

反応予測と反応データベース - その活用・問題点・展望

科学技術振興事業団 さきがけ研究 2 1
理化学研究所 有機合成化学研究室
佐藤寛子

Reaction Prediction Using Reaction Database

Hiroko Satoh

PRESTO, Japan Science and Technology Corporation
and RIKEN (The Institute of Physical and Chemical Research)

はじめに

化学者は実験や観測から多くの実験事実や経験を積み上げ、そこから得られる法則や規則を駆使し、あるいは例外を見付けることで新たな実験事実や経験をさらに積み上げ研究を展開している。これらの化学情報は沢山の書物や近年では様々なデータベースに格納され、化学者は本を読み、データベースを検索することで情報を得ている。これが化学が経験の学問であると言われる所以であり、化学情報は化学の発展に欠くことのできない貴重な知的資源である。

化学情報のコンテンツは広範囲かつ多様である。例えばデータベースとしては、化合物、化学反応、スペクトル(NMR, 質量分析等)、結晶構造、高分子、材料、文献情報等多岐に渡る。各データベースコンテンツと利用目的に応じた検索・可視化ソフトウェアメディアも研究開発されている。また WWW を介した化学情報検索・設計・予測・解析ソフトウェア利用のための研究開発も盛んになりつつある。学术论文の電子化による WWW 投稿や購読、さらに自動情報編集等の研究も近年急速に発展している。

このように、コンピュータ化学研究者等の地道な努力とコンピュータの進歩により化学情報や種々の化学ソフトウェアが化学者の研究活動に貢献できるようになってきた。しかし広範囲に渡る化学情報を取扱うための要素技術とその応用研究は多くの課題を抱えており、真に役立つものとして化学者に十分に認知されるまでには残念ながら至っていない。逆に言えば、将来、化学研究へ大きく貢献できる可能性を孕んだ分野であるとも言える。

我々は、「積み上げた事実情報から法則や規則を見出す化学者の思考過程を形式的にコンピュータ上でシミュレートする」発想のもと、反応データベースから反応予測のための法則や規則を見出す研究を進めている。反応予測は反応設計や構造推定とも密接な関係にあり、これらの統合化も念頭に置き研究を進めている。本稿では、主として反応予測の観点で、一方で反応設計や構造推定との関連にも触れながら、現在の反応データベースの有効活用の可能性と問題点と将来展望について、我々のこれまでの研究成果と併せて述べたいと思う。

市販の反応データベース

現在市販されている代表的な反応データベース・検索システムとしては ISIS¹, Crossfire (Beilstein データベース等が検索できる)², CASREACT³ 等がある。これらのデータベースはいずれも文献情報をデータソースとし、それぞれ数十万～数百万件の反応データを格納している。

一般に反応データベースに格納されている情報は、反応物、試薬、触媒、溶媒、生成物、温度などの反応条件と文献情報である。反応物と生成物は原子間の結合情報と各原子種を組合わせた分子構造情報として記述されている。原子の位置は x-y 座標で記述され 3 次元構造情報は含まない。ただし立体化学については、入力描画における up/down が結合情報として記述されている。試薬、触媒、溶媒については、テキスト形式、あるいは反応物等と同様の分子構造記述形式で情報が格納されている。

反応データベースと反応予測

反応データベースを活用した反応予測研究について

我々は現在、反応予測を目的とし、これまでに蓄積された大量の反応情報の種々の性質の潜在的な傾向をニューラルネットワークやパターン認識手法などを用いて系統的に解析することで反応の本質や規則性・法則性を見出す研究を進めている。以下に概略を説明する。

反応の複雑さと多様さゆえ、任意の化学反応を予測できる一般理論はいまだ確立されていない。本研究の目的は膨大な反応情報をもとに、反応の法則や規則性を見出す化学者の思考過程を模倣した反応モデルの構築とそれを活用した反応予測である(Fig.1).⁴ 具体的には、まず大量の反応情報として反応データベースを使用する (Fig.1-(1)). 反応情報は反応の支配因子に基づいて特性値化する。特性値化した反応を系統的に分類する (Fig.1-(2)). 分類の識別子と成り得た特性値で反応を説明する反応モデルを構築する (Fig.1-(3)). ここでは、コンピュータの特色である前提を維持した演算の高速性と網羅性、正確な数値処理、大容量の記憶容量を十分に活かすことで、系統的で定量的な予測能力を有し、かつ化学者の定性的な直観や思考と相補的に発展できるモデル化を目指す。これを主軸として活用し、定量的反応予測システムを構築する。

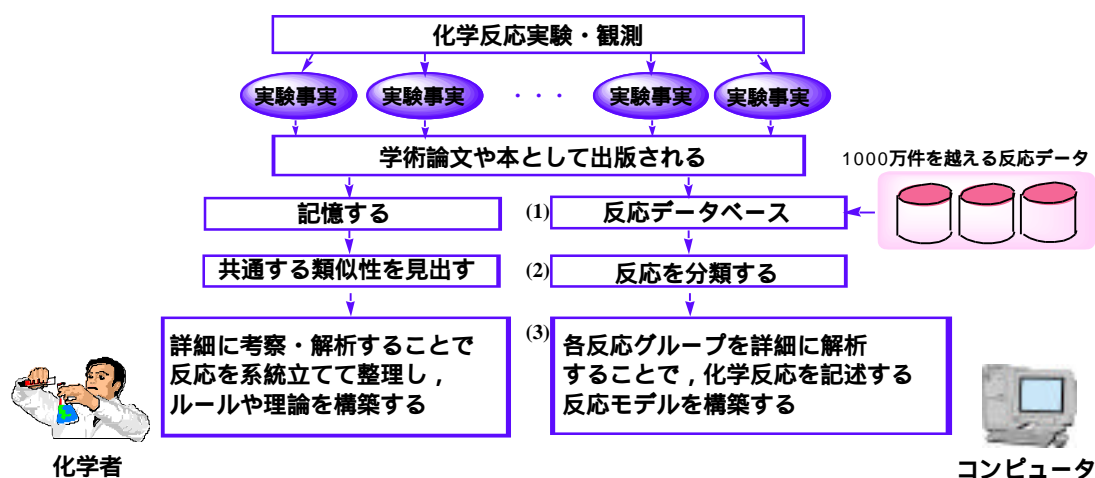


Fig. 1 化学者の思考過程を形式的にシミュレートした反応モデル構築と反応予測

これまでの研究について以下に述べる。

1. 反応の支配因子の特性値化⁴⁻⁶ 反応の支配因子の特性値化方法として、仮想反応相手との静電的・立体的相互作用に基づいて分子を特性値化する FRAU (Field characterization for Reaction Analysis and Understanding)を開発した。^{4,5} 立体構造の規範的コード化方法として、二面角に基づき任意の原子周りの立体化学環境を規範的にコード化する手法 CAST (Canonical representation of STereochemistry) を開発した。⁶ いずれも有機化合物全般に適用できる汎用的な特性値化およびコード化法である。

2. 系統的分類研究^{4, 5, 7-9} 開発した特性値化法や種々の物理化学パラメータを識別子とした反応分類研究を進めている。分類研究を通じ、反応部位の電子的特性値の変化が分類に有効であるとの重要な知見を得た。^{4, 7} また、FRAU の特性値を識別子とし種々の反応試薬を機能別に分類することにも成功した。^{4, 5, 8} さらに、進行した反応の反応物の反応しなかった部位に着目することで、反応性と反応条件を分類する方法を開発した。⁹ CAST コードによる3次元部分構造分類も可能とした。

3. 予測モデル・システムの開発^{8, 10-12} 反応予測システムの足掛かりとして、平面部分構造の変化により反応を記述した、合成経路設計システム AIPHOS¹⁰ の知識ベースを活用し、任意の反応物と反応条件から可能な反応生成物を予測するシステム SOPHIA (System for Organic reaction Prediction by Heuristic Approach)を開発した。^{11, 12} 次に数値的に予測するモデルとして、FRAU で特性値化した反応試薬分子を基に、試薬の機能を数値的に予測するニューラルネットワークモデルを構築し、⁸ さらに予測結果の化学実験による検証とフィードバック研究を行っている。一方、CAST コードによる3次元部分構造分類結果を基に、NMR 化学シフト予測システムを開発中である。¹³

反応データベースの問題点について

上述の我々の研究は、「現在の反応データベースの最大限の活用」と、「現在の反応データベースに不足している情報の構築」の2つの方向に大きく分けることができる。反応データベースは、システムとして成熟してきたものとの見方もあるが、反応予測に活用する場合には、以下に述べる種々の問題点や情報の不足がある。

1. データの間違い データの信頼性は使用者にとって非常に重要である。我々の所有する市販データベース(約 60 万件)に3割近い間違いのあることが約 300 件の無作為抽出調査で判明した。データの見直しと改良を提案したが、60 万件ものデータの見直しは、かなりの時間とコストを要するため難しいと考えられる。自動エラーチェックなどの機能も必要であろう。

2. データの質が均一でない 文献データをデータソースとしているため、各反応の行なわれた目的の違い(反応開発, 条件検討, 合成等)等によるデータの質の違いが生じる。例えば、副生成物の記述にばらつきがある。上述の SOPHIA システムにおいては、副生成物情報を反応データが相互に補うデータ処理と、反応部位の全ての組合わせの切断と再結合処理により、個別の反応データのみからでは得られない副反応情報を得る工夫をしている。^{11, 12}

3. 一段階反応でない場合が多い 反応データの多くは多段階反応であり、一段階反応と多段階反応との別が記述されていない場合がある。これは文献情報をデータソー

スとしていることが理由の一つである。つまり、文献では数段階の反応を一つの反応スキームとしてまとめて書かれていることが多いからである。合成経路設計には多段階反応データは有用であるが、反応予測には有用ではない。

4．系統的な情報が少ない 現在の反応データ件数は数百万件にのぼるが、反応にバリエーションのある反面、種々の条件で検討した結果などの系統的なデータが少ない。

5．立体構造情報がない 分子構造は平面構造情報として格納されている。原子間の結合関係の情報は分子の基本情報として有用であるが、立体構造は反応の重要な支配因子であり、反応予測のために必要である。しかしながら up/down で結合を記述した平面構造情報から、正確な立体化学を反映した立体構造を自動的に組み上げる手法は未だ確立されていない。また、自動的に可能なコンホマーを発生する手法も確立されていない。一方我々は、任意の部位（反応部位など）周りの立体化学環境の類似と相違を規範的に認識する手法として新しい分子コード化法 CAST を開発し、CAST で分子を記述した三次元構造データベースを構築している。^{6,13}

6．進行しなかった反応情報を含まない 進行しなかった反応をどれだけ知っているかは有機合成研究を成功させるカギの一つであるといわれる。これらの情報のデータベース化は重要であるが、データの性質上、一般に公にされることはない。進行した反応の反応物の反応しなかった部位に着目した反応性・反応条件分類は、現在の反応データベースから、進行しなかった反応に相当する情報を導き出すための研究である。⁹

反応データベースの将来展望

上述した通り、現存の反応データベースは、反応予測に活用するには種々の問題点を有する。これらの実現には反応データベースの質・内容・量の向上と活用のための新しいアプローチが必要であろう。以下に、将来の反応データベースについての展望をあげてみたい。

1．ロボット合成による反応データソースの拡充 ロボット合成 - 自動分離 - 自動構造解析 - 自動登録の一連の処理が可能になれば、一律の処理のもとで得られた質の均一な、系統的で信頼性の高い反応データを得ることができると考える。これが実現すれば、文献情報をデータソースとすることにより生じる問題点の多くを解決することができる。例えば上述のデータのエラー、質の不均一性、一段階反応として格納されていない、系統的情報の不足、進行しなかった反応情報がない、等の問題が解決できる。

自動構造解析のための、NMR 等のスペクトルデータをもとにした自動構造推定システムも種々開発されており、¹⁴ 現在、平面構造レベルでの構造推定・NMR スペクトル予測は企業等において実際に使用されている。我々は、これらを立体化学を考慮した実用に適う精度に向上させるための研究開発を現在進めている。^{6,13}

2．理論化学計算結果データベース 量子化学計算や分子力場計算などの理論化学計算手法は、コンピュータの性能と計算手法等の向上により、今後も高精度計算の適用範囲がさらに広がると期待される。これまで非経験的分子軌道法についての文献データベースは構築されてきているが、理論化学計算から得られる種々の物理化学データに関するものはない。これらの物理化学データを我々のアプローチにより活用すれば、理論化学的観点からの系統的な反応分類・モデル化が可能となる。このアプローチは、情報

化学と理論化学とを融合することで，化学反応の新しい系統的整理・知識化・解釈をもたらす可能性をもつ．現在，山口大学の堀らにより遷移状態データベース構築が進められており，¹⁵ その内容の重要性とともに，情報化学との融合という観点からも期待される．また，分子配座の自動発生が可能になれば，1.で述べたロボット合成によるデータソースの拡充と組み合わせることで三次元構造反応データベースも同時に実現することが可能となる．

3．進行しなかった反応データベース　進行しなかった反応データを集めてデータベース化することは可能であろうか？　一つは‘ネガティブ反応投稿雑誌’のようなものを作り，データソースとすることが考えられる．その性質上，一般にこのようなネガティブデータは公にされないが，ネガティブデータを公表する必要性は唱えられている．¹⁶ 将来の一つの方向として考えることができるかもしれない．二つには，1.で述べたロボット合成によるネガティブデータの獲得である．これも将来期待できる方向と考える．

以上述べてきた重要な種々のデータがフレキシブルにリンクされた質の高い統合的反應データベースを，反応予測の観点から展望する．このような質と内容のよい反応データベースは，それを基にした反応予測，反応設計，NMR 化学シフト予測，自動構造推定等のシステムにおける多様な情報表現・処理手法の研究開発の発展と結びつくことで，合成化学における，より高度なコンピュータ利用を可能とするであろう．

References

1. MDL Information Systems, Inc.
2. MDL Information Systems, Inc. (Beilstein database 開発は Beilstein Institute)
3. Chemical Abstracts.
4. 佐藤寛子,「化学と工業」化学のフロンティア'99 - はばたけ若き研究者たち - , vol. 52, 2月号, 146-150, (1999).
5. Satoh, H., Itono, S., Funatsu, K., Takano, K., Nakata, T., *J. Chem. Inf. Comput. Sci.*, **39**, 671-678, (1999).
6. Satoh, H., Koshino, H., Funatsu, K., Nakata, T., *J. Chem. Inf. Comput. Sci.*, **40**, 622-630, (2000).
7. Satoh, H., Sacher, O., Nakata, T., Chen, L., Gasteiger, J., Funatsu, K., *J. Chem. Inf. Comput. Sci.*, **38**, 210-219, (1998).
8. Satoh, H., Funatsu, K., Takano, K., Nakata, T., *Bull. Chem. Soc. Jpn.*, in press.
9. 佐藤寛子, 船津公人, 中田 忠, 第21回情報化学討論会, 東京, 11月, 1998.
10. 船津公人, 佐々木慎一「AIPHOS コンピュータによる有機合成経路探索」共立出版, 1994.
11. Satoh, H., Funatsu, K., *J. Chem. Inf. Comput. Sci.*, **35**, 34-44, (1995).
12. Satoh, H., Funatsu, K., *J. Chem. Inf. Comput. Sci.*, **36**, 173-184, (1996).
13. a)佐藤寛子, 越野広雪, 船津公人, 鶴澤 洵, 中田 忠, 日本農芸化学会 2000 年度大会, 東京, 3月, 2000., b) 越野広雪, 佐藤寛子, 船津公人, 中田 忠, 鶴澤 洵, 日本農芸化学会 2000 年度大会, 東京, 3月, 2000.
14. SpecInfo (Chemical Concept), ACD (Advance Chemistry Development), CHEMICS (豊橋技術科学大学, 佐々木, 船津ら) などがある
15. 堀 憲次, 2000 計算化学討論会, 東京, 6月, 2000.
16. Sierra, M. A., Torre, M. C., *Angew. Chem. Int. Ed.*, **39**, 1538-1559, (2000).