Structure-Based versus Property-Based Approaches in the Design of G-Protein-Coupled Receptor-Targeted Libraries

Konstantin V. Balakin,* Stanley A. Lang, Andrey V. Skorenko, Sergey E. Tkachenko, Andrey A. Ivashchenko, and Nikolay P. Savchuk

Chemical Diversity Labs Inc., 11558 Sorrento Valley Road, San Diego, California 92121

Received June 11, 2003

In this work, two alternative approaches to the design of small-molecule libraries targeted for several G-protein-coupled receptor (GPCR) classes were explored. The first approach relies on the selection of structural analogues of known active compounds using a substructural similarity method. The second approach, based on an artificial neural network classification procedure, searches for compounds that possess physicochemical properties typical of the GPCR-specific agents. As a reference base, 3365 GPCR-active agents belonging to nine different GPCR classes were used. General rules were developed which enabled us to assess possible areas where both approaches would be useful. The predictability of the neural network algorithm based on 14 physicochemical descriptors was found to exceed the predictability of the similarity-based approach. The structural diversity of high-scored subsets obtained with the neural network-based method exceeded the diversity obtained with the similarity-based approach. In addition, the descriptor distributions of the compounds selected by the neural network algorithm more closely approximate the corresponding distributions of the real, active compounds than did those selected using the alternative method.

INTRODUCTION

The superfamily of GPCRs is a diverse group of transmembrane proteins involved in signal transduction.[1] GPCRs initiate a cascade of cellular responses to diverse extracellular mediators and play a fundamental role in physiology and pathophysiology. Nearly 40% of marketed drugs act by modulating some GPCR function.[2] For this reason, GPCRs are potential targets for therapeutic intervention in many diseases. In addition, identifying the function of GPCRs for which a ligand has not yet been identified (orphan GPCRs) could provide a path to the discovery of new cellular components that are important in human physiology.[3] Indeed, simply being able to remove these novel proteins from the orphan GPCR list would, in itself, have a major impact on the diagnosis and treatment of a wide range of human diseases.

The key to harnessing the therapeutic benefits of GPCRs is in the ability to identify highly selective pharmaceutical ligands for these receptors and to then elucidate their functions and evaluate their clinical potential. It is apparent that for preclinical discovery the optimal ligands will be small molecules that possess many of the features required for the final orally available small-molecule drug. Specifically, optimal ligands will possess high affinity and specificity for the target protein and will have reasonable membrane permeability to maximize the probability of significant biological activity in whole cell assays. Availability of such GPCR-specific small-molecule libraries is a necessity in the modern pharmaceutical industry, and the crucial issue is determining which strategy is the most appropriate in the design of such libraries.

* Corresponding author phone: (858)794-4860; fax: (858)794-4931; e-mail: kvb@chemdiv.com.

Successful approaches to the design of focused compound libraries have recently been reported. They range from 2D simulation algorithms to the analysis of ligand−receptor spatial arrangement and neural network learning QSAR systems. With such a variety of tools, it now appears possible to design libraries that are enriched in compounds possessing the desired target-specific properties.

Similarity searches are now a standard tool for target-specific library design.[4,5] The underlying theory is that, given a compound with a desired biological activity, compounds that are "similar" to it in structure are likely to have similar activity. In a common practice of focused library design, an investigator provides a set of chemical structures as a "probe," searches over a database of compounds, finds those that are most similar, and submits them for testing. Similarity searching can be done on the basis of either 2D or 3D structure. 2D similarity searches are computationally very inexpensive and rapid; their popularity is related to the trend toward applying high-throughput medicinal chemistry methodologies.

Artificial neural networks (ANN) have long been used to solve classification problems.[6] Several recent papers describe the successful employment of different neural network approaches to distinguish among different categories of compounds. Ajay et al.[7] used a Bayesian neural network for discriminating drugs from nondrugs. Using this method, they correctly classified 80% of the compounds from the MDDR. The same group of researchers[8] described a solution in the design of a CNS-active library based on a similar neural network classification procedure. Appearing in back-to-back issues with the Ajay et al.[7] article was a contribution from Sadowski and Kubinyi[9] in which the authors described the development of a feed-forward neural network method for discriminating drugs from nondrugs. Our recent observations

10.1021/ci034114g CCC: $25.00
Published on Web 08/07/2003

Table 1. Database of GPCR Ligands

receptor type	ligand type	compounds
tachykinin NK1	antagonists	940
5-HT1a	agonists	618
dopamine D2	agonists and antagonists	423
sigma receptors	antagonists	362
corticotropin-releasing factor	antagonists	301
dopamine D3	antagonists	245
δ-opioid	agonists	185
κ-opioid	agonists	147
$\alpha-2$ adrenoceptor	antagonists	144
total database		3365

indicate that the presence of a combination of some specific physicochemical features can accurately differentiate the GPCR ligands from compounds belonging to other target-specific classes.[10] Using these findings, a neural network classification model was created with excellent discriminatory power. In a control experiment with a random selection of GPCR(+) and GPCR(−) test compounds, up to 92% of GPCR ligands and 93% of non-GPCR ligands were correctly predicted. The ability of a neural network to optimize a large number of input parameters is a very useful feature of this approach; it enabled us to name the procedure a "property-based design", analogous to the terminology proposed by van de Waterbeemd et al.[11] Regarding these "properties", the authors intended that physicochemical as well as pharmacokinetic characteristics of compounds be taken into consideration at the early stages of the drug design process.

After we effectively discriminated the GPCR ligands from non-GPCR active compounds,[10] we attempted to solve another, more difficult problem: whether the specific GPCR classes could be differentiated from each other. Two alternative methodologies were studied to address this problem: a fragment similarity-based search and a neural network classification procedure based on a preselected set of physicochemical descriptors. The goal was to assess the usefulness of these approaches in designing small-molecule libraries that would show preferential and specific GPCR activity. The term, "specific GPCR activity", refers to the ability of a compound to be a successful ligand for a specific GPCR. These observations can be generalized to aid in the selection of an optimal methodology for a target-specific library design; it is not restricted by the GPCR-targeted libraries studied here. In addition, the results can be used for profiling the bioactivity of compounds based on comparison with the structures of known agents possessing a certain biological activity.

METHODS

Database. A database of 3365 known GPCR ligands belonging to nine particular target-specific classes was used as a source of structural information (Table 1). Compounds in the stages of (pre)clinical trials, marketed drugs, and compounds with proven in vivo GPCR activity were included in this data set. An important remark should be made about the Sigma receptor family. The signal transduction mechanisms for this receptors have not been fully elucidated. Several papers reported that Sigma receptors may be coupled to G-proteins.[12,13] Recently, these receptors were revealed to have no seven transmembrane moiety in its deduced amino acid sequence.[14] It was also demonstrated that the Sigma receptors are not directly coupled to G proteins.[15] However, we believe it was useful to evaluate the developed classification algorithms for their ability to discriminate between the closely related compound categories participating in eukaryotic signal transduction mechanisms. Such comparison is particularly interesting as many Sigma receptor ligands have structures similar to the structures of GPCR ligands.

This database was used in all the similarity- and neural network-based experiments. All compounds were selected from the Ensemble database[16] which is a licensed database of known pharmaceutical agents compiled from the patent and scientific literature. Molecules were filtered based on molecular weight range (150−850) and atom type content (only C, N, O, H, S, P, F, Cl, Br, and I allowed).

Similarity Searches. The counts of the numbers of fragment substructures common to a pair of molecules provided a computationally efficient and effective basis for quantifying the degree of structural resemblance between the two molecules under consideration. In this work, a conventional screen search was used to identify the matches to the queried substructures that occur in the reference training set, and then a similarity measure based on the screens common to the target and to each of these matches was used to rank the substructure-search output in order of decreasing similarity to the query. Specifically, the similarities were calculated using the cosine coefficient.

Definitions of Similarity by ChemoSoft. ChemoSoft software [CDL, Inc.] was used for all similarity and diversity calculations performed in this study. The usefulness of these algorithms for selecting compounds for bioscreening are reported in a previous paper.[17] Only the principal features are presented here.

To calculate similarity, ChemoSoft divides a molecule into structural fragments (screens). Screens are simple structural fragments, centroids, with the topological distance equal to one bond length between the central atom and the atoms maximally remote from it. Screens are stored in an internal database. If a new screen appears during calculation, it is added to the database automatically. Given the fact that the internal database contains N screens that define the total data set, each compound in the data set can be represented as an N-dimensional vector, elements of the vector being 0 if the screen is absent in the compound, or 1 if the screen is present. These screens are used to define the similarity as *SIMILARITY(I,J)* for any pair of compounds. The angle cosine between the corresponding bit-string vectors M_I and M_J are used as a similarity measure of compounds *I* and *J*. The cosine is a frequently used similarity measure for a bit vector. The number of components *F* of these vectors is equal to the number of screens in the set, and the *M(K)* component is equal to 1, if the *K*th fragment is present in the compound, and to 0 in the opposite case:

$$SIMILARITY(I,J) = \frac{\sum_{K=1}^{F} M_I(K) M_J(K)}{\sqrt{\sum_{K=1}^{F} M_I(K)^2 \sum_{K=1}^{F} M_J(K)^2}} \quad (1)$$

Table 2. Descriptors Used in This Study

descriptor	definition	absolute average t'-value
	(a) Atom and Bond Count Descriptors[a]	
a_nN	no. of nitrogen atoms	8.64
a_nO	no. of oxygen atoms	7.29
a_aro	no. of aromatic atoms	11.72
a_Hdon	no. of hydrogen bond donors	8.64
a_Hacc	no. of hydrogen bond acceptors	9.20
	(b) Physical Properties	
MW[a]	molecular weight	10.80
$LogD_{74}$[a]	log of the 1-octanol/water partition coefficient (pH 7.4)	7.00
Apol[b]	sum of atomic polarizabilities	9.19
	(c) Jurs Charged Partial Surface Area Descriptors[b]	
JPNSA-1	partial negative surface area: sum of the solvent-accessible surface areas of all negatively charged atoms	11.76
JFNSA-1	descriptor obtained by dividing descriptor JPNSA-1 by the total molecular solvent-accessible surface area	8.88
JFPSA-1	descriptor obtained by dividing the partial positive surface area (sum of the solvent-accessible surface areas of all positively charged atoms) by the total molecular solvent-accessible surface area	8.88
JPNSA-3	atomic charge weighted negative surface area: sum of the product of solvent-accessible surface area × partial charge for all negatively charged atoms	10.45
JWNSA-2	descriptor obtained by multiplying the total charge weighted negative surface area (partial negative solvent-accessible surface area multiplied by the total negative charge) by the total molecular solvent-accessible surface area and dividing by 1000	11.01
JWPSA-3	descriptor obtained by multiplying the atomic charge weighted positive surface area (sum of the product of solvent-accessible surface area × partial charge for all positively charged atoms) by the total molecular solvent-accessible surface area and dividing by 1000	7.18

[a] Calculated by ChemoSoft software. [b] Calculated by Cerius[2] software.

The Tanimoto similarity coefficient is frequently used instead of cosine coefficients. It has been shown, however,[18] that using the cosine and Tanimoto values leads to highly correlated similarity values.

To calculate the similarity value for the *I*th compound in a test set to a specified set *L* consisting of more than one structure, the average similarity is used

$$S_{I,set} = \Sigma S_{IJ}/N \qquad (2)$$

where $S_{I,set}$ is the average similarity of the *I*th compound in the collection with the data set *L*, S_{IJ} is the similarity of the *I*th compound in the collection with the *J*th compound in *L*, and *N* is the number of compounds in this set.

Molecular Descriptors. Sixty molecular descriptors encoding important molecular properties, such as lipophilicity, charge distribution, topological features, steric, and surface parameters, were explored. These descriptors were calculated for the entire 3365-compound database with the Cerius[219] and ChemoSoft software tools. For calculation of $LogD_{74}$, the SLIPPER program,[20] integrated into the ChemoSoft environment, was used. After removing the highly correlated ($R > 0.9$) descriptors, the total number was reduced to 47. To further reduce the number of descriptors that could contain redundant information, the following statistical analysis was performed. To measure the ability of molecular descriptors to discriminate between two categories of compounds, t'-statistics has been used. Nine calculations were performed, where one category was represented by compounds belonging to a particular GPCR-ligand class and the second category by the total database. The difference in the means of the distributions for each descriptor can be expressed in units of standard error according to the formula

$$t'_i = (X_{tot} - X_i)/\sqrt{(\sigma_{tot\tau}{}^2/n_{tot} + \sigma_i{}^2/n_i)} \qquad (3)$$

where for each of the two compound sets *X* is the mean, σ^2 is the variance, *n* is the sample size, and indices "*tot*" and "*i*" denote the total data set and the particular target-specific set, respectively; $n_{tot} = 3365$; n_i values are shown in Table 1. For the compound sets studied here, the trend is toward normal distribution; accordingly, t'-statistics are applicable for testing the significance of the difference between the two distributions for each descriptor. Average absolute values of the t'-parameters were used for comparative evaluation of the descriptors. For each GPCR-activity class, a more accurate descriptor set (a set with maximal t'-values for a particular class) could be applied to achieve a higher level of discrimination; however, for reasons of simplicity and to demonstrate the possibility for generalization, we used a unified set of 14 descriptors that possess a reasonably high level of discrimination ability ($t' > 7{,}0$) for all neural network operations. The final set of descriptors is shown in Table 2.

Detailed analysis of the selected descriptors is not reported here because of the volume of the data. We will note, however, that the measured descriptors encode several key molecular features that determine both its potential ability to bind to a specific receptor and its potential pharmacokinetic behavior. These features are molecular size (MW), lipophilicity (a_aro, $LogD_{74}$), electronic (Apol) and surface properties (number of H-bond donors/acceptors, Jurs surface

descriptors). In the course of neural network modeling of such a complex set of factors, simultaneous optimization can be referred to as property-based design, which is opposite to a similarity-based algorithm. In this sense, a comparison of these two approaches is a comparison of structure-based and property-based strategies in GPCR-specific library design.

Neural Network Modeling. The NeuroSolution 4.0 program[21] was used for all neural network operations. Feed-forward nets were constructed that consist of 14 input neurons (descriptors in Table 2), one hidden layer with four processing elements, and two output neurons. The networks were trained with the molecular descriptors as input values and the scores as output values. The final score was calculated by subtracting the "activity" score from the "nonactivity" score. The back-propagated nets were trained following the momentum learning rule as implemented in the NeuroSolution program. The training was performed over 1000 iterations.

Methodology of Scoring. The general method used as a basis for comparing the two alternative approaches consisted of dividing each group of GPCR ligands into training and test sets, applying a scoring procedure specific to each method, ranking the compounds according to their predicted activity, and assessing the quality of the discrimination between compounds belonging to one specific class and all other compounds. Nine independent series of experiments corresponding to the nine target-specific classes studied in this work were carried out; the tested compounds were sorted according to their ability to bind to a specific GPCR. Accordingly, within each series of experiments, two principal categories of compounds were considered: "actives" and "nonactives". For example, in the experiments related to discrimination of tachykinin NK1 receptor-specific compounds, 940 "actives" and 2425 "nonactives" were considered (see Table 1). These experiments model the real procedure of target-specific library design in which a subset of molecules having a high potential to be successful ligands to a particular target is searched in a database of diverse compounds. Three randomizations corresponding to the complementary training/cross-validation/testing sets were used for each of the nine target-specific classes of GPCR ligands.

Before each scoring procedure, the total randomized sets of 3365 molecules were subdivided into three parts: (1) the training set of 1682 compounds (50% of the total number of compounds), (2) the cross-validation set of 841 compounds (25%), and (3) the test set of 842 compounds (25%). The random sets were generated in such a manner as to contain the percentage of actives in each part of the selection proportional to the relative size of each set (50% of actives in the training, 25% of actives in both cross-validation and test sets). For the similarity search, the compounds of the test set were sorted according to their similarity to the actives of the training set. For the neural network classification modeling, the neural networks were trained with the compounds of the training set and the cross-validation set was used to avoid overtraining. The assumption was made that all the molecules within each target-specific class do not possess activity against any of the other eight GPCRs used in the study.

Ranking of Compounds. After all the similarity values or neural network scores were calculated for a test set, they were sorted from high to low similarity/score. Ranks were then assigned: the molecule with the highest similarity/score is rank 1, the next highest, rank 2, etc. Only the ranks of the compounds from this study were used, since the distribution of absolute scores varied between the two approaches.

Measurement of Performance. After the molecules had been ranked, the performance of the scoring schemes was determined using the modified performance measures reported by Kearsley et al.[22,23] These measures are based on a simulated screening experiment on a database of molecules that contains some number of active molecules. The molecules are scored, ranked, and then "assayed" in the order of descending score. The total number of active molecules found can be plotted against the total number of molecules tested (or the position in the ranked list). The measures used in the publications mentioned above are the *global enhancement* (GE) and *initial enhancement* (IE) factors. If A_{50} is the number of molecules that must be tested to find half the active molecules, then the global enhancement is the ratio of the actual A_{50} to the A_{50} expected for the random case:

$$\mathrm{GE} = A_{50,\mathrm{rand}}/A_{50} \tag{4}$$

If R_{10} represents the number of active molecules found after testing the first 10% of molecules in the ranked list, then the initial enhancement indicates how many more active molecules are found than would be expected in a random distribution:

$$\mathrm{IE} = R_{10}/R_{10,\mathrm{rand}} \tag{5}$$

These parameters depend on the relative content of active molecules in the initial test set and the test can lead to incorrect performance measures when the percentage of actives exceeds 3−5%. In this work, the parameters of the *general enhancement efficiency* (*GEE*) and the *initial enhancement efficiency* (*IEE*) were used to introduce a constraint on the number of active compounds where $A_{50,\mathrm{id}}$ is

$$\mathrm{GEE} = [(A_{50,\mathrm{rand}} - A_{50})/(A_{50,\mathrm{rand}} - A_{50,\mathrm{id}})] \times 100\% \tag{6}$$

$$\mathrm{IEE} = [(R_{10} - R_{10,\mathrm{rand}})/(R_{10,\mathrm{id}} - R_{10,\mathrm{rand}})] \times 100\% \tag{7}$$

equal to half the active compounds and $R_{10,\mathrm{id}}$ is equal to 10% of the ranked compounds. The difference between the corresponding measures is illustrated in Figure 1, where a hypothetical curve for the accumulation of actives vs ranked is depicted. Two limiting cases are also shown: "ideal", where all the actives would be at the front of the list, and "random", where all the actives would be randomly distributed throughout the list. If $A_{50} = A_{50,\mathrm{rand}}$, or $R_{10} = R_{10,\mathrm{rand}}$, the discrimination efficiency and the values of parameters 5 and 6 are equal to 0 and, vice versa, if $A_{50} = A_{50,\mathrm{id}}$, or $R_{10} = R_{10,\mathrm{id}}$, the discrimination efficiency is 100%. The more closely the curve approximates the ideal line, the better is the discrimination efficiency.

In a compound selection program where only a small fraction of compounds are to be selected, the initial enhancement efficiency is a useful measure; when a large fraction

Table 3. Diversity Parameters[a] for the Target-Specific Classes Studied in This Work and the Corresponding Performance Values Observed for the Two Alternative Scoring Procedures

target-specific class	compounds	diversity[b]	screens[c]	heterocycles	GEE_{50} FS	GEE_{50} ANN	IEE_{10} FS	IEE_{10} ANN
tachykinin NK1	940	0.747	2537	134	72.0	98.8	22.9	98.6
5HT1a	618	0.774	1730	118	47.4	67.4	25.0	37.5
dopamine D2	423	0.764	1657	113	63.4	61.3	18.3	31.7
sigma receptor	362	0.724	1019	74	86.7	93.3	44.4	55.6
corticotropin-releasing factor	301	0.765	1189	71	99.0	96.2	65.8	56.1
dopamine D3	245	0.730	898	54	90.5	87.6	50.0	48.6
δ-opioid	185	0.688	667	33	97.4	88.6	76.1	48.7
κ-opioid	147	0.600	676	37	98.2	82.7	85.3	43.1
α2-adrenoceptor	144	0.818	798	54	52.6	79.8	23.6	35.8
average	374	0.734	1241	76	78.6	84.0	45.7	50.6

[a] Calculated by ChemoSoft software tool (CDL, Inc.). [b] Cosine coefficients are calculated, and the sums of nondiagonal similarity matrix elements are used in the ChemoSoft program as a diversity measure. [c] Screens are simple structural fragments, centroids, with the topological distance equal to one bond length between the central atom and the atoms maximally remote from it.

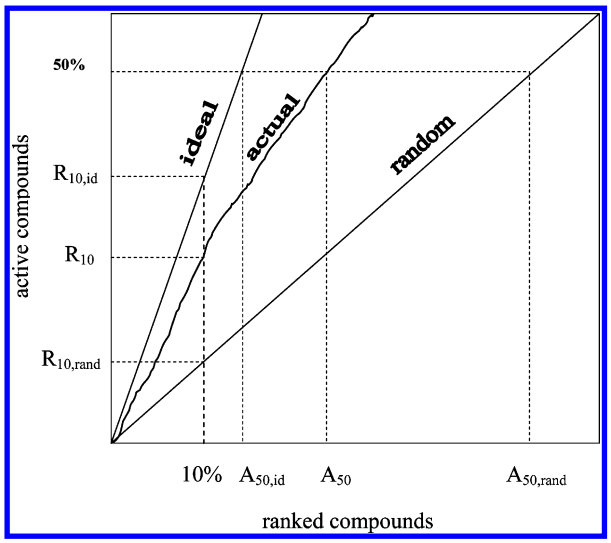


Figure 1. Graph showing hypothetical data on the accumulation of actives vs ranked compounds.

of compounds is to be selected, the global enhancement efficiency is more relevant.

RESULTS AND DISCUSSION

Comparing Results of the Similarity Search and ANN-Based Procedures. Figure 2 shows the graphs for the accumulation of actives versus ranked for the nine target-specific series studied here. These graphs contain data averaged over three independent randomizations (see Methods). Table 3 lists the values of performance for fragment similarity- (FS) versus neural network-based scores. Average values of the performance parameters show that the ANN-based strategy, in general, is better than the similarity-based algorithm in its discrimination ability. Within each particular target-specific series, distinct differences are observed in the discrimination efficiency achieved by the two methods employed. As can be clearly seen, however, the ANN-based approach provides more consistent results. Accordingly, application of the neural network method can be expected to yield more reliable results in terms of an increased hit rate in the synthesized focused libraries.

The discrimination ability observed in these experiments permits us to successfully address the problem of designing a target-specific library. For example, if a collection of 50 000 available compounds contains 1% (typical level of hit rate) that are active against a GPCR target such as studied here, it would be interesting to assess the probable percentage of active compounds contained in the first 10% of the selection after scoring and ranking procedures.

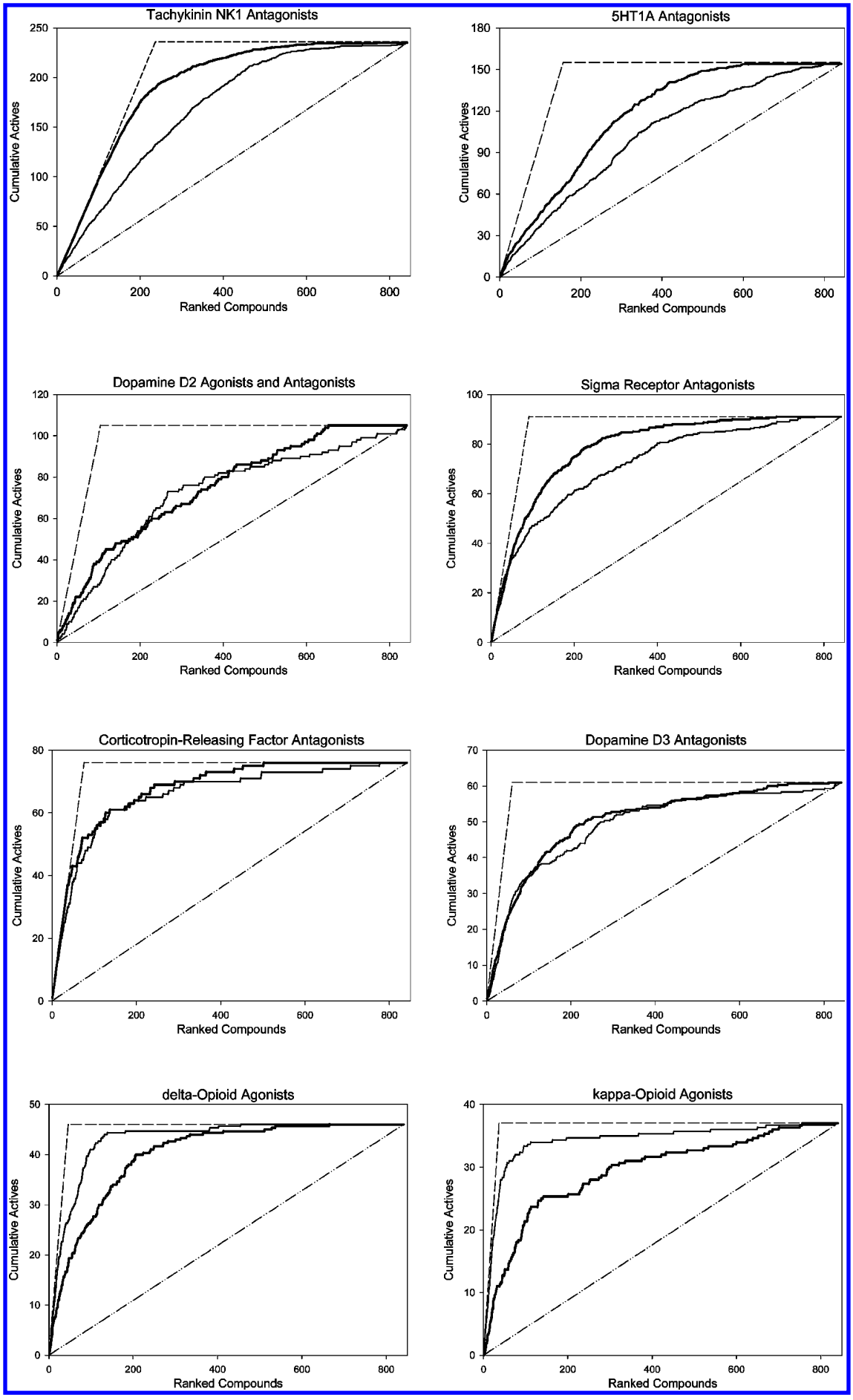
Table 4. Percentage of Active Compounds in the First 10% of the Ranked Database in a Hypothetical Target-Specific Library

IEE_{10}	percentage of actives (100% − 5000 compds)
0 (random set)	1.0
30	3.6−4.0
40	4.5−4.9
50	5.3−5.7
60	6.2−6.6
70	7.1−7.5
80	8.0−8.4
90	8.9−9.3

Table 4 shows the approximate percentages of active compounds in the first 10% of the ranked lists for several discrimination efficiency values typical of those obtained in the course of the experiments described. The tolerance ranges were adjusted according to the findings observed in this work. As can be clearly seen, the IEE_{10} values greater than 40−50% (corresponding to $GEE_{50} \geq 80-90\%$) indicate a reasonably high percent of actives in the final selection, and IEE_{10} values higher than 90% indicate hit rates increased by an order of magnitude. Such results demonstrate that the described scoring methodologies, particularly those based on a neural network scheme, are very useful tools for a significant enhancement of the target-specific content of the focused selections.

Diversity Parameters of High-Scoring Subsets. For optimal screening performance, a target-specific library should be as diverse as possible. To that end, it is important to assess the substructural diversity of the active compounds that occur in the high-scoring subsets of the ranked test sets. Figure 3 shows that the diversity of subsets consists of only active compounds as a function of the number of actives in the ranked list for the two target-specific series. Clearly the diversity achieved using the property-based approach is higher in both cases; a similar tendency was observed for all the target-specific classes studied. These findings appear to be related to the underlying principles of a fragment similarity search based on the selection of structurally similar compounds. In this approach, the dissimilarity, or diversity, of the high-scoring compounds tends to be minimal.

Correlation of the Rankings between the Two Methods. Little to no correlation was found between the two methods


Tachykinin NK1 Antagonists
5HT1A Antagonists
Dopamine D2 Agonists and Antagonists
Sigma Receptor Antagonists
Corticotropin-Releasing Factor Antagonists
Dopamine D3 Antagonists
delta-Opioid Agonists
kappa-Opioid Agonists
Cumulative Actives
Ranked Compounds

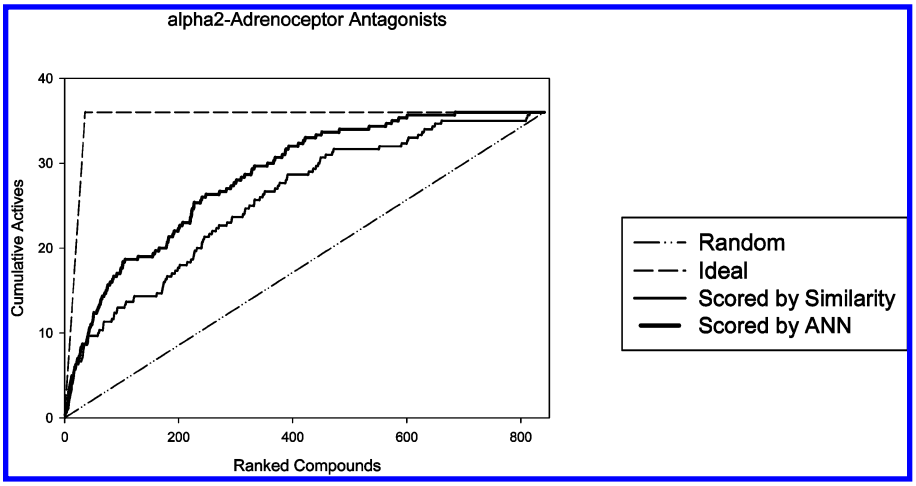


Figure 2. Graph showing accumulation of actives vs ranked for nine target-specific series studied in this work. The solid line is assigned to the ANN ranking procedure and the dashed line is assigned to the fragment similarity-based ranking procedure.

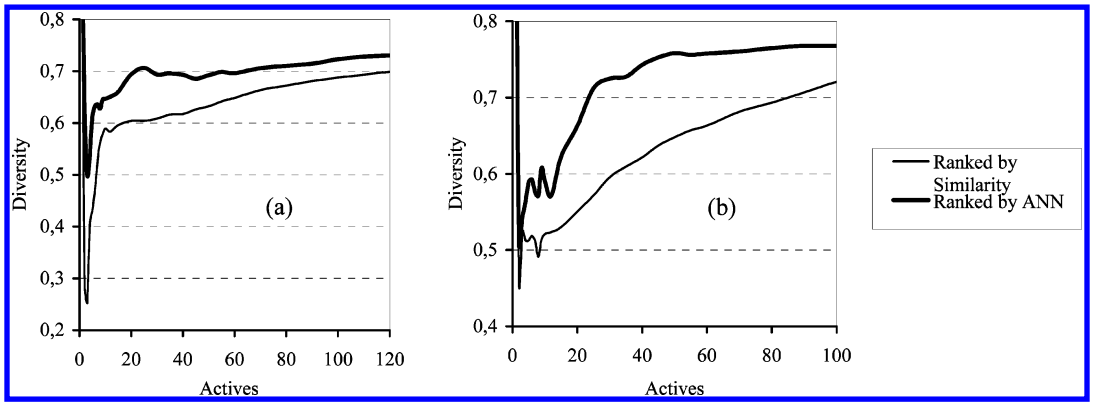


Figure 3. Substructural diversity of the first actives in the ranked lists for Tachykinin NK1 (a) and 5HT1a (b) series.

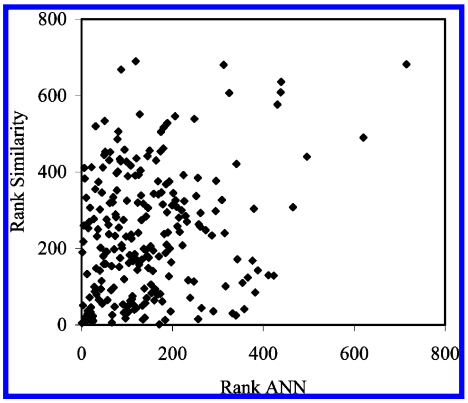


Figure 4. Correlation of ranked actives (test set) for two methods. Tachykinin NK1 series (randomization 1) as an example.

with respect to the rankings of active compounds from the test sets. An example from the tachykinin NK1 series is shown in Figure 4. Evidently, the scattered actives do not fall near the diagonal; in other words, very different rankings are assigned to the same active compounds whether similarity- or ANN-based scoring algorithms are applied. The correlation coefficient between the rankings is 0.22—clear evidence that the equal discriminatory power of the structure-based and property-based approaches does not suggest that the identity (and even close similarity) of the resulting compound selections will be the same. In other words, these methods can be considered complementary, the choice of method depending on the desired outcome in the design of the target-specific library.

Optimization of Molecular Parameters. To assess the ability of the applied methods to optimize the key molecular properties of the generated selections, we analyzed the power of the distributions of several important molecular parameters. Two 210-compound subsets (dopamine D3 series) were scored and ranked with the similarity-based and neural network-based approaches and then compared with the full set of the initial dopamine D3-specific data set (245 compounds). Given that the discrimination ability of the two alternative scoring procedures is approximately equal for this target-specific class (see Figure 2), we selected the dopamine D3 series as the more appropriate comparative. The diagrams below (Figure 5) depict the molecular property distribution profiles obtained for the three compound sets: (i) compounds scored and ranked with neural net (210 compounds, the first 25% of the ranked test set; orange line, triangles), (ii) compounds scored and ranked with similarity method (210 compounds, the first 25% of the ranked test set; pink line, rectangles), and (iii) full initial dopamine D3-specific data set (245 compounds; blue line, diamonds).

Among the compounds selected by the property-based neural network algorithm, the descriptor distributions of the compounds were found to be definitely closer to the corresponding distributions of the really active compounds than those selected using the structural similarity approach. This observation was especially evident for such key parameters as molecular weight and the number of hydrogen-bond acceptors, JPNSA-1 and JFPSA-1. The effect of optimization was negligible in the case of $LogD_{74}$ and also

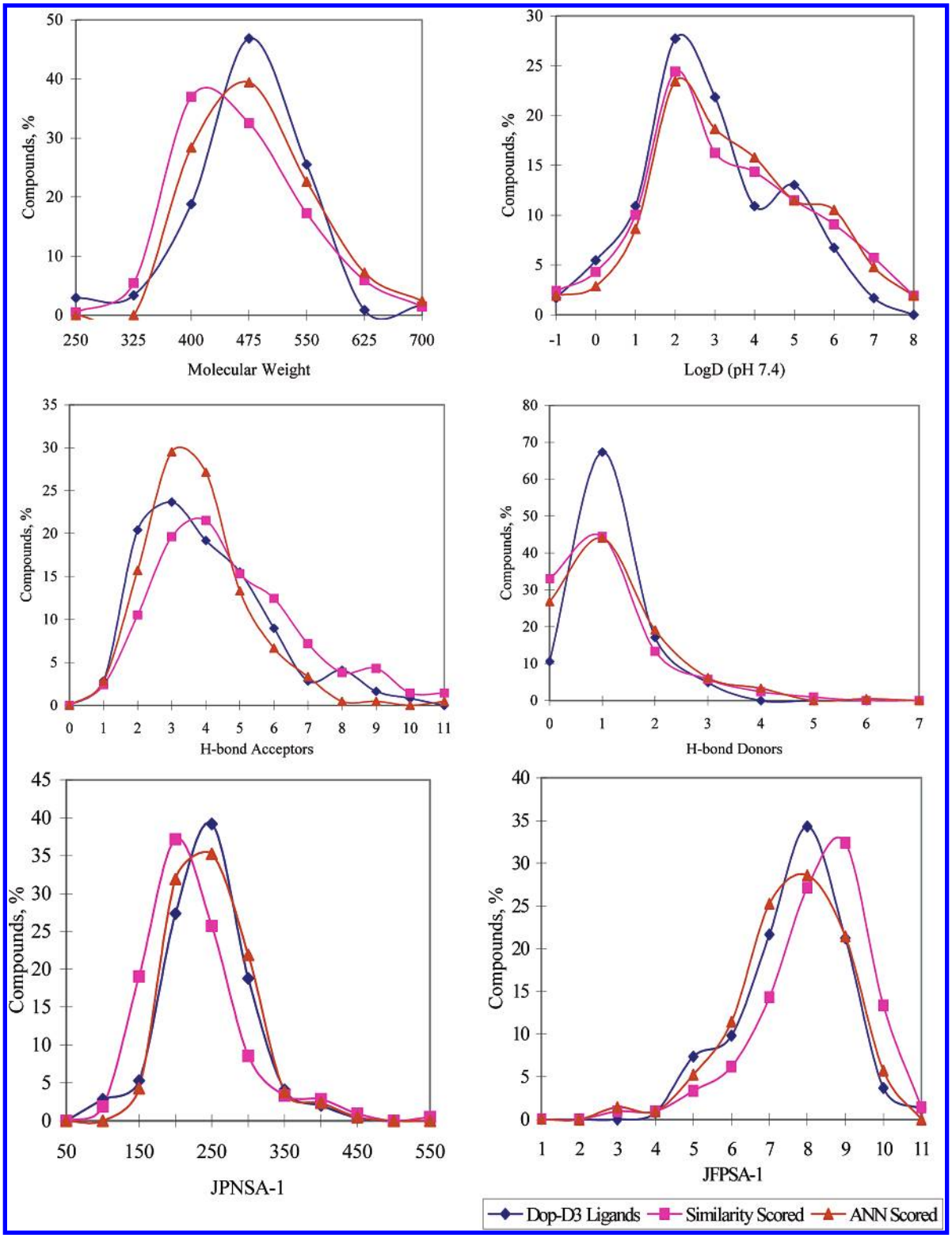


Figure 5. Molecular parameter distributions within the first 25% of the test set (210 compounds) scored and ranked using two alternative algorithms (dopamine D3 series), in comparison with the corresponding distributions for the full initial data setof dopamine D3 receptor ligands (245 compounds).

in the case of the number of hydrogen-bond donors. This finding can be explained by the fact that these parameters are strongly dependent on the presence of specific substructural elements in a molecule, and thus their distributions can also be correlated for the two approaches under consideration. Because of the inherent ability of the applied neural network classification algorithm, based on properties of active/nonactive compounds from the training set, to optimize molecular parameters of the final high-scoring selection, it is anticipated that the observed tendencies are more general in nature.

Properties of Initial Reference Data Sets and Their Influence on Discriminatory Power. In a real situation of target-specific library design there is often a need for prior knowledge of the general rules that govern the ability of the discrimination algorithm to be applied. For that reason, we

Table 5. Correlation Coefficients between the Performance and Diversity Parameters of the Initial Target-Specific Sets

	compounds	diversity	screens	heterocycles
	Fragment Similarity-Based Approach			
GEE_{50}	−0.41	−0.69	−0.47	−0.61
IEE_{10}	−0.59	−0.79	−0.66	−0.77
	ANN-Based Approach			
GEE_{50}	0.08	−0.20	−0.04	−0.24
IEE_{10}	0.68	−0.05	0.60	0.38

conducted a study of the initial properties of the reference data sets and their influence on the ability of the two alternative methods to discriminate. Table 5 shows the correlation between the diversity parameters of the initial target-specific data sets (see Table 3) and the corresponding parameters related to the ability to discriminate within these data sets.

Two main conclusions can be derived from these data. (1) For the fragment similarity-based approach, a clear dependence of GEE_{50} and IEE_{10} values on the diversity of the initial target-specific data sets was observed (correlation coefficients are −0.69 and −0.79, correspondingly). Similar, although less expressed, negative correlations were observed within the number of compounds, screens, and heterocyclic fragments in a data set. Thus it seems that this method will provide good results only if the reference data set of the known active compound prototypes contains a large number of congeneric compounds whose structural diversity is low. (2) The ability of the neural network procedure depends mainly on the size of the reference database, not on diversity. Thus a clear positive correlation with the number of compounds (0.68) in the initial data set was observed for the IEE_{10} parameter. Interestingly, no significant dependencies of GEE_{50} parameter on the diversity parameters of the initial data sets were observed. These observations suggest where the limitations of such an approach lie. The neural network-based method is particularly applicable when a considerable body of quality structural data is available on the ligands to a particular target.

CONCLUSIONS

In this work, nine experimental series with large target-specific data sets of GPCR ligands described in medicinal chemistry literature were carried out. In the course of these experiments designed to model the real process of focused library design, the comparative ability of the two alternative strategies was assessed: the fragment similarity-based search and the neural network classification procedure that is based on a preselected set of physicochemical descriptors. Having evaluated the comparative ability of these two methods to discriminate between different GPCR-specific classes of ligands, we conclude that the ANN-based strategy, in general, is superior to the similarity-based algorithm in these cases. Although within each particular target-specific series distinct differences were observed in the discriminatory ability of the two methods, the ANN-based approach provided more reproducible results. An increased number of congeneric structures in the initial reference data set demonstrate the better performance of the fragment similarity approach. The application of the ANN-based strategy should be favored when a considerable body of quality structural data is available on the ligands to a particular target.

Particularly beneficial for the new lead discovery programs is the ability of the neural network algorithm to generate compound libraries with increased structural diversity as compared with the similarity-based approach. This feature suggests the enhanced probability of finding novel lead chemotypes in these libraries. Another useful property of the high-scoring compound selections generated by the ANN-algorithm is that the descriptor distributions of the compounds are definitely closer to the corresponding distributions of the really active compounds than those selected using the structural similarity approach. The explanation appears to be that the neural network classification algorithm has an inherent ability to optimize molecular parameters of the high-scoring selection.

It can be also concluded that there exists a potential to combine both methods to achieve even more optimal results, as little to no correlation between the rankings of active compounds from the test sets by the two methods was observed. We are currently working on enhancing predictability using the combination of these approaches. There is clear indication that the modified methodology permits to overcome some limitations observed for the described strategies applied separately.

Beyond these observations, it should be noted that the neural network classification methodology has significant applications in various fields of medicinal chemistry. Their increasing popularity can be explained by two main factors: the growing availability of quality structural data on pharmaceutical agents and their targets and the increasing computational power of today's computers. However, several examples of successful applications of this methodology already reported[7−9] were related mainly to the solution of general problems such as generation of drug-like compound libraries. In the present work, we successfully demonstrated the power of the neural network classification approach in the design of highly specific compound libraries targeted for several therapeutically significant GPCRs. Applied at the stage of primary bioscreening, this property-based approach offers a very promising methodology for target-specific library design in the search for novel ligand chemotypes.

On the other hand, as a result of their availability and simplicity, 2D fragment similarity methodologies remain very useful when applied in conjunction with high-throughput technologies in chemistry and biology. This assertion is especially relevant to the design of orphan-GPCR-targeted libraries because of their high degree of structural homology with known active agents. Specifically, the fragment similarity approach is the method of choice when a rapid search of structures possessing a high level of similarity to a predefined set of compounds is required. It is also an optimal methodology for extending existing target-specific libraries.

We anticipate that these results will improve the search for novel GPCR-active agents and contribute to the field of focused library development.

REFERENCES AND NOTES

(1) Gudermann, T.; Nurnberg, B.; Schultz, G. G-Protein-Coupled Receptors and G-Proteins as Primary Components of Transmembrane Signal Transduction. Part 1. G-Protein-Coupled Receptors: Structure and Function. *J. Mol. Med*. **1995**, *73*, 51−63.

(2) Drews, J. Drug Discovery: a Historical Perspective. *Science* **2000**, *287*, 1960−1964.

(3) Howard, A. D.; McAllister, G.; Feighner, S. D.; Liu, Q.; Nargund, R. P.; Van der Ploeg, L. H. T.; Patchett, A. A. Orphan G-Protein-Coupled Receptors and Natural Ligand Discovery. *Trends Pharm. Sci.* **2001**, *22*, 132−140.
(4) *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G. M., Eds.; John Wiley & Sons: New York, 1990.
(5) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(6) Hertz, J.; Krogh, A.; Palmer, R. G. *Introduction to the Theory of Neural Computation*; Addison-Wesley: Redwood City, CA, 1991.
(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) Ajay, A.; Bemis, G. W.; Murcko, M. A. Designing Libraries with CNS Activity. *J. Med. Chem.* **1999**, *42*, 4942−4951.
(9) Sadowski, J.; Kubinyi, H. A Scoring Scheme for Discriminating between Drugs and Nondrugs. *J. Med. Chem.* **1998**, *41*, 3325−3329.
(10) Balakin, K. V.; Tkachenko, S. E.; Lang, S. A.; Okun, I.; Ivashchenko, A. I.; Savchuk, N. P. Property-Based Design of GPCR-Targeted Library. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1332−1342.
(11) van de Waterbeemd, H.; Smith, D. A.; Beaumont, K.; Walker, D. K. Property-Based Design: Optimization of Drug Absorption and Pharmacokinetics. *J. Med. Chem.* **2001**, *44*, 1313−1333.
(12) Itzhak, Y. Multiple Affinity Binding States of the Sigma Receptor: Effect of GTP−Binding Protein-Modifying Agents. *Mol. Pharmacol*. **1989**, *36*, 512−517.
(13) Connick, J. H.; Hanlon, G.; Roberts, J.; France, L.; Fox, P. K.; Nicholson, C. D. Multiple Sigma Binding Sites in Guinea-pig and Rat Brain Membranes: G-protein Interactions. *Br. J. Pharmacol*. **1992**, *107*, 726−731.
(14) Tokuyama, S.; Hirata, K.; Yoshida, A.; Maruo, J.; Matsuno, K.; Mita, S.; Ueda, H. Selective Coupling of Mouse Brain Metabotropic Sigma Receptor with Recombinant Gi1. *Neurosci. Lett*. **1999**, *268*, 85−88.
(15) Hong, W.; Werling, L. L. Evidence That the Sigma(1) Receptor is Not Directly Coupled to G Proteins. *Eur. J. Pharmacol*. **2000**, *408*, 117−125.
(16) Ensemble database of biologically active compounds, Prous Science: 2002. URL: http://www.prous.com.
(17) Trepalin, S. V.; Gerasimenko, V. A.; Kozyukov, A. V.; Savchuk, N. Ph.; Ivashchenko, A. A. New Diversity Calculations Algorithms Used for Compound Selection. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 249−258.
(18) Holliday, J. D.; Ranade, S. S.; Willet, P. A Fast Algorithm for Selecting Sets of Dissimilar Molecules from Large Chemical Databases. *Quant. Struct.-Act. Relat.* **1995**, *14*, 501−506.
(19) Accelrys, Inc., 2000. URL: http://www.accelrys.com.
(20) Raevsky, O. A.; Trepalin, S. V.; Trepalina, H. P.; Gerasimenko, V. A.; Raevskaya, O. E. SLIPPER-2001 − Software for Predicting Molecular Properties on the Basis of Physicochemical Descriptors and Structural Similarity. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 540−549.
(21) NeuroDimension, Inc., 2001. URL: http://www.nd.com.
(22) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical Similarity Using Physicochemical Property Descriptors. *J. Chem. Inf. Comput. Sci*. **1996**, *36*, 118−127.
(23) Hull, R. D.; Fluder, E. M.; Singh, S. B.; Nachbar, R. B.; Kearsley, S. K.; Sheridan, R. P. Chemical Similarity Searches Using Latent Semantic Structural Indexing (LaSSI) and Comparison to TOPOSIM. *J. Med. Chem.* **2001**, *44*, 1185−1191.

CI034114G