ELSEVIER

Teaser To be able to predict chemical reactions is of the utmost importance for the pharmaceutical industry. Recent trends and developments are reviewed for reaction mining, computer-assisted synthesis planning, and QM methods, with an emphasis on collaborative opportunities.

Computational prediction of chemical reactions: current status and outlook

Ola Engkvist[1], Per-Ola Norrby[2], Nidhal Selmi[1], Yu-hong Lam[3], Zhengwei Peng[3], Edward C. Sherer[3], Willi Amberg[4], Thomas Erhard[4] and Lynette A. Smyth[4]

[1] Discovery Sciences, Innovative Medicines and Early Development Biotech Unit, AstraZeneca R&D Gothenburg, SE-43183 Mölndal, Sweden
[2] Pharmaceutical Sciences, Innovative Medicines and Early Development Biotech Unit, AstraZeneca R&D Gothenburg, SE-43183 Mölndal, Sweden
[3] Modeling and Informatics, MRL, Merck & Co., Rahway, NJ 07065, USA
[4] AbbVie Deutschland GmbH & Co. KG, Neuroscience Discovery, Medicinal Chemistry, Knollstrasse, 67061 Ludwigshafen, Germany

Over the past few decades, various computational methods have become increasingly important for discovering and developing novel drugs. Computational prediction of chemical reactions is a key part of an efficient drug discovery process. In this review, we discuss important parts of this field, with a focus on utilizing reaction data to build predictive models, the existing programs for synthesis prediction, and usage of quantum mechanics and molecular mechanics (QM/MM) to explore chemical reactions. We also outline potential future developments with an emphasis on pre-competitive collaboration opportunities.

Introduction

Small organic molecules are the bread and butter of drug discovery. To synthesize these small organic molecules, reaction predictions are practiced routinely by medicinal chemists, who make diverse sets of molecules on a small scale to efficiently probe the structure–activity relationship (SAR) through the design–make–test–analyze cycle, and by process chemists, who intend to discover the most efficient, cost-effective, and environmentally green routes to synthesize late-stage drug candidates in larger quantities. As such, the effectiveness of reaction prediction is a key factor contributing to the efficiency and success of drug discovery and development. Therefore, it is no surprise that there are many *in silico* tools available to assist chemists in reaction prediction and that this area has remained active in terms of research and development, especially in recent years. We have come together in a precompetitive fashion to further discussion of how the larger community can drive additional development in this space through data sharing and collaboration.

Corresponding author: Engkvist, O. (Ola.Engkvist@astrazeneca.com)

Ola Engkvist was awarded his PhD in computational chemistry by the University of Lund in 1997, and continued with postdoctoral research at the University of Cambridge and the Czech Academy of Sciences. Between 2000 and 2004, he worked for two biotech companies before joining AstraZeneca in Gothenburg, Sweden, in late 2004. He is currently a team leader in cheminformatics within the Discovery Sciences sector. His research interests include cheminformatics, molecular modeling, machine learning, phenotypic screening, and open innovation.

Yu-hong Lam was awarded his M. Chem. and PhD by the University of Oxford under the guidance of Veronique Gouverneur working in organofluorine chemistry. After he graduated, he carried out postdoctoral research at UCLA with Ken Houk in computational catalyst design and reaction discovery. He is currently a senior scientist at Merck Research Laboratories, and his research involves the collaborative application of modeling and informatics to streamline organic synthesis.

Willi Amberg studied at ETH Zfirich and was awarded his PhD in organic chemistry in 1989, followed by postdoctoral research at MIT and Scripps Research Institute. In 1992, he joined BASF Pharma, working in oncology and cardiovascular research, before taking up his current position as a group leader in neuroscience at Abbott GmbH and later at AbbVie GmbH. Ludwigshafen, Germany. His research activities are mainly focused on therapies for schizophrenia and Alzheimer's disease, and his interests include medicinal chemistry and drug design.

https://doi.org/10.1016/j.drudis.2018.02.014

There are several types of question to be addressed by reaction predictions: (i) forward reaction prediction: given a set of reaction building blocks, what could be the potential products? Which one might be the major product? What might be the most favorable reaction condition(s) for the putative major product? What is the potential yield of the putative major product? (ii) Retrosynthetic analysis: given a desired molecule, what are the possible synthetic route(s) to make this molecule based on available reaction building blocks on hand? How can we rank and filter these possible synthetic routines according to user-defined criteria? (iii) Reaction mechanism elucidation: given an overall reaction, what could the fundamental mechanistic reaction steps be? What are the major factors determining product yield or stereo- and regioselectivity?

Here, we discuss tools and methods to address these three types of question, with a focus on: (i) the latest machine learning (ML) approaches for both forward reaction prediction and retrosynthetic analysis; (ii) the utility of retrosynthetic analysis tools in the eyes of medicinal and process chemists; and (iii) the state of the art and outstanding problems in the application of quantum chemical calculations to elucidate reaction mechanisms, origins of selectivity, and spectroscopic properties.

Reaction knowledge mining

Background

Here, we focus on recent development in cheminformatics to better use historical reaction data for predicting synthetic pathways for novel molecules. With more sophisticated methods to extract data from in-house and literature sources, reaction knowledge mining is entering the 'big-data' era. This, in combination with ML methods, is creating a step change in the application of reaction knowledge mining.

Data sources, standardization, extraction, and reaction classification

As depicted in Fig. 1a, reaction data already published in scientific journals and patent literature are generally extracted, curated, aggregated, and hosted by data vendors and made available for users to access through vendors' proprietary tools (e.g., SciFinder from Chemical Abstracts Services and Reaxys from Elsevier). The vendor-provided reaction databases are not discussed here, because they have been recently reviewed elsewhere [1]. In general, end users do not have direct access to the full set of vendor reaction data for reaction knowledge mining. Only recently have several academic groups published reaction mining and predictive modeling works based on the reaction data content in *Reaxys*. Their work is discussed in the section titled 'Predictive reaction modeling'.

For proprietary reaction content generated by biotech, pharma, and chemical companies, it is common to have corporate electronic laboratory notebooks (ELN) for data and intellectual property (IP) capture (Fig. 1). Given that ELN applications are mainly designed for data and IP capture, they are not ideal environments for knowledge mining in general.

To perform in-depth reaction knowledge mining to address specific scientific questions using cheminformatics tools, the reaction data have to be hosted in an IT environment that is easy to access and of high performance (Fig. 1). AstraZeneca reported the successful extraction of its MedChem ELN pages and loaded them into an internal reaction DataMart to support web searches and other external applications [2]. In 2013, Roche reported that they had collaborated with both Elsevier and NextMove to extract reaction data content from more than ten internal reaction databases and its corporate chemistry ELN, and combined them with public reaction content from Elsevier to form an integrated reaction DataMart hosted behind the Roche firewall. Roche scientists can use a customized version of *Reaxys* to search and browse all of these reaction data sources in an integrated and streamlined manner. In addition, the integrated DataMart provides a larger and richer set of reactions to enable more powerful and effective knowledge mining [3].

The *HazELNut* suite of tools from NextMove Software is commonly used to extract reaction content from vendor-provided ELN systems, perform format conversion and data curation to fix common data entry issues seen in ELNs, and add additional annotation, such as reaction classification (Fig. 1b, [4,5]). These operations directly benefit downstream operations, such as knowledge mining and predictive model building. In addition to commercial software tools, there are open-source software tools available for basic reaction analysis [6].

Once the corporate ELN content is extracted and stored in a minable format, knowledge mining can be applied to address questions such as: (i) how many syntheses have been attempted using named reactions (e.g., Suzuki aryl C–N coupling reaction and Buchwald–Hartwig aryl C–N coupling reactions)? (ii) What are the distributions and trends observed in terms of success rate in these reactions? And, (iii) how frequently has a reaction building block been used for named reactions, and what were the associated reaction success rates [2,5]? This type of information can be readily used by chemists to make more-informed decisions during compound design and building-block selection. More in-depth analysis of knowledge mining has led to the publication of a set of most robust and commonly used reactions by the medicinal chemistry community [7] and extracted reaction rules to support either retrosynthetic analysis or reaction-based virtual library enumeration [8].

Scientific journals and patent literature are biased against negative data [9,10] and the same is expected to be true for the published reaction content. By contrast, corporate ELNs do contain negative data (failed reactions). Without such a bias against negative data, ELN reaction content is expected to be more suited for knowledge mining and predictive model building. However, even with millions of reaction records inside a typical corporate ELN system, the vast chemical reaction space (defined by reaction type, reactants, products, and more variables in reaction conditions) is still only sparsely explored [11]. Recent advances in miniaturization (down to the nanomolar scale) and workflow streamlining have demonstrated the potential to explore reaction space in a more systematic and well-controlled way with higher throughput [12]. The age of big-data might have finally arrived for organic synthesis [13]. It is also encouraging that a nonproprietary format has been developed (RInChI) for handling chemical reactions [14].

Predictive reaction modeling: machine learning

Given the increased availability of reaction data, reflected both in the number of different reactions and various successful conditions for a specific reaction, it is not surprising that there have been