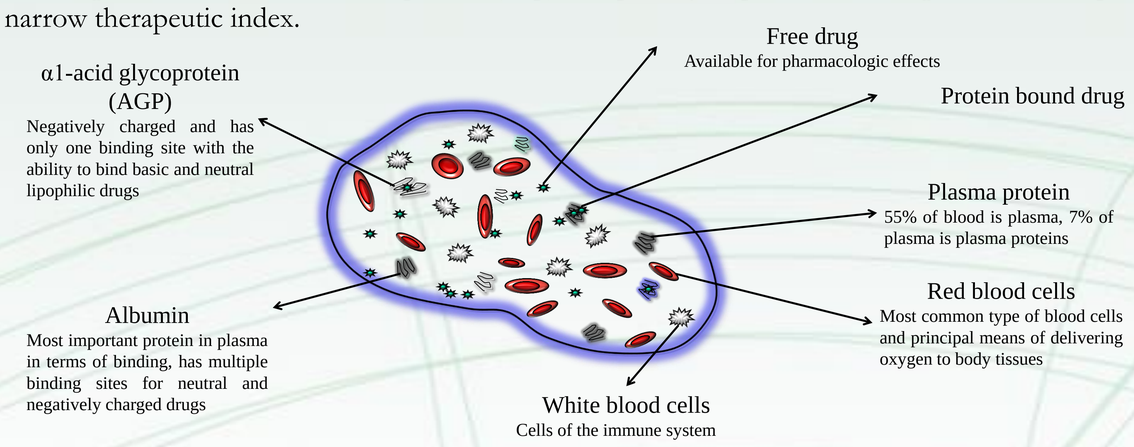
Computational models for improved estimation of highly plasma-protein-bound compounds

Jayeeta Ghosh, Michael Lawless, Marvin Waldman, Robert D. Clark, and Walter S. Woltosz

Simulations Plus, Inc., 42505 10th Street West, Lancaster, CA 93534, USA (www.simulations-plus.com)

Introduction

Plasma protein binding is an important parameter for characterizing the pharmacokinetics of drug candidates. Binding affinities vary tremendously among compounds, affecting the free fraction in blood, which consequently affects such pharmacokinetic properties as volume of distribution, clearance, bioavailability, and elimination. Only unbound drug can interact with target proteins, so knowing the fraction unbound in plasma (Fup) is especially important for drugs with low Fup and narrow therapeutic index.

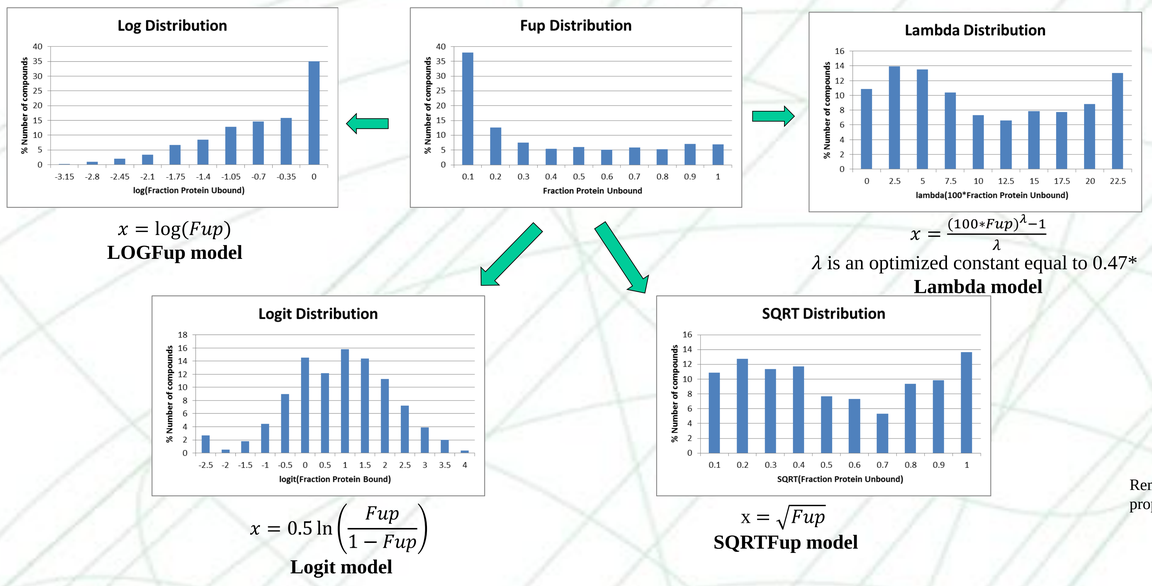


Schematic representation of whole blood

This study was undertaken to find the best regression model for estimating Fup by examining four data transformation techniques, with a particular focus on accurately predicting Fup for highly bound compounds without compromising the quality of estimations for the rest of the data set.

Model Building Process

Fup values for 791 compounds were collected from literature [1,2] and carefully curated. The data set spans a broad chemical space, including small drug-like compounds, pesticides used on food, and high-production volume chemicals. It is somewhat unbalanced in terms of binding, 38% of the compounds having Fup values less than 0.10. The large number of compounds with low Fup makes building useful models challenging, which led us to explore four different response transformations: logarithmic, Box-Cox [3], square root, and logit (for example [4]).



$x = \log(Fup)$
LOGFup model

$x = \frac{(100*Fup)^{\lambda}-1}{\lambda}$
λ is an optimized constant equal to 0.47*
Lambda model

$x = 0.5 \ln\left(\frac{Fup}{1-Fup}\right)$
Logit model

$x = \sqrt{Fup}$
SQRTFup model

Distributions of original and transformed data

Artificial neural network ensemble (ANNE) regression models were constructed using the ADMET Modeler™ module in ADMET Predictor™. A test set of 102 compounds was created using a Kohonen map and each model was trained on the remaining 689 compounds. A range of model architectures with varying numbers of neurons and descriptors was created for each transformation. The best model for each was selected based on training and test set statistics.

The performance of each model was evaluated for the low range (Fup < 0.1) as well as over the entire range of binding values. The RMSLE (root mean square log error) [5] is an important parameter in the low unbound range, as illustrated in the table below.

$$RMSLE = \sqrt{\frac{\sum_i (\log(pred_i) - \log(exp_i))^2}{Nobs}}$$

Molecule	Fup Pred.	Fup Exp.	Error	Log Error
Example 1	0.052	0.102	0.05	0.29
Example 2	0.903	0.953	0.05	0.02

Results

The Box-Cox transformation (Lambda model) and SQRTFup model were very similar because the λ value obtained (0.47) is very close to 0.5 (which is equivalent to square root transformation). Therefore, we do not report detailed results for the Box-Cox transformation.

All models identified lipophilicity as an important factor in plasma protein binding because S+logP (Simulations Plus' proprietary model for octanol-water partition coefficient) is a key descriptor in all models. Other important descriptors were fractions cationic and anionic; molal volume; dipole moment; and the fraction of aromatic bonds.

Models	# Neurons	# Descriptors	Key Descriptors
PrUnBnd	5	8	S+logP, Fcation, EqualEta, Fanion, VMcGowan, F_AromB, T_Dipole, T_Grav3
LogPrUnBnd	5	10	S+logP, Fcation, Fanion, F_AromB, EqualEta, N_IsolLP, VMcGowan, T_Grav3, MaxQ, T_Dipole
SQRTFup	4	9	S+logP, Fcation, Fanion, F_AromB, EqualEta, T_Dipole, N_IsolLP, VMcGowan, T_Grav3
Logit	6	10	S+logP, N_Iodine, N_AromR, Fanion, N_Sulfur, HBDoch, QAvgPos, EEM_XFh, EqualEta, M_RNG
Complexity of each model with best descriptors			

*In their original publication, Box and Cox proposed finding the optimal λ by a maximum-likelihood approach. We optimized λ by minimizing the Kolmogorov-Smirnov statistic (maximum distance between Cumulative Distribution Functions, CDFs) between the observed and normal CDFs of the λ-transformed data set .

Performance of Models

All predictions were transformed back to the original scale in order to compare the models in terms of RMSLE (root mean square log error), RMSE (root mean square error), MAE (mean absolute error, and R^2 (coefficient of determination) on the same scale. We considered statistics for the data set as a whole and for the highly bound compounds in particular in selecting the best model.

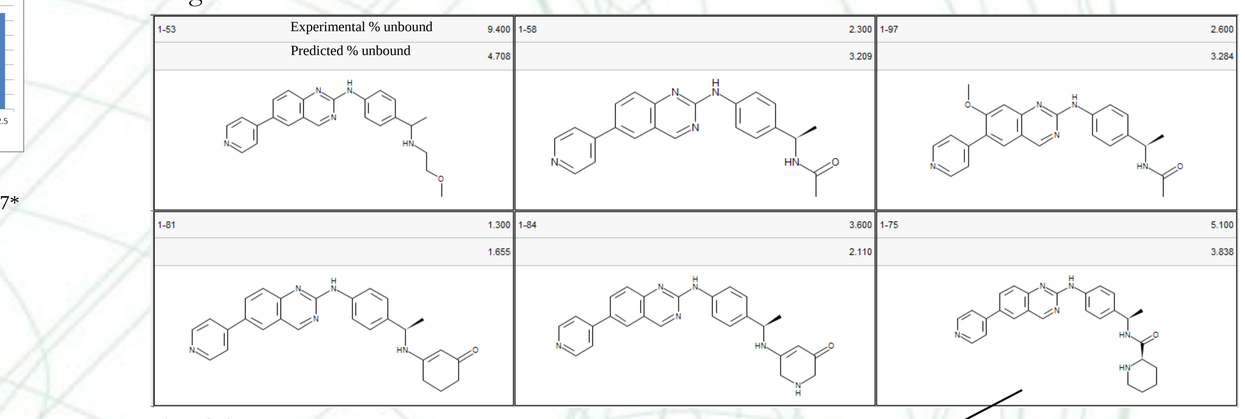
		Fup	LOGFup	SQRTFup	Logit
RMSLE	Whole set	0.78	0.43	0.46	0.45
	highly bound	1.19	0.57	0.67	0.62
RMSE	Whole set	0.176	0.202	0.179	0.176
	highly bound	0.137	0.077	0.105	0.104
MAE	Whole set	0.131	0.137	0.128	0.120
	highly bound	0.097	0.046	0.067	0.060
R^2	Whole set	0.71	0.66	0.71	0.72
	highly bound	0.13	0.18	0.17	0.17

Performance for each type of transformed regression models
Best performance statistics are circled for each statistic

The LOGFup and Logit based models have the best statistics. The LOGFup model performed best on highly bound compounds (RMSLE 0.57 and 0.62, RMSE 0.077 and 0.104, for LOGFup and Logit models, respectively), whereas the Logit model produced somewhat better statistics overall (RMSLE 0.43 and 0.45, RMSE 0.202 and 0.176, for LOGFup and Logit models, respectively). All regression models showed good predictivity on three external test sets [6-8] and two in-house test sets (data not shown). Our primary interest was on predicting values for highly bound compounds well, so the log transformed model was preferred to those obtained for the other transforms.

Example application

The LOGFup model was tested on 23 kinase inhibitors from Vasbinder *et al.* [9], whose measured % unbound values were not used in the model training. The predicted values are very close to the experimental ones (RMSE ~2.2% and MAE ~ 1.7%). Descriptor Sensitivity Analysis [10] suggested that one of the lead compounds in the study could be modified to a lower predicted plasma protein binding.

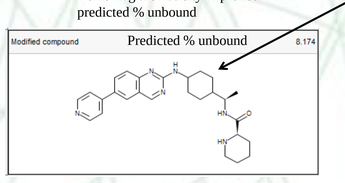


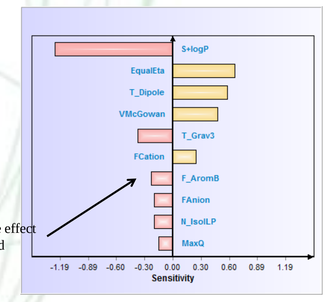
Removing aromaticity changed other properties in a way that improve Fup

Green: better
Red: worse

	1-75	Modified compound
S+logP	3.851	3.645
EqualEta	0.162	0.158
T_Dipole	2.39	2.851
VMcGowan	534.82	587.08
T_Grav3	18.686	18.863
FCation	0.695	0.761
F_AromB	0.605	0.447
FAnion	0	0
N_IsolLP	1	1
MaxQ	0.217	0.217

Removing aromaticity improved predicted % unbound





Aromaticity has inverse effect on predicted % unbound

Six lead compounds reported by Vasbinder *et al.*
Experimental and predicted percent unbound values are shown

Conclusion

In some cases, data sets like the one described here can be modeled more effectively by transforming the response data used to build the model. We compared models using different data transformation methods in terms of the whole data set as well as the highly bound portion of the data set.

The computational models developed in this study do a good job of estimating Fup. Overall, predictions for Fup are slightly better for the Logit model but the LOGFup model performed better on the highly bound compounds. Hence, the log transformation was used for the S+PrUnbnd model in ADMET Predictor™ version 7.0. This model can help in selecting lead candidates as well as in estimating parameters for PBPK studies.

References

1. R. S. Obach *et al.*, *Drug Metabolism and Disposition*, 36, (7) 1385-1405, 2008.
2. B. A. Wetmore *et al.*, *Toxicological Sci*, 125(1) 157-174, 2012.
3. G. E. P. Box and D. R. Cox, *J. Royal Statistical Society. Series B*, 26(2) 11-252, 1964.
4. X. Zhu *et al. Pharm. Res.* 30 1790-1798, 2013
5. S. Jachner *et al., J. Stat. Software*, 22(8) 2008
6. K. Yamazaki and M. Kanaoka, *J. Pharma. Sci*, 93(6) 1480-1494, 2004
7. J. R. Votano *et al., J. Med. Chem*, 49(24) 7169-7181, 2006.
8. Z. Zhivkova and I. Doytchinova, *J. Pharm. Sci*, 101(12) 4627-4641, 2012.
9. M. M. Vasbinder *et al., J. Med. Chem*, 56 (5) 1996-2015, 2013
10. R. Fraczkiewicz *et al. CICSJ Bulletin* 27 (4) 96-102, 2009

simulationsplus, inc.