Transforming data into knowledge for intelligent decision-making in early drug discovery

Paricharak, S.A.

Summary

This thesis describes various analyses of life science data with the aim of achieving efficiency gains in future experimental campaigns and novel insights into compound mode-of-action (i.e., the protein target modulated for the desired phenotypic effect). The increase in publicly available life science data has created opportunities for bioactivity modeling, and the role cheminformatics and bioinformatics play in the latter is discussed.

Chapter one describes the relevance of computational drug discovery. The fundamentals of bioactivity modeling are explained in detail, followed by an introduction of the topics discussed in this thesis. *Chapter two* is a literature review on data-driven approaches used for compound library design, hit triage and bioactivity modeling in high-throughput screening. High-throughput screening campaigns are routinely performed in pharmaceutical companies to explore activity profiles of chemical libraries for the identification of promising candidates for further investigation. In particular, the remarkable progress in the activity modeling area since the recent introduction of large-scale bioactivity-based compound similarity metrics is discussed in detail, outlining its significance in the field.

In *Chapter three*, the relevance of chemical space for bioactivity modeling is inspected. Chemicals can be described in terms of a set of characteristics (descriptor) that computers can easily use to assess similarity between molecules. Chemical diversity is a widely applied concept used to select structurally diverse subsets of molecules, often with the objective of maximizing the number of hits in biological screening. The extent to which the descriptors used in this study correlated in their assessment of molecular diversity across a number of compound sets ranging in size, diversity and origin is outlined in detail. Descriptors based on atom topology are shown to correlate well in rank-ordering compounds, whereas shape-based descriptors show weak correlation with other descriptor types. Finally, the descriptor "Bayes Affinity Fingerprints" which is based on predicted bioactivity profiles of compounds is shown to be most effective in selecting compound sets that are diverse in bioactivity space.

Chapter four illustrates the application of a computational method geared toward systematic compound prioritization, aimed at increasing the efficiency of compound screening campaigns over high-throughput screening campaigns performed currently in the pharmaceutical industry. The screening strategy described in this chapter consisted of the iterative selection of compounds chemically and biologically similar to actives identified in

multiple rounds of testing and was retrospectively validated on Novartis high-throughput screening data. Large efficiency gains were observed across assays covering a wide range of assay biology: by only screening 1% of the full screening collection, a consistent retrieval of diverse sets of compounds belonging to the top 0.5% was achieved. Employing this method can potentially lead to considerable savings in both time and resources.

Chapter five describes the data-driven derivation of an "informer compound set". Once screened, this set provides the most information on which yet untested compounds from the remainder of a large compound collection to screen next, irrespective of biological target. The derivation of this informer set involves the concept of *active learning*, which attempts to maximize the predictive power of the informer set. A retrospective validation of this set was performed on public high-throughput screening data, and an improvement in early retrieval of active compounds is observed for 38 out of 46 assays, increasing the success rate of smaller follow-up screens.

The final research chapter, *Chapter six*, represents a case study in the context of mode-of-action analysis of anti-malarial compounds identified in phenotypic screens by GlaxoSmithKline. Here, the application of two machine learning methods (Bayesian target prediction and proteochemometrics modeling) is illustrated for simultaneous polypharmacology and affinity predictions. Overall, 534 compounds were identified as dihydrofolate reductase inhibitors by the target prediction algorithm, while the proteochemometrics modeling approach identified 25, with an overlap of 23 compounds between both methods.

Finally, *Chapter seven* draws general conclusions from this thesis and provides future perspectives where some of my views on (early) drug discovery in academia and the pharmaceutical industry are discussed.