

AUTOMATED DETECTION OF STRUCTURAL ALERTS (CHEMICAL FRAGMENTS) IN (ECO)TOXICOLOGY

Alban Lepailleur^{a,b}, Guillaume Poezevara^{a,c}, Ronan Bureau^{a,b,*}

Abstract: This mini-review describes the evolution of different algorithms dedicated to the automated discovery of chemical fragments associated to (eco)toxicological endpoints. These structural alerts correspond to one of the most interesting approach of *in silico* toxicology due to their direct link with specific toxicological mechanisms. A number of expert systems are already available but, since the first work in this field which considered a binomial distribution of chemical fragments between two datasets, new data miners were developed and applied with success in chemoinformatics. The frequency of a chemical fragment in a dataset is often at the core of the process for the definition of its toxicological relevance. However, recent progresses in data mining provide new insights into the automated discovery of new rules. Particularly, this review highlights the notion of Emerging Patterns that can capture contrasts between classes of data.

MINI REVIEW ARTICLE

Introduction

Nowadays, the understanding of the chemical risks to health and the environment represents a hot topic and *in silico* toxicology is extremely appealing because of its high throughput, its inexpensiveness, and its capacity to reduce the use of animals. By the way, according to regulators, *in silico* techniques, like Quantitative Structure-Activity Relationships (QSARs), read-across, and structural alerts have a foot in the door in the assessment of chemicals.¹⁻³ The definition of structural alerts corresponds to one of the most interesting approach whose main advantage is the identification of chemicals with common mechanism of action. Indeed, investigators have always been interested in the structural and physicochemical basis of the biological behavior of chemicals. A well-known example is the Tennant and Ashby's set which defines structural alerts for DNA reactivity based on the analysis of *in vitro* mutagenicity and *in vivo* carcinogenicity data.⁴ This set of alerts has been largely superseded by others that incorporated and extended them. To date, one of the most advanced lists for evaluating the mutagenic and carcinogenic potential of chemicals is the list proposed by Benigni and Bossa,⁵ which has been implemented as a rule-based system in Toxtree⁶ and in the OECD QSAR Toolbox.⁷ Derek Nexus^{8, 9} is another example of expert system which associates structural alerts (generalized structural features in this case, like substituted vinyl ketone, 2,5-Dihalothiophene, or alkylating agent) with various toxicological endpoints (e.g. mutagenicity/carcinogenicity, skin/ocular irritation, or skin sensitization). These systems do not discover new associations,

but rather store knowledge from human experts and the scientific literature, and often use a reasoning model to make a prediction. However, the expand of the knowledge base is very time consuming since it requires strong investment of domain experts and a detailed analysis of the literature. Thus, the evolution of artificial intelligence and data mining tools should benefit to the reduction of time and efforts needed to identify new structural alerts, sometimes beyond the limits of human perception. This mini-review describes different programs leading to automated detection of such information from a dataset partitioned into two data classes. These programs range from commercially available expert systems to fundamental research tools, based on algorithms that are sometimes under development.

It is impossible to make an exhaustive view of all the approaches leading to structural alerts in this mini review. We have made the difficult choice to discard the methods involving predefined fragments (chemical fingerprint) or a fix length of chemical fragments (hashed chemical fingerprint). So, some very interesting works like those carried out by Scheiber et al. (ECFP4)¹⁰ and Pauwels et al. (Pubchem fingerprint)¹¹ are not described in this review. Two exceptions were done for the studies associated to the recent notion of Emerging Patterns (*vide infra*).

Expert systems based on data mining approaches

CASE,¹² *MultiCASE*¹³

Historically, CASE is the first program which extracts chemical fragments from the comparative analysis of two chemical datasets. This approach identifies the most relevant descriptors, using automated algorithms, and creates expert systems capable of recognizing the existence of structural alerts in new chemicals. The program consists in tabulating, for each molecule, different fragments by breaking up the molecule into linear subunits containing between 3 and 12 interconnected heavy atoms. All fragments belonging to an active molecule are labeled active while those belonging to an inactive molecule are labeled inactive. Once all molecules have been entered, a statistical analysis of the fragment distribution is made. A binomial distribution is assumed, and each type of fragment is considered

^aNormandie Univ, France

^bUNICAEN, CERMN (Centre d'Etudes et de Recherche sur le Médicament de Normandie, FR CNRS INC3M - SF ICORE, Université de Caen Basse-Normandie, U.F.R. des Sciences Pharmaceutiques), F-14032 Caen, France

^cUNICAEN, GREYC (Groupe de Recherche en Informatique, Image, Automatique et Instrumentation de Caen, CNRS UMR 6072, Université de Caen Basse-Normandie), F-14032 Caen, France

* Corresponding author.

E-mail address: ronan.bureau@unicaen.fr (Ronan Bureau)

irrelevant if its distribution among actives and inactives is the same as that of the total sample of molecules. Any significant discrepancy from a random distribution of subunits between the active and inactive chemical derivatives is taken as an indication that the fragment is relevant. It is labeled as activating (biophore) if its distribution is skewed toward active molecules and inactivating (biophobe) otherwise. For example, from a training set consisting of 39 cyclic *N*-nitrosamines tested on rats (27 active carcinogens and 12 inactive compounds), two biophores and one biophobe were found by the program to have a better than 98% chance of being related to activity (Figure 1). For a new molecule, the program compares its fragments to those that are held in memory and defines a probability that the new molecule is active or not.

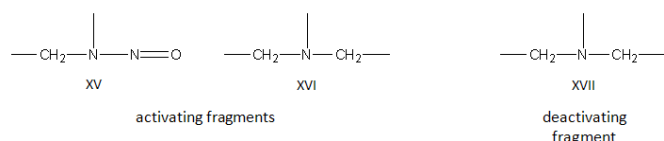


Figure 1. Fragments related to the rat carcinogenicity of cyclic *N*-nitrosamines according to Klopman *et al.*¹²

Concerning MultiCASE, the program is based on a hierarchical statistical analysis of a database. Like CASE, the program aims to discover chemical fragments that appear mostly in active molecules. It starts by identifying the statistically most significant fragment existing within the learning set. This first fragment is labeled as the top biophore, responsible for the activity of the largest possible number of active molecules. The active molecules containing this fragment are then removed from the database and the remaining ones are submitted to a new analysis leading to the identification of the next most significant fragment. This procedure is repeated until either the activity of all the molecules in the learning set have been accounted for or no additional statistically significant substructure can be found. For each set of molecules containing a specific biophore, MultiCASE identifies additional parameters, named modulators, which consists of the presence of certain fragments or the value of calculated parameters such as HOMO and LUMO energies, octanol–water partition coefficient, and so on. Finally, MultiCASE proposes fragment-based QSAR models by applying a QSAR methodology for each group of molecules containing a specific biophore.

We can also mention that the same team has implemented the CASE/MultiCASE biophores in a genetic artificial neural network (GA-ANN). This new computer program was called Expert System Prediction (ESP).¹⁴ The purpose was to evaluate the significance of the biophores from a different point of view. The neural network learns the relationships between the patterns (represented in the form of a pattern vector) and the activities of the chemicals, and this knowledge is later used for activity prediction of new molecules. The effectiveness of the ESP approach was illustrated by studying the carcinogenicity of a diverse set of chemicals.

PASS^{15, 16}

The PASS program is based on a regression approach that provides predictions from the SAR analysis of a training set containing more than 30000 compounds. This noncongeneric database encompasses more than 500 different biological activities. The molecules were represented by "Multilevel Neighborhoods of Atoms" (MNA) descriptors,¹⁷ which are based on their 2D representation. Briefly, an MNA descriptors set is subdivided on levels and generated recursively. A zero-level MNA descriptor describes the

atom itself and any next level MNA descriptor is the substructure notation A(D₁D₂...), where A is the atom A descriptor, and D_i is the previous level MNA descriptor of the *i*th neighbor atom for the atom A. To estimate the activity for a new compound, its MNA descriptors have to be generated and then, the probabilities of belonging to the classes of active and inactive compounds are calculated.

Cat-SAR¹⁸

Starting from the Tripos Sybyl HQSAR module, each chemical is fragmented into all possible substructures. HQSAR allows the user to select attributes for fragment determination including atom count, bond types, atomic connections, hydrogen atoms, chirality, and hydrogen bond donor and acceptor groups. Fragments can be linear, branched, or cyclic moieties. Models developed contained fragments between three and seven atoms and considered atoms, bond types, and atomic connections. To ascertain an association between each fragment and activity (or inactivity), the first selection rule is the number of times a fragment is identified. The second rule relates to the proportion of active or inactive compounds that contribute to each fragment. Chemical fragments are considered meaningful if they are found in at least three compounds in the learning set and are comprised of either 90% or more active or inactive compounds. To make a prediction for a new compound, the cat-SAR program determined which fragments, from the model's pool of significant fragments, the test compound contains. If none were present, no prediction of activity was made. If one or more fragments were present, the number of active and inactive compounds containing each fragment was determined. The probability of activity or inactivity was then calculated based on the total number of active and inactive compounds that went into deriving each of the fragments. For example (Figure 2), 4-aminodiphenyl was predicted to be active as a mammary carcinogen with a probability of 100%. This prediction was based on the occurrence of four similar fragments derived from nine carcinogens and zero noncarcinogens in the model's learning set. However, the example clearly highlights a redundancy issue with this program.

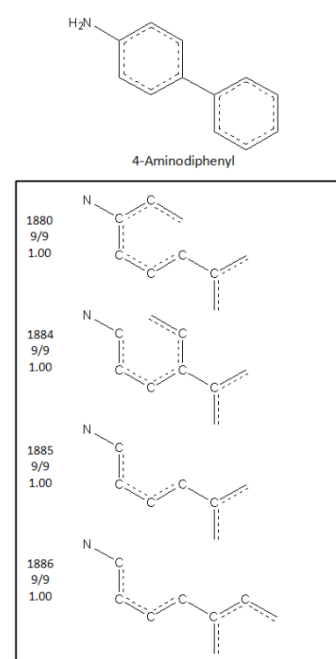


Figure 2. Illustration of significant fragments used to predict the carcinogenic potential of 4-aminodiphenyl adapted from Cunningham *et al.*¹⁸

*LAZAR/MOLFEA*¹⁹

LAZAR, a tool for predictive toxicology, uses a simplified version of the MOLFEA algorithm, an inductive database system tailored toward discovering substructures within sets of small molecules. The considered substructures are linear molecular fragments, i.e. sequences of atoms and connecting bonds, and are based on a subset of the SMARTS language. Inductive databases are databases that can be queried not only for data but also for patterns and regularities (chemical fragments in this case) that occur within the data and fulfill a specific constraint defined by the user. For example, a frequency constraint could require that the fragment occurs in at least (respectively at most) $x\%$ of the molecules belonging to a given data set. A query in MOLFEA is composed of several conditions, each of which has to be fulfilled in order to make it a solution fragment. To efficiently compute the fragments that satisfy a given query, the generality relation on fragments is exploited. This property imposes two borders, a lower and an upper. A lower border (called the S -set) on the space of possible solutions contains all maximally specific fragments that satisfy the constraint. It was called a border because all fragments more general than an element of S will also satisfy the constraint and all fragments that are not more general than at least one fragment in S will not satisfy the constraint. Dually, the frequency constraint imposes also an upper border (called the G -set in their publication) on the space of possible solutions. To update the borders, MOLFEA applies the following algorithm. First, those fragments that do not satisfy the minimum constraint are deleted. Second, the elements of S are updated using a levelwise search algorithm. This algorithm keeps track of a list of candidates C_i and a list of solutions L_i to the frequency constraint. Both lists are initialized with the maximally general element T and iteratively updated. During each iteration, the candidates (fragments) associating an atom type and a bond type at level i are computed by refining existing fragments at level $i-1$. Those candidates satisfying the frequency threshold are retained and used to generate the candidates in the next iteration. The process is continued until no more candidates can be generated. At this point, the S set is computed by taking the maximally specific elements among the solutions computed that are more specific than an element of G . As an example, the authors identified relevant fragments (Figure 3) from a data set containing 341 mutagenic and 343 nonmutagenic compounds using the query $(freq(f, mutagens) \geq t) \wedge (freq(f, nonmutagens) \leq t)$, where t varied from 0.01 (1%, 6 compounds) to 0.10 (10%, 68 compounds). Three machine learning algorithms (C4.5, PART, SVM) were used to learn SAR models from these fragments.

Developments of graph data mining algorithms based on a frequency constraint

The tools described above have been implemented in expert system software which are generally commercial products. However, due to the evolution of the modern information methods and technology, collecting, combining, storing, and mining huge amounts of data can be done at very low costs. Indeed, several works have been developed in informatics to extract the frequent subgraphs from a dataset. Apply to the field of the chemoinformatics, those works allow to extract the frequent substructures. The frequency constraint is popular in data mining for its anti-monotonic²⁰ property : if a substructure sub of size n (the size is the number of atoms) is not frequent in a dataset of molecular graphs, it means that all the substructure of size $n+1$ whose contain sub are not frequent in this dataset. The frequency is useful as it efficiently disregards infrequent substructures. However, as a simple remark, if according to the

frequency constraint there are too few compounds that contain oxygen, then peroxide containing substructures are not explored.

Several families of algorithms exist for extracting the frequent subgraphs from a dataset of graphs²¹. The Apriori approach²² used a Breadth-First search strategy to cross the search space associated to the frequent subgraphs. The discovery of the frequent atoms in the dataset is the first step, then it iterates until no more frequent substructures are discovered by increasing by one the size of the candidates. At each iteration, the substructures of size $n+1$ are obtained thanks to the fusion of the frequent substructures of size n . A fusion process merges two frequent substructures of size n whose differ by only one atom and provides all the substructures resulting from this fusion. By testing all the possible fusions, all the candidates of size $n+1$ are generated. Finally, the frequency is checked and only the frequent ones are kept. The initial algorithm (named AGM) to extract the frequent subgraphs from a dataset of graphs belongs to this family. The Pattern-Growth Based approach used a Depth-First search strategy to cross the search space of the candidates. A frequent substructure of size n is extended by adding a new chemical bond leading to a candidate substructure of size $n+1$. Only the substructures respecting the frequency constraint are stored. The first step of this method consists to collect all the frequent atoms, then each atom are extended until no more frequent subgraphs are generated. Only substructures with at least two atoms are stored. We described here applications exploiting the Pattern-Growth Based approach.

```

1.6274328192175296 * c:c:c:c:c:c:c:c:c
1.4455302626881337 * C-Cl
1.3226667063998578 * C-C-C-C-N-C
1.310524380418045 * C-C-C-O
0.9516054404252757 * C-C=C
0.8654786477941714 * c:c:c:c:c:n
0.8243351055367271 * C-C-C-C=C
0.8197902253156605 * C-C-C-N-C
0.7969086522621357 * c:c:c-C=O
0.7819601605449131 * C-N-C
0.7796980414561107 * N-N
0.7498673413287917 * C-C-C-C-O
0.7276759799450657 * C-C-N-N
0.727514353351238 * N-O
0.7167965121501293 * C-O-C
0.6784153103780268 * C
0.6744410897500348 * C-N-c:c:c:c:c:c
0.6744410897500348 * C-N-c:c:c:c:c:c
0.6716119052528489 * c:c-N
0.5686660779334143 * C-C-N

```

Figure 3. The 20 strongest activating fragments (threshold of 0.05) for Salmonella mutagenicity derived from linear SVM, and according to Helma *et al.*¹⁹

*Gaston*²³

The algorithm $gSpan^{24}$ is one of the first using the Pattern-Growth Based approach²⁵ to generate the substructures. It is often used in chemoinformatics for two reasons : (i) it uses the quickstart principle : the set of the frequent substructures could be partitioning into three subsets, the subset of the paths (an atom cannot be linked

to more than two other atoms), the subset of the trees (there is no cycle) and the subset of the graphs whose contain cycles and (ii) the authors make available its source code. For an illustration of the potential of Gaston, an elaborate method of graph-based chemical representation was developed by Kazius *et al.*,²⁶ and tested for extracting substructures from a mutagenicity data set (4337 entries). At the start of this process, the initial substructure is mapped everywhere it fits into every molecule in the data set. For each such mapping, the atoms at the neighboring positions of this substructure are stored. For each substructure in such a collection, its statistical association with mutagenicity, expressed as the p-value, was determined from the amounts of mutagens and nonmutagens. It was then determined which substructure was most strongly associated with mutagenicity, that is, which substructure possessed the lowest p-value. This substructure was then selected to split the chemical data set into two subsets (linear decision trees). Each split generated one subset of compounds that contain this substructure and another subset of compounds that lack this substructure. This latter subset was used to recompute the p-values of all substructures. From these p-values, the next most mutagenic substructure was determined and then used to split this chemical subset in two, and so on. After six splits, all compounds from the original database are divided over seven subsets. The result is illustrated in Figure 4. The statistics of mutagenicity prediction based on 10 fold cross-validation showed a sensitivity of 83% and a specificity of 74%.

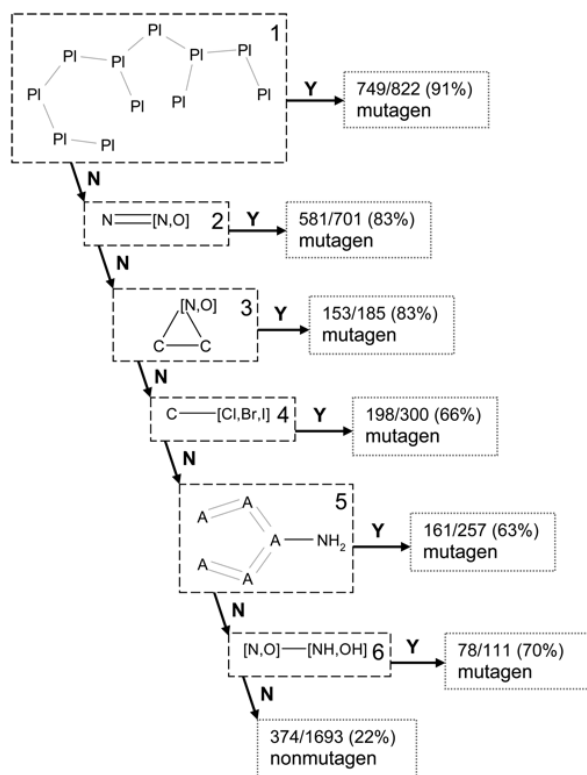


Figure 4. Decision list extracted from a mutagenicity data set according to Kazius *et al.*²⁶

MOSS/MoFa²⁷

Concerning this program, the association rules start from an algorithm similar to Eclat²⁸ (search trees). The process is well described in the publication and is illustrated on a set of molecules (Figure 5) with a minimum support (frequency) of 50%. First the sulfur atom is embedded forming the root of the search tree (Figure

6), and then the embeddings are extended in all possible ways. Of course, subtrees of the search tree are pruned if they refer to substructures not having enough support. This leads to the definition of six frequent substructures starting from the example molecules (Figure 7). A crucial step concerns the definition of the starting point (sulfur in this example). We can start with a different atom, as long as this atom is rare in the molecule, or a specific core like an aromatic ring with one or two side chains for instance. Contrast structures are then extracted corresponding to substructures that are frequent in a predefined subset of the molecules and infrequent in the complement of this subset. Experimental results concerned HIV-1 infection and they extracted nitrogen based, sulfur-based, and selenium based fragments particularly.

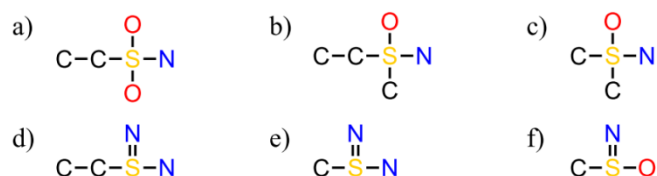


Figure 5. Set of six example molecules.

From frequent patterns to emerging patterns

In informatics, the recent studies of pattern mining have given more attention to the discovery of patterns that are "significant", "emerging", "dominant" and so forth, than simply frequent. Indeed, it appears that frequent patterns alone have a limited applicability and usability in terms of building predictive models.²⁹ Thus, methods of finding representative subsets of frequent patterns that could be effectively useful are appealing. In this part, we will focus on the notion of Emerging Patterns and its applications in the field of chemoinformatics.

Emerging patterns³⁰

Emerging Patterns (EPs) were introduced by Dong and Li. The emerging constraint captures differentiating characteristics between two classes of data. An EP is defined as an itemset which support (i.e. its number of occurrences in the dataset) increases significantly from one dataset D_1 to another D_2 . Dong and Li have proposed to use the growth rate measure to evaluate this increasing. The growth rate of a pattern *pat* from D_2 to D_1 is given by the ratio of the frequency of *pat* in D_1 over the frequency of *pat* in D_2 . In this way, EPs capture contrasts between two data classes. An interesting point on this project is to discover EPs with small support. The authors precise that it is a challenge due to two reasons: i) the useful anti-monotonic property no longer holds for EPs, and ii) there are usually too many candidates. Naïve algorithms, which consider all itemsets, are not viable since it would be too costly. However, such collections of itemsets have a nice property corresponding to the notion of closed intervals.³¹ If X and Z are in S and Y is a set such that $X \subseteq Y \subseteq Z$ then Y is in S . Thus, they described large interval-closed collections of itemsets using borders, defined as the pair of the sets of the minimal and the maximal itemsets. Clearly, borders are usually much smaller than the collections they represent. Such borders can be efficiently discovered by algorithms like Max-miner.³² For instance, on a mushroom dataset with a growth rate threshold of 2.5, 2²⁸ EPs are possible but they can be represented by only half a million borders. Otherwise, if the support of an EP in D_2 is null then this pattern was called a Jumping Emerging Pattern (JEP). A JEP is defined as the most expressive EP.³³

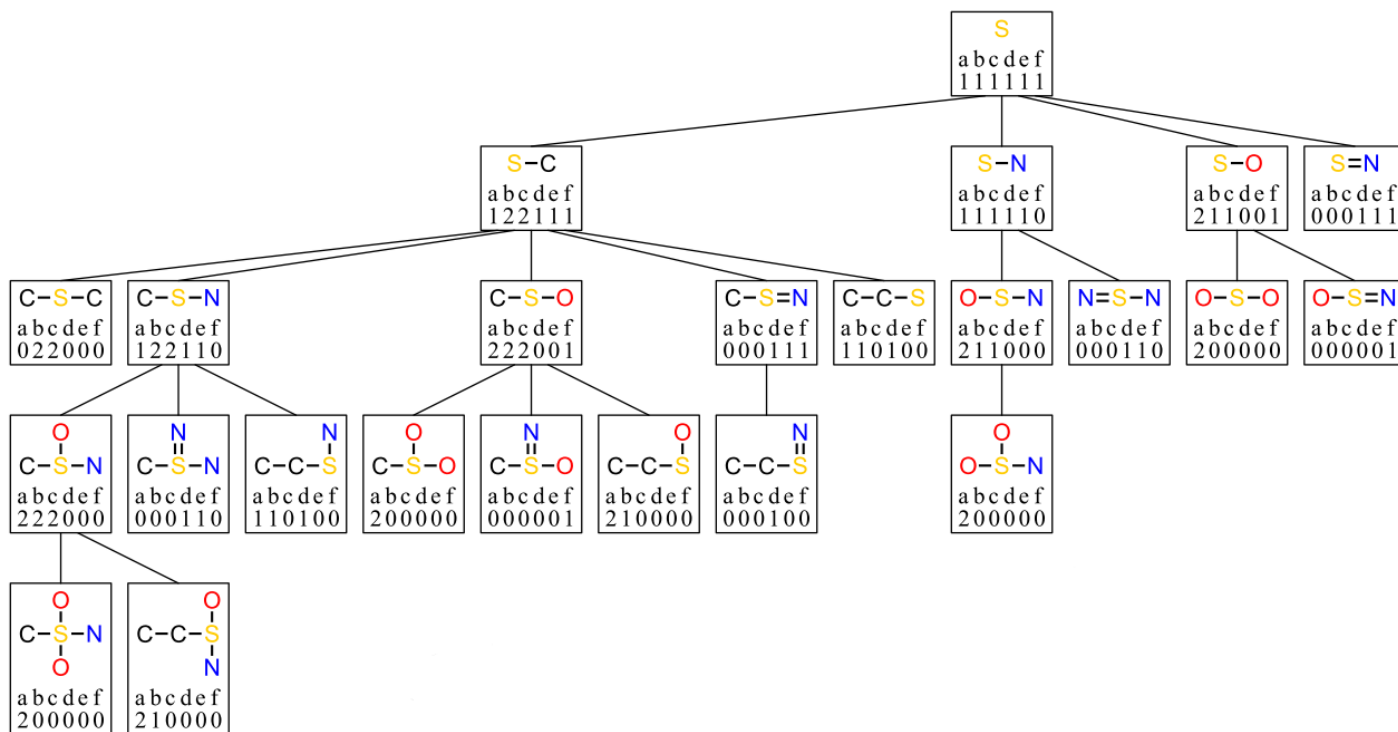


Figure 6. Search tree from the six previous molecules.

Table 1. Coverage rate of the JEFs on the H400 molecules of the learning set, and success rate of the prediction rule on the testing set.

		Frequency threshold (%)					
		5	4.3	3	2.6	1	0.6
Learning set	Support in H400 molecules	15	13	9	8	3	2
	Coverage rate on H400 (%)	34.3	41.5	60.4	62.9	81	84.3
	Coverage rate on H400 (dev)	6.43	4.9	3.83	2.7	1.27	0.74
Testing set	H400 success rate (%)	38.3	42.9	62.6	66.9	79	81.9
	H402 success rate (%)	95.8	94.3	85	81	55.8	47.1
	Overall success rate (%)	58.1	60.6	69.9	70.7	71	69.9

Table 2. FEMP *vs* RPMP in function of the growth rate.

Growth rate	Number of patterns			Length of patterns	
	FEMP	Closed FEMP	RPMP	Closed FEMP	RPMP
2	2 414 271 394	735	735	15.2	2.29
5	1 629 688 309	415	415	15.7	2.36
10	1 629 688 309	264	264	16.4	2.4
25	1 632 132 495	238	238	16.9	2.43
∞	1 632 131 769	236	236	16.9	2.43

Emerging Chemical Patterns³⁴

In 2006, Auer and Bajorath applied for the first time the concept of EPs in chemoinformatics. They introduced the notion of Emerging Chemical Patterns (ECP) as a novel approach to molecular classification. The authors used the subset of the JEP to conduct an experimental study. An hypergraph-based algorithm was applied to mine the JEPs from two classes of data, actives or inactives. The potential of this approach to classify derivatives was analyzed on four publicly available compound data sets. But in this case, they do not use molecular graphs, they used a set of sixty one 1D and 2D molecular descriptors with values ranges discretized into suitable intervals. On the basis of their results, ECPs are expected to broaden the spectrum of molecular classification methods, and complement computational methodologies like binary QSAR and decision trees.

Jumping Emerging Fragments³⁵

This study corresponds to the first application of EPs to a classification task in ecotoxicology. The authors assumed that the level of toxicity for a chemical may be influenced by the presence of a specific molecular fragment. Such a fragment has been called a Jumping Emerging Fragment (JEFs) since it has a strong foothold in the toxic chemicals and is missing from the non-toxic chemicals. A three-step algorithm that automatically extracts chemical fragments was designed. Let D be partitioned into two subsets D_1 and D_2 . The first step is to extract the frequent connected subgraphs in D_1 according to the frequency threshold. For this step, a Pattern-Growth Based algorithm was used instead of an Apriori Based algorithms.³⁶ According to an experimental comparison of four Pattern-Growth Based algorithms,³⁷ gSpan²⁴ was chosen for a question of memory and speed. The second step consists in defining for each graph G_D of D and for each connected graph G resulting from step I, if G is a subgraph of G_D . For that task, an in-house implementation of the J.R. Ullmann's algorithm³⁸ was used to solve the resulting multiple subgraph isomorphism problems. For the third step, the problem is described by items (presence or absence of each frequent connected graph) and Music-DFS algorithm³⁹ was used to discover JEFs. The authors applied this methodology to discover JEFs from H402 (harmful to aquatic life) to H400 (very toxic to aquatic life) chemicals, and parametrize the simplest possible decision rule based on these jumping fragments: a molecule is H400 just in case it contains a JEF. The analysis of the results obtained on the testing set, in function of the frequency and the coverage rate on H400 molecules (learning set) has shown that a JEF recorded at a high frequency threshold is meaningful to define the toxicity of a derivative (see the H402 success rate for 3-5%).

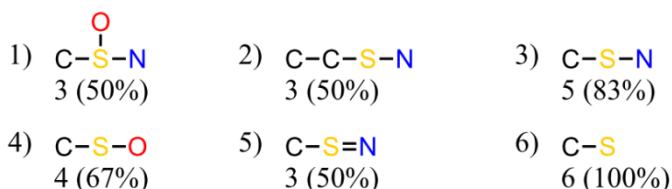


Figure 7. Frequent fragments found from the example of Borgelt *et al.*²⁷

Representative Pruned Molecular Patterns⁴⁰

In comparison with previous study, the authors consider now the conjunction of fragments, i.e. the combination of different moieties of a molecule in only one pattern. The new notion of Representative Pruned Molecular Patterns (RPMPs) was introduced. As a reminder,

when a dataset is partitioned into targeted examples and non-targeted ones (also called “classes”), the growth-rate of a pattern is defined as the ratio between its frequency in the target class over its frequency outside the target class. The first step consists in the enumeration of all the frequent and emerging molecular patterns (FEMPs) according to frequency and growth rate constraints. In practice, FEMPs are often numerous and include redundant information, but they could be condensed by applying the notion of closed pattern.⁴¹ A closed pattern is a pattern for which no element can be added without decreasing its extent, i.e. the set of molecules in which the molecular pattern occurs. Thus, by retaining only closed FEMPs the authors can condense the important set of FEMPs without losing information. However, closed patterns tend to be very long (number of fragments) since a large part corresponds to subfragments of a larger fragment. These redundant subfragments can be pruned without losing information. These resulting shorter representations have been called Representative Pruned Molecular Patterns (RPMPs). The illustrative example corresponds to a dataset of 295 chemicals annotated by their toxicity to aquatic life (223 toxic and 172 non-toxic chemicals). The Table 2 indicates the evolution of the number and length of the corresponding patterns in function of the growth rate values (minimum frequency threshold of 2.8%).

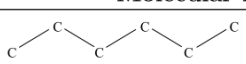
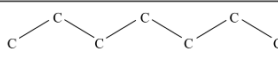
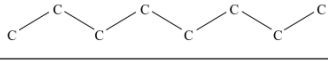
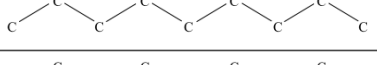
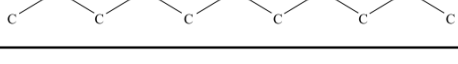
Growth-rate	Molecular fragment
2.7	
6.9	
11.5	
13.8	
∞	

Figure 8. Growth-rate values of the alkyl chains according to their length.

Besides, the interpretability of the RPMPs seems to be more obvious. A simple illustration deals with the impact of the length of an alkyl chain towards the ecotoxicity. The RPMPs show an evident relationship between the growth-rate values and the number of carbon atoms of the fragments (see Figure 8). The meaningful length of the alkyl chains begins for C6 (growth-rate of 2.7), it increases strongly for C7 (growth-rate of 6.9), to reach a maximum value for C11 (growth-rate of ∞ , corresponding to a Jumping Emerging Pattern). Thus, in terms of structure-activity relationships, this result highlights that the hydrophobicity of an alkyl chain is correlated with its length, and is in straight relation with its ecotoxic effect. Several other examples are given in the article.

Jumping Emerging Patterns⁴²

Recently, researchers of the Sheffield University collaborated with Derek Nexus developers to help automate the process of knowledge extraction from toxicity data sets. Their approach is based on the discovery of Jumping Emerging Patterns (JEPs). In this work, structural fingerprints are used as descriptors. The complete procedure for mining the JEPs is in 6 steps : i) generation of all atom pairs under user-defined constraints from the active compounds in the data set, ii) application of the Horizon-Miner algorithm to extract the maximal patterns for both the actives and the inactives, iii) application of the border-differential algorithm to mine the set of all possible

minimal JEPs, iv) reduction of the set of minimal JEPs to those that occur in distinct sets of actives, v) identification of the relationships between the supporting actives of minimal JEPs and arrangement of them into hierarchies, and vi) extraction of the maximum set of commonly occurring atom pairs from the set of actives that support each minimal JEP. The illustrative examples deal with Ames mutagenicity, oestrogenicity, and hERG channel inhibition end points. The method is effective to cluster 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. For example, Figure 9 highlights a JEP for well-known mutagenic alkylating agents.

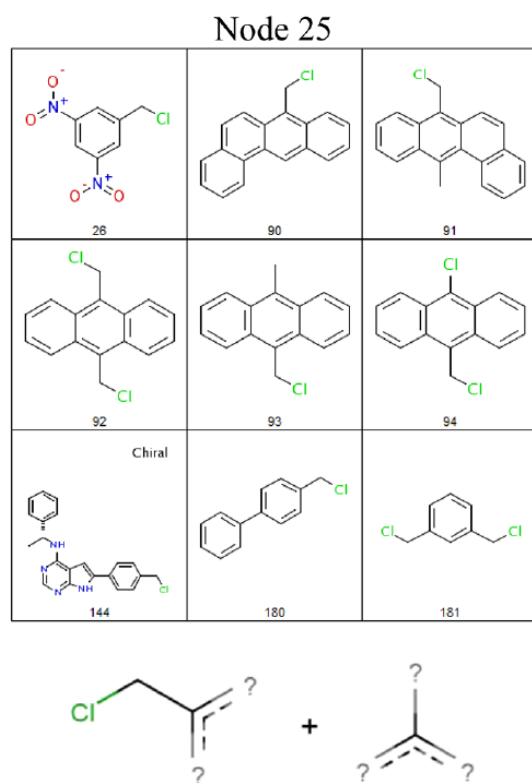


Figure 9. Example of JEPs (bottom) for mutagenic alkylating agents and supporting active compounds (top), according to Sherod *et al.*⁴²

Conclusion

The characterization of a chemical fragment or a chemical pattern associated to a toxicological profile is of first interest by considering the continuous development of chemoinformatics tools around this topic. Owing to the fact that a single chemical fragment is not always responsible for the overall toxicity of a chemical, the present objective is to analyze the combination of chemical fragments (chemical patterns) leading to an increase or a decrease of toxicity starting from a referential toxic fragment. The first described approach corresponds to the Klopman's method (CASE/MultiCASE) which extracts meaningful chemical fragments (named biophores) in function of their distribution between two datasets. Afterwards, these biophores were associated to QSARs or implemented in classification tools, leading to the first commercialized expert systems. Concerning the extraction of chemical fragments, a second generation of tools was more recently developed, corresponding to the Pattern Growth-Based algorithms (like Gaston). They led to very interesting results in terms

of characterization of the fragments and statistical results for the estimation of the toxicity. The last evolution corresponds to the search of representative subsets of frequent patterns. In this review, we emphasize the notion of Emerging Patterns (EP), whose extraction is based on the notion of contrast (growth rate) between two datasets. We are at the beginning for EPs but, their potential in terms of statistics and interpretability related to the toxicological profile of chemical derivatives is really promising.⁴³ The size of the EP set (number of patterns) in function of the size of the initial dataset is really an issue, and the notion of border does not seem to be sufficient to solve it. As described in the review, Representative Pruned Molecular Patterns (RPMPs) represents a first way to reduce this size without losing chemical information but to go further, an analysis of the relationships between the RPMPs must be carried out. This is underway.

Citation

Lepailleur A, Poezevara G, Bureau R (2013) Automated detection of structural alerts (chemical fragments) in (eco)toxicology. *Computational and Structural Biotechnology Journal*. 5 (6): e201302013. doi: <http://dx.doi.org/10.5936/csbj.201302013>

References

1. REACH. Registration Evaluation Authorization and restriction of Chemicals 2007. http://ec.europa.eu/environment/chemicals/reach/reach_intro.htm.
2. Rogers, M. D., The European Commission's White Paper "Strategy for a Future Chemicals Policy": A Review. *Risk Anal.* 2003, 23 (2), 381-388.
3. 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 (4), 247-255.
4. 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 (3), 229-306.
5. Benigni, R.; Bossa, C., Mechanisms of Chemical Carcinogenicity and Mutagenicity: A Review with Implications for Predictive Toxicology. *Chem. Rev.* 2011, 111 (4), 2507-2536.
6. Benigni, R.; Bossa, C., Structure alerts for carcinogenicity, and the Salmonella assay system: A novel insight through the chemical relational databases technology. *Mutat. Res. Rev.* 2008, 659 (3), 248-261.
7. 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 (3-4), 207-220.
8. Sanderson, D. M.; Earnshaw, C. G., Computer prediction of possible toxic action from chemical structure; the DEREK system. *Hum. Exp. Toxicol.* 1991, 10 (4), 261-273.
9. 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 (1-3), 267-279.
10. Scheiber, J.; Jenkins, J. L.; Sukuru, S. C.; Bender, A.; Mikhailov, D.; Milik, M.; Azzaoui, K.; Whitebread, S.; Hamon, J.; Urban, L.; Glick, M.; Davies, J. W. Mapping adverse drug reactions in chemical space. *J. Med. Chem.* 2009, 52, 3103-3107.

11. Pauwels, E.; Stoven, V.; Yamanishi, Y. Predicting drug side-effect profiles: a chemical fragment-based approach. *BMC Bioinformatics* 2011, *12*, 169.
12. Klopman, G., Artificial intelligence approach to structure-activity studies. Computer automated structure evaluation of biological activity of organic molecules. *J. Am. Chem. Soc.* 1984, *106* (24), 7315-7321.
13. Klopman, G., MULTICASE 1. A Hierarchical Computer Automated Structure Evaluation Program. *Quant. Struct.-Act. Rel.* 1992, *11* (2), 176-184.
14. Klopman, G.; Chakravarti, S. K.; Zhu, H.; Ivanov, J. M.; Saiakhov, R. D., ESP: a method to predict toxicity and pharmacological properties of chemicals using multiple MCASE databases. *J. Chem. Inf. Comput. Sci.* 2004, *44* (2), 704-15.
15. Poroikov, V. V.; Filimonov, D. A.; Borodina, Y. V.; Lagunin, A. A.; Kos, A., Robustness of biological activity spectra predicting by computer program PASS for noncongeneric sets of chemical compounds. *J. Chem. Inf. Comput. Sci.* 2000, *40* (6), 1349-55.
16. Lagunin, A.; Stepanchikova, A.; Filimonov, D.; Poroikov, V., PASS: prediction of activity spectra for biologically active substances. *Bioinformatics* 2000, *16* (8), 747-8.
17. Filimonov, D.; Poroikov, V.; Borodina, Y.; Glorizova, T., Chemical Similarity Assessment through Multilevel Neighborhoods of Atoms: Definition and Comparison with the Other Descriptors. *J. Chem. Inf. Comput. Sci.* 1999, *39* (4), 666-670.
18. Cunningham, A. R.; Moss, S. T.; Iype, S. A.; Qian, G.; Qamar, S.; Cunningham, S. L., Structure-activity relationship analysis of rat mammary carcinogens. *Chem. Res. Toxicol.* 2008, *21* (10), 1970-82.
19. 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* (4), 1402-1411.
20. Agrawal, R.; Srikant, R. Fast Algorithms for Mining Association Rules in Large Databases. In *Proceedings of the 20th International Conference on Very Large Data Bases.*, Morgan Kaufmann Publishers Inc.: San Francisco, USA, 1994, pp. 487-499
21. Corneil, D. G.; Krueger, R. M. A Unified View of Graph Searching. *SIAM J. Discret. Math.* 2008, *22*, 1259-1276.
22. Inokuchi, A.; Washio, T.; Motoda, H. An Apriori-Based Algorithm for Mining Frequent Substructures from Graph Data. In *Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, Springer-Verlag: London, UK, 2000; pp 13-23.
23. Nijssen, S.; Kok, J. N. A quickstart in frequent structure mining can make a difference. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM: New York, USA, 2004; pp 647-652.
24. Yan, X.; Han, J. gSpan : Graph-Based Substructure Pattern Mining. In *ICDM*, 2003; pp 721-724.
25. Han, J.; Cheng, H.; Xin, D.; Yan, X., Frequent pattern mining: current status and future directions. *Data Min. Knowl. Disc.* 2007, *15*, 55-86.
26. Kazius, J.; Nijssen, S.; Kok, J.; Back, T.; Ijzerman, A. P., Substructure mining using elaborate chemical representation. *J. Chem. Inf. Model.* 2006, *46* (2), 597-605.
27. Borgelt, C.; Berthold, M. R., Mining molecular fragments: finding relevant substructures of molecules. *Proc. of the IEEE International Conference on Data Mining* 2002, 51-58.
28. Zaki, M.J.; arthasarathy, S.; Ogihara, M.; Li, W. New Algorithms for Fast Discovery of Association Rules. Technical Rapport, Rochester USA, 1997.29.
29. Karunaratne, T. Is Frequent Pattern Mining useful in building predictive models? *ECML/ PKDD workshop of Collective Learning and Inference on Structured Data* 2011.
30. Dong, G.; Li, J., Efficient Mining of Emerging Patterns: Discovering Trends and Differences. In *KDD*, 1999; pp 43-52.
31. Mannila, H.; Toivonen, H. Multiple uses of frequent sets and condensed representations. In *Proc. of the International Conference on Knowledge Discovery in Databases* 1996, 89-194.
32. Bayardo, J., Efficiently Mining Long Patterns from Databases. In *SIGMOD Conference*, 1998; pp 85-93.
33. Li, J.; Dong, G.; Ramamohanarao, K. Making use of the most expressive jumping emerging patterns for classification. *Knowl. Inf. Syst.* 2001, *3*, 1-29.
34. Auer, J.; Bajorath, J., Emerging chemical patterns: a new methodology for molecular classification and compound selection. *J. Chem. Inf. Model.* 2006, *46* (6), 2502-14.
35. Lozano, S.; Poezevara, G.; Halm-Lemeille, M. P.; Lescot-Fontaine, E.; Lepailleux, A.; Bissell-Siders, R.; Cremilleux, 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* (8), 1330-1339.
36. Poezevara, G.; Cuissart, B.; Crémilleux, B., Discovering Emerging Graph Patterns from Chemicals. In *Lecture Notes in Computer Science*, Springer-Berlin; Heidelberg, Eds. 2009; Vol. 5722, pp 45-55.
37. Worlein, M.; Meinel, T.; Fischer, I.; Philippsen, M., A quantitative comparison of the subgraph miners mofa, gspan, FFSM, and gaston. In *Proc. of the 9th European conference on Principles and Practice of Knowledge Discovery in Databases*, Springer-Verlag: Porto, Portugal, 2005.
38. Ullman, J., An algorithm for subgraph isomorphism. *J. Am. Chem. Soc.* 1976, *23*, 31-42.
39. Soulet, A.; Klema, J.; Cremilleux, B., Efficient mining under rich constraints derived from various datasets. In *Proceedings of the 5th international conference on Knowledge discovery in inductive databases*, Springer-Verlag: Berlin, Germany, 2007.
40. Cuissard, B.; Poezevara, G.; Cremilleux, B.; Lepailleux, A.; Bureau, R., *Emerging Patterns as Structural Alerts for Computational Toxicology*. Dong, G. & Bailey, J., Eds.; Taylor & Francis Group: 2012.
41. Calders, T.; Rigotti, C.; Boulicaut, J. F. A Survey on Condensed Representations for Frequent Sets. In *Computer Science, Constraint-Based Mining and Inductive Databases, volume 3848 of Lecture Notes*. Boulicaut, J. F., De Raedt, L. & Heikki Mannila, Eds.; Springer Berlin / Heidelberg: 2005; pp 64-80.
42. 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.
43. Dong, G.; Bailey, J., *Contrast Data Mining: Concepts, Algorithms and Applications*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series: 2012.

Competing Interests:

The authors have declared that no competing interests exist.



© 2013 Lepailleux et al.

Licensee: Computational and Structural Biotechnology Journal.

This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are properly cited.