available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/ejps

Development and validation of *in silico* models for estimating drug preformulation risk in PEG400/water and Tween80/water systems

Patrizia Crivori[a,*], *Amedea Morelli*[b], *Daniele Pezzetta*[b], *Maurizio Rocchetti*[a], *Italo Poggesi*[a,1]

[a] *Pharmacokinetic & Modeling/Modeling, Accelera, Nerviano Medical Sciences, viale Pasteur 10, 20014 Nerviano, Italy*
[b] *Preclinical Profiling, Accelera, Nerviano Medical Sciences, viale Pasteur 10, 20014 Nerviano, Italy*

ARTICLE INFO

Article history:
Received 10 October 2006
Received in revised form
19 April 2007
Accepted 25 June 2007
Published on line 21 August 2007

Keywords:
PEG400
Tween80
Solubility
In silico model
Formulation
QSPR
PLSD
E-state indices
ACDLabs
ChemGPS
Volsurf
Solubility prediction

ABSTRACT

Solubility is one of the most important properties of drug candidates for achieving the targeted plasma concentrations following oral dosing. Furthermore, the formulations adopted in the *in vivo* preclinical studies, for both oral and intravenous administrations, are usually solutions. To formulate compounds sparingly soluble in water, pharmaceutically acceptable cosolvents or surfactants are typically employed to increase solubility. Compounds poorly soluble also in these systems will likely show severe formulation issues. In such cases, relatively high amount of compounds, rarely available in the early preclinical phases, are needed to identify the most appropriate dosing vehicles. Hence, the purpose of this study was to build two computational models which, on the basis of the molecular structure, are able to predict the compound solubility in two vehicle systems (40% PEG400/water and 10% Tween80/water) used in our company as screening tools for anticipating potential formulation issues. The two models were developed using the solubility data obtained from the analysis of approximately 2000 chemically diverse compounds. The structural diversity and the drug-like space covered by these molecules were investigated using the ChemGPS methodology. The compounds were classified (high/low preformulation risk) based on the experimental solubility value range. A combination of descriptors (i.e. log*D* at two different pH, E-state indices and other 2D structural descriptors) was correlated to these classes using partial least squares discriminant (PLSD) analysis. The overall accuracy of each PLSD model applied to independent sets of compounds was approximately 78%. The accuracy reached when the models were used in combination to identify molecules with low preformulation risk in both systems was 83%. The models appeared a valuable tool for predicting the preformulation risk of drug candidates and consequently for identifying the most appropriate dosing vehicles to be further investigated before the first *in vivo* preclinical studies. Since only a small number of 2D descriptors is need to evaluate the preformulation risk classes, the models resulted easy to use and characterized by high throughput.

* *Corresponding author*. Tel.: +39 0331 581476; fax: +39 0331 581105.
E-mail address: patrizia.crivori@nervianoms.com (P. Crivori).
[1] Current address: Clinical Pharmacokinetics/Modeling and Simulations, Psychiatry Center of Excellence for Drug Discovery, GlaxoSmithKline R&D, Verona, Italy.
0928-0987/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.ejps.2007.06.008

1. Introduction

Solubility in a vehicle is one of the prerequisites needed for the development of new molecules in the drug industry. In principle, solutions are the most widely used formulations for the first *in vivo* preclinical studies. Also, solubility, together with permeability to the gastrointestinal mucosa, is one of the major determinants of the oral absorption that allows the achievement of the target plasma levels following oral dosing; therefore, it is one of the most important physicochemical properties considered in the early optimization process of oral drugs (Lipinski et al., 1997). Consequently, *in vitro* screens and *in silico* models have been implemented in drug discovery to estimate the aqueous solubility of new compounds and to design optimized drug libraries (Lipinski et al., 1997; Huuskonen et al., 2000b; Jorgensen and Duffy, 2000; Ran et al., 2001; Tetko et al., 2001a; Engkvist and Wrede, 2002; Gao et al., 2002; Butina and Gola, 2003; Cheng and Merz, 2003; Bergström et al., 2004).

In the pharmaceutical industry, cosolvents and/or surfactants are selected on the basis of safety (i.e. tolerability and reduced effects on pharmacokinetic and pharmacological properties) and technical aspects (i.e. producing chemically and physically stable solutions) to enhance the aqueous solubility of compounds poorly soluble in water (Yalkowsky, 1999a,b; Sweetana and Akers, 1996). Molecules that cannot be dissolved also in these conditions are unlikely to be easily formulated using standard vehicles. In such cases, the identification of the most appropriate dosing vehicles requires relatively high amount of compounds, rarely available in early preclinical phases, and labor-intensive investigations, which can cause a costly delay of the first *in vivo* preclinical studies.

On the basis of these considerations, the importance of an early screening assessing the potential preformulation risk of compounds was definitely recognized. In our company, medium throughput preformulation screens are applied for measuring the solubility of compounds in two commonly used systems; i.e. 40% polyethylene glycol 400 (PEG400) in water and 10% polysorbate 80 (Tween80) in water. The compositions of the mixtures were selected to obtain meaningful solubility information for the first *in vivo* studies and clear indications on the most appropriate modifier class to be further investigated in the subsequent formulation assays. In these screens, compounds soluble (i.e. solubility greater than 1.5 mg/ml) in at least one of the two systems can be easily formulated using standard vehicles. Thus, they are classified as molecules with low risk of formulation. On the contrary, compounds poorly soluble in both the mixtures (i.e. solubility less than 1.5 mg/ml) are likely to show high risk of formulation. In such cases, the use of different formulation strategies or the deprioritization of the first *in vivo* studies must be considered.

These *in vitro* screens and their results were thoroughly scrutinized and it was demonstrated that they provide an added value in the early discovery process. Due to this, it was felt appropriate to evaluate the possibility of using the historical data for developing *in silico* screens that are low-resource demanding and applicable to virtual compounds.

Whilst many groups of scientists focused their attention exclusively on developing models for anticipating water solubility issues (Lipinski et al., 1997; Huuskonen et al., 2000b; Jorgensen and Duffy, 2000; Ran et al., 2001; Tetko et al., 2001a; Engkvist and Wrede, 2002; Gao et al., 2002; Butina and Gola, 2003; Cheng and Merz, 2003; Bergström et al., 2004), the present paper describes one of the few models trying to predict potential formulation problems. Only recently Rytting et al. (2004, 2005) reported a series of quantitative structure–property relationships for predicting drug solubility in 25%, 50% and 75% PEG400/water mixtures.

Since our *in vitro* standardized and validated preformulation screens involved single-point medium throughput solubility determinations, the outcome of these assays was classes of preformulation risk. Hence, such qualitative data were considered in the models development. In a first step, the drug-like properties and structural diversity of the datasets utilized in this study were investigated using the ChemGPS (chemical global positioning system) (Oprea et al., 2000; Oprea and Gottfries, 2001a) tool combined with Volsurf descriptors (GPSVS) (Cruciani et al., 2000; Oprea et al., 2001b). Then, by selecting a reasonable set of 2D descriptors, the models were obtained by correlating the preformulation risk classes of about 2000 compounds to the molecular properties (2D descriptors) using partial least squares discriminant (PLSD) analysis. The performance of each PLSD model was evaluated by predicting the preformulation risk classes of 518 drug-like compounds not included in the original training sets. Moreover, we challenged our models by evaluating the performance of the combined models on the test set. Furthermore, the molecular properties that more likely influenced the drug solubility in the two different vehicles were also investigated.

2. Materials and methods

2.1. Materials

Dimetilsulfoxide (DMSO), polyethylene glycol 400 (PEG400), polyoxyethylene-sorbitan-mono-oleate (polysorbate80, Tween80), polytetrafluoroethylene (Teflon), acetonitrile and 12 commercial drugs were purchased from Sigma–Aldrich, Inc. (St. Louis, MO, USA). All other compounds were synthesized in-house. Only the free forms of the compounds were tested experimentally (i.e. no salts or other derivatives were involved in the current study). The compounds and reagents used were of analytical grade.

2.2. Experimental method

Seventy microliters of a 10 mM solution in dimethylsulphoxide (DMSO) of each compound was distributed twice in two 96-well plates (PP microplate 96K U form GREINER). Fifty microliters suspension of Teflon powder (12 μm diameter) in DMSO (20 mg/ml) was dispensed in each well of the microplates. Then, the microplates were inserted into the rotoevaporator (Christ Alpha RVC + Alpha 1–4) for solvent evaporation. The cakes formed after evaporation were re-suspended with 100 μl of 40% PEG400 in water (volume/volume) in one microplate and with 100 μl of 10% Tween80 in water (volume/volume) in the other. The microplates were stirred for 2 h on the Micromix 5 stirrer