Development of a Virtual Screening Method for Identification of "Frequent Hitters" in Compound Libraries

Olivier Roche,[†] Petra Schneider,[†] Jochen Zuegge,[†] Wolfgang Guba,[†] Manfred Kansy,[†] Alexander Alanine,[†] Konrad Bleicher,[†] Franck Danel,[†] Eva-Maria Gutknecht,[†] Mark Rogers-Evans,[†] Werner Neidhart,[†] Henri Stalder,[†] Michael Dillon,[‡] Eric Sjögren,[‡] Nader Fotouhi,[§] Paul Gillespie,[§] Robert Goodnow,[§] William Harris,[∥] Phil Jones,[∥] Mikio Taniguchi,[⊥] Shinji Tsujii,[⊥] Wolfgang von der Saal,[#] Gerd Zimmermann,[#] and Gisbert Schneider*[,†]

F. Hoffmann-La Roche AG, CH-4070 Basel, Switzerland, Roche Bioscience, 3401 Hillview Avenue, Palo Alto, California 94304, Hoffmann-La Roche, Inc., Nutley, New Jersey 07110-1199, Roche Products, Ltd., Welwyn Garden City, Hertfordshire AL7 3AY, U.K., Nippon Roche K.K., 200 Kajiwara, Kamakura City, Kanagawa Prefecture 247-8530, Japan, and Roche Diagnostics GmbH, Sandhoferstrasse 116, DE-68305 Mannheim, Germany

Received May 18, 2001

A computer-based method was developed for rapid and automatic identification of potential "frequent hitters". These compounds show up as hits in many different biological assays covering a wide range of targets. A scoring scheme was elaborated from substructure analysis, multivariate linear and nonlinear statistical methods applied to several sets of one and two-dimensional molecular descriptors. The final model is based on a three-layered neural network, yielding a predictive Matthews correlation coefficient of 0.81. This system was able to correctly classify 90% of the test set molecules in a 10-times cross-validation study. The method was applied to database filtering, yielding between 8% (compilation of trade drugs) and 35% (Available Chemicals Directory) potential frequent hitters. This filter will be a valuable tool for the prioritization of compounds from large databases, for compound purchase and biological testing, and for building new virtual libraries.

Introduction

In conjunction with high-throughput screening technology, the virtual screening concept provides an entry point for accelerating hit finding and lead structure generation.[1] Virtual screening and filtering methods are now routinely utilized during the early phases of the drug design and discovery process. Nevertheless, the selection of any promising molecules critically depends on the quality of the available screening libraries. On the basis of expert knowledge, several filters have already been developed to improve the quality of these compound libraries.[2−4] The aim of such filtering systems is to remove undesired compounds as early as possible during the drug discovery process in order to maximize the likelihood for a lead reaching clinical trials. Crude filters to remove compounds with chemically reactive[5] or toxic chemical groups[6] have been used for a while. Over the past few years, additional filters focusing on more subtle properties have been developed, where the classification of "drugs" and "nondrugs" is a recurring theme.[7−10] Oral bioavailability,[11] aqueous solubility,[12] and metabolic clearance[13] have also been the subject of intense research for implementation of new virtual screening routines. In this work, we have tried to rationalize another piece of medicinal chemists' knowledge: the recognition of "frequent hitters". These compounds show up as hits in many different biological assays covering a wide range of targets, which can happen for two main reasons: (i) the activity of the compound is not specific for the target ("promiscuous compounds"); (ii) the compound perturbs the assay or detection method, e.g., colored or fluorescent molecules. In both cases, such molecules are usually poor starting points for lead optimization programs and can cause an expenditure of money and loss of time without any benefits. Sometimes medicinal chemists are able to identify frequent hitters by obvious undesired structural features or properties. Rationalizing these characteristics and automating the frequent hitter identification process can increase efficiency and thus assist in the selection of promising hit or lead candidates.

The first fundamental step of our analysis was to compile a reliable data set. One possible approach is to follow the "likeness concept" requiring sets of both frequent hitters and "nonfrequent hitters".[2−4] In the second step, specific molecular features had to be determined, which are suited to discriminate between the two classes of compounds. Substructure analysis in combination with multivariate statistical analysis of molecular descriptors represents a practical approach to this task. In the third step, a predictive scheme was elaborated. For this purpose we explored a wide range of molecular descriptors and various prediction techniques. Finally, we employed our new virtual screening tool to identify potential frequent hitters in several compound databases: the Available Chemicals Directory (ACD),[14] the World Drug Index (WDI),[15] and the MedChem Database.[16]

* To whom correspondence should be addressed. Phone: +41 61 68 79696. Fax: +41 48 68 87408. E-mail: gisbert.schneider@roche.com.
[†] F. Hoffmann-La Roche AG.
[‡] Roche Bioscience.
[§] Hoffmann-La Roche, Inc.
[∥] Roche Products, Ltd.
[⊥] Nippon Roche K.K.
[#] Roche Diagnostics GmbH.

10.1021/jm010934d CCC: $22.00
Published on Web 12/04/2001

Materials and Methods

Data Compilation. A diverse collection of data is needed to establish a scoring scheme for the classification of frequent hitters and nonfrequent hitters. In the ideal case, these data accurately represent the distributions of the two classes of molecules in a chemical space. The compilation of such a database is a time-consuming process because all the assay results from various drug discovery projects must be reviewed and the consistency of the data must be evaluated. The origin of our frequent hitter data is diverse. First, a raw collection of 2408 nonredundant molecules was compiled from different sources. A major subset (930 compounds) was compiled by retrieving the best 1000 hits for 161 different high-throughput screening assays performed at Roche, where 930 of them showed up in at least eight different assays. The cutoff of eight represents a relatively conservative threshold. The second major subset (1389 compounds) stemmed from our in-house depository, containing compounds that had been requested by at least six different drug discovery projects. Additional compounds were contributed by all Roche research sites, stemming from various medicinal chemistry projects. To refine this raw data collection, a panel of 11 independent teams of medicinal chemists from all major Roche research centers was asked to identify structures in the list that may be "frequent hitters" based on their intuition and expert knowledge. This was also done to exclude molecules that may be false-annotated in the database or show up in various assays because of degradation or impurity. The votes are thus based on these chemists' expertise. A compound was defined as a frequent hitter if it obtained at least 9 out of 11 possible votes (>80% agreement). This procedure resulted in a set of 479 structures defined as frequent hitter compounds. A diverse set of 423 drugs extracted from the Roche human drug database was used as a reference for nonfrequent hitter molecules. This set has been chosen assuming that it is unlikely for a commercial drug to be a frequent hitter. The number of 423 nonfrequent hitters is given by the fact that the set of >700 drugs in the Roche human drug database contains 423 drugs that represent all therapeutic areas without over-representing individual targets or disease areas. The final data set contained 902 structures (479 frequent hitters and 423 nonfrequent hitters).

Substructure Analysis. Unique substructures describing frequent hitters and nonfrequent hitters were identified by using the commercially available software tool LeadScope (release 2RC1).[17] LeadScope describes compounds in terms of approximately 27 000 predefined structural features and displays their distribution by separate histograms for the frequent hitters and the nonfrequent hitters. Structural elements that seem to be unique for the data sets were selected manually from the program output.

Descriptor Generation. In total 345 molecular one- and two-dimensional descriptors were calculated for each molecule. A total of 74 general properties, molecular indices, and attributes were generated by the program TSAR 3.21.[18] For prediction of the octanol/water partition coefficient (log P_c), the routine of Meylan and Howard was applied.[19] In addition, 120 atom types were calculated following the definition of Ghose and Crippen (GC descriptors).[20] And finally, 150 topological atom-pair descriptors were generated with the program CATS.[21] Different combinations of these descriptors were analyzed to determine their relevance and usefulness in the context of this work.

Statistical Methods. Linear and nonlinear multivariate analysis was applied to find the most predictive molecular descriptors and to form a frequent hitter prediction model. In the first step projection techniques were used to extract relevant molecular features; in the second step classification models were developed on the basis of these features.

1. Linear Methods. Principal component analysis (PCA) was performed to extract a small set of orthogonal factors describing the data distribution. It helps to understand possible relationships between distributions of compounds, which facilitates the identification of outliers and data clusters.[22,23] After this step, the partial least squares (or "projection to latent structures") (PLS) method (a multivariate linear regression technique) was used to elaborate a prediction scheme.[23,24] Several different sets of descriptors were tested. Both PCA and PLS studies were performed with the default options available in the SIMCA-P 8.0 software.[25]

2. Nonlinear Methods. To see whether a nonlinear projection reveals separate clusters of frequent hitters and nonfrequent hitter molecules, a self-organizing map (SOM) analysis of the high-dimensional descriptor space was performed. The SOM technique can be used to generate a topology-preserving nonlinear mapping of a high-dimensional space to a low-dimensional space.[26] The SOM served the purpose of visualizing data distributions in descriptor space. It complements the linear PCA projections.

Three-layered supervised neural network systems were used to find a classifier separating frequent hitters from nonfrequent hitters. Their architecture contained an input layer (fan-out units), one hidden layer (sigmoidal units), and a single sigmoidal output unit. The neural networks were trained using an evolutionary algorithm implementing adaptive step-size control, as detailed elsewhere.[27] The mean square error (mse) was the objective function that had to be minimized during network training. The number of optimization cycles (generations) was 100 per network, and the population size per generation was 500. In different training runs, the number of hidden layer units was systematically varied to find an appropriate setting. The desired output value (target value) of the neural network was 1 for frequent hitters and 0 for nonfrequent hitters. A 10-times cross validation was performed with random 80% (training) + 20% (test) splits of the data. For further validation, the dataset was randomly divided into three parts: a training set (60%), a testing set (20%), and a validation set (20%). We used the two first sets as in the 80% + 20% split for cross validation, and the last set served as an additional independent test set. Both the SOM and the neural networks were implemented by us using the C programming language.[28] The standard Matthews correlation coefficient for binary data, cc, was used to estimate reclassification (training data) and classification (test data) ability of the neural networks.[29] It is defined as

$$\mathrm{cc} = \frac{NP - OU}{\sqrt{(N+O)(N+U)(P+O)(P+U)}}$$

where P, N, O, and U are the number of true positive, true negative, false positive, and false negative predictions, respectively. A perfect prediction gives a correlation coefficient of 1. Neural network output values were converted to binary numbers using a threshold of 0.5. To reduce the "overlearning" (or "overtraining") effect during neural network training, the process was terminated when cc reached an optimum for the test set (forced stop). In addition, the least complex network was selected for the final prediction model, i.e., the one containing the smallest number of hidden neurons, leading to high-prediction accuracy. Further details about neural network theory and development can be found elsewhere.[30]

Preparation of Databases. The prediction model was applied to five databases: ACD,[14] WDI,[15] MedChem,[16] Trade Drugs, and a set of therapeutic drugs from the Roche human drug database. The Trade Drugs data set is composed of molecules from the WDI database that have an available trade name. Many of the therapeutic drugs from the Roche collection are also listed in a compilation by Dollery.[31] The following three-step procedure was applied to harmonize the data sets:

1. Remove all redundancies within each database.
2. Remove all counterions.
3. Remove the compounds for which the GC descriptor calculations fail.

We obtained, respectively, 183 221, 55 750, 36 418, 3344, and 703 compounds from ACD, WDI, MedChem, Trade Drugs, and the Roche human drug database ("therapeutic drugs").

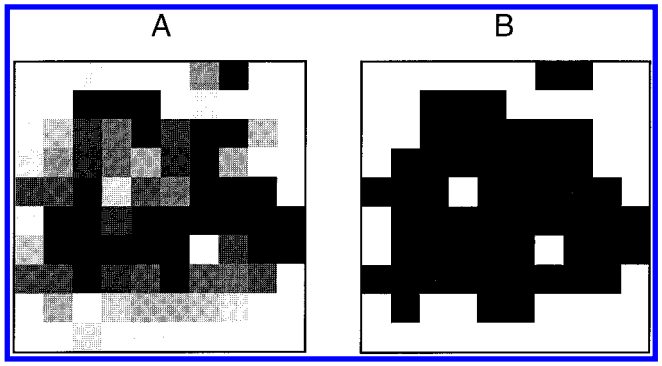

Figure 1. Self-organizing map (SOM) projection of the compound distribution in a high-dimensional space spanned by 120 GC descriptors, where 10 × 10 clusters were formed: (A) density of frequent hitters (white, none; black, many); (B) binary classification of chemical space; frequent hitter area in black, nonfrequent hitter area in white. Note that the map forms a torus.

Results and Discussion

Several classification, feature extraction, and modeling methods were used to establish a computer-based prediction tool detecting frequent hitter molecules. These investigations were based on a set of 479 frequent hitters and 423 nonfrequent hitters.

Substructure Analysis. The software tool LeadScope was used to identify substructure elements that allow for a straightforward discrimination between frequent hitters and nonfrequent hitters. Several substructures were found to be more abundant in the frequent hitter set than among the nonfrequent hitters: benzene, 1-alkylamino, 4-heteroiminomethyl (0/30); imine, *N*-heteroaryl- (0/35); hydrazone, phenyl (2/45); amine, dimethyl, aryl (2/32); benzene, 1,2-dihydroxy (10/67). Numbers in parentheses give their numbers of occurrence in the nonfrequent hitter set and the frequent hitter set, respectively. We have not found a single meaningful substructure that is common to all frequent hitters. None of these structures alone or in combination are sufficient to unambiguously define a frequent hitter compound. This observation suggests that there seems to be no obvious relationship between the presence of a single particular substructure in a compound and its potential to be a frequent hitter. The underlying structure−activity relationship seems to be more complicated, such as a combination of different features of the molecule. A more extensive analysis based on additional molecular descriptors was therefore required.

Self-Organizing Map (SOM) Assessment of Molecular Descriptors. We followed the SOM approach to evaluate the usefulness of different sets of molecular descriptors. A two-dimensional SOM projection of our high-dimensional data will display clusters of data if the molecular descriptors are appropriate for the classification task given. Several SOMs were developed using different descriptor sets: (i) all 345 descriptors, (ii) CATS topological atom pairs, (iii) GC descriptors, and (iv) CATS and GC descriptors together. Clear cluster formation was observed using the GC descriptors alone (Figure 1A). From the SOM projection we concluded that the discrimination between frequent hitters and nonfrequent hitters should be feasible using GC descriptors. This conclusion is supported by the fact that the binary classification ability of the SOM shown in Figure 1B yields a Matthews correlation of cc = 0.8 (32). Only 11% of the frequent hitters fall into the "nonfrequent hitter area" (white area in Figure 1B), and 9% of the nonfrequent hitters are found in the "frequent hitter area" (black area in Figure 1B) of the map. This raw classification accuracy was obtained with equal weights on the individual descriptor contributions.

PCA and PLS Analyses. PCA was performed to complement the SOM analysis and to identify both strong and weak outliers among the data. The apparent cluster formation of frequent hitters and nonfrequent hitters revealed by SOM projection (Figure 1) is less striking in the plane spanned by the two dominant principal components (Figure 2). Several outliers were found, as denoted in Figure 2. Subsequent PLS analysis was performed to find a linear predictive model. Two models were built from the raw data, including the outliers found by PCA, and from the cleaned data set. The resulting models did not show relevant differences in the assessment of the relevant variables with or without the outliers, suggesting that the solution found by PLS is reliable and not critically influenced by extreme raw data values. PLS confirms the strong importance of the GC descriptors for classification, which was already found by SOM projection. The most important descriptors extracted with the variable influence on projection parameter (VIP) included in SIMCA-P are listed in Table 1.[23] The best linear prediction tool derived from the PLS (which is based on three relevant latent variables) reaches a Matthews correlation of cc = 0.8 with the complete data set, using a threshold of 0.5 for the conversion of the model output to binary prediction classes. This value is identical to the binary SOM classification accuracy (Figure 1B). A total of 92% of the frequent hitters and 88% of the nonfrequent hitter examples were correctly classified. In the next step, artificial neural networks were developed to see whether a simple nonlinear model would produce similar or even better results and to compare these models to the PLS system.

Neural Network Analysis. We have trained three-layered supervised neural networks with a single hidden layer containing two, three, or four neurons. A total of 10 independent runs were performed to cross-validate the predictions using test sets composed of 20% randomly chosen compounds of the complete data. All neural network models were based on the GC descriptors only. The simplest network (two hidden neurons) seems to be the best suited among the networks tested. Figure 3 displays the evolution of the averaged mse, cc, and learning step size during the 10 training runs performed with this network architecture. This network reaches an average test data Matthews correlation of $cc_{test} = 0.81$, where 90% of frequent hitters and 91% of the nonfrequent hitters were correctly predicted. In Figure 4 the distribution of raw prediction scores for the test set produced by this network is shown. Compared to the linear PLS model ($cc_{training} = 0.8$), the nonlinear neural network system led to improved prediction accuracy ($cc_{training} = 0.89$, $cc_{test} = 0.81$). A great advantage of the PLS system is that because of its linearity, it can easily be analyzed as to which input variables are most important. It is also possible to gain access to the "important variables" when using neural networks, yet some more sophisticated approach is required.[7] We have not performed such an analysis for the present application. If one is only interested in a crude, binary frequent hitter classifier, the neural network approach seems to outperform the PLS model. The binary classification ability of the neural network system was further assessed by randomly dividing the

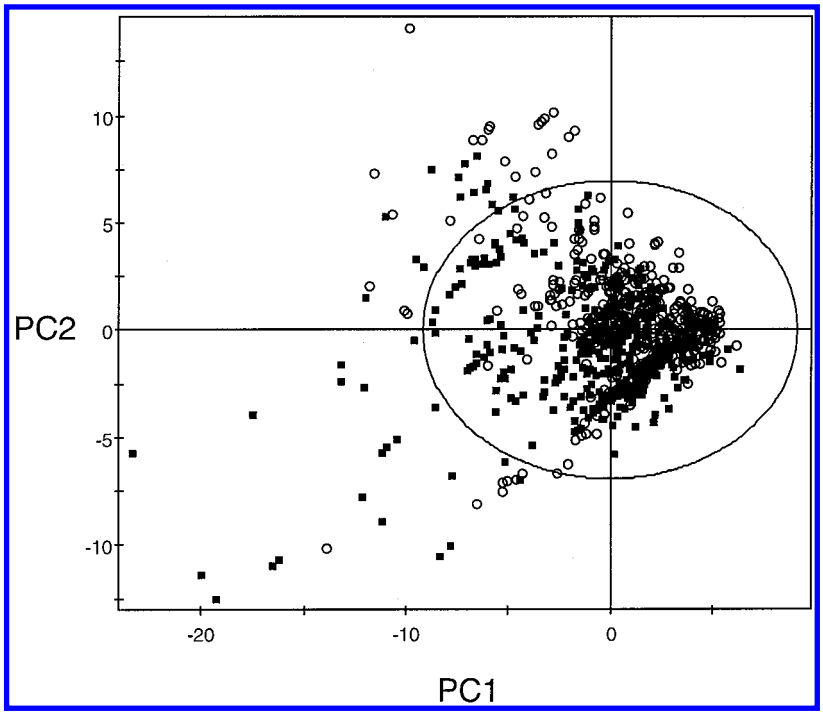

Figure 2. Score plot of the principal component projection of the frequent hitter set (902 compounds) on the plane formed by the two most important principal components (PC1 and PC2) derived from the GC descriptors. Frequent hitters are depicted as white circles, nonfrequent hitters as black squares. The ellipse represents the projection of the 95% confidence region limit in the two-dimensional score plot.

Table 1. Ten Most Relevant Descriptors According to VIP Analysis

variables[a]	VIP
aromatic−OH	2.72
C═C(−X)−C	2.68
X	2.61
C_{sp^3}, having 1 X attached to next carbon	2.03
C═C(−C)═C	1.99
F−C_{sp^3}−X	1.91
Cl−C_{sp^3}−X	1.91
Cl−C_{sp^3}−not X	1.62
F−C_{sp^3}−not X	1. 62
Br−C_{sp^3}−not X	1. 62

[a] X represents any heteroatom (O, N, S, P, Se, halogens); ═ represents double bonds.

dataset in three parts: a training set composed of 60% of the compounds, a testing set composed of 20%, and a validating set composed of the last 20%. By use of these training data for network optimization, the Matthews' correlation was $cc_{test} = 0.80$ for both the testing and the independent validating part. This result indicates that the network found a stable solution that seems to be valid not only for the training data. When both training and test sets were used for network training, the prediction accuracy was increased to $cc_{validation} = 0.83$. The fact that the augmented training set led to an improved neural network model suggests that the overall size of our frequent hitter database needs to be increased to obtain more generalizing solutions.

The neural network containing two hidden neurons was also trained with the complete data set to establish the final filter system. It yields a Matthews correlation $cc_{training} = 0.92$ with 96% of the frequent hitters correctly reclassified and only 4% of the nonfrequent hitters classified as frequent hitters (false positives). A clear discrimination between the two classes of molecules can be observed in the scores histogram (Figure 4). Although in total approximately 90% of the data were correctly

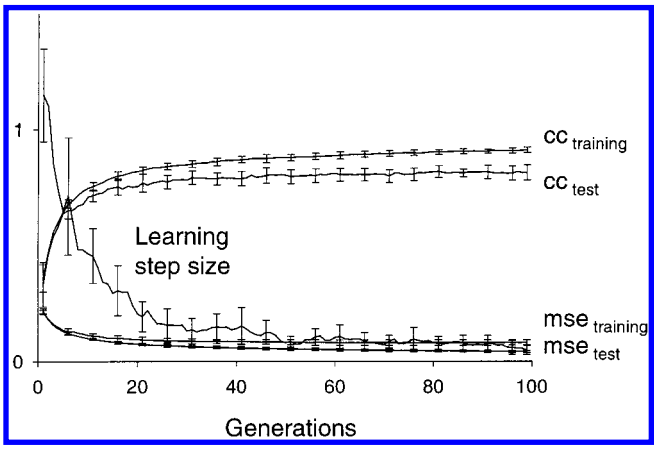

Figure 3. Evolution of the mean square error (mse), Matthews correlation (cc), and the learning step size during neural network optimization for the training and the test data. The values were averaged over 10 independent runs; the error bars give standard deviations.

classified by this system, we observed a small percentage of potentially false positives (drugs predicted to be frequent hitters) that had to be analyzed in more detail.

Analysis of False Positives. Figure 5 shows the structures of some of the therapeutic drugs predicted to be frequent hitters, together with their prediction score. For most of these drugs, the appropriate classification will depend on the context. We have to deal with a "twilight zone" in our frequent hitter definition. For example, if we want to target the central nervous system, dopamine-like molecules will certainly not be considered as frequent hitters; rather, they might be regarded as "privileged structures". A similar argument can be made for the polyiodo compounds dextrotiroxina and dextrotiroxina-sodica, which target the thyroid. One possible way to clarify this point would be to form a new set of frequent hitters containing only one representa-

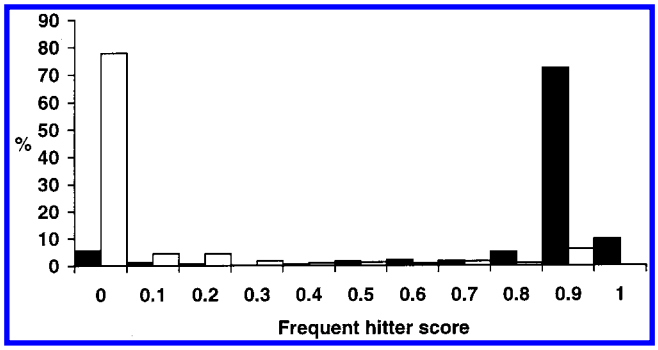

Figure 4. Distribution of the neural network output (prediction score) values for the test data set (10-times cross-validation result). The target values were 1 for frequent hitters (black bars) and 0 for nonfrequent hitters (white bars).

tive for a class of similar receptors and to repeat the analysis. Molecules such as tocopherol, calciferol, and idebenone, on the other hand, can be more easily classified as being "real" frequent hitters. Clearly, the term "frequent hitter" (as we defined it here) is not a synonym for "undesired structure". The instances shown in Figure 5 stress the assumption that there probably is no strict definition of a frequent hitter; rather, an appropriate classification of a molecule depends on its target. This is why this frequent hitter filter should be used as a new flagging routine and not as an elimination criterion in the virtual screening cascade. It may be used for compound prioritization, thereby complementing existing tools such as the "druglikeness filter", the "rule-of-five", toxicity flagging, etc.

Table 2. Fractions of "Druglike" Molecules and "Frequent Hitters" Predicted for Several Compound Databases and Correlation of the "Druglikeness" and "Frequent Hitter" Scores (r^2)

database	compounds	frequent hitters, %	druglike, %	r^2
ACD	183221	35	26	0.08
WDI	55750	22	81	0.03
MedChem	36418	16	52	0.01
Trade Drugs	3344	13	76	0.05
therapeutic drugs[a]	703	8	81	0.02

[a] From Roche database.

Application to Database Filtering. We have used the frequent hitter filter to get an idea of the proportion of potential frequent hitters in different databases. Predictions were made for ACD, WDI, MedChem, Trade Drugs, and our in-house therapeutic drugs database. To see whether there is a correlation with a neural network predicting "druglikeness", we trained a neural network to separate "drugs" from "nondrugs", consequently following the original idea of Sadowski and Kubinyi.[8] Instead of the conventional back-propagation-of-errors approach, an evolution strategy was used for network training.[8,30] We were able to confirm their results yielding 81% overall correct predictions compared to 80%. This accuracy level was confirmed by a series of similar experiments based on in-house sets of "drug" and "nondrug" data (results not shown). Table 2 gives the

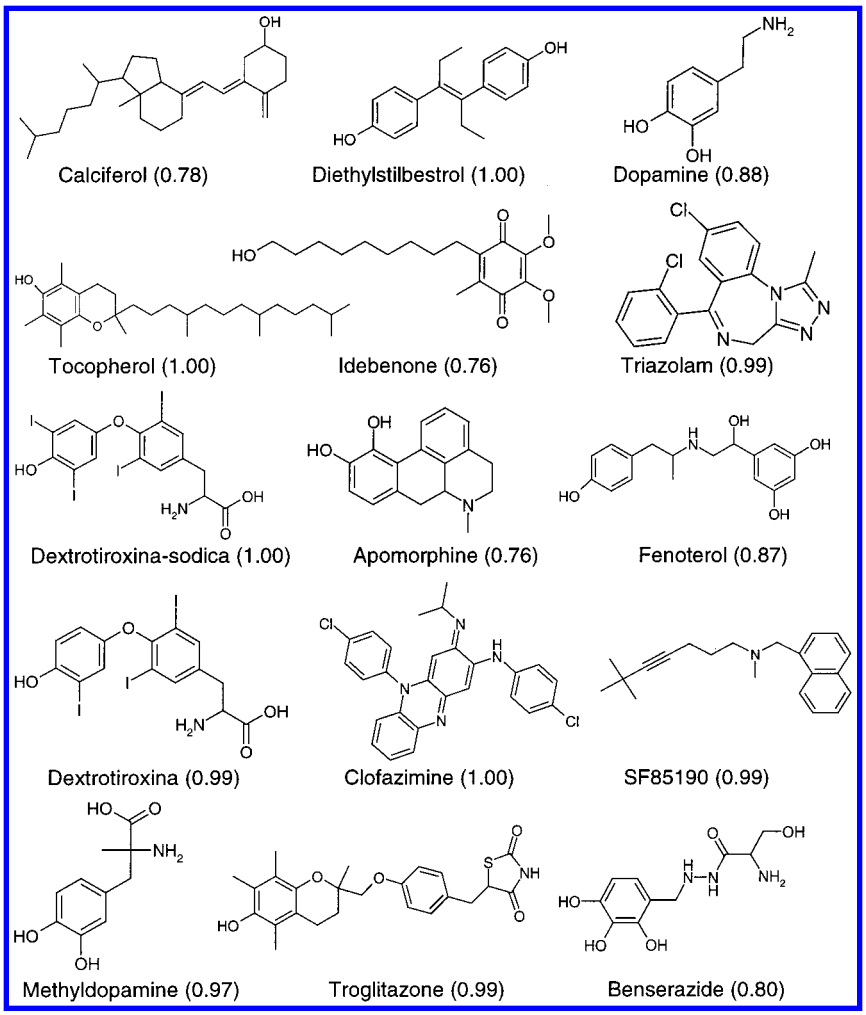

Figure 5. Structures of 15 drug molecules classified as frequent hitters by the prediction routine.

fractions of druglike molecules and frequent hitters predicted by our models for each database. The largest fraction of frequent hitters is reported for the ACD collection (35%) containing all kinds of commercially available chemicals. The predicted fraction of druglike structures in ACD is only 26%. It turns out that with an increasing druglikeness of the database, a decreasing fraction of frequent hitters is predicted. The two compilations of drug molecules (trade drugs from WDI and therapeutic drugs) gain approximately 80% druglikeness and only around 10% frequent hitters. One might assume that the druglikeness score and the frequent hitter score are correlated. Actually, the squared correlation coefficients underscore that no relationship exists between the two scores (Table 2). This is a surprising finding because both filter systems, the druglikeness filter and our new frequent hitter filter, are based on the identical set of GC descriptors providing the neural network input. The models clearly differ in the descriptor-weighing scheme, which makes them independent tools for analyzing distinct structural features.

Conclusions

A fast automatic scoring scheme has been established and parametrized for the discrimination between frequent hitters and nonfrequent hitters. It succeeded in correctly classifying approximately 90% of the compounds, which is an astonishing result because our set of frequent hitters represents a heterogeneous group formed by compounds that perturb the assays, compounds that bind nonspecifically to different targets, and even potentially privileged structures. We also observed that there is no strict consensus among medicinal chemists about what structural attributes characterize a frequent hitter molecule. Nevertheless, our analysis revealed certain features that either qualify or disqualify a compound as a frequent hitter. This new filter can be applied to prioritize compounds from large databases, for purchase or biological testing, and also in the construction of new virtual libraries.

Acknowledgment. The authors thank Hans-Joachim Böhm, Malcolm Page, Stephen Pickett, Brad Sherborne, Sung-Sau So, and Martin Stahl for helpful discussions and encouragement. Jens Sadowski is thanked for leaving us his compilation of drugs and nondrugs.

References

(1) Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual Screening—an overview. *Drug Discovery Today* **1998**, *3*, 160−178.
(2) Wang, J.; Ramnarayan, K. Toward designing drug-like libraries: a novel computational approach for prediction of drug feasibility of compounds. *J Comb. Chem.* **1999**, *1*, 524−533.
(3) Teague, S. J.; Davis, A. M.; Leeson, P. D.; Oprea, T. The design of leadlike combinatorial libraries. *Angew. Chem., Int. Ed.* **1999**, *38*, 3743−3748.
(4) 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.
(5) Rishton, G. M. Reactive compounds and in vitro false positives in HTS. *Drug Discovery Today* **1997**, *2*, 382−384.
(6) Sadowski, J. Database profiling by neural networks. In *Virtual Screening for Bioactive Molecules*; Böhm, H.-J., Schneider, G., Eds.; Wiley-VCH: Weinheim, New York, 2000; pp 117−129.
(7) Ajay, A.; Walters, W. P.; Murcko, M. A. Can we learn to distinguish between "drug-like" and "nondrug-like" molecules? *J. Med. Chem.* **1998**, *41*, 3314−3324.
(8) Sadowski, J.; Kubinyi, H. A scoring scheme for discriminating between drugs and nondrugs. *J. Med. Chem.* **1998**, *41*, 3325−3329.
(9) Frimurer, T. M.; Bywater, R.; Naerum, 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.
(10) Clark, D. E.; Pickett, S. D. Computational methods for the prediction of "drug-likeness". *Drug Discovery Today* **2000**, *5*, 49−58.
(11) 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 Delivery Rev.* **1997**, *23*, 3−25.
(12) Huuskonen, J.; Rantanen, J.; Livingstone, D. Prediction of aqueous solubility for a diverse set of organic compounds based on atom-type electrotopological state indices. *Eur. J. Med. Chem.* **2000**, *35*, 1081−1088.
(13) Zuegge, J.; Schneider, G.; Coassolo, P.; Lavé, T. Prediction of hepatic metabolic clearance—comparison and assessment of prediction models. *Clin. Pharmacokinet.* **2001**, *40*, 553−563.
(14) *ACD: Available Chemicals Directory*, version 99/1; MDL Information Systems: San Leandro, CA, 1999.
(15) *WDI: World Drug Index*, version 00/2; Derwent Publications, 2000.
(16) *MedChem Database*, version 00; BioByte Corp., 2000.
(17) Roberts, G.; Myatt, G. J.; Johnson, W. P.; Cross, K. P.; Blower, P. E. LeadScope: Software for exploring large sets of screening data. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1302−1314.
(18) *TSAR* version 3.21; Oxford Molecular Ltd. (The medawar center, Oxford science park, Oxford OX4 4GA, England).
(19) Meylan, W. M.; Howard, P. H. Atom/fragment contribution method for estimating octanol−water partition coefficients. *J. Pharm. Sci.* **1995**, *84*, 83−92.
(20) Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. Atomic physicochemical parameters for three-dimensional structure directed quantitative structure−activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163−172.
(21) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. "Scaffold-hopping" by topological pharmacophore search: a contribution to virtual screening. *Angew. Chem., Int. Ed.* **1999**, *38*, 2894−2895.
(22) Otto, M. *Chemometrics—Statistics and Computer Application in Analytical Chemistry*; Wiley-VCH: Weinheim, New York, 1999.
(23) Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Wold, S. *Introduction to Multi- and Megavariate Data Analysis Using Projection Methods (PCA & PLS)*; Umetrix: Umeå, Sweden, 1999.
(24) Wold, S. Exponentially weighted moving principal component analysis and projection to latent structures. *Chemom. Intell. Lab. Syst.* **1994**, *23*, 149−161.
(25) *SIMCA-P*, version 8.0; Umetrics AB (Box 7960, S-907 19): Umeå, Sweden.
(26) Kohonen, T. Self-organized formation of topologically correct feature maps. *Biol. Cybern.* **1982**, *43*, 59−69.
(27) Schneider, G.; Wrede, P. Artificial neural networks for computer-based molecular design. *Prog. Biophys. Mol. Biol.* **1998**, *70*, 175−222.
(28) Kernighan, B. W.; Ritchie, D. M. *The C programming language*, 2nd ed; Bell Telephone Laboratories, Incorporated, 1988.
(29) Matthews, B. W. Comparison of predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **1975**, *405*, 442−451.
(30) Hertz, J.; Krogh, A.; Palmer, R. G. *Introduction to the Theory of Neural Computation*; Addison-Wesley: Redwood City, CA, 1991.
(31) Dollery, C., Ed. Therapeutic drugs, 2nd edition; Churchill Livingstone: Edinburgh, 1999.
(32) Lee, M. L.; Schneider, G. Scaffold architecture and pharmacophoric properties of natural products and trade drugs: Application in the design of natural product-based combinatorial libraries. *J. Comb. Chem.* **2001**, *3*, 284−289.

JM010934D