JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

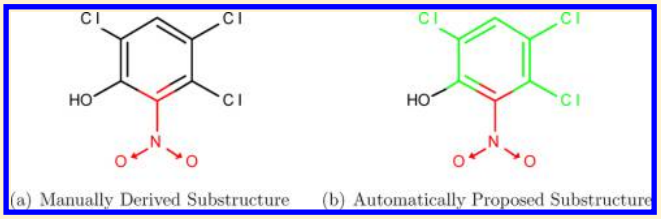
# Computational Derivation of Structural Alerts from Large Toxicology Data Sets

Ernst Ahlberg,*[,†] Lars Carlsson,[†] and Scott Boyer[†,‡]

[†]Drug Safety and Metabolism, AstraZeneca Research & Development, Pepparredsleden 1, 43183 Mölndal, Sweden

Supporting Information

**ABSTRACT:** Structural alerts have been one of the backbones of computational toxicology and have applications in many areas including cosmetic, environmental, and pharmaceutical toxicology. The development of structural alerts has always involved a manual analysis of existing data related to a relevant end point followed by the determination of substructures that appear to be related to a specific outcome. The substructures are then analyzed for their utility in posterior validation studies, which at times have stretched over years or even decades. With higher throughput methods now being employed in many areas of toxicology, data sets are growing at an unprecedented rate. This growth has made manual analysis of data sets impractical in many cases. This report outlines a fully automatic method that highlights significant substructures for toxicologically important data sets. The method identifies important substructures by computationally breaking chemical structures into fragments and analyzing those fragments for their contribution to the given activity by the calculation of a *p*-value and a substructure accuracy. The method is intended to aid the expert in locating and analyzing alerts by automatic retrieval of alerts or by enhancing existing alerts. The method has been applied to a data set of AMES mutagenicity results and compared to the substructures generated by manual curation of this same data set as well as another computationally based substructure identification method. The results show that this method can retrieve significant substructures quickly, that the substructures are comparable and in some cases superior to those derived from manual curation, that the substructures found covers all previously known substructures, and that they can be used to make reasonably accurate predictions of AMES activity.




## 1. INTRODUCTION

"Structural alerts" are a widely employed tool in computational toxicology and have been instrumental in identifying potentially hazardous chemical structures in a number of contexts including cosmetics, industrial chemicals, and drug discovery. Structural alerts have served as an aid to quickly identify chemicals that should be either prioritized for testing or for elimination from further consideration and use. Structural alerts exist now for a number of toxicologically relevant end points, and due to their ease of implementation and interpretability, their use is set to expand.

Perhaps the most simple method to extract substructural alerts from data is to utilize chemical and toxicological expert knowledge. One of the first examples of this was the derivation of "Ashby Tennant" alerts for bacterial mutagenicity assays.[1,2] The extraction of structural alerts using chemical and toxicological expert knowledge or any other manual technique is time-consuming and risks generating subjective interpretations of the data because it is heavily dependent on the skill and expertise of the chemist or toxicologist. Additionally, it is not always clear to the end-user what the significance of the alert is, i.e., with what certainty will a new substance containing the alerting substructure actually exhibit the given activity and whether there are other substructures in the molecule that will modify this risk. With the development of higher throughput methods for the assessment of toxicologically relevant end points, the need for computational aids to the derivation of more information-rich structural alerts is acutely needed and will aid in increasing the speed and fidelity of the test−learn−test cycle.

Thus, there is a clear need for a method that (1) identifies and displays contributing substructures to both positive and negative experimental outcome, (2) assesses and displays the statistical contribution of these substructures to that outcome, and (3) is flexible and adaptable to any new chemistry upon which data are available. These three qualities are available, to a limited extent in the combination of existing commercial tools[3−8] but not in a single method. This report therefore pulls together these requirements into a single method.

Flexible assessment of substructural contributions to experimental outcome is not new. The best methods available today grow substructure graphs from the atom-types or "nodes" by computing frequent "cliques", where a clique is a set of pairwise adjacent nodes in the graph.[9] To exemplify this, one can take an aliphatic chain in which each pair of *C*−*C* atoms are cliques or one can take an epoxide where all three atoms *C*1*OC*1 are connected to each other. The clique-based techniques start with the individual nodes in the graph and grow the substructures by combining nodes until no more

**Received:** May 28, 2014

   dx.doi.org/10.1021/ci500314a | *J. Chem. Inf. Model.* XXXX, XXX, XXX−XXX

substructures can be found that obey the user-specified occurrence threshold, i.e., the fraction of the total set of compounds that must share the same substructure. This is an exhaustive search of substructures in the data and is well suited for finding substructures, but it comes with a high computational effort. This effort is due to the requirement for all permutations to be analyzed, resulting in an algorithm with exponential complexity.

There are also methods that utilize maximum common substructure, MCS, computations, but those are primarily designed to identify privileged structures, i.e., the scaffold from which compounds are built. In such cases, MCS computations are applied after clustering of the compounds,[10] and the substructures therefore describe chemical classes of compounds in the data and not necessarily substructures which contribute to a given activity.

This report presents a new method to efficiently mine toxicological data for substructures that have the ability to explain a given activity. We explore the possibility that it is sufficient to grow substructures using the signature descriptor[11,12] generating circular substructures exhaustively in a systematic fashion. The method described here constantly monitors the contribution of the substructures to the given activity. The method is then compared on the following criteria: ability to capture information compared to manual curation, predictivity of the resulting substructures, and computational efficiency.

## 2. METHOD

**2.1. Data Set.** The method reported here is demonstrated using AMES mutagenicity data consisting of two activity classes, "positive" and "negative". The training data is described in Kazius et al.[13] and the external validation data of 880 compounds was reported by Young et al.[14] The compounds present in both data sets have been removed from the external validation data.

**2.2. Derivation of Substructures.** The substructures used in this work are the "atom signatures" presented by Faulon et al.[11,12] In this report, the term signature will be used for atom signatures. Signatures are chemical substructures, originating from each atom in a compound, represented as trees where the nodes are atoms and the edges are bonds. The root of the tree corresponds to an atom in the molecule, and the next layer of the tree represents the sorted neighbors of that atom. In this way, the tree is built layer by layer, generating a canonical representation of the substructure. The number of added layers is said to be the height of the signature, where height zero represents the root atom only. Signatures have been calculated using the Chemistry Development Toolkit, CDK.[15]

**2.3. Assessment of Substructure Significance.** For classification, the substructures are used as outcome indicators for compounds belonging to a certain class. An outcome for a specific substructure can be said to be the occurrence of that substructure in a number of compounds with a certain activity and the occurrence in a number of compounds without that activity. To determine whether such an outcome is likely to come from the same binomial distribution as the underlying data, the $p$-value is used. To calculate the $p$-value for this type of data set, the outcome has to be related to the occurrence of the activity in the total number of compounds, i.e., even in compounds where the substructure does not exist. For example, a full data set with $n$ compounds has $m$ compounds with a specific activity label where $m \leq n$. In the data set, a substructure is found in $n'$, $n' \leq n$, compounds, and the amount of those compounds with the specific activity label is $m'$, $m' \leq m$. Given that a set of compounds has a specific substructure, the *substructure frequency* describes how accurately the substructure separates data, i.e., the number of compounds that have both a specific activity and the specific substructure divided by the number of compounds that have the specific activity, $(m'/n')$. The corresponding $p$-value is $\sum_{i=m'}^{n'}((n'!)/(i!(n'-i)!)) \times (m/n)^i \times (1-(m/n))^{(n'-i)}$, where the $p$-value is the probability of obtaining the outcome or any other less probable outcome, given that it originates from the specific binomial distribution of the underlying data.

For a substructure to make a statistically significant contribution to activity, it is required that the $p$-value is below this level. Furthermore, the $p$-value alone is not sufficient for a substructure to accurately describe an activity. Depending on the skew of the data, the accuracy and $p$-values required for a significant call will intersect differently. For example, in a 50/50 distribution of positives and negatives, the minimum occurrence for which a substructure can be significant depending on whether its $p$-value is 5, see Figure 1A. For the majority class of the 70/30 example, the accuracy and $p$-value are intersected at 50 compounds in the data, see Figure 1B, whereas the minority class is intersected already at three compounds, Figure 1C. This means that, in addition to the level of significance, a lowest level of substructure frequency needs to be imposed on the substructures.

For data where one activity is overrepresented, unbalanced data, it is possible to obtain significant substructures with only two or three occurrences in the data. To avoid this, a lower bound on the number of compounds in which a substructure exists has to be used, denoted minimum occurrence. The minimum occurrence defaults to 5, which derives from the 50/50 distributed data and a $p$-value of 0.05, see Figure 1A.

**2.4. Data Set Processing.** The algorithm takes a data set with a discrete activity, thresholds for the $p$-value, minimum occurrence, and the substructure frequency. In the initial step, all height zero signatures are computed from the compounds in the data set. For each signature, the value of $n'$ is recorded together with the $m'$ values for each activity class in the data set. A signature is evaluated if the number of occurrences, $n'$, is above the given threshold. If so, the substructure frequency of the signature is computed for each activity class compared to the all of the other activities $(m'/n')$, in this case between "positive" and "negative". If the substructure frequency for an activity is above the substructure frequency threshold, the $p$-value is computed. If the $p$-value is below its threshold, the signature is labeled significant in discriminating the activity. If the signature is significant, the search for significant substructures is terminated in that direction. For all signatures that passed the occurrence threshold but has not been labeled significant, the search is extended to the next height by going back to the compounds and calculate the *height* + 1 signatures for the atoms that correspond to the analyzed signatures. The above procedure is repeated exhaustively, searching all compounds and atoms until no signature can fulfill the criteria and be significant. The possibility to adjust for false discoveries due to multiple hypothesis testing has been added as a postprocessing filter to the method using the Benjamini Hochberg procedure.[16]

A toy example is visualized in Figure 2, and it is only intended to show the process, not to be a complete calculation. Starting with computation of all h 0 signatures for all


dx.doi.org/10.1021/ci500314a | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

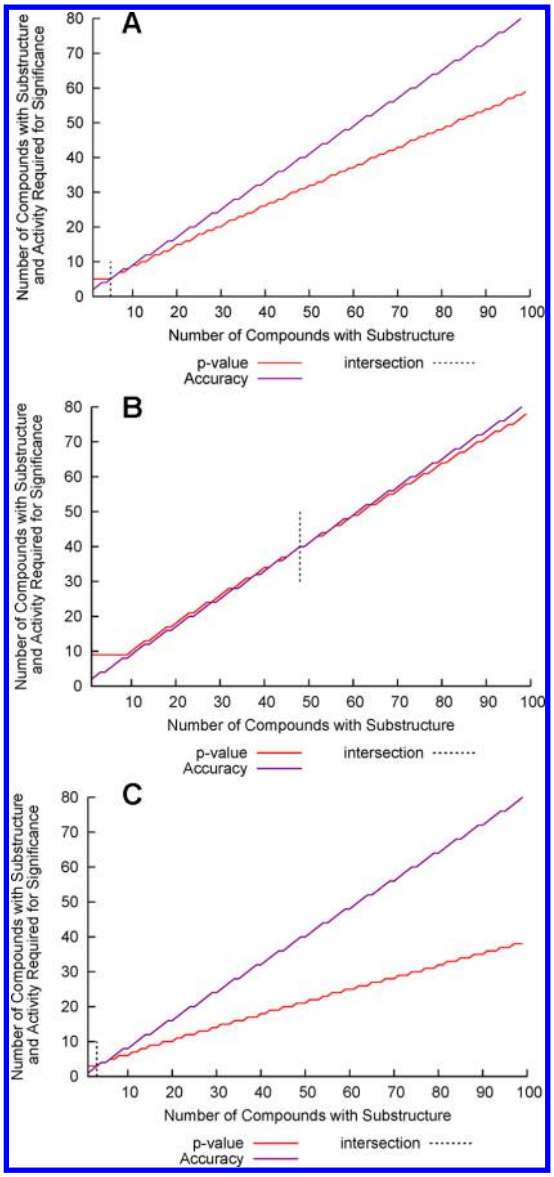



**Figure 1.** Lowest level of compounds with substructure needed for significance with respect to *p*-value and substructure frequency vs the number of compounds in the data set. The "threshold" level for the minimum number of compounds is given by the dashed vertical line, indicating the intersection of the *p*-value and the substructure frequency lines. (A) Data with 50% positive and negative compounds. (B) Majority class of data with 70/30 distribution. (C) Minority class of data with 70/30 distribution

compounds, the generated signatures are then tested for significance, none is significant at evaluation 0, so for all atoms in all compounds the h 1 signatures are calculated. At evaluation 1, one signature is significant and one has too few associated compounds. For the other signatures, the process is repeated, calculating the h 2 signatures for the atoms that correspond to the remaining h 1 signatures. This is sufficient because the signatures are hierarcical and the height h a signature contain the h (a-1), which means that only atoms that could be described with the h (a-1) can be described by the h a signature.

**2.5. Processing and Visualization of Results.** To easily visualize the substructures, Ogham[17] was used to create molecular figures to the output. For each significant signature, the signature together with the activity, positive count, total count, *p*-value, and substructure frequency is written to a spreadsheet table, together with the visualization of the fragment on one of the molecules, where positive fragments are red and negative fragments are green. An example of the output is given in Table 1. When the significant signatures are used for prediction, the overall result is reported together with a tagged structure data file,[18] SD-file, where the significant signatures have been marked. The SD-files for the data with matching substructures together with visualization of the matched substructures are provided in the Supporting Information.

The method uses OpenBabel[19] for molecular transformation, but the SMiles ARbitrary Target Specification, SMARTS,[20] matching for the clique-based method, was conducted using OEChem.[21] All computations have been performed on a desktop machine with four 3 GHz Intel Xeon cores and 4G RAM running CentOS release 5.5 AMD64. The implementation is serial, thus only using one core.

## 3. RESULTS

The algorithm described here has been compared to two other algorithms, the first one being manual annotation presented by Kazius et al.[13] and the second is the clique-based method PAFI.[22] The clique-based method retrieved frequent subgraphs from the data. The frequent subgraphs have been automatically converted to SMARTS, and based on the SMARTS, significant substructures have been retrieved based on *p*-value, substructure frequency, and occurrence. The data analysis has been performed according to the procedure described by Kazius et al.,[13] where compounds that contain no significant substructures have been classified as negative.

The work flow is demonstrated using the AMES mutagenicity data set described in Kazius et al.,[13] as a training set. An external validation set of 880 compounds reported by Young et al.[14] has also been used. The compounds present in both data sets have been removed from the external validation set.

**3.1. Comparison to Manually Derived Substructures.** To compare the accuracies of the substructures retrieved by Kazius et al.,[13] the significant signatures from the proposed method were applied to the training set, predicted on the training set and on the external validation set. Because the manually derived substructures only predict AMES positive compounds, whereas the proposed method predicts both positive and negative compounds, a third label, inconclusive, has been used to represent compounds not distinctly predicted positive or negative. Thus, the accuracies of the methods are reported in two ways, first using only the accuracy of the compounds distinctly hit by either positive or negative substructures and second where all compounds with a positive fragment have been classified as positive and the rest have been classified to be negative. When the criteria used by Kazius et al.[13] were applied to the algorithm presented here, 70% substructure frequency and a *p*-value of 0.1, their method produced slightly better overall predictive accuracies. However, when the criteria were more strict, 80% substructure frequency and a *p*-value of 0.05, the results were comparable. In a third test, the Benjamini Hochberg procedure for controlling the false discovery rate when doing multiple hypothesis testing was applied to the 80% test case. All results are presented in Tables 2 and 3. For the 70% case, 17600 signatures were analyzed and 985 of those found significant, and for the 80% case, 24300


dx.doi.org/10.1021/ci500314a | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

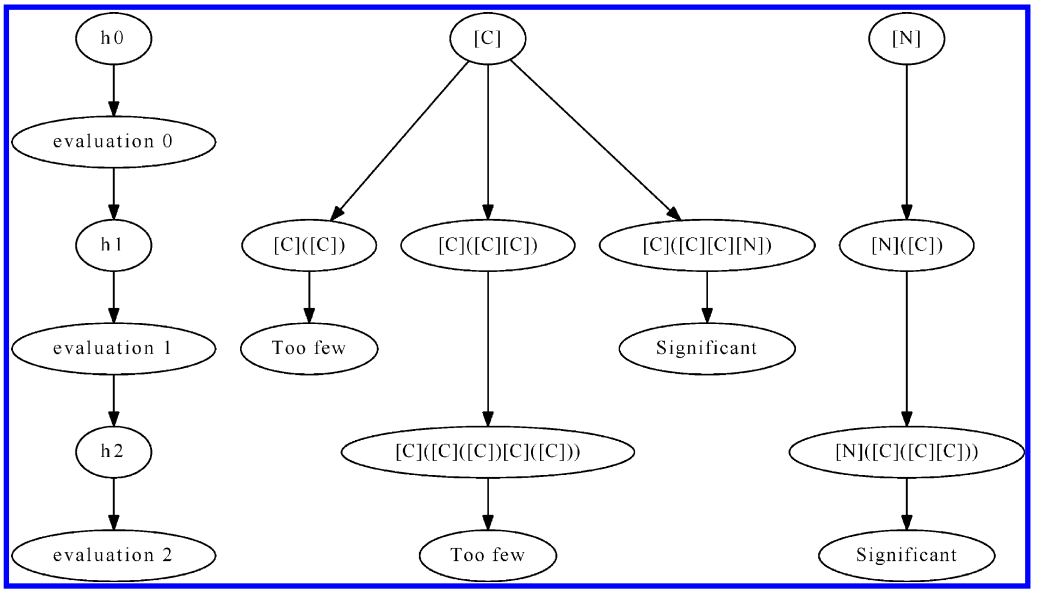
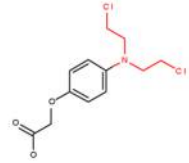
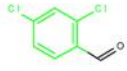
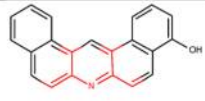



**Figure 2.** A toy example describing the proposed algorithm.

**Table 1. Example of the Output from the Substructure Significance Analysis**

| Experimental Activity | Number of Positive Compounds | Total Number of Compounds | *p*-value | Substructure Accuracy | Representation of Significant Substructure |
|---|---|---|---|---|---|
| Significant Substructure | | | | | |
| [Cl]([C]([C]([N]))) | | | | | |
| POS | 58 | 61 | 4.3e-12 | 95.1 |  |
| Significant Substructure | | | | | |
| [Cl]([C](p[C](p[C])p[C](p[C][Cl]))) | | | | | |
| NEG | 6 | 44 | 7.9e-9 | 86.4 |  |
| Significant Substructure | | | | | |
| [N](p[C](p[C]p[C])p[C](p[C]p[C])) | | | | | |
| POS | 144 | 150 | 3.9e-30 | 96.0 |  |

signatures were evaluated and 1137 found significant. When applying the Benjamini Hochberg procedure, the number of significant signatures was reduced to 133. A set of example compounds predicted with both the manually derived substructures and the proposed method is presented in Table 4 and shows differences and similarities between the methods. The two first examples in Table 4 show examples of where the two methods agree on the substructures and example, 3−5 show cases where the manually annotated aromatic amine SMARTS can be improved using the proposed methods. The


dx.doi.org/10.1021/ci500314a | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

**Table 2. Results on the Training Data from the Comparison between Manually Derived Substructures and the Substructures Generated Using Significant Signatures, Where INC is Inconclusive[a]**

| | proposed method substructure frequency 70% | | proposed method substructure frequency 80% | | proposed method substructure frequency 80% Benjamini Hochberg | | manually derived substructures | |
|---|---|---|---|---|---|---|---|---|
| training set | POS | NEG | POS | NEG | POS | NEG | POS | NEG |
| predicted POS | 1578 | 277 | 1830 | 276 | 1640 | 324 | 1944 | 418 |
| predicted NEG | 117 | 891 | 77 | 957 | 65 | 609 | | |
| predicted INC with hits | 586 | 590 | 191 | 227 | 78 | 78 | | |
| predicted INC without hits | 85 | 130 | 268 | 428 | 583 | 877 | 422 | 1470 |
| INC Excluded | | | | | | | | |
| overall accuracy | | 86.24 | | 88.76 | | 85.25 | | 82.30 |
| sensitivity, specificity | 93.10 | 76.28 | 95.96 | 77.62 | 96.19 | 65.27 | 100.0 | 0.00 |
| INC Included | | | | | | | | |
| overall accuracy | | 74.87 | | 80.06 | | 75.32 | | 80.25 |
| sensitivity, specificity | 91.46 | 54.08 | 85.92 | 73.36 | 72.61 | 78.71 | 82.16 | 77.86 |

[a]The INC label is further divided into two subcategories, with and without hits. The INC with hits is where the method has found both positive and negative substructures in the compound, and the INC without hits is where the method has found no significant substructure in the compound. Furthermore the % substructure frequency in the table is the required substructure frequency for each significant substructure.

**Table 3. Results on the External Validation Set from the Comparison between Manually Derived Substructures and the Substructures Generated Using Significant Signatures, Where INC is Inconclusive[a]**

| | proposed method substructure frequency 70% | | proposed method substructure frequency 80% | | proposed method substructure frequency 80% Benjamini Hochberg | | manually derived substructures | |
|---|---|---|---|---|---|---|---|---|
| external validation set | POS | NEG | POS | NEG | POS | NEG | POS | NEG |
| predicted POS | 307 | 64 | 352 | 51 | 362 | 52 | 438 | 75 |
| predicted NEG | 49 | 146 | 40 | 137 | 31 | 96 | | |
| predicted INC with hits | 175 | 78 | 101 | 34 | 60 | 15 | | |
| predicted INC without hits | 19 | 41 | 57 | 107 | 97 | 166 | 112 | 254 |
| INC Excluded | | | | | | | | |
| overall accuracy | | 79.67 | | 84.31 | | 84.66 | | 85.38 |
| sensitivity, specificity | 86.38 | 67.86 | 89.80 | 72.87 | 92.11 | 64.86 | 100.0 | 0.00 |
| INC Included | | | | | | | | |
| overall accuracy | | 74.97 | | 79.29 | | 77.82 | | 78.73 |
| sensitivity, specificity | 89.45 | 50.76 | 82.36 | 74.16 | 76.73 | 79.64 | 79.64 | 77.20 |

[a]The INC label is further divided into two subcategories, with and without hits. The INC with hits is where the method has found both positive and negative substructures in the compound, and the INC without hits is where the method has found no significant substructure in the compound. Furthermore the % substructure frequency in the table is the required substructure frequency for each significant substructure.

last example shows a compound which the proposed method has labeled inconclusive with hits, showing that there are both significantly positive and negative fragments present. The full set of compounds is presented in the Supporting Information. The training time for the proposed method was roughly 15 s in both the 70% and 80% cases on the computer specified previously.

**3.2. Comparison to PAFI.** To perform a more thorough validation of the method, the training set has been divided into 10 subsets using stratified sampling and evaluated using cross validation, where the significant substructures have been retrieved from the remainder of the training set and tested on the subset.

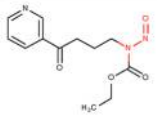
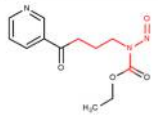
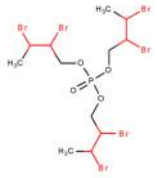
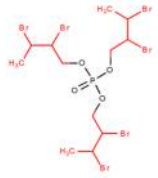
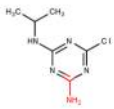
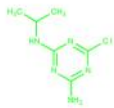
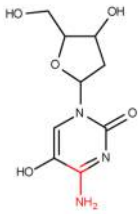
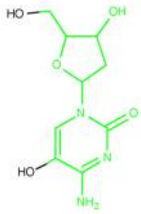
The proposed method and PAFI have been compared with varying number of minimum occurrences (5, 10, 20, 50, and 100) and a $p$-value of 0.05. The averaged results from the cross-validation study are presented in Table 5. In addition, Table 6 presents results on the external validation set which are all on a par with the results generated by the manually derived substructures. For PAFI, the training time and the prediction time is very similar. In Tables 5 and 6, for 5 minimum occurrences, results for PAFI are missing due to insufficient memory. The tables also show that the possibility to mine chemical data using a low occurrence threshold results in increased accuracy. The computational effort for training and predicting in Tables 5 and 6, respectively, shows that the computational effort of the method proposed here is considerably lower than the computational effort for any of the clique-based method. The relatively high computational effort of the clique-based method is also reflected in the finding that the average number of generated significant substructures is very high, sometimes expanding up to 1.35 million. The corresponding value for the method presented here is 1015.

## 4. DISCUSSION

The aim of this work was to develop a method for automated retrieval of significant substructural alerts.The proposed algorithm can be applied to any data set consisting of molecules and activity. A user-friendly method is presented which gives a


dx.doi.org/10.1021/ci500314a | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

**Table 4. Example Compounds and Corresponding Hits for the Manually Derived Substructures and the Proposed Method**

| ID | Experimental Activity | Manually Derived Substructures | Proposed Method |
|---|---|---|---|
| 1 | POS |  |  |
| 2 | POS |  |  |
| 3 | NEG |  |  |
| 4 | NEG |  |  |
| 5 | NEG |  |  |
| 6 | NEG |  |  |

good picture of the surrounding chemistry and can automatically be applied for prediction of novel compounds.

The results from the comparison between the manually derived substructures and the method presented show that the proposed method is able to identify substructures corresponding to 32 of the 34 manually derived substructures and that the overall accuracies are similar. The two missed substructures are only present in three compounds in the training data and can therefore not be identified by the proposed method. Because the proposed method identifies both active and inactive fragments, it is now possible to study substructures that deactivates mutagenicity, for example, compound 3 in Table 4. The refinement of the method to accurately take deactivating fragments into account when assessing data is ongoing work. The effort needed to do this work manually is very high and time-consuming, whereas the presented method completed the task in under a minute. As the methods are different, there are also differences in the results, as can be seen in Table 4. There the major difference between the methods is that the proposed

**Table 5. Averaged Results from the Cross Validation Study of the AMES Data, Where Minimum Occurrences, Method, Average Number of Substructures, Test Accuracy, and Training Time is Presented[a]**

| minimum occurrences | method | mean number of substructures | accuracy test set | training time (min) |
|---|---|---|---|---|
| 5 | proposed method | 1015 | 77.46 | 0.20 |
| 5 | PAFI | – | – | – |
| 10 | proposed method | 519 | 77.20 | 0.17 |
| 10 | PAFI | 1355873 | 74.42 | 1275.63 |
| 20 | proposed method | 203 | 76.09 | 0.17 |
| 20 | PAFI | 216538 | 74.38 | 416.04 |
| 50 | proposed method | 52 | 74.14 | 0.16 |
| 50 | PAFI | 48866 | 73.91 | 126.99 |
| 100 | proposed method | 19 | 70.66 | 0.16 |
| 100 | PAFI | 9545 | 73.15 | 25.68 |

[a]Lack of data due to insufficient memory, and thus a failed analysis, is represented by –.

**Table 6. Results on the External Validation Set for the Computational Methods[a]**

| minimum occurrences | method | accuracy | prediction time (min) |
|---|---|---|---|
| 5 | proposed method | 78.58 | 0.05 |
| 5 | PAFI | – | – |
| 10 | proposed method | 78.07 | 0.03 |
| 10 | PAFI | 76.34 | 1753.96 |
| 20 | proposed method | 78.53 | 0.03 |
| 20 | PAFI | 77.00 | 518.88 |
| 50 | proposed method | 76.67 | 0.02 |
| 50 | PAFI | 76.90 | 126.81 |
| 100 | proposed method | 75.16 | 0.02 |
| 100 | PAFI | 75.28 | 25.25 |

[a]Displaying minimum occurrences, method, overall predictive accuracy, and prediction time. Missing data in the table is the result of the complexity of the algorithms resulting in insufficient memory to complete the analysis.

method, as applied here, will not detect similarities between fragments and join them. For example, the manually derived substructure aliphatic halide will hit any Cl, Br, or I attached to an aliphatic carbon to cover the same functionality with the proposed method, with the current atom labeling, three substructures are needed, one for each halide. This will in some cases result in a low number of occurrences of the substructure in the data.

The prediction examples 1 and 2 in Table 4 shows that similar fragments are picked up by both methods. Examples 3, 4, and 5 all show cases where the manually derived substructure *Specific Aromatic Amine* is not specific enough to distinguish between positive and negative compounds. In this way, the proposed method can be used to aid the understanding of the underlying relationship by displaying substructures that


dx.doi.org/10.1021/ci500314a | J. Chem. Inf. Model. XXXX, XXX, XXX–XXX

deactivate the compound. The sixth and last example in Table 4 shows a compound for which both positive and negative substructures are found. Such compounds are difficult to classify, and to improve the techniques further, these are the compounds that need attention together with the set of inconclusive compounds without hits.

Comparing the proposed method with and without false discovery rate detection shows that the number of significant substructures can be drastically reduced and still maintain some predictive ability on the data. In this case, however, the overall sensitivity has suffered and 34% of the compounds did not get a prediction. In early drug development, it is very important to flag potential risks and the prediction will always be put in perspective of the underlying data and analyzed together with experts in the field. In this case, the false discovery rate detection was applied as a filter after completing the proposed procedure. One possibility to improve would be to make use of the hierarchy built into the signatures and either analyze each height as a family of tests or use the parent signature, *height* − 1, to group the tests. This needs to be investigated further.

When comparing the proposed method with a clique-based technique, the results show that the proposed method can mine the data quickly at a low occurrence threshold, Table 5, and that the number of significant substructures generated is much lower for this method compared to the clique-based method, making manual inspection of substructures feasible. As we have demonstrated, the predictive strength of the proposed method is on par with the manually derived substructures and the clique-based method but with significantly lower computational effort. One reason for this is that the substructures are generated systematically with tracking of the originating atom, and that once a structure is significant, the search is terminated for that atom. Another reason for the predictive strength of the proposed method is that chemically important features are usually dependent on the local surroundings which are well represented with the signatures.

Regarding the computational effort, the clique-based method has an exponential complexity with respect to the number of atoms, whereas the corresponding complexity of the signature algorithm is claimed to be polynomial,[23] which means that it is possible to analyze large data sets quicker with the proposed method. The number of generated substructures is also quite high for the clique-based method which results in slow predictions, because when predictions are made, all matches of substructures on compounds needs to be found.

The comparison with other methods has been performed solely on open source methods. There are, however, commercial software packages like Computer Automated Structure Evaluation, CASE,[24] or mCASE,[5] which are somewhat similar to the proposed method. CASE retrieves substructures based on the KLN chemical structure code,[25] which is essentially long paths of atoms that may contain only one major branch. After the substructure retrieval, CASE performs a binomial test on the retrieved substructures with respect to the biological activity to identify significant substructures. The mCASE software performs the CASE analysis multiple times in a hierarchical fashion, separating the data in each step. This has the advantage that it is not necessary to treat the whole data set throughout the analysis but rather separated subsets. There is, however, a risk associated with this type of analysis because each subset must then contain a significant amount of a feature for that feature to be included in the model. In this way, features that may be applicable to the whole data set will only affect the model for some subclasses. To assess the differences between the methods, a detailed comparison will be needed, but that is not in the scope of this article.

The comparison to the manually derived substructures described by Kazius et al.[13] shows that the proposed method can be used to replace or complement manual searches for significant substructures. The proposed method can also be used as a preprocessing step for the molecular generation algorithms based on signatures.[26]

## 5. CONCLUSIONS

This paper illustrates a complete automated work flow for substructure pattern generation. The method treats the data objectively and generates a set of significant substructures according to the user defined constraints. The results are comparable with results obtained by hand annotation but with the significant difference that this method is automated and objective to the findings in the data.

The main advantages with this method are that the part of finding the significant substructures is deterministic, objective, automated, and the results are easily visualized. The proposed method is also well suited for use in an automated learning system, analyzing and re-evaluating the significant substructures on the fly as new data become available.

Finally, the proposed method clearly identifies "deactivating" fragments in molecules that contain traditional "alerting" substructures, giving the user a clear view of the significance of an alert in the local chemical environment.

## ■ ASSOCIATED CONTENT

### Supporting Information

Complete results from the simulated data runs together with the visualizations of the comparison between the Kazius toxicophores and the proposed method. In addition, the substructures extracted by the proposed method have been added. This material is available free of charge via the Internet at http://pubs.acs.org.

## ■ AUTHOR INFORMATION

### Corresponding Author

*Phone: +46317064681. E-mail: ernst.ahlberg@astrazeneca.com.

### Present Address

‡(S.B.) Computational Toxicology, Swedish Toxicology Sciences Research Center, Forskargatan 20, 151 36 Södertälje, Sweden.

### Notes

The authors declare no competing financial interest.

## ■ REFERENCES

(1) Ashby, J.; Tennant, R. W. Chemical structure, Salmonella mutagenicity and extent of carcinogenicity as indicators of genotoxic carcinogenesis among 222 chemicals tested in rodents by the U.S. NCI/NTP. *Mutat. Res. Rev. Genet.* **1988**, *204*, 17−115.

(2) Ashby, J.; Tennant, R. W. Definitive relationships among chemical structure, carcinogenicity and mutagenicity for 301 chemicals tested by the U.S. NTP. *Mutat. Res. Rev. Genet.* **1991**, *257*, 229−306.

(3) Johnson, D. E.; Blower, P. E.; Myatt, G. J.; Wolfgang, G. H. I. Chem-tox informatics: data mining using a medicinal chemistry building block approach. *Curr. Opin. Drug Discovery Dev.* **2001**, *18*, 76−79.

(4) *Leadscope—Chemoinformatics Platform for Drug Discovery*; http://www.leadscope.com/index.php (accessed Feb 7, 2014).
(5) Klopman, G. MULTICASE. 1. A Hierarchical Computer Automated Structure Evaluation Program. *Quantum Struct.−Act. Relat.* **1992**, *11*, 176−184.
(6) *MultiCASE Inc*; http://www.multicase.com/products/prod01.htm (accessed Oct 9, 2010.
(7) Sanderson, D.; Earnshaw, C. Computer Prediction of Possible Toxic Action from Chemical Structure; The DEREK System. *Hum. Exp. Toxicol.* **1991**, *10*, 261−273.
(8) *DEREK*; Lhasa LTD, 2009; http://www.lhasalimited.org/products/derek-nexus.htm.
(9) Deshpande, M.; Kuramochi, M.; Wale, N.; Karypis, G. Frequent Substructure-Based Approaches for Classifying Chemical Compounds. *IEEE Trans. Knowl. Data Eng.* **2005**, *17*, 1036−1050.
(10) Nicolaou, C. A.; Tamura, S. Y.; Kelly, B. P.; Bassett, S. I.; Nutt, R. F. Analysis of Large Screening Data Sets via Adaptively Grown Phylogenetic-Like Trees. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1069−1079.
(11) Faulon, J.-L.; Visco, D. P. J.; Pophale, R. S. Signature Molecular Descriptor. 1. Using Extended Valence Sequences in QSAR and QSPR Studies. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 707−720.
(12) Faulon, J.-L.; Churchwell, C. J. Signature Molecular Descriptor. 2. Enumerating Molecules from Their Extended Valence Sequences. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 721−734.
(13) Kazius, J.; McGuire, R.; Bursi, R. Derivation and Validation of Toxicophores for Mutagenicity Prediction. *J. Med. Chem.* **2005**, *48*, 312−320.
(14) Young, S.; Gombar, V.; Emptage, M.; Cariello, N.; Lambert, C. Mixture Deconvolution and Analysis of Ames Mutagenicity Data. *Chemom. Intell. Lab. Syst.* **2002**, *60*, 5−11.
(15) Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK) An Open-Source Java Library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493−500,12653513.
(16) Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc., Ser. B: Stat. Methodol.* **1995**, *57*, 289−300.
(17) *Ogham 2D Chemical Structure Layout and Rendering*; http://www.eyesopen.com/docs/ogham/1.7.0/html/index.html (accessed on Jan 3, 2011).
(18) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244−255.
(19) *The Open Babel Package*; http://openbabel.org/wiki/Main_Page (accessed on Mar 2, 2014).
(20) *SMARTS*; http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (accessed Aug 7, 2010).
(21) *Openeye Scientific Software*; http://www.eyesopen.com (accessed Aug 7, 2010).
(22) Kuramochi, M.; Karypis, G. An Efficient Algorithm for Discovering Frequent Subgraphs. *IEEE Trans. Knowl. Data Eng.* **2004**, *16*, 1038−1051.
(23) Faulon, J.-L. Isomorphism, Automorphism Partitioning, and Canonical Labeling Can Be Solved in Polynomial-Time for Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 432−444.
(24) 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*, 7315−7321.
(25) Klopman, G.; McGonigal, M. Computer simulation of physical−chemical properties of organic molecules. 1. Molecular system identification. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 48−52.
(26) Ahlberg Helgee, E.; Carlsson, L.; Boyer, S. A Method for Automated Molecular Optimization Applied to Ames Mutagenicity Data. *J. Chem. Inf. Model.* **2009**, *49*, 2559−2563,19877655.