Chapter 5.18
Learning and Prediction of Complex Molecular Structure–Property Relationships:
Issues and Strategies for Modeling Intestinal Absorption for Drug Discovery

Rahul Singh
San Francisco State University, USA

ABSTRACT

The problem of modeling and predicting complex structure-property relationships, such as the absorption, distribution, metabolism, and excretion of putative drug molecules is a fundamental one in contemporary drug discovery. An accurate model can not only be used to predict the behavior of a molecule and understand how structural variations may influence molecular property, but also to identify regions of molecular space that hold promise in context of a specific investigation. However, a variety of factors contribute to the difficulty of constructing robust structure activity models for such complex properties. These include conceptual issues related to how well the true bio-chemical property is accounted for by formulation of the specific learning strategy, algorithmic issues associated with determining the proper molecular descriptors, access to small quantities of data, possibly on tens of molecules only, due to the high cost and complexity of the experimental process, and the complex nature of bio-chemical phenomena underlying the data. This chapter attempts to address this problem from the rudiments: the authors first identify and discuss the salient computational issues that span (and complicate) structure-property modeling formulations and present a brief review of the state-of-the-art. The authors then consider a specific problem: that of modeling intestinal drug absorption, where many of the aforementioned factors play a role. In addressing them, their solution uses a novel

DOI: 10.4018/978-1-60960-818-7.ch5.18

characterization of molecular space based on the notion of surface-based molecular similarity. This is followed by identifying a statistically relevant set of molecular descriptors, which along with an appropriate machine learning technique, is used to build the structure-property model. The authors propose simultaneous use of both ratio and ordinal error-measures for model construction and validation. The applicability of the approach is demonstrated in a real world case study.

INTRODUCTION

The recent past in human history has been witness to several significant events in the evolution of our understanding at the intersection of biology and medicine. Among others these include, the elucidation of the structure of the DNA, understanding the cell-cycle, cloning of proteins, advances in structure-elucidation techniques, development of rational drug design especially against well identified targets like angiotensin converting enzyme and protein kinases, and most recently, the sequencing of the human genome and mapping of the genomic DNA (Lander, 2001).

Considering the fact that all known commercial drugs today, interact with no more than 500 distinct targets, advances in genomics promise to provide a proliferation of targets that may not only lead to newer or improved therapeutics, but also open exciting avenues like individualized medicine. Somewhat simultaneously, recent developments in industrial robotics, combinatorial chemistry, and high-throughput screening have significantly increased the number of lead compounds that can be synthesized in pharmaceutical drug-discovery settings (Flickinger, 2001; McKinsey Lehman Brothers report 2001). Taken together, these factors may be assumed to point to both advancements in treatment and eradication of diseases as well as a significant reduction in the time-to-market (currently approximately 14 years on average per drug) and cost (currently 100-897 million dollars per drug, depending on the business model) of drug discovery.

Unfortunately, the trends from *pharmaceutical science* and industry differ considerably. A detailed study involving the pharmaceutical sector (McKinsey Lehman Brothers report 2001) accessed the impact of genomics on biopharmaceutical drug development. Broadly speaking, this study found that the cost and number of failures in drug discovery can be expected *to increase* in the immediate future. This startling result can be explained due to two factors. First, once a target is identified, it needs to be validated to establish its role in a disease. Moreover, its interactions with other genes/targets have to be identified as well, for example, by elucidating the pathways it is involved in. However, validation remains a complex, non-standardized process and the advancements in genomics have, till date, been more effective in increasing our capabilities in identifying new targets, rather than in validating them. This has typically resulted in many insufficiently validated targets being considered for *drug discovery*. Second, newer targets often require that newer classes of molecules be designed to interact with them. However, owing to the structural novelty of such molecules, historical data on their *pharmacokinetics* (influence of the human biological system on the drug molecule), pharmacodynamics (influence of the drug molecule on the human body), or toxicity profiles is scarce. Appropriate pharmacokinetics, pharmacodynamics, and toxicity characteristics are essential for a successful drug. However, these properties are typically tested for, in the later stages of drug discovery due to the associated time and cost. In turn, this leads to the increased possibility of late stage attrition if the pharmacology of a molecule is found to be undesirable.

It is increasingly being recognized that computational approaches can play a significant role in biology and drug discovery, not only at the level of data management, sequence comparison and analysis, and systems biology but also in model-