Available online at www.sciencedirect.com
SCIENCE DIRECT®
European Journal of Medicinal Chemistry 38 (2003) 427–431
EUROPEAN JOURNAL OF MEDICINAL CHEMISTRY
www.elsevier.com/locate/ejmech

Prediction of oral bioavailability by adaptive fuzzy partitioning

Marco Pintore [a], Han van de Waterbeemd [b], Nadège Piclin [a], Jacques R. Chrétien [a,*]

[a] BioChemics Consulting, Innovation Center, 16, rue Leonard de Vinci, 45074 Orleans cedex 2, France
[b] Pfizer Global Research and Development, PDM, IPC 351, Ramsgate Road, Sandwich, Kent CT13 9NJ, UK

Received 14 October 2002; received in revised form 9 January 2003; accepted 27 January 2003

Abstract

An adaptive fuzzy partition (AFP) algorithm was applied on two bioavailability data sets subdivided into four ranges of activity. A large set of molecular descriptors was tested and the most relevant parameters were selected with help of a procedure based on genetic algorithm concepts and stepwise method. After building several AFP models on a training set, the best ones were able to predict correctly 75% of the validation set compounds. Furthermore, an improvement of about 15% in the validation results was got, on the same data set, as regard to other prediction methods. The importance to work with data sets including a large molecular diversity, and to use tools able to manage it, was also shown. The prediction power was increased up to 25% employing a data set with a better-optimised molecular diversity.

Keywords: Database mining; Fuzzy logic; Oral bioavailability; Molecular diversity

1. Introduction

Combinatorial chemistry technologies and high throughput screening (HTS) are simultaneously used by all major pharmaceutical companies in order to increase the possibility of finding new leads [1,2]. However, many lead compounds fail to progress into the clinic phase because they do not show satisfactory pharmacokinetic properties, such as oral absorption, bioavailability, volume of distribution, and metabolic stability (ADME properties) [3,4]. But screening large series of compounds for these properties, especially using animal tests, requires too many resources that limit its utility in Drug Discovery. Furthermore, extrapolations from in vitro or in vivo data to human bioavailability often lead to inaccurate predictions. Computational models able to pre-screen ADME properties can serve as a filter in the design and screening of combinatorial libraries [5–7].

Several structure-bioavailability relationship models have been proposed in the last few years [7–9]. Probably the best known of these studies is the 'rule-of-five' developed by Lipinski et al. [10]. Based on four molecular descriptors, this empirical approach generates an alert about potential absorption problems if two of the following conditions are satisfied: (1) molecular weight >500; (2) number of hydrogen-bond acceptors >10; (3) number of hydrogen-bond donors >5; (4) calculated log $P>5.0$.

More recently, Veber et al. proposed two other descriptors and suggested that 'compounds which meet only the two criteria of (1) 10 or fewer rotatable bonds and (2) polar surface area equal to or less 140 $Å^2$ (or 12 or fewer H-bond donors and acceptors) will have a high probability of good oral bioavailability in rat' [11]. However, further validation has to be furnished in the evaluation of human bioavailability.

Yoshida and Topliss proposed a classification model to predict bioavailability by using a data set of 272 drugs, finger-print and pharmacokinetics descriptors, with help of a method named ORMUCS (ordered multicategorical classification method using the simplex

* Corresponding author.
E-mail address: jacques.chretien@univ-orleans.fr (J.R. Chrétien).

0223-5234/03/$ - see front matter
doi:10.1016/S0223-5234(03)00052-7

technique) [12]. This approach, after dividing the bioavailability data in 4 classes, allowed the authors to get a correct classification rate of 71% for the training set and 60% for the 40 compounds included in the test set.

A quantitative structure-bioavailability relationship (QSBR) model was developed by Andrews et al. on a data set including 591 compounds [13]. A stepwise regression procedure was used to relate oral bioavailability in humans and structural fragments in drugs. Compared to the 'rule-of-five' of Lipinski, this model allowed to reduce sensibly the prediction of false negatives and positives.

Finally, Bains et al. proposed evolutionary and adaptive methods for classifying drug bioavailability into 'high' and 'low' classes [14], and showed that obtaining predictive models on the basis of the molecular structure alone is solvable.

Innovative concepts for correlating molecular structures with biological activities are represented by fuzzy logic (FL) [15]. In fact, FL methods based on the possibility to handle the 'concept of partial truth', provide interesting solutions to classification problems within the context of imprecise categories, in which ADME properties can be included. Fuzzy classification represents the boundaries between neighbouring classes as continuous, assigning to the compounds a degree of membership of each class. FL has been widely used in the field of process control, where the idea is to convert human expert knowledge into fuzzy rules, and it is able to extract relevant structure-activity relationships (SAR) from a database, without a priori knowledge. Furthermore, these methods were successfully applied in the fields of olfaction [16], toxicity [17], and medicinal chemistry. For example, a data set of about 400 molecules actives in the central nervous system and divided in eight receptor classes, was correctly classified with a prediction ratio of 83% [18].

A first trial to apply FL in bioavailability classification was implemented by Hirono et al. [19], which analysed a data set of 188 compounds, divided into three classes, by a fuzzy adaptive least-squares method. The prediction results were relatively satisfactory, but three separated models were derived to classify non-aromatic, aromatic, and heteroaromatic molecules.

The aim of the current project was then to apply a FL procedure, called adaptive fuzzy partition (AFP) [16], to two bioavailability data sets, in order to build robust and general screening models. A large set of molecular descriptors was examined and the most relevant parameters were selected by a procedure combining genetic algorithm concepts and a stepwise technique (GA/SW) [20]. To evaluate the prediction ability of AFP as regard to other classification techniques, the prediction results were compared with the scores obtained by Yoshida and Topliss on the same bioavailability data set [12].

2. Materials and methods

2.1. Compound selection

A first data set of 272 drugs was built by exploiting data, concerning bioavailability in healthy human subjects, included in the work of Yoshida and Topliss [12]. Bioavailability (F) is defined as the percentage of an administered dose of a parent compound reaching the systemic circulation after oral administration. The drugs were divided into four classes according to the following bioavailability intervals reported in the original work: $F \leqslant 20\%$ for class 1; $20 < F < 50\%$ for class 2; $50 \leqslant F < 80\%$ for class 3; $F \geqslant 80$ for class 4. For simplicity, this compound series was named 'Topliss data set', and the same training and validation sets defined in the original work were adopted.

Another set of 235 molecules, not included in the Topliss data set, were extracted from four Refs. [21–24]. The same bioavailability ranges previously defined were used to classify the drugs. A test set of 75 molecules was randomly extracted from this data set and each class included at least 17 compounds. The remaining molecules were added to the Topliss data set and the global molecular series of 432 compounds was named 'reference data set'.

2.2. Molecular descriptor selection

General molecular descriptors have proved a good compromise for data mining in large databases in terms of efficiency. The advantage of these descriptors is their ability to take into account the main features of each molecule. The bioavailability data set was distributed within a hyperspace defined by 164 molecular descriptors, including constitutional, information, topological, electrotopological, physicochemical, and electronic parameters [25–27]. More details about the different molecular descriptors used can be found in Ref. [18].

To select the best parameters for separating and assessing the data set compounds, a method based on GA was used [28]. GA are very effective for exploratory search, applicable to problems where little knowledge is available, but they are not particularly suitable for local search. Then, a stepwise approach was combined with GA in order to reach local convergence, as it is quick and adapted to finding a solution in 'promising' areas already identified [20,29].

Finally, a specific index was derived by the fuzzy clustering method [15] to evaluate the fitness function. Furthermore, to prevent over-fitting and a poor generalisation, a cross validation procedure was included in the algorithm during the selection procedure, randomly dividing the database into training and test sets. The fitness score of each set of descriptors derives from the

Table 1
Representation of the most relevant descriptor selected for by GA/SW for the Topliss and reference data set

ID	Symbol	Definition	Family
Topliss data set			
1	X2	simple 2nd order chi index	topological
2	Xp6	simple 6th order path chi index	topological
3	SssssC	sum of all (>C<) E-state values	electrotopological
4	Shvin	sum of all H E-state values for (=CH–)	electrotopological
5	Shbint2	E-state of internal H bonds (2 path length)	electrotopological
6	Shbint5	E-state of internal H bonds (5 path length)	electrotopological
7	Idwbar	Bonchev–Trinajsti mean information content	information
8	Idcbar	Idwbar based on path 2 counts	information
9	TTs(4)simple	total simple topological index of 4th order	topological
10	Δ log D	log D(pH 6.5) − log D(pH 7.4)	physicochemical
Reference data set			
1	Qsv	polarity parameter	electronic
2	NumHBd	number of H-bond donors	constitutional
3	Xp10	simple 10th order path chi index	topological
4	Xvch3	valence 3rd order chain chi index	topological
5	Dxvp5	difference valence path 5th order chi index	topological
6	SssssC	sum of all C E-state values	electrotopological
7	SsssNH	sum of all NH E-state values	electrotopological
8	Shbint7	sum of E-state related to H-bonds	electrotopological
9	log P	lipophilicity at pH 7.0	physicochemical

combination of the scores of both the training and test sets.

More details about the strategy of molecular descriptor selection proposed and the proprietary software used can be found in Ref. [20].

2.3. Adaptive fuzzy partition

AFP is a supervised classification method implementing a fuzzy partition algorithm [30]. It models relations between molecular descriptors and chemical activities by dynamically dividing the descriptor space into a set of fuzzy partitioned subspaces. The aim of the algorithm is to select the descriptor and the cut position which allow to get the maximal difference between the two fuzzy rule scores generated by the new subspaces. The score is determined by the weighted average of the chemical activity values in an active subspace A and in its neighbouring subspaces.

Indicating with $P(x_1,\ldots x_n)$ a molecular vector in a n-dimensional descriptor hyperspace, a *rule* for a subspace S_k is defined by [31]:

if x_1 is associated with $\mu_{1k}(x_1)$ **and** x_2 is associated with $\mu_{2k}(x_2)\ldots$ **and** x_N is associated with $\mu_{Nk}(x_N) \Rightarrow$ the score of the activity O for P is O_k (1)

where x_i represents the value of the ith descriptor for the molecule P, μ_{ik} is the membership function related to the descriptor i for the subspace k, and O_k is the biological activity value related to the subspace S_k. The 'and' of the fuzzy rule is represented by the *Min* operator [32], which selects the minimal value amidst all the μ_{ik} components.

All the rules created during the fuzzy procedure are considered to establish the model between descriptor hyperspace and biological activities. After establishing the AFP model, a centroid defuzzification procedure [33] determines the chemical activity of a test molecule. All the subspaces k are considered and the general formula to compute the score of the activity O for a generic molecule P is:

$$O(P) = \frac{\sum_{k=1}^{N_{subsp}} (Min_i^N \mu_{ik}(x_i)_P)(O_k)}{\sum_{k=1}^{N_{subsp}} (Min_i^N \mu_{ik}(x_i)_P)} \quad (2)$$

where N is the total number of descriptors and N_subsp represents the total number of subspaces.More details about the AFP method can be found in Ref. [16].

3. Results and discussion

A first bioavailability model was established on the Topliss data set. The GA/SW technique was applied on the 232 compounds of the training set and ten relevant descriptors, with a wide-spanning diversity, were selected (Table 1). They include connectivity and information indices, three descriptors relative to hydrogen bonding, and a lipophilicity parameter. Several AFP models were developed on the training set compounds, distributed in the 10D descriptor hyperspace, and

validated by comparing the predicted and experimental classes of the 40 compounds included in the validation set. For each compound, the method allowed to get its degrees of membership of the different bioavailability classes within a 0–1 range. The validation results for the best AFP model are shown in Table 2.

Comparing the validation data associated with the AFP model and the results reported in the paper of Yoshida and Topliss, AFP method improved the power prediction of about 10–15% on the training and validation set molecules. Furthermore, also the prediction ratio on the compounds correctly classified ±one class was sensibly increased, from 95 to 99.8%; amidst the 40 test set compounds, only clomethiazole, affected to class 1, was completely missed.

Then, the AFP model built on this first training set was applied on the test set defined in Section 2.1. Table 2 shows a validation score of 40% and, more particularly, the less bioavailable classes are very badly predicted. The reason of these results can be found in the poor superposition between the structural space covered by the Topliss training set and the test set. After projecting both sets in a self organising map (SOM) [34] derived from the 10D descriptor hyperspace, a statistical analysis show that the Topliss training set recovers only 60% of the zones occupied by the test set.

The next phase of the work consisted then in selecting a new training set, more representative of the drug population. The 432 molecules included in the reference data set were submitted to a rational selection based on GA/SW and SOM, in order to define the training and validation sets used to build and validate the SAR models. The GA/SW procedure selected nine relevant descriptors reported in Table 1. Most descriptors are topological and electro-topological parameters. Furthermore, it is very interesting to observe the presence of log P and the number of H-bond donor descriptors, which are two parameters included in 'the rule-of-5' of Lipinski. After projecting the related 9D hyperspace in a SOM chart, 352 compounds were selected by exploiting all the regions of the chart and included in the training set. The best AFP models developed show scores of about 70% by predicting the training and validation sets (Table 2). These results are satisfactory, on account of the high variability affecting the experimental procedures used to calculate the bioavailability scores for this work data set.

But, more important, the test set validation shows that the predictive power, by using the training set derived by SOM from the reference data set, was improved of about 25% compared to the models previously developed with the molecules included only in the Topliss data set. In fact, the main object of the SAR procedures does not consist in getting impressive scores by predicting training and validation sets, but in developing robust models able to predict correctly also test sets never involved in the building procedures.

4. Conclusion

Combinatorial chemistry and HTS techniques allow to generate each year an impressive amount of new lead compounds, which will be successively converted in compounds with an in vivo therapeutic efficiency worthy of a potential drug candidate. But most of these drugs are intended for oral therapy. Then, there is a need to incorporate, already in the first phases of the drug discovery, procedures able to predict oral availability of a given molecule, for not wasting considerable resources.

Systematic experimental determination of oral activity would require as many resources as optimising in vitro activity. Then, computational methods can offer interesting alternatives, by analysing large amounts of information and automatically extracting new knowl-

Table 2
Statistical values defining the robustness of the AFP model developed on the Topliss and reference data sets

Class	Training set (%)	Validation set (%)	Test set (%)
Topliss data set-272 compounds			
1	73 (43)	63 (53)	19
2	70 (52)	71 (53)	19
3	87 (84)	75 (64	38
4	89 (89)	85 (64)	53
All classes	*82 (71)*	*75 (60)*	*40*
Reference data set-432 compounds			
1	69	70	64
2	62	58	55
3	71	69	62
4	81	78	70
All classes	*70*	*68*	*64*

The scores obtained by Yoshida and Topliss on the 'Topliss' data set are indicated between brackets. Better results derived from the reference data set underline the necessity of a rational selection of the training set.

edge from databases. Amidst these methods, FL concepts constitute an interesting approach to the virtual screening of chemical libraries and they were already successfully applied in the fields of medicinal chemistry, ecotoxicity and olfaction [16–18].

In this study, an AFP algorithm was applied on two data sets for oral bioavailability, divided into four ranges of activity. The AFP method consists in modelling molecular descriptor—activity relationships by dynamically dividing the descriptor hyperspace into a set of fuzzy subspaces. After computing a large set of molecular descriptors, the most relevant parameters were selected with help of a procedure based on GA concepts and stepwise method.

Several AFP models were built and, for the best one, the experimental classes were predicted correctly for 75% of the validation set compounds. Furthermore, the prediction score was improved of 15%, on the same data set, as regard to other prediction methods [12]. The importance to work with data sets including a large molecular diversity and to use tools able to manage it was clearly demonstrated. The prediction power was increased up to 25% by employing a training set with a better-optimised molecular diversity.

These preliminary results are very encouraging, also considering the high variability affecting the experimental procedures in the area of ADME and the complexity of the phenomena related. Further improvements of these models can be obtained by increasing the molecular diversity in the data sets.

Furthermore, work is underway to evaluate different descriptors, such as fingerprints and, above all, Volsurf parameters [35]. In fact, the latter descriptors are linked to main properties that influence bioavailability, such as transcellular permeability and intrinsic solubility [36], and that are badly encoded by the descriptors used in this work.

References

[1] G. Lowe, Chem. Soc. Rev. 24 (1995) 309–317.
[2] S. Borman, C&EN 77 (1999) 33–48.
[3] P.J. Eddershaw, A.P. Beresford, M.K. Bayliss, DDT 5 (2000) 409–414.
[4] A.P. Li, DDT 6 (2001) 357–366.
[5] P. Stenberg, K. Luthman, K. Artursson, J. Control Release 65 (2000) 231–243.
[6] S. Ekins, C.L. Waller, P.W. Swaan, G. Cruciani, S.A. Wrighton, J.H. Wikel, J. Pharm. Tox. Methods 44 (2000) 251–272.
[7] D.E. Clark, S.D. Pickett, DDT 5 (2000) 49–58.
[8] S. Ekins, J. Rose, J. Mol. Graph. Model. 20 (2002) 305–309.
[9] G. Klopman, L. Stefan, R.D. Saiakhov, Proceedings of the 223rd American Chemical Society National Meeting, ACS, Washington DC, USA, 2000.
[10] C.A. Lipinski, F. Lombardo, B.W. Dominy, P.J. Feeney, Adv. Drug Deliv. Rev. 46 (2001) 3–26.
[11] D.F. Veber, S.R. Johnson, H.-Y. Cheng, B.R. Smith, K.W. Ward, K.D. Kopple, J. Med. Chem. 45 (2002) 2615–2623.
[12] F. Yoshida, J.G. Topliss, J. Med. Chem. 43 (2000) 2575–2585.
[13] C.W. Andrews, L. Bennett, L.X. Yu, Pharm. Res. 17 (2000) 639–644.
[14] W. Bains, R. Gilbert, L. Sviridenko, J.-M. Gascon, R. Scoffin, K. Birchall, I. Harvey, J. Caldwell, Curr. Opin. Drug Discovery Dev. 5 (2002) 44–51.
[15] L.A. Zadeh, in: J. Van Ryzin (Ed.), Classification and Clustering, Academic Press, NY, USA, 1977, pp. 251–299.
[16] M. Pintore, K. Audouze, F. Ros, J.R. Chretien, Data Sci. J. 1 (2002) 99–110.
[17] M. Pintore, N. Piclin, E. Benfenati, G. Gini, J.R. Chrétien, Env. Tox. Chem. 22 (2003), in press.
[18] M. Pintore, O. Taboureau, F. Ros, J.R. Chrétien, Eur. J. Med. Chem. 36 (2001) 349–359.
[19] S. Hirono, I. Nakagome, H. Hirano, Y. Matsushita, F. Yoshi, I. Moriguchi, Biol. Pharm. Bull. 17 (1994) 306–309.
[20] F. Ros, M. Pintore, J.R. Chrétien, Chemometr. Intell. Lab. 63 (2002) 15–26.
[21] R.J. Bertz, G.R. Granneman, Clin. Pharmacokinet. 32 (1997) 210–258.
[22] L.S. Goodman, L.E. Limbird, P.B. Milinoff, A.G. Gilman, J.G. Hardman (Eds.), Goodman and Gilman's: the pharmacological basis of therapeutics, ninth ed., Mc-Graw-Hill, New York, NY, USA.
[23] M.A. Navia, P.R. Chaturvedi, DDT 1 (1996) 179–189.
[24] W.K. Sietsema, Int. J. Clin. Ther. Tox. 27 (1989) 179–211.
[25] A. Sabljic, in: W. Karcher, J. Devillers (Eds.), Practical Applications of Quantitative Structure-Activity Relationships (QSAR) in Environmental Chemistry and Toxicology, Kluwer Academic Publishers, Dordrecht, 1990, pp. 61–82.
[26] J.C. Dearden, in: W. Karcher, J. Devillers (Eds.), Practical Applications of Quantitative Structure-Activity Relationships (QSAR) in Environmental Chemistry and Toxicology, Kluwer Academic Publishers, Dordrecht, 1990, pp. 25–59.
[27] L.B. Kier, L.K. Hall, Molecular Structure Description—The Electrotopological State, Academic Press, San Diego, CA, USA, 1999.
[28] R.L. Haupt, S.E. Haupt, Practical Genetic Algorithms, Wiley, NY, USA, 1998.
[29] R. Leardi, A.L. Gonzales, Chemometr. Intell. Lab. (1998) 195–207.
[30] Y. Lin, G.J. Cunninghan, J. Intell. Fuzzy Syst. 2 (1994) 243–250.
[31] M. Sugeno, T. Yasakawa, IEEET. Fuzzy Syst. 1 (1993) 7–31.
[32] D. Dubois, H. Prade, in: G. Shafer, J. Pearl (Eds.), Readings in Uncertain Reasoning, Morgan Kaufmann, San Francisco, CA, USA, 1990, pp. 742–761.
[33] M.M. Gupta, J. Qi, Fuzzy Set Syst. 40 (1991) 431–450.
[34] T. Kohonen, Self-Organizing Maps, Springer, New York, NY, USA, 2001.
[35] G. Cruciani, P. Crivori, P.-A. Carrupt, B. Testa, J. Mol. Struct. (Theochem.) 503 (2000) 17–30.
[36] T.I. Oprea, I. Zamora, A.L. Ungell, J. Comb. Chem. 4 (2002) 258–266.