Development of a Method for Evaluating Drug-Likeness and Ease of Synthesis Using a Data Set in Which Compounds Are Assigned Scores Based on Chemists' Intuition

Yuji Takaoka,* Yutaka Endo, Susumu Yamanobe, Hiroyuki Kakinuma, Taketoshi Okubo, Youichi Shimazaki, Tomomi Ota, Shigeyuki Sumiya, and Kensei Yoshikawa

Molecular Simulation Group, Research Center, Taisho Pharmaceutical Co., Ltd., 1-403 Yoshino-cho, Kita-ku, Saitama-shi, 331-9530 Saitama, Japan

Received March 2, 2003

The concept of drug-likeness, an important characteristic for any compound in a screening library, is nevertheless difficult to pin down. Based on our belief that this concept is implicit within the collective experience of working chemists, we devised a data set to capture an intuitive human understanding of both this characteristic and ease of synthesis, a second key characteristic. Five chemists assigned a pair of scores to each of 3980 diverse compounds, with the component scores of each pair corresponding to drug-likeness and ease of synthesis, respectively. Using this data set, we devised binary classifiers with an artificial neural network and a support vector machine. These models were found to efficiently eliminate compounds that are not drug-like and/or hard-to-synthesize derivatives, demonstrating the suitability of these models for use as compound acquisition filters.

INTRODUCTION

In recent years, the development of combinatorial chemistry and high throughput screening techniques has resulted in the ability to obtain hit compounds at ever-increasing numbers. One of the primary areas of current research is the speed with which we can evaluate and optimize the potency and properties of these compounds or derivatives thereof to find good lead compounds. At this point, if a compound we focus on needs to be modified due to undesirable properties or potential toxicity, or if we must find similar compounds with other skeletons due to the difficulty of synthesizing derivatives, the "hit to lead" process requires significantly more time. This requires not only compound diversity in a screening library for increasing the rate of hits in any assay but also good ADME properties (absorption, distribution, metabolism, and excretion) for each compound, freedom from potential toxicity and the ease with which derivatives can be synthesized. Since we cannot measure these properties for each of the enormous numbers of candidate compounds to be incorporated into our screening library, these properties need to be predicted *in silico*.

Many types of computational methods have recently been developed to predict ADME/Tox properties, ranging from filters using simple descriptors such as Lipinski's rule of five,[1] to statistical models such as QSPR (quantitative structure−property relationship) and expert systems based on accumulated experience.[2−8] However, due to insufficient experimental data, both the reliability and applicability of these statistical models for novel compounds remain limited.[9] The concept of "drug-likeness," which should comprise all ADME/Tox properties, has also been discussed and modeled.[10−17] Most reports that attempt to discriminate between drug-like and non drug-like molecules use databases such as the CMC (MDL Comprehensive Medical Chemistry),[18] the MDDR (MDL Drug Data Report),[19] and the WDI (Derwent World Drug Index)[20] to specify molecules considered drug-like, while the ACD (MDL Available Chemicals Directory)[21] is used to specify non drug-like molecules. However, the WDI also clearly contains what we believe to be non drug-like molecules, while the ACD contains some drug-like molecules. Generally, keywords such as dye or radiopaques are used to remove nondrug molecules from drug databases. But such approaches are inadequate for the very same reasons that prevent the accurate exclusion of all non drug-like molecules from our screening library using characteristically unfavorable substructures. Furthermore, since many compounds listed in the ACD are small reactive agents, models developed using these databases may proceed primarily by differentiating reagents, making them less useful for filtering out non drug-like molecules from a screening library.

To resolve these difficulties, we created a data set in which 3980 diverse compounds were assigned scores by five chemists, on the assumption that their collective experience should capture the concept of drug-likeness. The scores were assigned in pairs, assessing not only drug-likeness but also ease of synthesis, another important trait of a good lead compound. Using this data set, we constructed binary classifiers with an artificial neural network (ANN) and a support vector machine (SVM) and evaluated their efficiency in filtering out compounds that are non drug-like and/or hard-to-synthesize derivatives.

METHODS

Compound Selection and Scoring. The compounds to be checked by chemists were selected as follows: First, about 5000 compounds were selected from our corporate database by centroid clustering using a set of numerical descriptors

* Corresponding author phone: +81-48-669-3060; fax: +81-48-663-2145; e-mail: yuji.takaoka@po.rd.taisho.co.jp.

10.1021/ci034043l CCC: $25.00
Published on Web 05/21/2003

Table 1. Molecular Descriptors Calculated by MOE[22]

symbol	description
petitjeanSC	Petitjean graph shape coeffecient
weinerPath	Wiener path number
weinerPol	Wiener polarity number
balabanJ	Balaban's connectivity topological index
a_nN	number of nitrogen atoms
a_nO	number of oxygen atoms
a_aro	number of aromatic atoms.
a_ICM	atom information content (mean)
a_acc	number of hydrogen bond acceptor atoms
a_don	number of hydrogen bond donor atoms
b_rotR	Fraction of rotatable bonds
chi0v	atomic valence connectivity index (order 0)
chi1v_C	carbon valence connectivity index (order 1)
KierA1	first alpha modified shape index
KierA2	second alpha modified shape index
KierA3	third alpha modified shape index
reactive	indicator of the presence of reactive groups
weight	molecular weight (including implicit hydrogens) with atomic weights
vdw_vol	van der Waals volume calculated using a connection table approximation.
vsa_acid	approximation to the sum of VDW surface areas of acidic atoms
vsa_base	approximation to the sum of VDW surface areas of basic atoms
SlogP	log of the octanol/water partition coefficient by Crippen
TPSA	polar surface area
PEOE_RPC+	relative positive partial charge
PEOE_RPC-	relative negative partial charge

(the second set described in the following "descriptors" section). This hierarchical clustering method defines the distance between two clusters as the squared Euclidean distance between their means, so that the selected compounds (cluster centers) tend to be separated by equal distances in chemical space. This nature is suitable for collecting a very diverse set of compounds, including many singular compounds, despite the fact that drug-like compounds in general compound libraries are densely distributed in the same regions in chemical space. Next, compounds clearly having reactive or toxic substructures, such as acyl halides, Michael acceptors, nitroso, and so on were removed to avoid rendering statistical models trivial. The remaining 3980 compounds were assigned scores by five chemists.

Each of the compounds was assigned to one of four categories of drug-likeness and one of three categories of ease of synthesis, with points assigned as follows:

Drug likeness
- A: drug-like (3 points)
- B: drug-like, if forced to categorize (2 points)
- C: non drug-like, if forced to categorize (1 point)
- D: non drug-like (0 points)

Ease of synthesis
- A: easy (2 points)
- B: possible (1 point)
- C: hard (0 points)

Descriptors. We tested 6 types of descriptors/fingerprints for explanatory variables, each of which was used for model building independently. The first was an arbitrarily selected set of 25 descriptors (Table 1) calculated using an MOE software package[22] based on two-dimensional molecular structures. The second was PC (principal component) axes 1 through 9, which cover more than 90% of the information provided by the 25 descriptors' space. The third was a fingerprint that indicates the presence or absence of certain types of atom defined by Wildman and Crippen.[23] We excluded hydrogens and atom types that cannot be defined without hydrogen. Unusual atoms (for example, Si, Cu) were packed into 1 bit. Thus, the length of the fingerprint is 63 bits. We also counted how many atoms of each type were present in a molecule and used this frequency as a fourth explanatory variable. The fifth and the sixth explanatory variables, respectively, were fingerprint and frequency values (number of appearance of each atom/substructure type correspond to each bit) for the MDL MACCS key (166 bits).[24]

Statistical Methods. If the scores are consistent among the chemists, compounds could be sorted in terms of drug-likeness or ease of synthesis according to the assigned scores, and we could apply QSAR methods such as partial least squares to build a model that could judge compounds' drug-likeness and ease of synthesis quantitatively. However, as is shown later, this is not the case. Nevertheless, quite a number of compounds obtained the consent of all five chemist as being non drug-like or hard to synthesize. To create a filter that eliminates unfavorable molecules, we only need to distinguish these compounds from all others including controversial one. For this purpose, we employed binary classification machines, ANN and SVM.

A three-layer network with a back-propagation algorithm was implemented through software developed in-house. The number of neurons in the input layer was set equal to the number of explanatory variables, while that of the output layer was set to one. The number of hidden layer units was varied to find the optimal value. All input values were scaled to values between 0.1 and 0.9, and output values were likewise appropriately rescaled. In the training phase, weight differentials for all links were scaled so that the maximum change did not exceed 0.02. The training step was truncated at certain points to avoid over-learning. Every N-step, weights of which absolute value exceed the threshold (set to 0.01) were reduced, correspond to "forgetting," which has the effect of diminishing useless neurons and links, creating a simpler network.[25,26] The frequency of the forgetting procedure was also optimized.

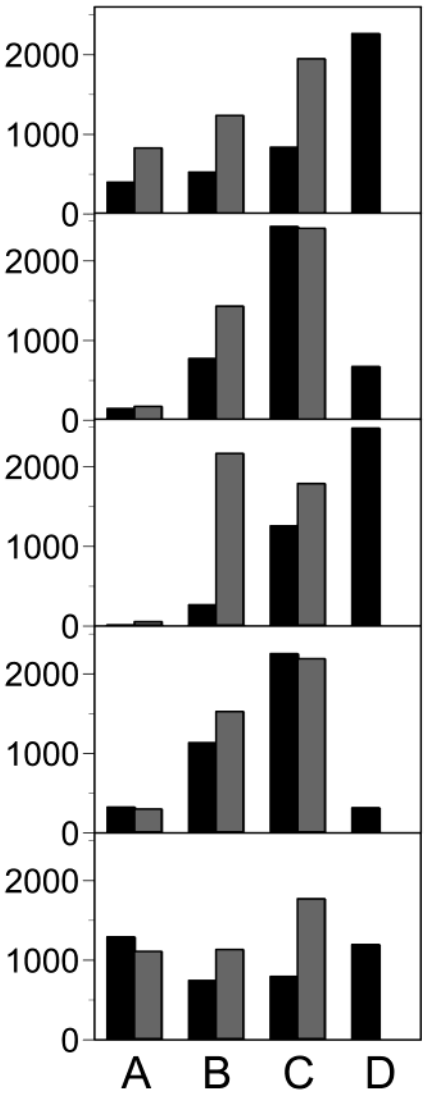

Figure 1. Histograms of number of compounds assigned to each category of drug-likeness (black) and ease of synthesis (gray) by each of five chemists. For drug-likeness: A: drug-like; B: drug-like if forced to categorize; C: non drug-like if forced to categorize; D: non drug-like. For ease of synthesis: A: easy; B: possible; C: hard.

SVM is generally regarded to be one of the best learning machines for pattern recognition,[27] and has been used in many areas, including drug design.[28−30] Details of the specific methods are described elsewhere.[27,31] In brief, they involve the optimization of Lagrangian multipliers α_i with constraints $0 \le \alpha_i \le C$ and $\sum \alpha_i y_i = 0$ to yield a decision function

$$y = \mathrm{sgn}\left(\sum_{SV} \alpha_i y_i K(x_i, x) - h\right) \quad (1)$$

where y_i are input class labels that take -1 or 1, x_i is a set of descriptors, and $K(x_i, x)$ is a kernel function. The sign function sgn(u) returns 1 when $u > 0$, and -1 when $u \le 0$.

We used the LIBSVM implementation developed by Chang and Lin.[32] Four types of kernel functions (linear, polynomial, radial bases function, and sigmoid) were tested. The upper bound parameter C and the inverse of kernel width γ were varied to determine the most accurate model.

RESULTS AND DISCUSSIONS

Score Distribution. Figure 1 shows histograms of the number of compounds assigned to each category by five chemists. The number of compounds assigned scores for the same category for drug-likeness and ease of synthesis reflects good correlation between the paired values for all the chemists except one, who classified many compounds as possible to synthesize but non drug-like and assigned only four compounds to the drug-like category (A). For both properties, the categories to which the molecules were assigned varied considerably among the chemists, which are also shown in Tables 2 and 3.

Table 2. Correlation Coefficient between Scores Assigned by Each Chemist on Drug-Likeness

	chemist A	chemist B	chemist C	chemist D	chemist E
chemist A	1.00	0.55	0.63	0.57	0.58
chemist B		1.00	0.51	0.50	0.50
chemist C			1.00	0.52	0.54
chemist D				1.00	0.54
chemist E					1.00

Table 3. Correlation Coefficient between Scores Assigned by Each Chemist on Ease of Synthesis

	chemist A	chemist B	chemist C	chemist D	chemist E
chemist A	1.00	0.50	0.40	0.40	0.56
chemist B		1.00	0.42	0.47	0.52
chemist C			1.00	0.40	0.48
chemist D				1.00	0.48
chemist E					1.00

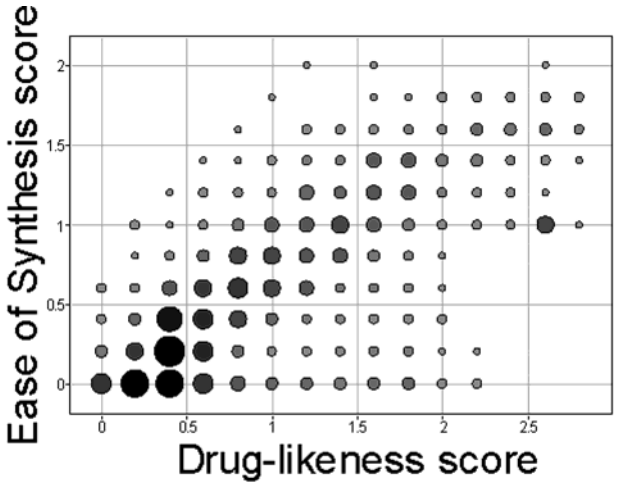

Figure 2. Distribution of average scores assigned by five chemists for drug-likeness and ease of synthesis. The number of compounds having the same combined scores for both properties were counted and expressed as the radii of circles.

Figure 2 shows the distribution of average scores assigned by five chemists for drug-likeness (d) and ease of synthesis (e). The number of compounds having the same combined score for both properties was counted and expressed as the radii of circles. The most common score, (d,e) = (0.4,0.2), was assigned to 192 (4.8%) compounds. Since the centroid clustering method tend to leave outliers as singletons, a lot of "odd" compounds had been selected. Consequently, many compounds were given relatively low scores, fulfilling our purpose that we should collect as many unfavorable compounds as possible. Two properties are correlated, which might be partly because the chemists have scored these two properties at one time. Nevertheless, some compounds were assigned relatively high drug-like scores but low ease of synthesis scores. The characteristics of these compounds will be discussed further below.

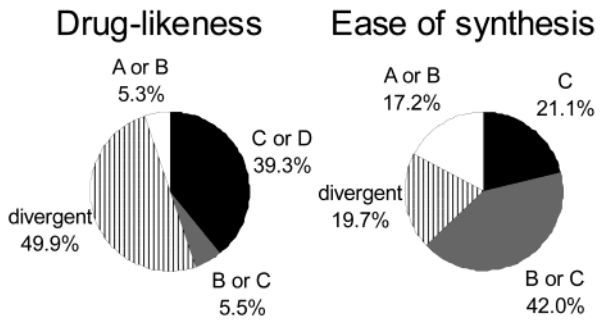

Figure 3. Proportion of compounds classified based on the assigned categories. A or B: placed into categories A or B including all B; B or C: placed into categories B or C except all B and all C; C: placed by all chemists into category C; C or D: placed into categories C or D, including all C; divergent: placed into more than three categories.

Table 4. Percentage of Well-Classified Compounds with ANN Models Averaged for Three Data Sets

descriptor	learning step	# hidden unit	reconstruction step	learning set	test set
Drug Likeness					
MACCS	300	4	80	85.6	76.5
MACCS frequency	5000	4	80	78.0	75.8
Crippen	300	11	80	81.0	73.0
Crippen frequency	3000	12	80	73.4	73.1
MOE descriptors	10000	13	80	76.6	60.8
MOE PC9	10000	10	80	73.0	73.4
Ease of Synthesis					
MACCS	100	3	80	84.5	78.2
MACCS frequency	3000	5	100	90.4	83.3
Crippen	100	12	80	81.2	75.5
Crippen frequency	3000	6	200	84.3	80.3
MOE descriptors	1000	7	80	84.2	82.0
MOE PC9	5000	40	80	81.3	77.6

Since the objective is to create a filter that eliminates only non drug-like molecules, compounds that all chemists consider non drug-like should comprise a set of non drug-like molecules. For this reason, we classified a compound as non drug-like if all five chemists assigned it a C (non drug-like, if forced to categorize) or D (non drug-like). Assessments of ease of synthesis vary widely depending on an individual chemist's experience, but a compound may be a lead even if only one chemist believed it offered the potential for synthesis of derivatives. Once again, we classified a compound as hard to synthesize if all five chemists assigned it a C (hard to synthesize).

Based on these conditions, the number of non drug-like compounds and the number of hard-to-synthesize compounds are 1563 (39.3%) and 840 (21.1%), respectively. Proportion of compounds classified to other categories is shown in Figure 3. We selected 800 compounds at random from among non drug-like molecules and another 800 from the other categories to create a learning set for model building. With respect to ease of synthesis, we selected 400 compounds from among hard-to-synthesize molecules and another 400 from the other categories to create a learning set. For both properties, three different learning sets were selected to create three different data sets. The classification models were evaluated using the remaining compounds (test set).

Table 4 presents a summary of ANN results, showing percentages of well-classified compounds averaged for three data sets in case when the cut-off threshold for the output score was set to 0.5. For each descriptor, the model rendering the highest test set predictions is shown. Discrepancies in predicted values among the three data sets were less than 5% except for four cases, in which we did not pursue the optimal condition due to low accuracy. Changing the frequency of forgetting had relatively little effect on accuracy. Table 5 presents a summary of SVM results. Discrepancies in predicted values in three data sets were less than 3%, except for three cases. The value of C, the upper bound parameter, had relatively little effect on accuracy, provided it remained in the range vicinity of $\sim 10^0$. However, if we reduce it by a factor of 10, accuracy tends to decline; if we increase it by a factor of 10, SVM tends to overlearn the learning data set and renders poor accuracy for the test data set. Hereafter, we analyze and compare each of the best models devised using ANN and SVM.

Table 5. Percentage of Well-classified Compounds with SVM Models Averaged for Three Data Sets

descriptor	C	γ	Kernel function	learning set	test set
Drug Likeness					
MACCS	1	0.04	rbf[a]	90.9	76.5
MACCS frequency	0.1	0.01	polynomial(3,1)[b]	88.3	78.5
Crippen	1	0.1	rbf	84.0	74.6
Crippen frequency	1	0.02	rbf	88.5	74.1
MOE descriptors	1	0.1	rbf	84.9	75.0
MOE PC9	1	0.2	rbf	78.4	73.1
Ease of Synthesis					
MACCS	1	0.02	rbf	88.3	79.6
MACCS frequency	3	0.003	rbf	92.3	81.7
Crippen	1	0.1	rbf	87.9	77.1
Crippen frequency	1	0.05	polynomial (2,10)	96.0	78.2
MOE descriptors	1	0.04	rbf	85.5	79.9
MOE PC9	1	0.1	rbf	92.3	78.9

[a] rbf: radial bases function $K(x_i, x) = \exp[-\gamma|x_i - x|^2]$. [b] $K(x_i, x) = (\gamma x_i x + 1)^3$.

Drug-Likeness. Figure 4a shows the distributions of ANN scores for both non drug-like compounds (1563) and all others (2417) when the MACCS fingerprint was used as explanatory variables. Based on this distribution, the relationship between sensitivity (true positive rate) and specificity (true negative rate), known as the ROC (receiver operating characteristic) curve, is plotted (Figure 4b). Since the purpose of this study is to devise a filter that eliminates only non drug-like molecules, it is preferable to leave a certain number of "bad" compounds in the library than to risk filtering out any "good" compounds. Thus, we placed greater emphasis on specificity than sensitivity, setting the cut-off threshold to 0.4. With this condition, the predictive accuracies of the learning set and test set were 80.1% and 79.1%, respectively.

The overall accuracy of SVM predictions exceeded those based on ANN. Differences between learning set accuracy and test set accuracy were greater than with ANN. This was observed whenever we considered the case of optimal test set prediction. The variances between three data sets are much lower than with ANN, indicating the robustness of the data sets. We took the case where MACCS frequency was used as explanatory variables and evaluated its efficiency as a filter to eliminate non drug-like molecules.

Figure 5a shows the ratios of compounds eliminated by classification machines functioning as filters against the average drug-like scores assigned by five chemists. Since

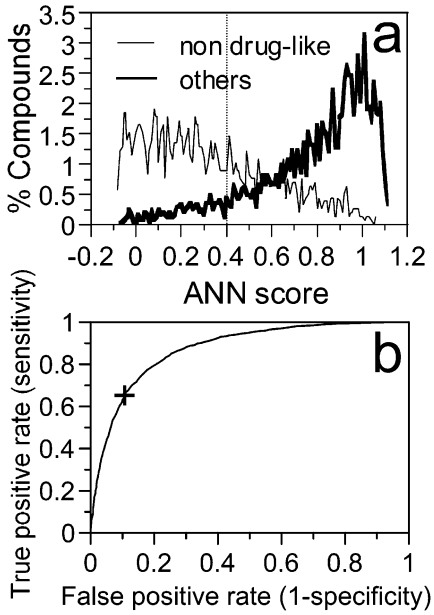

Figure 4. (a) Distributions of ANN scores for both non drug-like compounds and all others normalized by number of compounds in each class, when MACCS fingerprint was used as an explanatory variable. The vertical dotted line at 0.4 indicates the cut-off threshold. (b) ROC curve for the ANN model. The cross mark indicates the point where the cut-off threshold is set to 0.4.

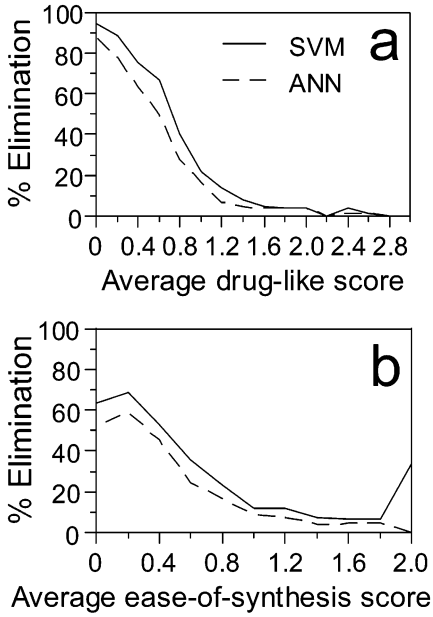

Figure 5. (a) The ratios of compounds eliminated by filters developed by ANN (dashed line) and SVM (solid line) using drug-like scores against the average drug-like score assigned by five chemists. (b) The ratios of compounds eliminated by filters developed by ANN (dashed line) and SVM (solid line) using drug-like scores against the average ease of synthesis score assigned by five chemists.

we shifted the cut-off threshold for ANN to eliminate fewer compounds, the ratio of filtered compounds with the ANN model is slightly less than the SVM model over all score ranges, but only low score compounds are effectively eliminated by both models. Figure 5b also gives the ratios of compounds eliminated by the same drug-like filters against the average ease-of-synthesis scores. Only three compounds were assigned A (2 points) by all five chemists for ease of synthesis, one of which was eliminated by the drug-like filter

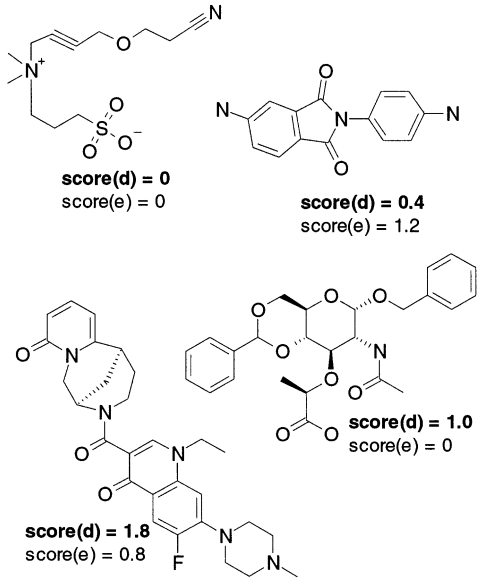

Figure 6. Examples of filtered (eliminated) compounds by the ANN model, with average drug-like score (d) and average ease of synthesis score (e) assigned by five chemists.

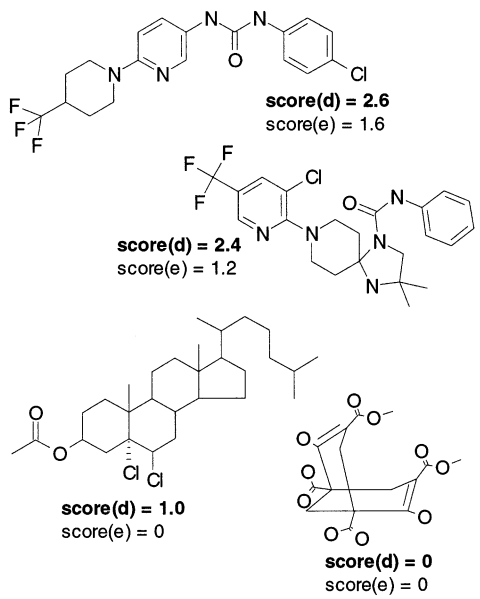

Figure 7. Examples of passed compounds (compounds retained in the database) by the ANN model, with the average drug-like scores (d) and average ease of synthesis scores (e) assigned by five chemists.

developed by SVM, resulting in a high ratio of elimination at a total score of 10. With this exception, the graph shows that the drug-like filter also effectively filters out hard-to-synthesize compounds, demonstrating that most non drug-like compounds are hard-to-synthesize.

Figure 6 presents examples of filtered compounds with the ANN model, while Figure 7 presents compounds passed with the same filter. One of the disadvantages of these statistical methods is our inability to isolate the rationales for such determinations. However, examining structures of well- and misclassified compounds revealed certain common structural features. For example, chemists tend to regard compounds that contain atoms with three ring bonds as non drug-like. Of 348 compounds having such atoms among the original 3980 compounds, 195 compounds (56%) were deemed non drug-like by all five chemists, while only one

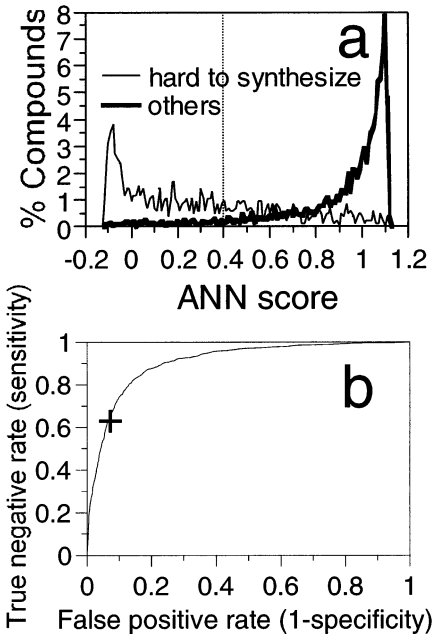

Figure 8. (a) Distributions of ANN scores for both hard-to-synthesize compounds and all normalized by number of compounds in each class, when MACCS frequency was used as an explanatory variable. The vertical dotted line at 0.4 indicates the cut-off threshold. (b) ROC curve for the ANN model. The cross mark indicates the point where the cut-off threshold is set to 0.4.

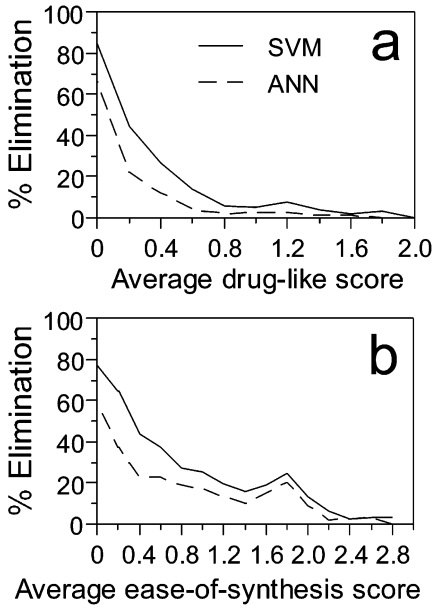

Figure 9. (a) The ratios of compounds eliminated by filters developed by ANN (dashed line) and SVM (solid line) using ease of synthesis scores against average ease of synthesis scores assigned by five chemists. (b) The ratios of compounds eliminated by filters developed by ANN (dashed line) and SVM (solid line) using ease of synthesis scores against average drug-like scores assigned by five chemists.

compound was deemed drug-like. This may be why this classification machine filtered out compounds with relatively good scores, as shown in Figure 6.

Ease of Synthesis. Figure 8a shows distributions of ANN scores for both hard-to-synthesize compounds (840) and all others (3140). Figure 8b shows the ROC curve plotted from this distribution. Again, since the purpose of this work is to devise a filter that eliminates only compounds that are clearly

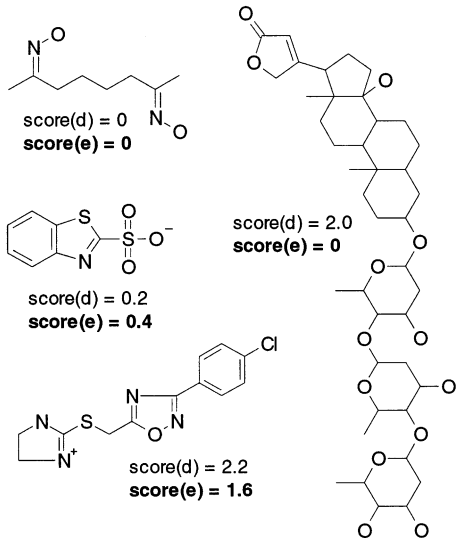

Figure 10. Examples of filtered compounds (compounds eliminated from the database) by the ANN model, with average drug-like scores (d) and average ease of synthesis scores (e) assigned by five chemists.

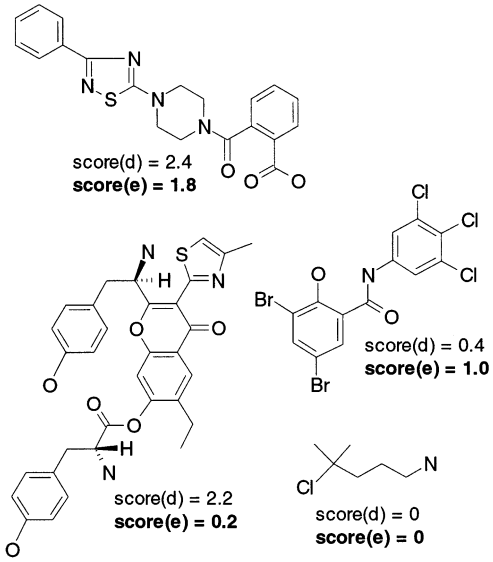

Figure 11. Examples of passed compounds (compounds retained in the database) by the ANN model, with average drug-like scores (d) and average ease of synthesis scores (e) assigned by five chemists.

hard to synthesize, we set the cut-off threshold to 0.4. Given this condition, the overall predictive accuracy of the learning set and test set were 84.9% and 87.6%, respectively. The SVM-based model accounting for MACCS frequency also yielded the most accurate model.

Figure 9a shows the ratios of compounds eliminated by classification machines functioning as filters against the average ease-of-synthesis scores assigned by five chemists. The graph shows higher efficiency than with the drug-like filter, partly because this model distinguishes between compounds of score 0 and all others, while the drug-like filter tries to discriminate between non drug-like compounds assigned the maximum score of 5 and other compounds assigned the minimum score of 2.

Figure 9b presents the ratios of compounds eliminated by ease-of-synthesis filters against the average drug-likeness scores, showing that about 20% of the compounds whose

total drug-like scores fell in the middle were removed. This offer a sharp contrast to the case in which drug-like filters quite effectively eliminate hard-to-synthesize molecules. One of the reasons for this is that there are some compounds that were given relatively high drug-like scores but low ease of synthesis scores (Figure 2). Representative examples of such compounds are sugars, nucleic acids, and steroids, which are regarded to be hard-to-synthesize by most of the chemists, but whose drug-likeness is controversial. Examples of filtered compounds with the ANN model are presented in Figure 10, while passed compounds with the same filter are presented in Figure 11.

CONCLUSION

Five chemists assigned a pair of scores to each of 3980 diverse compounds, with the component scores of each pair corresponding to drug-likeness and ease of synthesis, respectively. Using these data, we developed classification models that selectively filter out non drug-like compounds and hard-to-synthesize compounds. We believe these models can serve as suitable compound acquisition filters.

REFERENCES AND NOTES

(1) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug. Deliv. Rev.* **1997**, *23*, 3−25.

(2) Butina, D.; Segall, M. D.; Frankcombe, K. Predicting ADME properties *in silico*: methods and models. *Drug Discovery Today* **2002**, *7*, S83−S88.

(3) Basak, S. C.; Grunwald, G. D.; Gute, B. D.; Balasubramanian, K.; Opitz, D. Use of statistical and neural net approaches in predicting toxicity of chemicals. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 885−890.

(4) Wang, J.; Lai, L.; Tang, Y. Structural features of toxic chemicals for specific toxicity. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1173−1189.

(5) Dearden, J. C.; Barratt, M. D.; Benigni, R.; Bristol, D. W.; Combes, R. D.; Cronin, M. T. D.; Judson, P. N.; Payne, M. P.; Richard, A. M.; Tichy, M.; Worthl, A. P.; Yourick, J. J. The development and validation of expert systems for predicting toxicity; ECVAM − The European Centre for the Validation of Alternative Methods. This workshop was organized jointly with the European Chemicals Bureau (ECB), Joint Research Centre, Ispra (VA), Italy, 1999.

(6) Pearl, G. M.; Livingston-Carr, S.; Durham, S. K. Integration of computational analysis as a sentinel tool in toxicological assessments. *Curr. Top. Med. Chem.* **2001**, *1*, 247−255.

(7) Darvas, F.; Dormán, G.; Papp, A. Diversity measures for enhancing ADME admissibility of combinatorial libraries. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 314−322.

(8) Egan, W. J.; Merz, K. M.; Baldwin, J. J. Prediction of drug absorption using multivariate statistics. *J. Med. Chem.* **2000**, *43*, 3867−3877.

(9) Beresford, A. P.; Selick, H. E.; Tarbit, M. H. The emerging importance of predictive ADME simulation in drug discovery. *Drug Discovery Today* **2002**, *7*, 109−116.

(10) Ajay; Walters, W. P.; Murcko, M. A. Can we learn to distinguish between "drug-like" and "non drug-like" molecules? *J. Med. Chem.* **1998**, *41*, 3314−3324.

(11) Clark, D. E.; Pickett, S. D. Computational methods for the prediction of "drug-likeness". *Drug Discovery Today* **2000**, *5*, 49−58.

(12) Frimurer, M.; Bywater, R.; Nærum, L.; Lauritsen, L. N.; Brunak, S. Improving the odds in discriminating "drug-like" from "non drug-like" compounds. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1315−1324.

(13) Muegge, I.; Heald, S. L.; Brittelli, D. Simple selection criteria for drug-like chemical matter. *J. Med. Chem.* **2001**, *44*, 1841−1846.

(14) Walters, W. P.; Murcko, M. A. Prediction of "drug-likeness". *Adv. Drug Deliv. Rev.* **2002**, *54*, 255−271.

(15) Xu, J.; Stevenson, J. Drug-like index: A new approach to measure drug-like compounds and their diversity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1177−1187.

(16) Sadowski, J.; Kubinyi, H. A scoring scheme for discriminating between drugs and nondrugs. *J. Med. Chem.* **1998**, *41*, 3325−3329.

(17) Wagener, M.; van Geerestein, V. J. Potential drugs and nondrugs: prediction and identification of important structural features. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 280−292.

(18) CMC; Comprehensive Medicinal Chemistry is available from MDL Information Systems Inc., San Leandro, CA, 94577, and contains drugs already on the market.

(19) MDDR; MACCS-II Drug Data Report is available from MDL Information Systems Inc., San Leandro, CA, 94577, and contains biologically active compounds in the early stages of drug development.

(20) WDI; World Drug Index is available from Derwent Information, London, U.K. Website: http://www.derwent.com.

(21) ACD; Available Chemicals Directory is available from MDL Information Systems Inc., San Leandro, CA, 94577, and contains specialty bulk chemicals from commercial sources. Website: http//:www.mdli.com.

(22) MOE 2002.03. Molecular Operating Environment; 2002.03. ed.; Chemical Computing Group Inc. Copyright 1997−2002: 1010 Sherbrooke Street West, Suite 910 Montreal, Quebec, Canada, H3A 2R7.

(23) Wildman, S. A.; Crippen, G. M. Prediction of physicochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 868−873.

(24) MACCS key; MDL Information Systems Inc.: 14600 Catalina Street, San Leandro, CA 94577.

(25) Aoyama, T.; Ichikawa, H. Reconstruction of weight matrices in neural networks − A method of correlating outputs with inputs. *Chem. Pharm. Bull. (Tokyo)* **1991**, *39*, 1222−1228.

(26) Isu, Y.; Nagashima, U.; Aoyama, T.; Hosoya, H. Development of neural network simulator for structure−activity correlation of molecules (NECO). Prediction of endo/exo substitution of norbornane derivatives and of carcinogenic activity of PAHs from ^{13}C-NMR shifts. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 286−293.

(27) Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*; Cambridge University Press: 2000.

(28) Czermiński, R.; Yasri, A.; Hartsough, D. Use of support vector machine in pattern classification: application to QSAR studies. *Quant. Struct.-Act. Relat*. **2001**, *20*, 227−240.

(29) Bao, L.; Sun, Z. R. Identifying genes related to drug anticancer mechanisms using support vector machine. *FEBS Lett.* **2002**, *521*, 109−114.

(30) Yuan, Z.; Burrage, K.; Mattick, J. S. Prediction of protein solvent accessibility using support vector machines. *Proteins* **2002**, *48*, 566−570.

(31) Chang, C.-C.; Lin, C.-J. Training ν-support vector classifiers: Theory and algorithm. *Neural Comput*. **2001**, *13*, 2119−2147.

(32) Chang, C.-C.; Lin, C.-J. LIBSVM 2.0: Solving different support vector formulations **2000**, http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

CI034043L