Greater Than the Sum of Its Parts: Combining Models for Useful ADMET Prediction

Sean E. O'Brien[†] and Marcel J. de Groot*[,‡]

Department of Medicinal Informatics Structure and Design, Pfizer Global Research and Development, Sandwich, Kent, UK, and Accelrys Inc., 10188 Telesis Court, Suite 100, San Diego, California 92121

Received September 14, 2004

In silico ADMET (absorption, distribution, metabolism, excretion, and toxicity) models are important tools in combating late-stage attrition in the drug discovery process. This work shows how ADMET models can be combined to tailor predictions depending on one's needs. We demonstrate how the judicious use of data and considered combination of predictions can produce models that provide truly useful answers. This approach is illustrated with the prediction of hERG channel blocking and cytochrome P450 2D6 inhibition, where combination of two predictive models (with >80% of compounds correctly predicted) resulted in models with even better predictive values (with >90% of compounds correctly predicted for those classes of interest).

Introduction

Investigations into the causes of late-stage failures in drug development, performed in the 1990s, revealed that poor pharmacokinetics and toxicity were often responsible.[1,2] In an effort to reduce the time and expense of the drug discovery process it soon became apparent that early consideration of these areas was essential.[1−3] The need to know the absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties of drug candidates has propelled the development of numerous high-throughput screening methods. These screens provide data on synthesized molecules, but there is an increasing need for robust and accurate computational tools to aid in the design of new compounds and libraries with desirable ADMET profiles. Prioritization of hit and lead candidates can also be influenced by the prediction of their ADMET properties.[4,5] Therefore the drive toward the reduction of attrition is one in which in silico modeling can have a major impact.

The choice of how to build an ADMET model varies depending on the case and the circumstances − no one technique is consistently better than all others. Irrespective of the method used, the quality of the predictions must be assessed. There are a several ways in which this can be done,[6] but ultimately the model must be useful in the drug discovery context. Therefore, it is essential to determine how the model will be applied before proceeding.[7] The data source and validity are all important.[2] Curation of data according to quality and relevance will enhance the probability of obtaining good models. For example, when training models on high-quality, in-house data one can be confident that the predictions will be applicable to in-house chemistry. Once models are obtained, the predictions (from two or more models) can be combined to further suit the application.[8−10]

The data used to build ADMET models often come from high throughput in vitro screens. These data are effectively categorical (e.g. YES/NO, Positive/Negative) in nature, so classical QSAR (quantitative structure activity relationship) methodologies, predicting activity values, are rarely suitable. However, there are numerous techniques and descriptors available for modeling categorical data that have been shown to be useful for ADMET modeling.[6,7,11,12] In this work we report models built using both Neural Networks and Bayesian statistics, but the processes described can be applied to alternative modeling methods. We examine the results provided by both methods and demonstrate how combining the respective models can enhance results. In addition, we show how the resulting predictions can be tailored to suit the individual project or task requirements. The approach is illustrated with two significant targets for ADMET prediction: toxicity caused by blocking of the human ether-a-go-go related gene (hERG) K^+ channels, and inhibition of cytochrome P450 2D6 (CYP2D6).

Block of hERG K^+ channels by a variety of drugs has been linked to acquired long QT syndrome, a disorder of cardiac repolarization that predisposes to lethal arrhythmias. This is because hERG channels mediate the rapidly activating delayed rectifier K current (I_{Kr}) in the heart.[13] A number of structurally diverse compounds have been removed from the market due to drug-induced long QT.[14] Therefore identification of potential hERG channel blockers early in the drug discovery process is desirable. An assay that measures displacement of dofetilide from hERG binding can be used to determine the likelihood of a potential interaction with the hERG channel.[15]

Cytochrome P4502D6 is a polymorphic member of the P450 super-family.[16] It is absent in 5−9% of the Caucasian population, resulting in diminished metabolism of numerous drugs. When one compound inhibits CYP2D6, the subsequent decrease of metabolism of another compound can lead to unexpected drug−drug interactions.[17] This is due to accumulation of the latter

* To whom correspondence should be addressed. Phone: +44-1304-648746; Fax: +44-1304-651820; E-mail: marcel.degroot@pfizer.com.
[†] Accelrys Inc.
[‡] Pfizer Global Research and Development

10.1021/jm049254b CCC: $30.25
Published on Web 02/01/2005

Table 1. Model Statistics for Dofetilide Displacement Assay Prediction

model	sensitivity, %	specificity, %	concordance, %	Kappa
Neural Network	86	83	85	0.69
Bayesian	84	80	82	0.63
recover +ve	92	75	82	0.64
recover −ve	79	89	85	0.68
consensus-overall	79	75	76	0.62
consensus-predicted[a]	91	87	89	0.77

[a] 87% of positives and 86% of negatives predicted.

compound as it is not being metabolized. Therefore, inhibition of CYP2D6 is an unwanted feature in a drug candidate. Fluorescence screens can be used to determine the degree of CYP2D6 inhibition.[18]

Results and Discussion

Dofetilide Displacement. A Neural Network and a Bayesian model for the assay were constructed as detailed in Materials and Methods. All the results and statistics presented in this section relate to the activity predictions for the validation set of 11996 compounds (5000 positive, 6996 negative). The statistical measures and terminologies are defined in Materials and Methods. The results for both individual models are comparable, with the Neural Network being slightly better (Table 1). They both show a high overall concordance and predict a greater proportion of positives accurately than negatives.

Although the Neural Network and Bayesian models have similar prediction accuracies, they do not predict all the same compounds correctly. There is a large degree of overlap between both predictions, but there are compounds that are accurately classified by one method and not the other. This difference enables us to combine predictions from both models in each of three ways. The models can be combined such that:

1. A positive prediction is returned if *either* model predicts a compound to be positive. This is the "recover +ve" model as it tends to return more true positives.

2. A negative prediction is returned if *either* model predicts a compound to be negative. This is the "recover −ve" model as it tends to return more true negatives.

3. A prediction is only given if *both* models agree. This is the "consensus model". There are two sets of statistics that can be found for this model: the first pertaining to the accuracy of prediction of all the molecules in the validation set (consensus-overall); the second reflecting the accuracy of prediction of the set of compounds for which a classification is actually made i.e. all compounds for which both individual models agree (consensus-predicted).

The statistics for all of the above combinations are shown in Table 1 and depicted in Figure 1.

This results in a set of models that recover different sets of compounds. The recover +ve model greatly increases the number of positive compounds that are retrieved from the validation set. Only 8% of molecules that displace dofetilide are not found. However, this model also increases the false positive rate (by 5−8%), and 25% of negative compounds are misclassified. This model is useful when the aim is to discard any compound that blocks the hERG channel, as few positive compounds are progressed in the drug discovery process.

The recover −ve combination only classifies a compound as positive if both models agree. This cuts down the false positive rate (by 6−9%) but does mean that 21% of potential hERG binders are missed. This model is useful when the aim is not to remove potentially valuable compounds. The remaining molecules cover the widest chemical space that will not block the hERG channel.

The consensus model (consensus-overall) shows the lowest concordance of all the models (76%). However, when a prediction is made, 91% of positive compounds and 87% of negative compounds are accurately classified (Table 1). This consensus model provides significantly increased confidence in prediction, but 14% of the molecules have unassigned activities (as conflicting predictions have been made by the two individual models). Thus false positive and false negative rates are reduced compared to the individual Neural Network and Bayesian results. As a measure of the goodness of the model, the Kappa value of 0.62 for the consensus-overall model is lower than all the other models, but a Kappa of 0.77 for the compounds that were actually predicted

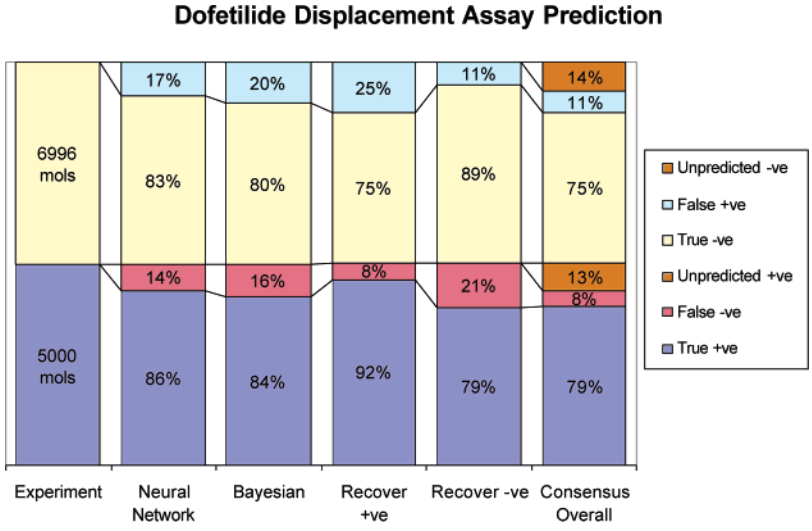

Figure 1. Statistics for dofetilide displacement assay prediction models.

Table 2. Selected Model Statistics for CYP2D6 Inhibition

model	sensitivity, %	specificity, %	concordance, %	Kappa
Neural_5	79	87	86	0.58
Neural_3	91	80	82	0.53
Bayesian	92	78	80	0.51
recover +ve[a]	95	73	77	0.47
recover −ve[b]	77	90	88	0.62
consensus-overall[a]	87	74	76	0.60
consensus-predicted[a,c]	100	99	99	0.97

[a] Model combines Neural_3 and Bayesian predictions. [b] Model combines Neural_5 and Bayesian predictions. [c] 87% of positives and 75% of negatives predicted.

is significantly higher. The consensus can be used in isolation for increased accuracy in prediction or as an indication of confidence in classification alongside the other, previously described, models.

CYP2D6 Inhibition. Two Neural Networks and one Bayesian model for the assay were constructed as detailed in Materials and Methods. All the results and statistics presented in this section relate to the validation set of 600 compounds (106 positive, 494 negative).

The use of five nodes for the Neural Network generates a model (Neural_5) that has a high concordance of 86% but predicts the negative compounds significantly better than the positive ones (Table 2). In contrast, Neural_3 and the Bayesian model classify the positive compounds with greater than 90% accuracy and produce few false negatives. The reduced concordance is due to a decreased specificity and the balance of the validation set: it contains nearly five times more negative than positive compounds.

As with the dofetilide assay models, there is a large degree of overlap between the predictions made by all models, but there are compounds that are accurately classified by one method and not the other. Hence, they can be combined to create a more useful model for CYP2D6 inhibition. As we have three predictions, we can have all the combinations shown in the previous section for any two of the three models (e.g. recover +ve models made from the combinations of Neural_3 and Neural_5 or Neural_3 and Bayesian or Neural_5 and Bayesian).

In addition these combinations can be made using all three models (e.g. recover −ve model that predicts a molecule as negative if *any* one of three models classifies a compound as negative). As there are three different predictions we can also construct a "voting model". In this scenario, a compound is given the value assigned to it by the majority of the models. All the above combinations were assessed, and the best results for each category are shown in Table 2 and Figure 2. The statistics for the voting model are not displayed in this paper, as it shows no discernible advantage over any of the other featured models. The recover +ve and the consensus models result from the combination of Neural_3 and the Bayesian classifications and the recover −ve is derived from the Neural_5 and the Bayesian models.

The recover +ve model has a low concordance of 77%, but it correctly classifies 95% of CYP2D6 inhibitors. However, the false positives are significant and only 74% of negative compounds are correctly classified. The ability of the recover +ve model to identify almost all the inhibitors of CYP2D6 makes this extremely powerful.

The recover −ve model correctly assigns 90% of negative compounds and produces a 10% false positive rate. This is better than all the individual models and is useful when wishing to identify definite inhibitors.

The consensus model is constructed from the Neural_3 and Bayesian models. It has a worse concordance than all other models and fails to predict 13% of positives and 25% of negatives. However, within the limits of the consensus model, it provides exceedingly accurate predictions. Every CYP2D6 inhibitor is correctly predicted, and only 5/368 noninhibitors are misclassified (Table 2). This model gives predictions in which we can have extremely high confidence (but only when a classification is made). It can also be used as a stand-alone predictor or in conjunction with another CYP2D6 model to provide confidence in prediction.

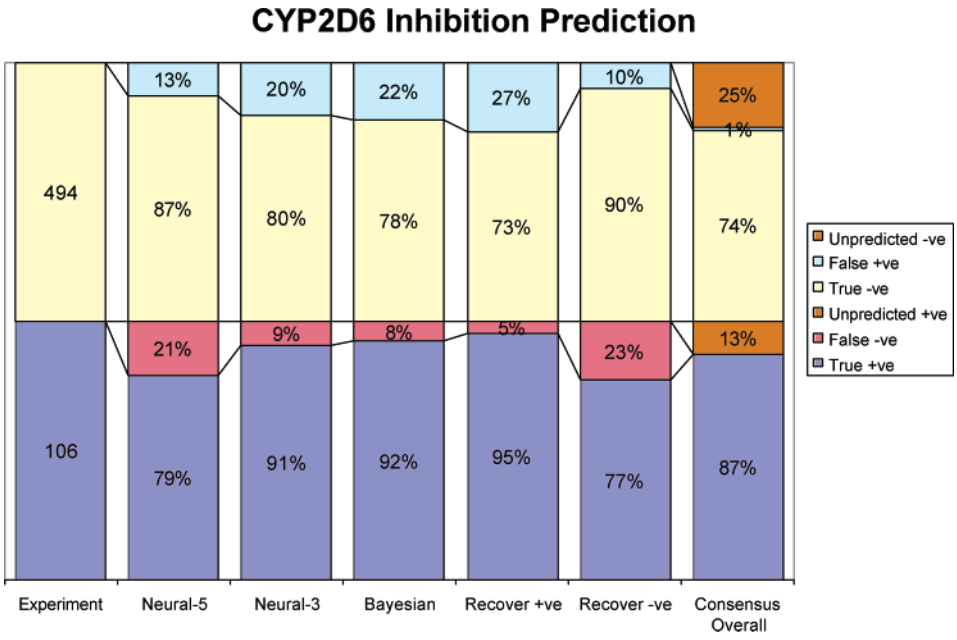

Figure 2. Statistics for selected CYP2D6 inhibition prediction models.

Conclusions

We show that there are several ways in which ADMET models can be used to tailor predictions depending on one's needs, i.e. give the best overall models, or focus on recovering the positive or negative predictions. The selection of which actual model, or combination of models, should be used will be influenced by the nature of the project goals and the quality of the available data. In the absence of a completely accurate classification method, the judicious use of data and considered combination of predictions, models that provide truly useful answers can be produced. In the two examples shown here, combination of two predictive models (>80% of compounds correctly predicted) resulted in models with even better predictive values. These combined models allow the correct prediction of >90% of compounds for those classes of interest, either positive or negative.

Materials and Methods

Dofetilide Displacement. In-house data[15] on 58963 compounds were split randomly 80/20 (46967/11996) to obtain training and validation sets, respectively. The training set comprised of 20200 compounds that displaced dofetilide and 26767 that did not.[19] The values used to determine classification into positive or negative were decided in consultation with project biologists. We refer to the former as positives and the latter as negatives. The validation set contained the same ratio of positive to negative molecules as the training set.

CYP2D6 Inhibition. IC_{50} data[18] from 2410 compounds were obtained from both in-house and CEREP (BioPrint) measurements and were split randomly 75/25 (1810/600) for training and validation sets. The training set comprised of 431 compounds that inhibit CYP2D6 and 1979 that do not.[20] We refer to the former as positives and the latter as negatives. The validation set contained the same ratio of positive to negative molecules as the training set. The values used to determine classification into positive or negative were decided in consultation with project drug metabolism experts.

Statistics are shown for the validation sets throughout this work. For each data set, models were constructed using the Neural Network module in Cerius[2 21] and the Bayesian modeling module in Pipeline Pilot.[22]

The back-propagation Neural Networks as implemented in Cerius[2] allow the mixing of numerical and categorical descriptors as well as 2-D fingerprints. In this instance, the E-state keys[23] and Barnard 4096-bit fingerprints[24] were used as descriptors for each molecule. The model building and prediction were performed through the use of the binary data files (BDF), as this permitted the easy manipulation of 46967 compounds. The number of nodes in the hidden layer was varied between 1 and 10, and the most useful models were chosen: one for hERG (five nodes) and two for CYP2D6 (three nodes and five nodes, referred to as Neural_3 and Neural_5, respectively). Additional alteration of parameters did not improve the models, so all the remaining default settings were used. During model generation, 10% of the training set is reserved as a test set. The minimization method uses this test set to decide when it would be prudent to stop training and also which values of the weights visited during training will be final. This measure attempts to prevent overtraining of the Neural Network. The generation of the hERG model took 1 day on a 600 MHz SGI Octane while the construction of the CYP2D6 models took a few minutes. Subsequent predictions using these models, including descriptor calculation, proceed at about 40 molecules per second.

The Bayesian models were constructed using the modified naïve Bayesian statistics implemented in Pipeline Pilot. Model building was carried out using the datasets described above and using FCFP_6 (functional class fingerprints, where atoms are abstracted to the role they play in the molecule), AlogP, Molecular Weight, and the counts of hydrogen bond acceptors and donors. A more extensive description of this Bayesian implementation is available elsewhere.[25] Model construction took several minutes on a dual-CPU Windows 2000 server (2.4 GHz). Predictions using these models have a throughput of approximately 25 molecules per second.

We evaluated each of the models using the following statistical measures:

1. Sensitivity: the percent of positives correctly predicted positive.

2. Specificity: the percent of negatives correctly predicted negative.

3. Concordance: the percent of compounds correctly classified.

4. Kappa: a weighted kappa statistic.[26,27] When Kappa equals 0, the model is equivalent to that expected by chance. When it equals 1, there is perfect agreement between actual and predicted values. The stronger this agreement, the higher the value of Kappa.

In addition, the figures illustrate the proportion of false positives (negative compounds classified as positive) and the proportion of false negatives (positive compounds classified as negatives) returned by each model.

References

(1) van de Waterbeemd, H.; Gifford, E. ADMET In Silico Modelling: Towards Prediction Paradise. *Nat. Rev. Drug Discuss.* **2003**, *2*, 192−204.

(2) Lombardo, F.; Gifford, E.; Shalaeva, M. In Silico ADME Prediction: Data, Models, Facts and Myths. *Mini Rev. Med. Chem.* **2003**, *3*, 861−875.

(3) Oprea, T. I. Virtual Screening in Lead Discovery. *Molecules* **2002**, *7*, 51−62.

(4) Hou, T.; Xu, X. Recent Development and Application of Virtual Screening in Drug Discovery: An Overview. *Curr. Pharm. Des.* **2003**, *10*, 1011−1033.

(5) Pirard, B. Computational Methods for the Identification and Optimisation of High Quality Leads. *Comb. Chem. High Throughput Screening* **2004**, *7*, 271−280.

(6) Migliavacca, E. Applied Introduction to Multivariate Methods Used in Drug Design. *Mini Rev. Med. Chem.* **2003**, *3*, 831−843.

(7) O'Brien, S. E.; de Groot, M. J. Recursive Partitioning, Models and Statistics: What Can We Extract from Categorical Data? Presented at the 227th National Meeting of the American Chemical Society, Mar 27−Apr 1, 2004 Anaheim, CA.

(8) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P. et al. Random Forest: A Classification and Regression tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947−1958.

(9) van Rhee, M. A. Use of Recursion Forests in the Sequential Screening Process: Concensus Selection by Multiple Recursion Trees. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 941−948.

(10) Banik, G. M. In Silico ADME-Tox predcition: The more, the merrier. *Curr. Drug Discuss.* **2004**, 31−34.

(11) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, 2000.

(12) Moore, A. Carnegie Mellon, School of Computer Science, http://www-2.cs.cmu.edu/~awm/tutorials/list.html.

(13) Sanguinetti, M. C.; Jiang, C.; Curran, M. E.; Keating, M. T. A mechanistic link between an inherited and an acquired cardiac arrthytmia: HERG encodes the IKr potassium channel. *Cell* **1995**, *81*, 299−307.

(14) Fermini, B.; Fossa, A. A. The impact of drug-induced QT interval prolongation on drug discovery and development. *Nat. Rev. Drug Discuss.* **2003**, *2*, 439−447.

(15) Pfizer Patent, WO 03/021271, 2003.

(16) Nelson, D. R.; Koymans, L.; Kamataki, T.; Stegeman, J. J.; Feyereisen, R. et al. P450 Superfamily: Update on New Sequences, Gene Mapping, Accession Numbers and Nomenclature. *Pharmacogen.* **1996**, *6*, 1−42.

(17) de Groot, M. J.; Ackland, M. J.; Horne, V. A.; Alex, A. A.; Jones, B. C. Novel Approach to Predicting P450 Mediated Drug Metabolism. The Development of a Combined Protein and Pharmacophore Model for CYP2D6. *J. Med. Chem.* **1999**, *42*, 1515−1524.

(18) Cohen, L. H.; Eremley, M. J.; Raunig, D.; Vaz, A. D. N. In Vitro Drug Interactions of Cytochrome P450: An Evaluation of

Fluorogenic to Conventional Substrates. *Drug Metabol. Dispos.* **2003**, *31*, 1005−1015.
(19) The cutoff for activity was 20 μM (IC_{50}).
(20) The cutoff for activity was 3 μM (IC_{50}).
(21) Accelrys Inc., Cerius^2; version 4.91, San Diego, CA.
(22) Scitegic Inc. Pipeline Pilot; version 3.0.6.0, San Diego, CA.
(23) Kier, L. B.; Hall, L. H. An Electrotopological-State Index for Atoms in Molecules. *Pharm. Res.* **1990**, *7*, 801−807.
(24) Downs, G. M.; Gill, G. S.; Willett, P.; Walsh, P. Automated descriptor selection and hyper structure generation to assist SAR studies. *SAR QSAR Environ. Res.* **1995**, *3*, 253−264.
(25) Maliski, E. G.; Gallant, P.; Rogers, D. Classification of Kinase Inhibitors Using a Bayesian Model. *J. Med. Chem.* **2004**, *47*, 4463−4470.
(26) Cohen, J. A Coefficient of Agreement for Nominal Data. *Educ. Psych. Measurement* **1960**, *20*, 37−46.
(27) Stokes, M. E.; Davis, C. S.; Koch, G. G. *Categorical Data Analysis Using the SAS System*; 2nd ed.; SAS Publishing: Cary, NC, 2000.

JM049254B