Solubility Prediction by Recursive Partitioning

Xiaoyang Xia,[1,3] Edward Maliski,[1] Janet Cheetham,[2] and Leszek Poppe[2]

Received June 13, 2003; accepted June 19, 2003

Purpose. To build and test a computational model for predicting small molecule solubility, to improve the cost-effectiveness of the selection of vendor compounds suitable for nuclear magnetic resonance (NMR) screening.

Methods. A simple recursive partitioning decision tree–based classification model was generated utilizing "off-the-shelf" commercial software from Accelrys Inc., with a training set of 1992 compounds based on a series of calculated topologic and physical properties. The predictive ability of the decision tree was then assessed by employing it to classify a test set of 2851 vendor compounds, and the classification was subsequently used to guide the purchase of 686 compounds for the purpose of NMR screening.

Results. When the decision tree was used to guide purchasing, the percentage of "acceptable" compounds suitable for NMR screening doubled compared with the use of a simple cLogP cutoff, improving the successful selection rate from 25% to 50%.

Conclusions. A simple recursive partitioning decision tree may successfully be used to improve cost-effectiveness by reducing the wastage associated with the unnecessary purchase of vendor compounds unsuitable for NMR screening because of insolubility.

KEY WORDS: solubility prediction; decision tree; recursive partitioning; NMR screening.

INTRODUCTION

In the biopharmaceutic context solubility is a critical factor in the selection of compounds suitable for research.

Solubility is known to be dependent on temperature, polarity, and molecular size. Molecular shape has also been implicated as a factor in determining the readiness of solvation, but the relationship is clearly not straightforward.

Many pharmaceutical tests, such as high-throughput screening of compounds, depend on the solubility of test compounds [in a solvent such as dimethyl sulfoxide (DMSO) or aqueous buffer] to obtain results in an effective and accurate manner. Attempting to perform tests on insoluble compounds is wasteful of the tests and an unnecessary expense in the acquisition process.

To avoid this solubility pitfall (which seems to be an industry-wide problem), an efficient virtual compound-screening method is required that can evaluate the likely solubility of compounds before their purchase. Such a method could streamline the screening process and significantly reduce the monetary outlay required for screening.

For example, Amgen employs nuclear magnetic resonance (NMR) as part of its discovery process. In order for an NMR spectrum to distinguish binding, the compounds used for screening must be soluble and be in monomer form. Although many computational models are available for predicting solubility (1–13), most have the following disadvantages:

1. They contain training datasets with only a small number of drug-like molecules.
2. They cannot predict whether a given compound exists as a monomer or as an aggregate.

The computational model reported in this paper is derived from a recursive partitioning decision tree, available as a commercial software package. The model helps to prioritize compounds before purchase and yields a larger proportion with improved levels of solubility as determined by NMR.

The recursive partitioning decision tree is a useful statistical classification tool, allowing (in this instance) a relatively small group of molecular compounds with well-characterized properties to be used to generate a set of classification rules. Those rules may then be applied to assist in the selection of compounds from a less well-characterized superset.

Fig. 1 illustrates a fragment of an example decision tree derived from a set of data with known parameters $X_1, X_2, X_3 \ldots X_n$ falling into classes $A, B, C, \ldots Z$.

At each decision point in the process, the data are split into two subgroups based on (usually) the value of a particular parameter. Each subgroup is then similarly split until further splitting is not possible or a threshold value for stopping is reached. The decision as to which parameter to use, and the appropriate value to use to determine a split, are obtained by an automated statistical analysis of the entire data set.

Using the recursive partitioning decision tree model, we have found that the increase in the rate of successful selection of suitable vendor compounds is sufficiently large (rising from 25% using a simple cLogP cutoff threshold to 50% using the model) that this approach is likely to have a significant impact on the cost-effectiveness of compound purchasing. The increase is achieved using off-the-shelf software suitable for in-house research, utilizing a single workstation, and without the necessity of specialist tuning of the default parameters.

MATERIALS AND METHODS

Hardware and Software

The hardware comprised a Silicon Graphics® Octane™ workstation with a single R10000™ cpu running at 250 MHz.

The software included the operating system—IRIX64 Release 6.5—and the Accelrys Inc. application Cerius^2 version 4.5.

[1] Research Informatics, Amgen Inc., Thousand Oaks, California 91320.

[2] Molecular Structure, Amgen Inc., Thousand Oaks, California 91320.

[3] To whom correspondence should be addressed. (email: xxia@amgen.com)

ABBREVIATIONS: AlogP98, the log of the partition coefficient (see cLogP), atom-type value; cLogP, the log of the partition coefficient—the equilibrium concentration in a nonpolar solvent (octanol) divided by the equilibrium concentration in a polar solvent (usually water), expressed as a logarithmic value; NMR, nuclear magnetic resonance; QSAR, quantitative structure–activity relationship.

0724-8741/03/1000-1634/0 © 2003 Plenum Publishing Corporation

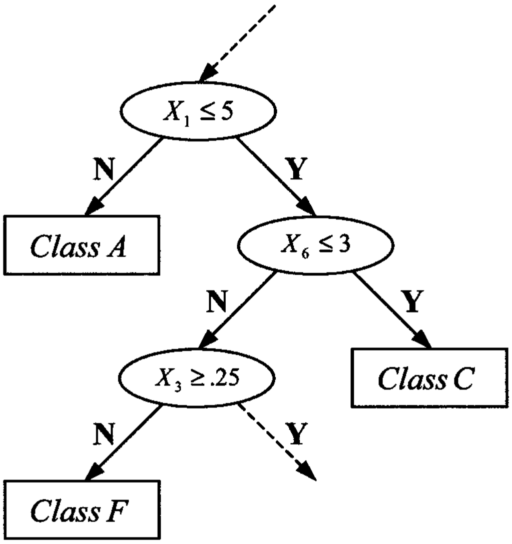


Fig. 1. Fragment of an example recursive partitioning decision tree.

Molecular Descriptors

Molecular descriptors (QSAR and Daylight Fingerprint 1024; see Table I a–f) were provided with Cerius^2, although the use of the Daylight Fingerprint required the purchase of a separate license from Daylight Inc.

The QSAR set includes the commonly used Lipinski descriptors (8) (see Table Ia,b) in addition to 18 spatial indices, 28 topological indices, and nine information content indices.

The spatial indices (Table Ic) are descriptors pertaining to the space occupied by molecules, measured in terms of points, lines, surfaces, volumes, and so on. They include molecular volume and the Stanton and Jurs set (14), which combines shape and electronic information to characterize molecules by calculating the mapping of atomic partial charges on solvent-accessible surface areas of individual atoms. The FPSA descriptors are derived from the division of the PPSA descriptors by the SASA descriptor.

The topological indices (Table Id) are two-dimensional (2D) descriptors based on graph theory concepts (15–17). They help to differentiate molecules according to size, degree of branching, flexibility, and overall shape. They comprise Kier's Shape and Alpha-Modified Shape Indices, the Molecular Flexibility Index, the Kier and Hall Molecular Connectivity Index, the Kier and Hall Valence-Modified Connectivity Index, the Kier and Hall Subgraph Count Indices, the Wiener Index, the Zagreb Index, and one of two Balaban Indices.

The information content indices (Table Ie) are derived from a view of molecules as structures that can be partitioned into subsets of elements that are in some sense equivalent. The definition of "equivalent" depends on the particular descriptor.

They comprise the Multigraph Information Content Indices (consisting of information content, bonding information content, complementary information content, structural information content), the edge adjacency/magnitude, the edge distance/magnitude, the information of atomic composition index, the vertex adjacency/magnitude, and the vertex distance/magnitude.

Finer detail may be obtained from Accelrys' product documentation.

Selection of the Training Set

We tested the solubility of 1992 compounds using NMR-based target-binding assays and then used these compounds as the training set to build a predictive model. Of the 1992 compounds, 981 were found to be soluble, and 1011 were insoluble. Computational experiments were devised to assess the diversity of the compounds in the training set. Compound diversity was defined by considering responses to the following questions.

How Many Chemical Classes Are in the Training Set?

ClassPharmer™ from Bioreason Inc. was used to tally the chemical classes. Homogeneity level and redundancy level were set to medium for generating the classes.

Do Similar Compounds Dominate the Training Set?

The similarity of pairs of compounds was measured by performing a near-neighbor calculation on each molecule in the training set as follows: For a given compound A, the Tanimoto distance between A and each of the other compounds in the data set was calculated. Daylight Fingerprint (18) was used as the structural descriptor. Any compound for which the Tanimoto distance to A was less than 0.25 was defined as A's near neighbor (NN). The count of near neighbors thus defined for the data set is summarized in Table II.

Molecular Descriptors and Model Building

Once diversity was established, for each compound the following descriptors were calculated and used as independent variables within Cerius^2: Daylight Fingerprint (1024 bits) (18) and 60 QSAR default descriptors (19), brief details of which have already been given (Table I). A recursive partitioning decision tree was constructed using Cerius^2 version 4.5 (19) with the default settings: weight classes equally; score splits using Gini impurity; perform moderate pruning; nodes must contain a minimum of 1% of the samples; limit knots per variable to 20; maximum tree depth 10. Fig.2 shows a sample screenshot of the defaults.

The process is quite straightforward and does not require specialized knowledge. The application is started, the file containing the training set data is read in, the QSAR and Daylight Fingerprint descriptors are selected, the relevant properties are calculated, and the statistical method—recursive partitioning decision tree (RP)—is selected. The set of molecules to be used for training is selected (invariably the entire training set), and the model begins training using just the default parameters. The whole process is finished in under 1 min.

Once complete, the decision tree is available for viewing in Cerius^2 and may be applied to a test set of molecules. As each molecule is classified, its probability of belonging to a particular class is displayed along with its position on the tree (the leaf node). A "real world" example of such a tree is given in Fig. 3.

Prediction

The tree was applied to sets of vendor compounds to determine their solubility, and compounds predicted to be soluble were purchased. Near-neighbor calculations, similar

Table I. QSAR Molecular Descriptors Utilized for Training

Descriptor class	Descriptor(s)	Description
(a) Physicochemical	AlogP98	The log of the partition coefficient
(b) Structural	Rotbonds	Number of rotatable bonds
	Hbond acceptors	Number of hydrogen bond acceptors
	Hbond donors	Number of hydrogen bond donors
	MW	Molecular weight
(c) Spatial	V_m	Molecular volume
	Jurs:	
	PPSA-1	Partial positive surface area
	PPSA-2	Total charge-weighted positive surface area
	PPSA-3	Atomic charge-weighted positive surface area
	PNSA-1	Partial negative surface area
	PNSA-2	Total charge-weighted negative surface area
	PNSA-3	Atomic charge-weighted negative surface area
	DPSA-1	Difference in charged partial surface areas
	DPSA-2	Difference in total charge-weighted surface areas
	DPSA-3	Difference in atomic charge-weighted surface areas
	FPSA-1	Partial positive surface area/SASA*
	FPSA-2	Total charge-weighted positive surface area/SASA*
	FPSA-3	Atomic charge-weighted positive surface area/SASA*
	FNSA-1	Partial negative surface area/SASA*
	FNSA-2	Total charge-weighted negative surface area/SASA*
	FNSA-3	Atomic charge-weighted negative surface area/SASA*
	TPSA	Total polar surface area
	SASA*	Total molecular solvent-accessible surface area
(d) Topological:	Kier and Hall	Shape
	Kappa-1	First order
	Kappa-2	Second order
	Kappa-3	Third order
		Alpha-modified shape
	Kappa-1-AM	First order
	Kappa-2-AM	Second order
	Kappa-3-AM	Third order
	PHI	Molecular flexibility index
		Molecular connectivity index
	CHI-0	Zero order
	CHI-1	First order
	CHI-2	Second order
	CHI-3_P	Third order, path
	CHI-3_C	Third order, cluster
	CHI-3_CH	Third order, chain
		Valence-modified connectivity index
	CHI-V-0	Zero order
	CHI-V-1	First order
	CHI-V-2	Second order
	CHI-V-3_P	Third order, path
	CHI-V-3_C	Third order, cluster
	CHI-V-3_CH	Third order, chain
		Subgraph count indices
	SC-0	Zero order
	SC-1	First order
	SC-2	Second order
	SC-3_P	Third order, path
	SC-3_C	Third order, cluster
	SC-3_CH	Third order, chain
	Wiener	Wiener index
	Zagreb	Zagreb index
	JX	Balaban index
(e) Information Content:	Multigraph	
	IC	Information content
	BIC	Bonding information content
	CIC	Complementary information content
	SIC	Structural information content
	E-ADJ-mag	Edge adjacency/magnitude
	E-DIST-mag	Edge distance/magnitude

Table I. Continued

Descriptor class	Descriptor(s)	Description
	IAC-total	Information of atomic composition index
	V-ADJ-mag	Vertex adjacency/magnitude
	V-DIST-mag	Vertex distance/magnitude
(f) Fingerprint	DYFP-1024	Daylight Fingerprint 1024

*SASA, solvent-accessible surface area.

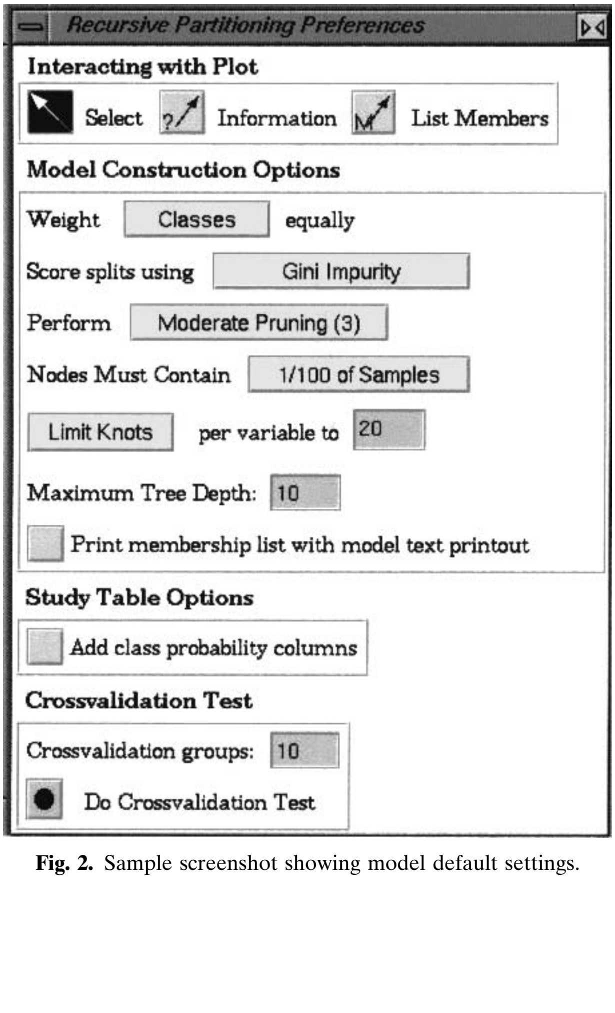
to those described for the training set, were also performed on the purchased compounds to determine their diversity, both in comparison with each other and in comparison with the compounds in the training set. The results are listed in Table II.

The resulting sets of purchased compounds were then assessed using NMR to determine solubility. We define acceptable solubility as (a) being soluble to 1 μM in phosphate-buffered saline (PBS: 0.9% NaCl, 10 mM sodium phosphate/pH 7.2) and 6% dimethyl sulfoxide (DMSO) and (b) providing an acceptable NMR spectrum.

RESULTS AND DISCUSSION

It is essential that a predictive model is built from a heterogeneous training data set, i.e., a data set that is representative, reliable, and informative. Because the data set was generated in our NMR laboratory in a consistent manner, reliability is not an issue.

However, steps need to be taken to ensure that the compounds in the data set represent a variety of chemical classes of interest.



Recursive Partitioning Preferences

Interacting with Plot

Model Construction Options

Weight Classes equally

Score splits using Gini Impurity

Perform Moderate Pruning (3)

Nodes Must Contain 1/100 of Samples

Limit Knots per variable to 20

Maximum Tree Depth: 10

□ Print membership list with model text printout

Study Table Options

□ Add class probability columns

Crossvalidation Test

Crossvalidation groups: 10

● Do Crossvalidation Test

Fig. 2. Sample screenshot showing model default settings.

The first step we took was to confirm diversity, and the results showed the data set contained 299 chemical classes, and more than half of the compounds did not have near neighbors based on the 0.25 Tanimoto distance cutoff (Table II).

The second step was to develop a model using a "machine learning" approach. We elected to use a decision tree as our modeling tool for the following reasons:

1. Rapid training, even with large numbers of variables or records. In this case, we had over 1000 independent variables.
2. Insensitivity to outliers. Unlike some methods, such as regression, that tend to draw the outliers closer to the model, with this approach any outliers usually do not change the split position.
3. Ready comprehension of the model. Splitting rules are displayed, and the most important variables are usually at the top of the tree.

The third step was the deployment of the tree. When the tree was applied to a set of 2851 preselected vendor compounds, all of which possessed a desired scaffold, a total of 686 compounds from two sets (Purchased Set 1 and Purchased Set 2) were identified as suitable for purchase. Table II shows that the purchased compounds are significantly diverse from each other (rows 2 and 3) and from compounds in the training set (rows 4 and 5). We were encouraged by the fact that the tree did not just select compounds similar to the training set.

Previously we had used cLogP < 3.5 as the cutoff when selecting compounds for purchasing and screening. By that approach an average of 23% of purchased compounds was found to be soluble as indicated by NMR spectra. When we tested the two new sets proposed by the recursive partitioning tree, an average twofold enrichment rate (~50%) of soluble compounds was observed in all purchases. The results are summarized in Table III.

Examples

To make the demonstration of examples more manageable, the complex decision tree in Fig. 3 is divided into two separate trees based on the first split (AlogP98): Fig. 4a shows the path followed when, for a given compound, AlogP98 < 3.47, and Fig. 5a shows the path followed when, for a given compound, AlogP98 ≥ 3.47. By convention, a "True" response to any given split follows the branch to the right, and a "False" response to any given split follows the branch to the left. In both figures, irrelevant leaf nodes have been colored gray to more readily contrast the relevant

Table II. Diversity (Near-Neighbor Summary) of the Training and Purchased Sets

	Total number of compounds	NN = 0	NN = 1–5	NN = 5–10	NN > 10	Number of chemical classes
Training set (self)	1992	1159 (58%)	716 (36%)	101 (5%)	16 (1%)	299
Purchased set 1 (self)	468	285 (61%)	170 (36%)	9 (2%)	4 (1%)	68
Purchased set 2 (self)	218	171 (78%)	47 (22%)	0	0	62
Purchased set 1 vs. training set	468	427 (91%)	39 (8%)	2 (1%)	0	
Purchased set 2 vs. training set	218	153 (70%)	54 (25%)	4 (2%)	7 (3%)	

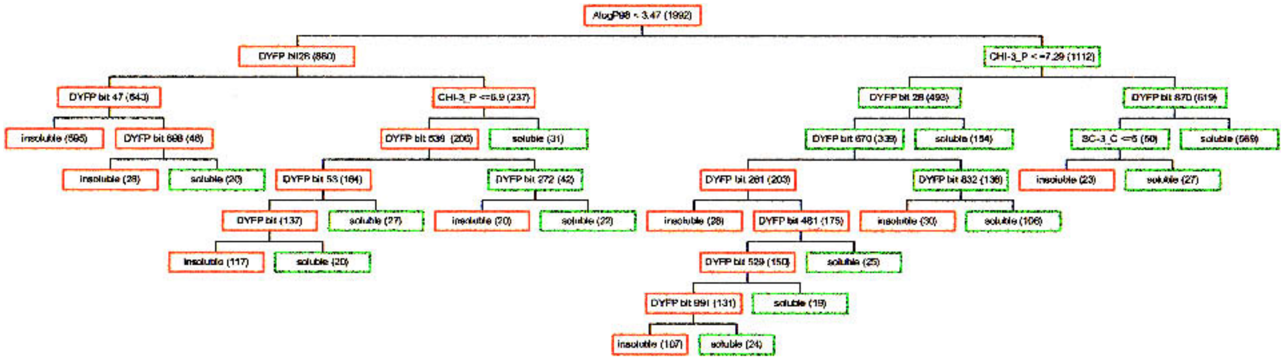


Fig. 3. Real-world example of a decision tree.

Table III. Comparison of Pass Ratios for Recursive Partitioning Decision Tree vs. cLogP

	Number of compounds purchased	Number of compounds passed	Pass ratio
Threshold: cLogP < 3.5			
Prior purchased set 1	210	48	23%
Prior purchased set 2	188	41	22%
Prior purchased set 3	2561	613	24%
Threshold: recursive partitioning decision tree			
Purchased set 1	468	232	50%
Purchased set 2	218	128	59%

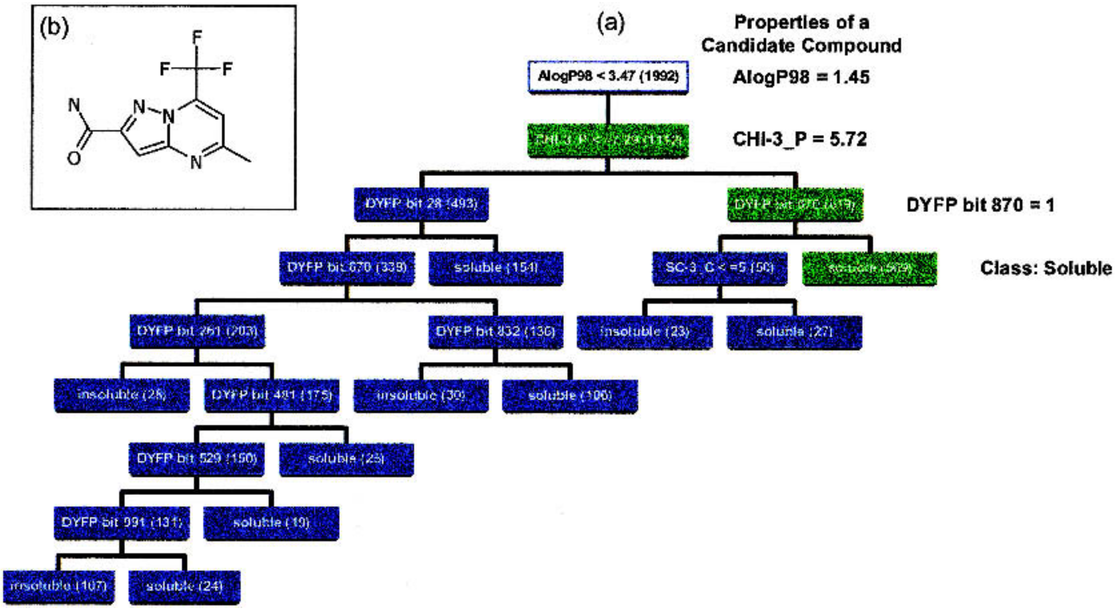


Fig. 4. Decision path (a) and structure (b) for a candidate compound that is classified as soluble.

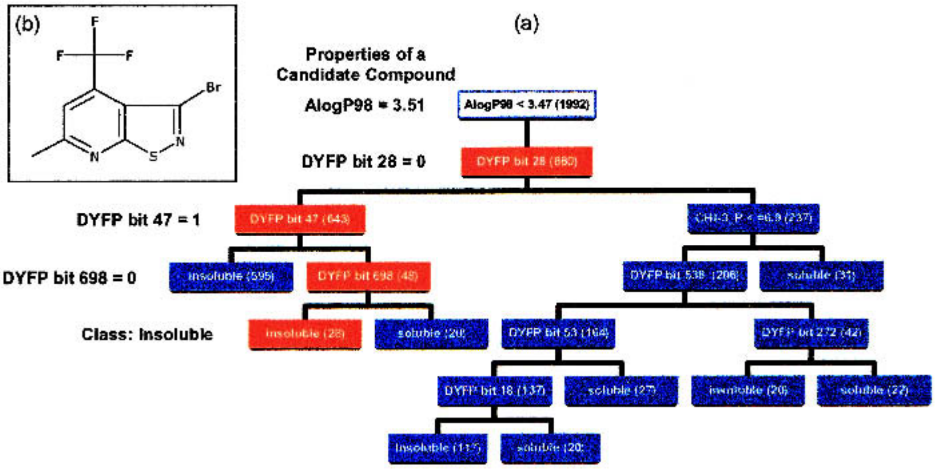


Fig. 5. Decision path (a) and structure (b) for a candidate compound that is classified as insoluble.

path followed in classifying a candidate compound. Figures 4b and 5b show the respective structures of the example classified compounds.

CONCLUSIONS

The recursive partitioning decision tree generated with the aid of the Cerius^2 package has successfully improved the prediction of solubility and helped to avoid the unnecessary purchase of compounds unsuitable for NMR screening. The model was easy to construct, understand, and deploy. The only disadvantages of this method—which are true for all decision trees—are:

1. The tree model can be subject to local minima. Every tree is dependent on the first split and is not necessarily guaranteed to arrive at the best model.
2. The tree is susceptible to instability triggered by small changes because decision boundaries are rectilinear. By adding or deleting a few records, one may obtain a very different tree model. Our decision to apply diversity criteria to our model does, however, reduce the risk of undue sensitivity.

The success rate of the predictions may appear to be low, but it is sufficient for our needs, and it is, to some degree, expected.

The model demonstrates that it may be usefully applied to guide compound purchasing, although the specific decision tree generated by the training set used here is almost certainly applicable only to this particular instance.

Although it doubles the previous successful selection rate for suitably soluble compounds, the model is clearly not capable, in its present form, of predicting solubility class in 100% of cases. This may result from a variety of factors, including—but not limited to—the size of the training set, the range of parameters available for classification, measures of dissolution and other crystal-related phenomena, and the statistical analysis leading to the first split.

The evident complexity of the generated decision tree hints at the complexity of the process of solvation and may indicate that more (or at least different) parameters may need to be included in the model in order to improve still further its predictive capabilities.

Further study is in progress, and results will be reported in due course.

ACKNOWLEDGMENTS

This work was supported by research funding generously provided by Amgen Inc. (Thousand Oaks, CA). We thank Mr. Peter Brooks and Mr. Ken Kashani for preparation of the manuscript.

REFERENCES

1. C. A. S. Bergstrom, U. Norinder, K. Luthman, and P. Artursson. Experimental and computational screening models for prediction of aqueous drug solubility. *Pharm. Res.* **19**:182–188 (2002).
2. W. Jorgensen and E. M. Duffy. Prediction of drug solubility from structure. *Adv. Drug Deliv. Rev.* **54**:355–366 (2002).
3. N. R. McElroy and P. C. Jurs. Prediction of aqueous solubility of heteroatom-containing organic compounds from molecular structure. *J. Chem. Inf. Comput. Sci.* **41**:1237–1247 (2001).
4. Y. Ran and S. H. Yalkowsky. Prediction of drug solubility by the general solubility equation (GSE). *J. Chem. Inf. Comput. Sci.* **41**:354–357 (2001).
5. A. Katritzsky and D. B. Tatham. Correlation of the solubility of gases and vapors in methanol and ethanol with their molecular structures. *J. Chem. Inf. Comput. Sci.* **41**:358–363 (2001).
6. J. Huuskonen. Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. *J. Chem. Inf. Comput. Sci.* **40**:773–777 (2000).
7. J. Huuskonen, M. Salo, and J. Taskinen. Aqueous solubility prediction of drugs based on molecular topology and neural network modeling. *J. Chem. Inf. Comput. Sci.* **38**:450–456 (1998).
8. C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeny. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **23**:3–25 (1997).
9. T. M. Nelson and P. C. Jurs. Prediction of aqueous solubility for a diverse set of heteroatom-containing organic compounds. *J. Chem. Info. Comput. Sci.* **34**:601–609 (1994).
10. M. Kalmet. Linear solvation energy relationships: an improved equation for correlation and prediction of aqueous solubilities of aromatic solutes including polycyclic aromatic hydrocarbons and polychlorinated biphenyls. *Prog. Phys. Org. Chem* **19**:295–317 (1993).
11. G. Klopman, S. Wang, and D. M. Balthasar. Estimation of aqueous solubility of organic molecules by the group contribution approach. Application to the study of biodegradation. *J. Chem. Inf. Comput. Sci.* **32**:439–445 (1992).

12. N. Bodor and N.-J. Huang. A new method for the estimation of the aqueous solubility of organic compounds. *J. Pharm. Sci.* **81**: 954–960 (1992).
13. N. Bodor, A. Harget, and N.-J. Huang. Neural network studies. 1. Estimation of the aqueous solubility of organic compounds. *J. Am. Chem. Soc.* **113**:9480–9483 (1991).
14. D. T. Stanton and P. C. Jurs. Development and use of charged particle surface area structural descriptors in computer assisted quantitative structure–property relationship studies. *Anal. Chem.* **62**:2323–2329 (1990).
15. L. B. Kier and L. H. Hall. *Molecular Connectivity in Chemistry and Drug Research*. Academic Press, New York, 1976.
16. L. B. Kier and L. H. Hall. Molecular connectivity VII: Specific treatment of heteroatoms. *J. Pharm. Sci.* **65**:1806–1809 (1976).
17. L. B. Kier and L. H. Hall. *Molecular Connectivity in Structure-Activity Analysis*. Wiley, New York, 1986.
18. Daylight Chemical Information Systems Inc. Mission Viejo, CA. *www.daylight.com.*
19. Cerius^2 is a product of Accelrys Inc., San Diego, CA. *www.accelrys.com.*