Simulation of Sequential Screening Experiments Using Emerging Chemical Patterns

Jens Auer and Jürgen Bajorath*

Department of Life Science Informatics, B-IT, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstr. 2, D-53113 Bonn, Germany

Abstract: A method called "Emerging Chemical Patterns" (ECP) has recently been introduced as a novel approach to binary molecular classification (for example, "active" versus "inactive"). The underlying pattern recognition algorithm was first introduced in computer science and then adopted for applications in medicinal chemistry and compound screening. A special feature is its ability to accurately classify molecules on the basis of very small training sets containing only a few compounds. This feature is highly relevant for virtual compound screening when only very few experimental hits are available as templates. Here we adopt ECP calculations to simulate sequential screening using an experimental high-throughput screening (HTS) data set containing inhibitors of dihydrofolate reductase. In doing so, we focus on minimizing the number of database compounds that need to be evaluated in order to identify a substantial fraction of available hits. We demonstrate that iterative ECP calculations recover on average between ~19% and ~39% of available hits in the data set while dramatically reducing the number of compounds that need to be tested to between ~0.002% and ~9% of the screening database.

Key Words: Pattern recognition, data mining, descriptors, molecular similarity, molecular classification, structure-activity relationships, virtual screening, iterative screening.

1. INTRODUCTION

Machine learning methods have become widely used tools for compound classification. Accordingly, a number of different machine learning algorithms have been introduced for chemical database mining [1-6]. Often, but not always, such methods are used for "binary" or so-called "class label" predictions, for example, "active" versus "inactive". Irrespective of algorithmic details, machine learning depends on the availability of compound sets for training and their performance is often much influenced by the composition of training sets. In addition, there are other issues that are often critical for the successful application of machine learning techniques. For example, predictive models are usually difficult to generate when only a limited number of molecules are available for training and their performance can be greatly affected by data heterogeneity or data sets with a substantial unbalance of class labels. Only a few methods, for example, Bayesian models, have been developed to handle noisy data [6].

We aimed at developing a methodology to build classification models on the basis of very limited training data even in the presence of data noise that is typically found in experimental screening sets. In general, only small numbers of active compounds are available during the early stages of HTS campaigns or hit-to-lead and lead optimization programs. Therefore, we have studied and adopted a concept from computer science called "emerging patterns" [7]. Our ECP approach systematically explores molecular feature patterns and selects patterns that occur with high frequency in one class of compounds (e.g., the "active" class), but not in another (e.g., the "inactive" class) [8]. Molecular features are generated through calculations of large numbers of molecular structure and property descriptors [9]. Such discriminating patterns are then used for class label predictions. In our initial study, we could demonstrate that ECP calculations could distinguish with high accuracy between active compounds at different potency levels, i.e. micromolar versus nanomolar potency [8]. These calculations on different activity classes succeeded on the basis of learning sets consisting of only three to 10 highly potent and weakly potent compounds belonging to the same class. Here we further extend our analysis and investigate whether ECP analysis is capable of supporting iterative biological screening campaigns where often only a few hits are available initially and where screening data are noisy.

Virtual screening (VS) and HTS are complementary in nature [10,11]. The complementarity of computational and experimental screening is best exploited when implementing so-called "iterative" or "sequential" screening schemes [11, 12]. In sequential screening, VS methods such as cluster analysis or similarity searching are applied to pre-select small subsets from large compound libraries for experimental evaluation. As VS templates, already known experimental hits are used or, alternatively, sets of known active compounds from patents or the literature. The underlying idea is to enrich small database subsets with novel hits and establish an iterative computational and experimental screening protocol. The subsets are experimentally screened and newly identified hits are taken into account as additional information during subsequent rounds until a sufficiently large number of novel hits are obtained [10]. Combining VS and HTS in se-

*Address correspondence to this author at the Department of Life Science Informatics, B-IT, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstr. 2, D-53113 Bonn, Germany; Tel: +49-228-2699-306; Fax: +49-228-2699-341; E-mail: bajorath@bit.uni-bonn.de

1573-4064/08 $55.00+.00

quential screening can dramatically reduce the number of compounds that need to be tested. For example, comparing the results of several sequential screening campaigns suggests that at least 50% of hits available in screening libraries can be identified by experimental testing of only 10%-20% of all database compounds [12].

In this study, we have simulated sequential screening trials using our ECP methodology and an experimental HTS data set in order to investigate whether it would indeed be possible to achieve meaningful predictions and enrichments of active compounds when only a few experimental hits are used as VS templates. From a computational point of view, this presents a challenging scenario consistent with tasks for which ECP was developed. In addition, this situation is mirroring early stages of HTS campaigns and is therefore of practical relevance.

2. METHODOLOGY

The computer scientific foundations of ECP are fairly complex. However, in the following section we will describe the methodology in an intuitive manner. The classification technique utilizes combinations of discrete molecular descriptors. In general terms, molecular descriptors represent features relating to molecular structure and properties, mathematically often of greatly varying complexity [9]. Relatively simple examples include molecular mass, the numbers of hydrogen bond donors and aromatic atoms in a molecule, or solvent accessible molecular surface area. The majority of descriptors is numerical in nature and adopts continuous or discrete value ranges. Currently, more than 5,000 different descriptors have been catalogued [9] but no single descriptor captures enough chemical information to, for example, classify compounds according to biological activity. Thus, combinations of varying numbers of molecular descriptors are typically employed for compound classification. The ECP algorithm automatically identifies such combinations from different sample sets of molecules to generate classifiers for class-label prediction.

2.1. Data Mining in Descriptor Spaces

In a pictorial view, descriptors constitute a multi-dimensional space where each molecule is associated with a vector depending on its descriptor values. Fig. (**1**) shows a simple space built from two descriptors, D_1 and D_2. Twenty

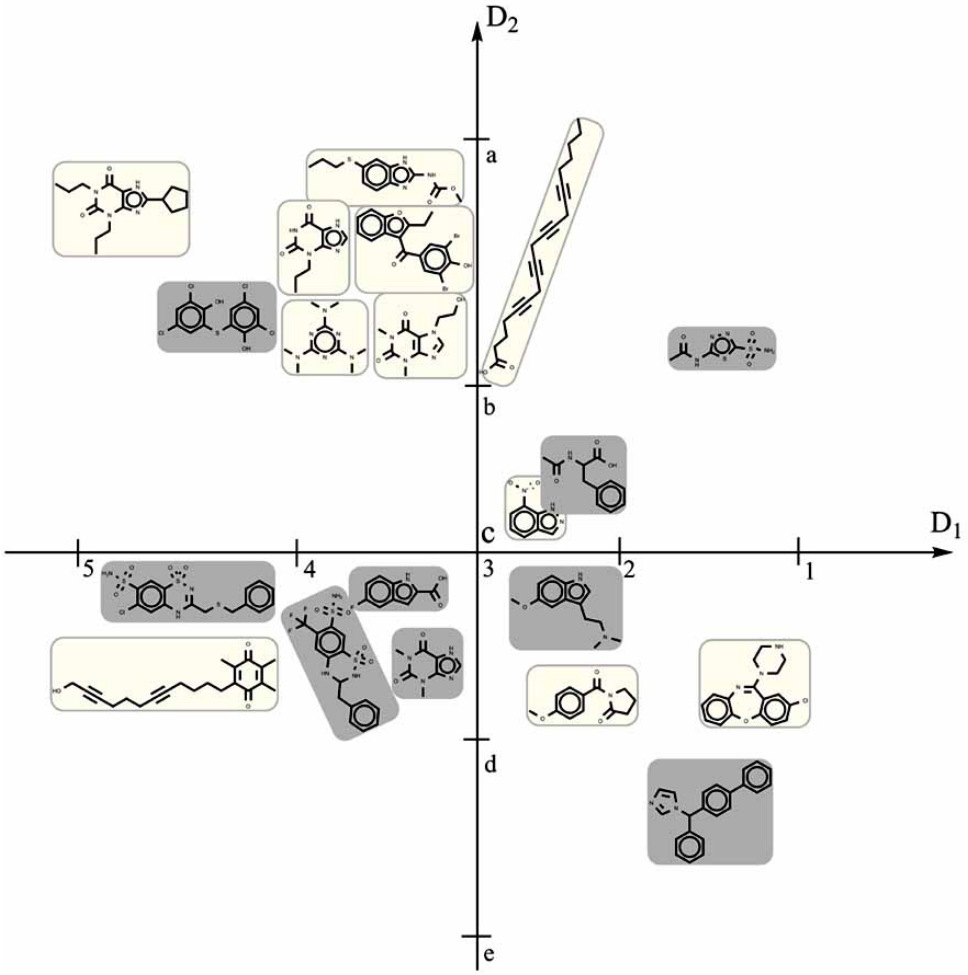

Fig. (1). Descriptor space. For demonstration purposes, a simple chemical space is shown that is formed by only two descriptors, D1 and D2. Descriptor value ranges are divided into intervals. A total of 20 compounds belonging to two different classes (light and dark gray, respectively) are projected into this space. In Figs. (**1**-**3**), compound structures are only used for illustrative purposes and their specific identities are not relevant for further discussion.

molecules belonging to two classes are projected into this space. A region of space where compounds from one of the classes dominate is outlined by combinations of descriptor values typical for this class. The discriminatory power of such combinations can be related to the "purity" of the sub-region: the more one class of compounds dominates, the more typical a combination of features becomes for the dominating class. The sharpest distinction can be made if one class is completely absent in a sub-space, e.g. in the sub-space defined by the intervals [3,4] from descriptor D_1 and [a,b] from descriptor D_2. In the following, such combinations will be written as {D_1:[3,4], D_2:[a,b]}. This region contains five compounds from one class, which is nearly half of the population of this class, but no compounds from the other class. Thus, in order to be as discriminatory as possible, one searches for regions in descriptor space that cover as many compounds as possible from one class, but the least number of compounds from the other class (or classes). However, in high-dimensional spaces with multi-valued descriptors, the number of descriptor value combinations grows exponentially. Therefore, we need data mining algorithms that are capable of extracting most discriminatory descriptor combinations from high-dimensional space representations.

2.2. Emerging Patterns

For the following descriptions, we consider molecules organized in a table with all descriptors, one per column. A molecule can then be associated with descriptor value combinations $\{D_1 : V_1, \ldots, D_n : V_n\}$. An arbitrary set of descriptor value pairs is called a "pattern" and the number of compounds sharing a pattern p is defined as the "support" $\sup(p)$ of a pattern p. For two classes, the fraction of their supports of a pattern p in each class is called its "growth rate" *growth (p)*. Using this terminology, we can now formally describe discriminatory regions in descriptor space as patterns having high support in one class and low support in the other, which corresponds to patterns having a high growth rate. In data mining, such patterns are called "emerging patterns" [7]. In order to reduce the number of emerging patterns to be considered, we focus on a sub-set of emerging patterns where the support in one class is zero. These patterns are called "jumping emerging patterns" (JEP) and have highest discriminatory power [13].

As already mentioned, the number of possible patterns can become exceedingly large for high-dimensional spaces formed by descriptors with continuous value ranges. A possibility to reduce the number of patterns is to concentrate on patterns that are most general, i.e. patterns representing a large number of compounds. Following our terminology, these would be emerging patterns having highest support, as illustrated in (Fig. **2**). Both patterns $\{D_1 : [2,3], D_2 : [a,b]\}$ and $\{D_1 : [3,4], D_2 : [a,b]\}$ are emerging patterns capturing one and five compounds from one of the two classes, respectively. However, seven of eight compounds are also captured by the smaller pattern $\{D_2 : [a,b]\}$, which has higher support than both individual patterns and is thus more general and preferred. In order to further reduce the number of emerging patterns, we select only patterns containing a minimal number of items. In (Fig. **2**), we can find two JEP capturing the single compound in the lower left: $\{D_1 : [4,5]\}$ and $\{D_1 : [4,5], D_2 : [b,c]\}$. Since both patterns have the same support of 1/3, we cannot use support here as a measure of generality. However, the second pattern contains more items and places more constraints on matching compounds, making it more special and less general than the first one. In other words, every compound that fits the second pattern also fits the first one, which therefore represents a more general pattern. Considering such differences in generality, we can further reduce the number of JEP for analysis by select-

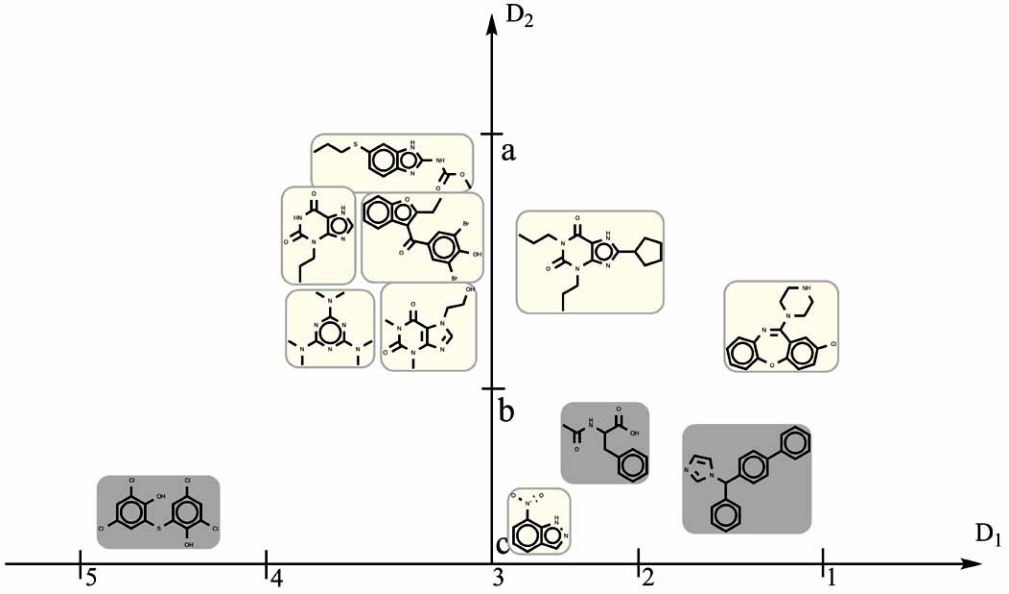

Fig. (2). Illustration of most expressive jumping emerging patterns. Shown are a total of 11 compounds from two classes distributed in a chemical space formed by two descriptors D_1 and D_2. Patterns are described in the text.

ing JEP for which (1) no superset pattern with more descriptor value pairs has higher support and (2) no subset is also a JEP. Patterns meeting those two criteria are called "most-expressive JEP" [13]. For the analysis of molecular patterns, we have defined ECPs as the most expressive JEP from descriptor-dependent feature analysis [8]. Such patterns can be mined in molecular datasets using different types of methods including a hypergraph-based algorithm used in our implementation [14].

2.3. Discretization of Descriptor Values

The generation of ECPs and other emerging patterns requires the availability of finite and discrete descriptor values and intervals. Most descriptors can adopt many different values or continuous value ranges and, therefore, they must be subjected to discretization procedures to divide their value ranges into suitable intervals. Methods of rather different complexity are available to transform multi-valued or continuous descriptors into discrete domains that take training set distributions into account [8]. The same discretization procedure must be applied to training and test compounds. For the studies reported herein, we introduce a straightforward discretization approach based on statistical properties of the training set compounds. For a training set with two classes, we compute the mean μ and standard deviation σ of a descriptor in each class and use the three ranges $(\mu-\sigma,\mu+\sigma]$, $(-\infty,\mu-\sigma]$ and $(\mu+\sigma,+\infty)$ as discrete intervals. We eliminate descriptors which show no variation or for which the coefficient of variation $c_v=\frac{\sigma}{\mu}>1$. Descriptors with no variance do not have any generalization ability and are thus not useful for classification. The removal of descriptors with large variation is rationalized by the fact that descriptor value distributions are class-selective with the majority of descriptor values found in small class-specific value ranges [15]. These value ranges are particularly attractive for the generation of patterns. On the other hand, descriptors having a coefficient of variation greater than one are considered high variance descriptors because their standard deviation is larger than the mean. Therefore, it is likely that many irrelevant database compounds populate these ranges, which severely restricts the predictive value of such descriptors. Consequently, they are also omitted.

2.4. Classification and Virtual Screening

For JEP, different classification schemes have been introduced in computer science [13,16]. Our procedure for molecular classification using ECP is outlined in Fig. (**3**). A set of ECP is computed from the training set and used to classify test compounds. When classifying a compound, we accumulate the support of all ECP from each class that are found in the test compound. The molecule is then predicted to belong to the class producing highest accumulated support. In our initial study, it became apparent that very small

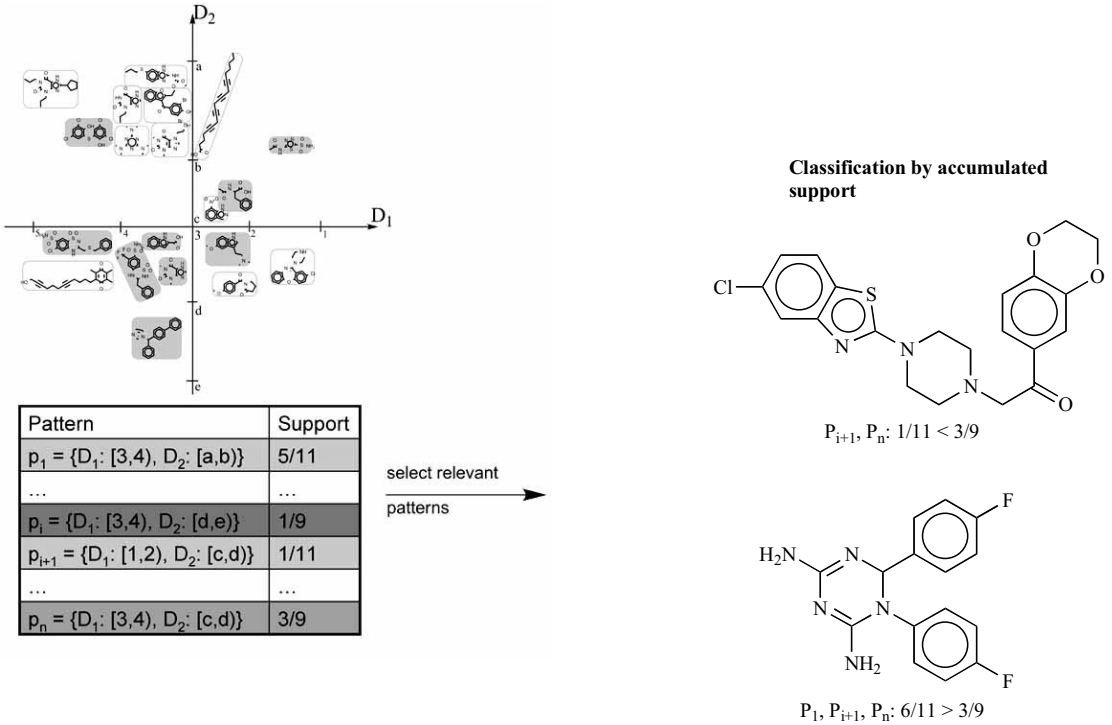

Fig. (3). Summary of ECP classification. Test compound are classified on the basis of highest accumulated class support, as discussed in the text.

training sets containing only three, five, or 10 compounds per class were sufficient to achieve ECP prediction accuracies on different compound classes of, on average, 70%-80% [8].

In this study, we go beyond compound classification and apply ECP to virtual screening. This application presents additional challenges because small numbers of active molecules must be distinguished from very large numbers of irrelevant database compounds. Therefore, we introduce an ECP-based compound ranking by calculating for each test compound the difference of accumulated support from ECP learning sets of active and inactive molecules and using it as a score.

3. HTS DATA AND SCREENING CALCULATIONS

For simulation of sequential screening trials, we used an HTS data set consisting of 50,000 test compounds produced in the search for novel inhibitors against the well-known enzyme and drug target dihydrofolate reductase (DHFR) [17]. Among the 50,000 compounds there were a total of 32 confirmed competitive DHFR inhibitors with K_i values ranging from 26 nM to 11 μM [17,18]. This data set was made publicly available in the context of the first McMaster University data mining and docking competition as a training data set [19]. Fig (**4**) shows the 12 most active inhibitors that were identified in this screen. As can be seen, these compounds had diverse structures. Compared to typical benchmark settings using constructed databases as decoys and optimized literature compounds as potential hits, the analysis of HTS data has two intrinsic advantages, although these data are generally prone to noise and experimental errors [11]: first, the HTS data sets contain confirmed inactive compounds and, second, active compounds are typical screening hits, rather than leads added to constructed databases. Both aspects make simulations on HTS data more similar to practical screening applications involving VS calculations.

We simulated sequential screening experiments by performing the following steps. Initially, ECP training was carried out on sets of five randomly selected DHFR inhibitors and 20 inactive compounds taken from the HTS data. Compounds were randomly selected computationally using a random number generator. For each individual calculation, as described below, a training set was randomly selected. For discretization and feature selection, a previously reported set of 61 1D/2D descriptors with minimized descriptor correlation was used as descriptor pool [20]. The set exclusively consisted of descriptors with pair-wise correlation coefficients of 0.8 or smaller and contained four topological indices, logP(o/w), 16 atom or bond counts, three partial charge descriptors, and 37 complex molecules surface property descriptors approximated from 2D molecular representations. Thus, the selected descriptor set included 2D and implicit 3D molecular descriptors, but no structural or fragment-type descriptors.

The resulting ECP classifiers were then applied to rank all remaining compounds in the HTS data set. Following the sequential screening paradigm, we then selected the top-scoring 10, 100, or 500 compounds from the HTS set and examined this selection set for new hits, thereby mimicking experimental evaluation of the top-ranked compounds. From each selection set, the top-ranked 10 compounds plus all remaining hits (for selection sets of 100 and 500 molecules) were then added to the training set in order to re-build and refine the classifier for the next iteration. For each selection set size (10, 100, or 500 molecules), 100 individual trials using different training sets were carried out in order to produce a statistically relevant sample. In each case, a total of nine sequential screening iterations were carried out such that the maximum number of "tested" compounds was smaller than 10% of the entire HTS data set for the largest selection set of 500 compounds.

4. RESULTS AND DISCUSSION

Only five active molecules were used to derive ECP classifiers in the presence of varying numbers of inactive database compounds. These were fewer active compounds than typically required for classification methods such as, for example, decisions trees or Bayesian models, which we previously compared to ECP [8]. Since we randomly selected for each calculation five active database compounds for ECP training, the DHFR HTS set contained only a total of 27 hits. Thus, it provided an equally interesting and challenging scenario for investigating whether ECP calculations were sufficiently sensitive to select very small numbers of active molecules from large numbers of inactive database compounds. For each selection set size, we carried out 100 calculations with different training sets in order to (a) determine the top performance and potential of the ECP methodology and (b) estimate the expected performance level in iterative screening independent of the composition of learning sets. This was done because compound classification calculations are generally much influenced by compound-class specific features and training set composition [10,11]. For ECP, this analysis was particularly relevant since we only used five active compounds for training, which put high weight on the characteristics and contributions of each individual molecule.

4.1. Pattern Distribution and Composition

First we analyzed the patterns that were derived from the learning sets as a basis for classification. On average, 29 to 30 (of 61 available) descriptors were utilized in each ECP calculations, thus only about half the descriptor basis set. Therefore, large numbers of descriptors were not required for pattern derivation. For active learning set compounds, on average ~10,700 patterns consisting of 7.5 descriptor value pairs were produced. By contrast, for inactive compounds on average only ~170 patterns emerged with 3.3 descriptor value pairs per pattern. Thus, active compounds generated significantly more and larger patterns than inactive ones. These findings can be rationalized by considering that during learning each active compound must be distinguished from all inactive molecules and vice versa. Since there were considerably more inactive than active compounds in each training set, active molecules required more descriptors to distinguish themselves from inactive compounds. Therefore, the large difference in the number of patterns between active and inactive molecules is due to the fact that the number of potential patterns grows exponentially in descriptor spaces of increasing dimensionality. It follows that the deliberately unbalanced composition of our learning sets was reflected in

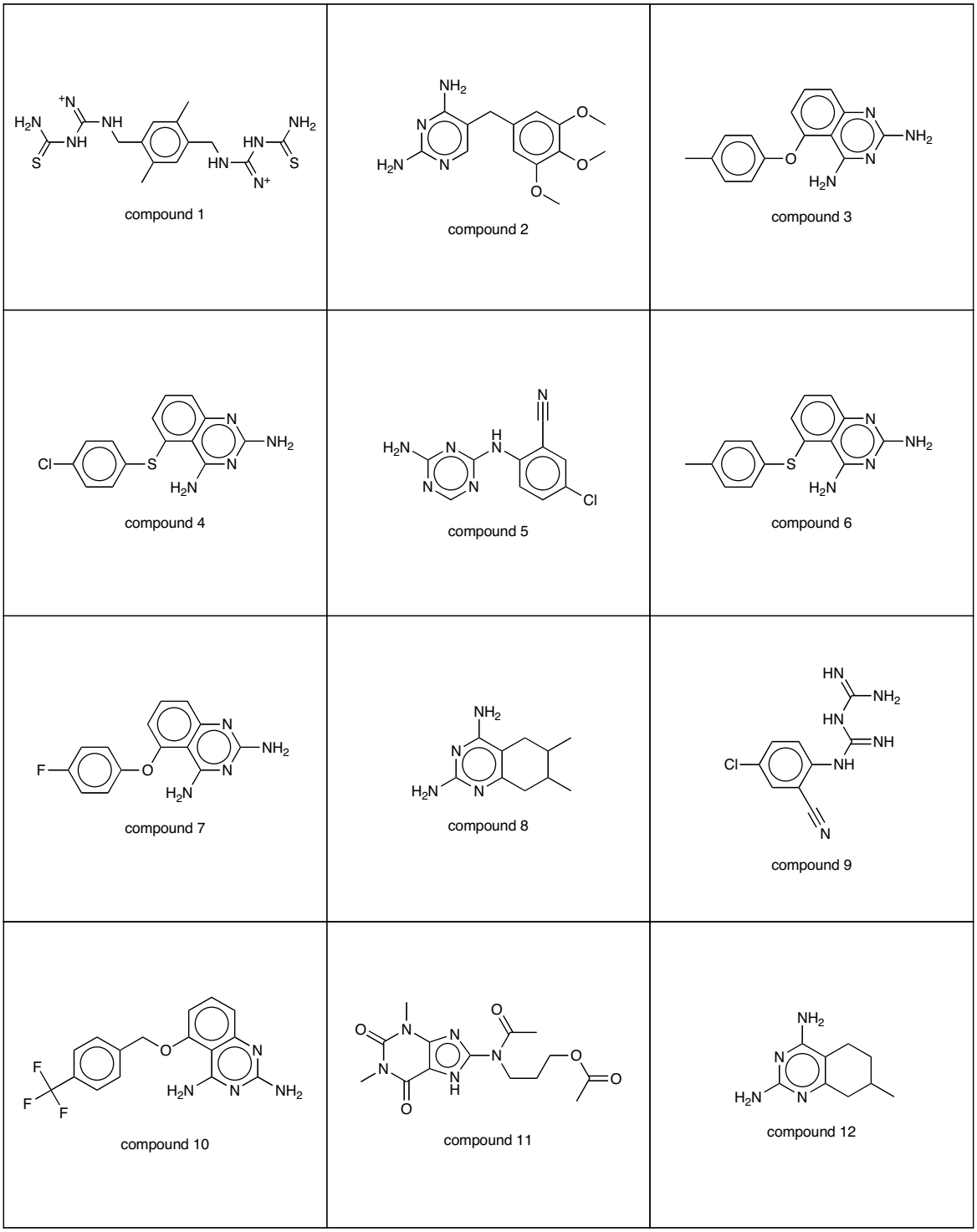

Fig. (4). Structures of the 12 most active DHFR inhibitors. The comparison reveals the presence of different scaffold types among the HTS hits.

large differences in the numbers of "active" and "inactive" patterns, consistent with our expectations. For iterative screening applications, learning sets should contain more inactive than active compounds because many more inactive molecules are available.

4.2. Individual ECP Trials

Next we analyzed the results of ECP calculations using compound selection sets of different size. Selection set size influenced the calculations because different numbers of active molecules were added to the training sets during each

of the nine iterations. Selection sets of increasing size, from 10 to 100 and 500 compounds, typically contained an increasing number of hits that were added to the training set for the next iteration. The more active compounds are available, the better the basis for training becomes, but the trade-off for improved training is the larger number of databases compounds that need to be "tested". Tables (**1**, **2**, and **3**) report the top ten ECP trials for selection sets of 10, 100, and 500 compounds, respectively. The results reflect these trends. For selection sets of only 10 database compounds, the best individual trials recovered 33% of the hits (Table **1**), which corresponded to nine of 27 available active compounds. The top ten trials showed the same retrieval characteristics because they identified the same sets of active compounds, despite different learning set composition (whereas other trials within the top 20 list produced different sets). During iterative screening, the number of cumulatively identified hits increased from three in the first trial to nine after the last.

Table 1. Top 10 ECP Trials for Selection Sets of 10 Database Compounds

Iteration		Trial									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	TR	35	35	35	35	35	35	35	35	35	35
	ACT	3	3	3	3	3	3	3	3	3	3
	RR	11.1%	11.1%	11.1%	11.1%	11.1%	11.1%	11.1%	11.1%	11.1%	11.1%
2	TR	45	45	45	45	45	45	45	45	45	45
	ACT	5	5	5	5	5	5	5	5	5	5
	RR	18.5%	18.5%	18.5%	18.5%	18.5%	18.5%	18.5%	18.5%	18.5%	18.5%
3	TR	55	55	55	55	55	55	55	55	55	55
	ACT	5	5	5	5	5	5	5	5	5	5
	RR	18.5%	18.5%	18.5%	18.5%	18.5%	18.5%	18.5%	18.5%	18.5%	18.5%
4	TR	65	65	65	65	65	65	65	65	65	65
	ACT	7	7	7	7	7	7	7	7	7	7
	RR	25.9%	25.9%	25.9%	25.9%	25.9%	25.9%	25.9%	25.9%	25.9%	25.9%
5	TR	75	75	75	75	75	75	75	75	75	75
	ACT	7	7	7	7	7	7	7	7	7	7
	RR	25.9%	25.9%	25.9%	25.9%	25.9%	25.9%	25.9%	25.9%	25.9%	25.9%
6	TR	85	85	85	85	85	85	85	85	85	85
	ACT	7	7	7	7	7	7	7	7	7	7
	RR	25.9%	25.9%	25.9%	25.9%	25.9%	25.9%	25.9%	25.9%	25.9%	25.9%
7	TR	95	95	95	95	95	95	95	95	95	95
	ACT	8	8	8	8	8	8	8	8	8	8
	RR	29.6%	29.6%	29.6%	29.6%	29.6%	29.6%	29.6%	29.6%	29.6%	29.6%
8	TR	105	105	105	105	105	105	105	105	105	105
	ACT	9	9	9	9	9	9	9	9	9	9
	RR	33.3%	33.3%	33.3%	33.3%	33.3%	33.3%	33.3%	33.3%	33.3%	33.3%
9	TR	115	115	115	115	115	115	115	115	115	115
	ACT	9	9	9	9	9	9	9	9	9	9
	RR	33.3%	33.3%	33.3%	33.3%	33.3%	33.3%	33.3%	33.3%	33.3%	33.3%

"TR" reports the size of the training sets, "ACT" the total number of active compounds retrieved (with newly identified ones added to the training set for the next round), and "RR" the cumulative recovery rates of active compounds. The same abbreviations are used in Tables 2-4.

Table 2. Top 10 ECP Trials for Selection Sets of 100 Database Compounds

Iteration		Trial									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	TR	36	36	35	35	35	35	35	36	36	38
	ACT	4	4	0	0	0	0	0	3	3	4
	RR	14.8%	14.8%	0.0%	0.0%	0.0%	0.0%	0.0%	11.1%	11.1%	22.2%
2	TR	49	48	45	45	45	45	45	47	46	49
	ACT	8	7	0	0	0	0	0	10	4	7
	RR	29.6%	25.9%	0.0%	0.0%	0.0%	0.0%	0.0%	37.0%	14.8%	25.9%
3	TR	62	58	56	56	56	56	56	57	56	59
	ACT	13	7	1	1	1	1	1	10	4	8
	RR	48.1%	25.9%	3.7%	3.7%	3.7%	3.7%	3.7%	37.0%	14.8%	29.6%
4	TR	72	70	68	68	68	68	68	68	66	69
	ACT	13	11	4	4	4	4	4	11	4	8
	RR	48.1%	40.7%	14.8%	14.8%	14.8%	14.8%	14.8%	40.7%	14.8%	29.6%
5	TR	83	80	79	79	79	79	79	78	77	79
	ACT	14	11	5	5	5	5	5	11	5	8
	RR	51.9%	40.7%	18.5%	18.5%	18.5%	18.5%	18.5%	40.7%	18.5%	29.6%
6	TR	93	90	91	91	91	91	91	88	88	90
	ACT	14	11	8	8	8	8	8	11	7	9
	RR	51.9%	40.7%	29.6%	29.6%	29.6%	29.6%	29.6%	40.7%	25.9%	33.3%
7	TR	103	101	106	102	102	102	102	98	99	101
	ACT	14	12	9	9	9	9	9	11	8	10
	RR	51.9%	44.4%	33.3%	33.3%	33.3%	33.3%	33.3%	40.7%	29.6%	37.0%
8	TR	113	111	117	113	113	113	113	108	111	112
	ACT	14	12	10	10	10	10	10	11	11	11
	RR	51.9%	44.4%	37.0%	37.0%	37.0%	37.3%	37.0%	40.7%	40.7%	40.7%
9	TR	123	121	128	124	124	124	124	118	121	122
	ACT	14	12	11	11	11	11	11	11	11	11
	RR	51.9%	44.4%	40.7%	40.7%	40.7%	40.7%	40.7%	40.7%	40.7%	40.7%

For selection sets of 100 compounds (Table **2**), the best cumulative recovery rate of individual trials was 52% corresponding to 14 of 27 hits. Here retrieval characteristics substantially differed and in some instances, individual trials produced good results, although active compounds could not be recovered during the first one or two iterations. In these cases, the training sets were expanded by addition of 10 inactive compounds per iteration, which then led to the identification of hits. This was due to the generation of patterns that became increasingly characteristic for active compounds because more inactive molecules needed to be discriminated. These findings further illustrated the usefulness of unbalanced training sets if only a few hits were available for learning and because inactive compounds contain information for ECP classification. Best recovery rates were obtained for selection sets of 500 compounds, as expected (Table **3**). Here the top ten trials retrieved between 15 and 19 hits compounds, producing a top recovery rate of 70%. Thus, at top performance levels, iterative ECP calculations were capable of producing significant retrieval rates for selection sets of varying size.

Table 3. Top 10 ECP Trials for Selection Sets of 500 Database Compounds

Iteration		Trial									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	TR	39	38	39	37	38	41	43	41	36	41
	ACT	5	6	6	3	5	8	8	6	7	9
	RR	18.5%	22.2%	22.2%	11.1%	18.5%	29.6%	29.6%	22.2%	25.9%	33.3%
2	TR	52	48	51	51	51	53	55	55	47	53
	ACT	9	6	9	10	8	10	10	11	9	11
	RR	33.3%	22.2%	33.3%	37.0%	29.6%	37.0%	37.0%	40.7%	33.3%	40.7%
3	TR	62	58	62	64	64	64	68	65	57	64
	ACT	9	6	10	13	11	14	13	11	9	12
	RR	33.3%	22.2%	37.0%	48.1%	40.7%	51.9%	48.1%	40.7%	33.3%	44.4%
4	TR	75	69	72	74	78	75	78	76	68	74
	ACT	12	7	10	13	15	15	13	12	10	13
	RR	44.4%	25.9%	37.0%	48.1%	55.6%	55.6%	48.1%	44.4%	37.0%	48.1%
5	TR	90	80	84	85	88	85	90	86	82	84
	ACT	17	8	13	14	15	15	15	12	14	13
	RR	63.0%	29.6%	48.1%	51.9%	55.6%	55.6%	55.6%	44.4%	51.9%	48.1%
6	TR	101	93	96	96	99	95	100	98	93	94
	ACT	18	11	15	15	16	15	15	14	15	13
	RR	66.7%	40.7%	55.6%	55.6%	59.3%	55.6%	55.6%	51.9%	55.6%	48.1%
7	TR	111	105	106	107	109	106	110	109	103	105
	ACT	18	13	15	16	16	16	15	15	15	14
	RR	66.7%	48.1%	55.6%	59.3%	59.3%	59.3%	55.6%	55.6%	55.6%	51.9%
8	TR	121	117	117	117	119	116	120	119	113	115
	ACT	18	16	16	16	16	16	15	15	15	14
	RR	66.7%	59.3%	59.3%	59.3%	59.3%	59.3%	55.6%	55.6%	55.6%	51.9%
9	TR	132	128	128	127	129	126	130	129	123	126
	ACT	19	18	17	16	16	16	15	15	15	15
	RR	70.4%	66.7%	63.0%	59.3%	59.3%	59.3%	55.6%	55.6%	55.6%	55.6%

4.3. Overall ECP Performance in Simulated Sequential Screening

Since we carried out 100 ECP trials with randomized learning sets for each selection set size, we could statistically estimate the performance level of iterative screening independent of learning set composition (also taking learning sets into account that failed to recover active compounds or identified only a few). The results are reported in (Table **4**). For selection sets of 10, 100, and 500 database compounds, average recovery rates of 19%, 26%, and 39% were observed, respectively. For the smallest selection set, this corresponded to the identification of approximately five hits when evaluating only 115 database compounds (~0.2% of the HTS set) and for the largest selection set, 10 to 11 hits based on the evaluation a total of 4525 compounds (~9%). The steady increase in average recovery rates over the nine iterations indicated that additional hits could be retrieved by adding more screening cycles. An attractive feature of iterative ECP calculations was the enrichment of hits among selection sets of only 10 compounds, where evaluation of 115 database compounds was sufficient to produce on average five hits, given initial learning sets containing only five active com-

Table 4. Average ECP Results Over 100 Independent Screening Trials

Iteration	10				100				500			
	tested	TR	ACT	RR	tested	TR	ACT	RR	tested	TR	ACT	RR
1	35	35	1.32	4.9%	125	36.00	2.41	8.9%	525	37.40	3.94	14.6%
2	45	45	1.78	6.6%	225	46.84	3.78	14.0%	1025	49.09	6.29	23.3%
3	55	55	2.73	10.1%	325	57.26	4.69	17.4%	1525	60.01	7.47	27.7%
4	65	65	3.53	13.1%	425	67.74	5.41	20.0%	2025	70.61	8.26	30.6%
5	75	75	4.00	14.8%	525	78.00	5.80	21.3%	2525	81.24	8.94	33.1%
6	85	85	4.50	16.6%	625	89.00	6.40	23.7%	3025	91.80	9.52	35.3%
7	95	95	4.90	18.2%	725	99.00	6.60	24.6%	3525	102.15	9.95	36.9%
8	105	105	5.10	18.9%	825	109.00	6.90	25.6%	4025	112.40	10.20	37.8%
9	115	115	5.20	19.4%	925	119.00	7.10	26.3%	4525	122.71	10.60	39.1%

Averages are reported for all nine iterations and selection sets of 10, 100, and 500 database compounds. The total number of tested database compounds ("tested") is given and "RR" provides averaged cumulative recovery rates.

pounds. On average, selecting 10 times more database compounds gave two additional hits and testing 50 times more compounds doubled the number of recovered hits relative to the smallest selection sets. These results were well in accord with our previous findings that ECP was capable of successfully operating on the basis of very few active compounds and revealed an additional aspect, the presence of high sensitivity and specificity of ECP calculations especially for small compound selection sets. This trend was further supported when we investigated the recovery of the subset of the 12 most potent hits in the DHFR HTS set, as shown in Fig. (**4**). Our learning sets included on average only 1.5 of these active compounds and the selections set of 10, 100, and 500 compounds contained on average about four, five, and six of these hits, respectively. Thus, in small selection sets, the largest relative enrichment of potent hits was observed.

Taken together, our findings indicated that ECP produced recovery rates that were at least comparable to clustering or other classification methods used for sequential screening when approximately 10% of the screening data set was tested. However, ECP calculations already recovered approximately 20% of available hits when only about 100 of 50,000 screening set compounds were evaluated, and these hits were among the most potent ones available in the HTS set. Therefore, the application of ECP is thought to further reduce compound selection set sizes in iterative screening trials, which adds to the sequential screening paradigm.

5. PERSPECTIVE AND CONCLUSIONS

This study introduces ECP as a new methodology in simulated sequential screening and shows the potential of modern data mining techniques in pharmaceutical research. Although this screening paradigm is currently far from being established in pharmaceutical research, it is increasingly considered as a complement or an alternative to brute force HTS [10-12]. In simulated sequential screening trials, ECP calculations generated a steady increase in the recovery of active compounds and already produced multiple hits by iteratively selecting as little as 0.2 % of the HTS data set. Clearly, with 50,000 compounds the DHFR test set studied here was smaller than many currently used HTS compound sets that are frequently an order of magnitude larger in size. However, the size of screening sets is not a limiting factor for ECP analysis and given the observed sensitivity and specificity of the calculations, there are no reasons to expect that substantially different results would be obtained for differently sized screening sets (but results differ for differently sized training sets). ECP calculations are particularly attractive for sequential screening applications when complete recovery of available active compounds is not the primary goal of the screening efforts, but rather rapid recovery of novel hits. Our results further support the view that iterative computational and experimental screening can streamline biological screening efforts and greatly reduce the experimental and data analysis requirements, including secondary assays to eliminate false-positives. If a practical ECP-supported sequential screening application on the DHFR set would have produced results similar to our simulations, it would have been possible to replace HTS analysis of this data set with series of low-throughput assays to identify multiple hits. On the basis of our findings, we conclude that ECP analysis should merit further consideration in HTS data mining and sequential screening.

REFERENCES

[1] Tamura, S.Y.; Bacha, P.A.; Gruver, H.S.; Nutt, R.F. *J. Med. Chem.,* **2002**, *45*, 3082.
[2] Keseru, G.M.; Molnar, L.; Greiner, I. *Comb. Chem. High Throughput Screen.,* **2000**, *3*, 535.
[3] Stockfisch, T.P. *J. Chem. Inf. Comput. Sci.,* **2003**, *43*, 1608.
[4] Harper, G.; Bradshaw, J.; Gittin, J.C.; Green, D.V.S.; Leach, A.R. *J. Chem. Inf. Comput. Sci.,* **2001**, *41*, 1295.
[5] Jorissen, R.N.; Gilson, M.K. *J. Chem. Inf. Model.,* **2005**, *44*, 549.
[6] Labute, P. *Pac. Symp. Biocomput.,* **1999**, *4*, 444.
[7] Dong, G.; Li, J. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*,

Chaudhuri, S.; Fayyad, U.; Madigan, D., Eds; ACM Press: New York, **1999**; pp. 43.
[8] Auer, J.; Bajorath, J. *J. Chem. Inf. Model,* **2006**, *46*, 2502.
[9] Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*, Wiley-VCH: Weinheim, **2000**.
[10] Bajorath, J. *Nat. Rev. Drug Discov.,* **2002**, *1*, 882.
[11] Parker, C.N.; Bajorath, J. *QSAR Comb. Sci.,* **2006***, 25,* 1153.
[12] Engels, M.F.M.; Venkatarangan, P. *Curr. Opin. Drug Discov. Devel.,* **2001***, 4*, 275.
[13] Li, J.; Dong, G.; Ramamohanarao, K. *Knowl. Inf. Syst.,* **2001**, *3,* 131.
[14] Bailey, J.; Manoukian, T.; Ramamohanarao, K. In *Proceedings of the 3rd IEEE International Conference on Data Mining*, IEEE Computer Society: Los Alamitos, **2003**; pp. 485.
[15] Eckert, H.; Bajorath, J. *J. Med. Chem.,* **2006**, *49,* 2284.
[16] Li, J.; Dong, G.; Ramamohanarao, K.; Wong, L. *Machine Learn.,* **2004**, *54*, 99.
[17] Zolli-Juran, M.; Cechetto, J.D.; Hartlen, R.; Daigle, D.M.; Brown, E. D. *Bioorg. Med. Chem. Lett.,* **2003**, *13*, 2493.
[18] Elowe, N.H.; Blanchard, J.E.; Cechetto, J.D.; Brown, E.D. *J. Biomol Screen.*, **2005**, *10*, 653.
[19] Parker, C.N. *J. Biomol Screen.*, **2005**, *10*, 647.
[20] Xue, L.; Godden, J. W.; Stahura, F.L.; Bajorath, J. *J. Chem. Inf. Comput. Sci.*, **2003**, *43*, 1151.

Received: 09 May, 2007 Revised: 15 November, 2007 Accepted: 19 November, 2007