VOCCluster: Untargeted Metabolomics Feature Clustering Approach for Clinical Breath Gas Chromatography - Mass Spectrometry Data

Yaser Alkalifah1, Iain Phillips1, Andrea Soltoggio1, Karen Darnley3, William H. Nailon3, Duncan McLaren3, Michael Eddleston2, C. L. Paul Thomas4 and Dahlia Salman4,*

1Department of Computer science, Loughborough University, Loughborough, LE11 3TU, UK
2Pharmacology, Toxicology & Therapeutics Unit, University of Edinburgh, Edinburgh, UK
3Edinburgh Cancer Centre, NHS Lothian, Edinburgh, UK
4Department of Chemistry, Loughborough University, Loughborough, LE11 3TU, UK

*To whom correspondence should be addressed.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Metabolic profiling of breath analysis involves processing, alignment, scaling and clustering of thousands of features extracted from Gas Chromatography Mass spectrometry (GC-MS) data from hundreds of participants. The multi-step data processing is complicated, operator error-prone and time-consuming. Automated algorithmic clustering methods that are able to cluster features in a fast and reliable way are necessary. These accelerate metabolic profiling and discovery platforms for next generation medical diagnostic tools.

Results: Our unsupervised clustering technique, VOCCluster, prototyped in Python, handles features of deconvolved GC-MS breath data. VOCCluster was created from a heuristic ontology based on the observation of experts undertaking data processing with a suite of software packages. VOCCluster identifies and clusters groups of volatile organic compounds (VOCs) from deconvolved GC-MS breath with similar mass spectra and retention index profiles. VOCCluster was used to cluster more than 15,000 features extracted from 74 GC-MS clinical breath samples obtained from participants with cancer before and after a radiation therapy. VOCCluster was able to cluster those features into 1081 groups (including endogenous, exogenous compounds and instrumental artifacts) with an accuracy rate of 96% (± 0.04 at 95% confidence interval). Results were evaluated against a panel of ground truth compounds, and compared to other clustering methods used in previous metabolomics studies such as DBSCAN and OPTICS.

Availability: The source code and the data used in this paper are available for download at https://github.com/Yaser218/Untargeted-Metabolomics-Clustering.

Contact: D.Salman@lboro.ac.uk

1 Introduction

Breathomics, the analysis of volatile organic compounds (VOCs) in breath, offers a promising approach for the non-invasive study of metabolic processes and derangements (Lourenço and Turner, 2014). Much has been made of its potential for the development of new and enhanced diagnostic approaches (Amann and Smith, 2013). Non-targeted metabolomic studies with breathomics use Gas Chromatography-Mass Spectrometry (GC-MS) as the gold standard analytical technique (Watson and Sparkman,
The combination of high-resolution separations, low limits of detection (picogram level) and mass spectral fragmentation patterns provides efficient Class 2 compound identification (Rathbun-Parr et al., 2016). Amann and Smith (2013) provide an excellent introduction to the theory and practice of GC-MS in breathomics. Despite the high fidelity of GC-MS breathomics data, it is not yet possible to adopt and follow the guidelines and recommendations for metabolomic characterisation proposed and adopted widely in metabolic studies involving blood-plasma or urine (Brown et al., 2005). The reasons for this “arrested-development” arise from the nature of breath samples and the inherent variability of GC-MS data.

Breath samples are not stable and consequently cannot be stored for significant lengths of time (Kang and Thomas, 2016). This means that pooled samples and batch processing are not currently possible. Further, breath data are acquired throughout the study and contain the artefacts that arise from instrument degradation and maintenance cycles. A putative workflow has described how these attributes of breath samples and subsequent may be managed (Guadarrama-Hoyas et al., 2013). This previous work also described how deconvolution of the mass spectra obtained from co-eluting VOCs and their subsequent registration through retention indexing could be used to assign unique identifiers to unknown VOCs and thereby facilitate multi-variate analysis. However this work was incomplete as it did not address adequately the variability of the mass spectrometric component in GC-MS breath data; and, it did not solve the impracticality of scaling the workflow to encompass the many tens of thousands of breath features that are generated from a modestly sized study (cohort size: n=20 to 50).

Deconvolution generates a retention-indexed (RI) mass spectra and peak areas for each isolated feature (supplementary data, Figure 1). An ideal analysis would result in any given VOC always generating identical mass spectra and consequently being assigned the same identifier for its deconvolved feature in every sample it is isolated. However small variations in the intensities of the mass-to-charge ratios (m/z) of the fragment ions that constitute a mass spectrum results in the same VOC feature being potentially assigned a different identity. Consequently, post-processing of breath data currently requires specialist and expert evaluation of every extracted feature to verify that each VOC is correctly classified with a single identifier across the breath data matrix as a prelude to multivariate modelling. This methodology is not sustainable when studying untargeted metabolomics involving thousands of VOCs isolated from the breath samples of the study cohort. Further, the endeavour turns rapidly into an exercise in human endurance with the further introduction of variability and misclassification errors arising from operator fatigue or knowledge gaps.

The current study sought to capture the heuristic ontology used by breathomics researchers in clustering deconvolved breathomics features, so that features that arise from identical chemical species are assigned to the a unique identifier and clustered together correctly, regardless of variability in signal attributes. Their knowledge was used to iteratively develop a computational clustering method that produces the essential feature clustering step for GC-MS breathomics data that resolves the scalability roadblock and enhances the reliability of the clustering encoding. This is the vital step in the preparation of the “discovery” data modelling.

This research was fostered by previous classification approaches that proposed supervised pattern recognition methods for the classification of breathomics GC-MS data. Van Berkel et al. applied Support Vector Machine (SVM) algorithm to breath data to distinguish between smoking and non-smoking subjects (n = 22); by creating predictive models with significant generalisation power despite working with small data set or data with large variation. However, SVM is only feasible when classification labels are available, and this is not necessarily the case when working with blind clinical trials where labels are revealed once the potential biological candidates are discovered. Unsupervised clustering techniques were also used to identify trends and correlations in metabolomics GC-MS data without prior assumptions. They can be used to identify groups of correlated metabolites from different samples prior to clustering them together as one group. Table 1 illustrates examples of current unsupervised methods that were used for features clustering in untargeted metabolomic profiling studies.

<table>
<thead>
<tr>
<th>Method</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSCout (Tikunov et al., 2012)</td>
<td>A software tool for analysis of untargeted profiling using the subtractive fuzzy clustering method.</td>
</tr>
<tr>
<td>MetaAnalyzer (Xia et al., 2013)</td>
<td>It is a web server designed to perform comprehensive metabolomic data analysis, visualisation and interpretation. It supports such techniques as: PCA, PLS-DA, hierarchical clustering and a number of machine learning methods.</td>
</tr>
<tr>
<td>MSerry (Nicola et al., 2012)</td>
<td>It performs unsupervised and untargeted data mining on GC-MS data.</td>
</tr>
<tr>
<td>Carotta (Hauschild et al., 2015)</td>
<td>It is a cluster analysis framework using unsupervised statistical learning, it is using hierarchical clustering to identify groups of VOCs and groups of patients with a similar VOCs.</td>
</tr>
<tr>
<td>NCMS (Smith et al., 2006)</td>
<td>It involves nonlinear retention time alignment, matched filtration, peak detection, and peak matching. Hundreds of endogenous metabolites can be identified and used as standards to calculate a nonlinear retention time correction profile for each sample.</td>
</tr>
<tr>
<td>ChlMSID (Depke et al., 2017)</td>
<td>It is an algorithm that uses cosine similarities to build a distance matrix for all spectra, density based (specifically OPTICS) and hierarchical clustering to group structurally similar compounds.</td>
</tr>
</tbody>
</table>

Overall, the available algorithms are either designed for targeted analysis, where only a panel of compounds is searched for and clustered together, or untargeted analysis which is user dependent and the results (i.e. number and size of clusters) are influenced by given parameters. Untargetted analysis produces additional difficulties as the parameters are often estimated and decided by the user and could be unconsciously influenced by the needed results (Winie et al., 2015). Density-based spatial clustering of applications with noise (DBSCAN), ordering points to identify the clustering structure (OPTICS) and hierarchical clustering methods are examples of the most commonly used clustering methods in untargeted metabolomics data analysis. All of these algorithms do not require k, an estimate of the number of the clusters, as input and are therefore useful in untargeted studies where the number and size of clusters are unknown.

DBSCAN was reported by Ester et al. (1996) as a density-based clustering process in a time determined similarity matrix. It is used for clustering tandem mass spectra data for both metabolomics and proteomics fields (Rieder et al., 2017) Kriegel et al. (2011). However, some report challenges with DBSCAN when the data are highly dimensional with a varying density profiles between clusters. OPTICS clustering was used by Depke et al. (2017) to develop ChlMSID algorithm which was applied to cluster metabolites mass spectra from a P. aeruginosa cell extract. The ChlMSID workflow was reported to provide correctly grouped metabolites with common functional elements such as peptides. However, this algorithm was not developed or used for highly dimensional variable breath data. Hierarchical clustering (HC) was based on the consolidation of peak lists by calculating the Euclidian distances of metabolites to produce groups of same multivariate similarity (Fiehn et al., 2000; De Souza et al., 2006). De Souza et al. (2006) used HC for GC-MS mass spectra data.
collected for the metabolic profiling of Leishmania parasites. However, it requires the scientist to make an arbitrary decision of how and where to cut the presented dendrogram, which becomes harder with hundreds of thousands of VOCs detected per cohort.

Despite both supervised classification and unsupervised clustering techniques described above, the algorithms are specific to the nature of data and research question. These approaches may not necessary enable mechanistic molecular identity or a specific metabolic pathway recognition. Therefore, there is a need for an algorithm that is automated, fast, not user dependent and can cluster thousands of similar VOCs from hundreds of samples (despite RI and m/z variation between samples) to use for multi-variate modelling.

The proof of concept of an algorithm VOCCluster, a novel GC-MS feature-clustering algorithm coded in Python, is presented. VOCCluster can identify clusters of same VOCs from different deconvolved breath samples in a non-supervised manner. VOCCluster employs cosine distance between VOCs’ mass spectra within a RI region from different samples to establish a distance measurement Cosine distance has been used as a robust similarity measure when compared with other measuring distances (Stein and Scott, 1994; Wan et al., 2002; Liu et al., 2007). As opposed to other clustering techniques, VOCCluster continually monitors the clustering of features and will reassess the cluster membership as the algorithm progresses. VOCs can therefore be removed and recluttered, which involves additional computation processes, but the candidate clusters results are potentially better. In this paper, VOCCluster was tested to process clinical breath samples obtained from participants with cancer before and after a radiation therapy dose. This research formed part of the TOXI-triage clinical trial (Touk-triage, 2018). To compare the performance of VOCCluster, DBSCAN and OPTICS were tuned and applied to the same data and the results of all three algorithms were compared and evaluated.

The rest of the paper is organised as follows: Section 2 introduces the steps taken by VOCCluster to cluster VOCs into groups, Section 3 showcases VOCCluster outcome results and evaluates its accuracy by comparing a panel of VOCs clusters with the “ground truth” that was generated by the expert as well as results obtained from DBSCAN and OPTICS clustering tools, and Section 4 concludes the results of this paper.

2 Materials and methods

VOCCluster measures mass spectra similarities for VOCs from different samples. In order to do this, a distance matrix is formed. This is supplemented by an examination of the RI for each VOC and together these are used to cluster VOCs into groups, each group containing a single VOC from each sample. Figure 1 illustrates and simplifies VOCCluster workflow stages within the algorithm.

2.1 Retention Index variance calculation

As there is a natural and expected variation in conditions between samples, there is a need to calibrate the expected range for RI for each VOC in a study. VOCCluster calculates a RI variation between breathomatic samples using an input list of targeted compounds (targetedVOCs) covering the busy range of the chromatogram. The list can be formatted by selecting one compound from any breath sample (for each targeted compound) in the given dataset. The targetedVOCs (e.g. table 1 in supplementary data) should include compounds that exist in a majority of samples including an estimated RI range for each one. For example, spiked internal standard compounds and common endogenous and exogenous compounds.

Calculating the RI variations for the targeted compounds helps to build a reliable distance matrix and consequently more accurate clustering outputs. It is important to note that the targetedVOCs were only used for this purpose without being used for the actual clustering process. The

VOCCluster also calculates the distances among all VOCs in each group at this stage to find the minimum available similarity between two VOCs at the same group. This is needed to determine the minimum threshold epsilon (ε) similarity of two VOCs to be assigned as same VOCs, which will be detailed later in section 2.3. In other words, ε is used as an input parameter for the clustering process.

VOCCluster generates a report at this step which includes clusters of all the targeted compounds that need to be evaluated by a human operator to examine if the samples were aligned correctly. The distance matrix is then built and the clustering process is executed. Otherwise, any samples with RI alignment errors that are observed by the operator have to be reprocessed and aligned again for the RI variations to be recalculated.

Fig. 1. A workflow summarising the different clustering processes within the VOCCluster algorithm. VOCCluster starts by calculating the RI variations for a given input and consequently the cosine similarity distance will be calculated between all the features in the given dataset. Clustering each feature into its related group will take place in the last three stages.
2.2 Distance matrix of mass spectra

Using this information VOCCluster is able to narrow the choice of potentially similar VOCs in other samples as it will only search for VOCs within the calculated RI range. The next step is to calculate the cosine distance (Equation 2, (Stein and Scott, 1994)) between a particular VOC in a sample with other VOCs in different samples within the RI range. VOCCluster determines the relative intensities for every m/z prior to calculating the cosine correlation, each m/z intensity for each VOC was normalised to the most intense m/z intensity for that VOC:

\[
D(A, B) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}
\]

(2)

where \(A_i\) and \(B_i\) are the normalized m/z intensities of mass A and B respectively. The distances between assigned VOCs are stored in a matrix. A similarity value of zero was given when compared VOCs were outside the RI range, otherwise the cosine distance was calculated. The size of the cosine matrix was \(m \times n\), where \(m\) is the number of VOCs in the collection of samples. Table 2 and figure 2 in supplementary data show an example of a distance matrix.

2.3 VOCCluster clustering methodology

To introduce the notion of VOCCluster clustering technique, we first make the following observation: consider a set of breath samples (\(N\)) that contain a set of VOCs, which will be called points at this stage, and they are listed in a dataset for clustering purposes. In breathomics studies, \(N\) is known and therefore, the number of points \(p\) and the established-distance \(C_{\text{max}}\) must not exceed \(N\). The similarity between two mass spectra can be examined by \(\epsilon\) which was calculated previously.

Breathomics datasets may contain clusters of varying density and in some cases, there are no density drops between close or even overlapped clusters. For example, (Heptane, 2,4-dimethyl-) (A) and (Octane) (B) (figure 2) may appear in some breath samples within the calculated RI thresholds and have similar mass spectrum profiles except for 2-3 ions of higher m/z values. Consequently, the clusters of these two different groups are difficult to be differentiated as their RIs and mass spectra overlap.

Density-based clustering algorithms (e.g. DBSCAN and OPTICS) fail in these circumstances. This fact is illustrated in figure 3, where A and B are density-based clusters with respect to \(\epsilon\). Part of cluster A is classified with cluster B as a result of the high similarity between the points of the two clusters which should be in separate clusters. Example of such cases can also be found in the supplementary data (figure 3).

Consequently, in VOCCluster several distance parameters are calculated and processed at the same time. Furthermore, a clustered point can be re-clustered into another cluster, if that point is highly similar to a point in the new cluster while still taking \(\epsilon\) and other parameters into consideration.

2.4 VOCCluster algorithm detail

As the algorithm starts the clustering process, a candidate point is selected and then the points with the highest similarity values are collected and processed first. This will ensure that points grouped in a cluster all have the highest probability of similarity to that point and each other. VOCCluster assigns cluster membership to each point in the dataset. Outlier points will not be clustered. For the purpose of clustering, VOCCluster classifies points in the dataset as established, core, border and outlier points. The definition of each point is illustrated in figure 4. VOCCluster stores information about each point before it was allocated to a cluster. This information will be used to assess the distance similarity of a point before it is re-clustered into another cluster. This information consists of three values for each clustered point: core-distance, core-point and established-distance, introduced in the following definitions:

- **core-distance** of a point for a point \(p\) is the similarity distance value between \(p\) and the highest similarity point at the same cluster. For example, the core-distance value for point E in figure 4 is 0.93, which is the distance between E and the closest point in the cluster which is C as C was the highest similar point to E. core-distance for non-clustered points will be zero.
- **core-point** for a point \(p\) is the point with the highest similarity to \(p\). For example, point C is the core-point for point E and D in figure 4. Unclustered points do not have a core point.
- **established-distance** for a point \(p\) is the similarity distance between \(p\) and the established point of a cluster. For example, the established-distance for point C, E, D and B is A, where A is the first point that was processed in the cluster. Unclustered points do not have an established-distance.
VOCCluster selects a unclustered point, $p$, from the dataset to start a new cluster. $p$ will be assigned as an established point and $p$ will be its own core point. All of the neighbours of $p$ with respect to $\varepsilon$ are examined and stored in potentials list even if already clustered. The neighbour points in the potentials list are ordered based on the highest similarity value to $p$. The most similar point in the potentials list, $q$, will be examined to find its neighbours. The potentials list will be updated with the neighbours of $q$. Points that are already in this list and have higher similarity to $q$ than $p$ will be updated. The process keeps repeating and the cluster grows until no points in the potentials list can be added or $C_{\text{max}}$ is fulfilled. After that, a new unclustered point will be selected again to start a new cluster and the process repeated.

However, when a clustered point, $w$, is added into the potentials list, VOCCluster will test the two distance values of the overlapped $w$. If $w$ has a core-distance value to a point in the new cluster greater than core-distance value to a point in the previous cluster and the establish-distance of $w$ in the new cluster is greater than the establish-distance of $w$ in the previous cluster, then, it will be moved to the new cluster. All of the points that share the same core-point, $w$, in the previous cluster will be reclustered with $w$. Otherwise, $w$ will be left as it was. Applying this sophisticated movement of points between clusters improves the accuracy of clustering because a point will be allocated to the most probable similarity point in the dataset. All code used in these experiments is available at https://github.com/Yaser218/Untargeted-Metabolomics-Clustering.

The novelty of VOCCluster comes from the employment of the core-distance and established-distance mechanisms. These enable a VOC to be reassessed and cluster membership to be corrected if another VOC has higher similarity to the those in the cluster. This overcomes problems with existing techniques where once clustered a VOC will not be reclustered. This makes VOCCluster less sensitive to the order of the VOCs in the dataset. Figure 3 demonstrated misclustering as part of the H cluster was assigned to the G cluster. This makes VOCCluster less sensitive to the order of the VOCs in the data.

2.5.2 Sensitivity and specificity of clusters

The performance was evaluated by determining the accuracy of VOCs clusters according to the following:

- True positive (TP): point is clustered in the correct group.
- False Positive (FP): point is clustered in an incorrect group.
- False Negative (FN): point should have been clustered but was not.
- True Negative (TN): a sample doesn’t contain a point and this point wasn’t clustered. This is limited to feature relating to points that exist in other samples.

A list of 27 ground truth compounds, covering a range of chemical functional groups and masses, was prepared and used for comparison with the automated clusters generated by VOCCluster. The ground truth compounds were acetone, methane-d, toluene, octamethylcyclotetrasiloxane, octamethylcyclohexasiloxane, benzophenone, cyclotrisiloxane, benzaldehyde, benzene, heptane, decanal, ethylbenzene, hexanal, furfural, benzofuran, acetic acid, styrene, thiophene, 3-methyl-1-hexanol, 2-ethyl-heptane, 2,4-dimethyl-, 2,4-dimethyl-1-heptene.

The accuracy of each cluster was calculated as:

\[
\frac{(TP + TN)}{n}
\]

Furthermore, the mean of the distance similarity between VOCs in a cluster was calculated for each of the ground truth compounds and an overall similarity between the clustered VOCs was determined.

3 Results and evaluation

3.1 RI alignment accuracy and variation calculation

Accuracy of the manual retention indexing alignment step was checked using the panel of targeted VOC’s illustrated in figure 5. The RI-variation report was generated and entails the absolute clusters of those targeted VOCs. Correction of the RI alignment is important for the clustering phases, the smaller the RI-variation the faster and more accurate the clustering functions because choices of compared VOCs are cut back.

The calculated $\Delta RI$ (figure 5) varied over the RI range. For examples $\Delta RI$ for acetone and cyclohexanol, dodecamethyl were 24 and 49 RI units, respectively. Therefore, the use of one or an estimated $\Delta RI$’s value for the entire analysis is not an accurate assumption and it is important to calculate the variation of the RIs per compound to help improve the accuracy of the VOCCluster clustering functions. The mean, standard deviation and coefficient variation (%) was calculated for

1 from SpectralWorks, https://www.spectralworks.com/
Rs per each compound within the targeted VOCs panel, any sample with incorrect alignment was highlighted and corrected manually by re-deconvoluting/aligning the data using AnalyzerPro software. Once alignment for that sample is corrected, the $\Delta RIs$ is recalculated by V OCCluster.

3.2 V OCCluster clustering

The dataset of the $n = 74$ contained 15,307 deconvoluted features. The generated distance matrix contains many zeros, this is a result of either no obvious statistical correlation between the VOCs or their Rs were not compatible.

The clustering function produces a list of values whose length is equal to the dataset length. Each value in the list represents a cluster number for the same index VOC in the dataset. For example, the same cluster value was given when two points were assigned to be the same VOC as they had the exact same parameters and features. The average similarity between points in all clusters was 96.4%.

3.3 Performance evaluation

V OCCluster clustered all 15,307 deconvoluted features in around 3 hours. A similar number of samples is estimated to take a minimum of 12 weeks to cluster manually. The accuracy of the clustering was evaluated based on the accuracy of merging the extracted VOCs into their clusters.

The accuracy of the algorithm varies between compounds depending on the chemical nature of the compound, intensity, extracted m/z profile for that compound etc. For example, cyclotetrasiloxane, octamethyl-, toluene and siloxanes were the most accurate clusters as they had similar mass spectra profiles. The accuracy of detecting various compounds was calculated per compound and is shown next to each box plot.

The clustering function produces a list of values whose length is equal to the dataset length. Each value in the list represents a cluster number for the same index VOC in the dataset. For example, cyclotetrasiloxane, octamethyl-, (f) Cyclotetrasiloxane, dodecamethyl-. Calculated $\Delta RIs$ were calculated per compound and is shown next to each box plot. Hence a different mass spectrum profile to the decanal cluster of higher concentration. The average extracted ions for the low concentration compounds was 8 ions while it was 18 ions for the high concentration compounds.

Decanal is another challenging compound to detect as it is often misclassified when at low concentrations. Here some m/z values are missing and hence a different mass spectrum profile to the decanal cluster of higher concentration. The average extracted ions for the low concentration compounds was 8 ions while it was 18 ions for the high concentration. However, this needs further clarification by chemists as calibration curves and lower limit of detection for decanal on the GC-MS analytical method are needed to verify the above proposition.

V OCCluster was more accurate that DBSCAN and OPTICS using the same dataset. DBSCAN and OPTICS were tested with different $\epsilon$ and minPts values for a number of ground truth compounds. Both of the algorithms have been built in Python (available at https://github.com/Yaser218/Untargeted-Metabolomics-Clustering) and use the same distance matrix that is used with V OCCluster. Table 3.3 illustrates the number and accuracy of the clusters for each algorithm. It is clear that these algorithms are sensitive to the given parameters. For example, DBSCAN generated 681 clusters and the accuracy was 78.3% with an $\epsilon = 90$ for the ground truth compounds. Toluene-D8 at this stage well separated and the accuracy was 100%. At the same position, acetone was clustered with an accuracy of 61.6% where about 21 FPs samples have been noticed. Once the $\epsilon$ value was increased to be about 98%, the accuracy of Acetone was improved to reach about 84% and only 8 samples were FPs. In contrast to Acetone, the accuracy of detecting Toluene-D8 decreased to be 64.9% and 26 samples were FNs. More details about DBSCAN and OPTICS results can be found in section 5 in the supplementary data.

It’s also important to note that V OCCluster, with minimum tuning, can be used to cluster features and peaks for other GC-MS metabolomics data.

### Table 2. A summary of V OCCluster performance accuracy based on the “ground truth” VOC panel.

<table>
<thead>
<tr>
<th>VOC</th>
<th>Manual TP</th>
<th>Manual TN</th>
<th>V OCCluster TP</th>
<th>V OCCluster TN</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acetone</td>
<td>63</td>
<td>9</td>
<td>65</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>Methane-d, trichloro</td>
<td>74</td>
<td>0</td>
<td>74</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Toluene</td>
<td>74</td>
<td>0</td>
<td>74</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Toluene-D8</td>
<td>74</td>
<td>0</td>
<td>71</td>
<td>0</td>
<td>3*</td>
</tr>
<tr>
<td>Cyclotetrasiloxane, octamethyl-</td>
<td>74</td>
<td>0</td>
<td>70</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Cyclopentasiloxane, dodecamethyl-</td>
<td>74</td>
<td>0</td>
<td>73</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Cyclotetrasiloxane, hexamethyl-</td>
<td>74</td>
<td>0</td>
<td>74</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cyclotetrasiloxane, dodecamethyl-</td>
<td>74</td>
<td>0</td>
<td>73</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Benzodiphenol</td>
<td>57</td>
<td>17</td>
<td>53</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>(1,1',3,1'-Terphenyl)-2'-ol</td>
<td>53</td>
<td>21</td>
<td>52</td>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td>Benzaldehyde</td>
<td>74</td>
<td>0</td>
<td>74</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Benzen</td>
<td>70</td>
<td>4</td>
<td>70</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Noranal</td>
<td>68</td>
<td>6</td>
<td>67</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Heptanal</td>
<td>49</td>
<td>25</td>
<td>48</td>
<td>4</td>
<td>21</td>
</tr>
<tr>
<td>Decanal</td>
<td>59</td>
<td>15</td>
<td>57</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Ethylbenzene</td>
<td>66</td>
<td>8</td>
<td>58</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>n-Prune</td>
<td>65</td>
<td>9</td>
<td>63</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Hexanal</td>
<td>47</td>
<td>27</td>
<td>46</td>
<td>1</td>
<td>26</td>
</tr>
<tr>
<td>Furfural</td>
<td>38</td>
<td>36</td>
<td>35</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Benzenzfuran</td>
<td>64</td>
<td>10</td>
<td>64</td>
<td>3*</td>
<td>0</td>
</tr>
<tr>
<td>Acetic Acid</td>
<td>46</td>
<td>28</td>
<td>44</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Styrene</td>
<td>74</td>
<td>0</td>
<td>74</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Thiofene, 3-methyl-</td>
<td>49</td>
<td>25</td>
<td>44</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>1-Hexene, 2-ethyl-</td>
<td>59</td>
<td>15</td>
<td>52</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>Heptane, 2,4-dimethyl-</td>
<td>65</td>
<td>9</td>
<td>58</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>Octane</td>
<td>31</td>
<td>43</td>
<td>22</td>
<td>1</td>
<td>43</td>
</tr>
<tr>
<td>2,4-Dimethyl-1-heptene</td>
<td>63</td>
<td>11</td>
<td>62</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

| Mean VOC Cluster Accuracy | 96% |

*Another compound co-eluted with Toluene-D8 for those 3 samples and changed the profile of the extracted mass spectra (addition of ions such as m/z 159).
Table 3. DBSCAN and OPTICS results using different $\epsilon$ threshold when applied to the radiation clinical GC-MS data set. Accuracy, percentage of clustered points from all samples and number of clusters ($n$) were influenced with both algorithms by the given $\epsilon$ value.

<table>
<thead>
<tr>
<th>$\epsilon$</th>
<th>DBSCAN clustered accuracy (%)</th>
<th>OPTICS clustered accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>80</td>
<td>612</td>
<td>89.8</td>
</tr>
<tr>
<td>90</td>
<td>681</td>
<td>71.0</td>
</tr>
<tr>
<td>98</td>
<td>788</td>
<td>82.6</td>
</tr>
</tbody>
</table>

matrices such as saliva, skin and urine. Overall, VOCCluster provides more accurate results compared with the available approaches in the literature. VOCCluster demonstrates a new computational approach to clustering VOCs from breath data being able to cluster the 15,000 features in our 74 clinical GC-MS data in less than 3 hours on a commodity computer.

4 Conclusion

The VOCCluster algorithm provides untargeted metabolomics feature classification for breath GC-MS data. VOCCluster was used for a clinical breath data set ($n = 74$) obtained from cancer patients before and after radiation therapy as part of the TOXI-triage clinical trial. Mass spectra similarities, RI range, cosine similarity, and a new clustering principle were optimised and applied to a clinical data set. The accuracy of VOCCluster clustering outcome was evaluated and compared to a manual VOC panel “ground truth”, DBSCAN and OPTICS. The approach resulted in an accurate clustering of VOCs (96%, ± 0.04 at 95% confidence interval). This was superior to existing computational techniques and faster than manual processing.

5 Acknowledgements

The authors of this manuscript would like to acknowledge the TOXI-triage project and the clinical research nurse and radiotherapy staff for their help with obtaining the clinical data sets. TOXI-triage received funding from the European Union’s Horizon 2020 Innovation action Programme H2020-EU.3.7 - Secure societies - Protecting freedom and security of Europe and its citizens under grant agreement No 653409. The authors would also like to acknowledge SpectralWorks limited for providing us with free AnalyserPro software that was used to deconvolve the clinical breath data used in this study.

References


Tikunov, Y. et al. (2012). MSChis: A tool for unsupervised mass spectra extraction of chromatography-mass spectrometry ion-wise aligned data. Metabolomics, 8(4), 714–718.


