

化学研究における実践的活用を指向した 化学反応データベースの検証

佐藤 寛子^{a*}, 中田 忠^b

^a 国立情報学研究所 知能システム研究系, 〒 101-8430 東京都千代田区一ツ橋 2-1-2

^b 理化学研究所 有機合成化学研究室, 〒 351-0198 埼玉県和光市広沢 2-1

*e-mail: cheminfo@nii.ac.jp

(Received: June 27, 2003; Accepted for publication: July 18, 2003; Published on Web: September 8, 2003)

本論文では, 化学反応データベースの信頼性についての検証結果が述べられている. 近年, 化学反応データベースを自動的に活用し, 化学研究への実践的活用を目指す化学反応予測や合成経路設計の研究が報告されている. これらのコンピュータシステムにとって, その基盤となる化学反応データベースの質と内容は極めて重要である. そこで, データベースの質の基本であり, かつ, 検索対象として利用する場合にも重要であることから, データの信頼性についての検証が行なわれている. 60万件の化学反応データから抽出した329件の反応データのうち151件に誤データが見つげられている. 誤データの事例と現行の一般的な反応データベースの抱える課題について考察が行なわれる.

キーワード: 化学反応データ, データベース, 化学反応予測, 合成経路設計

1 はじめに

近年のコンピュータやネットワークの急速な普及により, 化学を取り囲むコンピュータ事情も大きく様変わりした. 化学反応データベースもその1つである. 一昔前までは図書館で Chemical Abstracts や Beilstein を調べていたものが, 昨今ではデータベース化された情報とグラフィカルな検索システムとインターネット通信により, 実験室や居室から, 容易に化学反応データを閲覧することができるようになった.

有機合成化学の分野において化学反応情報は, 時として研究の成否の行方を左右する. すなわち, キーとなる反応を主軸とした効率的な合成経路を設計するにあたって, 既知の化学反応をどれだけ知っているか, また, 検索できるかは, 重要な要素である. この意味から, 有機化学においても「情報を制するものが研究を制する」との声が聞かれるようになってきている.

一方, こうしてコンピュータ上に蓄積されるようになった化学反応情報を自動的に処理し, 化学合成経路

設計や化学反応予測に繋げることで, 化学研究の支援と, 設計・予測先導型の新しい研究分野を開拓しようとする研究が行なわれるようになった. 世界で最初の合成経路設計システム LHASA[1] や反応予測システム CAMEO[2] などでは, 合成や反応についての経験的な規則をプログラムに組込むアプローチがとられている. これらのシステムは論理的処理が可能である利点をもつが, 膨大な規則や例外, 新しいデータが追加されるたび毎に1つずつプログラム化しなければならない欠点をもつ. そこで, 新規に報告された反応情報も含めた迅速な活用が可能である新しいアプローチとして, 蓄積された化学反応情報そのものを自動的に処理し, 合成設計や反応予測に活用できる知識ベースを誘導する方法が考案された. 合成経路設計システム AIPHOS[3], WODCA[4], 反応予測システム SOPHIA[5] などがこれにあたる. 反応予測システム EROS[6] も反応データベースから直接知識ベースを誘導するが, あらかじめ同一タイプの反応のみを集めた化学反応データベースをユーザーが構築しなければならないという点で, 知

識の半自動誘導型システムと位置づけられる。また、AIPHOSは、こうした自動誘導型知識ベースの利用とあわせて、Coreyらの逆合成設計の思想 [7]にもとづく論理型の合成設計も行うことで、現実的で、かつ論理的な合成経路設計を目指す点に特色がある。

また、化学反応情報を体系的に分類し、化学反応の系統的理解と予測につなげようとする研究も行なわれている。化学反応分類研究としてはGasteigerらにより先駆的な研究が行なわれており、Michael反応などの特定の反応タイプに属する化学反応をあらかじめ集めた反応データセットの中で分類が行なわれている [8, 9]。佐藤、船津らは、報告された反応情報そのものを体系的に分類することを指向し、反応タイプなどの既存の枠を設定しない、より汎用的なデータセットと反応表現による分類研究を報告している [10]。さらに、化学反応情報から、より高度な情報を引き出すための表現法の開発や予測モデル化などへの発展的な研究についても報告している [11–15]。

このように、化学反応情報を活用する研究は着々と前進している。化学反応情報を基盤とした合成設計・反応予測等の研究は、実験事実を基盤とし、かつ理論化学計算結果をも情報として取り込むこともできるため、現実性と理論性を兼ね備えた予測や設計を可能とする。また、不要な副反応生成物を出さない反応の予測や環境にやさしい触媒や試薬の設計、さらに、設計や予測を行った上で実験を行う研究形態への道の開拓等の21世紀の化学にとっても重要な課題である環境問題に対するひとつの解を与えることも期待される。これらの意味から、今後さらなる発展の望まれる領域であると見なすことができる。

このための基盤となる化学反応情報の内容と質は極

めて重要である。検索利用の場合、通常、研究者は検索結果を自ら検証するとともに、重要なものについては文献へと辿り確認を行うため、反応情報が多少の誤りを含んでいたとしても、深刻な事態に陥ることは少ない。しかしながら、化学反応情報を自動的に利用する場合には、データの誤りや不均一性などが結果を大きく左右することが時として起こりうる。また、化学研究への実践的活用を鑑みた場合に必須の情報が化学反応データベースに含まれていない、もしくは適切に表現されていない、といった問題も起こりうる。現在入手可能な化学反応データベースは、これらの要求に対し、はたしてどこまで耐えうるのであろうか。現存の化学反応データベースは将来の活用にも供することのできる知的基盤となりえるのであろうか。

そこで本研究では、実際に市販されている化学反応データベースについて、最も基本となるデータの信頼性について検証した。今後の課題の提言も含めて報告する。

2 化学反応データの検証方法

化学反応データベースとして、主要データベースの1つであり、化学系企業等で幅広く利用されているISISデータベースのChemInform (ISIS-ChemInform)を用いた [16]。本DBに格納されている約60万件の化学反応情報から、理化学研究所の有機合成化学研究室の合成化学者が報告したデータ329件を抽出し、個々のデータの報告者がオリジナル情報と照合し、データベース情報の正しさの検証を行った。

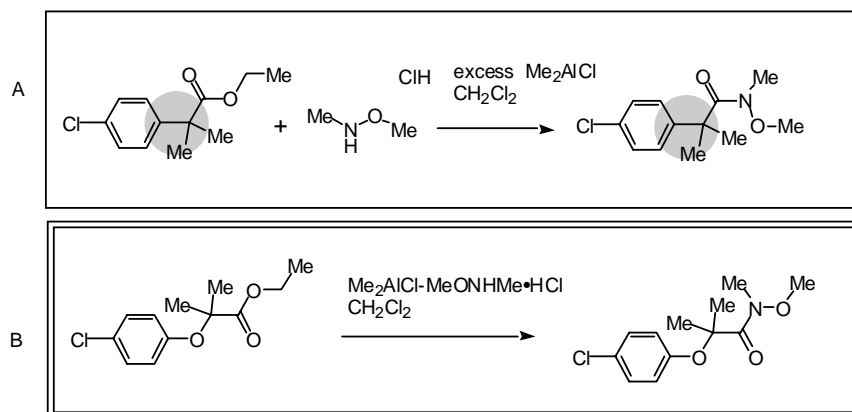


Figure 1. An example of error on planar structures.

3 結果

3.1 概要

検証の結果、329件中151件(46%)に何らかの誤りが見出された。Table 1に、エラー項目別にまとめた結果を示す。これらの間違いはいずれも、入力の際のタイプミスや解釈ミスであり、文献の間違ひではなかった。一件の化学反応データに複数の種類の間違いが含まれる場合もあるため、項目別の誤データ件数の総数は、間違いの見付かったデータ件数151とは一致しない。

Table 1. The results from the verification of a reaction database.

エラー項目	件数
反応段階数	83
化学反応式	68
反応部位	61
反応物・反応生成物などの分子構造	50
反応条件	31
収率	18

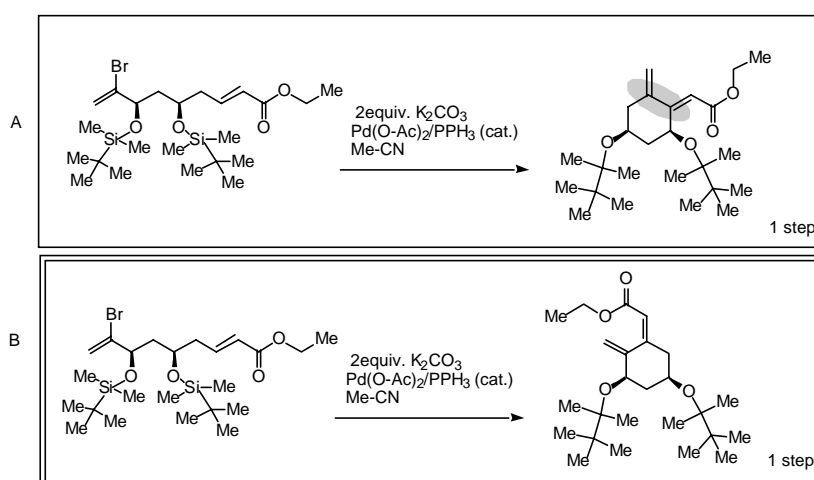


Figure 2. An example of error on substitution.

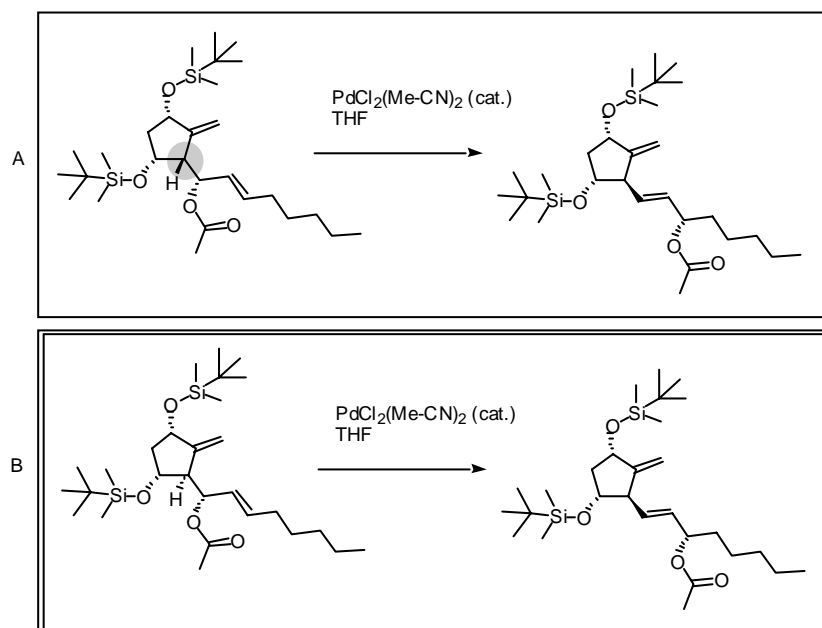


Figure 3. An example of error on stereochemistry.

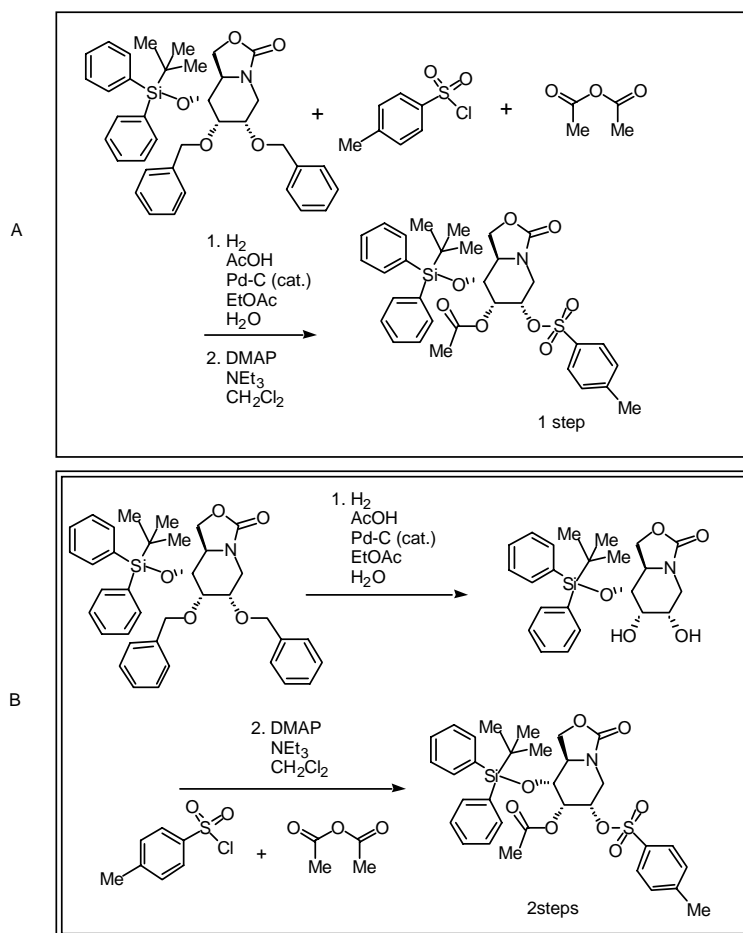


Figure 4. An example of error on the number of reaction steps and schemes.

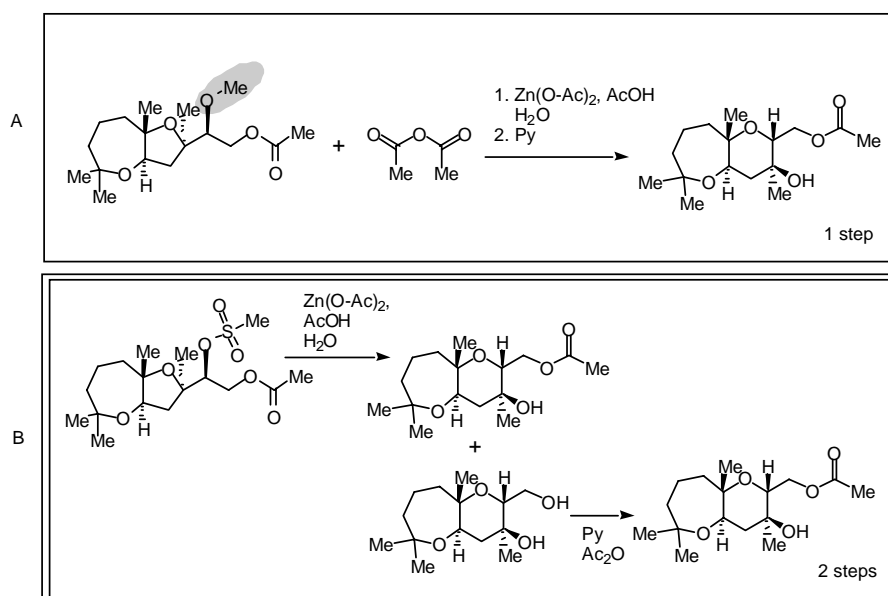


Figure 5. An example of error on planar structures, the number of reaction steps, and schemes.

分子構造は化学反応データの最も基本かつ重要な情報であるが、329件中50件(15%)のデータに誤りが見付かった。最も誤りの件数の多かったのが反応段階数であり、329件中83件(25%)であった。つぎに、個々のエラー内容の実例を示す。

3.2 実例

分子構造の間違ひの実例を Figures 1-3 に示した。各図とも、A がデータベース内のデータであり、B が正しいデータである。間違ひの箇所をグレーで示した。

Figure 1 は平面構造式の間違ひ例である [17]。反応物、反応生成物ともにエーテル構造が欠落していた。Figure 2 は置換位置の間違ひ例である [18]。反応生成物の置換位置が間違ひている。Figure 3 は立体化学の間違ひ例である [19]。ISIS データベースでは分子構造中に up/down の記述がされているが、1箇所のみ down で表されるべき箇所が up として描かれていた。

Figures 4, 5 は反応スキップ数と反応スキームの間違ひ例である [20, 21]。Figure 5 では反応物の平面構造にも間違ひが見られる [21]。Figures 4, 5 では、2段階目の反応物が2段階分の反応条件とあわせて1スキームで描かれている。Figure 5 の例では、1段階目で望みのエステル体と、副生成物としてアルコール体を得られたため、これらアルコール体を2段階目でエステル体へと変換したのが正しいデータである。

より複雑な間違ひの例を Figure 6 に示す。この例ではまず、反応物と反応生成物の平面構造式が間違ひている [22]。また、実際は2段階の反応であり、Aにおいて反応物の1つとして描かれているケトンは、実際には2段階目の反応物である。さらに、平面構造と反応段階数を修正した反応 B は文献では「No Reaction」と報告されている。すなわち、反応の進行の有無についての事実が間違ひて登録されていたことになる。文献中では、本反応では異なる反応生成物を得られたと報告されている (Figure 7)。

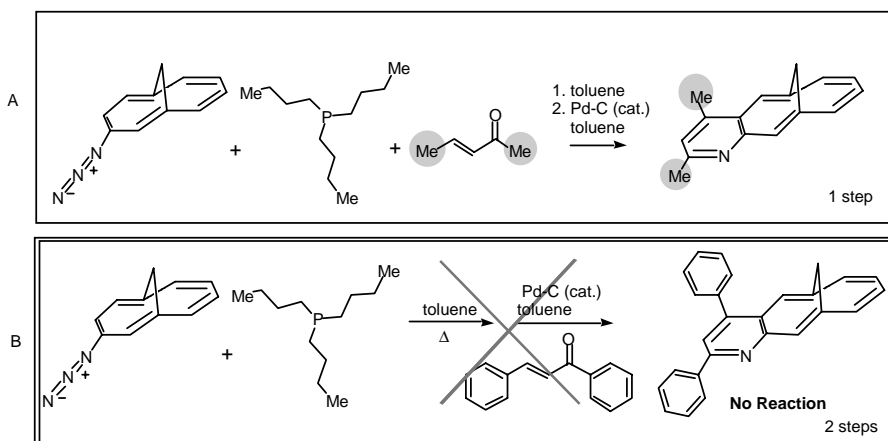


Figure 6. An example of complicated error data.

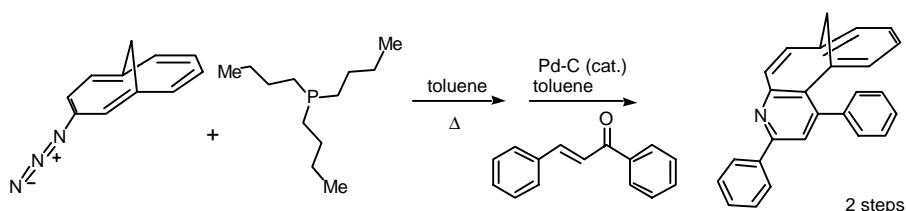


Figure 7. A reaction scheme reported in the journal.

反応部位，文献，収率，反応条件についての間違いの事例は示していないが，例えば収率の場合には，数値が間違っている場合や，ジアステレオマー等も含む複数の反応生成物が得られた場合の生成比が逆転している等の間違いが見られた。

4 議論

誤データの割合 4.6% は，検証前に想定した数値を遥かに越えるものであった。ISIS-ChemInform は，認知度や化学系企業等での普及率の高さからも，現在までに報告され市販されている反応データベースの中でも標準以上のレベルを有するシステムであると位置づけることができる。その ISIS-ChemInform ですらも，これだけの高い割合で誤データを有しているという事実は，ほぼ同様の作成手順を踏んでいると考えられる一般的な他の化学反応データベースにおいても同様の傾向を内包しているであろうことを示唆している。この数字は，反応データベースを検索対象として利用する際にも，半数近い確率で誤データが含まれる可能性を意識する必要があることを示している。

反応予測・合成経路設計システムにおける利用を指向した場合，間違った情報をもとに予測や設計が行なわれる可能性が高いことを示しており，深刻である。大多数の正しいデータによって少数のデータの間違いが浮き彫りにされることは，大量の情報を活用する場合の利点であるが，半数近くの間違いの割合は，この利点が活かされない場合があることを示している。中でも分子構造の間違いは致命的であり，いずれのシステムにとっても予測・設計の間違いを引き起こし，出力結果の信頼性が根本から問われることに繋がる。反応段階数の間違いと不均一性は，1 段階の反応スキームについての反応生成物と生成比を予測する必要のある反応予測システムにおいて特に深刻な問題であるが，Figures 4, 5 に示した例では反応スキームを構成する反応物構成も間違っているため，合成経路設計においても同様に問題となる。

対策としてはまず，現行の反応データベース作成手順の中での誤データ検査の過程を，より厳しくすることであろう。誤データの種類のうち，平面構造式や立体配置などの分子構造の間違いや反応条件，収率，文献の間違いは，データベース作成者側の多重検査により，かなりの程度まで誤データ数を減らすことができると考えられる。反応部位や反応段階数には反応の解

釈の違いによる不均一性が生じる場合があるが，一定基準を設けることで，ばらつきを最小限に抑えた質の高いデータベースを構築することができるのではないだろうか。反応段階の記述については，反応予測や合成経路設計における活用を考えた場合，鍵となる反応物と反応生成物に，その間の多段階の反応条件のみを記したものと，多段階の反応それぞれを，1 段階毎に記したものと 2 通りのデータが必要である。

日々増加する化学反応情報に対応するためには，反応データの検査を可能な限り自動的に行う機能を開発することも必要であろう。いずれにせよ，提供する反応データに対し，最大限の品質を保証することは，反応データベース作成者に求められる重要な義務である。

現行の化学反応データベースは検索を主要目的として開発されてきたものである。そのためもあり，反応予測や合成経路設計を指向した場合，情報の質と内容として種々の不十分な点が浮き彫りにされてくる。例えば，進行しなかった反応の情報，一律な環境下での一連の系統的な反応の検討結果，理論計算の結果等がこれに相当する。この点については，いずれ別の機会に論じることとする。

5 まとめ

反応予測・合成経路設計システムの知的基盤として，化学反応情報を化学研究に実践的に活用するためには，質・内容ともに多くの課題を解決する必要がある。今回報告した検証の結果も，現行の反応データベースの抱える問題を示唆している。反応予測・合成経路設計システムにおいては，現行の反応データベースを最大限有効利用する工夫もなされているが，それだけでは解決できない問題も多い。コンテンツや反応情報の表現方法も含め，反応予測・合成経路設計システムと連携した合目的な化学反応データベースの構築が必要である。

本研究は，科学技術振興事業団さきがけ研究 2-1 のもとで行なわれました。化学反応データの検証にあたって，理化学研究所有機合成化学研究室の研究員の方々に多大なご協力をいただいたことを感謝いたします。また，検証結果の編集やデータ収集の一連の作業を担当して下さった曾麗氏に感謝いたします。

参考文献

- [1] E. J. Corey, W. T. Wipke, *Science*, **166**, 178-192 (1969).
- [2] T. D. Salatin, W. L. Jorgensen, *J. Org. Chem.*, **45**, 2043-2057 (1980).
- [3] K. Funatsu, S. Sasaki, *Tetrahedron Comput. Method.*, **1**, 27-38 (1988).
船津公人, 佐々木慎一, 「コンピュータ・ケミストリーシリーズ (2)AIPHOS-コンピュータによる合成経路探索」, 共立出版 (1994).
- [4] J. Gasteiger, W. D. Ihlenfeldt, P. Röse, *Recl. Trav. Chim. Pays-Bas*, **111**, 270-290 (1992).
- [5] H. Satoh, K. Funatsu, *J. Chem. Inf. Comput. Sci.*, **35**, 34-44 (1995).
H. Satoh, K. Funatsu, *J. Chem. Inf. Comput. Sci.*, **36**, 173-184 (1996).
- [6] P. Röse, J. Gasteiger, *Anal. Chem. Acta*, **235**, 163-168 (1990).
P. Röse, J. Gasteiger, *Software Development Chemistry*, **4**, 275-288 (1990).
- [7] E. J. Corey, X-M, Cheng, *The Logic of Chemical Synthesis*, Wiley-Interscience (1995).
- [8] L. Chen, J. Gasteiger, *Angew. Chem.*, **108**, 844 (1996).
Angew. Chem. Int. Ed. Engl., **35**, 763-765 (1996).
- [9] L. Chen, J. Gasteiger, *J. Am. Chem. Soc.*, **119**, 4033-4042 (1997).
- [10] H. Satoh, O. Sacher, T. Nakata, L. Chen, J. Gasteiger, K. Funatsu, *J. Chem. Inf. Comput. Sci.*, **38**, 210-219 (1998).
- [11] H. Satoh, S. Itono, K. Funatsu, K. Takano, T. Nakata, *J. Chem. Inf. Comput. Sci.*, **39**, 671-678 (1999).
- [12] H. Satoh, K. Funatsu, K. Takano, T. Nakata, *Bull. Chem. Soc. Jpn.*, **73**, 1955-1965 (2000).
- [13] H. Satoh, H. Koshino, K. Funatsu, T. Nakata, *J. Chem. Inf. Comput. Sci.*, **40**, 622-630 (2000).
- [14] H. Satoh, H. Koshino, K. Funatsu, T. Nakata, *J. Chem. Inf. Comput. Sci.*, **41**, 1106-1112 (2001).
- [15] H. Satoh, H. Koshino, T. Nakata, *J. Comput. Aided Chem.*, **3**, 48-55 (2002).
- [16] MDL-Information Systems Inc.
- [17] T. Shimizu, K. Osako, T. Nakata, *Tetrahedron Lett.*, **38**, 2685-2688 (1997).
- [18] K. Nagasawa, Y. Zako, H. Ishihara, I. Shimizu, *Tetrahedron Lett.*, **32(37)**, 4937-4940 (1991).
- [19] M. Nakazawa, Y. Sakamoto, T. Takahashi, K. Tomooka, K. Ishikawa, T. Nakai, *Tetrahedron Lett.*, **34(37)**, 5923-5926 (1993).
- [20] S. Takahashi, H. Kuzuhara, *J. Carbohydr. Chem.*, **17(1)**, 117-128 (1998).
- [21] K. Nagasawa, N. Hori, R. Shiba, T. Nakata, *Heterocycles*, **44(1)**, 105-110 (1997).
- [22] N. Kanomata, H. Kawaji, M. Nitta, *J. Org. Chem.*, **57**, 618-625 (1992).

Verification of a Chemical Reaction Database –Is It Sufficient for Practical Use in Chemical Research?

Hiroko SATOH^{a*} and Tadashi NAKATA^b

^aIntelligent Systems Research Division, National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda, Tokyo 101-8430, Japan

^bSynthetic Organic Chemistry Laboratory, RIKEN
2-1 Hirosawa, Wako, Saitama 351-0198, Japan

**e-mail: cheminfo@nii.ac.jp*

The accuracy of the chemical reaction data of a reaction database is verified to determine the validity for practical use in chemical research. Reaction databases have been traditionally used in data-search, but recently, there have been some approaches deriving knowledge for synthetic design and reaction prediction systems. The design and prediction based on the approaches are providing valid scientific technologies that could provide a new chemical research style in which design and prediction are done before experiments. The technologies must give an answer to the serious issues concerning the environments of Earth, and are expected to reduce the number of experiments, predict a synthetic route producing no useless side-reaction products, and design environmentally friendly catalysts and reagents for replacing to hazardous and toxic ones. In the reaction prediction and synthetic design systems based on a reaction database, the quality and contents of the reaction database are of critical importance. Low quality and lack of contents may lead to wrong outputs from the systems. High accuracy of reaction data is particularly essential for both database search and knowledge derivation, and a reaction database is accordingly verified in order to determine the correctness of the data. The verification is done using 329 sampling reaction data from 600,000 data in a commercially available database, and 151 error data are found. The types of error concern planar and/or stereochemical structures, the number of reaction steps, reaction schemes, reaction sites, reaction conditions, product ratios, and article information. This paper describes the results from the verification and discussion of the problems for practical use in the current available reaction databases.

Keywords: Chemical reaction data, Database, Chemical reaction prediction, Organic synthetic design