European Journal of Medicinal Chemistry 54 (2012) 679–689

Contents lists available at SciVerse ScienceDirect

European Journal of Medicinal Chemistry

journal homepage: http://www.elsevier.com/locate/ejmech

Original article

Is chemical synthetic accessibility computationally predictable for drug and lead-like molecules? A comparative assessment between medicinal and computational chemists

Pascal Bonnet*

Janssen Research & Development, Division of Janssen Pharmaceutica N.V., Turnhoutseweg 30, 2340 Beerse, Belgium

ARTICLE INFO

Article history:
Received 24 February 2012
Received in revised form
4 June 2012
Accepted 12 June 2012
Available online 21 June 2012

Keywords:
De novo design
Drug discovery
Synthetic accessibility
Synthesis design
Virtual screening

ABSTRACT

The design of lead and drug-like molecules with expected desired properties and feasible chemical synthesis is one of the main objectives of computational and medicinal chemists. Prediction of synthetic feasibility of *de novo* molecules is often achieved by the use of *in-silico* tools or by advices received from medicinal and to a lesser extent from computational chemists. However, the validation of predictive tools is often performed on selection of compounds from external databases. In this study, we compare the synthetic accessibility (SA) score predicted by SYLVIA and the score estimated by medicinal chemists who synthesized the molecules. Therefore, we solicited 11 bench-based medicinal and computational chemists to score 119 lead-like molecules synthesized by same medicinal chemists. Their scores were compared with score calculated from SYLVIA software. Irrespective of the starting material database, we obtained a good agreement between average of medicinal and computational chemist scores for the ensemble of compounds; as well as between all chemists and SYLVIA SA scores with a correlation of 0.7. Furthermore, analysis of the marketed drugs since 1970 shows some consistency in average SYLVIA SA scores. Compounds entered in different phases of clinical trials show some large variation in synthetic accessibility scores due to natural-derived molecular scaffolds.

Here, we proposed that the selection of compounds based on synthetically accessibility should not be done solely by one individual chemist to avoid personal gut-feeling appreciation from its experience but by a group of medicinal and computational chemists. By assessing synthetic accessibility of hundreds of compounds synthesized by medicinal chemists, we show that SYLVIA can be used efficiently to rank and prioritize virtual compound libraries in drug discovery processes.

1. Introduction

Computer-aided ligand design is heavily used by pharmaceutical companies in many drug design projects since it generates novel ideas to be exploited by medicinal chemists. The increase of computational power such as grid computing, processor performance and more recently cloud computing allows the generation of thousands *in silico* lead-like molecules in a reasonable time frames. The recent *in silico* tools developed by many software and pharmaceuticals companies can generate large number of *de novo* ligands with novel chemical structures such as VEHICLe [1], TIN [2] or GDB [3,4]. The virtual compounds generated for a specific project are then filtered by desired criteria such as Lipinski's rules [5], affinity prediction, ADME-tox parameters or pharmacophoric features. However, the remaining prioritized compounds need to be synthesized and validated in a relevant biological assay. In many cases, the ultimate prioritization step performed by medicinal chemists is assessed by the ease of synthesis of the compounds.

Many *de novo* rationale design tools such as LUDI [6], LigBuilder [7,8], SkelGen [9], HOOK [10], BREED [11], SPROUT [12,13], FLexNovo [14] and PhDD [15] are aimed to generate large number of diverse ligands. Even though the generated molecules fulfill the expected requirements for binding affinity or drug-likeness, their chemical structure is often so complex that their synthesis cannot be executed in a fast and easy way. However, to validate *in silico* models, *de novo* molecules have to be synthesized and tested on biological systems. *De novo* methods generate large number of molecules which are often arduous to synthesize due to their unavailable starting materials, stereochemistry, ring complexity and substituents arrangement. Therefore, tools to compute synthetic accessibility are needed to further filter out *de novo* ligands. Synthetic accessibility corresponds to the ease of synthesis of organic compounds according to their synthetic complexity

* Corresponding author. Tel.: +32 14 605 961; fax: +32 14 605 403.
E-mail address: pbonnet@its.jnj.com.

0223-5234/$ – see front matter © 2012 Elsevier Masson SAS. All rights reserved.
http://dx.doi.org/10.1016/j.ejmech.2012.06.024

which combines starting materials information and structural complexity. To evaluate synthetic accessibility score, several methods have been developed which takes into account molecular complexity, information of starting materials or retrosynthetic analysis. The advantage of the methods that predict synthetic feasibility from structural complexity [16,17,18] is usually the speed of the calculations since they can easily process tens to hundreds of thousands of molecules in a reasonable time frame. Recently, several methods have been developed such as CAESA [19], RECAP [20], WODCA [21], LHASA [22], RASA [23], RSsvm [24] and SYLVIA [25] that perform retrosynthesis and/or synthetic accessibility prediction for compound libraries.

To validate these methods, several attempts have been made where experienced chemists evaluate ease of synthesis of large data set of compounds. It has been shown in several examples that experienced medicinal chemists don't score synthetic accessibility of compounds in a consensus manner [25–28]. Takaoka *et al.* [26] use a data set of 3980 diverse compounds and five chemists assigned two scores corresponding to ease of synthesis and compound drug-likeness. From these data the group developed predictive statistical models to rank novel compounds and to filter out hard-to-synthesize compounds. The models can be used to prioritize compounds acquire by external providers. In the project of Lajiness *et al.* [27] thirteen medicinal chemists reviewed 22,000 compounds divided into 11 lists of about 2000 compounds for their "attractiveness". They have shown that the chemists are not consistent in rejecting undesirable compounds. Same conclusion was obtained when the chemists reviewed again a set of identical 2000 compounds. Podolyan *et al.* [24] presented two support vector machines-based models; RSsvm, a statistical model trained on a set of reactions and information of starting materials and DRsvm which takes into account synthetic information of nearest neighbors and is therefore not tied to a specific set of reactions or starting materials. To validate the SAscore, a score that estimates synthetic accessibility [28], Ertl *et al.* asked 9 experienced chemists to score 40 diverse molecules selected from the PubChem database. A very good enrichment ($r^2 = 0.89$) was obtained between consensus estimated score from medicinal chemists and calculated score from SAscore. The synthetic accessibility score (SAscore) is calculated from a combination of fragment contributions and a complexity penalty. In addition, Boda *et al.* [25] used a dataset of 100 diverse molecules extracted from the Journal of Medicinal Chemistry, which were estimated by 5 medicinal chemists. The weights of each individual component to calculate the total synthetic score of SYLVIA were estimated by linear regression analysis using the average scores provided by the medicinal chemists. The reliability of the method was estimated by comparing the average computational scores and chemist estimations, a good correlation of 0.89 was obtained from this analysis. Recently, a retrosynthesis-based scoring method called RASA [23] (Retrosynthesis-based Assessment of Synthetic Accessibility) was trained on 100 compounds extracted from the CMC (Comprehensive Medicinal Chemistry) database. Five chemists were selected to assess independently the synthetic accessibility of the compounds using publically available information. The weights of the three individual components contained in the scoring function were derived from linear regression analysis. To validate the scoring function, the former 5 chemists were asked to score 30 new compounds extracted from the CMC and 5 other chemists were asked to estimated synthetic accessibility of 25 additional compounds. Good correlation coefficients of 0.81 and 0.79 were obtained between the calculated RASA scores and the estimated scores respectively.

Validation of synthetic accessibility score has been performed on compounds, not made by medicinal chemists involved into the assessment, but rather extracted from external libraries such as MDL Drug Data Report (MDDR) [29], PubChem [30] or ZINC [31] databases. However, to correctly assess the ability of medicinal chemists to estimate synthetic accessibility of molecules, we validate their perception using a library of 119 compounds synthesized by the experienced bench-based medicinal chemists themselves and perform a cross-evaluation between all the medicinal chemists. At least one chemist knows about the number of synthetic steps, synthetic feasibility and starting material availability. Therefore the prediction of synthetic accessibility of the compounds is performed on a dataset where knowledge of synthesis is known by the chemists. Since *in silico* ligands are often proposed by computational chemists, we solicit 4 computational chemists to score the compounds having an average of 11 years of experience in the drug design field. Furthermore to check the consistency of scoring molecules by all the chemists, i.e. the medicinal and computational chemists, we randomly include many times same molecules in the library. In this study, SYLVIA software was used to calculate synthetic accessibility score.

Synthetic accessibility is of high importance in early drug discovery stage but also in manufacturing processes of drug molecules. GMP (Good Manufacturing Practice) batch productions require being consistently reliable and reproducible in large scale chemical synthesis which usually prevent difficult synthetic routes to achieve low-cost manufacturing processes, high synthesis purity, quantity and quality. However, drugs with difficult synthesis steps were approved in the last few years such as Fuzeon (enfuvirtide) or several natural products [32]. The recent 2010 approval of oncology drug Halaven (eribulin) [33,34] by the FDA (Food and Drug Administration) [35] shows that highly synthetically complex drugs [36,37] can still be marketed despite economically manufacturing process challenges. In this study, we also analyzed the synthetic feasibility of marketed drugs using SYLVIA software as well as compounds in different clinical phases. The synthetic accessibility of large molecules, often derived from natural products is not covered due to their specific synthesis approach, and only small drug-like molecules are included in this study.

In addition to others molecular informatics tools [38], the evaluation of synthetic accessibility of virtual ligand database could be very useful to distinguish difficult versus easy-to-make compounds. The synthetic accessibility score can be used to prioritize compounds in order to get compound synthesized and tested more rapidly. In this context, the global drug design cycle time could be accelerated substantially.

2. Methods

2.1. Datasets

11 chemists, including 7 medicinal chemists and 4 computational chemists with several years of experience in drug discovery agreed to score 119 lead-like molecules based on their synthetic accessibility (SA). In previous studies, the comparison of synthetic accessibility and feasibility of compounds by medicinal chemists was always performed on a dataset obtained from external libraries. In this study, each molecule included in the dataset was made by one of the selected medicinal chemists; therefore at least one medicinal chemist has experience of known chemical synthesis route and available building blocks for each compound. All selected compounds have a molecular weight greater than 300 Da and are not reaction intermediates bearing any protecting group. Compounds made by each chemist were extracted from the Johnson & Johnson corporate database. SYLVIA synthetic accessibility score (SSc) was calculated on all compounds and divided into 4 bins ($SSc \leq 3$, $3 < SSc \leq 4$, $4 < SSc \leq 5$, $SSc > 5$) where high score indicates difficult compounds to synthesize. For each bin, the