAUSSIAN94 プログラムを用いて振動数と赤外強度を計算した。これらの分子について我々は以前に赤外スペクトルの精密測定から幾つかの興味ある現象を見出していた。たとえば、 CH_4 、 CH_3Cl 、 CH_2Cl_2 、 CHCl_3 の系列では C-H 伸縮振動の赤外強度が大幅に変化するが、この現象は単純な電荷理論では説明できないこと [43]、また CH_2Cl_2 については気体の赤外スペクトルに異常な形状が観測されること [44] を観測しており、これらの実験事実が理論計算でどこまで説明できるかに興味があったからである。

4.1 方法

本研究で計算した分子はメタン 5 種 (CH_4 、 CH_3D 、 CH_2D_2 、 CHD_3 、 CD_4)、塩化メチル 2 種 (CH_3Cl 、 CD_3Cl)、ジクロロメタン 3 種 (CH_2Cl_2 、 CHDCl_2 、 CD_2Cl_2)、クロロホルム 2 種 (CHCl_3 、 CDCl_3)、四塩化炭素 1 種 (CCl_4) の計 13 種である。これらの分子について GAUSSIAN94 プログラム [45] を用い、分子構造を最適化した後、基底関数と計算法の多数の組み合

わせについて振動数と赤外強度を計算した。

用いた基底関数は 6-31G、6-31G(d)、6-31G(d,p)、6-31G(3df,3pd)、6-31+G、6-31+G(d)、6-31+G(d,p)、6-31+G(3df,3pd)、6-31++G、6-31++G(d)、6-31++G(d,p)、6-31++G(3df,3pd)、6-311G、6-311G(d)、6-311G(d,p)、6-311G(3df,3pd)、6-311+G、6-311+G(d)、6-311+G(d,p)、6-311+G(3df,3pd)、6-311++G、6-311++G(d)、6-311++G(d,p)、6-311++G(3df,3pd)、D95、D95(d)、D95(3df,3pd)、cc-pVDZ、cc-pVTZ、cc-pVQZ、AUG-cc-pVDZ、AUG-cc-pVTZ、AUG-cc-pVQZ の 33 種類であり、計算法は RHF、B3LYP、BLYP、BPW91、MP2(Full) の 5 種類である。コンピュータは IBM RS/6000 SP システム (Power2-135MHz) を用いた。計算値と比較すべき実測値として、振動数の実測値は文献値 [46] を、赤外強度の実測値は我々の文献値 [43] を用いた。

4.2 結果と考察

ここで取り上げた分子の基準振動の数は、四面体 T_d 型分子 (CH_4 、 CD_4 、 CCl_4) では 4 個 (その内、赤外活性振動は 2 個)、三角錐 C_{3v} 型分子 (CH_3D 、 CHD_3 、 CH_3Cl 、 CD_3Cl 、 CHCl_3 、 CDCl_3) では 6 個 (すべて赤外活性)、 C_{2v} 型分子 (CH_2D_2 、 CH_2Cl_2 、 CD_2Cl_2) では 9 個 (赤外活性は 8 個)、 C_s 型分子 (CHDCl_2) では 9 個 (すべて赤外活性) である。したがって、13 種の分子の基準振動の総数は合計 84 個 (赤外活性振動の総数は 75 個) になり、これだけの多数の振動について基底関数 33 種類、計算法 5 種類の組み合わせによる振動数と赤外強度の計算結果を本論文に示すことは不可能なので Appendix Tables 1-13 に示す [47]。

非経験的分子軌道法で振動数と赤外強度を計算する際には、分子振動の調和性が仮定されている。一方、現実の分子は非調和振動を行っており、したがって C-H 伸縮振動のような振動の非調和性の影響が大きい振動では振動数の計算値は実測値より高めになる。赤外強度に対する振動の非調和性の影響は現在まで未解明であるが、何らかの影響があると考えられる。

この振動の非調和性などの影響を補正するために、振動数については分子軌道法による計算値を補正する scale factor という概念が導入され、種々の scale factor の値がこれまでに提案されている [48–52]。すなわち、計算値と実測値の比が多数の分子の多数の振動について一定値に近ければ、この比を補正係数と考えて計算値から実測値を算出しようという考え方である。赤外

強度については振動数以上に振動の非調和性の影響が大きく、計算値が実測値を再現する可能性は低いと考えられる。そこで、ここで計算した多数の振動について振動数と赤外強度の計算値と実測値の比を求め、この scale factor を計算値に補正して実測値がどの程度の誤差で再現できるかを調べた。この振動数比と赤外強度比の結果も付録の Appendix Tables 1-13 にまとめて示す。

84 個の振動について基底関数 33 種類と計算法 5 種類の組み合わせ、計 165 種類の計算法による振動数の計算値/実測値の比 (以下では振動数比と略記) の総平均は 1.0205、標準偏差は 0.0374 となり、振動数の計算値はほぼ実測値を再現していることがわかる。しかし、振動形ごとに振動数比をみると顕著な傾向が認められる。各振動ごとに 165 種類の計算法について求めた振動数比の平均値と標準偏差を求めた結果を Table 4 に示す。

84 個の振動に共通して、振動数の高い振動ほど振動数比も高いという傾向が認められる。分子軌道法による振動数の計算値を補正する scale factor として種々の値がこれまでに提案されている [48–52]。従来の scale factor は振動数に関わらず一定値を用いるのに対し、最近、多数の分子についての DFT 法による振動数の計算結果から、振動数と 1 次式の関係の scale factor を用いる linear scaling 法が提案された [51, 52]。我々の結果は RHF と MP2 を含む計算法全般についてこの方法と同様な傾向があることを示している。

振動形ごとに振動数比をまとめると、C-H 伸縮振動では 1.0425-1.0626、C-D 伸縮振動では 1.0169-1.0461、C-H および C-D 変角振動では 1.0067-1.0485、C-Cl 伸縮振動では 0.9588-0.9815、C-Cl 変角振動では 0.9805-1.0030 となり、それぞれかなり狭い範囲に収まる。特に興味深いのは塩素原子が関与する振動の振動数比であり、振動数が高い C-Cl 伸縮振動よりも振動数が低い C-Cl 変角振動の振動数比が大きくなった。

この結果は振動数が高いほど振動数比も高い関係にあるとする上記の linear scaling 法とは逆の結果である。我々の結果は塩素化メタン分子 13 種の 84 個の振動についての結果であるのに対し、linear scaling 法の論文でははるかに多種多様な分子のはるかに多数の振動を取り扱っており、単純な比較はできない。しかし、振動数が高い C-Cl 伸縮振動よりも振動数が低い C-Cl 変角振動の振動数比が大きいという我々の結果は、塩素を含む分子の振動の非調和性について有益な情報を提供していると思われる。

Table 4. Uncertainty class (UC), observed frequencies (cm^{-1}), means and standard deviations of frequency ratios

	UC	Obsd Freq	Mean	Std Dev		UC	Obsd Freq	Mean	Std Dev		
CH ₄	v ₄	C	1306	1.0476	0.0475	CH ₂ Cl ₂	v ₄	B	282	0.9981	0.0604
	v ₂	A	1534	1.0312	0.0367		v ₃	B	717	0.9776	0.0649
	v ₁	A	2917	1.0425	0.0252		v ₉	B	758	0.9646	0.0930
	v ₃	A	3019	1.0456	0.0250		v ₇	B	898	1.0150	0.0460
CH ₃ D	v ₆	C	1155	1.0451	0.0455		v ₅	B	1153	1.0320	0.0523
	v ₃	C	1300	1.0485	0.0473		v ₈	B	1268	1.0384	0.0521
	v ₅	C	1471	1.0346	0.0382		v ₂	C	1467	1.0067	0.0459
	v ₂	A	2200	1.0393	0.0259		v ₁	B	2999	1.0486	0.0324
	v ₁	E	2945	1.0442	0.0259		v ₆	B	3040	1.0626	0.0328
	v ₄	B	3017	1.0463	0.0260		CHDCl ₂	v ₆	B	283	0.9902
CH ₂ D ₂	v ₄	C	1033	1.0340	0.0433	v ₅		B	692	0.9815	0.0621
	v ₇	C	1090	1.0419	0.0470	v ₉		B	738	0.9665	0.0868
	v ₉	C	1234	1.0435	0.0470	v ₄		C	778	1.0147	0.0481
	v ₅	C	1333	1.0283	0.0366	v ₈		A	890	1.0215	0.0568
	v ₃	C	1436	1.0353	0.0397	v ₇		A	1223	1.0342	0.0523
	v ₂	C	2202	1.0169	0.0253	v ₃		C	1282	1.0286	0.0470
	v ₈	C	2234	1.0461	0.0262	v ₂		B	2249	1.0418	0.0319
	v ₁	C	2974	1.0439	0.0258	v ₁	B	3024	1.0565	0.0322	
CHD ₃	v ₆	C	3013	1.0476	0.0260	CD ₂ Cl ₂	v ₄	C	282	0.9893	0.0605
	v ₃	C	1003	1.0400	0.0469		v ₃	B	687	0.9807	0.0642
	v ₆	C	1036	1.0330	0.0429		v ₇	D	712	1.0106	0.0467
	v ₅	C	1291	1.0349	0.0400		v ₉	B	727	0.9700	0.0857
	v ₂	C	2142	1.0245	0.0255		v ₅	C	826	1.0243	0.0523
	v ₄	C	2263	1.0327	0.0259		v ₈	B	957	1.0321	0.0593
	v ₁	C	2993	1.0462	0.0259		v ₂	D	1052	1.0297	0.0478
CD ₄	v ₄	B	996	1.0390	0.0468		v ₁	B	2205	1.0344	0.0325
	v ₂	B	1092	1.0251	0.0366	v ₆	C	2304	1.0457	0.0321	
	v ₁	B	2109	1.0201	0.0254	CHCl ₃	v ₆	B	261	0.9957	0.0608
	v ₃	A	2259	1.0345	0.0260		v ₃	C	363	1.0018	0.0605
CH ₃ Cl	v ₃	A	732	0.9721	0.0701		v ₂	B	680	0.9730	0.0597
	v ₆	A	1017	1.0249	0.0459		v ₅	B	774	0.9588	0.0987
	v ₂	A	1355	1.0385	0.0479		v ₄	B	1220	1.0290	0.0563
	v ₅	A	1452	1.0344	0.0424		v ₁	B	3034	1.0570	0.0340
	v ₁	E	2937	1.0523	0.0299	CDCl ₃	v ₆	C	262	0.9888	0.0606
	v ₄	B	3039	1.0521	0.0301		v ₃	C	369	0.9805	0.0590
CD ₃ Cl	v ₃	A	701	0.9736	0.0662		v ₂	B	659	0.9753	0.0604
	v ₆	A	768	1.0169	0.0458		v ₅	B	749	0.9626	0.0901
	v ₂	A	1029	1.0295	0.0494		v ₄	B	914	1.0177	0.0644
	v ₅	A	1060	1.0269	0.0420		v ₁	C	2266	1.0430	0.0334
	v ₁	A	2160	1.0238	0.0292	CCl ₄	v ₂	C	217	1.0030	0.0645
	v ₄	A	2283	1.0412	0.0298		v ₄	C	314	0.9997	0.0587
Uncertainty class (UC) taken from ref [46].							v ₁	C	459	0.9753	0.0720
							v ₃	E	776	0.9717	0.1060

一方、標準偏差は C-Cl 伸縮振動が最大で、C-Cl 変角振動や C-H または C-D 変角振動がそれに次ぎ、C-H または C-D 伸縮振動で最小、という傾向が見られる。このことは、種々の計算法のどれを用いても C-H または C-D 伸縮振動は振動数比がほぼ一定値になるのに対し、塩素原子が関与する振動の振動数比は計算法によってかなりばらつくことを意味する。

すなわち、塩素原子が関与する振動の振動数比は基底関数により大きく変動し、6-31G、6-311G、D95 などの基底関数を使った場合は分極関数を含む基底関数を使った場合に比べかなり低い値になる。一方、分極関数を含む基底関数を使った場合には基底関数の影響はごくわずかであり、塩素原子が関与する振動の振動数の計算には分極関数を含む基底関数を使う必要があることが分かる。また、塩素原子が関与する振動の振動数は計算法によっても大きく変動し、RHF や MP2 に比べて DFT 法は低目の振動数を与える傾向が認められるが、この傾向は文献 [51, 52] においても指摘されている。

赤外強度についても、振動数の場合と同様、各振動ごとに 165 種類の計算法について求めた強度比の平均値と標準偏差を求めた結果を Table 5 に示す。赤外強度の計算値は振動数の場合よりも実測値との差が大きい。赤外不活性および観測されない振動を除く 72 個の振動についての強度比の平均値は 1.6693、標準偏差は 2.2624 となる。上記のように、調和振動の近似の下で計算される振動数あるいは赤外強度の計算値と、振動の非調和性が含まれている実測値との比の大半は分子振動の非調和性によるものと考えられる。

したがって、上記の振動数比 1.0205 は振動数に対する非調和性の影響が僅か 2.05 % にすぎないことを意味する。それに対して、強度比 1.6693 という結果は赤外強度に対する非調和性の影響が非常に大きいこと、およびその影響が赤外強度を減少させる方向にあることを意味している。振動数に対する非調和性の影響に関する研究は多いが、赤外強度に対する非調和性の影響に関する研究はないので、この結果を検証できる実験データはないが、興味深い結果である。

赤外強度の計算値と実測値の違いのもう 1 つの原因として考えられるのが電子相関の効果である。本研究の計算結果から計算法ごとに強度比の平均値を求めると、RHF は 2.1788、B3LYP は 1.4878、BLYP は 1.7240、BPW91 は 1.5467、MP2 は 1.4600 となり、電子相関が取り入れられていない RHF では計算値が実測値の 2 倍以上の値になる。赤外強度を MP3 以上の

電子相関補正法で計算した論文は見あたらないので、DFT 法や MP2 における強度比がほぼ 1.5 という値が振動の非調和性によるものか、電子相関によるものかどうかは不明であるが、この点も興味ある結果である。

赤外強度の場合は振動数の場合と異なり、実測強度の測定誤差が大きな振動が多く、また強度比の標準偏差が大きい振動も多いので、強度に対する振動形の影響は認めにくい。しかし、注意すべきは CHCl_3 の ν_6 、 CDCl_3 の ν_6 と ν_1 、 CCl_4 の ν_4 における標準偏差の異常な大きさである。このことはこれらの赤外強度が計算法によって大きく変動することを意味する。事実、標準偏差が最大の CDCl_3 の ν_1 では強度比は MP2/D95 での 72.9 から MP2/cc-pVTZ での 0.000 まで大きな差がある。この振動は C-D 伸縮振動であり、塩素原子が関与する振動ではない。しかし、この分子では塩素原子が 3 個存在するため、C-D 結合の伸縮振動における電子密度の変化量 (赤外強度に比例する) が基底関数、特に電子相関の効果によって大きく影響されるためであると考えられる。

分子軌道法による理論計算が赤外スペクトルをどの程度まで再現するかという観点からは、振動数や赤外強度の計算値の絶対値が実測値と大きくかけ離れていても問題ではなく、重要なことは計算値と実測値の比率が一定であることである。すなわち、その比率が一定であれば比率を scale factor として計算値を補正すれば実測値が再現できることになる。

この観点から各計算法の精度を比較するために、基底関数と計算法の組み合わせについて、全振動 (赤外強度については不活性および観測できない振動は除く) についての振動数比と強度比の平均値と標準偏差を計算した。ただし、実測値には測定誤差が存在するので、この誤差に見合った重みを考慮して平均値と標準偏差を計算した。すなわち、振動数の場合は文献 [46] に記載されている振動数の不確かさのクラスによって測定誤差 $\Delta\nu$ を A: 1.0、B: 2.0、C: 4.5、D: 10.5 cm^{-1} とし、これに基づく重み w_i を

$$w_i = (1/\Delta\nu)^2$$

により計算した。

一方、赤外強度の場合は強度の実測値 I とその測定誤差 ΔI に対して重み w_i は

$$w_i = (I/\Delta I)^2$$

により計算した。これらの重み w_i を用い、振動数比または強度比 x_i の平均 M と標準偏差 SD を

$$M = \sum w_i x_i / \sum w_i$$

$$SD = [\sum w_i (M - x_i)^2 / \sum w_i]^{1/2}$$

Table 5. Observed intensities (km/mol), their errors (km/mol), means and standard deviations of intensity ratios

		Obsd Int	Error	Mean	Std Dev			Obsd Int	Error	Mean	Std Dev
CH ₄	v ₄	34.00	2.00	1.5545	0.6922	CH ₂ Cl ₂	v ₄	0.60	0.10	0.8921	0.3646
	v ₃	74.00	5.00	1.1845	0.3685		v ₃	8.00	0.40	1.7307	0.4624
CH ₃ D	v ₆	15.70	2.00	0.7410	0.3285	v ₉	95.00	8.00	1.6580	0.2374	
	v ₃	12.20	1.50	1.4853	0.6584	v ₇	1.20	0.10	1.5378	0.9025	
	v ₅	4.10	1.00	0.7697	0.3773	v ₈	26.60	1.20	1.9008	0.5058	
	v ₂	7.20	1.00	1.1487	0.4001	v ₂	0.60	0.10	0.2190	0.3512	
	v ₁	7.20	2.00	1.2204	0.3830	v ₁	6.90	0.50	1.2068	0.5075	
	v ₄	53.00	2.00	0.5452	0.1730	v ₆	0.00				
CH ₂ D ₂	v ₄	5.20	1.50	1.4577	0.6316	CHDCl ₂	v ₆	0.60	0.10	0.9143	0.3763
	v ₇	8.10	2.00	1.4943	0.6740		v ₅	5.70	0.40	2.0031	0.5601
	v ₉	11.10	2.00	1.4773	0.6523		v ₉	72.00	3.00	1.8017	0.3914
	v ₃	4.40	1.50	1.5018	0.6967		v ₄	2.80	1.00	0.8065	0.2446
	v ₂	5.60	1.50	0.8440	0.3020		v ₈	29.00	0.70	1.3647	0.3941
	v ₈	12.20	2.00	1.0631	0.3607		v ₇	17.00	0.40	2.0500	0.5569
	v ₁	13.80	2.00	1.1684	0.3687		v ₃	0.60	0.10	1.5514	0.6371
	v ₆	26.00	2.00	1.1002	0.3494		v ₂	2.40	0.10	1.7031	0.7487
CHD ₃	v ₃	7.00	0.50	1.4838	0.6657	v ₁	3.10	0.10	1.2846	0.6501	
	v ₆	8.10	0.50	1.8310	0.7745	CD ₂ Cl ₂	v ₄	0.60	0.10	0.9332	0.3768
	v ₅	7.10	0.50	1.7175	0.8150		v ₃	8.00	1.50	1.5128	0.3973
	v ₂	2.50	1.00	0.8239	0.3025		v ₇	0.00			
	v ₄	22.70	1.50	1.1348	0.3858		v ₉	67.00	3.00	1.7922	0.4022
	v ₁	18.30	1.50	1.2269	0.3890		v ₈	50.00	2.00	1.6114	0.3760
					v ₂		0.20	0.10	3.6932	2.9995	
CD ₄	v ₄	20.00	1.00	1.5942	0.7056	v ₁	4.30	0.20	1.5944	0.5997	
	v ₃	35.00	1.50	1.0963	0.3736	v ₆	0.00				
CH ₃ Cl	v ₃	22.00	1.00	1.3736	0.4285	CHCl ₃	v ₆	0.05	0.03	4.3290	6.1091
	v ₆	4.10	0.50	1.8184	0.8486		v ₃	0.50	0.10	0.7144	0.5820
	v ₂	7.50	0.50	2.2169	0.8416		v ₂	4.40	0.60	1.3646	0.5768
	v ₅	12.00	1.00	1.0708	0.2770		v ₅	226.00	1.10	1.5551	0.2475
	v ₁	18.00	1.00	1.4064	0.2831		v ₄	31.40	2.60	1.7132	0.4066
	v ₄	9.70	1.00	1.5198	0.7351		v ₁	0.32	0.10	2.2949	2.5719
CD ₃ Cl	v ₃	15.30	1.00	1.5084	0.4997	CDCl ₃	v ₆	0.05	0.03	4.2205	5.9592
	v ₆	1.10	0.50	1.6358	1.0940		v ₃	0.30	0.20	1.2813	1.0256
	v ₂	10.70	1.00	1.7684	0.6245		v ₂	5.20	0.10	1.0554	0.4471
	v ₅	6.90	1.00	0.9723	0.2640		v ₅	157.00	7.00	1.7301	0.4563
	v ₁	11.30	0.50	1.5961	0.3439		v ₄	87.50	1.70	1.4365	0.3701
	v ₄	5.10	0.50	1.5924	0.7755		v ₁	0.06	0.04	6.4807	12.0997
					CCl ₄	v ₄	0.20	0.05	6.7870	7.4114	
						v ₃	322.00	60.00	1.6749	0.2921	

Observed intensities and their errors taken from ref. [43].

により計算した。この標準偏差を計算法の精度の目安となる計算誤差とみなした。種々の計算法の中で最適法を決めるために、この標準偏差を全振動について平均した。また、最適の計算法の選定に当たっては計算時間も評価基準の1つになるので、全分子の計算時間

(CPU時間)を平均した。このようにして求めた各計算法の標準偏差と計算時間を Table 6a, b に示す。この結果から計算法の影響について幾つかの傾向を見いだすことができる。

Table 6a. Standard deviations of frequency (SDF) and intensity (SDI) ratios and CPU times (T, sec)

	RHF			B3LYP			BLYP		
	SDF	SDI	T	SDF	SDI	T	SDF	SDI	T
6-31G	0.0543	4.9585	162.0	0.0620	0.8870	417.5	0.0758	0.8995	398.0
6-31G(d)	0.0215	2.3211	235.8	0.0261	0.9114	694.9	0.0426	1.8811	645.0
6-31G(d,p)	0.0201	2.2419	212.7	0.0246	0.8272	755.8	0.0416	1.7879	732.7
6-31G(3df,3pd)	0.0181	0.9604	2455.8	0.0214	0.4714	5241.3	0.0373	1.3340	4218.6
6-31+G	0.0541	5.2328	182.3	0.0622	1.0530	588.0	0.0763	0.9934	545.3
6-31+G(d)	0.0216	2.0680	275.3	0.0266	0.6031	962.4	0.0432	1.9392	894.4
6-31+G(d,p)	0.0197	1.9505	257.6	0.0251	0.5638	1037.2	0.0420	1.9195	985.6
6-31+G(3df,3pd)	0.0185	0.8047	3215.5	0.0218	0.7091	6508.6	0.0380	1.7851	5398.4
6-31++G	0.0537	4.9785	198.9	0.0622	0.9183	627.1	0.0767	0.9479	577.9
6-31++G(d)	0.0214	2.0093	280.5	0.0266	0.6094	1008.7	0.0432	1.9663	931.2
6-31++G(d,p)	0.0196	1.8694	262.9	0.0251	0.5754	1083.5	0.0421	2.0024	1024.4
6-31++G(3df,3pd)	0.0181	0.8122	3320.9	0.0217	0.6969	6762.8	0.0381	1.8894	5578.7
6-311G	0.0511	2.9923	226.7	0.0647	1.0823	649.6	0.0797	2.4455	591.9
6-311G(d)	0.0231	2.6583	324.7	0.0278	1.7864	1106.9	0.0441	4.4690	987.0
6-311G(d,p)	0.0207	1.8726	318.3	0.0259	1.5580	1152.7	0.0426	4.4133	1023.6
6-311G(3df,3pd)	0.0198	0.9954	3945.0	0.0225	1.3899	7014.0	0.0397	2.5214	5689.2
6-311+G	0.0480	4.2303	269.5	0.0613	0.7404	899.2	0.0772	1.2433	832.8
6-311+G(d)	0.0228	1.5954	411.2	0.0266	0.8633	1495.3	0.0428	2.8614	1319.8
6-311+G(d,p)	0.0204	1.4028	1058.6	0.0246	0.8431	1542.1	0.0416	3.0586	1369.6
6-311+G(3df,3pd)	0.0195	0.7765	5390.5	0.0224	1.0011	8661.6	0.0395	2.4934	6853.3
6-311++G	0.0476	4.3371	275.6	0.0611	0.6981	952.0	0.0760	1.1935	884.0
6-311++G(d)	0.0225	1.5250	418.5	0.0264	0.8512	1553.0	0.0423	2.8163	1374.4
6-311++G(d,p)	0.0203	1.3453	410.5	0.0245	0.8370	1614.2	0.0409	2.8303	1439.4
6-311++G(3df,3pd)	0.0195	0.7879	8851.9	0.0223	1.0253	8925.2	0.0389	2.1609	7084.0
D95	0.0556	7.1656	193.6	0.0590	5.5569	485.0	0.0680	6.2332	453.3
D95(d)	0.0216	8.3824	402.1	0.0225	4.2274	804.2	0.0371	4.4182	744.3
D95(3df,3pd)	0.0184	0.9399	3022.4	0.0211	0.4982	5870.1	0.0375	1.1316	4803.9
cc-pVDZ	0.0161	1.3587	468.6	0.0251	0.7192	1205.1	0.0381	1.8259	1148.9
cc-pVTZ	0.0187	1.0471	3016.8	0.0262	0.5896	6480.7	0.0423	1.9499	5688.1
cc-pVQZ	0.0188	0.8817	37667.9	0.0244	0.5736	40888.3	0.0409	1.8497	34250.1
AUG-cc-pVDZ	0.0159	0.9569	815.4	0.0266	0.8379	2568.7	0.0444	2.0494	2459.5
AUG-cc-pVTZ	0.0185	0.8538	15834.2	0.0270	0.8558	21269.5	0.0420	2.1714	17523.1
AUG-cc-pVQZ	0.0186	0.8231	187538.6	0.0243	0.7902	207582.3	0.0398	2.0411	141621.1

Table 6b. Standard deviations of frequency (SDF) and intensity (SDI) ratios and CPU times (T, sec)

	BPW91			MP2		
	SDF	SDI	T	SDF	SDI	T
6-31G	0.0618	0.7662	422.2	0.0631	2.0596	401.6
6-31G(d)	0.0279	1.1719	674.5	0.0119	1.2526	674.7
6-31G(d,p)	0.0274	1.0858	760.2	0.0124	1.8623	1062.7
6-31G(3df,3pd)	0.0241	0.7282	4531.1	0.0140	0.4840	23526.0
6-31+G	0.0622	0.8435	582.5	0.0627	2.0183	671.7
6-31+G(d)	0.0288	1.0948	942.1	0.0121	0.9868	1401.1
6-31+G(d,p)	0.0282	1.0756	1025.7	0.0133	0.9977	1752.7
6-31+G(3df,3pd)	0.0245	0.9432	5554.1	0.0146	0.7267	35256.2
6-31++G	0.0621	0.7912	616.6	0.0627	1.8551	693.6
6-31++G(d)	0.0287	1.0713	984.8	0.0128	0.9131	1675.7
6-31++G(d,p)	0.0282	1.0599	1079.2	0.0144	0.9577	1811.5
6-31++G(3df,3pd)	0.0244	0.9405	5751.3	0.0143	0.7240	36377.7
6-311G	0.0640	2.0806	619.0	0.0584	1.5001	998.9
6-311G(d)	0.0288	3.1829	1020.4	0.0134	1.3524	2143.1
6-311G(d,p)	0.0282	2.9490	1090.0	0.0162	2.0261	2335.0
6-311G(3df,3pd)	0.0251	1.4643	5796.5	0.0120	1.4083	39071.5
6-311+G	0.0611	0.9072	866.7	0.0543	1.3863	1842.6
6-311+G(d)	0.0276	1.6028	1385.4	0.0126	0.5928	3467.3
6-311+G(d,p)	0.0271	1.5587	1436.4	0.0160	0.7798	20654.4
6-311+G(3df,3pd)	0.0250	1.1392	6919.5	0.0122	0.9132	38668.9
6-311++G	0.0608	0.8459	921.0	0.0538	1.3970	2480.4
6-311++G(d)	0.0275	1.5724	1439.6	0.0126	0.5751	3587.5
6-311++G(d,p)	0.0270	1.5485	1512.3	0.0162	0.7331	20804.4
6-311++G(3df,3pd)	0.0250	1.1446	7191.3	0.0121	0.9301	39308.2
D95	0.0603	5.7000	479.4	0.0607	8.6427	1023.4
D95(d)	0.0252	4.2460	781.0	0.0144	5.2768	15689.8
D95(3df,3pd)	0.0244	0.5802	4904.1	0.0150	0.6629	22594.7
cc-pVDZ	0.0289	1.1041	1187.8	0.0183	1.6280	24458.2
cc-pVTZ	0.0285	1.0033	5796.2	0.0145	0.5230	31159.1
cc-pVQZ	0.0273	0.9122	33913.6	0.0147	0.5230	226014.8
AUG-cc-pVDZ	0.0302	1.0043	2557.5	0.0195	0.5668	7349.4
AUG-cc-pVTZ	0.0292	1.1123	17806.9	0.0114	0.7771	76835.4
AUG-cc-pVQZ	0.0270	0.9855	138130.3	0.0143	0.3287	549134.6

まず第1は振動数の計算誤差に対する基底関数の影響である。たとえば、計算法 RHF において基底関数を 6-31G、6-31G(d)、6-31G(d,p)、6-31G(3df,3pd) と変えると計算誤差は 0.0543、0.0215、0.0201、0.0181 と大幅に減少する。これに対して、同じ RHF において基底関数を 6-31G、6-31+G、6-31++G、6-311G、6-311+G、6-311++G と変えても計算誤差は 0.0543、

0.0541、0.0537、0.0511、0.0480、0.0476 と変化は小さい。これは上記のように、本研究の塩素原子を含む分子では分極関数の影響が大きいが、diffuse 関数の影響は小さいことを示している。

また、電子相関の効果として RHF を MP2 に変えると計算誤差は全般的にはほぼ半減しているが、分極関数を含まない 6-31G、6-31+G、6-31++G、6-311G、

6-311+G、6-311++G では RHF と MP2 とで計算誤差が減少せず、むしろ MP2 の方が若干ではあるが誤差が大きい。これは電子相関の補正の効果を引き出すためには分極関数を取り入れた基底関数を用いる必要があることを示しているが、MP2 で誤差が増えた原因は不明である。

第 2 点は振動数の計算誤差に対する計算法の影響である。Table 6a, b の結果から各計算法における計算誤差の平均を算出すると、RHF では 0.0266、B3LYP では 0.0325、BLYP では 0.0482、BPW91 では 0.0344、MP2 では 0.0237 となる。この結果から、まず本研究の場合には RHF が DFT 法より良い精度を与えることが分かる。多数の分子についての振動数の計算結果から、RHF が最も精度が悪く、MP2 がそれに次ぎ、DFT が最も良い精度であるとの報告がある [49]。彼らと我々では分子の数と種類が違うので結果の違いは当然であるが、上記のように DFT 法は塩素原子の関与する振動の振動数を低く与える傾向があり、これが誤差を大きくしている。

また、MP2 は RHF より計算時間がほぼ 2 倍程度かかるのに計算精度はそれほど向上しないことが分かる。さらに、DFT 法の中ではこれまで B3LYP が比較的よい結果を与えるとされているが、本研究の結果では BPW91 も B3LYP に匹敵する誤差になった。この結果は文献 [52] に示されている B3LYP と B3PW91 が最も誤差が小さいという結果と対応している。

第 3 点は計算精度と計算時間の関係であり、全般的に基底関数や計算法の精度を上げれば、計算時間は長くなるが振動数の計算誤差は減少するという傾向が認められる。しかし、大幅な計算時間がかかる cc-pVQZ、AUG-cc-pVTZ、AUG-cc-pVQZ などの基底関数を用いた場合に必ずしも最高の計算精度が得られるとは限らない点である。Table 6a, b の結果において、各計算法の中で最高の計算精度が得られた基底関数は RHF では AUG-cc-pVDZ、B3LYP では D95(3df,3pd)、BLYP では D95(d)、BPW91 では 6-31G(3df,3pd)、MP2 では AUG-cc-pVTZ であり、これらの基底関数の場合の計算時間は各計算法の中で最長ではない。したがって、計算精度が最高で、かつ計算時間も適切な最適な計算法が存在することが分かる。

第 4 点は赤外強度に関する全般的傾向として、上記の振動数に関する傾向ほど明確ではないが、ほぼ同様の傾向が認められる。この場合に特筆すべき点として、計算法 B3LYP、BPW91 では基底関数 6-31G(3pd,3df) との組み合わせにおいて赤外強度の計算誤差が最小と

なっている。振動数の再現においては DFT 法、特に B3LYP がよい結果を与えることはこれまでの研究から判明しているが、赤外強度の再現に関しては種々の計算法を比較した研究はこれまで皆無である。本研究では塩素化メタン分子しか扱っていないので、一般の分子についてこの結論があてはまるかどうかは不明であるが、比較的計算時間の短い基底関数 6-31G(3pd,3df) と B3LYP、BPW91 の組み合わせが赤外強度の計算法として最適であることは興味深く、他の多くの分子についての検討が待たれる。

Table 6a, b の結果から最適の計算法を決定することができるが、計算の目的によって最適の計算法を選定する基準が異なる。赤外スペクトルを理論計算する目的はいろいろあるが、実測の赤外スペクトルに現れているピークがどの振動モードによるかを定める振動の帰属を行う場合には、振動数の計算値が実測値にできるだけ近く、計算値の絶対的な精度が高い計算法であることが要求される。このような場合は一般に少数の分子について計算を行うことが多いので計算時間はあまり問題にならない。

そこで、この観点から Table 6a, b の結果の中から標準偏差の最も小さい計算法を探すと、振動数に対しては AUG-cc-pVTZ の基底関数を用いて MP2 計算を行うと標準偏差 0.0114、すなわち相対誤差 1.14 % で実測振動数が再現できることになる。この相対誤差 1.14 % という数値は C-H 伸縮振動が現れる 3000 cm^{-1} 領域において平均 34 cm^{-1} の計算誤差を与えることを意味するので、実測の赤外ピークの帰属には絶対的な精度として満足できる計算法とはいえない。

また、赤外強度に対しては AUG-cc-pVQZ の基底関数を用いて MP2 計算を行うと標準偏差 0.3287、すなわち相対誤差 32.87 % で実測強度が再現できることになる。赤外強度の測定誤差は実測強度の通常 10 % 程度なので、強度の計算誤差 32.87 % も絶対的な誤差としては必ずしも満足できる値ではない。したがって、このような実測ピークの帰属を行う場合には、これらの計算法を用いて振動数と赤外強度を計算し、本研究で行ったように振動数だけでなく赤外強度の計算値も考慮に入れて慎重に帰属を行うことが必要であると結論できる。

以上の場合には計算値の絶対的な精度の高さが要求されるので、現状の計算法でこのような問題を解決することは難題である。しかし、分子の環境の変化などに由来する実測スペクトルの変化を解析する場合には、振動数や赤外強度の相対的な変化が問題となるので、

上記ほどの精度の高さは必要ない。たとえば、孤立分子と分子錯体について振動数を分子軌道法で計算し、マトリックス単離法で観測した赤外スペクトルとの比較から、分子錯体の存在およびその構造を推定した研究がある [53–55]。このような問題の場合には振動数の絶対値は問題でなく、相対的なシフトが計算で説明できればよいので、上記の計算法で十分解決可能と考えられる。

赤外スペクトルを理論計算する 3 番目の目的として、緒言で記した未知化学種の同定がある。この場合には無数の化学種の中から実測の赤外スペクトルに一致する計算スペクトルを示す化学種を探索することが必要となる。スペクトルの一致は振動数や赤外強度の数値ではなくスペクトルの図形として判断するので、計算精度はあまり問題にならない。本研究で検討した計算法の中で最も近似の低い計算法は基底関数 6-31G と計算法 RHF の組み合わせであるが、この場合の振動数の計算誤差 5.43 % はスペクトル図形の再現には十分な精度である。

一方、赤外強度の計算誤差 495.85 % は測定誤差と比較すると満足できる精度ではない。しかし、未知化学種の赤外スペクトルを測定する場合、スペクトル中の最大ピークが通常 5 % 程度の透過率になるようセルの厚さを調節して測定が行われるので、その絶対的な強度は問題にならない。したがって、このような目的では相対的な強度が再現できれば十分であり、この点で上記の計算誤差はこの目的には十分な計算精度である。むしろ、この場合は莫大な数の化学種について赤外スペクトルを計算する必要があるため、最大の選定基準は計算時間である。上記の計算法での平均計算時間は 162.0 秒であり、多数の化学種の赤外スペクトルを計算できる方法とはいえない。この目的には同程度の計算精度でかつ計算時間が圧倒的に短い計算法の開発が望まれる。

4.3 理論的方法の問題点

さらに、実測の赤外スペクトルを再現するための理論計算には問題が山積している。第 1 はスペクトル中のピークの半値幅を計算できる理論が現在まで開発されていないことである。分子軌道法で計算される赤外強度は実測スペクトルではピークの面積強度に対応する。現実の分子の赤外スペクトルでは一般にピークの半値幅は振動により大きく異なる。したがって、実測の赤外スペクトルを再現するためにはピークの半値幅

を理論計算する必要がある。しかし、従来の赤外スペクトルの半値幅の理論では分子は剛体として取り扱われているため、ピークによる半値幅の違いを説明できない。

第 2 の問題は赤外スペクトルの測定が最も多く行われる液体や固体状態のスペクトルを再現するための溶媒効果の理論が不完全なことである。我々の報告 [44] にあるように、ジクロロメタン分子は水素結合のような強い分子間相互作用をすることは考えられないが、気体と液体の赤外スペクトルには大きな変化が観測される。このような溶媒効果を説明する理論として古典的な誘電体モデルがあり、それに基づいて振動数や赤外強度の溶媒効果の計算法として、Onsager Reaction Field Model、Self Consistent Isodensity Polarized Continuum Model などが開発され、GAUSSIAN プログラムに組み込まれている [45]。しかし、これらの理論モデルにおいても分子の内部振動が考慮されていないため、ジクロロメタンで観測されるピーク、すなわち、振動による溶媒効果の違いが説明できない。

第 3 は実測の赤外スペクトルには倍音や結合音がかなりの強度で現れたり、また、フェルミ共鳴の影響により赤外強度が理論値から大きく変動することが多い。しかし、これら倍音、結合音、フェルミ共鳴を含めて赤外スペクトルを再現するための理論が不完全である。

第 4 は、未知化学種の同定においては実測の赤外スペクトルを与えるような候補分子について赤外スペクトルを理論計算する必要があるが、現状では無数の化学種の中から候補分子を探索する方法がないことである。このような諸々の問題を考えると、緒言で記したような理論的方法によって赤外スペクトルから未知化学種を同定することは現状ではきわめて困難であると結論できる。

5 考察 各種の方法の比較

本研究では、コンピュータを用いて赤外スペクトルから化学構造を推定する方法として、スペクトルデータベースを利用する帰納的方法、線形学習機械やニューラルネットワークを利用する経験的方法、非経験的分子軌道計算を利用する理論的方法、など様々な方法を検討した。しかし、どの方法を用いても、現状のコンピュータの能力では赤外スペクトルから化学構造を推定するには限界があることが分かった。

まず、帰納的方法の限界の原因はデータベースの規模を無限大ではなくとも実用的規模にまで拡大することの不可能さであり、如何なる未知化合物でも検索できるデータベースを構築することが未来永劫に不可能であることは原理的に明白である。また、経験的方法の限界の原因は、赤外スペクトルのパターンと化学構造の関係の複雑さにある。この複雑さには、分子振動の中で赤外活性を支配する選択律の問題だけでなく、分子の非調和性やフェルミ共鳴も現実の赤外スペクトルパターンには大きな影響を与えており、測定で得られた赤外スペクトルパターンから経験的な規則によって化学構造を推定することがやはり原理的にきわめて困難であることに起因している。

この点で赤外スペクトルからの構造推定手段として今後最も期待できるのが非経験的分子軌道計算などを利用する理論的方法であり、コンピュータの演算能力が飛躍的に向上する将来においてはこの方法が主流になると思われる。しかし、測定された未知化合物の赤外スペクトルに対して、無限大の集合の化合物について理論スペクトルを計算することは将来においても不可能であり、何らかの方法による候補化合物の絞り込みが必要になると思われる。

そのための手段として、データベースを利用する帰納的方法や経験則を利用する経験的方法等との組み合わせが考えられる。すなわち、未知化合物の実測赤外スペクトルをデータベースで検索して出力された類似化合物の共通構造を調べて、未知化合物の官能基を拾い上げる。さらに、未知化合物の実測赤外スペクトルに対して経験則やニューラルネットワークを適用して含まれる官能基を拾い上げる。両方の結果を総合して、未知化合物の構造を組み立て、候補化合物を何組か用意する。これらの候補化合物について非経験的分子軌道法などにより赤外スペクトルを理論計算し、実測スペクトルを最もよく再現するものを正解とする、という方法である。

このような帰納的方法、経験的方法、理論的方法など各種の方法を組み合わせた赤外スペクトルの自動解析法が実現する将来には、赤外スペクトル測定の汎用性と相俟って、赤外スペクトルが NMR スペクトルと並んで、あるいはそれ以上に多くの問題解決に貢献する時代が到来すると思われる。

6 結言

我々は 30 年以上に亘る一連の研究の中で、スペクトルからの構造推定手段として、NMR スペクトルに対して赤外スペクトルの復権を何とかして図りたいという思いで、様々な方法を検討してきた。しかし、帰納的方法、経験的方法、理論的方法、どの方法を用いても、現状のコンピュータの能力では赤外スペクトルから構造を推定するには限界があることが分かった。いつの日か我々の夢が実現する時代が来ることを期待する。

一連の本研究を行うにあたり、SDBS の構築やニューラルネットワークシステムの評価などに貢献していただいた沢山の方々に感謝する。

参考文献

- [1] N. B. Colthup, *J. Opt. Soc. Am.*, **40**, 39 (1950).
- [2] O. Yamamoto, K. Someno, N. Wasada, J. Hiraishi, K. Hayamizu, K. Tanabe, T. Tamura, M. Yanagisawa, *Anal. Sci.*, **4**, 233 (1988).
- [3] 田辺和俊, 佐伯慎之助, 田村禎夫, *分析化学*, **23**, 626 (1974).
- [4] K. Tanabe, S. Saeki, *Anal. Chem.*, **47**, 118 (1975).
- [5] K. Tanabe, T. Tamura, J. Hiraishi, S. Saeki, *Anal. Chim. Acta*, **112**, 211 (1979).
- [6] 田村禎夫, 平石次郎, 田辺和俊, 佐伯慎之助, *分析化学*, **28**, 31 (1979).
- [7] 田村禎夫, 田辺和俊, 平石次郎, 佐伯慎之助, *分析化学*, **28**, 591 (1979).
- [8] 田村禎夫, 田辺和俊, 平石次郎, 佐伯慎之助, *分析化学*, **29**, 279 (1980).
- [9] 田村禎夫, 田辺和俊, 平石次郎, 佐伯慎之助, *分析化学*, **30**, 439 (1981).
- [10] S. Saeki, K. Tanabe, T. Tamura, M. Tasumi, I. Suzuki, *Appl. Spectrosc.*, **35**, 148 (1982).
- [11] K. Tanabe, T. Tamura, J. Hiraishi, *Bunseki Kagaku*, **31**, E27 (1982).

- [12] K. Tanabe, T. Tamura, J. Hiraishi, S. Saeki, I. Suzuki, M. Tasumi, *Bunseki Kagaku*, **31**, E177 (1982).
- [13] 田辺和俊, 田村禎夫, 平石次郎, 佐伯慎之助, 鈴木功, 田隅三生, *分光研究*, **32**, 241 (1983).
- [14] K. Tanabe, J. Hiraishi, S. Saeki, I. Suzuki, M. Tasumi, *Computer Enhanced Spectrosc.*, **2**, 13 (1984).
- [15] K. Tanabe, J. Hiraishi, T. Tamura, O. Yamamoto, M. Yanagisawa, N. Wasada, *Computer Enhanced Spectrosc.*, **2**, 97 (1984).
- [16] S. Saeki, K. Tanabe, *Appl. Spectrosc.*, **38**, 693 (1984).
- [17] 田辺和俊, 平石次郎, 田村禎夫, 山本修, 柳沢勝, 和佐田宣英, *分析化学*, **33**, 95 (1984).
- [18] 田辺和俊, 平石次郎, 田村禎夫, 山本修, 柳沢勝, 和佐田宣英, *分析化学*, **33**, 285 (1984).
- [19] D. C. Hassel, E. M. Bowman, *Appl. Spectrosc.*, **52**, 18A (1998).
- [20] 田辺和俊, 田村禎夫, 都築誠二, *分析化学*, **35**, 518 (1986).
- [21] 田辺和俊, 田村禎夫, 都築誠二, *分析化学*, **36**, 339 (1987).
- [22] 田辺和俊, 田村禎夫, *Chem. Express*, **7**, 549 (1992).
- [23] 田辺和俊, 田村禎夫, *分析化学*, **41**, 515 (1992).
- [24] 田辺和俊, 長塚義隆 (訳), *化学者のためのニューラルネットワーク入門*, 丸善 (1996).
J. Zupan, J. Gasteiger, *Neural Networks for Chemists*, VCH (1993).
- [25] 田辺和俊, *Neurosim/L* によるニューラルネットワーク入門, 日刊工業新聞社 (2003).
- [26] K. Tanabe, T. Tamura, H. Uesaka, *Applied Spectrosc.*, **46**, 807 (1992).
- [27] 田辺和俊, 田村禎夫, 上坂博亨, *J. Chem. Software*, **1**, 15 (1992).
- [28] K. Tanabe, H. Uesaka, *Computer-Aided Innovation of New Materials*, **2**, 943 (1993).
- [29] 田辺和俊, 田村禎夫, 上坂博亨, *Chem. Express*, **8**, 117 (1993).
- [30] 田辺和俊, 上坂博亨, 井上常史, 高橋博之, 田中壮一郎, *分析化学*, **43**, 765 (1994).
- [31] 田辺和俊, 上坂博亨, 福士真, *化学とソフトウェア*, **17**, 89 (1995).
- [32] 田辺和俊, 上坂博亨, 相川克明, 天野敏男, *化学とソフトウェア*, **17**, 127 (1995).
- [33] 田辺和俊, 上坂博亨, 南幸男, *化学とソフトウェア*, **19**, 3 (1997).
- [34] 松本高利, 田辺和俊, 佐伯和光, 天野敏男, 上坂博亨, *分析化学*, **48**, 483 (1999).
- [35] 松本高利, 田辺和俊, 上坂博亨, *化学とソフトウェア*, **21**, 53 (1999).
- [36] E. W. Robb, M. E. Munk, *Mikrochim. Acta [Wien]*, **1**, 131 (1990).
- [37] M. E. Munk, M. S. Madison, E. W. Robb, *Mikrochim. Acta [Wien]*, **2**, 505 (1991).
- [38] R. J. Fessenden, L. Gyoegy, *J. Chem. Soc. Perkin Trans.*, **2**, 1755 (1991).
- [39] D. Ricard, C. Cachet, D. Cabrol-Bass, T. P. Forrest, *J. Chem. Inf. Comput. Sci.*, **33**, 202 (1993).
- [40] C. Klawun, C. L. Wilkins, *J. Chem. Inf. Comput. Sci.*, **36**, 69 (1996).
- [41] C. Sosa, H. B. Schlegel, *J. Chem. Phys.*, **86**, 6937 (1987).
- [42] G. L. Fox, H. B. Schlegel, *J. Chem. Phys.*, **92**, 4351 (1990).
- [43] K. Tanabe, *Spectrochim. Acta*, **31A**, 1611 (1975).
- [44] S. Saeki, K. Tanabe, *Spectrochim. Acta*, **25A**, 1325 (1969).

- [45] M. J. Frisch, G. W. Trucks, H. B. Schlegel, P. M. W. Gill, B. G. Johnson, M. A. Robb, J. R. Cheeseman, T. Keith, G. A. Petersson, J. A. Montgomery, K. Raghavachari, M. A. Al-Laham, V. G. Zakrzewski, J. V. Ortiz, J. B. Foresman, J. Cioslowski, B. B. Stefanov, A. Nanayakkara, M. Challacombe, C. Y. Peng, P. Y. Ayala, W. Chen, M. W. Wong, J. L. Andres, E. S. Replogle, R. Gomperts, R. L. Martin, D. J. Fox, J. S. Binkley, D. J. Defrees, J. Baker, J. P. Stewart, M. Head-Gordon, C. Gonzalez, and J. A. Pople, *GAUSSIAN 94 Revision D.4*, Gaussian, Inc., Pittsburgh, PA (1994).
- [46] T. Shimanouchi, *Tables of Molecular Vibrational Frequencies, Consolidated Volume I, Nat. Stand. Ref. Data Ser., Nat. Bur. Stand. (US)*, **39** (1972).
- [47] Appendix Tables 1-13 は <http://sccj.net/publications/JCCJ/v4n1/a01/appendix.html> からダウンロードできる。
- [48] J. B. Foresman, A. Frisch, *Exploring Chemistry with Electronic Structure Methods*, 2nd Ed., Gaussian Inc., Pittsburgh, PA (1996), p.64.
- [49] A. P. Scott, L. Radom, *J. Phys. Chem.*, **100**, 16502 (1996).
- [50] S. Kudoh, M. Takayanagi, M. Nakata, *Chem. Phys. Lett.*, **322**, 363 (2000).
- [51] H. Yoshida, A. Ehara, H. Matsuura, *Chem. Phys. Lett.*, **325**, 477 (2000).
- [52] N. Akai, S. Kudoh, M. Takayanagi, M. Nakata, *J. Photochem. Photobiol., A*, **146**, 49 (2001).
- [53] E. D. Jemmis, K. T. Giju, K. Sundrarajan, K. Sankaran, V. Vidya, K. S. Viswanathan, J. Leszczynski, *J. Mol. Struct.*, **510**, 59 (1999).
- [54] S. Kudoh, K. Onoda, M. Takayanagi, M. Nakata, *J. Mol. Struct.*, **524**, 61 (2000).
- [55] F. Ito, T. Nakanaga, Y. Futami, S. Kudoh, M. Takayanagi, M. Nakata, *Chem. Phys. Lett.*, **343**, 185 (2001).

Computer Estimation of Chemical Structures from Infrared Spectra

Kazutoshi TANABE^{a*}, Takatoshi MATSUMOTO^b, Shoji ITOH^c, Hiroyuki UESAKA^d,
Seiji TSUZUKI^e, Tadao TAMURA^e, Shinnosuke SAEKI^e and Shuichiro ONO^a

^aFaculty of Social System Science, Chiba Institute of Technology
Tsudanuma 2-17-1, Narashino 275-0016, Japan

^bInstitute of Multidisciplinary Research for Advanced Materials, Tohoku University
2-1-1 Katahira, Aoba-ku, Sendai 305-8577, Japan

^cInstitute of Engineering, University of Tsukuba
Tennoudai 1-1-1, Tsukuba 305-0006, Japan

^dFaculty of Regional Science, Toyama University of International Studies
Higashikuromaki 65-1, Ohshima-cho, Kamishinkawa-gun, Toyama 930-1292, Japan

^eNational Institute of Advanced Industrial Science and Technology
1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, Japan

**e-mail: tanabe@pf.it-chiba.ac.jp*

Various computer-aided methods are used to estimate chemical structures from infrared spectra: an inductive method using a spectral database, an empirical method using a knowledge base, and a theoretical method using computational chemistry. With an aim to increase an ability for estimating structures from infrared spectra, various computer methods were intensively investigated. At first an inductive method using a spectral database was studied, and several problems were pointed out. Next a linear learning machine and a neural network were applied to estimating functional groups from infrared spectra as an empirical method, and several problems were also mentioned. Third a non-empirical molecular orbital calculation as a theoretical method was applied to examine the agreement between calculated and experimental vibrational frequencies and infrared intensities, and several problems were also underlined. It was concluded that there are many difficulties in estimating chemical structures from infrared spectra with the power of existing computers.

Keywords: Infrared spectra, Structure estimation, Spectral database, Neural network, Non-empirical molecular orbital calculation