Genetic Programming and Other Machine Learning Approaches to Predict Median Oral Lethal Dose (LD_{50}) and Plasma Protein Binding Levels (%PPB) of Drugs

Francesco Archetti[1,2], Stefano Lanzeni[1], Enza Messina[1], and Leonardo Vanneschi[1]

[1] D.I.S.Co., Department of Computer Science and Communication
University of Milan-Bicocca, p.zza Ateneo Nuovo 1, 20126,Milan, Italy
{archetti,lanzeni,messina,vanneschi}@disco.unimib.it
[2] Consorzio Milano Ricerche
via Leopoldo Cicognara 7, 20100, Milan, Italy
archetti@milanoricerche.it

Abstract. Computational methods allowing reliable pharmacokinetics predictions for newly synthesized compounds are critically relevant for drug discovery and development. Here we present an empirical study focusing on various versions of Genetic Programming and other well known Machine Learning techniques to predict Median Oral Lethal Dose (LD_{50}) and Plasma Protein Binding (%PPB) levels. Since these two parameters respectively characterize the harmful effects and the distribution into human body of a drug, their accurate prediction is essential for the selection of effective molecules. The obtained results confirm that Genetic Programming is a promising technique for predicting pharmacokinetics parameters, both from the point of view of the accurateness and of the generalization ability.

1 Introduction

Because of recent advances in high throughput screening (HTS), pharmaceutical research is currently changing. In the traditional drug discovery process, when a target protein is identified and validated, the search of lead compounds begins with the design of a structural molecular fragment with therapeutic potency. Libraries of millions of chemical compounds similar to the identified effective fragment are then tested and ranked according to their specific biological activity. After these tests some candidate drugs are selected from the library for more specific development (see figure 1.a). In order to have a real pharmacological value, compounds have not only to show a good target binding, but also have to reach the target in vivo. In other words, it is necessary that compounds follow a proper route into the human body without causing toxic behaviors. It is interesting to remark that, both in 1991 [22] and in 2000 [7], a considerable fraction of attritions in pharmacological development were generated at the level of pharmacokinetics and toxicology, producing an unacceptable burden on the budget of drug companies (see figure 1.b). The necessity of deeply characterizing the behaviors of the pharmacological molecules in terms of adsorption, distribution, metabolism, excretion and toxicity processes (collectively referred to as ADMET [4]) makes the

E. Marchiori, J.H. Moore, and J.C. Rajapakse (Eds.): EvoBIO 2007, LNCS 4447, pp. 11–23, 2007.

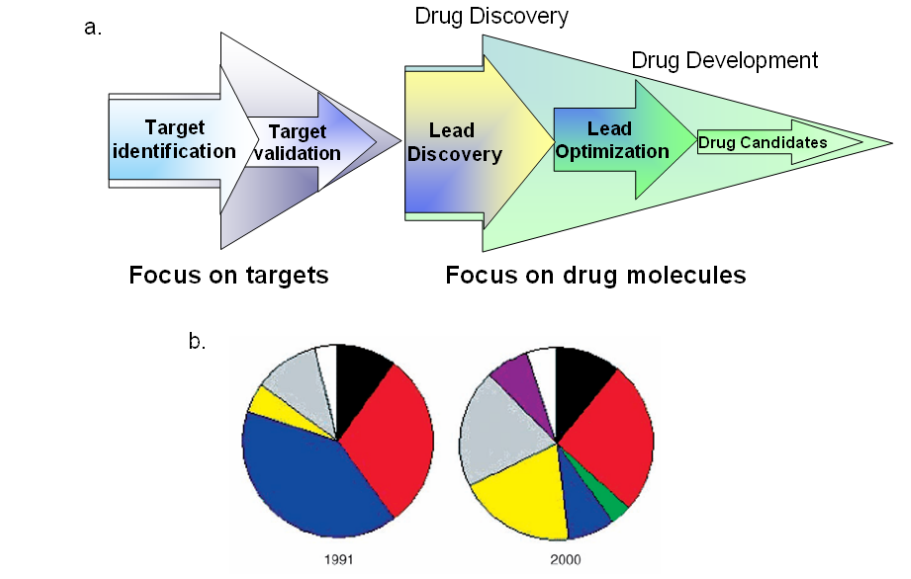

Fig. 1. a) The process of drug discovery from target protein identification to candidate drugs: the identification and validation of the target are followed by lead discovery and optimization. **b)** Reasons for failure in drug development in 1991 and 2000: clinical safety (black), efficacy (red), formulation (green), PK/bioavailability (blue), commercial (yellow), toxicology (gray), cost of goods (purple) and others (white).

development of computational tools applicable for pharmacokinetic profiling, enabling both the generation of reliable predictive models and the management of large and heterogeneous databases of outmost relevance [4] [25]. Reliable prediction tools allow the risk reduction of late-stage research failures, while reducing the number of cavies used in pharmacological research. In this paper, we empirically show that Genetic Programming (GP) [12] is a promising and valuable tool for predicting the values of Median Oral Lethal Dose (LD50) and Plasma Protein Binding (%PPB) levels. LD50 is one of the parameters measuring the toxicity of a given compound. More precisely, LD50 refers to the amount of compound required to kill 50% of the cavies. It is usually expressed as the number of milligrams of drug related to one kilogram of mass of cavies (mg/kg). Depending on the specific organism (rat, mice, dog, monkey and rabbit usually) and on the precise way of supplying (intravenous, subcutaneous, intraperitoneal, oral generally) chosen, it is possible to define a wide spectrum of LD50 experimental protocols. We consider the LD50 measured using rats as model organisms and supplying the compound orally. %PPB corresponds instead to the percentage of the initial drug dose which binds plasma proteins [13]. This measure is fundamental, both because blood circulation is the major vehicle of drug distribution into human body and because only free (unbound) drugs permeate the cellular membranes and reach the targets.

This paper is structured as follows: section 2 describes the mostly employed methods in literature for LD50 and %PPB levels predictions. In section 3, we describe the ML