JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

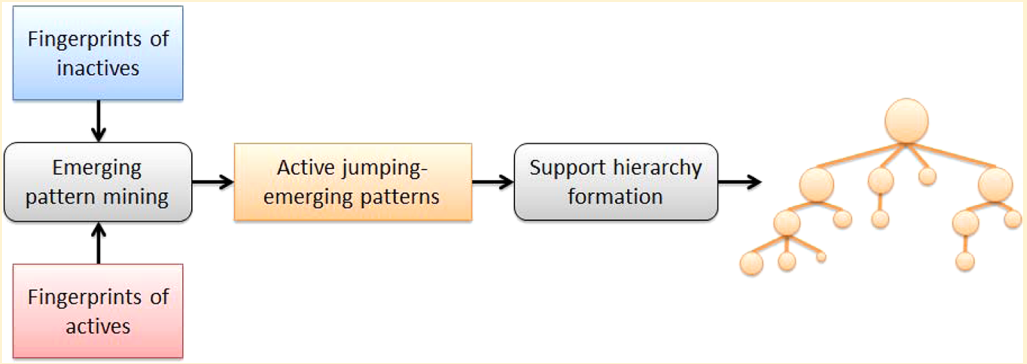
pubs.acs.org/jcim

Automating Knowledge Discovery for Toxicity Prediction Using Jumping Emerging Pattern Mining

Richard Sherhod,[†] Valerie J. Gillet,*[,†] Philip N. Judson,[‡] and Jonathan D. Vessey[‡]

[†]Information School, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield S1 4DP, U.K.

[‡]Lhasa Limited, 22-23 Blenheim Terrace, Woodhouse Lane, Leeds, LS2 9HD, U.K.



ABSTRACT: The design of new alerts, that is, collections of structural features observed to result in toxicological activity, can be a slow process and may require significant input from toxicology and chemistry experts. A method has therefore been developed to help automate alert identification by mining descriptions of activating structural features directly from toxicity data sets. The method is based on jumping emerging pattern mining which is applied to a set of toxic and nontoxic compounds that are represented using atom pair descriptors. Using the resulting jumping emerging patterns, it is possible to cluster toxic compounds into groups defined by the presence of shared structural features and to arrange the clusters into hierarchies. The methodology has been tested on a number of data sets for Ames mutagenicity, oestrogenicity, and hERG channel inhibition end points. These tests have shown the method to be effective at clustering the data sets around minimal jumping-emerging structural patterns and finding descriptions of potentially activating structural features. Furthermore, the mined structural features have been shown to be related to some of the known alerts for all three tested end points.

INTRODUCTION

Having the ability to predict potential toxic and environmental effects of chemical compounds is of growing importance to the pharmaceutical, cosmetics, and agrochemical industries.[1−3] The motivation for the development of computational tools arises from their low cost compared to in vivo or in vitro experiments; the possibility of applying the methods to compounds that have not yet been synthesized; and ethical factors such as reducing the need for animal testing.

The toxicity of a compound can be attributed to its physicochemical and structural properties so that many of the structure−activity relationship (SAR) methods that are used to predict therapeutic activities have also been applied to toxicity prediction. However, predicting toxic effects is particularly challenging for a number of reasons: multiple different end points exist; the same end point can arise through multiple mechanisms; and for many end points, such as carcinogenicity, the mechanisms are poorly understood.[4] Statistical methods of prediction such as quantitative structure−activity relationship (QSAR) modeling have limited success due to the poor quality and quantity of available toxicity data and the breadth of mechanisms that exist. Machine learning methods such as neural networks, random forests, and support vector machines have also been applied to toxicity prediction;[5] however, in many cases, it is not possible to extract an interpretable SAR from the models.

Expert systems, on the other hand, are based on knowledge from human experts encoded into multiple rules about relationships between structure and different toxic effects. These knowledge based systems can make predictions about many end points, provided that appropriate information is present within the knowledge base. The rules within an expert system's inference engine often include conditional clauses that link physicochemical and structural properties to a particular biological mechanism.[6] For example, the Derek Nexus system[7] encodes structural features that have been associated with particular toxicological effects as structural alerts alongside other parameters such as physicochemical properties and uses a reasoning model to weigh up multiple arguments both for and against toxicity. A disadvantage of expert systems is that the process of developing new structural alerts to expand the knowledge base requires considerable time and effort from domain experts and involves detailed analysis of relevant literature. The aim of this work on emerging pattern mining is to help automate the process of knowledge extraction

Received: June 1, 2012
Published: October 23, 2012

ACS Publications © 2012 American Chemical Society

dx.doi.org/10.1021/ci300254w | J. Chem. Inf. Model. 2012, 52, 3074−3087

Class 1

Entry	Properties				
1	*a*	*b*	*c*	*d*	*e*
2	*a*	*b*	*c*	*d*	
3	*a*	*b*	*c*		
4	*a*	*b*	*c*		*e*
5	*a*	*b*		*d*	*e*
6		*b*	*c*	*d*	

Class 2

Entry	Properties				
7	*a*		*c*	*d*	
8			*c*	*d*	*e*
9		*b*		*d*	*e*
10	*a*		*c*		*e*
11	*a*		*c*	*d*	*e*
12		*b*		*d*	

Figure 1. Hypothetical data set containing the emerging pattern {*a*, *c*}.

from toxicity data sets and thus reduce the time and effort required in identifying new structural alerts.

Emerging pattern (EP) mining aims to identify combinations of descriptors that are able to discriminate between classes of objects.[8] Emerging patterns may be extracted from any data that can be expressed as a series of discrete binary properties. A simple example of emerging pattern mining is the identification of patterns of characteristics of mushrooms that discriminate edible mushrooms from poisonous examples. An example of an emerging pattern in the edible class is the set of characteristics {*"no odor" "smooth stalk surface below ring" "one ring only"*}.[9]

Within chemistry, Auer and Bajorath[10] have applied emerging pattern mining to the prediction of biological activity based on physicochemical and molecular properties. Continuous value descriptors were discretized into bits, for example, discrete ranges of molecular weight and log P values. Emerging pattern mining was then used to identify combinations of bits (representing ranges of different physicochemical properties) that are more prevalent in active compounds than in inactives. More recently, Lozano et al.[11] described an approach to identifying "jumping fragments" present in toxicological data. "Jumping" is a term used in emerging pattern mining to indicate a pattern that is exclusive to the objects in one class. Using this terminology, Lozano et al. define a jumping fragment as one that is frequent in the actives while being absent from the inactives. Their method is graph-based and involves enumerating connected subgraphs directly from the active compounds and searching for each independently in the actives and inactives, retaining those that meet a set of defined criteria. This approach is similar to that described by Kazius et al. who enumerate subgraphs and select those which are most discriminative between mutagenic and nonmutagenic compounds.[12] However, Lozano et al.'s method does not make use of the emerging pattern mining algorithms that enable *combinations* of features to be identified.

Emerging pattern mining shares it roots with *formal concept analysis*,[13] otherwise known as *Galois lattice theory*[14] or association rule mining. However, whereas emerging pattern mining is a supervised technique and finds patterns that distinguish one class from another; association rule mining is unsupervised. An association rule is an implication of the form $X \Rightarrow Y$ where X and Y are disjoint subsets of all available items; X is called the antecedent and Y is the consequent. Jullian and Afshar's Knowledge Extraction and Management (KEM) software[15] is based on association rule mining. They describe an application of KEM to a data set of reproductive toxicants and innocuous compounds. The compounds are first fragmented into substructures of various sizes and each compound is encoded as binary properties indicating the presence or absence of the substructural fragments together with its activity, which is also encoded as a binary property. KEM can be used to find association rules that relate substructural fragments to activity by limiting the extracted rules to those containing the presence of activity as the consequent.

In other related work, Nicolaou et al.[16] describe an automated approach to identifying structural motifs associated with the active compounds in an HTS data set. The active compounds are clustered into nodes and a substructure common to the compounds in each node is identified. This process is repeated iteratively to produce a hierarchy of nodes by clustering the compounds in each newly created node and forming new common substructures. The inactive compounds are then used to populate the classes in a postprocessing step. The resulting clusters consisting of both actives and inactives can then be used to derive structure−activity relationships. Harper et al.[17] also describe a data driven clustering method aimed at identifying motifs that are common to actives compounds in an HTS data set. The motifs are based on reduced graphs which are generated for all compounds and are then ranked based on the activity values of the compounds that exhibit them. This approach is not intended as a predictive system and indeed many small active clusters will not be identified, instead it is intended as an aid to identifying gross features in a data set such as compounds synthesized in library design efforts or compounds that interfere with the assay due to the presence of large common motifs.

Here our focus is on using emerging pattern mining to identify structural features that may be associated with toxicity and which can be presented to knowledge base developers for further analysis. We use structural fingerprints as the descriptors and emerging pattern mining methods to identify combinations of bits that are present in the active compounds and absent from the inactives. The use of structural fingerprints enables the active compounds to be organized into hierarchies in which the structural descriptions identified become more detailed as a hierarchy is descended. The resulting hierarchies represent a form of supervised clustering in which the clusters are formed around features present in actives and absent from inactives, thus features that are common to both actives and inactives are ignored in the initial clustering (although they may be identified later). The hierarchies allow the knowledge base developers to browse through a set of compounds in a highly organized way and to choose a level of description that is consistent with any other knowledge that is available. Using structural fingerprints also permits the identification of patterns consisting of disconnected fragments.

Following a brief introduction to emerging pattern mining, we describe how we have adapted published emerging pattern mining algorithms to identify structural features in data sets of chemical

dx.doi.org/10.1021/ci300254w | J. Chem. Inf. Model. 2012, 52, 3074−3087

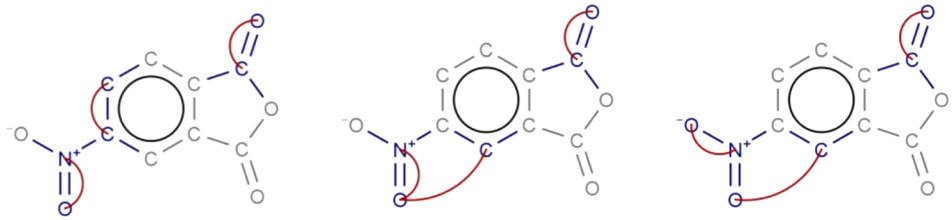
Class 2				
a		*c*	*d*	
		c	*d*	*e*
	b		*d*	*e*
a		*c*		*e*
a		*c*	*d*	*e*
	b		*d*	

→

Enumerated patterns				
a				
	b			
		c		
			d	
				e
a		*c*		
a			*d*	
a				*e*
	b		*d*	
	b			*e*
		c	*d*	
		c		*e*
			d	*e*
a		*c*	*d*	
a		*c*		*e*
a			*d*	*e*
	b		*d*	*e*
		c	*d*	*e*
a		*c*	*d*	*e*

<{minimal}, {maximal}>
<{*a*, *b*, *c*, *d*, *e*}, {*bde*, *acde*}>

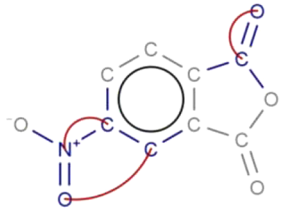
Figure 2. All possible subsets of the entries in the hypothetical class 2. The minimal and maximal borders are shown below, from which all of the subsets can be enumerated.



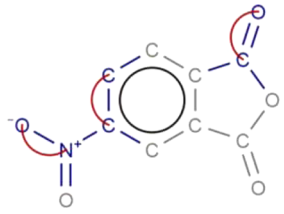
O(1,1)-2-N(3,1)
O(1,1)-2-C(3,1)
C(3,2)-2-C(2,2)

O(1,1)-2-N(3,1)
O(1,1)-2-C(3,1)
O(1,1)-4-C(2,2)

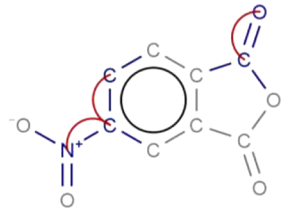
O(1,0)-2-N(3,1)
O(1,1)-2-C(3,1)
O(1,1)-4-C(2,2)



N(3,1)-2-C(3,2)
O(1,1)-2-C(3,1)
O(1,1)-4-C(2,2)



N(3,1)-2-O(1,0)
O(1,1)-2-C(3,1)
C(3,2)-2-C(2,2)



N(3,1)-2-C(3,2)
O(1,1)-2-C(3,1)
C(3,2)-2-C(2,2)

Figure 3. Six minimal JEPs each consisting of three atom pair descriptors and found in the same set of 29 active molecules are shown mapped onto one of the active molecules.

dx.doi.org/10.1021/ci300254w | *J. Chem. Inf. Model.* 2012, 52, 3074−3087

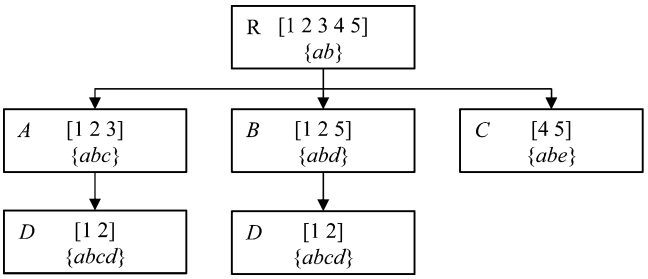


Figure 4. An example support hierarchy for five JEPs and the support sets extracted from the hypothetical example in Figure 1.

structures and how we link the patterns and the sets of active compounds that contain them into hierarchies. We then validate the methods on data sets in which the features associated with toxicity are already known and compare the patterns found with the structural alerts contained within Derek Nexus.

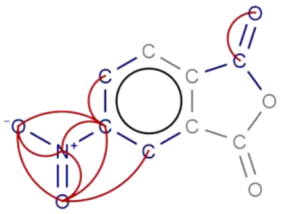
METHODOLOGY

The concept of an emerging pattern is illustrated for a hypothetical data set consisting of data entries in two classes shown in Figure 1, where each class consists of six entries, one per row. Each data entry is composed of a set of up to five binary *properties*, or *items* (*a, b, c, d, e*); if an item is present in a data entry then its label is shown in the corresponding row, otherwise the label is absent. A set of properties of any length (cardinality) is called an item set or pattern. Any pattern that is a proper subset (a subset of fewer items) or is equal to the item set of one or more of the data entries and occurs more frequently in one of the two classes is said to be "emerging" in that class. In Figure 1, the two highlighted items, *a* and *c*, represent an item set that occurs more frequently in data entries in class 1 relative to class 2 so that the pattern {*ac*} is emerging in class 1. Thus, an emerging pattern can be considered to be a characteristic of the entries in one class that distinguishes them from those in the other class, regardless of the discriminatory ability of any individual item in the pattern.

The proportion of data entries in a class that contain a pattern is referred to as its support; the support of pattern, pat, in class *D* is

$$\mathrm{Supp}(D)_{\mathrm{pat}} = \frac{D_{\mathrm{pat}}}{D_{\mathrm{total}}}$$

where D_{pat} and D_{total} are the number of entries in class *D* that contain the pattern and the total number of entries in *D*, respectively. Thus, an emerging pattern is one which has greater support in one class relative to the other. The emerging pattern



C(3,2)-2-C(2,2)
O(1,1)-4-C(2,2)
N(3,1)-2-C(3,2)
O(1,1)-3-C(3,2)
O(1,0)-3-C(3,2)
O(1,1)-2-N(3,1)
O(1,0)-2-N(3,1)
O(1,1)-3-O(1,0)
O(1,1)-2-C(3,1)

Figure 6. Nonminimal JEP consisting of the maximum common set of atom pairs, mapped to a supporting active molecule.

highlighted in Figure 1, {*ac*}, has support 0.67 in class 1, since it occurs in 4 of 6 data entries, whereas it has support of only 0.5 in class 2, since it occurs in 3 of 6 data entries. When an EP is present in only one class, it is known as a jumping-emerging pattern (JEP), i.e. it is jumping and emerging into the class in which it occurs.

Conceptually, the simplest method of identifying EPs and JEPs in one class compared to another is to enumerate all possible item sets in the class of interest and then to search for each item set in the data entries for each class and determine their relative supports. However, this process represents a huge combinatorial problem, even for low dimensional data sets. Dong and Li have developed two methods which have been adopted in this work to speed up the process of identifying JEPs considerably: the border description[8] and the border-differentials method.[18] The border description method allows a complete set of enumerated patterns to be defined from any data set represented by binary properties, without full enumeration of the patterns themselves. In any list of item sets, if an item set is as large as possible while not containing any other complete item set as a proper subset, that item set is said to be *minimal*; conversely, if an item set is as small as possible while not being present as a subset of any larger item set from the list, that item set is *maximal*. The minimal and maximal patterns of a list of item sets are the border descriptions and define the full list of item sets that can be generated from the list. As an example, full enumeration of the entries in the hypothetical class 2 in Figure 1 results in 19 unique patterns as shown in Figure 2: the individual items *a*, *b*, *c*, *d*, and *e* are highlighted in blue and are the minimal patterns

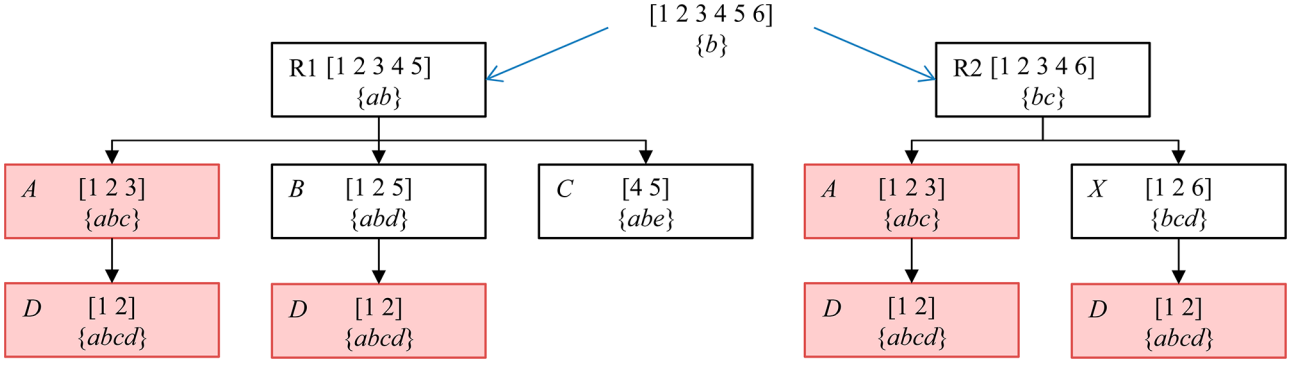


Figure 5. Example of two overlapping hierarchies for seven JEPs and their support sets. The highlighted nodes are common to both trees.

dx.doi.org/10.1021/ci300254w | J. Chem. Inf. Model. 2012, 52, 3074−3087

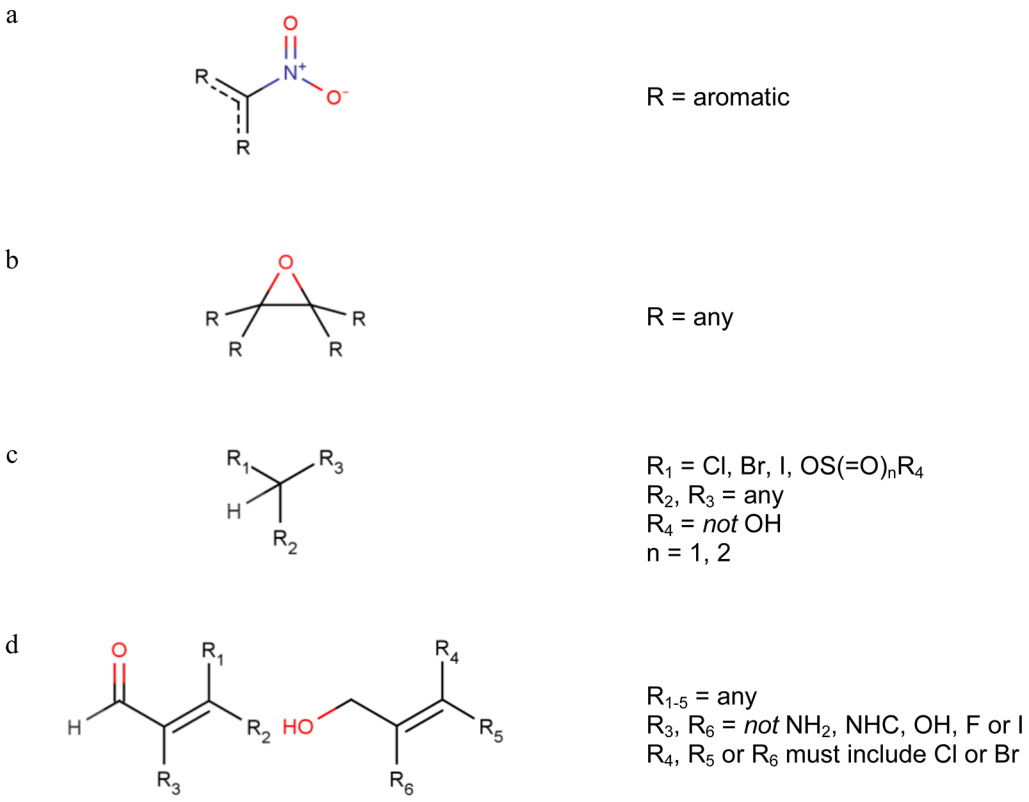


Figure 7. Alerts for Ames compounds: (a) aromatic nitro; (b) epoxide; (c) alkylating agents; (d) α,β-unsaturated aldehydes.

of the set; the item sets $\{bde\}$ and $\{acde\}$ are highlighted in red and are the maximal patterns. The border description in Figure 2 implies that any proper superset of the items d or e will occur in the fully enumerated list of patterns from class 2, providing it is also a subset of either of the item sets $\{bde\}$ or $\{acde\}$. The minimal patterns are the individual items represented in the set and we use the Horizon-Miner algorithm of Li et al.[19] to define the maximal border.

The border differential algorithm identifies the border description of the JEPs in one class by taking the differences in the border descriptions of the two classes. The border description of the JEPs consists of a set of minimal JEPs together with a set of maximal JEPs. If required, the full set of JEPs can be enumerated from these two limits.

Applying JEP Mining to Toxicity Data. We apply the JEP mining algorithm to compound sets in which the items (descriptors) are atom pairs of the form:

$$A(n_A, \pi_A) - X - B(n_B, \pi_B)$$

where A and B represent atoms, n is the number of non-hydrogen atoms connected to the atom; π is the number of π-bonds connected to the atom; and X is the path length measured as the number of atoms between A and B including the two atoms themselves. Given a data set consisting of known actives and inactives, the procedure for mining JEPs is as follows:

1. Generate all atom pairs under user-defined constraints from the active compounds in the data set and form atom pair fingerprints for both the actives and inactives
2. Apply the Horizon-Miner algorithm to extract the maximal patterns for both the actives and the inactives using the atom pair fingerprints
3. Apply the border-differential algorithm to mine the set of all possible minimal JEPs in the actives compared to the inactives
4. Reduce the set of minimal JEPs to those that occur in distinct sets of actives
5. Identify relationships between the supporting actives of minimal JEPs and arrange them into hierarchies
6. Extract the maximum set of commonly occurring atom pairs from the set of actives that support each minimal JEP, to form the largest and most descriptive representation of their common structural features.

In step 1, only those atom pairs that occur in the actives and which meet the defined constraints are put into the JEP mining algorithm. This is because the JEPs represent sets of atom pairs that are unique to the actives; atom pairs that occur only in inactives can only be relevant to the properties of the inactives and not the actives. The path length range used to generate the atom pairs is user-definable. For example, a range of 2−4 would include atom pairs with X = 2, 3, and 4. Also a threshold on absolute occurrence is applied to each atom pair. The rationale for the threshold is that a JEP that is supported by a very small number of actives is unlikely to be of interest and, for a JEP to be above a given support threshold, each atom pair must also occur above the threshold in the active compounds.

Step 4 is required since there can be multiple minimal JEPs that describe (or are supported by) the same set of compounds. For example, consider a hypothetical case in which $\{a\}$, $\{b\}$, and $\{c\}$ are not JEPs in class 1 but $\{ab\}$ and $\{bc\}$ are; then, both $\{ab\}$ and $\{bc\}$ will be minimal even if they are supported by the same set of data entries. This is illustrated in Figure 3 for atom pair JEPs mined from an Ames mutagenicity data set, where each of the six minimal JEPs is supported by the same set of

dx.doi.org/10.1021/ci300254w | J. Chem. Inf. Model. 2012, 52, 3074−3087

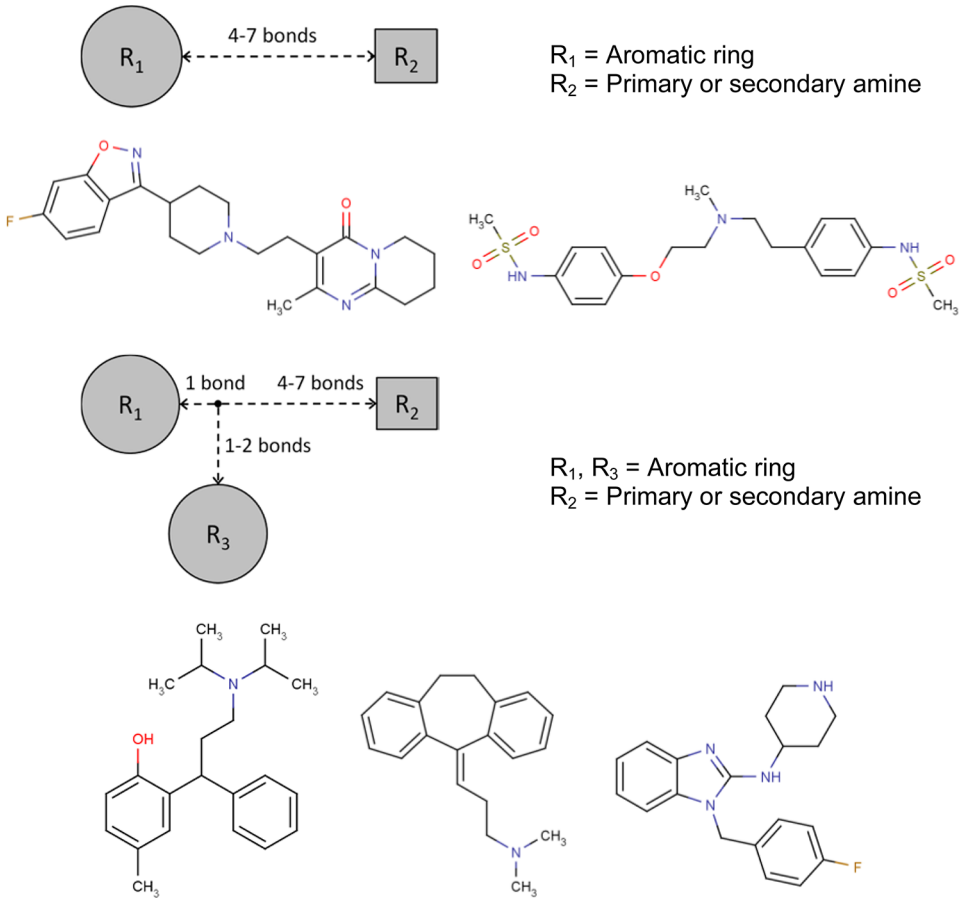


Figure 8. Descriptions of two hERG inhibition toxicophore alerts in Derek Nexus with example actives.

Table 1. Number of JEPs, Hierarchies (Trees), and Run Times for Different Atom Pair Path Length Ranges and Minimum Occurrence Thresholds[a]

		total		support ≥ 6		
	minimum occurrence	JEPs	trees	JEPs	trees	time (mm:ss)
path lengths: 2−4	1%	1041	199	206	40	02:41
	3%	953	174	206	40	01:40
	5%	799	150	195	33	00:41
path lengths: 2−5	1%	1874	268	448	74	46:38
	3%	1758	248	448	74	29:16
	5%	1498	206	418	57	10:54

[a]The columns headed "total" gives the total numbers of JEPs and trees found whereas the columns headed "support ≥ 6" gives the numbers of JEPs and trees supported by at least six actives.

active compounds. The atom pairs in each of the JEPs are shown mapped onto one of the compounds in the set (where an atom pair occurs multiple times, one potential mapping is chosen arbitrarily). Thus step 4 involves comparing the support sets for all minimal JEPs and retaining one minimal JEP only (note that the atom pairs common to all compounds in a support set, whether they form part of a JEP or not, are determined in a later step).

Many JEPs fall into families consisting of subset−superset relationships, with similar relationships existing between the compound sets that support them. In step 5, the JEPs within a family are arranged in a hierarchy with the smallest JEP forming the root node of a tree, together with the compounds that support it. The remaining JEPs are placed below the root node in order of increasing size. As a tree is descended, the JEPs become larger and therefore more descriptive while the compounds that support them become fewer. The formation of hierarchies is illustrated in Figure 4 based on the hypothetical example in Figure 1: the tree is composed of five JEPs from class 1 arranged in a tree of six nodes. Within the tree, each node represents a JEP and includes the data set entries that support that pattern, indicated using brackets, e.g. [1 2 3]. The JEP is shown below the support set and is indicated using braces, e.g. {*abc*}.

The hierarchies resulting from this process represent a form of *supervised clustering*, i.e. the clustering is based on structural features that only occur in the active compounds. The clusters found using JEPs are overlapping since a molecule can be present in more than one node of a hierarchical tree, and in more than one hierarchy. This is distinct from conventional *hierarchical clustering* methods, which do not typically result in overlapping clusters or multiple hierarchies. An example of two

dx.doi.org/10.1021/ci300254w | J. Chem. Inf. Model. 2012, 52, 3074−3087

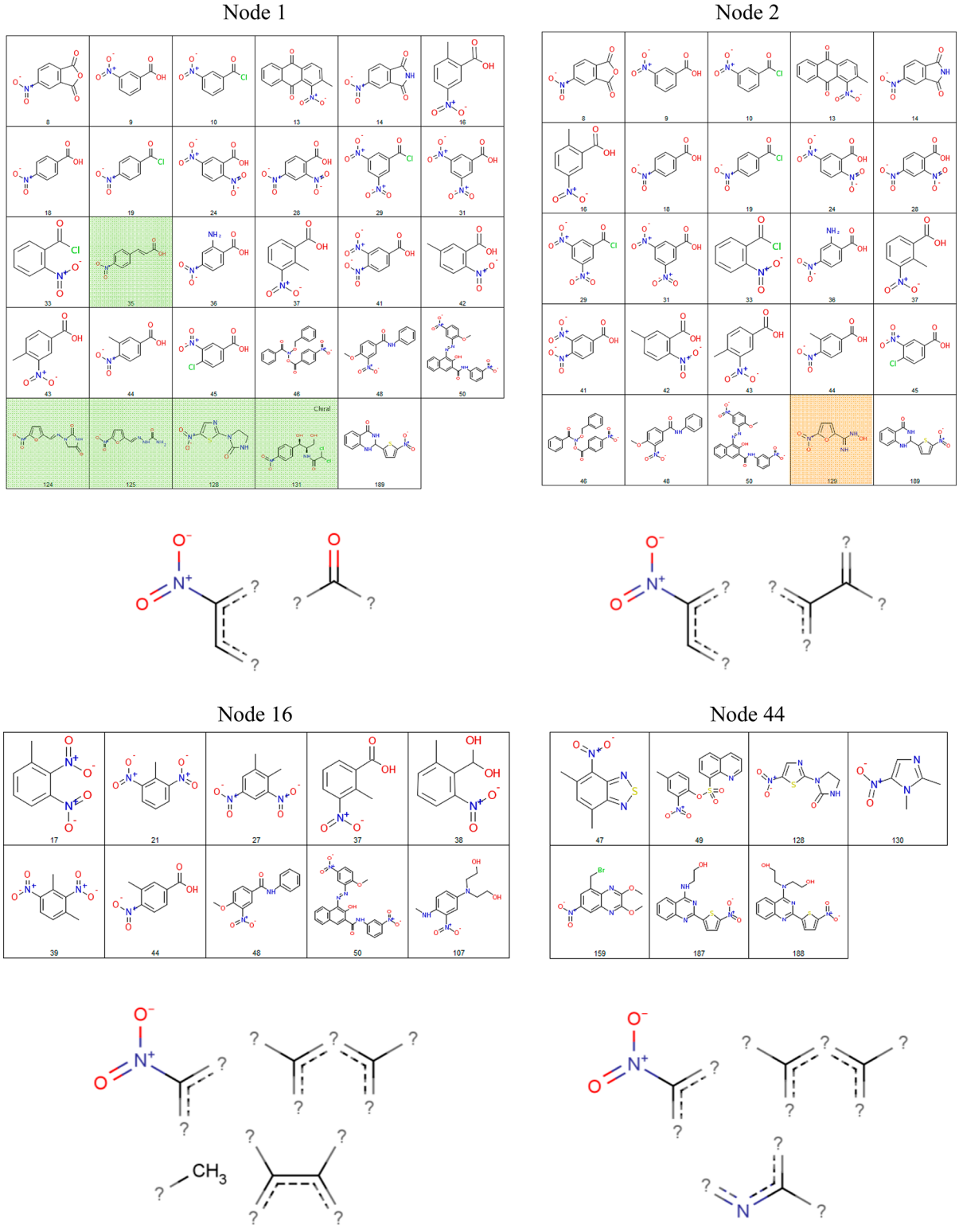


Figure 9. The JEPs and support sets for four root nodes that describe aromatic nitro compounds. The atom pairs comprising the JEPs are illustrated as substructures below each set of compounds. There is significant overlap between root nodes 1 and 2. The compounds that are unique to each cluster are shaded. The separation of aromatic nitro compounds into multiple overlapping nodes reflects the presence of aromatic nitro groups within the inactives—see text for further details.

overlapping trees is shown in Figure 5. Seven hypothetical JEPs labeled R1, R2, *A*, *B*, *C*, *D*, and *X*, and composed of abstract properties, *a*, *b*, *c*, *d*, and *e*, are shown arranged in two trees, one of six nodes and the other of five nodes. The tree with the root node R1, previously shown in Figure 4, shares the three highlighted nodes with the tree with root node R2. These three common nodes consist of two distinct JEPs; {*abc*} and {*abcd*}.

The final step (step 6) involves expanding the set of atom pairs in each of the minimal JEPs to include all atom pairs that are common

dx.doi.org/10.1021/ci300254w | *J. Chem. Inf. Model.* 2012, 52, 3074−3087

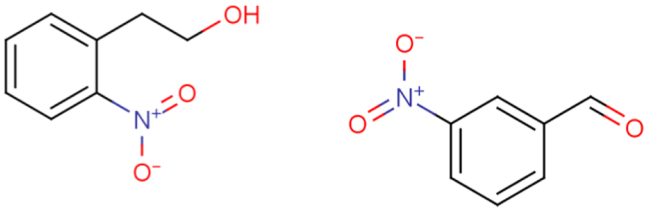


Figure 10. Two aromatic nitro compounds in the set of inactives.

to the supporting compounds. This is achieved by applying a bitwise AND operation to the atom pair fingerprints of the supporting compounds. As an example, there are nine atom pairs that are common to all 29 compounds that support each JEP in Figure 3, and this set includes the atom pairs in all six minimal JEPs together with three additional atom pairs, as shown in Figure 6.

Knowledge Extraction. Having arranged the compounds into hierarchical sets, the JEPs and their associated compounds can be browsed by knowledge base developers to assist in the time-consuming process of identifying new alerts. This would typically involve the identification of structural hypotheses based on a combination of the structural features identified in the data and prior knowledge. A search of the literature would then be made to identify compounds that fit the hypothesis and that either support or contradict it. If a biological mechanism for toxicity can be established, the hypothesis could lead to a new structural alert being formed, or an existing alert being extended, and the expansion of the knowledge base. The details of how this procedure would be carried out in practice will be described in a subsequent paper aimed at the toxicology community.

EXPERIMENTAL SECTION

Data Sets. The JEP mining has been applied to three data sets: an Ames mutagenicity data set; an oestrogen data set; and a hERG data set.

The Ames mutagenicity test[20] is an in vitro bacterial assay that measures the ability of a compound to cause mutations in several different strains of *Salmonella typhimurium*. Mutagenicity is considered to be an early alert for carcinogenicity. It is a well studied end point in toxicity prediction.[12,21,22] The Ames data set selected here consists of 195 active compounds (comprising alkylating agents; α,β-unsaturated aldehydes; epoxides; and aromatic nitro compounds) and 424 inactive compounds extracted from the Lhasa Limited Vitic 4 database. Figure 7 shows the set of alerts that correspond to these indicators of Ames activity where it can be seen that the activating features are believed to be relatively small substructural fragments.

Oestrogenicity is known to result from interactions with specific binding receptors and is dependent on more complex and often longer ranging structural features than give rise to Ames mutagenicity. The oestrogen data set was obtained from the Distributed Structure-Searchable Toxicity (DSSTox) network, hosted by the US EPA.[23] The data set is composed of 232 compounds with 131 classed as active (ER_RBA+) and 101 classed as inactive (ER_RBA−) based on oestrogen receptor binding affinity data, obtained through in vitro assay experiments. This data set has also been examined by Jullian and Afshar[15] using their closely related Galois lattice mining method.

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.[24] Molecules that bind to and block the channel can lead to arrhythmia of the heart with potentially fatal consequences. The structural features giving rise to hERG activity are thought to consist of both small well-defined activating substructures and features at relatively long distance separation. The knowledge base of Derek Nexus includes a number of alerts for hERG inhibition. However, in most cases, Derek Nexus bases its hERG predictions for the data set used in this study on the two

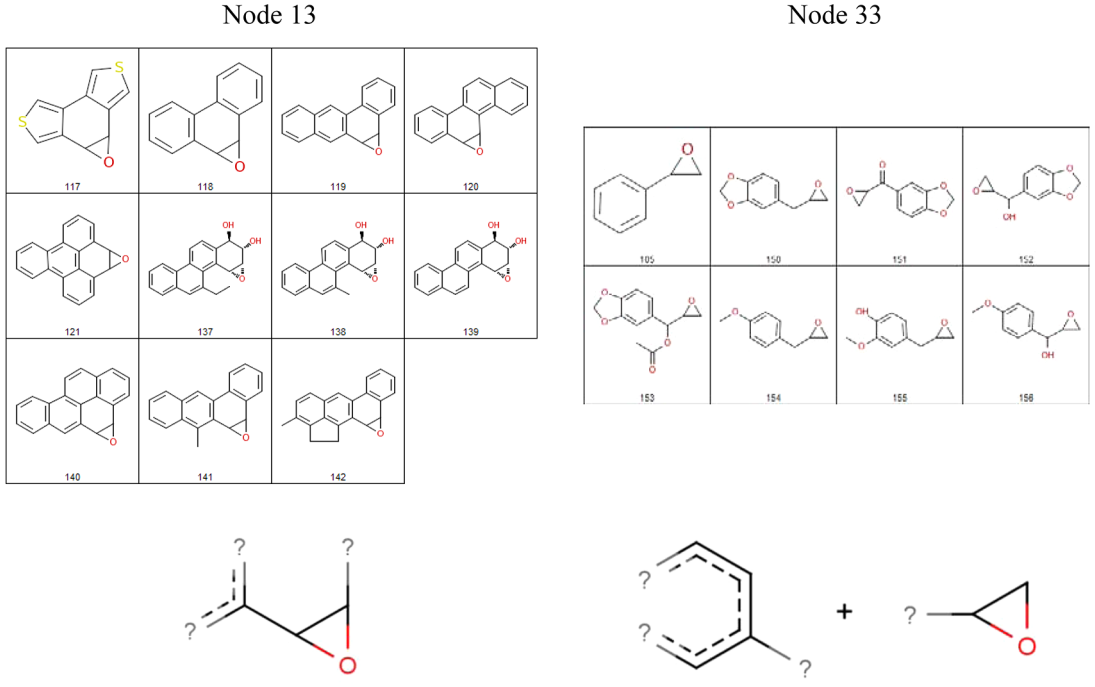


Figure 11. Supporting active compounds (top) and JEPs (bottom) for two root nodes representing epoxides. The atom pairs of the JEPs are illustrated as substructural fragments.

dx.doi.org/10.1021/ci300254w | J. Chem. Inf. Model. 2012, 52, 3074−3087

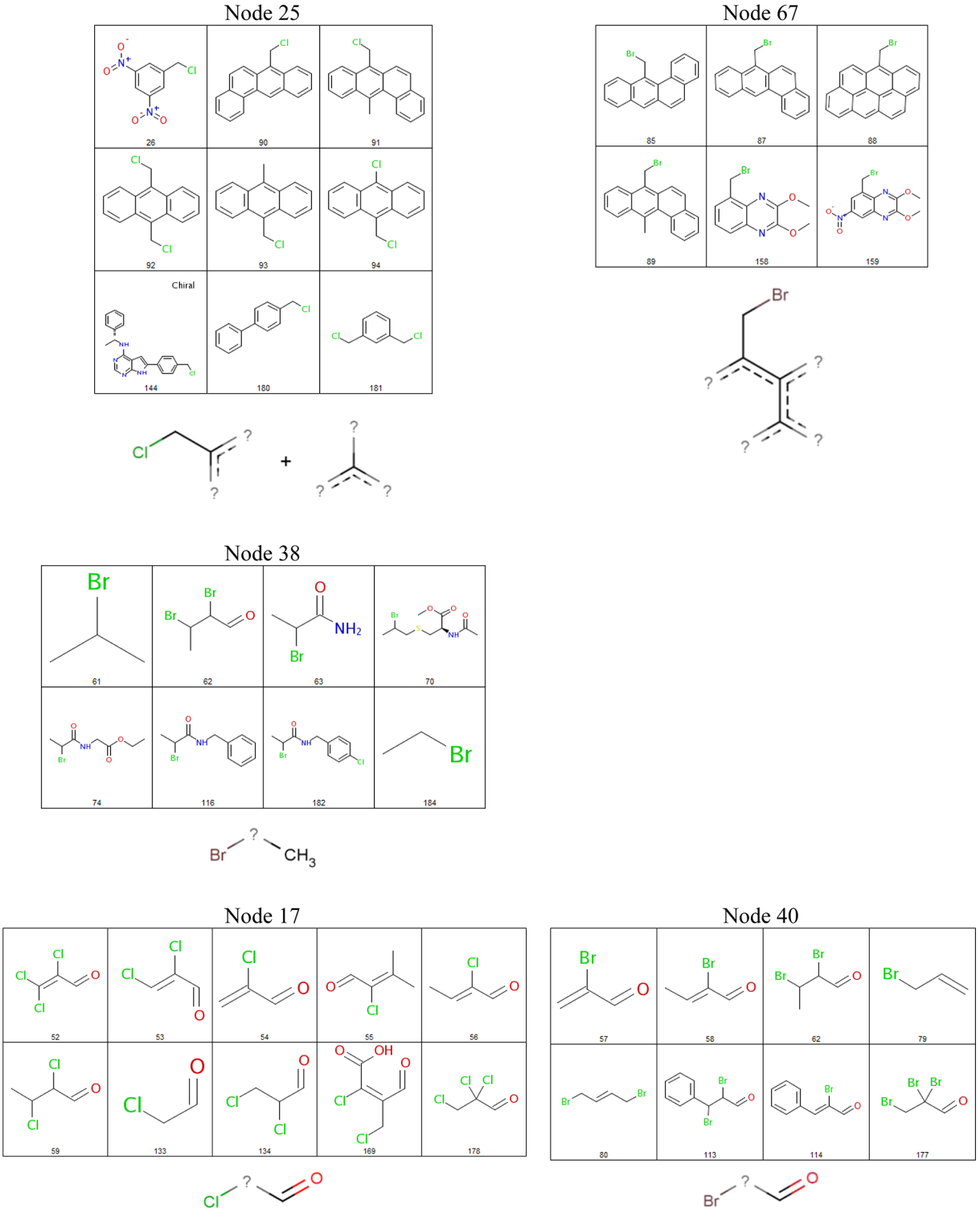


Figure 12. JEPs (bottom) and supporting active compounds for root nodes that represent alkylating agents.

alerts summarized in Figure 8, together with example compounds which exhibit the relevant alert and are known to be active. The hERG inhibition data set consists of 148 compounds extracted from Lhasa Limited's Vitic database. The compounds were classified using an IC_{50} threshold of 20 μM: compounds with IC_{50} values below 20 μM were classified as hERG active; while those with values of 20 μM or above were classified as inactive. This resulted in 114 active molecules and 34 inactives.

RESULTS

Ames Mutagenicity. JEPs and support hierarchies were mined from the Ames data set using atom pairs with two different path length ranges, 2−4 and 2−5, and at three different minimum occurrence thresholds, 1%, 3%, and 5%. The number of unique JEPs and support hierarchies produced, and the time taken to mine the JEPs, are shown in Table 1. The columns headed "total" refer to the total number of JEPs and

dx.doi.org/10.1021/ci300254w | J. Chem. Inf. Model. 2012, 52, 3074−3087

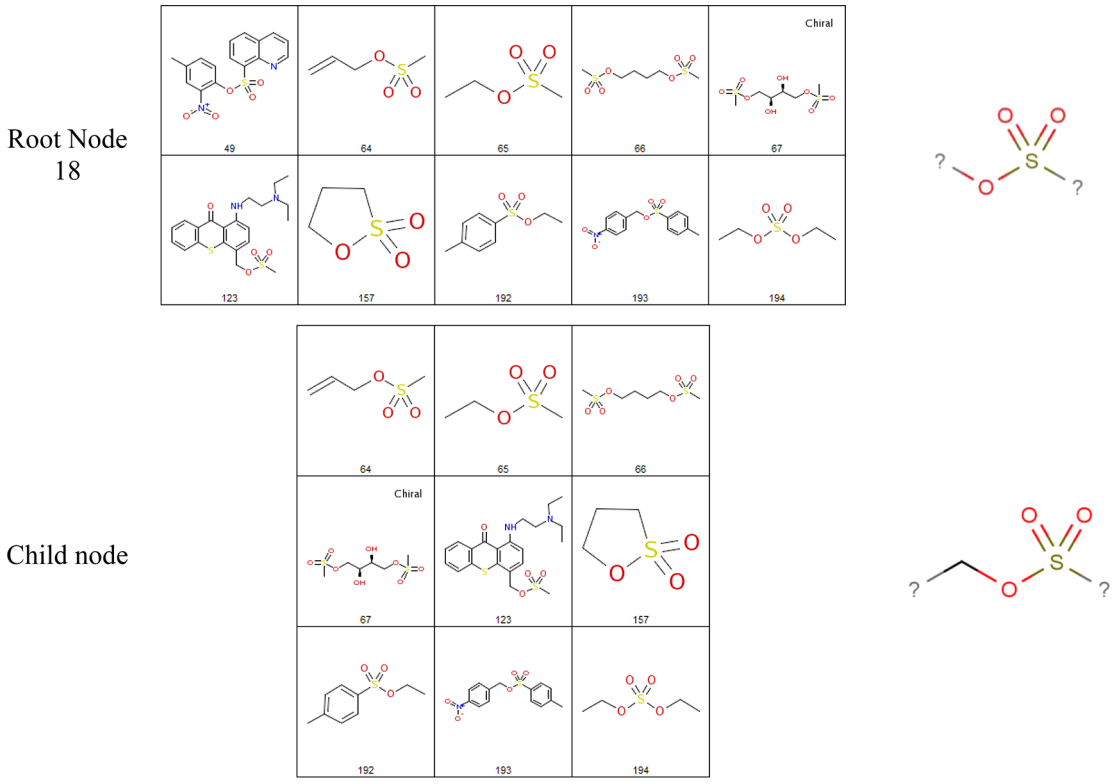


Figure 13. Node 18, a root node, supported by mostly alkylating agents. The JEP of the child node encodes an aliphatic carbon adjacent to the ester oxygen and excludes the aromatic compound, number 49, which is nonalkylating due to the moderating effect of the aromatic ring adjacent to the ester oxygen.

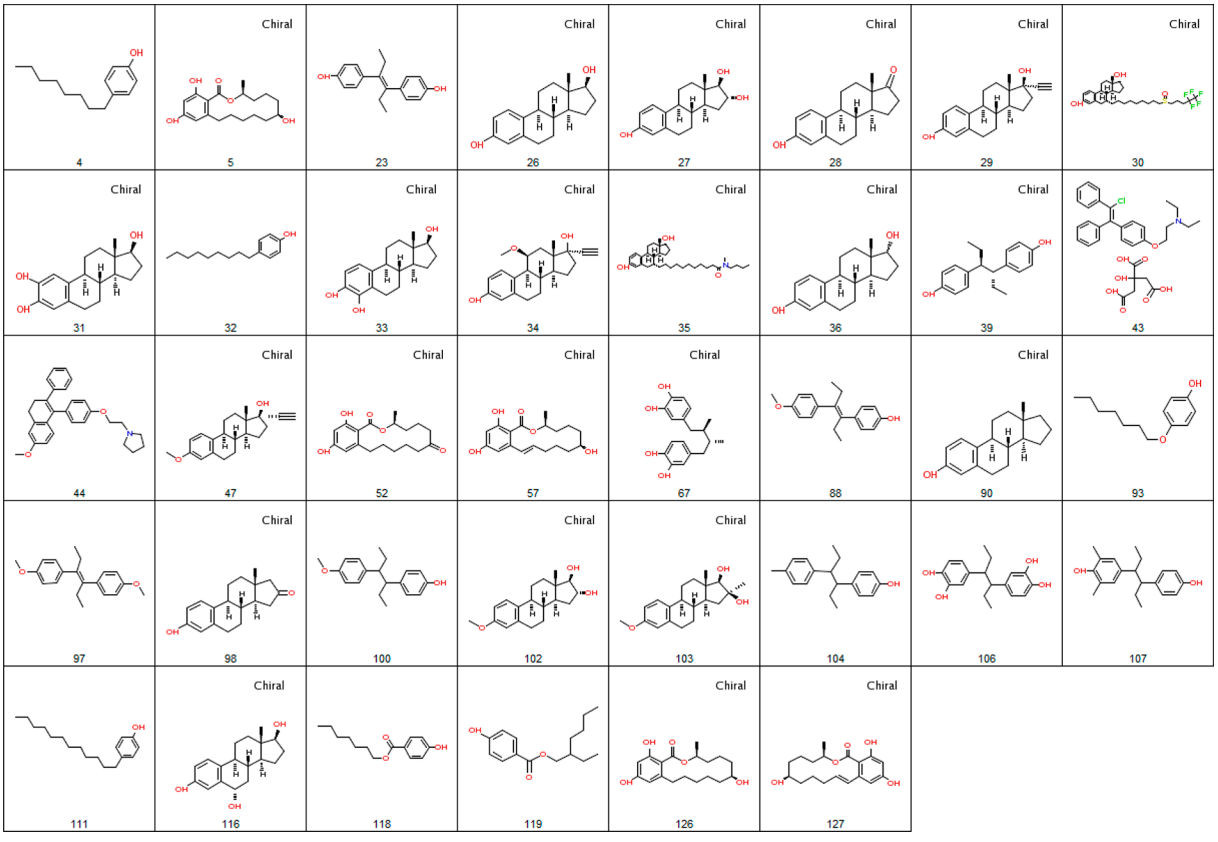


Figure 14. Compounds represented by the most supported root node identified for the oestrogen compounds.

dx.doi.org/10.1021/ci300254w | J. Chem. Inf. Model. 2012, 52, 3074−3087

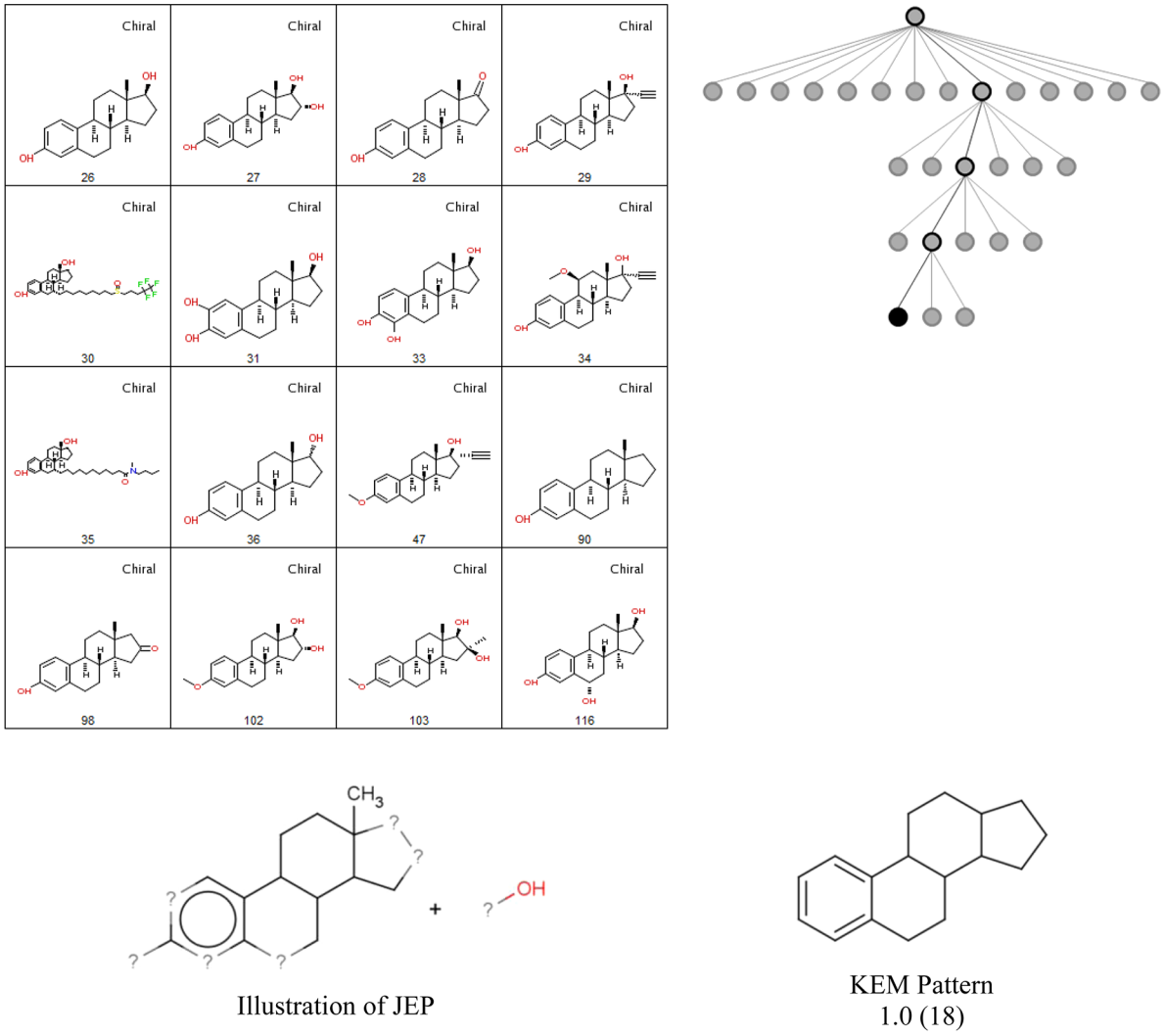


Figure 15. JEP that represents steroidal ring systems shown together with the compounds that support it. The graph shows the relationship of this node to the root node. The corresponding KEM association rule is shown bottom right. The substructure is present in 18 of the active compounds and none of the inactives, so according to KEM the probability of the substructure being associated with activity is 1.0.

trees (or hierarchies) identified, whereas the columns headed "support ≥ 6" refer to the number of JEPs that are supported by at least six active compounds. As the minimum occurrence threshold on the atom pairs is increased, the number of JEPs decreases as does the time taken to find them. This effect is most noticeable when the total number of JEPs is considered. The initial threshold that is set for atom pairs determines also the minimum potential value for final support: so processing time can be reduced by setting the initial threshold on atom pairs to the minimum acceptable for final support. Increasing the range of atom pairs increases the number of JEPs found and increases the time taken considerably. This is simply due to the increased number of descriptors.

The results described below are for atom pairs 2−5; minimum occurrence of 3%; and JEP support ≥ 6. Note that a JEP can have lower support than any of its component atom pairs so that although each atom pair must occur in at least six compounds (at 3% of 195) without a threshold on support JEPs could be found which occur in fewer than six compounds and which are not likely to be of interest. These parameters result in a total of 448 JEPs arranged into 74 trees. There is, however, considerable overlap between the JEPs (and their corresponding support sets).

Figure 9 shows the root nodes of four trees that represent structural variations of aromatic nitro compounds: the molecules that support each JEP are shown together with a substructural pattern that has been composed to represent the atom pairs in the JEP. JEPs are restricted to patterns that occur only in active compounds, and so numerous JEPs are formed that describe variations in the structural features surrounding the nitro group in the active molecules that are not present in the inactives. Figure 10 shows two inactive aromatic nitro compounds; the molecule on the right of Figure 10 contains a carbonyl group with only one heavy neighbor, rather than the secondary carbonyl group included in the JEP of root node 1.

Root node 1 is the highest supported JEP, and there is a large overlap in the actives that support root nodes 1 and 2; the active molecules that are unique to each set are shaded. There is overlap because the respective JEPs describe similar, fairly generic, structural features. The support sets of root nodes 16 and 44 are much smaller, with the corresponding JEPs describing more specific structural features. The JEP of root node 16 describes a point of substitution ortho to the nitro

dx.doi.org/10.1021/ci300254w | J. Chem. Inf. Model. 2012, 52, 3074−3087

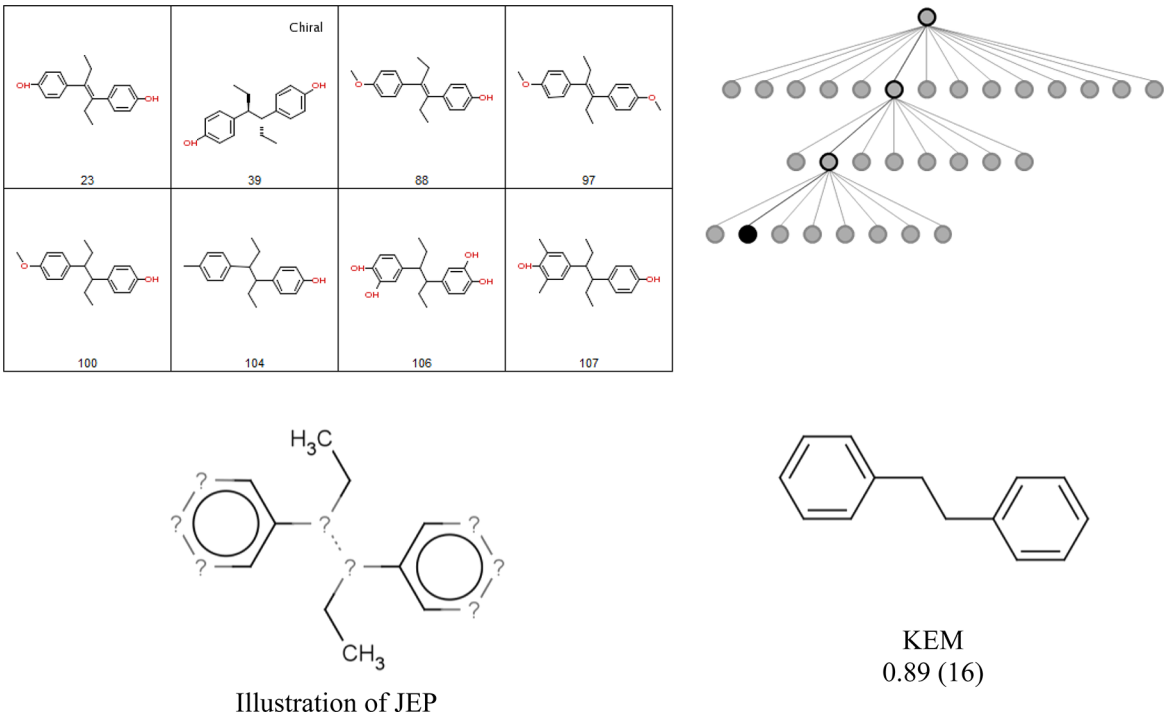


Figure 16. JEP that represents dihydroxydiphenyl-ethane and -ethene compounds shown together with the compounds that support it. The graph shows the relationship of this node to the root node. The corresponding KEM association rule is shown bottom right. The substructure is present in 16 of the active compounds and 2 of the actives, so according to KEM the probability of the substructure being associated with activity is 0.89.

group with a further meta or para substituent. The JEP also describes the presence of a terminal carbon that is mostly found as a methyl substituent on the ring. The JEP of root node 44 describes the presence of at least one aromatic nitrogen atom, in addition to details of ring substituents. In all cases, the most significant features of the aromatic nitro group alert are included, together with additional atom pairs that are required to fully distinguish active aromatic nitro compounds from inactives.

Figure 11 shows the two largest supported root nodes that describe epoxide containing compounds. Node 13 represents molecules where both carbons of the epoxide are substituted with one of the substituents being an aromatic atom. Node 33 represents molecules with a single substitution on the epoxide ring.

Figure 12 shows five root nodes that represent alkylating agents. Nodes 25 and 67 include quite similar structures with the main difference being the halogen atom. In both cases their associated JEPs include details of aromatic rings. The JEP of node 38 is very generic and simply describes a bromine atom two bonds away from a methyl group.

Figure 13 shows a further example of a root node composed largely of alkylating agents. The JEP describes a sulfonic ester group which is present in the Derek Nexus alert. However, the node also contains a molecule where the ester group is adjacent to an aromatic ring. A toxicologist would likely determine that this molecule is not an alkylating agent due to the moderating effect of the aromatic ring's adjacency to the ester oxygen atom. Although the JEP at the root node does not contain sufficient structural information to exclude the nonalkylating agent, the JEP of a child node of the root extends the substructure to include an aliphatic carbon adjacent to the ester oxygen which does therefore exclude the aromatic compound. In this case, a node below the root node maps more closely to the Derek Nexus alert and forms a better representation the known sulfonic ester toxicophore.

For the Ames data set, the JEP mining has been successful in producing clusters of compounds that are representative of the Derek Nexus structural alerts. Although a relatively large number of nodes are produced, in most cases it is the root level nodes that map most closely to the alerts and it is not necessary for a user to traverse the whole tree. In some cases, however, the more detailed descriptions below the root are more representative of the alerts. The presence in the inactives of key parts of alerts associated with activity, for example, aromatic nitro compounds and epoxide compounds, has resulted in longer patterns being identified by the algorithm that take account of the different environments in the actives. This level of detail might be difficult to see manually in a large data set.

Oestrogenicity. JEP mining was based on atom pairs of length 2−5. The root node with largest support represents 38 active compounds as shown in Figure 14. The associated JEP is composed of five atom pairs only, that describe aromatic and aliphatic carbon atoms at various path lengths and which are not sufficient to separate the different structural classes contained within the cluster. However, this node is the root of a hierarchical tree with the known different structural classes emerging as the tree is descended. Figure 15 shows a node consisting of a set of steroidal compounds, with the relationship of this node to the root illustrated on the right, and Figure 16 shows a set of 4,4′-dihydroxydiphenyl-ethane and -ethene compounds together with the relationship of this node to the root node.

This data set was also analyzed by Jullian and Afshar using their KEM program and some of the association rules that relate substructural fragments to activity are also shown. In Figure 15, the substructure that forms the association rule occurs in 18 molecules, and the probability of the rule being

dx.doi.org/10.1021/ci300254w | J. Chem. Inf. Model. 2012, 52, 3074−3087

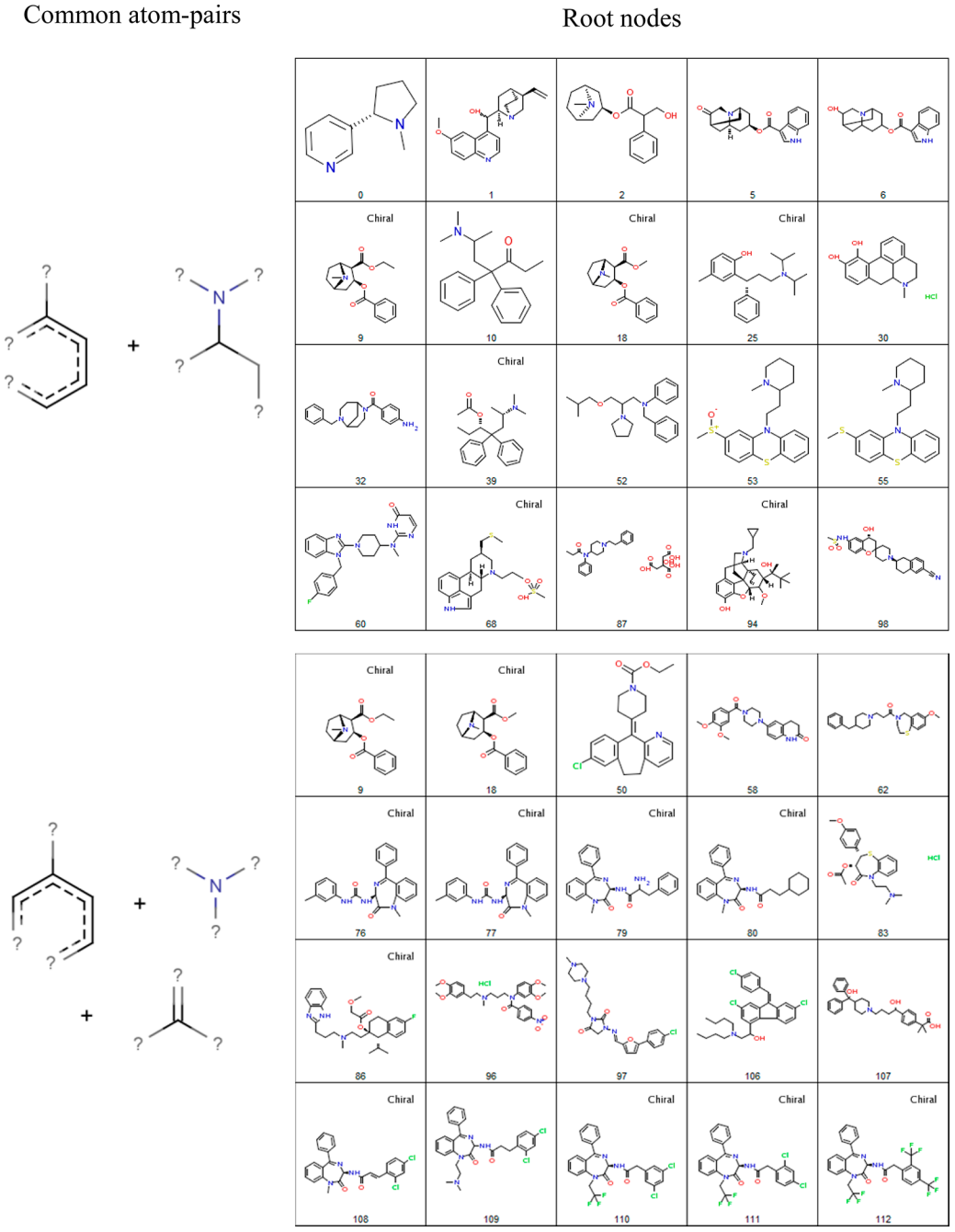


Figure 17. JEPs (left) and supporting active compounds (right) for two root nodes representing hERG channel inhibitors. The atom pairs of the JEPs are illustrated as substructural fragments.

associated with activity is equal to 1.0, indicating that it is absent from the inactives, that is, this fragment can be considered as a "jumping" fragment. The corresponding JEP includes features in addition to the steroidal ring system, namely the terminal methyl and the hydroxy group, and illustrates an advantage of mining and constructing structural features from a diverse set of small components (atom pairs) in a data driven process, rather than from predefined substructural fragments. In Figure 16, the corresponding KEM pattern occurs in 18 molecules of which 16 are actives. The probability of the rule being associated with activity is 0.89. This example demonstrates an advantage of association rule mining over JEP mining in which rules are identified that may also be present in inactives and are therefore more resistant to noise within the data.

For the oestrogenicity data set, the JEP mining has produced clusters of compounds that are similar to those produced by some of the association rules mined by Jullian and Afshar using KEM. A larger number of nodes are produced than for Ames mutagenicity, and few of the JEPs at the root level are capable of separating different structural classes. However, separation of these classes is achieved by more detailed JEPs further down the trees.

hERG Channel Inhibition. The JEP mining was based on atom pairs of length 2−8. The long-range descriptors were included because the known toxicophore alerts indicated that there could be up to eight atoms separating the aromatic ring and the trivalent nitrogen atom. Increasing the range of the atom pair descriptors typically leads to a significant increase in the time taken to mine the JEPs due to the increased number of

dx.doi.org/10.1021/ci300254w | J. Chem. Inf. Model. 2012, 52, 3074−3087

descriptors, as well as an increase in the number of JEPs identified. The data set is highly skewed toward actives with more than three active compounds for every inactive. So a threshold of 17% (equivalent to 20 compounds) was applied when generating the initial set of atom pairs.

The root node with largest support represents 65 of the active compounds. However, the JEP for this node did not distinguish between the different structural families within the actives: it consists of six atom pairs which describe a series of aromatic and aliphatic carbon atoms at varying numbers of bond distances and does not include the crucial nitrogen atom since the distance from the aromatic ring varies among the compounds in the set. Figure 17 shows two root nodes that have much lower support, but their JEPs successfully represent some of the key structural features of the alerts, including the trivalent nitrogen atom. Both JEPs include a number of longer atom pairs that link together the fragments shown in Figure 17. This provides some indication of the range of distances that are observed between the fragments.

For the hERG channel inhibition data set, the JEP mining has successfully produced clusters of compounds and JEPs that are representative of the crucial features of both of the Derek Nexus structural alerts. However, only poor separation of structural classes is achieved, even when traversing down the trees. Indeed, separation remains consistently poorer than for oestrogenicity, as the structural classes of hERG inhibitors are much harder to define due to the diversity of the input data. Overall, mining JEPs for hERG inhibition represents a middle-ground between Ames mutagenicity and oestrogenicity. Atom pair JEPs are capable of describing the alerts for hERG inhibition, as they are for Ames mutagenicity, but supervised clustering performed badly in this case because so few inactives were available for comparison.

■ CONCLUSIONS

We have developed a tool aimed at knowledge discovery that can assist knowledge base developers in the time-consuming process of compiling structural alerts for use in an expert system for toxicity prediction. We have validated the approach on data sets in which the features giving rise to the toxic effects are well understood. Although a large number of structural patterns can be generated, they are arranged hierarchically with a strict subset-superset relationship existing between nodes at different levels in the hierarchy, in terms of atom pairs included in the JEPs and the molecules that support them. The approach is not intended to generate definitive toxicological alerts wholly automatically, but to provide a tool for experts who are exploring data in order to develop alerts. Our future work is focused on mining emerging patterns as distinct from jumping emerging patterns, that is, patterns which are prevalent in the actives but which may also occur in inactives, albeit with lower frequencies.

■ AUTHOR INFORMATION

Corresponding Author

*Author to whom correspondence should be addressed. Tel.: +44-1142-222652. Fax:+44-1142-780300. E-mail: v.gillet@sheffield.ac.uk.

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Cronin, M. T. D.; Madden, J. C. *In Silico Toxicology: Principles and Applications*; Royal Society of Chemistry: Cambridge, UK, 2010.
(2) Greene, N. Computer systems for the prediction of toxicity: an update. *Adv. Drug Delivery Rev.* **2002**, *54*, 417−431.
(3) Dearden, J. C. *In silico* prediction of drug toxicity. *J. Comput-Aided Mol. Des.* **2003**, *17*, 119−127.
(4) Marchant, C. A. Computational toxicology: a tool for all industries. *WIREs Comput. Mol. Sci.* **2012**, *2*, 424−434.
(5) Lowe, R.; Glen, R.; Mitchell, J. Predicting Phospholipidosis Using Machine Learning. *Mol. Pharmaceutics* **2010**, *7*, 1708−1714.
(6) Simon-Hettich, B.; Rothfuss, A.; Stager-Hartmann, T. Use of computer-assisted prediction of toxic effects of chemical substances. *Toxicology* **2006**, *224*, 156−162.
(7) *Derek Nexus*, Lhasa Limited: Leeds, 2012.
(8) Dong, G.; Li, J. In Efficient mining of emerging patterns: discovering trends and differences, *The Fifth International Conference on Knowledge Discovery and Data Mining*, San Diego, 15−18 August 1999; Association for Computing Machinery Press: San Diego, 1999; pp 43−52.
(9) Dong, G.; Zhang, X.; Wong, L.; Li, J. In CAEP: classification by aggregating emerging patterns. *Second International Conference on Discovery Science (Discovery Science '99)*, Tokyo, Japan, 6−8 December 1999; Springer: Tokyo, Japan, 1999.
(10) Auer, J.; Bajorath, J. Emerging chemical patterns: a new methodology for molecular classification and compound selection. *J. Chem. Inf. Model.* **2006**, *46*, 2502−2514.
(11) Lozano, S.; Poezevara, G.; Halm-Lemeille, M. P.; Lescot-Fontaine, E.; Lepailleur, A.; Bissell-Siders, R.; Crémilleux, B.; Rault, S.; Cuissart, B.; Bureau, R. Introduction of jumping fragments in combination with QSARs for the assessment of classification in ecotoxicology. *J. Chem. Inf. Model.* **2010**, *50*, 1330−1339.
(12) Kazius, J.; Nijssen, S.; Kok, J.; Bäck, T.; Ijzerman, A. P. Substructure mining using elaborate chemical representation. *J. Chem. Inf. Model.* **2006**, *46*, 597−605.
(13) Wolff, K. E. In A first course in formal concept analysis - How to understand line diagrams. *The 7th Conference on the Scientific Use of Statistical Software*, Heidelberg, Germany, 14−18 March 1993; Gustav Fischer Verlag: Heidelberg, Germany, 1993; pp 429−438.
(14) Carpineto, C.; Romano, G. In Galois: an order-theoretic approach to conceptual clustering. *Tenth International Conference on Machine Learning*, Amherst, MA, USA, 27−29 June 1993; Morgan Kaufmann: Amherst, MA, USA, 1993; pp 33−40.
(15) Jullian, N.; Afshar, M. Novel rule-based method for multi-parametric multi-objective decision support in lead optimization using KEM. *Curr. Comput.-Aided Drug Des.* **2008**, *4*, 35−45.
(16) Nicolaou, C. A.; Tamura, S. Y.; Kelley, B. P.; Bassett, S. I.; Nutt, R. F. Analysis of large screening data sets via adaptively grown phylogenetic-like trees. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1069−1079.
(17) 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.
(18) Dong, G.; Li, J. Mining border descriptions of emerging patterns from dataset pairs. *Knowl. Inf. Sys.* **2005**, *8*, 178−202.
(19) Li, J.; Dong, G.; Ramamohanarao, K. Making use of the most expressive jumping emerging patterns for classification. *Knowl. Inf. Sys.* **2001**, *3*, 131−145.
(20) Mortelmans, K.; Zeiger, E. The Ames Salmonella/microsome mutagenicity assay. *Mutat. Res. Fundam. Mol. Mech. Mugag.* **2000**, *455*, 29−60.
(21) Langham, J. J.; Jain, A. N. Accurate and interpretable computational modeling of chemical mutagenicity. *J. Chem. Inf. Model.* **2008**, *48*, 1833−1839.
(22) Mazzatorta, P.; Tran, L.-A.; Schilter, B.; Grigoroz, M. Integration of Structure-Activity Relationship and Artificial Intelligence Systems To Improve in Silico Prediction of Ames Test Mutagenicity. *J. Chem. Inf. Model.* **2007**, *47*, 34−38.
(23) *OncoLogic*, 7.0; United States Environmental Protection Agency: Washington DC, 2012.
(24) Sanguinetti, M. C.; Tristani-Firouzi, M. hERG potassium channels and cardiac arrhythmia. *Nature* **2006**, *440*, 463−469.