ARTICLES

Evolving Interpretable Structure−Activity Relationships. 1. Reduced Graph Queries

Kristian Birchall,[†] Valerie J. Gillet,*[,†] Gavin Harper,[‡] and Stephen D. Pickett[‡]

Department of Information Studies, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield S1 4DP, United Kingdom, and GlaxoSmithKline, Medicines Research Centre, Gunnels Wood Road, Stevenage SG1 2NY, United Kingdom

Received February 13, 2008

A new machine learning method is presented for extracting interpretable structure−activity relationships from screening data. The method is based on an evolutionary algorithm and reduced graphs and aims to evolve a reduced graph query (subgraph) that is present within the active compounds and absent from the inactives. The reduced graph representation enables heterogeneous compounds, such as those found in high-throughput screening data, to be captured in a single representation with the resulting query encoding structure−activity information in a form that is readily interpretable by a chemist. The application of the method is illustrated using data sets extracted from the well-known MDDR data set and GSK in-house screening data. Queries are evolved that are consistent with the known SARs, and they are also shown to be robust when applied to independent sets that were not used in training.

INTRODUCTION

Machine learning methods have become popular tools for analyzing high-throughput screening (HTS) data sets.[1] The aim is to derive a model of activity using a training set of compounds with known activities which can then be used to predict the activities of previously unseen compounds. The compounds in the training set are usually assigned binary values, for example, active and inactive, whereas the predictions can either be binary (using a classification method) or quantitative (using a ranking method) where a score is assigned to each compound that reflects its probability of exhibiting activity. A variety of techniques have been applied, including support vector machines (SVMs), neural networks (NNs), substructural analysis (closely related to Naïve Bayesian classifiers), and decision trees.[2–7] While impressive prediction rates have been reported, these methods are limited in the structure−activity relationship (SAR) information that can be extracted from them. For example, SVMs and NNs are often referred to as "black boxes" so that it is generally not possible to interpret the models produced in terms of structural features that are responsible for activity. Substructural analysis does enable some structure−activity information to be extracted, for example, fragments that occur preferentially in active compounds are assigned high weights; however, the fragments are treated as being independent of one another and their co-occurrence is not taken into account. Decision trees enable rules to be accumulated through the hierarchy of nodes that lead to a leaf node; however, their predictive ability is generally inferior to other methods due to the order dependence of the splits and their tendency to overtrain. Their predictive performance can be improved by growing multiple trees using bootstrapping techniques, in what are known as random forests, and using a voting system.[8] However, the gain in accuracy comes at the cost of interpretability, while it is straightforward to extract rules from a single classification tree this is inherently more difficult when considering ensembles of trees.[9]

We have previously described the use of reduced graphs (RGs) for the representation and similarity searching of small molecules,[10–13] and closely related approaches have also been developed by other groups.[14–16] In an RG, groups of atoms are replaced by a single node with the topology of the original structure retained by connecting the nodes in the RGs appropriately. There are many ways in which chemical graphs can be reduced, and for similarity searching the aim is usually to capture the functionality of groups of atoms so that compounds with similar bioactivity are identified as similar, irrespective of their exact chemical scaffolds. Thus, graph reduction has focused on structural features that are thought to be relevant to drug-receptor binding such as hydrogen bond donors, hydrogen bond acceptors, aromatic rings, etc. This more abstract representation has been shown to offer advantages over more traditional 2D descriptors in scaffold hopping applications.[11,16]

Similarity searching has the advantage that it can be carried out using a single active compound; however, in this basic implementation it does not allow the user to take account of prior knowledge such as information relating to the SAR. Thus, Harper et al.[13] introduced the concept of RG queries based on SMARTS strings for more controlled searches. (It had been shown previously that RGs can be written as SMILES strings in which atoms not in the usual organic set, such as the transition metals, are used to represent the

* Corresponding author phone: +44-1142-222652; fax: +44-1142-780300; e-mail:v.gillet@sheffield.ac.uk.
[†] University of Sheffield.
[‡] GlaxoSmithKline.

10.1021/ci8000502 CCC: $40.75
Published on Web 07/17/2008

different node types). RG SMARTS queries allow the construction of flexible queries in a similar manner to that used for generic substructures, for example, a series of alternative node types can be specified at one position in the query to allow expressions such as "nonfeature ring or acceptor ring" etc. When used in substructure (subgraph) searches, the RG SMARTS queries allow the user to specify some characteristics of the molecules that they wish to be returned.

The RG has also been used to provide a view of HTS data using a method known as data-driven clustering.[13] In this approach, each molecule is represented by a number of *motifs* including the RG, near neighbors of the RG in which single nodes in the RG are deleted or changed, and Bemis and Murcko frameworks.[17] Molecules that share a common motif are clustered together. The clusters are then sorted with large clusters consisting predominantly of active compounds being presented to the user first. The interpretability of the RGs and frameworks allows the structural characteristics of the compounds to be seen immediately. This is in contrast to clustering based on conventional fingerprints.

The data driven clustering method is based on predefined descriptors. Here we present a more flexible approach to analyzing screening data in which an evolutionary algorithm (EA) is used to evolve RG SMARTS queries (subgraphs) that maximize the separation of active and inactive compounds. The queries are searched against compounds in a training set which are also represented by RGs; a compound is predicted active if its RG contains the query, otherwise it is predicted inactive. The EA aims to maximize the separation of actives and inactives in the training set with a test set being used to select the best query and so prevent overtraining. The RG queries enable structurally heterogeneous compounds to be grouped together, which is particularly useful when analyzing HTS data sets where active compounds may belong to different chemical series. The query returned by the EA encodes the structure−activity relationship in terms of topological pharmacophoric features that can be easily understood by a chemist. The query can also be used to select previously untested compounds for screening.

METHODOLOGY

Reduced Graph Definitions. We have used the RG definition described in Harper et al.[13] The full list of node types is shown in Table 1 and includes hydrogen bond donor, hydrogen bond acceptor, positively ionizable group, negatively ionizable group, aliphatic ring, and aromatic ring. The order of precedence for identifying nodes is "positively ionizable" before "negatively ionizable" before "donor" or "acceptor". If a node matches one of the earlier definitions in this list, then it is excluded from matching subsequent ones. Adjacent donor and acceptor features are merged into a distinct joint donor−acceptor feature so that the order of precedence between donor and acceptor need not be defined. The edges in the RG are labeled as single, except for edges that connect fused ring nodes which are labeled as double. The ring nodes and acyclic nodes are further characterized as positively ionizable, negatively ionizable, donor, acceptor, donor and acceptor, or no hydrogen bonding functionality, on the basis of user-defined SMARTS definitions. Terminal acyclic nodes that have no functionality are removed. This definition of the RGs gives rise to 18 node types.

Table 1. RG Node Types Together with the Atom Symbols Used in the SMILES

	Aromatic	Aliphatic	Acyclic
No Feature	Sc	Hf	Zn
Donor	Ti	Ta	Co
Acceptor	V	W	Ni
Donor/Acceptor	Cr	Re	Cu
Positively Ionisable	Mn	Y	Nb
Negatively Ionisable	Fe	Zr	Mo

The RGs are represented as SMILES with uncommon atom types used to identify each of the node types, single edges are represented by (implicit) single bonds, and double edges, which arise from fused rings, are represented by double bonds. The atom symbols used to represent the RG nodes in SMILES are shown in Table 1. Each node type is shown as a colored shape to provide a more graphical representation of the RGs as illustrated in Figure 1.

Encoding RG Queries as SMARTS. A subset of the features of SMARTS[18] has been adapted to represent RG queries and is presented in Table 2. It should be noted that the "=" symbol has a different meaning when applied to RGs compared to normal SMILES and SMARTS−here it is used to represent ring fusion so that the SMILES [Sc]=[Hf] represents a RG consisting of a featureless aromatic ring node fused to a featureless aliphatic ring, whereas [Sc][Hf] implies a single bond (nonfused) connection between the nodes. The logical operators and special atom and bond symbols included in the SMARTS definition allow further flexibility to be encoded within the RG queries. For example, the SMARTS query, [V,W]~[*][!Zn][Nb;D1], represents a RG query consisting of a node that is aromatic acceptor or aliphatic acceptor ([V,W]) connected (by a fused or nonfused connection(~)) to any node ([*]), which is in turn connected to a terminal acyclic positively ionizable node [Nb;D1] via any node type excepting a linker ([!Zn]).

Figure 2 illustrates the flexibility in search that is provided by the SMARTS. A set of hypothetical molecules (actives (A1 to A4) and inactives (I1 to I4)) is shown represented as RGs and SMILES together with several RG queries and their SMARTS descriptions. The molecules retrieved by each of the RG queries are also shown. In the graphical queries the disconnected operator is illustrated by "AND" (Query 6 and Query 7); series of alternative nodes are shown within a rectangle (for example, Query 3); and nodes that are combined with NOT logic are labeled "NOT" (for example, Query 5).

Queries 1 and 2 are simple RG subtrees that do not contain any special SMARTS features. It can be seen that neither has sufficient specificity to be able to capture the variation within the actives while excluding the inactives. Using the SMARTS OR operator increases the matching flexibility, although as

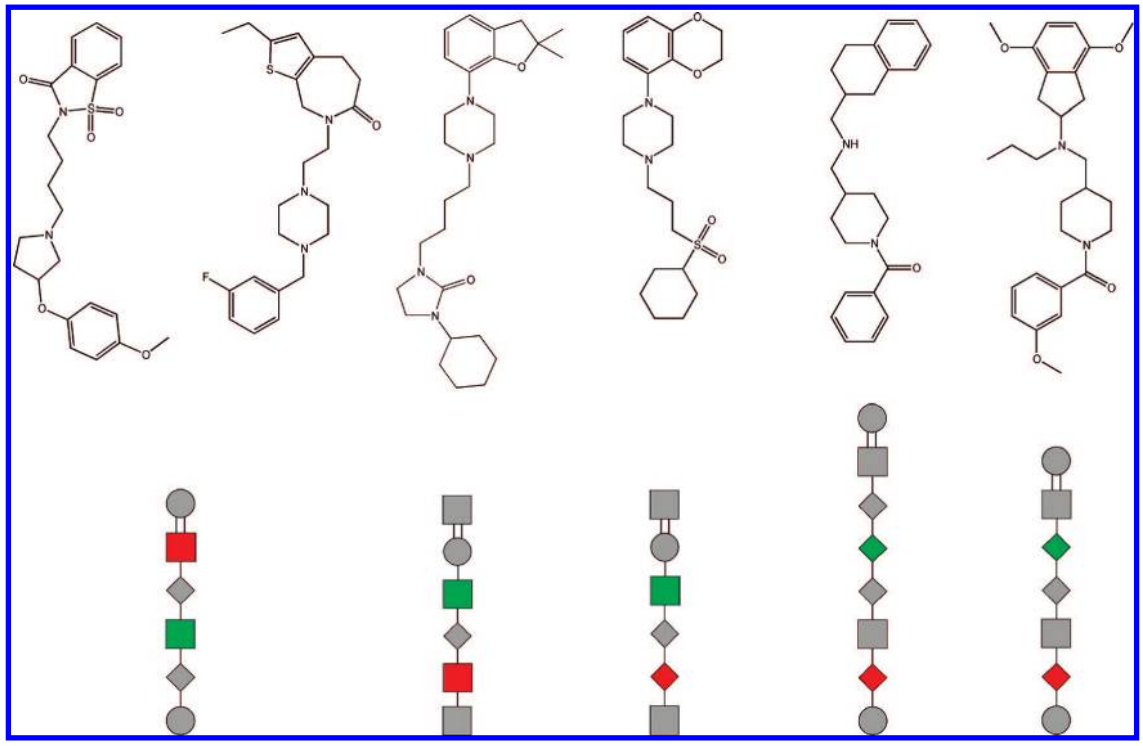

Figure 1. A series of 5HT1A agonists and their RG representations, the key to which is displayed in Table 1. The first two molecules are both represented by the same RG. The RGs of the second pair of molecules differ in the substitution of an aliphatic acceptor node for an acyclic acceptor node. The RGs of the last pair differ by the insertion/deletion of a linker node.

Table 2. SMARTS Features: The Subset of SMARTS Features That Has Been Adapted To Represent RG Queries

symbol	symbol name	atomic property
		Atom Primitives
*	wildcard	any atom
D<n>	degree	<n> explicit connections
		Bond Primitives
-	single bond	nonfused connection between RG nodes
=	double bond	fused connection between rg nodes
~	wildcard	can be a fused or nonfused connection
		Logical Operators
exclamation	!e1	not e1
comma	e1,e2	e1 or e2
		SMILES
.	DOT	disconnected
[and]	square brackets	used to enclose an atom - required with logical operators and primitives

illustrated by Query 3, this can lead to increases in both the number of actives and inactives. Restricting the ring node to be fused in Query 4 increases specificity by returning one fewer inactive compared to Query 3. Query 5 increases specificity further relative to Query 4 by excluding Inactive 3 through the use of NOT logic. The ability to specify disconnected queries allows the matching of RG features that are not necessarily adjacent. Thus Query 6 matches all the actives and also one inactive. Finally, the one remaining inactive can be excluded by restricting the degree of the additional node to terminal (D1) as shown in Query 7.

Overview of the Evolutionary Algorithm. The EA uses a tree-based chromosome representation rather than the linear strings used in a conventional genetic algorithm. The tree representation provides a natural mapping to RG queries (subgraphs) provided that the queries are also limited to trees (i.e., provided that cycles are not permitted in the queries). Cycles are relatively rare in RGs and only arise when a ring atom is common to three rings as in Figure 3. Limiting the RG queries to trees does not allow a complex ring system to be evolved; however, the queries can contain any number of fused rings provided that each fusion represents atoms that are shared by two rings only. Such queries can match more complex RGs, as shown in Figure 3. A further advantage of the tree representation is that the trees can vary in size and shape (both on initialization of the EA and through application of evolutionary operators) and therefore allow variably sized RG queries to be evolved.

The root node of the tree provides access to child nodes in a recursive manner. One possible mapping of an RG query to a tree-based chromosome is shown in Figure 4. The EA begins by generating a population of chromosomes in which trees are grown at random. The fitness of a chromosome is assessed by extracting the RG query from the tree and carrying out searches on the training set where the molecules in the training set are also represented as RGs. This is achieved by first parsing the tree into a SMARTS string (also shown in Figure 4) and then using the Daylight toolkit searching facilities.[18] Any compound whose RG representation contains the query is predicted as active, all other compounds are predicted as inactive, and the fitness of the query is determined by the classification rate; details will be given later. The EA then enters a series of iterations in which queries are chosen for modification, evolutionary operators are applied to produce child queries, and the new children are scored and inserted into the population. Each of these steps is described in more detail below. A number of related approaches involving the evolution of molecules have been described. [19–22]

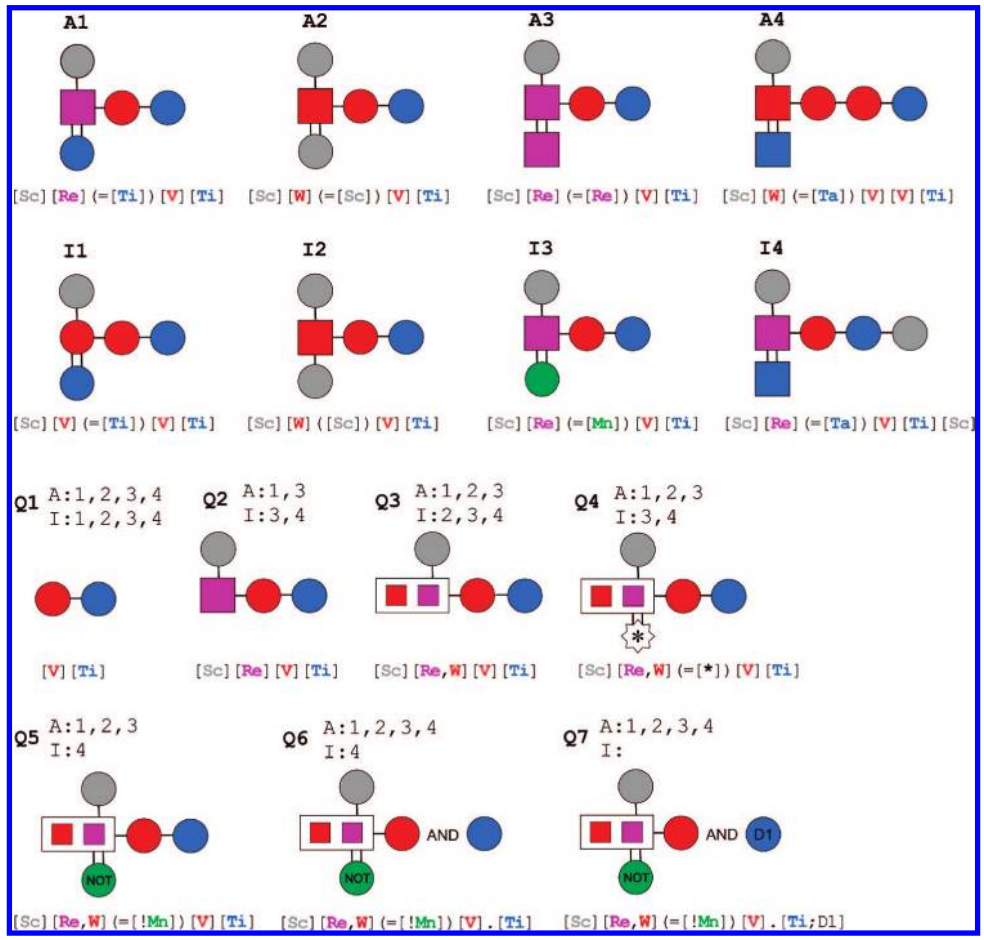

Figure 2. The effect of the SMARTS operators is demonstrated by applying the queries (*Q*) to the set of active (*A*) and inactive (*I*) RGs shown above. The RGs retrieved are listed alongside each query. Representations are given in both graphical and SMARTS format, color coded for ease of interpretation. Alternative nodes are shown enclosed by a rectangle; disconnected nodes are shown by the AND symbol; nodes combined with NOT logic are labeled "NOT"; a wild card node is labeled "*"; and the label D1 indicates a node is terminal.

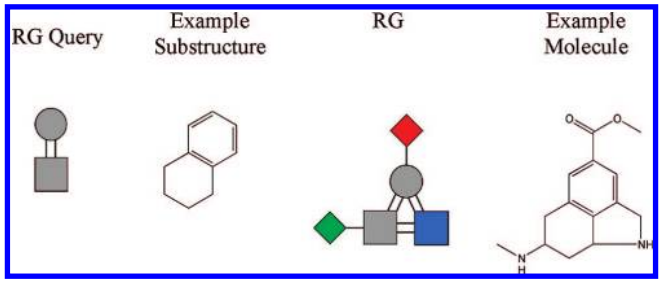

Figure 3. An RG query consisting of two fused rings is shown together with a substructure that it represents and a molecule whose RG contains a cycle that matches the RG query.

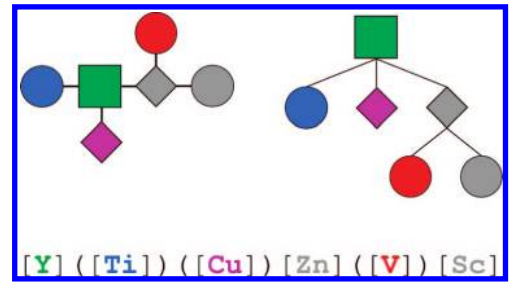

Figure 4. An RG query is shown (left) along with one possible mapping onto a tree (right) and the SMARTS that is parsed from the tree. The different RG node types are color coded to allow the correspondence between the various components to be identified.

Encoding SMARTS Operators in the Chromosomes. The SMARTS operators and atom primitives are encoded in the chromosome by attaching tags to nodes as shown in Figure 5. NOT and OR tags are used to indicate the SMARTS logical operators; the AND tag is used to indicate a disconnection (“.” in SMARTS notation); the D*x* tag (where *x* can take the value 1 or 2) is used to indicate the degree of a node; and the RF tag is used to indicate ring fusion (given as “=” in the SMARTS). The RF tag is applied to a child node only and indicates fusion to the parent node.

The data structure for a chromosome contains a pointer to a root node, from which the rest of the nodes can be accessed. Each node is allocated dynamically allowing the tree to vary in size through the addition and removal of nodes at run time. The data structure for a node contains three pointers to the nodes it is connected to (see Figure 6): the parent pointer denotes the connection to the appropriate node in the upper level of the tree and is null for the root node. The child pointer provides access to the lower levels of the tree and is null for terminal nodes. The sibling pointer provides access within the same level to sibling nodes. Each node data structure encodes information on the node type (a single RG node or a list of alternative RG nodes) and any tags (SMARTS operators) that are applied to that node. The maximum number of children on each node is user-definable.

Parsing a Tree into a SMARTS. The tree-based chromosome is parsed to generate a SMARTS query using a variant of the preorder class of tree reading algorithms[23] in which

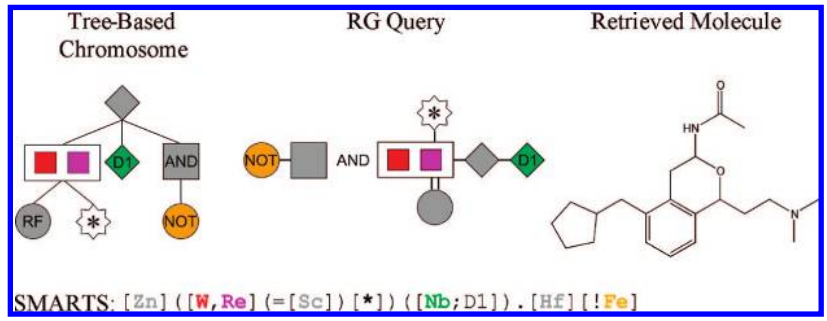

Figure 5. SMARTS operators are tagged to nodes in the tree.

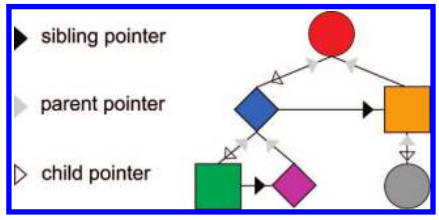

Figure 6. Connections in the tree-based chromosome are indicated by arrowheads, with any other pointers set to null. Any node in the tree can be accessed by following a series of node pointers held within each node. For example, to access the purple node starting at the red node (root) the sequence would be child pointer to blue, child pointer to green, and then sibling pointer to purple. From here, orange could be accessed by following the parent pointer to blue and the sibling pointer to orange.

```
Interpret(node)
{
  if node has sibling write "(" to SMARTS string
  write node type to SMARTS according to tag rules
  if node has child call Interpret(node->child)
  if node has sibling write ")" to SMARTS string
  if node has sibling call Interpret(node->sibling)
  exit iteration of Interpret(node)
}
```

Figure 7. Pseudocode illustrating how a tree-based chromosome is parsed into a SMARTS string.

reading begins at the root and then descends the tree depth−first starting with the left-most child and finishing with the right-most child. First, the root node is written to the SMARTS string, and then the recursive pseudocode shown in Figure 7 is applied, with the following additions to ensure syntactically correct SMARTS are produced. The RG node types are written enclosed in []'s (e.g., "[Zn]"). Sibling nodes indicate branching in the RG query and are written enclosed in ()'s. Alternative nodes that result from the OR tag are written without []'s and separated by commas e.g. "[Zn,Co]"). The AND tag results in a "." outside the []'s (e.g., ".[Zn]"). The RF tag is also written outside the []'s, as a "=" (e.g., "=[Sc]"). The NOT tag results in "!" being placed immediately before the tagged node (e.g., "[!Zn]").

Node-Tag Rules. Some node-tag combinations do not correspond to valid SMARTS, see Table 3, and therefore a set of rules has been devised to eliminate their occurrence both during chromosome initialization and following application of the evolutionary operators. For example, the RF tag is valid only if both the node it is applied to and its parent node are of type ring. If the node it is applied to also contains an OR tag which specifies both acyclic and ring node alternatives, then the RF tag is replaced by the "~" (wildcard) tag in the SMARTS to allow matching of both fused and nonfused connections in the RG SMILES.

Table 3. Prohibited Node-Tag Combinations

Prohibited Node-Tag Combinations
NOT & all node types
NOT & wildcard
AND & root
AND & terminal and wildcard
AND & terminal & NOT
RF & root
RF & AND
RF & no ring on node or parent
OR & wildcard set

Some tree-complexity rules are also implemented to avoid the evolution of SMARTS that, while being syntactically valid, are of limited use. For example, "[*].[*].[*].[*].[*]" is a valid SMARTS; however, it will match any RG with five or more nodes and so is not likely to be discriminating. Furthermore, such a query takes a disproportionately long time to search with. Thus trees with the following characteristics (>10 nodes, >3 AND tags, >4 wildcards, >4 NOT tags) are not interpreted into SMARTS, instead they are assigned a fitness of zero to minimize the chance of their undesirable features propagating through the population.

Population Initialization. The population of chromosomes is initialized by growing each tree to a depth specified by a random number in the range 1 to n where n is the maximum tree depth (the root node is at level 0). While n is user-definable, following consideration of the typical sizes of RGs, the default value was set to two. Growth proceeds by randomly choosing the number of children (0 to 3) for each node at the current level and then iterating on those children. If zero children are added at any level the process returns to the root node and continues with children only being added if the number chosen exceeds the number of children already present. Allocation of the node type is random. Node tag types are also randomly assigned. When an OR tag is applied, the list of nodes is extended randomly between 0 and 2 times, and a tag, selected at random, is added to each node with a probability of 0.5.

Evolutionary Operations. Six evolutionary operators are defined as illustrated in Figure 8: mutate, prune, excise, graft, grow, and crossover. Mutate can be the addition, removal, or substitution of a node tag or node type within an OR list. Prune involves the removal of one or more nodes, by randomly selecting a node and removing it and any attached subtree. Excise involves the removal of a parent node, with any children being transferred to the parents' parent, with the restriction that the number of children on the grandparent does not exceed the maximum permitted. Graft involves the insertion of a subtree between a randomly selected parent and child node pair. The subtree is grown using the process

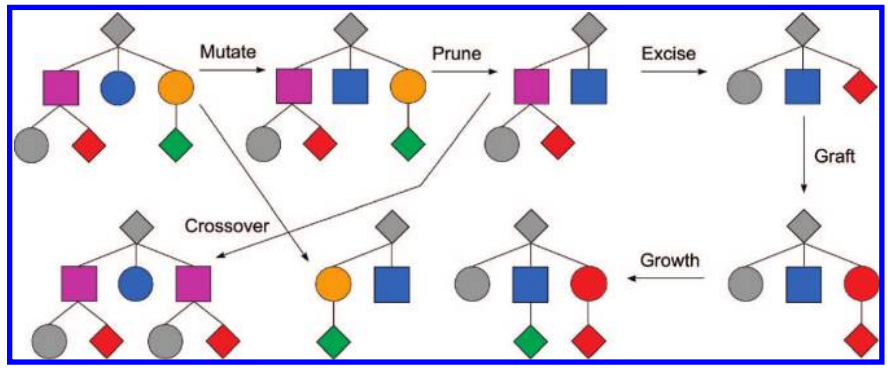

Figure 8. Schematic illustration of the evolutionary operators.

Actual Class	Predicted Class: Active	Predicted Class: Inactive
Active	*TA*	*FI*
Inactive	*FA*	*TI*

$$P = \frac{TA}{TA + FA}; R = \frac{TA}{TA + FI}; F = \frac{2PR}{P + R}; E = \frac{TA}{\left(\frac{TA + FI}{TA + FA + TI + FI}\right) \times (TA + FA)}$$

Figure 9. The contingency, or confusion, matrix shows the possible classification outcomes where *TA* indicates the true actives; *FA* the false actives; *TI* the true inactives; and *FI* the false inactives. Precision (*P*) is the proportion of predicted actives that are true actives. Recall (*R*) is the proportion of true actives predicted as active. The F-measure (*F*) is the average of recall and precision. Enrichment factors (*E*) are also reported in some cases where enrichment is the ratio of the true actives retrieved to the number that would be expected at random.

described for tree initialization. Grow involves the addition of one or more nodes, randomly to either a parent or terminal node. The growth proceeds as for tree initialization and randomly grows to extend the depth by one or two levels. Crossover involves the selection of two parent nodes followed by the exchange of their subtrees. Crossover can be applied within a single tree to produce a single offspring or between two trees to produce two offspring.

The validity of each new tree produced by the operators is checked against the node-tag rules and the tree-complexity rules. If an invalid tree is produced, then a new operator is applied to the original tree. A user-defined maximum number (default value set to 10) of unsuccessful tries are attempted before recording a fail. The probabilities of each of the operators being applied are user-defined.

Fitness Assessment and Parent Selection. The fitness of an individual is assessed by first parsing the tree to produce a SMARTS which is then used to classify compounds in a data set as predicted active (contain the RG query) or predicted inactive (do not contain the RG query). Any binary classifier has two implicit objective functions to optimize: performance on the class of interest (the actives here) and performance on the other class (the inactives). It is not usually possible to optimize both these objectives simultaneously, and the usual approach is to combine the two objectives into a single function which represents some compromise of the two. Here we use the F-measure[24] which is the harmonic mean of the precision (proportion of actives correctly predicted) and recall (proportion of true actives identified), see Figure 9.

Roulette wheel parent selection is implemented, in which the chance of an individual being selected as a parent is directly proportional to its fitness. This is equivalent to allocating each individual a slice of a roulette wheel proportional in size to its fitness; such that the fittest individuals have the largest sections, but even the least fit individuals have some chance of being selected when the wheel is spun. The ratio of the best individual's selection probability to the average selection probability of all individuals in the population is determined by the selection pressure. The number of parents selected is also user-defined as a percentage of the population to be replaced each generation. Following modification of the parents to produce offspring, the offspring then simply replace the least fit individuals in the population so that a constant population size is maintained.

Table 4. EA Parameters[a]

population size	100
no. of iterations	1000
parent selection	roulette wheel
selection pressure	1.25
no. levels in tree at initialization	3
max no. of child nodes	3
max attempts	10
convergence conditions (*x*, *n*, and *y*)	1, 400, 5
evolutionary operator rates:	
chance of mutation	30%
chance of modification	50%
chance of crossover	20%
chance of mutation occurring on each node	50%
chance of single vs two-parent crossover	50%
percentage of worst population members replaced each generation	10%

[a] See text for details. Modification refers to all genetic operators excluding mutation and crossover. Within this class of operators, one is chosen at random.

Preventing Overtraining. During training, the fitness of each query is evaluated on both a training set and a test set. Fitness in the training set is used for parent selection, that is, to control the evolution of the EA, whereas fitness in the test set is used to assess the generalizability of the trained queries and prevent overtraining of the queries on the training set. The EA proceeds until a convergence criterion has been fulfilled or a maximum number of iterations have been performed. Convergence is assessed by monitoring the change in the maximum fitness of the population in both data sets: a change of less than *x*% in the training set over the previous *n* generations (*x* and *n* are user-defined

Table 5. Hert Data Set[a]

activity class	actives	mean PS (SD)	training set	test set	validation set
5HT3 antagonists (5HT3 Ant)	752	0.33 (0.10)	251	251	250
5HT1A agonists (5HT1A)	827	0.32 (0.09)	276	275	276
5HT reuptake inhibitors (5HT-RT)	359	0.32 (0.11)	120	120	119
dopamine D2 antagonists (D2)	395	0.33 (0.10)	132	132	131
renin inhibitors (renin)	1130	0.51 (0.11)	377	377	376
angiotensin II AT1 antagonists (AT1 Ant)	943	0.39 (0.09)	314	314	315
thrombin inhibitors (thrombin)	797	0.36 (0.12)	266	266	265
substance P antagonists (Sub P)	1246	0.37 (0.10)	415	415	416
HIV 1 protease inhibitors (HIV1)	750	0.39 (0.11)	250	250	250
COX inhibitors (COX)	626	0.26 (0.09)	212	212	212
protein kinase C inhibitors (PKC)	453	0.29 (0.14)	151	151	151

[a] The mean pairwise similarities and standard deviations (Mean PS (SD)) are reported for each data set measured using Daylight fingerprints.[18] In each case, the active compounds were mixed with sets of 3000 inactive compounds also extracted from MDDR as described in the text.

Table 6. GSK Data Sets

target	inactive $pIC_{50} \leq 4.3$	low threshold $4.3 < pIC_{50} \geq 5.0$	med threshold $5.0 < pIC_{50} \geq 6.0$	high threshold $pIC_{50} > 6.0$	total
1A2	3603	157	215	131	4106
2C19	2470	727	734	175	4106
2C9	2275	449	948	434	4106
2D6	2783	407	531	385	4106
3A4	2028	605	1179	294	4106
hERG-S[a]	3201	1120	1003	403	5727
hERG-E	1715	130	430	111	2386

[a] The thresholds used for the hERG-S data set are 4.52, 5.03, and 6.0 due to assay sensitivity issues that resulted in some level of discontinuity in the pIC_{50} values.

values—see Table 4) or a drop in fitness in the test set of more than *y*% below the current training set fitness (user-defined value) indicate convergence. The latter indicates that overtraining is occurring. On termination of the EA, the best query is selected as that with greatest test set fitness found over all generations. A third independent data set (the validation set) is then used to score the best query.

EA Parameters. Preliminary experiments were carried out (results not shown) to establish appropriate parameters for the subsequent runs reported here. These are shown in Table 4.

Data Sets. The EA has been applied to publicly available data sets extracted from the MDL's Drug Data Reports Database (MDDR)[25] and to GSK in-house screening data sets. The MDDR data sets consist of the 11 activity classes shown in Table 5 as used by Hert et al.[26] While the MDDR has become a standard data set on which new methods are evaluated, its characteristics are somewhat different from those of real screening data sets. For example, while a high degree of confidence can be associated with the classification of active compounds in the MDDR (they are drugs that have been launched or are under development),[25] the compounds that are taken as "inactives" are compounds that have not been tested for activity in the classes being investigated rather than being true inactives. In contrast, for the in-house screening data all compounds have been tested for activity and have been assigned quantitative activity values.

The GSK data sets consist of compounds tested in seven developability assays against six different targets. The data sets were first cleaned by removing duplicates and compounds with percentage inhibition values that were either negative or above 100. The resulting number of compounds in each of the target classes is shown in Table 6. Five of the targets are the major cytochrome P450 (CYP450) enzymes and therefore give an indication of the potential bioavailability of the compounds tested.[27] The remaining target, hERG (protein derived from human Ether-a-go-go Related Gene) is a potassium ion channel, which is of particular importance in the maintenance of normal heart function.[28] Molecules that bind to and block the channel can lead to arrhythmia of the heart with potentially fatal consequences. There are two assays representing hERG activity: one assesses the disruption of activity from binding to the hERG channel via the consequential electrophysiological disruption (hERG-E), and the other measures binding to the hERG channel directly (hERG-S). In all cases, the activity values are recorded as pIC_{50} values. The compounds have been grouped into activity bands using the thresholds shown in Table 6.

RESULTS

Robustness of the EA. Three of the MDDR data sets were used to investigate the robustness of the EA: the renin inhibitors, the 5HT1A agonists, and the COX inhibitors. These were chosen to represent the range of different activity classes: the renin inhibitors are structurally very similar; the 5HT1A agonists are intermediate in structural diversity; and the COX inhibitors are structurally heterogeneous and therefore represent a difficult challenge, as acknowledged in previous studies.[29]

The variation in performance of the EA was tested over three different selections of active compounds and three different sets of inactives. Each activity class was divided at random into three equally sized subsets, see Table 5. Three sets of 3000 inactive compounds were selected at random

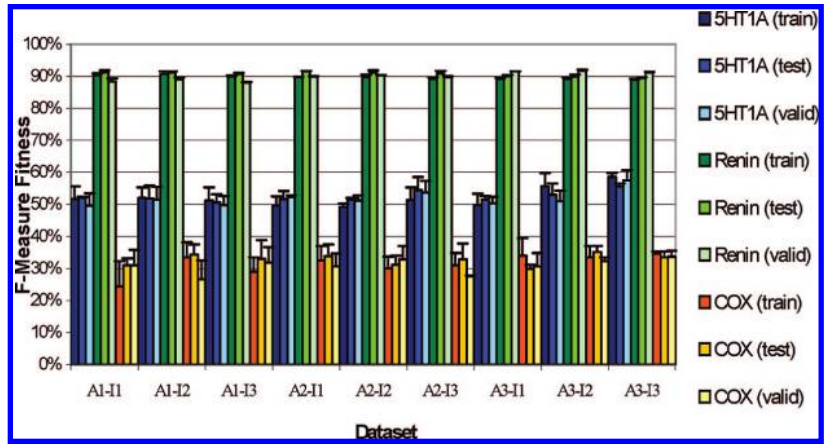

Figure 10. F-measure fitness values for the training, test, and validation data sets representing the renins, the 5HT1A agonists, and the COX inhibitors over the nine different data sets. The error bars indicate the standard deviation over three runs of the EA for each data set using different random seeds.

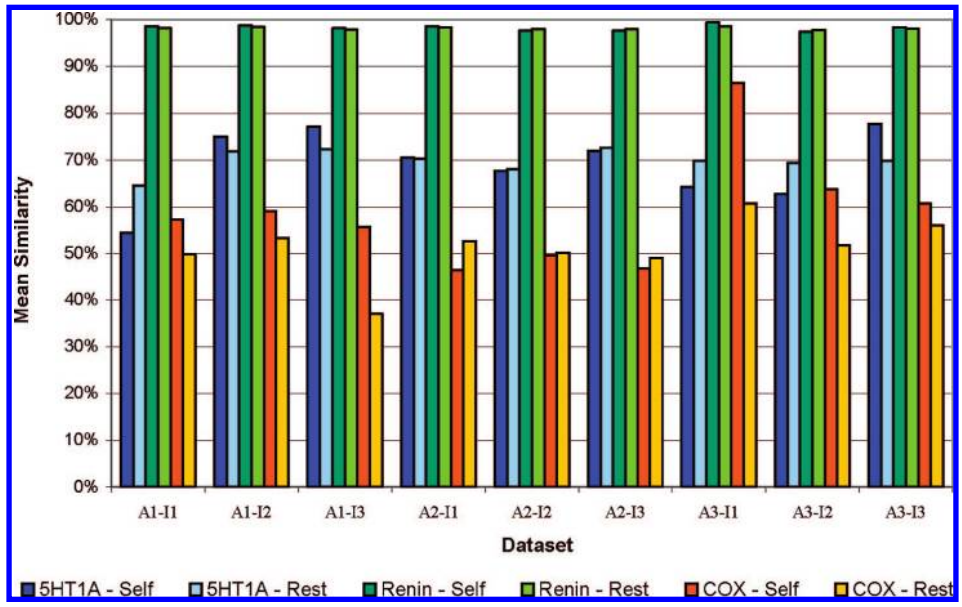

Figure 11. Phenotypic similarity of queries generated from repeat runs on a single data set (self) and queries generated on other data sets (rest).

and independently from the remainder of the MDDR after removal of all the Hert activity classes. The active and inactive subsets were paired to form training, testing, and validation sets, respectively. The inactive subsets were then exchanged so that each subset of inactives was used in training. Finally, the entire process was repeated twice more for different random partitionings of the actives. Thus a total of nine data sets was generated per activity class, each consisting of training, test, and validation sets. The data sets are labeled as *Ax-Iy* where *x* and *y* take the values 1, 2, or 3 to indicate the different subsets used in training.

As discussed in the Methods section, evolution in the EA is driven by performance on the training set with the EA configured to maximize the F-measure on the training set. The test set is used to select the best query generated throughout the course of the EA. The use of both training and test sets is designed to prevent overtraining. In a typical run, performance on both the training set and the test set increases as the EA progresses from initialization; however, at some point performance on the test set reaches a maximum and then begins to decline, while performance on the training set continues to increase. This indicates that the EA has begun to overtrain on the training set while losing the ability to generalize. The independent validation set mimics the application of the query to the selection of previously unseen compounds.

Results are shown in Figure 10 for the nine different data sets derived for each activity class. The error bars represent repeat runs of the EA using the same data set but varying the random seed. In all cases, the F-measure is shown for the training, test, and validation data sets.

The relative performance across the different activity classes corresponds to the differences in their structural diversities and is consistent with previous studies.[6,11,30] Hence, the performance on the renins is best (mean validation set F-measure 89.9 $\pm$ 1.4), COX is the worst (mean validation set F-measure 30.8 $\pm$ 2.3), and the performance on the 5HT1As is intermediate (mean validation set F-measure 51.9 $\pm$ 2.4). The renin results are very consistent over the different data sets and the different random seeds. Greater variation is seen across the different runs on the 5HT1A and COX activity classes; however, the variation using different data sets is comparable to varying the random seed on the same data set and therefore the subsequent

experiments are based on multiple runs using the same data sets. Both the reduced performance and the greater variation can be attributed to the greater structural diversity of these data sets. For all three activity classes, the performance in the test and training sets is better than that seen in the validation set as would be expected; however, the differences are remarkably small, indicating the excellent robustness of the models.

The queries can also be compared on the similarities of their phenotypes, i.e., the data set compounds they retrieve (predicted actives). A *retrieval profile* is derived for each query as a binary vector of length equal to the number of compounds in the data set. Each bit in the vector is set to "1" if the corresponding compound is predicted active; otherwise it is set to "0". The retrieval profiles of two queries can then be compared using the Tanimoto coefficient.

The similarities between the retrieval profiles of the queries are shown in Figure 11. The bars labeled *self* are the mean similarities between the three queries evolved for the same data set using different random seeds; the bars labeled *rest* are the mean similarities to all the other queries evolved on different data sets.

The renin queries are all very similar in terms of the compounds they retrieve, both for queries generated with different random seeds on the same data set and across different data sets. This is as expected since they all achieve high recall which leaves little scope for differences in the retrieval profiles. The 5HT1A and COX queries have lower levels of similarity, and compared to the renin queries there is more variation across the data sets. However, as before the difference between using different random seeds on one data set and different data sets is about the same.

Extracting SARs. The best query evolved for each activity class is shown in Figure 12 together with the corresponding confusion plot that quantifies its performance on the respective validation set. The renin query has recall of 85.6% and precision of 96.1% (325 out of 376 of the actives in the validation set are correctly identified with only 13 of the inactives being retrieved as false actives); the 5HT1A query is intermediate in performance (recall of 55.8%; precision 52.4%), and the COX query has lowest recall (23.1%), although it exhibits good precision (51.6%).

The RG queries are shown as SMARTS in Figure 12. However, these are difficult to interpret due to the complexities of the SMARTS notation and to the use of uncommon atom types to represent the different node types. Furthermore, the different data sets used in training can result in different queries being evolved. Therefore, the nine queries evolved for each activity class (one for each data set) have been analyzed to investigate the consistency in the structure−activity information generated over the different training sets. The queries are shown in Figure 13 in a common alignment (referenced using the labels A, B, C, etc. in the final column) to aid the comparison.

The 5HT1A and COX queries are of a similar length and mostly consist of three nodes, while the renin queries are typically larger. This corresponds to the general characteristics of the RGs generated for the different activity classes which are shown in Figure 14. The inactives range in size from around 2 to 20 nodes, with the modal peak around 7. The 5HT1A and COX RGs have similar distributions to the inactives and typically range in size from 2 to 10 nodes,

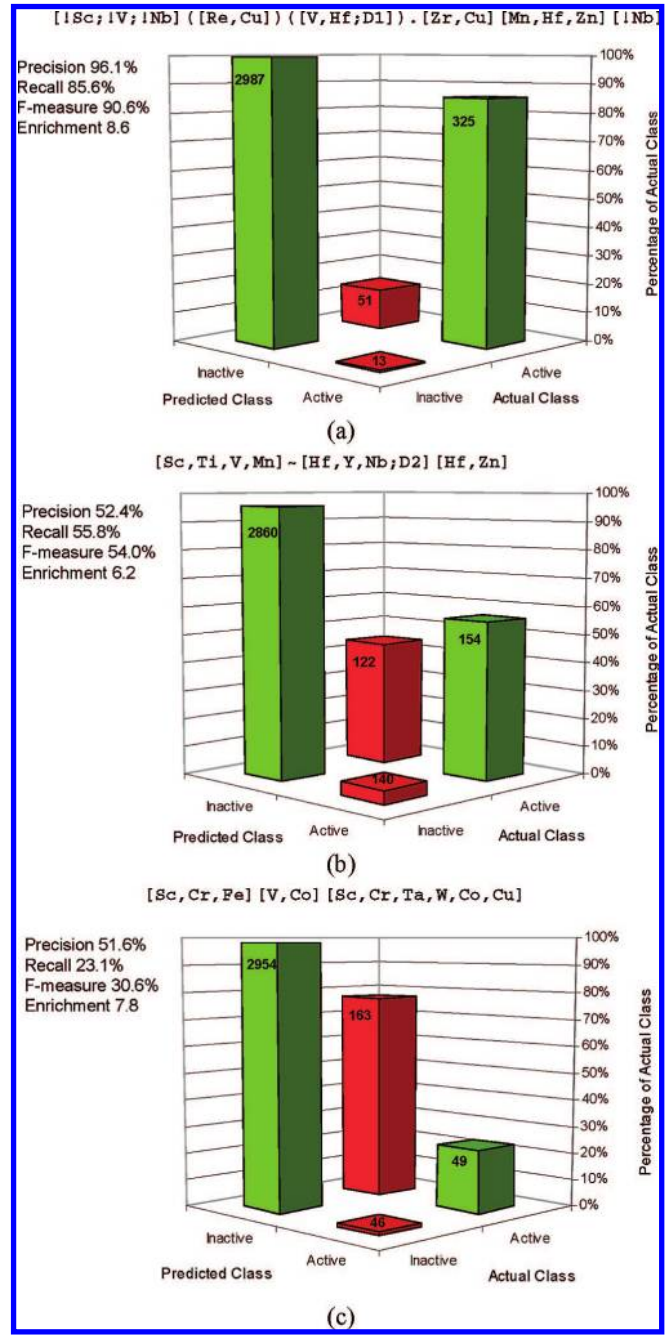

Figure 12. RG queries and confusion matrices for (a) renins; (b) 5HT1A agonists; and (c) COX inhibitors.

with modal peaks of 5 and 6, respectively; however, the renin RGs are mostly in the range of 5−20 nodes with a modal peak of 12. This difference in the characteristics of the RGs together with their structural homogeneity explains the excellent performance achieved for this activity class. Recently it has been suggested that a more realistic estimate of predictivity can be obtained if the decoys (or inactives) have similar physicochemical properties to the actives, i.e., each activity class should have its own carefully selected set of inactives. While we have not conducted these experiments on the MDDR, we have investigated the performance of the method on real screening data sets which do represent realistic challenges. Although the 5HT1A and COX queries are of a similar size, both to each other and to the inactives, the COX queries have lower recall than the 5HT1As and show greater variation in the queries. This is

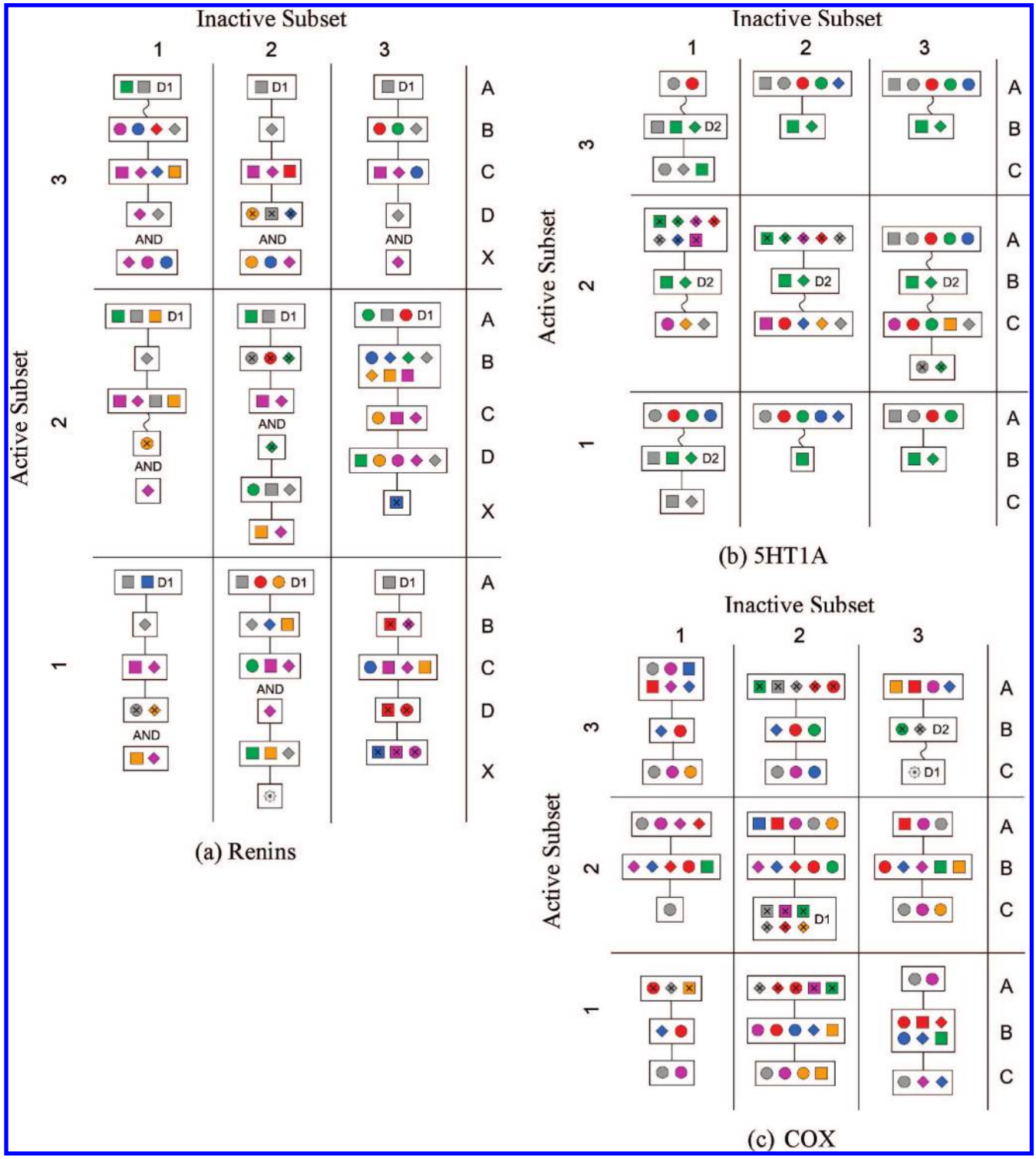

Figure 13. The nine queries evolved for the nine different training sets representing each activity class. The color scheme for the nodes is given in Table 1. The crosses within nodes indicate the NOT tag.

presumably due to the increased diversity of structures within the class such that it is not possible to describe them using a single RG query. This is also suggested by the phenotypic data in Figure 11 where the degree of overlap in the compounds retrieved in different runs is relatively low.

The aligned queries can be used to propose structural requirements for activity. For example, all the renin queries contain a terminal featureless aliphatic ring node [Hf;D1] at position A. Similarly, they all contain a linker node ([Zn]) at position B, although in some queries this is given alongside alternative nodes at the same position. Two nodes occur as alternatives at position C in all of the queries: aliphatic donor/acceptor and acyclic donor/acceptor. Thus it seems that the hydrogen bonding properties are more important at this position than whether the node is cyclic or acyclic. The graph reduction process, however, gives precedence to cyclic/acyclic nodes in preference to hydrogen bonding properties which introduces a somewhat artificial difference between the active compounds. This effect is removed in this case through the specification of alternative nodes. While six out

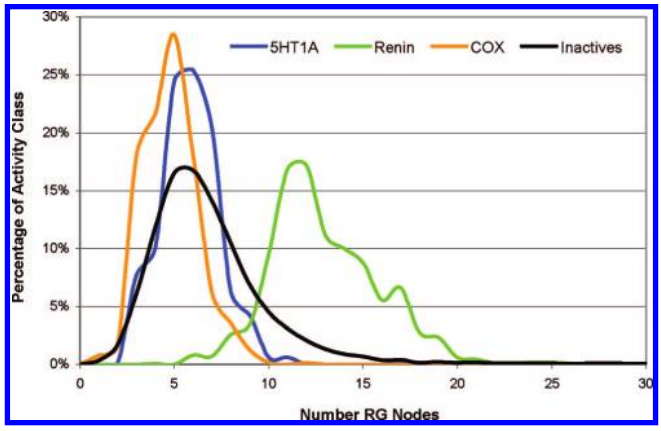

Figure 14. Distributions of the number of nodes in each RG are shown for the inactives and the activity classes renins, 5HT1A, and COX activity classes.

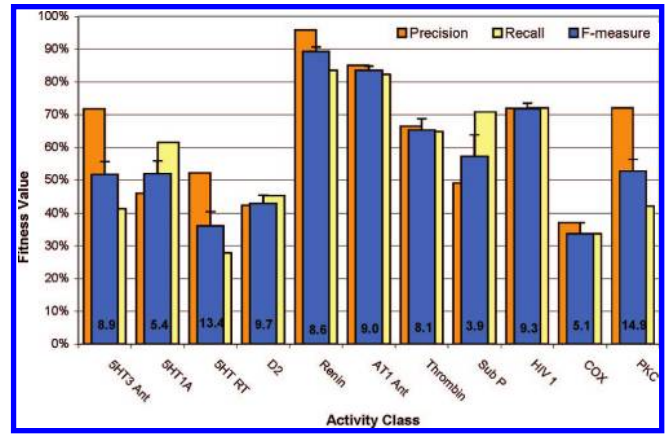

Figure 15. Validation set performance on the 11 activity classes extracted from the MDDR. Enrichment factors are shown as labels on bars.

of the seven renin queries that have nodes at position D include an acyclic joint donor/acceptor node ([Cu]) as a possible node (either explicitly or implicitly through the use of NOT logic), this has much lower specificity than the aliphatic ring at position A, i.e. there are more alternative nodes that could exist at this position and the NOT tags in four of the queries allow many different nodes to match. The features identified here are consistent with the published pharmacophore for peptide-based renin inhibitors and reflect the prevalence of these in the data set.[31] While more variation is seen across the queries evolved for the 5HT1A and COX activity classes, some well-known features are evident. The 5HT1A queries in particular contain features that are consistent with the established pharmacophore:[32] all queries were found to contain an aromatic featureless or aromatic acceptor node adjacent to an aliphatic positively ionizable node, with five of the nine queries also containing an acyclic featureless (linker) node. The published pharmacophore[32] also specifies an acceptor (imide group) that is not accounted for in the queries, which are limited to a maximum of three nodes. This reduced level of specificity is necessary for the queries to be able to describe the range of different chemical series present; a number of which do not contain the additional feature of the published SAR. Our companion paper[33] extends the methodology presented here to capture multiple SARs within a data set by evolving multiple queries with each query representing a subset of the actives compounds.

The structure−activity relationships encoded in the RG queries are focused on gross features of the molecules, for example, their binding characteristics rather than detailed substructural fragments, which allows commonalities to be perceived between compounds with different underlying skeletons. The resulting queries can indicate to the chemist where a compound should or should not be modified in order to maintain activity.

The MDDR Results. The quantitative performance of the EA across all the Hert activity classes is shown in Figure 15 as the F-measures averaged across the nine data sets for each activity class (the error bars indicate the standard deviations). The recall, precision, and enrichment factors (shown as labels on bars) are also shown to give a more detailed picture of query performance. In all cases the performance is shown for the independent validation data sets. The variation in performance across the activity classes follows the same trend as the mean intraclass similarities of the actives.[26]

As discussed earlier, a binary classifier has two conflicting objectives to optimize: maximizing the number of true positives while minimizing the number of false positives. By combining recall and precision using the F-measure a compromise in the two objectives is found. In the data sets considered so far, precision is often favored over recall such that the number of false positives is relatively low (Figure 15). A low false positive rate may be useful for extracting SAR since a more precise query is likely to be produced; however, it is generally undesirable for virtual screening where often a higher false positive rate would be tolerated in order that true positives are not missed. An extension of the single objective optimization approach is described in the companion paper[33] in which recall and precision are optimized independently to generate a family of RG queries that explore the tradeoff in recall and precision.

GSK Data Sets. *hERG-S Data Set.* The EA was trained to evolve queries of differing specificity using training sets based on different activity thresholds:

- Low activity threshold: the low, medium, and high activity compounds were combined to form the "actives", and all compounds with $pIC_{50} < 4.52$ were classed as inactive.
- Medium activity threshold: the medium and high activity compounds were combined as "actives", and all compounds with $pIC_{50} < 5.03$ were classified as inactive.
- High activity threshold: the high activity compounds ($pIC_{50} > 6.0$) only were used as "actives"; all other compounds were classified as inactive.

For each activity threshold, the actives were divided at random into three equally sized subsets (for training, testing, and validation), and the inactives were divided into two subsets (one for training and testing and one for validation). As before, the EA was configured to maximize the F-measure applied to the training set, the query output by the EA was that with maximum F-measure on the test set, and results are reported for the query applied to the independent validation set. The results are shown as confusion plots in Figure 16 along with the RG queries themselves. The F-measure fitness is greatest for the query derived using the low activity threshold which has recall of 67.6% and precision 60.5% (the number of compounds predicted active

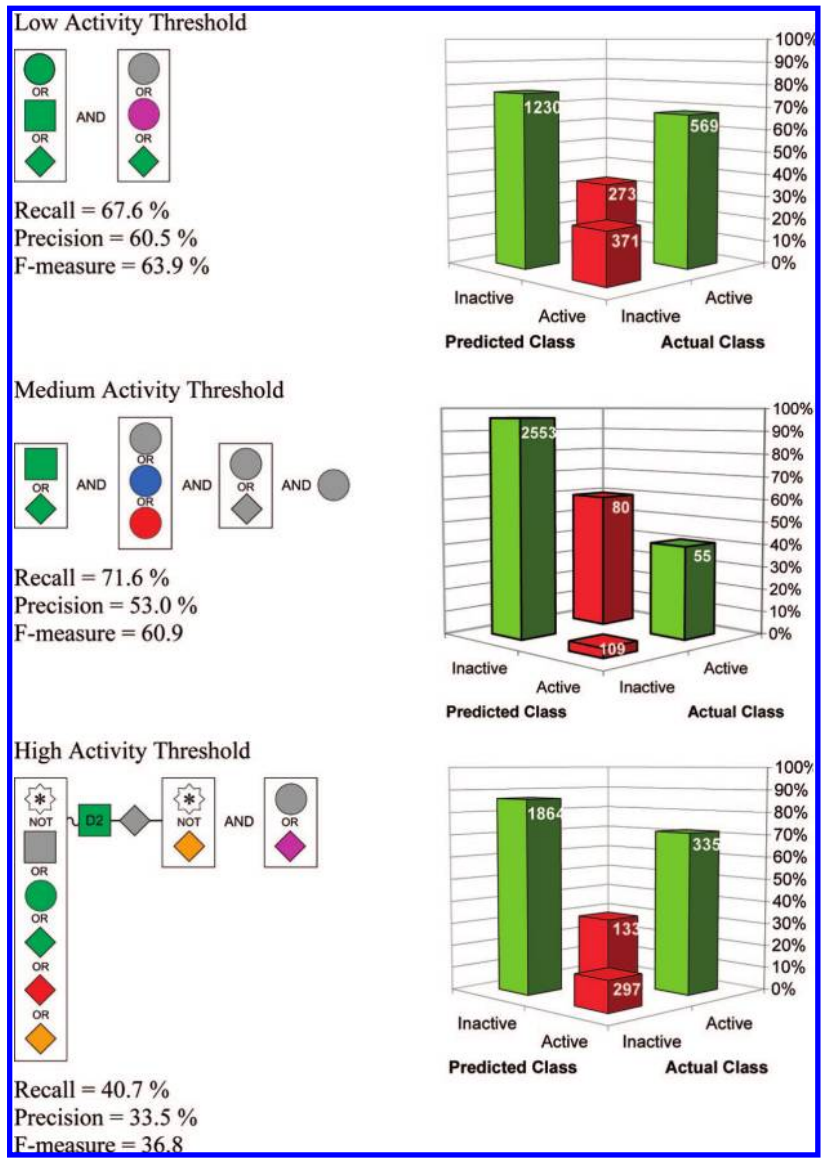

Figure 16. The RG queries evolved for the different activity thresholds are shown together with their performances on the respective validation sets.

is 940 of which 569 are true actives at this threshold, i.e. $pIC_{50} > 4.52$); the query derived using the medium activity threshold has recall of 71.6% and precision 53.0% (the number of compounds predicted active is 632 of which 335 are true actives, i.e. $pIC_{50} > 5.03$); and the query derived using the high activity threshold has recall of 40.7% and precision of 33.5% (the number of compounds predicted active is 164 of which 55 are true actives, i.e. $pIC_{50} > 6.0$). Figure 17 shows the distribution of the compounds predicted as active over the full range of activity values for each validation set. Thus, in Figure 17a, the compounds predicted as active when using the low activity threshold are distributed over the four categories: high activity, medium activity, low activity, and inactive. For each threshold, the proportion of misclassified molecules is greatest when the activity is closest to the threshold and least for molecules with activities that differ most from the classification threshold. Thus, the queries derived at the low and medium activity thresholds correctly classify the majority of the high activity compounds. As the threshold on activity is increased, the queries become more specific so that the query derived using the high activity threshold retrieves fewer compounds overall, and a large number of the high activity compounds are also misclassified. This is most likely due to the similarity of compounds in the high activity class to those in the medium activity class which were classified as inactive during training. In fact, the query is "successful" in predicting the majority of the medium active compounds as inactive; however, this is at the expense of false inactives in the high activity class. These experiments highlight the difficulty of choosing an appropriate threshold on activity when dealing with quantitative data.

The query derived using the medium activity threshold appears to be a good compromise: the recall of medium and high activity compounds is high and the false positives are evenly distributed between those in the low activity band which may also be of interest and those with pIC50 below 4.52. However, if the aim is to derive SAR information, then a more specific query may be desirable as demonstrated by an analysis of the queries themselves. The low-threshold query requires the presence of two disjoint nodes only, while the high-threshold query requires a four-node sequence and a further disconnected node. The inclusion of the SMARTS operators within the RG queries allows alternative nodes to be represented in the queries, and the variation in the nodes

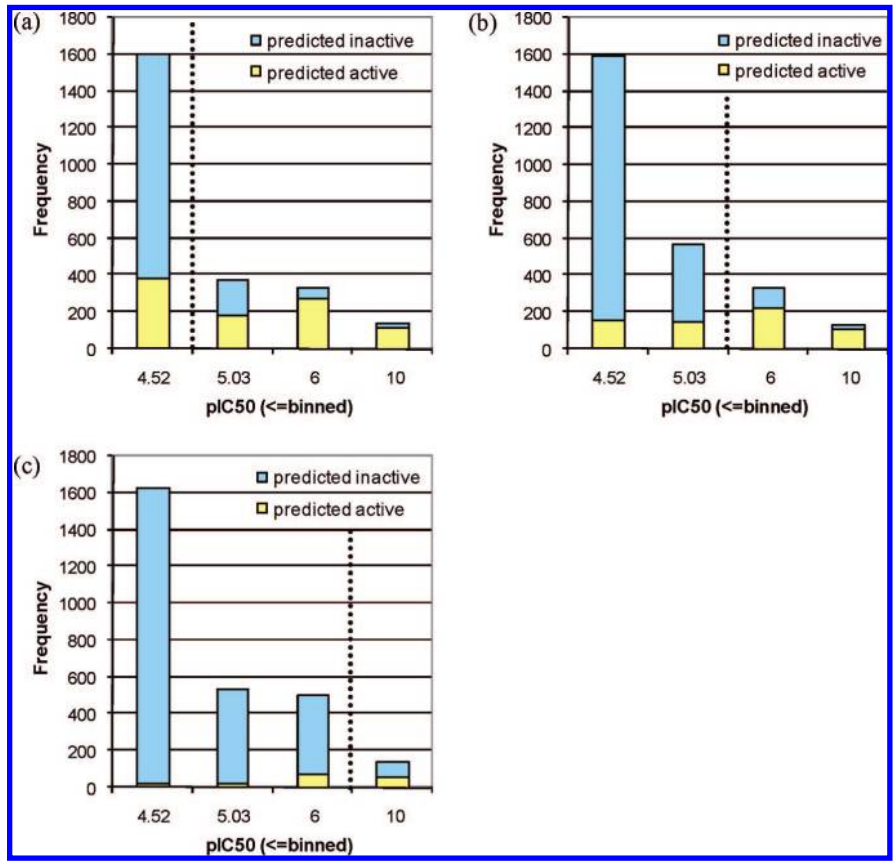

Figure 17. Distribution of predicted actives across the validation sets for queries derived using (a) low activity threshold; (b) medium activity threshold; and (c) high activity threshold. The validation sets are different in each case due to the different numbers of "actives" and "inactives" being used (see text for data set construction details).

is also an indicator of query specificity. Thus, the low-threshold query requires the presence of any positively ionizable group; in the medium-threshold query this node is restricted to an aliphatic or acyclic positively ionizable node; and in the high-threshold query the node must be an aliphatic positively ionizable group of degree two. All the queries include the specification of an aromatic node as an additional node. The SAR information thus obtained is consistent with current findings relating to the hERG pharmacophore.[34,35]

The relative importance of the alternative nodes can be determined by removing them one at a time from a query and examining the effect on performance. For example, where a query contains two alternative nodes at the same position, two simplified queries are generated by first removing one of the nodes, leaving the rest of the query intact, and then removing the other. Analysis of the medium-threshold query in this way showed that the most significant features are the aliphatic positively ionizable node and the featureless aromatic node which is consistent with the node definitions in the medium-threshold query; see Figure 18.

All GSK Data Sets. Quantitative results (recall and precision) are shown in Figure 19 for all the GSK data sets. The queries were evolved using the medium activity threshold values, and the classification rates are shown for the test and validation data sets. As for the MDDR data sets the results for the validation sets are remarkably similar to those seen in the test sets demonstrating the good generalizability of the models generated. The hERG-E data set shows comparable performance to the hERG-S data set.

The performance on the cytochrome P450 enzymes is reduced and reflects the none-specific nature of these

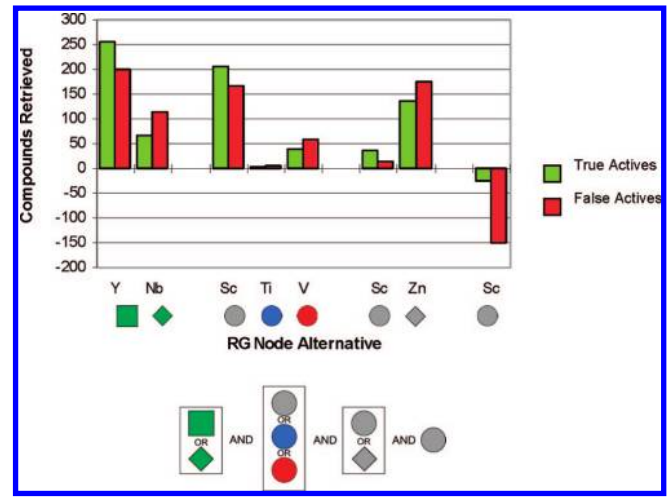

Figure 18. The importance of the nodes in an RG query can be analyzed by removing them one at a time and examining the effect on performance. Inclusion of the aliphatic +ve ionizable group leads to the retrieval of 250 true actives and around 200 inactives, whereas the acyclic +ve ionizable node retrieves around 50 true actives and 120 inactives. Thus the aliphatic +ve ionizable node is the more significant of the two which is also confirmed in the query derived using the high threshold on activity.

binding sites which are able to accommodate a wide variety of ligands. For example, 2D6 is one of the more studied P450s, and its characteristic pharmacophore has been reported to contain a positively charged nitrogen which is thought to be important for binding to negatively charged amino acid residues in the 2D6 active site.[36] Other features found to be important, particularly for 2D6

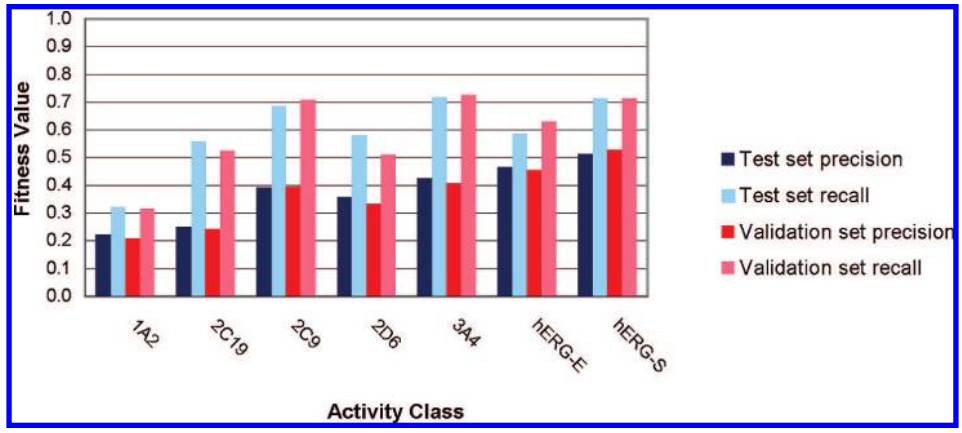

Figure 19. Test and validation set performance of queries derived on the GSK data sets at the medium activity cutoff.

inhibitors rather than substrates, are two to three hydrophobic regions (typically fulfilled by aromatic rings) in addition to a hydrogen bonding feature. [37] The set of queries derived from repeat runs on the 2D6 low activity cutoff data set are quite generic, only consisting of two disjoint nodes. One specifies any positively ionizable feature (either aromatic, aliphatic, or acyclic) with the other specifying either a featureless (hydrophobic) aliphatic or aromatic ring or an aromatic donor. The medium activity cutoff queries also contain a number of the same features, although with a higher level of specificity. The positively ionizable feature is now restricted to aliphatic and acyclic structural contexts, while the aromatic featureless node is restricted to being terminal. It is also notable that an acyclic donor occurs as an alternative to the aromatic featureless node in the majority of the queries. Some of the queries also specify that an acyclic featureless node is adjacent to the positively ionizable feature. The high activity cutoff queries are yet more specific, all giving information as to which features are required adjacent to the positively ionizable feature (aliphatic or acyclic) in addition to specifying two additional disjoint features. Most often the features adjacent to the positively group are either acyclic featureless or acyclic joint donor/acceptor; however, alternatives in some of the queries include aromatic and aliphatic donors along with aliphatic featureless aromatic joint donor/acceptors. One of the disjoint nodes consistently requires that a terminal featureless aromatic or aliphatic ring is present, although aliphatic and aromatic donors also occur as additional alternatives. The remaining disjoint node is typically a wildcard, which effectively adds a size constraint, ensuring that the query will only match RGs with four or more nodes.

Two factors may contribute to the poor performance of the 1A2 data set. One being the lack of available data on active compounds; at the medium activity threshold, there are only 115 active compounds in the training set compared with 1880 inactives, whereas the other data sets have a higher proportion of actives to inactives. The other may be the inability of the RGs to capture subtleties in the SAR which has been shown to be dependent on dipole moment and homo-lumo energy differences.[38]

CONCLUSIONS

A novel method for deriving structure-activity relationships is described based on the generation of RG queries. An evolutionary algorithm is used to evolve an RG query (or subgraph) that maximizes the separation of active and inactive molecules. The EA is designed to optimize the classification rate which is measured using the F-measure and represents a compromise in recall (number of actives retrieved by the query) and precision (the total number of compounds retrieved by the query). The method has been tested on data sets extracted from the MDDR as well as in-house screening data sets provided by GSK. The resulting queries encode the SARs in terms of topological pharmacophores which are readily interpretable by chemists. The EA has also been shown to have good predictive performance with classification rates on independent validation sets that are only slightly reduced relative to those used in training.

The EA is a traditional optimization method in that it seeks to optimize a single objective, in our case the F-measure. A limitation of the F-measure is that it is insensitive to how the false predictions are distributed between the classes. Thus for some data sets, relatively high recall is achieved albeit at the expense of low precision, whereas in other cases the queries are biased toward high precision and low recall. While the F-measure provides a balance between these two objectives, the user is unable to control where the balance lies. For a homogeneous set of actives, where the actives have features that easily distinguish them from the inactives, it can be possible to achieve both high recall and high precision, as was seen for the renin data set. However such data sets are unrealistic of typical screening data which is generally more heterogeneous with a tradeoff existing between precision and recall. Encoding the queries as RG representations facilitates the handling of heterogeneous data sets since they enable compounds which have similar binding properties but different two-dimensional skeletons to be represented by a single query. Further flexibility is also allowed through the incorporation of logical operators within the queries, for example, alternative nodes can be specified. However, for many data sets, it may not be possible to capture all the actives with a single query, for example those which contain compounds with different binding modes. In the companion paper, we extend the EA methodology to incorporate multiobjective optimization techniques in which recall and precision are treated independently thus removing the limitations of the single objective. We also describe how a third objective can be used to evolve queries that complement one another so that multiple structure−activity relationships can be derived.

ACKNOWLEDGMENT

We acknowledge Eleanor Gardiner for helpful comments on this manuscript, Daylight Chemical Information for software support, and MDL Information Systems Inc. for the provision of the MDDR database. The work was funded by GlaxoSmithKline and BBSRC via an industrial CASE studentship.

REFERENCES AND NOTES

(1) Harper, G.; Pickett, S. D. Methods for mining HTS data. *Drug Discovery Today* **2006**, *11*, 694–699.
(2) Saeh, J. C.; Lyne, P. D.; Takasaki, B. K.; Cosgrove, D. A. Lead hopping using SVM and 3D pharmacophore fingerprints. *J. Chem. Inf. Model.* **2005**, *45*, 1122–1133.
(3) Rogers, D.; Brown, R. D.; Hahn, M. Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening follow-up. *J. Biomol. Screening* **2005**, *10*, 682–686.
(4) Chen, B. N.; Harrison, R. F.; Pasupa, K.; Willett, P.; Wilton, D. J.; Wood, D. J.; Lewell, X. Q. Virtual screening using binary kernel discrimination: Effect of noisy training data and the optimization of performance. *J. Chem. Inf. Model.* **2006**, *46*, 478–486.
(5) Glick, M.; Jenkins, J. L.; Nettles, J. H.; Hitchings, H.; Davies, J. W. Enrichment of high-throughput screening data with increasing levels of noise using support vector machines, recursive partitioning, and Laplacian-modified naïve Bayesian classifiers. *J. Chem. Inf. Model.* **2006**, *46*, 193–200.
(6) Cannon, E. O.; Amini, A.; Bender, A.; Sternberg, M. J. E.; Muggleton, S. H.; Glen, R. C.; Mitchell, J. B. O. Support vector inductive logic programming outperforms the naïve bayes classifier and inductive logic programming for the classification of bioactive chemical compounds. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 269–280.
(7) Plewczynski, D.; Spieser, S. A. H.; Koch, U. Assessing different classification methods for virtual screening. *J. Chem. Inf. Model.* **2006**, *46*, 1098–1106.
(8) van Rhee, A. M. Use of recursion forests in the sequential screening process: Consensus selection by multiple recursion trees. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 941–948.
(9) Bruce, C. L.; Melville, J. L.; Pickett, S. D.; Hirst, J. D. Contemporary QSAR classifiers compared. *J. Chem. Inf. Model.* **2007**, *47*, 219–227.
(10) Gillet, V. J.; Willett, P.; Bradshaw, J. Similarity searching using reduced graphs. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 338–345.
(11) Barker, E. J.; Cosgrove, D. A.; Gardiner, E. J.; Gillet, V. J.; Kitts, P.; Willett, P. Scaffold-hopping using clique detection applied to reduced graphs. *J. Chem. Inf. Model.* **2006**, *46*, 503–511.
(12) Gardiner, E. J.; Gillet, V. J.; Willett, P.; Cosgrove, D. A. Representing clusters using a maximum common edge substructure algorithm applied to reduced graphs and molecular graphs. *J. Chem. Inf. Model.* **2006**, *47*, 354–366.
(13) Harper, G.; Bravi, G. S.; Pickett, S. D.; Hussain, J.; Green, D. V. S. The reduced graph descriptor in virtual screening and data-driven clustering of high-throughput screening data. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2145–2156.
(14) Takahashi, Y.; Sukekawa, M.; Sasaki, S. Automatic identification of molecular similarity using reduced-graph representation of chemical-structure. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 639–643.
(15) Rarey, M.; Dixon, J. S. Feature trees: A new molecular similarity measure based on tree matching. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 471–490.
(16) Steifl, N.; Watson, I. A.; Baumann, K.; Zaliani, A. Erg: 2D pharmacophore descriptions for scaffold hopping. *J. Chem. Inf. Model.* **2006**, *46*, 208–220.
(17) Bemis, G. W.; Murcko, M. A. The properties of known drugs.1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 28872893.
(18) Daylight Daylight Chemical Information Systems, Inc., 120 Vantis - Suite 550, Aliso Viejo, CA 92656, U.S.A. www.daylight.com at http://www.daylight.com.
(19) Globus, A.; Lawton, J.; Wipke, T. Automated molecular design using evolutionary techniques. *Nanotechnology* **1999**, *10*, 290–299.
(20) Nachbar, R. B. Molecular evolution: Automated manipulation of hierarchical chemical topology and its application to average molecular structures. *Genetic Programming Evolvable Machines* **2000**, *1*, 57–94.
(21) Brown, N.; McKay, B.; Gasteiger, J. The de novo design of median molecules within a property range of interest. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 761–771.
(22) Lameijer, E. W.; Kok, J. N.; Back, T.; Ijzerman, A. P. The molecule evoluator. An interactive evolutionary algorithm for the design of drug-like molecules. *J. Chem. Inf. Model.* **2006**, *46*, 545–552.
(23) Gusfield, D. *Algorithms on strings, trees and sequences: Computer science and computational biology*; Cambridge University Press: 1997.
(24) van Rijsbergen, C. J. *Information retrieval*; Butterworth: London, 1979.
(25) MDL Information Systems Inc. 2440 Camino Ramon, Suite 300, San Ramon, CA 94583.
(26) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177–1185.
(27) Zlokarnik, G.; Grootenhuis, P. D. J.; Watson, J. B. High throughput p450 inhibition screens in early drug discovery. *Drug Discovery Today* **2005**, *10*, 1443–1450.
(28) Sanguinetti, M. C.; Tristani-Firouzi, M. hERG potassium channels and cardiac arrhythmia. *Nature* **2006**, *440*, 463–469.
(29) Birchall, K.; Gillet, V. J.; Harper, G.; Pickett, S. D. Training similarity measures for specific activities: Application to reduced graphs. *J. Chem. Inf. Model.* **2006**, *46*, 577–586.
(30) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. New methods for ligand-based virtual screening: Use of data fusion and machine learning to enhance the effectiveness of similarity searching. *J. Chem. Inf. Model.* **2006**, *46*, 462–470.
(31) Dellaria, J. F.; Maki, R. G.; Bopp, B. A.; Cohen, J.; Kleinert, H. D.; Luly, J. R.; Merits, I.; Plattner, J. J.; Stein, H. H. Optimization and invivo evaluations of a series of small, potent, and specific renin inhibitors containing a novel leu-val replacement. *J. Med. Chem.* **1987**, *30*, 2137–2144.
(32) Chilmonczyk, Z. Ligand-5-ht1a receptor interaction. *Il Farmaco* **2000**, *55*, 191–193.
(33) Birchall, K.; Gillet, V. J.; Harper, G.; Pickett, S. D. Evolving interpretable structure−activity relationship models. 2. Using multi-objective optimization to derive multiple models. *J. Chem. Inf. Model.* **2008**, *48*, 1558–1570.
(34) Song, M.; Clark, M. Development and evaluation of an in silico model for hERG binding. *J. Chem. Inf. Model.* **2006**, *46*, 392–400.
(35) Aronov, A. M.; Goldman, B. B. A model for identifying hERG K+ channel blockers. *Bioorg. Med. Chem.* **2004**, *12*, 2307–2315.
(36) Wolff, T.; Distlerath, L. M.; Worthington, M. T.; Groopman, J. D.; Hammons, G. J.; Kadlubar, F. F.; Prough, R. A.; Martin, M. V.; Guengerich, F. P. Substrate-specificity of human-liver cytochrome-p-450 debrisoquine 4-hydroxylase probed using immunochemical inhibition and chemical modeling. *Cancer Res.* **1985**, *45*, 2116–2122.
(37) Ekins, S.; Bravi, G.; Binkley, S.; Gillespie, J. S.; Ring, B. J.; Wikel, J. H.; Wrighton, S. A. Three and four dimensional-quantitative structure activity relationship (3D/4D-QSAR) analyses of CYP2D6 inhibitors. *Pharmacogenetics* **1999**, *9*, 477–489.
(38) Chohan, K. K.; Paine, S. W.; Mistry, J.; Barton, P.; Davis, A. M. A rapid computational filter for cytochrome p450 1A2 inhibition potential of compound libraries. *J. Med. Chem.* **2005**, *48*, 5154–5161.

CI8000502