【解説】

天然有機化合物の NMR データ ベース “CH-NMR-NP”

早水紀久子

天然有機化合物の化学構造の決定には ^1H と ^{13}C -NMR スペクトルが重要な役割を演じ、近年は新規化合物の化学構造と NMR スペクトルが一体的に発表されることが多い。新規化合物、既知化合物、合成化合物の化学構造の考察には NMR データが重要な指標になっている。ここでは、天然有機化合物を対象にして世界で初めて構築された ^1H と ^{13}C -NMR のデータベース “CH-NMR-NP” (収録化合物数・約 22,100 件, 2011 年 1 月末) について、NMR の提供する化学構造情報ならびに既知の天然有機化合物の NMR データを短時間、総合的に検索入手できるシステムを紹介する。

天然有機化合物を構成する主要元素は炭素、水素、酸素、窒素であるにもかかわらず、多種多様な化学構造をもち、化学構造が決定されている天然有機化合物の数は 20 万件を優に超えている。その他の元素として、硫黄、リン、ハロゲンなどを含有する化合物も知られている。ここでは、天然有機化合物の NMR データベースについて述べる。一般有機化合物の NMR データベースは独立行政法人産業技術総合研究所 (産総研) の “SDBS-NMR”, バイオラッド社の KnowItAll, ACD のシフト予測システムに含まれる ^1H と ^{13}C -NMR データベース, CSEARCH NMR に含まれる ^{13}C -NMR シフト値な

どがあげられる。また化学構造式作画ソフトの ChemDraw では ^1H と ^{13}C のシフト値を推算するシステムが含まれている。推算システムは実測データに基づいて化学シフト値を算出している。類型的な化学構造については推算値の確度は良好であるが、特異な構造については信頼性は低い。データベースでは NMR スペクトルの測定値に基づいた生データを集積しているので、推算システムと似ているように見えるが同じものではない。

現在、有機天然化合物の構造決定には ^1H と ^{13}C -NMR が重要な役割を演じており、新規化合物の論文発表には必ず NMR データが記述されている。高磁場化に伴う NMR 装置の性能向上と構造決定のための二次元 NMR、特に COSY(DQF-COSY), HSQC, HMBC, NOESY などが日常的に測定され、解析のためのツールも利用できるようになったことから、スペクトル解析の信頼性が著しく向上している。アメリカ化学会の *Journal of Natural Products* 誌では Supporting Information に NMR のチャートを載せることを要求している。有機天然化合物研究において NMR スペクトルの重要性は強調する必要もないであろう。天然物研究で重要な役割を果たしている “Dictionary of Natural Products (DNP)” (CRC 社) は長い歴史があり、2010 年末で 234,000 件の化合物が収録されているが、NMR データは含まれていない。筆者

On an NMR Database for Natural Products “CH-NMR-NP”
Kikuko HAYAMIZU, (有)NMRDBTech

らは2001年5月から、天然有機化合物の ^1H と ^{13}C スペクトルを収録するNMRデータベース“CH-NMR-NP”の構築に取り掛かった。このプロジェクトのベースはSDBS-NMR構築の経験である。SDBS-NMRは1975年頃から市販試薬を中心に ^1H -NMR, ^{13}C -NMR, MS, IRスペクトルを測定して大型汎用計算機(MSP)にデータベースを構築するシステムとしてスタートした。MSP廃止後はパーソナルコンピュータ(PC)を用いてデータ追加を行ない、Unixでサービスするシステムに変換した。一方、天然有機化合物は分離精製した少量のサンプルのNMR測定から構造決定しているので、そのスペクトルは世界で唯一であり、しかも1回だけ測定されたものである可能性が高い。活性が高い天然物の場合には全合成が試みられ、構造確認にはNMRスペクトルが使われている。立体構造の確認あるいは訂正もあり、その趣旨の論文発表も行なわれている。言いかえると、天然有機化合物のNMRデータは文献からのみ収録することができる。

CH-NMR-NPはSDBS-NMR(2000年3月までに収録したデータをPCフォーマットに変換し、その著作権の50%を著者が保有する)のデータのうち天然物に関連する化合物のNMRデータ926件の著作権を産総研からスペクトルパターンを含めて買い取って含めてあるが、その他はすべて研究論文から収集したデータである。CH-NMR-NPについて、文献検索すればNMRデータを取得できるのでデータベースは不要であるという議論があり、文献検索のプロセスが学問を遂行するうえで重要であるという考え方もあることも確かである。しかし実際問題として、化合物検索から信頼性の高いNMRデータに到達するために費やす時間を考えてみよう。また、研究者(あるいは研究グループ)が対象とする化合物の種類は特定されているので、そのような範疇のデータだけを集めておけばよいという考え方もある。この場合は、予想を越えた化合物に出合ったときにはその重要性を見逃す可能性もある。

CH-NMR-NPでは信頼性の高いジャーナルに限定して、NMRスペクトルの質が確保できると判断した化合物を収録している。逆に言えば、NMRの質が悪い場合には決定された化学構造の信頼性も低い可能性がある。天然物研究の分野でのイソプレノイド、アルカロイドなどといった化合物区分を意識していないのは、もともと天然物の世界と馴染みが少なかった筆者の弱点であり、逆に推進力になっている。天然有機化合物討論会で4回にわたってCH-NMR-NPについて発表し、天然物研究分野でのコメントをいただいている^(1~4)。また情報化学

討論会ではシステム構築について発表している^(5~7)。

データベース設計の考え方

化学構造の作図と部分構造検索ができることが有機化合物のデータベースでは必須の条件である。筆者が従事していたSDBS-NMRは1999年3月末でのMSPの終了に伴ってデータベース構築システムをPCに新たに構築する必要に迫られた。この時点でPCをベースとした化合物データベース構築システムとして、MDL社のISIS/Baseを採用し、SDBS-NMRの構築システムを開発した⁽⁸⁾。2001年に筆者が天然物のデータベースを新規に計画したときに、コストとそれまでの経験を踏まえてISIS/Base2.4を採用した。PCソフト業界の再編成はめまぐるしく、今現在調べたところ、Excel 2007を駆動できるISIS/Base2.5はAccelrys社から販売されているようである。2010年に新規にデータベースを作成したいと計画するならばもっと広範囲のソフトからの選択が可能と考えている。ISIS/Baseは化学構造式をコンピュータで取り扱うロジック開発で先駆的な役割を果たし、化学構造式のモルファイル(MolFile)を作成した。MolFileとは、化学構造を原子種、位置座標、化学結合などに関する情報で記述したものであり、化学構造式を扱うPCソフトではすべてMolFileを基準とし、個別のシステムへのMolFileからの変換が可能になっている。NMRのデータベースではスペクトルの帰属を構造式に書き込むことが必要であるが、その点でもISIS/Drawは好都合である。

ISIS/Baseでのデータベースシステム構築は任意の大きさの箱を並べて個々の箱の定義をすることで可能なために、PC画面で視覚的にデータベースを作成できる。例を図1に示す。項目としては、化合物のNP番号(NO)、NMRスペクトルキー(CHS)、CHSは溶媒や温度、keto-enol互変異性体、conformerなど、同じNOに対して複数のCHSが存在する場合がある。帰属の入った化学構造式(Structure)が本データベースを特徴づける。帰属として炭素記号の代わりに数字を置くが、受け入れられるのは数字だけであるために、縮合環の4a, 8aなどは40, 80と置き換えている。また1', 2'などはやはり記述できないので、31, 32などと表記している。番号は論文準拠であり、 ^{13}C シフトの記述を中心にするために、化合物記述のルールには必ずしも従っていないので、正式な化合物名に与えられている番号と不一致のことも多い。化学シフト値と1:1に対応させるために、等価な炭素でない限り、すべての炭素に異なった番号を割り振っ

ID 181	CHS 24-181	NO 24181	CAS_NO
NAME			CRC Da2010/11/11
tumescenamide A			Source Streptomyces tumescens YM23-260
Structure

Data		MF C ₃₇ H ₅₇ N ₅ O ₈	MW 699.90
1,171.3,0,,,		Instrument	Ref
2,55.5,1,4.92,d 9.0 3.2,,		Solvent acetone-d ₆	500 RF
3,75.1,1,5.15,dq 3.2 6.1,,		Motohashi, K., et al, J. Antibiotics 63, 549 (2010)	
4,16.2,3,1.34,d 6.4,,		Reference	
,,7.04,d 9.0 NH,,		Seto, H. CA	haseto@nodai.ac.jp Mail
11,165.3,0,,,		Compound peptide	
12,128.0,0,,,		fermentation broth of a marine bacterium	
13,134.8,1,6.57,q 7.1,,		Remarks	
14,14.7,3,1.67,d 7.1,,		cyclo(Thr / alpha-amino-2-butenic acid / Tyr / Val / Leu)	
,,7.98,s NH,,		fatty acid	
21,169.5,0,,,			
22,56.7,1,4.45,m,,			
23,36.6,2,2.99,dd 13.7			
13.7,,			
2,29,4,2,4,12,0			

図1 ■ ISIS / Base で作成した
“CH-NMR-NP” の例

ている。

名前 (Name) には Trivial Name が付いてある場合は優先的に入力し、正式な化合物名が付記してある場合は ChemName の項目にも入力する。化学構造式だけで Name が与えられていない場合の採用率は低い。化学名でギリシャ文字は重要であるが、データベースでは非常に厄介な問題である。本システムでは α , β , δ は alpha, beta, gamma と置き換え、糖の結合位置を示す “→” は単に “-” で置き換えている。化学構造式から自動発生される分子式 (MF) と分子量 (MW) の項目がある。分子式は文献中の分子式と必ず照合しているが、文献に誤りがないとは言えない。分子量は産地 (同位元素比が同一でないなど) と測定精度によって変動する。化合物の分類 (Compound) は文献準拠であり、天然物有機化学を参照していない。原産物の学術名 (Source) とその特徴 (Remarks) を入力している。方式は論文に準拠しているので、データベース内での統一性はとっていない。NMR 測定条件に関する情報は、溶媒と温度 (Solvent) (室温付近の場合は省略), ¹H 周波数 (RF), シフト基準 (Ref),

測定装置 (非公開) である。引用文献名 (Reference) と論文代表者 (非公開) と E-mail アドレス (非公開) を含めてある。立体構造などの修正論文を見つけた場合にはデータベースを修正し、その論文名も同時に記載している。CAS 登録番号と DNP の番号 (CRC) (非公開) を徐々に取り込んでいる。ISIS/Base には単純な検索機能があるので、データベース構築上の重複チェックや統計などに便利に利用している。また Excel へエクスポートできるので、Excel の機能を利用してデータのチェックを行なっている。

NMR の帰属とシフト情報は、Excel で作成して ISIS/Base に貼りつけている。NMR 情報は帰属のための番号、対応する ¹³C シフト値、その炭素に結合している水素数 (0~3), ¹H シフト値、¹H の分裂情報と J 値、¹³C シフトの帰属が入れ替わる可能性がある場合のマーク、等価な炭素数である。

Input Search Conditions

Basic Properties

NP No.

Name

MF(C)

MF(H)

MF(N)

MF(O)

MF(others) ☐ Regular Expression

MW

Substructure

File Edit View Atom Bond Tools Templates Help

C H O N P S F Cl Br I +1 -1

Chemical Shifts (ppm)

C H

C H

C H

C H

C (allowance) H (allowance)

Number of the input box

Display results on each page

図2 ■ “CH-NMR-NP” の検索画面

図3 ■ NMR スペクトルの表示画面

データベースの検索・表示システム

パーソナルな目的でのデータベースであれば、ISIS/Base は文字情報や数値、部分化学構造の検索もできる

ので便利に使えるが、多数のユーザが共用したシステムにするためには、一般的なデータベース管理システム (DBMS) を使う必要がある。全世界から多数のアクセスを受けている SDBS-NMR では、筆者は Oracle を

DBMS として採用したが、コストを考えれば CH-NMR-NP で Oracle は利用できない。筆者らはオープンソースで Windows に対応する PostgreSQL を DBMS として利用することとした。英文字、数字および制御文字 (ASCII データ) の検索ができる汎用性の高い DBMS である。名前、分子式、化合物に関する分子量などの文字・数字データの検索と ^1H と ^{13}C の化学シフトの検索ができる。

化学構造の取り扱い方は次のようにした。システムを簡単化するために、帰属のついた化学構造式はイメージデータに変換し、表示用の Drawing のソフトを不要とした。検索用の化学構造は、帰属を取り除いた MolFile データ (原子種、位置座標、化学結合に関する情報) から検索用の英数字を基本にした文字列データ (SMILES) に変換して上述の DBMS に格納する。現在の SMILES には立体構造の記述はないので、立体構造を含んだデータは含まれていない。検索用の化学構造式の Drawing はフリーのソフト JChemPaint を使い、検索化合物のデータを SMILES に変換してから検索する。検索用の画面を図 2 に示す。検索は名前、原子数、分子式、分子量、 ^{13}C シフト、 ^1H シフト、NP 番号、部分化学構造式で可能であり、数値データは範囲で検索できる。ロジックはすべて “and” である。スペクトルの表示画面の例を図 3 に示す。左側の文字情報では、化合物名、分子式、分子量、化学名、分類などの化学情報、由来植物や動物名 (学術名と根、茎、葉などの特徴づける記述、培養など)、NMR 測定条件 (溶媒、 ^1H 周波数、シフト基準) などが表示される。CH-NMR-NP には、毒性を含む活性、利用可能性、融点、UV、IR 情報、旋光度、サンプルの形状 (結晶、粉末、液体など) などは収録されていない。

DNP にこれらの項目が記述されている。

^{13}C スペクトルパターンは、等価な炭素数を基準にした強度でシフト値の位置に棒グラフで表わした。帰属用の炭素番号を各線の上部に示してある。結合している水素の数に応じて S, D, T, Q をクリックすると、対応するスペクトル線を選択できる。DEPT (Distortionless Enhancement by Polarization Transfer. 炭素シグナルを CH_3 , CH_2 , CH , C に区別して測定できる) と類似した機能である。 ^{13}C と ^1H のシフト値データ、 ^1H の分裂パターンは右欄に示してある。番号は化学構造式の番号ならびにスペクトル図の番号と一致している。

天然物研究とデータベース

新規の天然物を発表した研究論文には、分子式、分子量、化学構造の他に、NMR データとスペクトル帰属が必ず記述されている。CH-NMR-NP では *Chem. Pharm. Bull.*, *J. Antibiot.*, *J. Nat. Prod.*, *Tetrahedron*, *Tetrahedron Lett.*, *Phytochemistry*, *Phytochem. Lett.*, *Org. Lett.*, *J. Org. Chem.*, *Eur. J. Org. Chem.* を中心に、2000 年以降の論文からデータを集めている。 ^{13}C シフト値はすべての炭素に付与されていること、 ^1H データが含まれていること、確からしい帰属があることが化合物採用の条件である。置換基が類似して骨格構造のシフト値にわずかな影響しか与えないような化合物の重複採用はしないようにしている。立体構造が異なる場合は記載されているすべての化合物を入力している。帰属の入れ替わりは、気がつけば正しい帰属に修正している。合成の論文に引用されている 1990 年代の論文や *JACS* (*J. Am. Chem. Soc.*) など上記以外のジャーナルの論文からも少

図 4 ■ 炭素数による化合物数の分布

図5 ■ 研究代表者の国と年度による収録化合物数の推移

数ではあるが化合物を収録している。2010年夏にアップロードしたデータ数は19,500件で、SDBS-NMRデータを含めると総数20,426件である。2011年2月には21,200件の文献データをアップロードした。19,500件の化合物の炭素数による分布を図4に示す。C₁₅、C₂₀、C₃₀の化合物数が多いという天然物特有の分布になっている。収録されている化合物で最大のものはC₁₆₄H₂₅₆Na₂O₆₈S₂ (分子量3,426)のマイトキシシで、既知の有機天然化合物では最大の海産毒素である。

天然物研究の潮流は全合成であるとか、国内外周辺から新規天然物を探索、分離するのは困難であるという話がささやかれ、研究論文を見ているそのような傾向が見られる。ここで、CH-NMR-NPに収録してあるE-mailアドレスと論文の発表年から研究代表者の所属する国と年度毎の化合物数の統計をとって図5に示す。E-mailアドレスでは.comや.netなどがあり、国を特定できないものが約15%含まれている。

図5からわかるように、年度毎の収録化合物数は増加の傾向にある。CH-NMR-NPはNMRのデータベースであるので、高磁場化とNMR装置の向上で精度のよいNMRデータが増えていることも原因と考えられ、新規化合物数が減少しているわけではなさそうである。では、新規な骨格構造をもつ化合物が増加しているのかという問いに対して、筆者は答えられない。国別の統計では、日本とアメリカの数の振れは少ないのに対して、中国から発表される新規化合物数は際立った伸びを示している。最近はこのジャーナルでも中国からの発表論文が非常に多い。図5には示していないが、10年間に発表された新規化合物数で言えば、日本(21%)、中国(14%)、アメリカ(9%)、台湾(6%)、ドイツ(5%)、韓国(3%)と続

き、アジアから発表される新規化合物数が50%をはるかに超えている。ラテン諸国からは長く論文発表が継続しているが、アラブ、アフリカの国々からの発表論文が散見するのは極く最近の傾向である。

個々の化合物のデータを集めてデータベース化する作業は、創造性のないルーチンワークであり、誰でもできると言われる。筆者は決して否定するつもりはないが、実際には誰でもができる仕事ではない。データの収録、集積が行なわれれば、広い活用が可能であり、ファクトデータベースは古くなって腐ることはない。一方、ソフトウェアはPC環境が変化するにつれて機能を増やすと称してバージョンアップがなされ、整合性がとれなくなり困惑することもある。データベース用のソフトについても同じことが言えるが、テキストデータとして取り出してフォーマット変換すれば、新規ソフトウェアに移植することもできる。データベースを利用して新規機能を付与したソフトウェアも比較的短時間のうちに作成できる。一方、データを評価したうえで、蓄積する作業を自動的に行なうことは難しい。各ジャーナルがNMRデータの記述方式を統一するなどということになれば可能かもしれないが、研究論文を作成・投稿する研究者の立場から言えば許容できないであろう。CH-NMR-NPは個人ベースでコストも必要最小限にとどめて作成してきたし、今後もできるだけ長期にわたって収録作業を継続したいと願っている。利用者の拡大と利用者からのコメントはデータベース活動に必要である。現在、CH-NMR-NPはインターネットで有料公開している。<http://61.194.57.7/CH-NMR-NP/>にアクセスすると詳しい情報が得られる。1年に2回更新している。

謝辞：CH-NMR-NP開発には降旗一夫、瀬戸治男氏から多くのコメントをいただいた。データベース収録作業に矢部篤子氏の多大な協力を得ている。また、検索表示システムは安田幸二氏が作成した。

文献

- 1) 早水紀久子、降旗一夫：第45回天然有機化合物討論会，145 (2003).
- 2) 早水紀久子、安田幸二、降旗一夫、瀬戸治男：第46回天然有機化合物討論会，251 (2004).
- 3) 早水紀久子、安田幸二、降旗一夫：第47回天然有機化合物討論会，451 (2005).
- 4) 早水紀久子、矢部篤子、安田幸二、降旗一夫：第51回天然有機化合物討論会，25 (2009).
- 5) 早水紀久子：第26回情報化学討論会，91 (2003).
- 6) 早水紀久子：第27回情報化学討論会，91 (2004).
- 7) 早水紀久子、矢部篤子、安田幸二：第33回情報化学討論会，78 (2010).
- 8) 早水紀久子：J. Comput. Aided Chem., 2, 1 (2001).