JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Discovering Structural Alerts for Mutagenicity Using Stable Emerging Molecular Patterns

Jean-Philippe Métivier,[†,‡] Alban Lepailleur,[†,§] Aleksey Buzmakov,[∥,⊥] Guillaume Poezevara,[†,‡,§] Bruno Crémilleux,[†,‡] Sergei O. Kuznetsov,[⊥] Jérémie Le Goff,[#] Amedeo Napoli,[∥] Ronan Bureau,[†,§] and Bertrand Cuissart*[,†,‡]

[†]Normandie Université, Caen, France

[‡]UNICAEN, GREYC, UMR CNRS 6072, F-14032 Caen, France

[§]UNICAEN, CERMN, UPRES EA 4258, FR CNRS 3038, F-14032 Caen, France

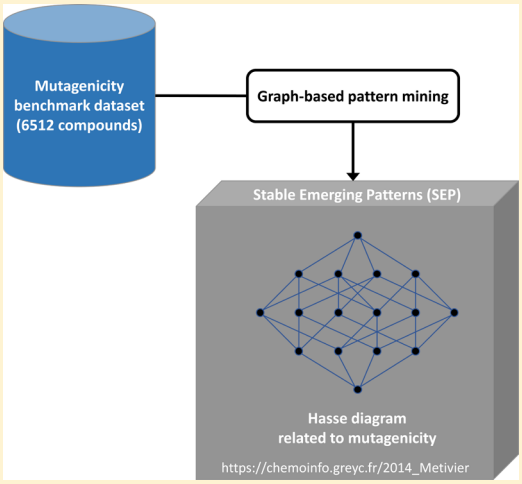
[∥]Laboratoire Lorrain de Recherche en Informatique et ses Applications (LORIA), University of Lorraine, Nancy, France

[⊥]National Research University Higher School of Economics (HSE), Moscow, Russia

[#]ADn'tox, Caen, France

Supporting Information

ABSTRACT: This study is dedicated to the introduction of a novel method that automatically extracts potential structural alerts from a data set of molecules. These triggering structures can be further used for knowledge discovery and classification purposes. Computation of the structural alerts results from an implementation of a sophisticated workflow that integrates a graph mining tool guided by growth rate and stability. The growth rate is a well-established measurement of contrast between classes. Moreover, the extracted patterns correspond to formal concepts; the most robust patterns, named the stable emerging patterns (SEPs), can then be identified thanks to their stability, a new notion originating from the domain of formal concept analysis. All of these elements are explained in the paper from the point of view of computation. The method was applied to a molecular data set on mutagenicity. The experimental results demonstrate its efficiency: it automatically outputs a manageable number of structural patterns that are strongly related to mutagenicity. Moreover, a part of the resulting structures corresponds to already known structural alerts. Finally, an in-depth chemical analysis relying on these structures demonstrates how the method can initiate promising processes of chemical knowledge discovery.



INTRODUCTION

In the pharmaceutical industry, it is widely recognized that early safety evaluation of candidate molecules is needed before significant investments of time and resources are made.[1,2] To this aim, the notion of predictive toxicology, which includes the application of computer technologies to detect relationships that connect chemical structures and toxicological activities in large biological and chemical data sets, is very appealing. The advantages of in silico techniques in comparison with in vitro and in vivo techniques can be summarized by their higher throughput, their cost effectiveness, and their potential to reduce the use of animals. In a regulatory framework, the use of toxicity prediction tools is encouraged to improve prioritization of data requirements and risk assessment not only for pharmaceuticals[3,4] but also for other chemical products such as cosmetics and agrochemicals.[5,6] In silico prediction methods can roughly be classified into two categories: knowledge-based expert systems and data-driven models. On the one hand, knowledge-based expert systems such as Derek Nexus,[7,8] HazardExpert,[9] and OncoLogic[10,11] do not discover new associations between chemicals and toxicity but rather formalize the knowledge of human experts and the scientific literature. On the other hand, data-driven models such as MultiCASE,[12] Topkat,[13] LAZAR,[14] and PASS[15] analyze existing data, identify chemical features that are relevant for the observed toxicological end points, and automatically build statistical models.

The definition of structural alerts corresponds to one of the most interesting approaches of predictive toxicology since it defines the key features of a molecule that are required to interact with a biological system and initiate a toxicology pathway. Its main advantage is the identification of chemicals with a common mechanism of action. The set of structural

Received: October 9, 2014
Published: April 14, 2015

DOI: 10.1021/ci500611v
J. Chem. Inf. Model. 2015, 55, 925−940

alerts developed by Ashby and Tennant[16] is a well-known example of such associations. This set defines structural alerts for DNA reactivity based on the analysis of in vitro mutagenicity data and in vivo carcinogenicity data. Other researchers have greatly extended this set of alerts, and one of the most advanced lists for evaluating the mutagenic and carcinogenic potential of chemicals to date has been proposed by Benigni and Bossa.[17] This list has been implemented as a set of rules in knowledge-based expert systems such as Toxtree[18] and the OECD QSAR Toolbox.[19] However, some limitations have been reported in the literature:[20,21] (i) updating the knowledge base is a very time-consuming process since it requires strong investment of domain experts and a detailed analysis of the scientific literature; (ii) the expert opinion can sometimes be prone to subjectivity, leading to inaccuracies; and (iii) a negative response cannot be interpreted as a lack of toxicity but simply as a lack of information with respect to the molecule of interest.

The evolution of artificial intelligence and data mining tools should answer some of the limitations mentioned above, particularly the time and effort needed to identify new structural alerts, sometimes beyond the limits of human perception. The calculation of the frequency of a chemical substructure in a data set is often at the core of the process for the definition of its toxicological relevance. The rationale for using a frequency constraint is that it is unlikely to generalize on a substructure that has been observed on a few chemicals. However, algorithms that enumerate frequent substructures from a set of molecules, such as Gaston[22] and gSpan,[23] often lead to the generation of too many such substructures.

To limit the number of generated substructures, methods for finding representative and significant structural patterns have been developed in recent years. For example, from a mutagenicity data set, Kazius et al.[24] determined the statistical association of each proposed frequent substructure with mutagenicity, expressed as the p value resulting from a statistical test. Even though it relies on manual annotations, this work has enabled the development of 29 approved toxicophores. Recently, Ahlberg et al.[25] proposed a framework that automatically derives potential structural alerts; it also relies on the p value of a statistical test to select the significant substructures. Even though the computation does not exhaustively enumerate all of the possible substructures, it constitutes a fast and automated way to derive toxicophores. These two works do not directly calculate significant molecular substructures: they compute significant atom signatures[26,27] from which the significant substructures are derived.

Helma et al.[28] used MOLFEA, a molecular feature miner, to discover linear molecular fragments (chains) that occur with a higher frequency in mutagenic compounds than elsewhere. MOLFEA uses a levelwise algorithm[29] enabling the extraction of linear substructures that are frequent in the set of mutagens but infrequent outside of it. However, the restriction to linear substructures disables the direct extraction of fragments containing a branching point or a ring. This technical limitation has been overtaken thanks to the design of general frequent subgraph mining algorithms such as Gaston[22] and gSpan.[23] Kazius et al.[30] applied this methodological advance to extract fragments; their work led to the discovery of six new structural alerts.

Emerging pattern mining is a contrast data mining technique[31] introduced by Dong and Li.[32] The emerging constraint captures characteristics that differentiate between two classes of data and was first applied in chemoinformatics by Auer and Bajorath in 2006.[33] They introduced the notion of emerging chemical patterns (ECPs) as a novel approach to molecular classification. To describe the molecules, they did not use molecular graphs but instead employed a set of physicochemical and molecular properties. The *jumping emerging patterns* (JEPs) correspond to a subset of the emerging patterns: a JEP denominates a pattern that is sufficiently present in one class and absent from the other. Closed JEPs, called JSM hypotheses, were used in predictive toxicology by Blinova et al.:[34] an itemset representation was used, with items staying for particular molecular fragments.

Recently, Sherhod and co-workers[35,36] applied the notion of emerging patterns to identify structural features contrasting mutagenic with nonmutagenic compounds. An emerging pattern here corresponds to a conjunction of structural features, a structural feature being either a functional group, a ring fragment, or an atom pair.[37] The functional groups and ring fragments are automatically computed from a molecular data set by keeping only the most meaningful parts of the molecules. The method has also been successfully used to investigate clusters of mutagenic compounds and to implement new structural alerts in the knowledge base of the Derek Nexus expert system.[38]

The current work introduces a method that computes the conjunctions of molecular fragments whose frequencies of occurrence in a data set are sufficiently discriminative between different subgroups of molecules (e.g., mutagens and nonmutagens) to be of interest. The method operates directly from the molecular graphs: it automatically enumerates the molecular fragments that are sufficiently frequent to be considered.

In our previous works,[39−41] we introduced a graph-based mining method for the extraction of emerging patterns from a data set of molecules. This method necessitates two combinatorial enumerations. First, the enumeration of the molecular fragments allows the identification of the frequent fragments that will be used as structural features. Then the enumeration of conjunctions of frequent fragments enables the discovery of conjunctions of molecular fragments whose occurrences are correlated to a subgroup of molecules, such as mutagens or nonmutagens.

In this paper, we rely on our original calculation of emerging graph patterns to mine a mutagenicity data set collected by Hansen et al.[42] The novelty of this work relies on the use of closed patterns: we focus on the extraction of closed patterns that are in a one-to-one correspondence to the related formal concepts. This relation gives a structure to the closed emerging patterns we find. Thanks to this structure and to formal concept analysis,[43−45] we are able to select the most consistent emerging patterns, named stable emerging patterns (SEPs). Moreover, we also provide an interactive visualization tool to easily explore and evaluate the structural alert candidates.

The computational method is detailed in Materials and Methods. The main results of an expert analysis demonstrate the practical interest of the computational method: the extracted structural patterns constitute an efficient basis for a process of chemical knowledge discovery.

MATERIALS AND METHODS

Notions. *Molecular Patterns.* As an input, we consider a data set of molecules in which the structure of each molecule is given by its usual graph model. A *molecular graph* consists of a

DOI: 10.1021/ci500611v
J. Chem. Inf. Model. 2015, 55, 925−940

set of vertices, the atoms, that interact by means of edges, the chemical bonds. A vertex of a molecular graph is labeled with the atomic number it represents, while the label of an edge indicates the bond order. A *molecular fragment* represents a connected part of a molecule. A fragment *occurs* within a molecule if there is an embedding of the fragment in the molecule that simultaneously satisfies the relational structure of the fragment (the presence and the absence of every edge), and the labeling schemes of the edges and atoms. The *extent* of a molecular fragment denotes the set of molecules of the input data set in which the molecular fragment occurs. Given a set of molecules $\mathcal{M}$, a fragment f is *closed* in $\mathcal{M}$ if there is no fragment that contains f and occurs in every molecule in the extent of f, i.e., one cannot extend the graph of a closed fragment while preserving its extent in $\mathcal{M}$.

Throughout the current study, we aim to discover *molecular patterns* that correspond to potential structural alerts. A molecular pattern is a set of molecular fragments; the *length* of a molecular pattern designates the number of fragments it contains. A molecular pattern *occurs* in a molecule if each of its fragments occurs in the molecule; the *extent* of a molecular pattern denotes the set of molecules of the input data set in which the molecular pattern occurs. The *frequency* of a pattern in a chemical data set quantifies the relative number of molecules in the data set in which the pattern occurs.

Given two different molecular patterns p and q, p is *included* in q if each fragment of p is contained in a fragment of q. Here we denote the inclusion of p in q as $p \subset q$. The inclusion relationship provides a *partial order* between the molecular patterns and turns the set of molecular patterns into a partially ordered set. When $p \subset q$, q *covers* p if there is no molecular pattern r such that $p \subset r \subset q$.[46] A finite partially ordered set is usually depicted by a *Hasse diagram*, in which every pattern is associated with its own region and every covering pair is joined by a line segment.[47] Figure 1 shows a Hasse diagram for a

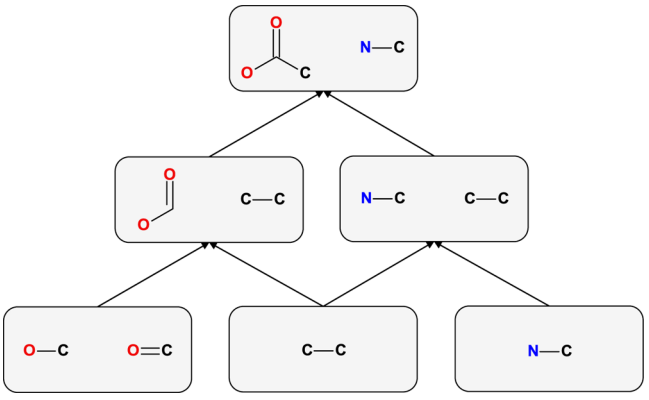


Figure 1. Example of a Hasse diagram for drawing the inclusions between molecular patterns.

partially ordered set of six molecular patterns. Each pattern is depicted as a rectangle, and the covering relation is indicated in a top-down manner: a molecular pattern q is linked to a molecular pattern p that lies under it if q covers p. For instance, the molecular pattern { C(=O)O and C−C } covers the molecular pattern at the bottom left ({CO and C=O}). From a Hasse diagram, it is easy to tell whether a pattern is included in another: $p \subset q$ if there is a sequence of connected line segments moving upward from p to q.

Closed Molecular Patterns. Given a set of molecules $\mathcal{M}$, a molecular pattern p is *closed* if there is no other molecular pattern that contains p and occurs in the same molecules as p, i.e., one cannot add any molecular fragment to p or extend an existing fragment of p while preserving all of its occurrences within the molecules of $\mathcal{M}$. As they lead to the chemically most interpretable information, we focus here on closed molecular patterns.

The following property stems from the definition of a closed molecular pattern: as soon as a molecular fragment f is an element of a closed molecular pattern p, any molecular fragment contained in f is also an element of p. Thus, the fact that a fragment f belongs to a pattern p loses its significance if f is contained in another fragment of p. Consequently, any fragment of a closed pattern p is pruned if it is a subfragment of another fragment of p; the resulting pattern is named a *pruned closed molecular pattern*. There is a one-to-one mapping between the pruned closed molecular patterns and the initial closed molecular patterns, and moreover, this correspondence preserves the extent.[40]

In this paper, we focus on computing and assessing the molecular patterns that are pruned closed molecular patterns (closed with respect to the considered learning molecular data set).[41] As demonstrated by Kuznetsov and Samokhin,[48] any element of a pruned closed pattern is necessarily a closed fragment. For the sake of simplicity, in the rest of this text, a pruned closed molecular pattern is called a molecular pattern.

Emerging Molecular Patterns. Since the entire set of molecular fragments is very large, it leads to a huge number of molecular patterns. To select meaningful patterns, one may consider a pattern only if it occurs sufficiently often in the molecular data set. However, a combination of frequent fragments does not necessarily lead to a relevant molecular pattern. For example, a molecular pattern made with the basic molecular fragment C−C does not carry alone any significance for studying important properties such as mutagenicity or acute toxicity.

To automatically discover structural alerts, it is highly appropriate to look for structural changes between different groups of molecules (e.g., between mutagens and nonmutagens). In particular, given a set of molecules, a molecular pattern that sufficiently occurs within the molecules of the given set and whose occurrences are significantly more frequent in the mutagens than in the nonmutagens stands as a potential structural alert related to the mutagenicity. The notion of a *frequent emerging molecular pattern* embodies this natural idea by using the growth rate measure. When a chemical data set is partitioned between targeted molecules and nontargeted ones (also named "classes"), the *growth rate* of a pattern p, denoted as $\rho(p)$, is defined as the ratio of the frequency of p in the targeted molecules to its frequency outside the targeted molecules.[32] Following our example, the growth rate of a molecular pattern is obtained by dividing its frequency in the mutagens by its frequency in the nonmutagens. A JEP is a pattern that has the noticeable property of occurring solely in molecules of the targeted class. By default, the growth rate of a JEP is denoted by the infinity symbol (∞). A frequent emerging molecular pattern denotes a molecular pattern that fulfills two constraints: a frequency sufficiently high to warrant further inductive usage and a growth rate sufficiently high to indicate a potential structural alert. Thus, being a frequent emerging molecular pattern depends on the settings of both the

DOI: 10.1021/ci500611v
J. Chem. Inf. Model. 2015, 55, 925−940

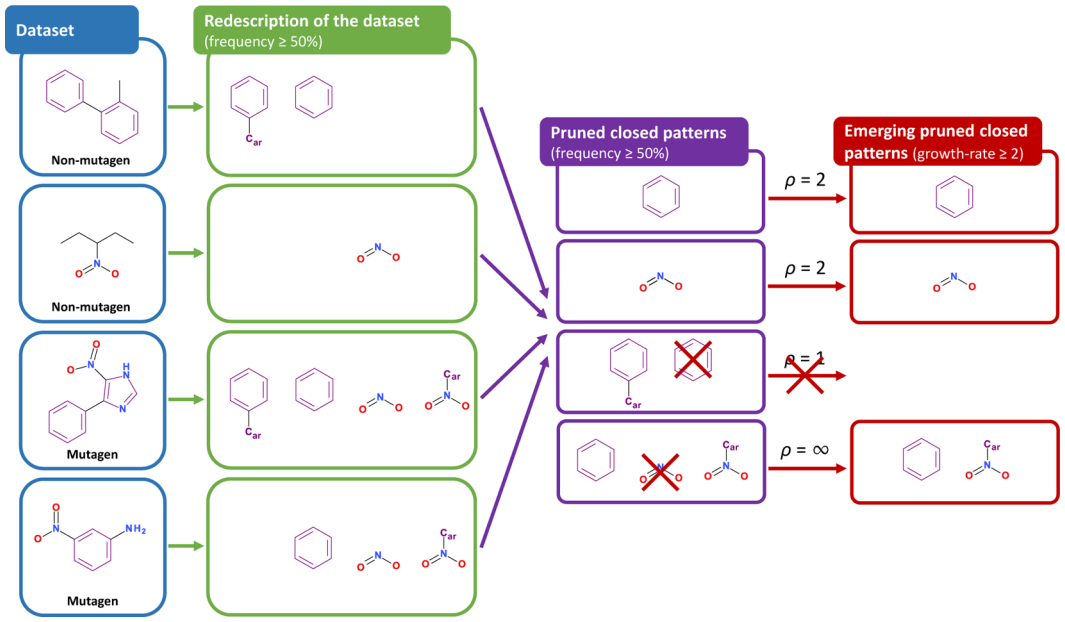


Figure 2. Illustration of going from a set of molecules to emerging pruned closed patterns.

minimum frequency threshold and the minimum growth rate threshold.

Illustration. Figure 2 illustrates the notion of a frequent emerging molecular pattern. As an input, this example considers the learning data set of four molecules depicted at the left of Figure 2, a minimum frequency threshold set to 50%, and a minimum growth rate threshold set to 2.

Since any frequent emerging molecular pattern is made up of frequent molecular fragments only,[40] it is sufficient enough to describe the molecules using the frequent fragments they contain. In the example, with the minimum frequency threshold set to 50%, a fragment that occurs twice or more among the four molecules is a frequent fragment; this results in four frequent fragments. These frequent fragments constitute the set of features used in an intermediate description of the molecules: every molecule of the data set is described by the frequent fragments it contains. From this binary description, one is able to generate every pattern that is frequent enough and is a closed pattern. In order to retain only meaningful pieces of information, these patterns are pruned: any fragment is removed from a pattern as soon as it is contained in another fragment of the pattern. In the example, the pruning step leads to the removal of the phenyl group from the third pattern and the removal of the nitro fragment from the fourth pattern.

In the example, with the minimum growth rate threshold set to 2, a frequent pattern that is at least twice as frequent among the mutagens as among the nonmutagens is an emerging pattern. Among the four frequent patterns, three are considered to be frequent emerging molecular patterns because their growth rates (denoted by ρ) exceed 2: each represents a conjunction of fragments that is frequent enough and whose occurrences in the learning set are discriminative enough to be of interest.

The following remark throws more light on the meaning of the growth rate: the growth rate of a pattern is directly related to the confidence of an association rule.[49] Let p be a pattern whose occurrences on a learning set indicate a relation to mutagenicity through a growth rate with the value $\rho(p)$. When we consider the association rule "$p \rightarrow$ mutagenicity", which states that "an occurrence of p implies mutagenicity", the *confidence* of this association rule corresponds to the conditional probability of being a mutagen among the molecules that contain the pattern p: the confidence of an association rule quantifies the validity of the related implication. The rule "$p \rightarrow$ mutagenicity" has a confidence c (measured on the learning set) that is related to $\rho(p)$ as follows:

$$c = \frac{\rho(p)}{\rho(p) + \frac{N^-}{N^+}} \quad (1)$$

where N^- and N^+ denote the numbers of nonmutagens and mutagens, respectively, in the learning set. As an illustration, if we consider the data set used in this article (detailed in Data Sets), the ratio of nonmutagens to mutagens is 0.89 in the learning set. Thus, a pattern whose growth rate is equal to 10 is associated with mutagenicity through an association rule whose confidence is $10/(10 + 0.89) = 0.91$. In other words, when the association rule "$p \rightarrow$ mutagenicity" applies in the learning set, it applies correctly in 91% of the cases.

Stable Emerging Patterns. The number of frequent emerging patterns is usually very high, and many of them are not significant and may result from artifacts of the data set. How can we select the most relevant patterns? In data mining there are a large number of measures for pattern ranking. One of them is stability, which originates from formal concept analysis.[43−45] The *stability* of a pattern p measures the relative number of subsets of the extent of p (i.e., subsets of molecules where p occurs) such that p is closed in these subsets. Intuitively, the stability of a pattern quantifies the degree to which a pattern depends on its extent.

Kuznetsov and co-workers[43,44,50] have shown that stability corresponds to the probability that pattern p is preserved if an arbitrary subset of molecules is removed from the data set. This gives us an intuition why stability is a useful measure for pattern selection. In fact, any pattern that we are going to find should be independent of any particular data set, and stability measures the extent to which the pattern is independent of the data set with respect to deletion of molecules.

DOI: 10.1021/ci500611v
J. Chem. Inf. Model. 2015, 55, 925−940

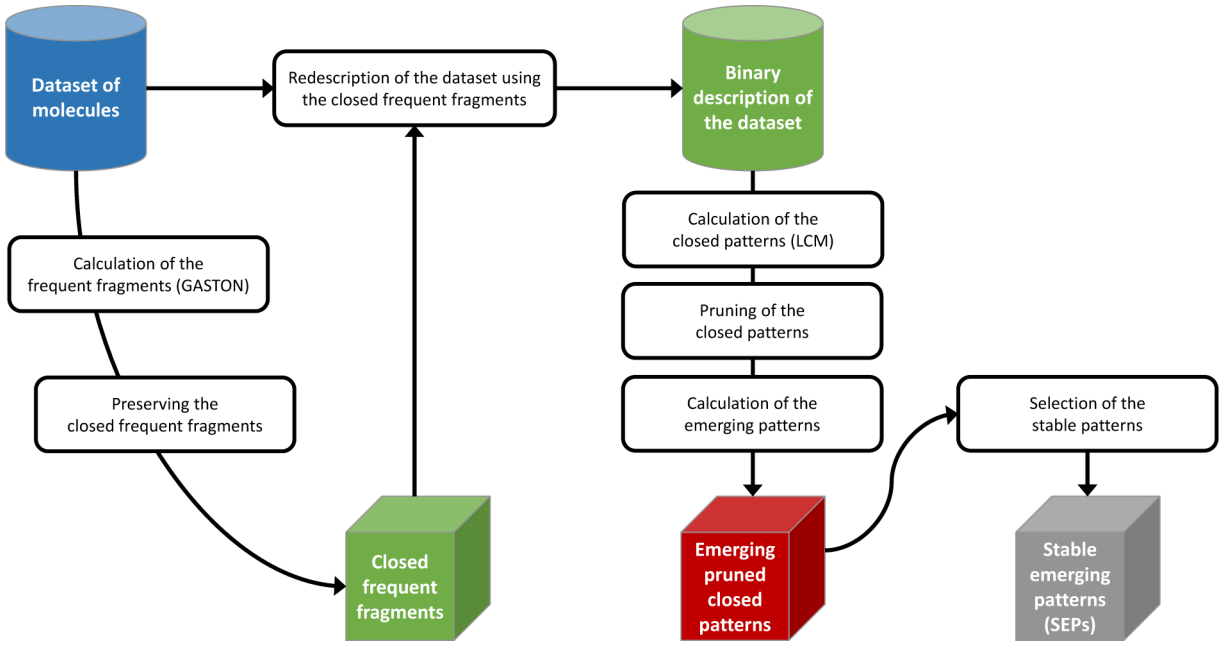


Figure 3. Workflow overview of the computation of the stable emerging patterns (SEPs).

Let us consider Figure 2. In this example, the phenyl group alone is a molecular pattern. To calculate the stability of this molecular pattern, one can reason as follows. From the initial data set of four molecules, 16 different subsets can be generated, including the empty set and the whole initial set. Among these 16 subsets, the phenyl group is considered as a molecular pattern if the subset fulfills two conditions. First, the phenyl group has to be considered as a fragment: it has to appear in at least one molecule of the subset. This condition is not fulfilled by two subsets: the empty set and the singleton subset with the second molecule alone. Second, the first molecule and the fourth molecule must be elements of the subset, as otherwise the phenyl group is not closed anymore. Indeed, without the fourth molecule in a subset, the closed molecular fragment should be a phenyl group associated by a single bond to an aromatic carbon and not a phenyl group alone. Similarly, without the first molecule in a subset, the (closed) pattern should be a phenyl group together with the fragment cN(=O)O. This second condition is fulfilled by four subsets. Thus, since the phenyl group is considered as a molecular pattern in four subsets generated from the initial data set, its stability is equal to 4/16 = 0.25.

Computational Method. This section details the calculation of the stable emerging patterns from a set of molecules partitioned into two subsets (e.g., mutagens and nonmutagens). A workflow overview of the method is provided in Figure 3. The method is a straight extension of the work introduced by Poezevara et al.[40] It relies on three main steps: the calculation of the closed frequent fragments, then the calculation of the frequent emerging pruned closed patterns, and finally the selection of the most stable emerging patterns.

Calculation of the Closed Frequent Fragments. First, the frequent molecular fragments are calculated by mining the training set of molecules provided as the input; the operation relies on a minimum frequency threshold. Gaston[22] is used to mine the chemical graphs. The efficiency of Gaston mainly relies on the adoption of the *quick-start principle* (see ref 22 for more details).

In our calculation, *a fragment never contains an incomplete chemical ring*. To exclude molecular fragments containing incomplete chemical rings, we use an approach similar to that of Borgelt.[51] As a pretreatment, any edge of a molecule that is included in at least one ring is tagged as "in a ring". As a filter, every frequent fragment output by Gaston is tested: if the fragment has at least one edge that is not in a ring but is tagged as "in a ring", the fragment is discarded from the list of the frequent molecular fragments.

The last operation in this step selects only frequent molecular fragments that are closed fragments. The fragments are grouped according to their extents, and then a fragment is kept only if it is not included in another fragment that shares the same extent. Any remaining fragment is a (closed) frequent molecular fragment. The inclusion test is handled using the Boost Graph Library (http://www.boost.org/doc/libs/1_55_0/libs/graph/), which is a refactoring of the work of Cordella et al.[52]

Calculation of the Emerging Pruned Closed Patterns. First, the frequent molecular fragments resulting from the previous step are used to describe the molecules of the training set: every molecule of the training set is described in terms of the set of frequent molecular fragments it contains. From this description, the closed molecular patterns are calculated using LCM (http://research.nii.ac.jp/~uno/codes.htm), which is an efficient implementation of the CbO algorithm.[53] It uses the *prefix-preserving extension*, which is an extension of a closed pattern to another closed pattern. This technique allows the algorithm to output closed patterns only. Every closed pattern is pruned: any of its molecular fragments is removed as soon as it is contained in another molecular fragment of this pattern. These pruned patterns are kept only if their growth rates exceed a threshold. This calculation identifies all of the emerging pruned closed patterns.

Selection of the Most Stable Molecular Patterns. The final step of the process consists of selecting the molecular patterns that are stable enough on the basis of a stability threshold. However, as the stability of a pattern (denoted as stab(p)) is

DOI: 10.1021/ci500611v
J. Chem. Inf. Model. 2015, 55, 925−940

hard to compute,[43−45] we use the estimate of stability computed from the difference in the extents for different patterns.[54]

Given two closed patterns p and q such that p is included in q, the *difference in the extents of p and q*, denoted as $d(p, q)$, corresponds to the difference in the cardinalities of the extents of p and q: $d(p, q) = |\text{extent}(p) - \text{extent}(q)|$. In order to bound the stability of a pattern p, the estimate uses the differences in the extents of p and q_i for every pattern q_i that covers p, i.e., $\{d(p, q_i)\colon q_i \text{ covers } p\}$:

$$1 - \sum_{q_i \text{ covers } p} 2^{-d(p,q_i)} \leq \text{stab}(p) \leq 1 - \max_{q_i \text{ covers } p} (2^{-d(p,q_i)}) \quad (2)$$

As we work with closed patterns, any pattern q_i that covers p has an extent smaller than the extent of p. Equation 2 relies on the following fact: for any pattern q_i that covers p, if q_i is a closed pattern, p is not closed in extent(q_i). Thus, p is not closed in any subset of extent(q_i) as soon as q_i covers p. If we exclude the extent of one pattern q_i that covers p, we have the upper bound of stab(p) given in eq 2. If we exclude the extents of all of the patterns covering p, since some of the subsets of the extent of p may be excluded several times, we have the lower bound of stab(p) given in eq 2.

Let us return to the example in Figure 2 and to the phenyl group as a pattern. The extent of the phenyl group contains three molecules. There are only two patterns that cover the phenyl group, and these two patterns have extents of size 2 (see Figure 4). By applying the latter formula, one obtains $1 - 2{\cdot}2^{-1} \leq \text{stab(phenyl)} \leq 1 - 2^{-1}$, i.e., the stability of the phenyl group lies between 0 and 0.5.

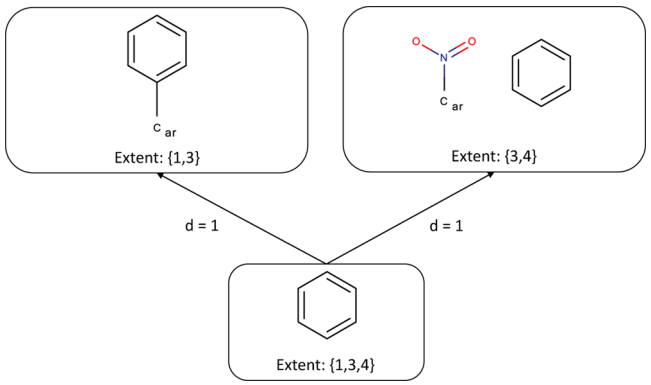


Figure 4. Selection of the most stable emerging patterns.

Data Sets. *The Hansen Data Set.* The mining process has been applied to a publicly available benchmark data set reported by Hansen et al.[42] The data set consists of 6512 compounds resulting from the compilation of Ames mutagenicity data described in the CCRIS,[55] Helma et al.,[28] Kazuis et al.,[24] Feng et al.,[56] VITIC,[57] and GENE-TOX databases.[58] To be classified as Ames-positive (i.e., a mutagen), a compound had to significantly induce a revertant colony growth in at least one of the strains of *Salmonella typhimurium*.[59] Even though the data set was already pretreated to remove duplicate structures and inorganic molecules, we cleaned the chemical data using Pipeline pilot (Accelrys Inc., San Diego, CA, USA) and ChemAxon (Chemaxon Ltd., Budapest, Hungary) components. The additional curation steps consisted of normalization of specific chemotypes (e.g., nitro group, organophosphate moiety, etc.), conversion of the structures to their aromatic form, and addition of hydrogens on the heteroatoms. This procedure resulted in a well-balanced data set containing 3503 mutagenic and 3009 nonmutagenic compounds.

The study of Hansen et al.[42] used a particular fivefold cross-validation scheme. The authors of the present paper have partitioned the data set into six parts. The first part gathers all of the compounds of the data set that are verifiable according to Derek Nexus[7,8] or MultiCASE;[12] the leftover compounds are distributed into the five other parts (with the same ratio of mutagens and nonmutagens). Within each of the five folds of cross-validation, the training set corresponds to the union of the first part with four of the five other parts and the test set is constituted by the remaining part. This validation scheme differs from the usual fivefold cross-validation because the first part of the partition is included in every training set.

External Test Set. It is now widely accepted that an external test is required to assess the predictivity of a classification model.[60,61] The constitution of a rigorous external test set, with no involvement in the model development, was considered as a part of this study. We collected every molecule from LeadScope (Leadscope Inc., Dublin, OH, USA) that is annotated with Ames mutagenicity data and does not belong to the Hansen data set. We curated the LeadScope chemical structures in the same way as for the Hansen data set, and we omitted molecules when inconsistent mutagenicity data were observed. This process resulted an external test set of 1178 molecules to measure the classification accuracy of our rules on unseen data.

EXPERIMENTAL RESULTS

Quantitative Assessments of the Stable Emerging Molecular Patterns. Throughout this section, key quantitative experimental facts are provided and discussed; they result from empirical investigations conducted on the data set described in the previous section. The whole experiment was performed thanks to the fivefold cross-validation scheme introduced by Hansen et al.[42] Every indicated result corresponds to an average calculated over the five folds, unless explicitly stipulated otherwise.

Closed Frequent Fragments. Setting the minimum frequency threshold to a low value results in a huge number of frequent fragments. For example, when the minimum frequency threshold was set to 0.31%, which corresponds to the necessity for a frequent fragment to occur in more than 17 molecules, three of the five folds of cross-validation each produced more than 30 billion frequent fragments (see the Supporting Information). This combinatorial explosion raises technical difficulties: it becomes impracticable to process such a huge number of fragments in order to investigate the subsequent patterns. Consequently, we limited the experimental study to frequency thresholds greater than or equal to 0.36%, which corresponds to the necessity for a frequent pattern to occur in at least 20 molecules.

Table 1 sets out the numbers of molecular fragments whose growth rates, denoted by ρ, exceed the indicated value as the minimum frequency threshold varies from 0.36% (at least 20 occurrences in different molecules of a training set) to 10% (at least 552 occurrences in different molecules). For example, when the minimum frequency threshold is set to 2%, 258.4 fragments are extracted from molecules of the training set and thus are considered as frequent; among these frequent fragments, 46.6 have $\rho > 2$, 12.8 have $\rho > 5$, and only 1.6 have $\rho > 10$.

DOI: 10.1021/ci500611v
J. Chem. Inf. Model. 2015, 55, 925−940

Table 1. Counts of Closed Frequent Fragments

frequency threshold	0.36%	1%	2%	5%	10%
no. of fragments					
total	2379.0	626.0	258.4	86.6	44.4
$\rho \geq 2$	532.8	143.4	46.6	14.4	12.0
$\rho \geq 5$	214.6	51.6	12.8	1.6	0.6
$\rho \geq 10$	91.2	17.8	1.6	0.0	0.0

As we aim to discover structural schemes that are related to mutagenicity, we are especially seeking fragments with high growth rates. In the present experimental context, these highly discriminative fragments (with $\rho \geq 10$) appear only when the frequency threshold is set to a low value (below 3%). Consequently, in the following, *the minimum threshold support is set to its lowest investigated value, 0.36%*: considering Hansen's data set together with its fivefold cross-validation scheme, this corresponds to the requirement that the fragment occur in at least 20 different molecules of a training set. Such a value for the frequency threshold allows us to find fragments and patterns that are neither too general nor too numerous. Moreover, as any further generalization is based on at least 20 molecules, it avoids conclusions relying on too few observations.

Table 2 reports the average numbers of frequent molecular fragments per molecule in a training set and in a test set; the

Table 2. Average Numbers of Closed Frequent Fragments per Molecule and the Related Cover Rates

	mutagens		nonmutagens	
	training set	test set	training set	test set
no. of fragments per molecule	30.58	34.84	31.97	38.74
cover rate	99.78%	99.52%	99.86%	99.36%

results are given separately for mutagens and for nonmutagens. Table 2 also provides the *cover rates* achieved by the frequent fragments: the cover rate achieved by a set of frequent fragments here denotes the portion of a set of molecules (mutagens or nonmutagens) that contains at least a frequent fragment. On average, a molecule includes more than 30 frequent molecular fragments, with very little difference between mutagens and nonmutagens. It follows very high cover rates that exceed 99% on any of the subsets of molecules. Very few molecules do not contain a frequent fragment: this population averages 19 molecules on a training set and 3 molecules on a test set. The latter are marginally small molecules compared to the molecules of the data set: on average these molecules contain 6.7 atoms against 18.4 atoms on the whole data set. As a conclusion, training and test sets taken from the Hansen's data set are properly covered by the frequent molecular fragments.

Closed Frequent Patterns. Table 3 displays the average cardinalities of molecular patterns for various minimum frequency thresholds and growth rate thresholds. For example, there are 222651 molecular patterns whose frequencies are over 0.36%; 1500 of these patterns have growth rates greater than 10. Comparison of the numbers of patterns exceeding a given frequency and a given growth rate with the corresponding numbers of fragments (see Table 1) shows that the number of frequent patterns is by far more important than the related number of frequent fragments. If we consider the growth rates

Table 3. Counts of Closed Frequent Patterns

frequency threshold	0.36%	1%	2%	5%	10%
no. of patterns					
total	222651.0	38889.6	8083.6	868.0	194.8
$\rho \geq 2$	12968.6	2217.4	534.8	75.8	41.4
$\rho \geq 5$	4564.2	690.2	122.4	4.2	1.2
$\rho \geq 10$	1499.8	189.0	22.8	0.0	0.0

exceeding 10, the frequent patterns are 16 times more numerous than the frequent fragments. In terms of association rules, there are 1500 different association rules that have a frequent pattern as their premise and conclude on the mutagenicity of a molecule with a confidence exceeding 91% (see eq 1), while there are only 91 rules having the same confidence using a frequent fragment as a premise.

Table 4 reports the cover rates related to these frequent molecular patterns having a frequency exceeding 0.36%, i.e., the

Table 4. Average Numbers of Frequent Molecular Patterns Per Molecule and the Related Cover Rates

	mutagens		nonmutagens	
	training set	test set	training set	test set
no. of patterns per molecule	1262.23	1518.64	2298.79	2816.33
cover rate	99.78%	99.52%	99.86%	99.36%

portions of mutagens and nonmutagens that contain at least one of these frequent molecular patterns. With the same frequency threshold, when a molecule contains a frequent fragment, it then contains the pattern that corresponds to this fragment alone. Therefore, a cover rate obtained with the frequent patterns is at least as important as the corresponding cover rate with the frequent fragments. Conversely, a molecular pattern is composed of fragments, and thus, its cover rate cannot exceed the fragment cover rate. Consequently, it is not surprising that the cover rates indicated in Table 4 are the same as the ones obtained with the frequent fragments (see Table 2).

Emerging Pruned Closed Patterns. Figure 5 details the cover rates related to the emerging molecular patterns according to different growth rate thresholds; the cover rates are given separately for mutagens and nonmutagens, and the figures correspond to averages measured on the five test sets. For instance, the molecular patterns with growth rates greater than or equal to 5 cover 74.04% of mutagens, while they cover only 31.96% of nonmutagens. As expected, the higher the minimum growth rate threshold is, the greater is the difference between the two cover rates. For example, when patterns having growth rates above 5 are considered, then the cover rate of mutagens is 2.32 times higher than the cover rate of nonmutagens. This ratio reaches 3.31 for patterns whose growth rates exceed 10. These results indicate that the emerging molecular patterns still occur discriminatively outside of their training set.

For several consecutive growth rate ranges, Table 5 reports facts about the frequent patterns whose growth rates belong to the given interval; the last column, with a growth rate value of ∞, is dedicated to the frequent patterns that occur only in mutagens. The three rows in the top portion of the table provide results computed for the training set averages calculated from the five training sets of the cross-validation. The rows entitled "ρ" and "related confidence" provide ranges

DOI: 10.1021/ci500611v
J. Chem. Inf. Model. 2015, 55, 925−940

Growth-rate threshold	*Mutagen cover rate*	*Non mutagen cover rate*
0	0.9986	0.9936
1	0.9773	0.8863
2	0.9202	0.6369
3	0.8582	0.4841
4	0.7888	0.3827
5	0.7404	0.3196
6	0.6863	0.2604
7	0.6204	0.2195
8	0.5757	0.1943
9	0.5277	0.1724
10	0.4928	0.1486

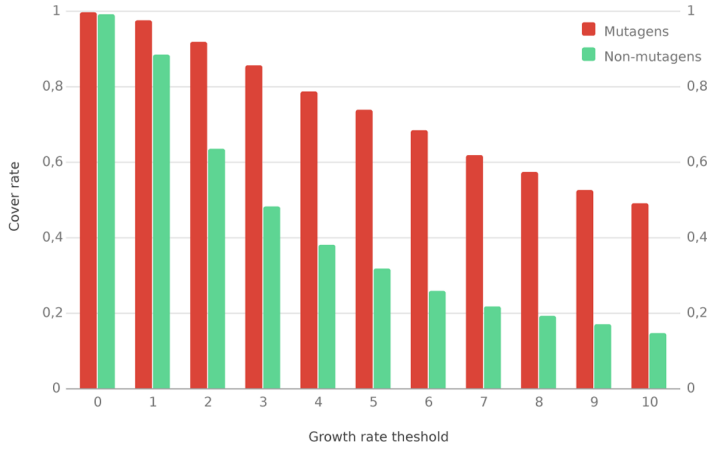


Figure 5. Cover rates obtained with the closed frequent patterns in a test set.

Table 5. Closed Frequent Patterns According to Their Growth Rates

		Computed from the Training Sets			
ρ	[2:5]	[5:10]	[10:20]	[20:∞[	∞
related confidence	[0.70:0.85]	[0.85:0.92]	[0.92:0.96]	[0.96:1.00[	1.00
no. of patterns	8404.40	3064.40	1094.00	405.80	370.20
		Measured in the Test Sets			
no. of matches	82658.20	26227.20	8699.00	2559.80	1887.20
no. of good matches	61631.60	21995.60	7726.20	2322.80	1790.80
measured confidence	0.74	0.84	0.89	0.91	0.95

Table 6. Best Values of the Stability Thresholds and the Related AUCs

fold	1	2	3	4	5	average
stability threshold	0.92	0.99	0.93	0.98	0.96	**0.96**
AUC	0.7792	0.7697	0.7739	0.7811	0.7789	**0.7766**

of growth rates from the training set and the confidences of the related association rules. The third row reports the numbers of patterns in the related growth rate ranges. The three rows in the bottom portion of the table deal with the application of the molecular patterns on a test set. If a given pattern occurs (at least once) in a given molecule, we name it a *match* (reported in the first row); and if the molecule is a mutagen, we name it a *good match* (reported in the second row). The last row indicates the average confidences of the related association rules measured on the test sets, corresponding to the proportions of good matches among the matches.

For example, there are 1094 patterns whose growth rates lie between 10 and 20. Each one of these 1094 patterns occurs on average in 7.95 molecules of a test set; thus, the number of matches is 8699. Among these 8699 occurrences, 7726 occur in a mutagen, thus, the level of confidence on the test sets is equal to 0.89. The application of the extracted molecular patterns as association rules on the test sets leads to very fair levels of confidence, ranging from 74% to 95%. Moreover, the level of confidence regularly increases together with the accounted growth rate. On average, 370 frequent molecular patterns occur only in mutagens of the training set. These patterns reach a particularly high level of confidence (95%) when they are applied to the molecules of a test set.

These quantitative facts constitute a strong advocation in favor of the emerging pruned closed patterns. Even when the growth rate threshold is high enough to be discriminative, the emerging pruned closed patterns still occur in a large portion of mutagens of a test set. Moreover, when the frequent emerging patterns are applied as association rules, they demonstrate high levels of confidence.

Stable Emerging Patterns. The previous section has provided experimental results that constitute a fair advocation for using emerging pruned closed patterns. Despite this fact, in order to obtain potential structural alerts, we rely on the stability measure to select candidates among the frequent emerging molecular patterns. This selection aims to provide a reasonable number of molecular patterns that are as independent of the constitution of the training set as possible.

Setting of the Minimum Stability Threshold. The selection based on the stability measure relies on a minimum stability threshold. To automatically set this parameter to its best value, we performed a cross-validation on each of the cross-validation folds of Hansen's data set (see the Supporting Information). To evaluate the impact of a value of the stability threshold, we measured this impact on the area under the curve (AUC) of a receiver operating characteristic (ROC) plot. An ROC plot is conceptually similar to an enrichment plot in that it shows the relationship between the true-positive rate and the false-positive rate.[62] The AUC of a ROC plot is a common way to quantitatively summarize the overall quality of the plot. On the basis of the AUC indicator, we aim here to discard as many frequent emerging molecular patterns as possible while conserving the ability to discriminate between mutagens and nonmutagens. Table 6 reports the best values of the stability and its related AUC for each of the original Hansen's folds. It also reports the average of these thresholds over all of the folds; the mean value will be used in the following. To maximize the

DOI: 10.1021/ci500611v
J. Chem. Inf. Model. 2015, 55, 925−940

AUC, the stability has to be set to a high value (over 0.90) on each cross-validation fold. On average, the value of the stability threshold is 0.96.

Stable Emerging Patterns. The quantitative assessment of the stable emerging molecular patterns relies on the measurements already used for assessing the SEPs. Table 7 displays the numbers of SEPs according to different frequency thresholds and different growth rate thresholds. Analogous counts without selection based on the stability are given in Table 3. A comparison of the results provided by Tables 3 and 7 indicates that the stability-based selection is very efficient. When the frequency threshold is set to 0.36%, the number of patterns is reduced by a factor of 15. At the same time, the selection resulting from a stability threshold tends to keep patterns with high values of the growth rate. For example, the frequent patterns having growth rates above 5 are reduced by only a factor of 7.5. The selection based on stability raises the portion of the strongly dicriminative molecular patterns among all of the frequent patterns; this fact fully coheres with the part such a selection has to play.

Table 7. Counts of Stable Emerging Patterns

frequency threshold	0.36%	1%	2%	5%	10%
no. of SEPs					
total	14943.0	9641.8	4387.8	792.4	183.2
$\rho \geq 2$	2167.2	1036.4	372.6	62.2	30.8
$\rho \geq 5$	616.8	261.0	71.8	3.8	0.8
$\rho \geq 10$	164.0	57.0	10.2	0.0	0.0

For a frequency threshold of 0.36%, Table 8 reports the cover rates related to the SEPs, i.e., the portions of mutagens and nonmutagens that contain at least one of these SEPs. A comparison with Table 4 indicates that the cover rates are highly maintained from the closed patterns to the SEPs. While the number of patterns has been divided by 15, the average number of occurrences per molecule is reduced by a factor varying from 5.2 (on mutagens of training sets) to 9.2 (on nonmutagens of test sets). These results indicate that the SEPs still properly describe the molecules of the data set.

Table 8. Average Numbers of SEPs Per Molecule and the Related Cover Rates

	mutagens		nonmutagens	
	training set	test set	training set	test set
no. of patterns per compound	242.86	263.35	364.04	307.71
cover rate	99.78%	99.51%	99.86%	99.36%

Figure 6 provides the cover rates obtained with the SEPs on a test set. For example, with the growth rate threshold set to 4, 76.10% of mutagens of a test set and 31.96% of nonmutagens contain at least one SEP. The comparison of these results with the ones in Figure 5 shows that the cover rate of mutagens slightly decreases when SEPs are used (the ratio varies between 1 and 85%), while the cover rate of nonmutagens decreases more significantly (the ratio varies from 1 to 60%). It follows that mutagens and nonmutagens are separated better by the set of SEPs than by the whole set of emerging pruned closed patterns when the growth rate threshold is set to a high value.

As a conclusion, the selection of the SEPs leads to a set of patterns that is more discriminative. Moreover, as such a selection noticeably decreases the number of patterns, it enables one to focus on the strongest chemical patterns and thus facilitates the examination of the selected patterns as potential structural alerts.

Contribution of the Stability. As seen previously, the stability greatly reduces the number of molecular patterns without jeopardizing the cover rate of molecules. To assess the contribution of stability in terms of discriminating power, molecular patterns and stable molecular patterns need to be compared in classification.

Molecular patterns can be used in association rules in which the premise is the presence of a pattern in a molecule and the conclusion is mutagenicity of a molecule with a confidence immediately correlated to the growth rate of the pattern in the premise. Given a growth rate threshold and a set of association rules, a naive classifier can be engineered to separate the molecules. Using a fivefold cross-validation (see the Supporting Information), the growth rate threshold is set to the value maximizing the accuracy (this value ranges from 3.39 to 4.05).

Table 9 reports the results in terms of accuracy, precision, recall, and AUC. The accuracy is a good prediction rate of a classifier. The precision is the number of mutagens among the predicted mutagen molecules: precision = 100% × TP/(TP + FP), where TP and FP are the numbers of true positives and false positives, respectively. The recall corresponds to the

Growth-rate threshold	*Mutagen cover rate*	*Non mutagen cover rate*
0	0.9986	0.9936
1	0.9728	0.8703
2	0.9031	0.5909
3	0.8332	0.4187
4	0.7610	0.3196
5	0.6951	0.2467
6	0.6309	0.1957
7	0.5455	0.1514
8	0.4973	0.1325
9	0.4470	0.1096
10	0.4226	0.0916

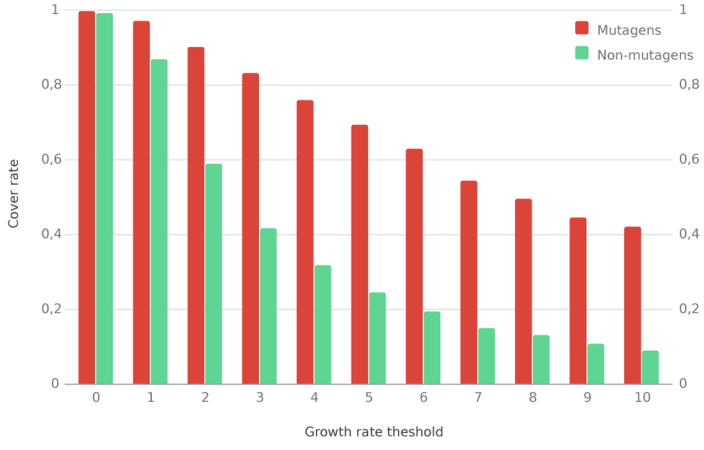


Figure 6. Cover rates obtained with the SEPs on a test set.

DOI: 10.1021/ci500611v
J. Chem. Inf. Model. 2015, 55, 925−940

Table 9. Result Table for the Classification of Hansen'S Dataset with Several Classifiers

	no. of patterns	accuracy (%)	precision (%)	recall (%)	AUC
EPCPs	222651	71.73	73.39	80.85	0.777
SEPs	14943	72.82	76.04	77.92	0.785

number of mutagens predicted among the whole set of mutagens: recall = 100% × TP/(TP + FN), where FN is the number of false negatives. The first row reports results obtained with emerging pruned closed patterns (EPCPs) in a naive classifier, and the second row gives the results obtained with SEPs.

The use of SEPs rather than EPCPs increases the accuracy of the naive classifier by more than 1%. It also increases the precision by about 3%, but at the cost of 3% in the recall. Nevertheless, these results on the naive classifier show that using SEPs improves the overall quality of the classifier. This can be explained by the fact that nonstable patterns do not generalize very well. Nonstable patterns are sensitive to their extension, so removing a few molecules may exclude them from the pattern set. This behavior can be related to labeling errors or statistical anomalies in the training set.

Seal et al.[63] published classification results using four genuine classifiers on Hansen's data set. One was a naïve Bayes classifier, which achieved an accuracy of 63.28%, and a sequential minimal optimizer achieved an accuracy of 66.43%. J48, which is a decision-tree-based classifier, reached an accuracy of 73.65%, and finally, a random forest classifier reached a high accuracy of 79.18%. Our results using SEPs are competitive with the ones published by Seal et al.[63] Indeed, they outperform the naïve Bayes classifier and the sequential minimal optimizer, give results similar to those of J48, but are less good than the random forest classifier. However, it is important to note that results of Seal et al.[63] did not use the same cross-validation. They used a fivefold cross-validation on all of the molecules, and thus, molecules from the static training set could be used in a test set. These molecules are easier to classify (see Hansen et al.[42]), so including them in the test set may boost the accuracy. Using the same type of cross-validation increases the accuracy of our approach to 74.57%, which ranks the SEPs as the second-best classifier in terms of accuracy.

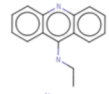
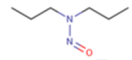
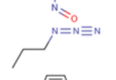
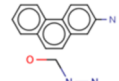
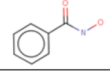
If we compare our results in terms of AUC, it is possible to complete our comparaison with the state of the art. Hansen et al.[42] and Xu et al.[64] used genuine classifiers (ranging from *k* nearest neighbors (*k*-NN) to support vector machine (SVM)) to separate mutagens from nonmutagens and reported as best results AUC values of 0.86 and 0.858, respectively. These results are better than the results returned by our naïve classifier, but nonetheless, our results in terms AUC indicate that the use of SEPs as a fingerprint in more sophisticated classification techniques is promising.

Expert Analysis of the Molecular Patterns. *Navigation Tool for Exploring the Emerging Molecular Patterns.* The previous section practically indicates that the successive application of a constraint of frequency, a constraint of emergence, and a constraint of stability leads to the automatic identification of promising molecular patterns. Nevertheless, in a process of chemical knowledge discovery, the emerging molecular patterns need a further manual examination by experts in the domain. The examination may produce definitions of new validated structural alerts, but it may also lead to a better understanding of the related activity (e.g., of the mutagenicity). Such work has to account for both the emerging molecular patterns and the relationship between them.

As the Hasse diagram provides an efficient way to explore a set of molecular patterns, we have implemented a tool that automatically builds the Hasse diagram related to a set of molecular patterns; this tool allows a user to navigate through the results thanks to an interactive Web site (https://chemoinfo.greyc.fr/2014_Metivier/). This Web site also provides a description of the molecular patterns with the associated numerical features (frequency, growth rate, and stability). Moreover, from the Web page dedicated to a given molecular pattern, one can directly access the list of each molecule that contains this pattern.

Analysis of the Frequent Molecular Fragments. In our methodology, molecules are redescribed by means of the occurrences of the closed frequent fragments (MFs). Selections of the MFs resulting from the mining process are displayed in Tables 10 and 11. These MFs were selected on the basis of their similarities to ToxAlerts toxicophores[65] for mutagenicity. ToxAlerts is an open expert-knowledge-based platform that contains more than 600 toxicophores from the literature for several end points such as mutagenicity, carcinogenicity, skin sensitization, idiosyncratic drug toxicity, and acute aquatic toxicity. In Tables 10 and 11, for every MF, the structure is given together with the support and the growth rate. The support of an MF (denoted here by *s*) corresponds to the number of molecules in which the MF occurs. The top 10 MFs (Table 10) are all JEPs, thus corresponding to structural features that are present in the toxic class but absent from the nontoxic one. Among them, we easily retrieve MFs associated with well-known toxicophores such as heteroaromatic nitro groups (MF_945, MF_954), polycyclic aromatic amines (MF_1666, MF_991), a nitrosamine group (MF_252), an azide group (MF_4), and a polycyclic planar hydrocarbon system (MF_1616). Since our mining process only preserves

Table 10. List of the Molecular Fragments Corresponding to JEPs

	JEPs	Support	ρ
MF_945		83	∞
MF_954		31	∞
MF_1666		27	∞
MF_991		32	∞
MF_252		27	∞
MF_4		28	∞
MF_1616		26	∞
MF_1211		25	∞
MF_87		25	∞
MF_414		37	∞

DOI: 10.1021/ci500611v
J. Chem. Inf. Model. 2015, 55, 925−940

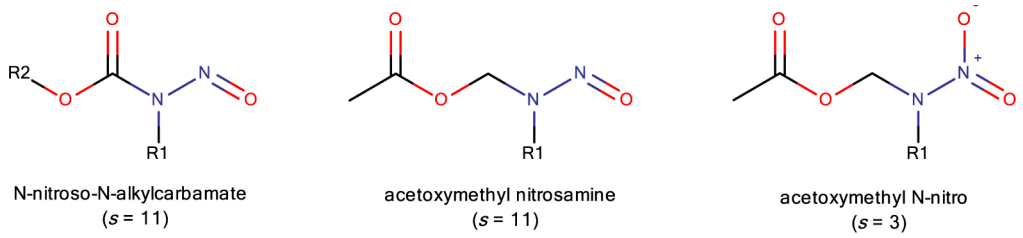


Figure 7. Three toxicophores generalized by MF_87.

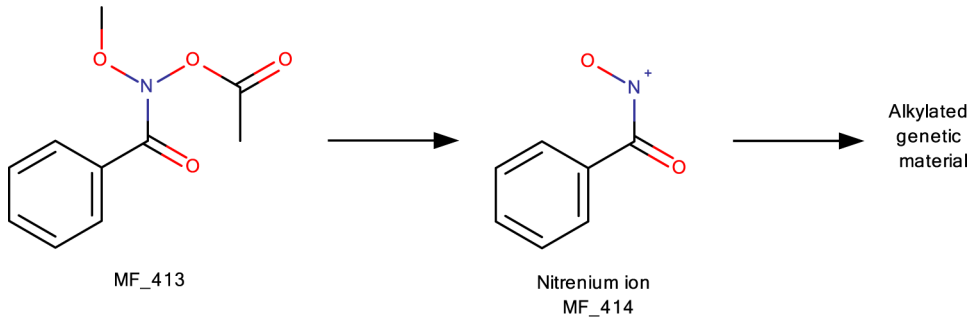
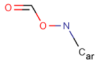
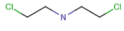
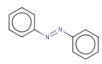


Figure 8. Detection of the nitrenium MF_414, metabolite of MF_413.

Table 11. Comparison of Molecular Fragments Corresponding to EPs with Known Structural Alerts from ToxAlerts

	EPs	Support	ρ	Structural Alert in ToxAlerts
MF_73		44	36.94	N-nitroso-N-alkylamides N-nitroso-N-alkylureas N-nitroso-N-alkylcarbamates
MF_0		67	27.92	Diazo
MF_1		55	22.76	Azide
MF_125		52	21.47	Aromatic and aliphatic aziridinyl derivatives
MF_156		47	12.60	Aromatic hydroxylamine ester
MF_72		27	9.09	Nitrosamine
MF_1651		178	6.79	Polycyclic aromatic hydrocarbons
MF_1841		26	6.59	Allylic halides
MF_824		48	6.01	Hydroxyl amine
MF_1667		226	5.84	Polycyclic aromatic hydrocarbons
MF_1831		36	5.33	Monohaloalkene
MF_75		162	5.19	Unsubstituted heteroatom-bonded heteroatom
MF_927		35	5.16	Aromatic N-acyl amine
MF_1298		27	4.94	Quinones
MF_920		964	4.70	Nitrosoarenes
MF_134		926	4.63	Aromatic nitro groups
MF_444		93	4.47	Nitrogen mustard

	EPs	Support	ρ	Structural Alert in ToxAlerts
MF_212		35	4.15	N mustard
MF_2188		29	4.12	Acyl halides
MF_2107		40	4.05	Alkyl ester of sulfonic and sulfuric acids
MF_34		74	3.68	Aliphatic azo
MF_1317		26	3.61	Heterocyclic polycyclic aromatic hydrocarbons
MF_1883		308	2.33	Aliphatic and aromatic epoxides
MF_916		656	2.17	Primary aromatic amine
MF_432		50	1.83	Alkyl carbamate
MF_41		151	1.54	Aromatic azo
MF_153		51	1.33	Aliphatic nitroso
MF_169		182	1.25	Tertiary aromatic amine
MF_2197		26	1.17	Aliphatic halogens
MF_1836		59	1.17	α,β-unsaturated carbonyl
MF_1470		43	0.82	Coumarins
MF_2220		1491	0.64	Simple aldehyde
MF_2097		47	0.58	Alkyl ester of phosphonic and phosphoric acids

the ring structures, some chemical functions can be truncated and do not always accord with chemical intuition. At first glance this can look like a limitation of the MFs automatically derived from a database compared with the domain expert rules. However, we can consider MF_1211, which highlights the capability of the method to extract generalized toxicophores. MF_1211 summarizes in only one toxicophore the mutagenic effects of polycyclic aromatic nitro groups and polycyclic aromatic amines. MF_87 is another example of such generalized toxicophores, but contrary to MF_1211, the related

DOI: 10.1021/ci500611v
J. Chem. Inf. Model. 2015, 55, 925−940

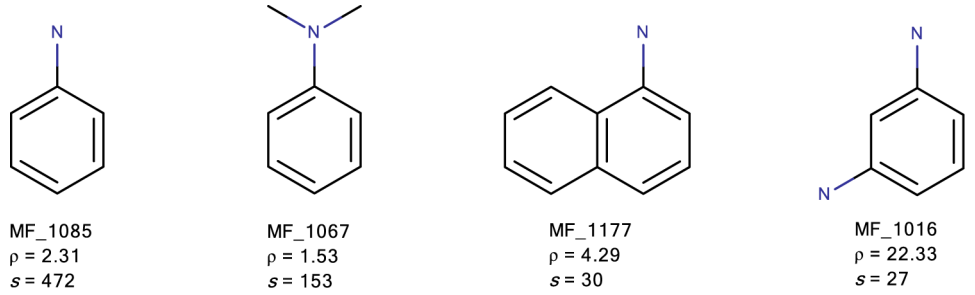


Figure 9. Growth rates associated with different aromatic amines.

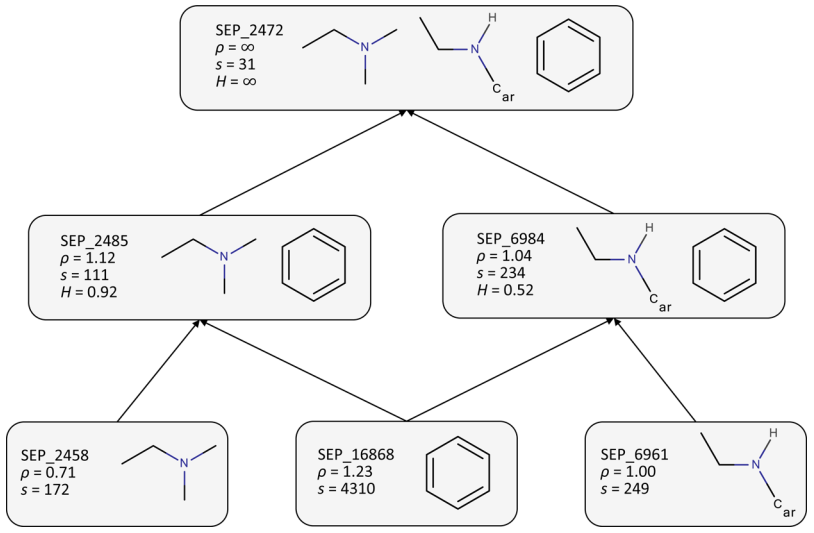


Figure 10. Example of a conjunction of SEPs leading to a JEP.

chemical functions are not obvious. By analyzing the extension of MF_87 we observed that it corresponds to the amalgam of *N*-nitroso-*N*-alkylcarbamate, acetoxymethyl nitrosamine, and acetoxymethyl *N*-nitro groups (Figure 7). We can also notice that these three groups were not supported by a sufficient number of examples in the training data set, preventing their individual extraction. The last JEP, MF_414, is of particular interest because it corresponds to a nitrenium ion, a DNA-alkylating agent resulting from the metabolization of MF_413 (Figure 8).

A limitation of JEPs is their inherent intolerance of noisy data, since the presence of even a small number of misclassified compounds can lead to the nondetection of very interesting patterns. Our method is also able to mine EPs that are more noise-tolerant in comparison with JEPs. EPs represent discriminating patterns that are common to toxic compounds but can also cover some nontoxic ones. Table 11 displays the statistical results for the most general EPs that are equivalent to 35 out of the 53 nonredundant well-known structural alerts.[65] These MFs are sorted according to their growth rates, allowing a domain expert to examine those with the highest growth rates first. Because of the straight relation between the growth rate and the contrast of the MF between the toxic and nontoxic classes, the top-ranked MFs represent the most probable toxicophores. In the following, we will discuss some representative, interesting, and sometimes intriguing results.

Let us first consider MF_73, which corresponds to a generalization of the *N*-nitroso-*N*-alkylamide, *N*-nitroso-*N*-alkylurea, and *N*-nitroso-*N*-alkylcarbamate toxicophores. MF_73 is supported by 43 toxics and only one nontoxic, resulting in an EP with a very high growth rate of approximately 36.9 denoting a very high discriminatory potency. The relation between the mutagenicity end point and the structural variations of the aromatic nitro compounds is also rightly captured by our method. While the heteroaromatic nitro compounds 2-nitrofurane and 2-nitrothiophene are both JEPs, the general toxicophore for an aromatic nitro group (MF_134) is basically an EP with a growth rate of 4.63. In regard to aromatic amines (Figure 9), significant differences were also observed as functions of the nature of the amine (MF_1085 vs MF_1067), the number of aromatic rings (MF_1085 vs MF_1177), and the number of amino groups (MF_1085 vs MF_1016). Similarly a relation between the nature of the nitroso substituents and the mutagenicity potency of the resulting compounds is emphasized: the growth rate increases in going from aliphatic nitroso compounds (MF_153, ρ = 1.33) to nitrosoarenes (MF_920, ρ = 4.70), and the highest growth rate for nitroso compounds is associated with polycyclic structures (MF_1201, ρ = 14.39; see the Supporting Information). Finally, we would like to mention an intriguing result. Although the alkyl esters of phosphonic and phosphoric acids are widely accepted as mutagenic toxicophores, the corresponding EP (MF_2097) exhibits a growth rate smaller than 1 (ρ = 0.58). This value indicates a greater extension of MF_2097 among the nontoxic class. The constitution of Hansen's data set does not allow its mutagenic potential to be highlighted.

In comparison with the toxicophores from ToxAlerts, those that are missing are not supported by a sufficient number of examples in the training data set to be extracted by our method.

DOI: 10.1021/ci500611v
J. Chem. Inf. Model. 2015, 55, 925−940

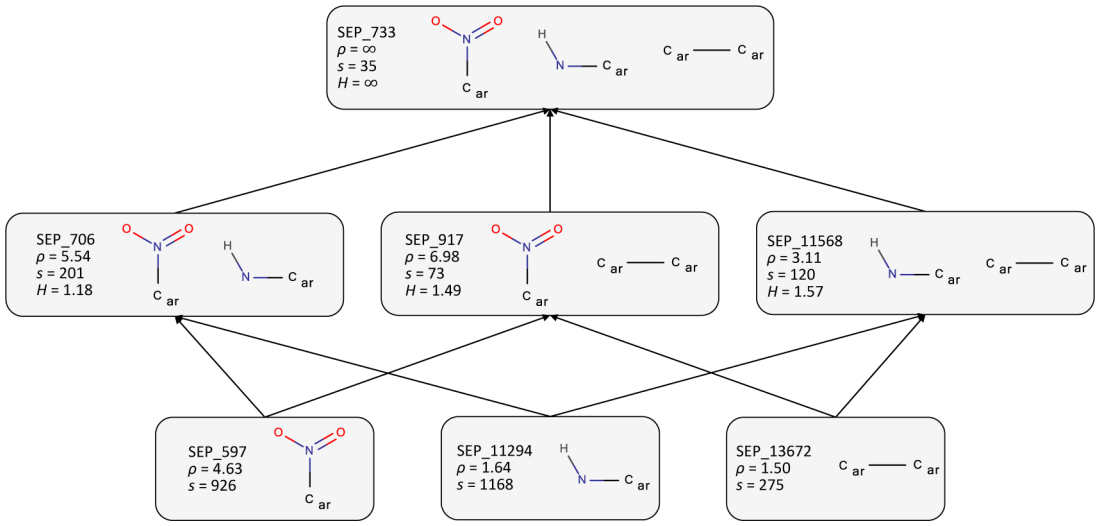


Figure 11. Example of stimulation of an aromatic nitro group.

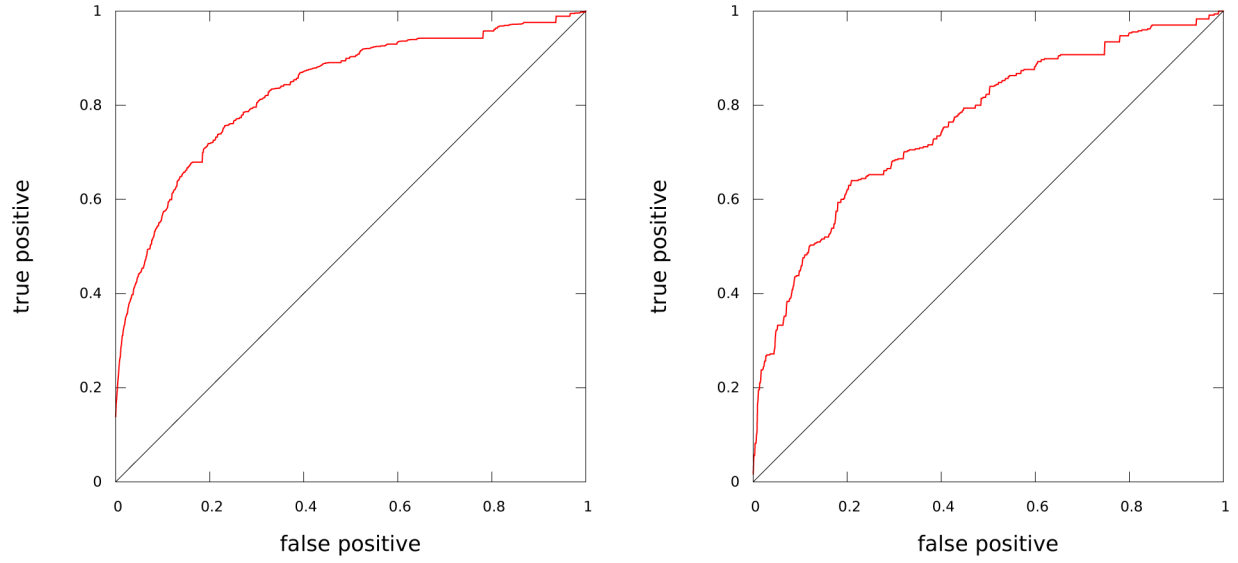


Figure 12. ROC curves for (left) the training set and (right) the external test set.

Indeed, the 18 missing toxicophores match from 0 to 19 compounds and did not satisfy the frequency constraint, which was set to 0.36% (i.e., an extent of 20 compounds). For example, among the missed toxicophores, the alkyl nitrite group, the α,β-unsaturated aliphatic alkoxy group, and the haloamines were respectively represented by 7, 13, and 6 mutagenic compounds and no nonmutagenic ones. Thus, with a reduction in the frequency threshold, these toxicophores would have been extracted.

Analysis of the Stable Emerging Patterns. In this section, we focus on the SEPs. As an additional measure, we can evaluate the following ratio:

$$\mathcal{H}(p) = \frac{\rho(p)}{\text{argmax}_{f_i \in f}\, \rho(f_i)} \tag{3}$$

where p is an SEP and f is the set of closed frequent fragments included in p. When $\mathcal{H}$ is greater than 1, the conjunction of fragments is more mutagenic than each of the individual fragments. Two hypotheses can explain this phenomenon. The first one is a conjunction of individually nonmutagenic fragments whose association leads to a mutagenic pattern. As an example (Figure 10), let us consider the conjunction of a tertiary amine (SEP_2458), an anilino fragment (SEP_6961), and a phenyl group (SEP_16868). The associations between the tertiary amine and the phenyl group (SEP_2485) and between the anilino fragment and the phenyl group (SEP_6984) do not lead to high growth rates (1.12 and 1.04, respectively), but the association of all three fragments leads to a JEP (SEP_2472, $\rho = \infty$, $s = 31$, $\mathcal{H} = \infty$).

The second hypothesis is stimulation[66] associated with some fragments, leading to an increase in the overall mutagenic property. For example (Figure 11), separately considered the aromatic nitro group already represents a constrasting molecular fragment in favor of mutagenicity (SEP_597, ρ = 4.63). The conjunction with a nitrogen (NH) connected to an aromatic group (SEP_11294) increases the growth rate (SEP_706, ρ = 5.54). The addition of a third fragment corresponding to two aromatic rings connected by a single bond (SEP_13672) even leads to a JEP (SEP_733, $\rho = \infty$, s = 35, $\mathcal{H}=\infty$). For most of the cases, this notion of stimulation is clearly pointed out.

DOI: 10.1021/ci500611v
J. Chem. Inf. Model. 2015, 55, 925−940

External Test of the SEPs. An independent external test set consisting of 1178 additional molecules selected from LeadScope was used to evaluate the generalization of the predicting rules in application, and the performances are shown in Figure 12. To be classified as a mutagen, a compound must exhibit at least one SEP. The area under the ROC curve and the maximum prediction rate were approximately 0.76 and 0.73, respectively. We observed a slight decrease in the performance in comparison with the training set (0.83 and 0.76, respectively). The maximum prediction rate was obtained with molecular patterns displaying growth rates greater than 4. By using SEPs with lower growth rates, we would detect a greater number of mutagens, but the number of false positives (i.e., nonmutagens classified as mutagens) would also increase. Implementation of the rules in more sophisticated classifiers, such as the *k*-NN algorithm, will improve the performance, as suggested by preliminary studies.

CONCLUSION

Stable emerging patterns (SEPs) have been designed to discover new relationships between molecular structural features and the toxicological behavior of a molecule. The computation of these patterns from a molecular data set has been achieved by means of a sophisticated workflow that integrates a graph mining tool with a well-established measurement of the contrast between classes and with the stability of a pattern, a new notion from the domain of formal concept analysis.

The methodology has been practically applied to a well-tried benchmark data set in order to study mutagenicity. The extracted SEPs have been assessed through both quantitative examination and chemical expertise. The results show that these patterns generalize very efficiently: their quality is preserved from the training set to the test set. Moreover, the SEPs cover a large scope of different relationships between a molecular structure and its mutagenicity. It follows that the SEPs, when they are used alone as association rules, reach a fair level of confidence on a test set. The chemical analysis has shown that SEPs demonstrate a high ability to express structural alerts. Several SEPs will be further studied, and they may define new structural alerts.

As a conclusion, SEPs represent an advance in the automatic extraction of structural relations between molecular structures and a given activity. As a technical innovation, they offer several promising future works. For example, these SEPs may enter into the description of a molecule used by a prediction tool, increasing both the efficiency of the predictions and their explanatory power. This prediction tool will help us to select molecules from an in-house chemical library to be biologically assessed using the Ames test.

ASSOCIATED CONTENT

Supporting Information

A short discussion about the tuning of the parameters of our method and a zip file containing the complete list of closed frequent fragments and the complete list of stable emerging patterns extracted from Hansen's data set. The archive contained in the zip file is also available free of charge at https://chemoinfo.greyc.fr/2014_Metivier. The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/ci500611v.

AUTHOR INFORMATION

Corresponding Author

*E-mail: bertrand.cuissart@unicaen.fr.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank the Regional Council of Basse Normandie and the European Regional Development Fund (ERDF) for their financial support. S.K. was also supported by the project "Data Mining Based on Applied Ontologies and Lattices of Closed Descriptions", funded by the Basic Research Program of the National University Higher School of Economics.

REFERENCES

(1) Pearl, G. M.; Livingston-Carr, S.; Durham, S. K. Integration of Computational Analysis as a Sentinel Tool in Toxicological Assessments. *Curr. Top. Med. Chem.* **2001**, *1*, 247−255.

(2) Muster, W.; Breidenbach, A.; Fischer, H.; Kirchner, S.; Müller, L.; Pähler, A. Computational Toxicology in Drug Development. *Drug Discovery Today* **2008**, *13*, 303−310.

(3) Kruhlak, N. L.; Contrera, J. F.; Benz, R. D.; Matthews, E. J. Progress in QSAR Toxicity Screening of Pharmaceutical Impurities and Other FDA Regulated Products. *Adv. Drug Delivery Rev.* **2007**, *59*, 43−55.

(4) Kavlock, R. J.; Ankley, G.; Blancato, J.; Breen, M.; Conolly, R.; Dix, D.; Houck, K.; Hubal, E.; Judson, R.; Rabinowitz, J.; Richard, A.; Setzer, R. W.; Shah, I.; Villeneuve, D.; Weber, E. Computational Toxicology—A State of the Science Mini Review. *Toxicol. Sci.* **2008**, *103*, 14−27.

(5) European Commission. REACH: Registration, Evaluation, Authorisation and Restriction of Chemicals (2007). http://ec.europa.eu/enterprise/sectors/chemicals/reach/index.htm (accessed Oct 1, 2014).

(6) Rogers, M. D. The European Commission's White Paper "Strategy for a Future Chemicals Policy": A Review. *Risk Anal.* **2003**, *23*, 381−388.

(7) Ridings, J. E.; Barratt, M. D.; Cary, R.; Earnshaw, C. G.; Eggington, C. E.; Ellis, M. K.; Judson, P. N.; Langowski, J. J.; Marchant, C. A.; Payne, M. P.; Watson, W. P.; Yih, T. D. Computer Prediction of Possible Toxic Action from Chemical Structure: An Update on the DEREK System. *Toxicology* **1996**, *106*, 267−279.

(8) Sanderson, D. M.; Earnshaw, C. G. Computer Prediction of Possible Toxic Action from Chemical Structure—The DEREK System. *Hum. Exp. Toxicol.* **1991**, *10*, 261−273.

(9) Smithing, M. P.; Darvas, F. HazardExpert—An Expert System for Predicting Chemical Toxicity. *ACS Symp. Ser.* **1992**, *484*, 191−200.

(10) Lai, D. Y.; Woo, Y.-T. A Mechanism-Based Expert System for Predicting the Carcinogenic Potential of Chemicals. In *Predictive Toxicology*; CRC Press: Boca Raton, FL, 2005; pp 385−413.

(11) Woo, Y.; Lai, D. Y.; Argus, M. F.; Arcos, J. C. Development of Structure−Activity Relationship Rules for Predicting Carcinogenic Potential of Chemicals. *Toxicol. Lett.* **1995**, *95*, 219−228.

(12) Klopman, G. J. Artificial Intelligence Approach to Structure−Activity Studies: Computer Automated Structure Evaluation of Biological Activity of Organic Molecules. *J. Am. Chem. Soc.* **1984**, *106*, 7315−7321.

(13) ADMET and Predictive Toxicology. http://accelrys.com/products/discovery-studio/admet.html (accessed Oct 1, 2014).

(14) Helma, C. Lazy Structure−Activity Relationships (lazar) for the Prediction of Rodent Carcinogenicity and *Salmonella* Mutagenicity. *Mol. Diversity* **2006**, *10*, 147−158.

(15) Lagunin, A.; Stepanchikova, A.; Filimonov, D.; Poroikov, V. PASS: Prediction of Activity Spectra for Biologically Active Substances. *Bioinformatics* **2000**, *16*, 747−748.

(16) Ashby, J.; Tennant, R. W. Definitive Relationships among Chemical Structure, Carcinogenicity and Mutagenicity for 301 Chemicals Tested by the U.S. NTP. *Mutat. Res.* **1991**, *257*, 229−306.
(17) Benigni, R.; Bossa, C. Mechanisms of Chemical Carcinogenicity and Mutagenicity: A Review with Implications for Predictive Toxicology. *Chem. Rev.* **2011**, *111*, 2507−2536.
(18) Benigni, R.; Bossa, C.; Jeliazkova, N.; Netzeva, T.; Worth, W. *The Benigni/Bossa Rulebase for Mutagenicity and Carcinogenicity—A Module of Toxtree*; Institute for Health and Consumer Protection, European Commission Joint Research Centre: Ispra, Italy, 2008.
(19) Van Leeuwen, K.; Schultz, T. W.; Henry, T.; Diderich, B.; Veith, G. D. Using Chemical Categories To Fill Data Gaps in Hazard Assessment. *SAR QSAR Environ. Res.* **2009**, *20*, 207−220.
(20) Guzelian, P. S.; Victoroff, M. S.; Halmes, N. C.; James, R. C.; Guzelian, C. P. Evidence-Based Toxicology: A Comprehensive Framework for Causation. *Hum. Exp. Toxicol.* **2005**, *24*, 161−201.
(21) Valerio, L. G. J. In Silico Toxicology for the Pharmaceutical Sciences. *Toxicol. Appl. Pharmacol.* **2009**, *241*, 356−370.
(22) Nijssen, S.; Kok, J. N. The Gaston Tool for Frequent Subgraph Mining. *Electron. Notes Theor. Comput. Sci.* **2005**, *127*, 77−87.
(23) Yan, X.; Han, J. gSpan: Graph-Based Substructure Pattern Mining. *Proc. 2002 IEEE Int. Conf. Data Min.* **2002**, 721−724 DOI: 10.1109/ICDM.2002.1184038.
(24) Kazius, J.; McGuire, R.; Bursi, R. Derivation and validation of toxicophores for mutagenicity prediction. *J. Med. Chem.* **2005**, *48*, 312−320.
(25) Ahlberg, E.; Carlsson, L.; Boyer, S. Computational Derivation of Structural Alerts from Large Toxicology Data Sets. *J. Chem. Inf. Model.* **2014**, *54*, 2945−2952.
(26) Faulon, J.; Visco, D.; Pophale, R. The signature molecular descriptor. 1. Using extended valence sequences in QSAR and QSPR studies. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 707−720.
(27) Faulon, J.; Churchwell, C.; Visco, D. The signature molecular descriptor. 2. Enumerating molecules from their extended valence sequences. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 721−734.
(28) Helma, C.; Cramer, T.; Kramer, S.; De Raedt, L. Data Mining and Machine Learning Techniques for the Identification of Mutagenicity Inducing Substructures and Structure Activity Relationships of Noncongeneric Compounds. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1402−1411.
(29) Mannila, H.; Toivonen, H. Levelwise search and borders of theories in knowledge discovery. *Data Min. Knowl. Discovery* **1997**, *1*, 241−258.
(30) Kazius, J.; Nijssen, S.; Kok, J. N.; Bäck, T.; Ijzerman, A. P. Substructure Mining Using Elaborate Chemical Representation. *J. Chem. Inf. Comput. Sci.* **2006**, *46*, 597−605.
(31) *Contrast Data Mining: Concepts, Algorithms, and Applications*; Dong, G., Bailey, J., Eds.; CRC Press: Boca Raton, FL, 2013.
(32) Dong, G.; Li, J. Efficient Mining of Emerging Patterns: Discovering Trends and Differences. *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Min., 5th* **1999**, 43−52 DOI: 10.1145/312129.312191.
(33) Auer, J.; Bajorath, J. Emerging Chemical Patterns: A New Methodology for Molecular Classification and Compound Selection. *J. Chem. Inf. Model.* **2006**, *46*, 2502−2514.
(34) Blinova, V.; Dobrynin, D.; Finn, V.; Kuznetsov, S.; Pankratova, E. Toxicology analysis by means of the JSM-method. *Bioinformatics* **2003**, *19*, 1201−1207.
(35) Sherhod, R.; Gillet, V. J.; Judson, P. N.; Vessey, J. D. Automating Knowledge Discovery for Toxicity Prediction Using Jumping Emerging Pattern Mining. *J. Chem. Inf. Model.* **2012**, *52*, 3074−3087.
(36) Sherhod, R.; Judson, P. N.; Hanser, T.; Vessey, J. D.; Webb, S. J.; Gillet, V. J. Emerging Pattern Mining To Aid Toxicological Knowledge Discovery. *J. Chem. Inf. Model.* **2014**, *54*, 1864−1879.
(37) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom pairs as molecular features in structure−activity studies: Definition and applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.
(38) Coquin, L.; Canipa, S. J.; Drewe, W. C.; Fisk, L.; Gillet, V. J.; Patel, M.; Plante, J.; Sherhod, R. J.; Vessey, J. D. New structural alerts for Ames mutagenicity discovered using emerging pattern mining techniques. *Toxicol. Res.* **2015**, *4*, 46−56.
(39) 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.
(40) Poezevara, G.; Cuissart, B.; Crémilleux, B. Extracting and Summarizing the Frequent Emerging Graph Patterns from a Dataset of Graphs. *J. Intell. Inf. Syst.* **2011**, *37*, 333−353.
(41) Cuissart, B.; Poezevara, G.; Crémilleux, B.; Lepailleur, A.; Bureau, R. Emerging Patterns as Structural Alerts for Computational Toxicology. In *Contrast Data Mining: Concepts, Algorithms, and Applications*; Dong, G., Bailey, J., Eds.; CRC Press: Boca Raton, FL, 2013; pp 269−282.
(42) Hansen, K.; Mika, S.; Schroeter, T.; Sutter, A.; ter Laak, A.; Steger-Hartmann, T.; Heinrich, N.; Müller, K.-R. Benchmark Data Set for in Silico Prediction of Ames Mutagenicity. *J. Chem. Inf. Model.* **2009**, *49*, 2077−2081.
(43) Kuznetsov, S. O. Stability as an Estimate of the Degree of Substantiation of Hypotheses on the Basis of Operational Similarity. *Autom. Doc. Math. Linguist.* **1990**, *24*, 62−75.
(44) Kuznetsov, S. O. On Stability of a Formal Concept. *Ann. Math. Artif. Intell.* **2007**, *49*, 101−115.
(45) Roth, C.; Obiedkov, S.; Kourie, D. G. On Succinct Representation of Knowledge Community Taxonomies with Formal Concept Analysis. *Int. J. Found. Comput. Sci.* **2008**, *19*, 383−404.
(46) Davey, B. A.; Priestley, H. A. *Introduction to Lattices and Order*, 2nd ed.; Cambridge University Press: Cambridge, U.K., 2002.
(47) Ganter, B.; Wille, R. *Formal Concept Analysis: Mathematical Foundations*; Springer: Berlin, 1999.
(48) Kuznetsov, S. O.; Samokhin, M. V. Learning Closed Sets of Labeled Graphs for Chemical Applications. *Lect. Notes Comput. Sci.* **2005**, *3625*, 190−208.
(49) Hastie, T.; Tibshirani, R.; Friedman, J. Unsupervised Learning. *The Elements of Statistical Learning*; Springer Series in Statistics; Springer: New York, 2009; Chapter 14, pp 485−585.
(50) Kuznetsov, S. O.; Obiedkov, S. A.; Roth, C. Reducing the Representation Complexity of Lattice-Based Taxonomies. *Lect. Notes Comput. Sci.* **2007**, *4604*, 241−254.
(51) Borgelt, C. Combining Ring Extensions and Canonical Form Pruning. *Proc. Workshop Min. Learn. Graphs, 4th* **2006**, 109−116.
(52) Cordella, L. P.; Foggia, P.; Sansone, C.; Vento, M. Performance Evaluation of the VF Graph Matching Algorithm. *Proc. Int. Conf. Image Anal. Processing* **1999**, 1172−1177 DOI: 10.1109/ICIAP.1999.797762.
(53) Kuznetsov, S. O. Learning of Simple Conceptual Graphs from Positive and Negative Examples. *Lect. Notes Comput. Sci.* **1999**, *1704*, 384−391.
(54) Buzmakov, A.; Kuznetsov, S. O.; Napoli, A. Scalable Estimates of Concept Stability. *Lect. Notes Comput. Sci.* **2014**, *8478*, 157−172.
(55) Chemical Carcinogenesis Research Information System. http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?CCRIS.
(56) Feng, J.; Lurati, L.; Ouyang, H.; Robinson, T.; Wang, Y.; Yuan, S.; Young, S. S. Predictive Toxicology: Benchmarking Molecular Descriptors and Statistical Methods. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1463−1470.
(57) Judson, P. N.; Cooke, P. A.; Doerrer, N. G.; Greene, N.; Hanzlik, R. P.; Hardy, C.; Hartmann, A.; Hinchliffe, D.; Holder, J.; Müller, L.; Steger-Hartmann, T.; Rothfuss, A.; Smith, M.; Thomas, K.; Vessey, J. D.; Zeiger, E. Towards the Creation of an International Toxicology Information Centre. *Toxicology* **2005**, *213*, 117−128.
(58) Genetic Toxicology Database (GENE-TOX). http://toxnet.nlm.nih.gov/newtoxnet/genetox.htm.
(59) Ames, B. N.; Lee, F. D.; Durston, W. E. An Improved Bacterial Test System for the Detection and Classification of Mutagens and Carcinogens. *Proc. Natl. Acad. Sci. U.S.A.* **1973**, *70*, 782−786.

DOI: 10.1021/ci500611v
J. Chem. Inf. Model. 2015, 55, 925−940

(60) Martin, T. M.; Harten, P.; Young, D. M.; Muratov, E. N.; Golbraikh, A.; Zhu, H.; Tropsha, A. Does Rational Selection of Training and Test Sets Improve the Outcome of QSAR Modeling? *J. Chem. Inf. Model.* **2012**, *52*, 2570−2578.
(61) Tropsha, A.; Golbraikh, A. Predictive QSAR Modeling Workflow, Model Applicability Domains, and Virtual Screening. *Curr. Pharm. Des.* **2007**, *13*, 3494−3504.
(62) Hanley, J. A.; McNeil, B. J. The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve. *Radiology* **1982**, *143*, 29−36.
(63) Seal, A.; Passi, A.; Jaleel, U. C. A.; Wild, D. J. In-Silico Predictive Mutagenicity Model Generation Using Supervised Learning Approaches. *J. Cheminf.* **2012**, *4*, No. 10.
(64) Xu, C.; Cheng, F.; Chen, L.; Du, Z.; Li, W.; Liu, G.; Lee, P. W.; Tang, Y. In Silico Prediction of Chemical Ames Mutagenicity. *J. Chem. Inf. Model.* **2012**, *52*, 2840−2847.
(65) Sushko, I.; Salmina, E.; Potemkin, V. A.; Poda, G.; Tetko, I. V. ToxAlerts: A Web Server of Structural Alerts for Toxic Chemicals and Compounds with Potential Adverse Reactions. *J. Chem. Inf. Model.* **2012**, *52*, 2310−2316.
(66) Bissell-Siders, R.; Cuissart, B.; Crémilleux, B. On the Stimulation of Patterns. *Lect. Notes. Comput. Sci.* **2010**, *6208*, 56−69.

■ NOTE ADDED AFTER ASAP PUBLICATION

There was an error in Figure 10 in the version published ASAP May 7, 2015; the corrected version was published ASAP May 12, 2015.

DOI: 10.1021/ci500611v
J. Chem. Inf. Model. 2015, 55, 925−940