Chemical Information in Organic Synthesis: From Data to Knowledge

Philippe Jauffret and Claude Laurenço

Laboratoire des Systèmes d'Information Chimique (UMR 5076 du CNRS)
ENSCM, 8 rue de l'Ecole Normale, 34296 Montpellier CEDEX 5

Abstract: The chemists widely use analogical reasoning to design a synthetic plan (sequence or tree of reactions leading from avalaible chemicals to the target molecule), trying to find and adapt previous experiments "similar" to the problem in hand. But the amount of avalaible information is so huge that it can be managed and used only trough sophisticated softwares. A first generation of "Computer Assisted Design Of Synthesis" systems appeared around 1970, surfing on the "Artificial Intelligence" wave. The most powerful of these systems exploited a reactional knowledge base. Results provided by such systems were often relevant, but these softwares rapidly showed limitations: in particular, "manually" built-up reactional knowledge bases were expensive, and difficult to maintain. Reaction databases have then been developed. Searches by substructure or by similarity allowed the chemists to make some analogical reasoning. Today, reaction databases are still the best tools avalaible for solving synthetic problems, but they are not the ultimate tool for the synthetic chemists. Even the most used of them, like Beilstein and ChemInform databases, have been shown to contain erroneous data, to suffer from heterogeneity and insufficient structuring [Coste 99]. Since 1990, a second generation of "Computer Assisted Design Of Synthesis" [Ihlenfeldt 95] systems involves automatic inductive acquisition of reactional knowledge starting from factual reaction databases. Our opinion is indeed that efficient tools for decision support have both to perform high-level reasoning and keep synchronized links with experimental data. Three examples of such a "data-mining" approach in the framework of organic synthesis will be presented and discussed from this point of view: GRAMS [Jauffret 00] and RESYN-MINING [Berasaluce 04] are two projects for reactional knowledge acquisition. COSYMA [Jauffret 03] deals with automatic extraction of synthetic strategy from synthetic pathways.

[Coste 99] J. Coste, O. Gien, A. Dietz et C. Laurenço "A propos de l'utilisation des bases de données de réactions" L'Actualité Chimique. 1999, num. de juillet, 27-32.
[Ihlenfeldt 95] W. D. Ihlenfeldt, J. Gasteiger, "Computer-Assisted Planning of Organic Syntheses: The Second Generation of Programs", Angew. Chem. Int. Ed. Engl., 1995, 34, 2613-2633
[Jauffret 00] JAUFFRET Ph., VOGEL H., SCHILDKNECHT S., KAUFMANN G. "Learning synthetic knowledge from reaction databases: dealing with experimental conditions", Proceedings of the 2000 International Chemical Information Conference, Ed. H. COLLIER, Pub Infonortics Ltd. Tetbury (England), 2000.
[Jauffret 03] JAUFFRET Ph., OSTERMANN C., KAUFMANN G. "Using the COSYMA system for the discovery of synthesis strategies by analogy", Eur. J. Org. Chem., 1983-1992 (2003).
[Berasaluce 04] S. Berasaluce, C. Laurenço, A. Napoli et G. Niel, "An experiment on knowledge discovery in chemical databases" Lecture Notes in Artificial Intelligence, 2004, 3202, 39-51.