

We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

5,500

Open access books available

136,000

International authors and editors

170M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com



Cheminformatics Applied to Analytical Pyrolysis of Lignocellulosic Materials

Jorge Reyes-Rivera

Abstract

Pyrolysis-Gas Chromatography/Mass Spectrometry has been used to characterize a wide variety of polymers. The main objective is to infer the attributes of materials in relation to their chemical composition. Applications of this technique include the development of new improved materials in the industry. Furthermore, due to the growing interest in biorefinery, it has been used to study plant biomass (lignocellulose) as a renewable energy source. This chapter describes a procedure for characterization and classification of polymeric materials using analytical pyrolysis and cheminformatics. Application of omics tools for spectral deconvolution/alignment and compound identification/annotation on the Py-GC/MS chromatograms is also described. Statistical noise is generated by production of numerous small uninformative compounds during pyrolysis. Such noise is reduced by cheminformatics here detailed and this facilitate the interpretation of results. Furthermore, some inferences made by comparison of the identified compounds to those annotated with a biological role in specialized databases are exemplified. This cheminformatic procedure has allowed to characterize in detail, and classify congruently, different lignocellulosic samples, even using different Py-GC/MS equipment. This method can also be applied to characterize other polymers, as well as to make inferences about their structure, function, resistance and health risk based on their chemical composition.

Keywords: Biomass pyrolysis, polymeric materials characterization, cheminformatics, multivariate comparative analysis, Py-GC/MS

1. Introduction

The largest repository of lignocellulosic biomass is generated by the cell walls of plants [1]. Its main chemical components are cellulose, hemicelluloses and lignin. The proportions are variable but close to 4:3:3, respectively, and the element content is 50% C, 6% H, 44% O y $\leq 0.4\%$ N, for resources such as wood [1]. Because biomass is a renewable resource, its study for the production of energy and value-added aromatic compounds has gained importance in recent decades [2, 3]. It has been considered that lignocellulosic biomass as a renewable energy source would satisfy around 25% of energy requirements [4]. Thus, CO₂ sequestered by plants during photosynthesis would balance the CO₂ generated by biofuels and their use would not contribute to global warming [5, 6]. On the other hand, after cellulose,

lignin is the most abundant polymer in nature and the main natural source of aromatic compounds [1, 7]. For this reason, lignin is important in the chemical industry and it has been projected as a replacement for aromatic polymers derived from fossil fuels [8].

Lignocellulosic biomass, like other non-volatile complex materials, cannot be directly analyzed in its original state by gas chromatography. Therefore, one of the most common methods for its analysis is the Pyrolysis-Gas Chromatography/Mass Spectrometry (Py-GC/MS). This method consists of the rapid heating of the materials under analysis (close 300°C), to break the covalent bonds and produce individual fragments. The compounds derived from pyrolysis pass through a capillary column of fused silica in a Gas Chromatograph using an inert gas as carrier (e.g., He). Then the fragments are separated based on their retention times. The selective fragmentation pattern caused by Electron Impact and the m/z ratio for each pyrolysis product are registered by a detector on a Mass Spectrometer. Finally, each compound is identified by comparing its mass spectrum to those in the reference electronic libraries (NIST, MONA, etc.) or to the mass spectra produced by analytical standards [9–12]. The sequential combination of these three processes in Py-GC/MS makes it a versatile and powerful tool for the analysis of lignocellulosic materials and other complex mixtures, such as polymers and copolymers [3, 13].

Analytical pyrolysis is currently implemented as a standard method for determining the ratio of H/G/S subunits in plant biomass, agricultural and industrial waste, soil samples and organic matter [6]. This technique has also been useful to elucidate the series of reactions and products derived from the pyrolysis of carbohydrates [14, 15] and lignins [16, 17]. It has been applied for monitoring changes during the delignification and bleaching process as well as for the characterization of different lignocellulosic materials [12]. In addition, it has been used to determine the S/G ratio in lignin of drought-resistant succulent species with results highly comparable to other characterization techniques [18]. On the other hand, its high sensitivity has enabled the detection of hundreds of chemical compounds, including less abundant monomers in lignin, such as acetylated subunits (i.e., sinapyl and coniferyl acetates [19]) and 5-hydroxyguaiacyl units [20]. Recently, Py-GC/MS applied to the analysis of cacti spines, with the use of cheminformatics, allowed a detailed characterization of lignocellulosic matrix, as well as the classification of the samples from a chemotaxonomic approach [21].

1.1 Advantages of Py-GC/MS

Different advantages confer great versatility of application to Py-GC/MS. Firstly, its efficiency, precision and relatively low operating costs [6] make it a suitable routine technique. In addition, it is a fast technique that requires a very small sample size [22, 23]. Volatilization of samples by pyrolysis minimizes the need for pre-isolation, even when analyzing macromolecules in complex mixtures [24]. Therefore, it can be used to analyze a wide variety of materials: e.g., fibers and textiles, wood, bark and paper, artistic materials, synthetic polymers and heteropolymers [12, 13]. Likewise, comparable and reproducible results can be obtained when the conditions of the analysis are kept constant: i.e., carrier gas, heating rate, maximum temperature, homogeneous particle size and removal of non-structural compounds [18, 21]. Therefore, samples with the same composition will produce the same derivatives of pyrolysis [13, 21]. On the other hand, the advantages of the coupled GC/MS system are associated with a high speed, specificity and sensitivity, in both the separation of the pyrolysis products and in their identification [9, 12]. In addition, Py-GC/MS allows the identification of compounds without the necessity of standards.

It enables the comparison to commercial or open access libraries, including some already curated for different classes of chemical compounds [21, 25–28]. Finally, the raw data generated can be exported for quantitative or qualitative analysis [29, 30].

1.2 Issues related to Py-GC/MS

Although the many advantages and applications of Py-GC/MS are evident, different authors consider some problematic aspects. The main ones are: 1) pyrolysis produces a large amount of compounds, therefore, is necessary to deal with the vast amount of information registered by the Mass Spectrometer. 2) Only one part of the compounds produced can be unambiguously identified. 3) Low availability of mass spectra in databases and reference libraries. 4) Altogether, this makes the interpretation of the results from analytical pyrolysis difficult. However, most of these problems can be solved if cheminformatics is applied to the data resulting from Py-GC/MS.

The following sections will describe the use of omics tools for the deconvolution of mass spectra, as well as the alignment and annotation of the compounds identified in the chromatograms (**Figure 1**). This process is useful to compare different samples obtained by Py-GC/MS, under the same operating conditions, even using different equipment. In addition, different multivariate methods will be described to minimize the statistical noise generated by numerous uninformative compounds (i.e., those derived from carbohydrates). Together, the use of omics tools and multivariate methods facilitate the interpretation of the results of analytical pyrolysis. The processes detailed here may also be applicable to Py-GC/MS analysis of materials other than lignocellulosics (i.e., polymers, copolymers, soil samples and organic matter). In addition, they can be applied to raw data generated by other chromatography systems coupled to mass spectrometry (i.e., GC/MS/MS, LC/MS, and LC/MS/MS), including different equipment and output formats.

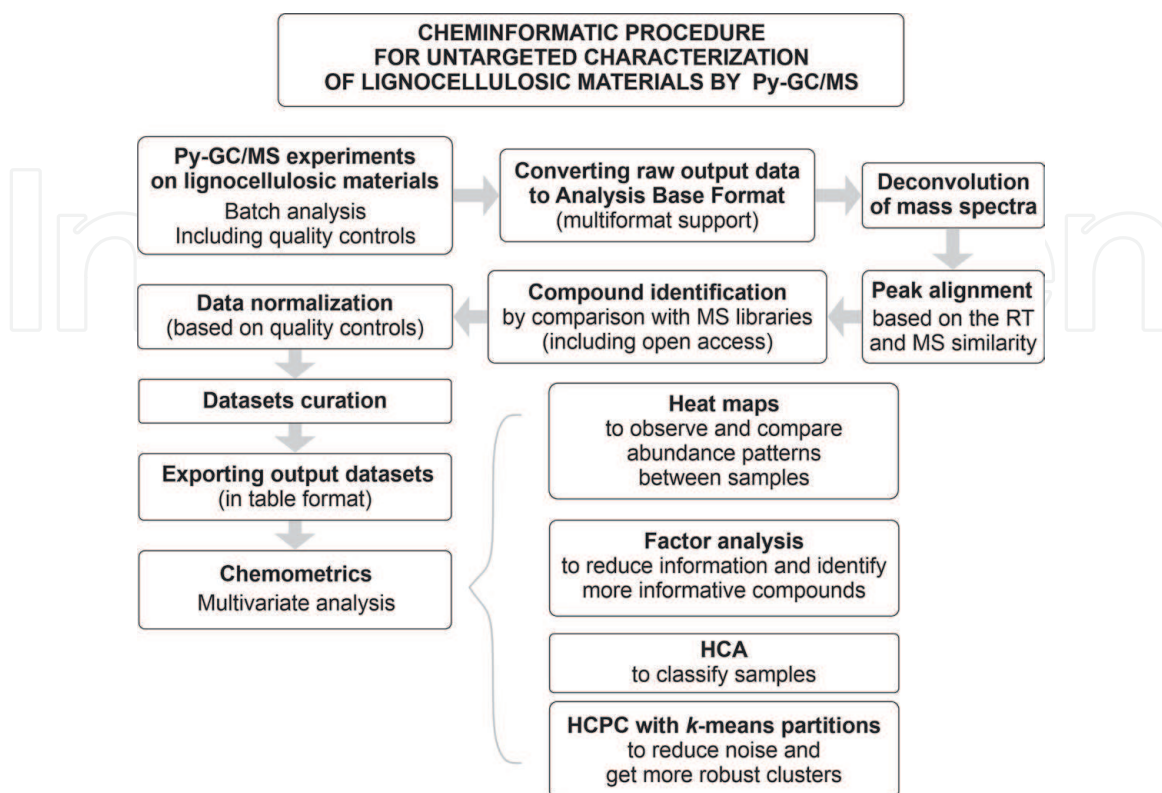


Figure 1. Untargeted cheminformatics workflow for analysis of lignocellulosic materials by Py-GC/MS.

1.3 Common problems in Py-GC/MS and contribution of cheminformatics for their solution

Some apparent methodological problems attributed to pyrolysis are associated with the conditions necessary for the analysis of specific materials. Lourenço *et al.* [12], point out that care must be taken with the pyrolysis temperature when analyzing materials rich in suberin, such as barks. The main problem is that suberin decomposes at temperatures in the range of 550–600°C [31]. Therefore, this is an aspect to take into account if it is required to know the composition of this polymer within lignocellulosic samples [12]. Another problem referred in various works is that Py-GC/MS cannot guarantee an entirely quantitative determination. However, some authors have successfully carried out quantitative analyses in the optimization of aromatic hydrocarbon production from biomass [29]. Also for the quantification of only small amounts of aromatic hydrocarbons, applying the external calibration method [3, 30].

The amount of information that is generated as a result of the entire process can be challenging aspect. One analysis of 45 minutes by Py-GC/MS on lignocellulosic samples can generate up to 2,729 mass spectra [21]. However, after cheminformatics and manual curation of the datasets, the authors were able to unambiguously recognize 451 compounds, including some putative isomers. Another common problem is the displacement of the peaks in the chromatograms for samples with different chemical composition. For example, the displacement of the peak corresponding to levoglucosan in Py-GC/MS chromatograms for syringil-rich wood [18]. The displacement is due to the absence of acetovanillone in the samples. Therefore, the peak of levoglucosan appears at a Retention Time (RT) of 22.72 min, while in species that produce acetovanillone it is observed at 23.55 min (**Figure 2**). The above effect is problematic when it is required to directly process a batch of several samples with differential compositions. There are two reasons: 1) the process would be very time consuming if several species are analyzed and all the peaks identified by Py-GC/MS are compared one by one (about 40 compounds per sample, using

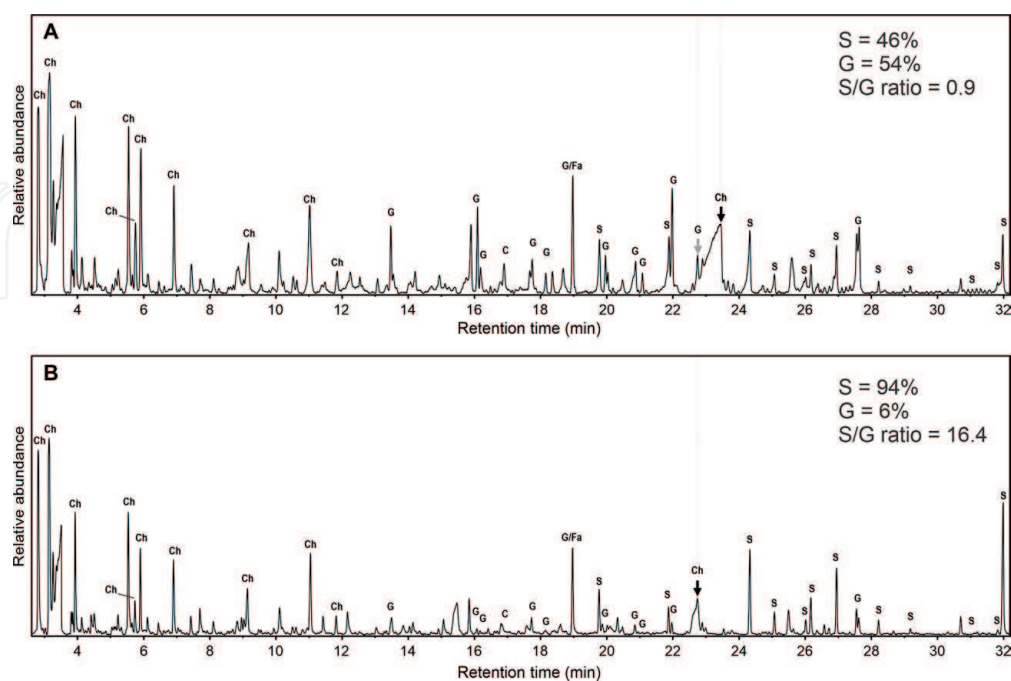


Figure 2.

Displacement of the peaks. Py-GC/MS chromatograms from extractives-free wood in cacti: A) Pilosocereus chrysacanthus and B) Ferocactus hamatacanthus. Displacement of levoglucosan (black arrows) is due to the absence of acetovanillone (gray arrows) in samples with 94% of syringil units [18]. The origin of the compounds is marked with letters: Ch, carbohydrates; G, guaiacyl subunits; S, syringil subunits, Fa, ferulates.

the native GC/MS software). This implies that the analysis has to be limited only to differences in the relative abundance, or the presence/absence, of only certain compounds. 2) If the raw datasets from the chromatograms are compared directly, using any multivariate method, the peak displacement would cause methodological bias because equivalent compounds are not being compared. Cheminformatics analysis solves this problem by automating the alignment of mass spectra and the identification of compounds for a batch of samples.

On the other hand, the high degree of degradation caused by the high temperatures used in pyrolysis represents, by far, the main problem of this technique. Therefore, this technique is considered to be of little use to characterize molecules larger than monomers or dimers in biopolymers such as lignin [6]. In addition, it is considered that the large number of derivatives makes the description of the chemical composition of sample difficult. Therefore, the detailed interpretation of the results is difficult and probably not necessary [3]. For example, when analyzing carbohydrate samples, low molecular weight derivatives can originate from hexoses or pentoses [12, 32]. The reason is that cellulose and hemicelluloses involve similar thermal degradation pathways, therefore a large part of the derivatives produced are the same [33, 34]. The reason is that cellulose pyrolysis causes the heterolytic cleavage of the glycosidic C—O bonds. In addition, it involves complex reactions and different pathways to give rise to anhydro sugars and numerous compounds with low molecular weight: i.e., acetic acid, 1-hydroxybutan-2-one, hydroxyacetaldehyde, 1-hydroxypropan-2-one and 2-furaldehyde [15, 35, 36]. A large part of these small compounds can also be originated from the decomposition of hemicelluloses. For example, 2-furaldehyde and acetic acid can be produced from the degradation of xylan [12, 37, 38]. On the other hand, there are contrary cases, but they also

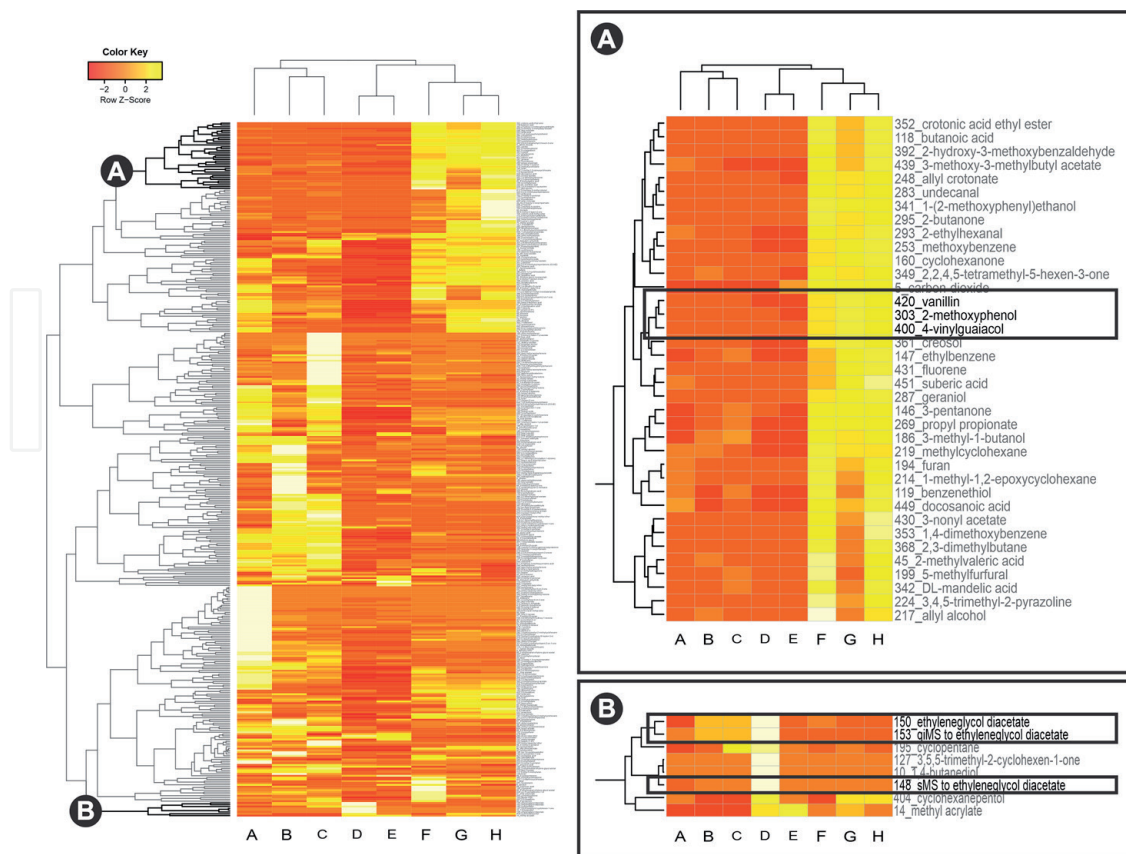


Figure 3. Complete profile of the compounds identified for eight samples of lignocellulosic materials. A) Cluster corresponding to Guaiacyl lignin derivatives. B) Abundance patterns for carbohydrates derivatives. Similar (sMS) or quasi identical (qiMS) mass spectra.

contribute to the ambiguity in the identification of the compounds and their origin. Particularly when different ions are produced by the same class of compounds. The case of pyrans and furans is an example of compounds with ambiguous origin; both, with different molecular ions, can derive from the degradation of cellulose or hemicelluloses [12]. In this sense, the use of cheminformatics makes it possible to identify the abundance patterns of the compounds in a batch of samples. Based on this, it can be inferred if there are coincidences in the behavior of the pyrolysis products (**Figure 3**). In this way, it is possible to infer whether different compounds have the same origin, or rule out differences due to the operating conditions of the method or the characteristics of the samples [21].

For example, 2,5-dimethylfuran and 4-methyl-2*H*-pyran correspond to different molecular ions, but have the same average mass (96.13 Da) with similar RT, 4.64 min and 4.74 min, respectively (*see* Supplementary Materials of [21]). Based on the observed abundance patterns, it can be deduced that they are related to

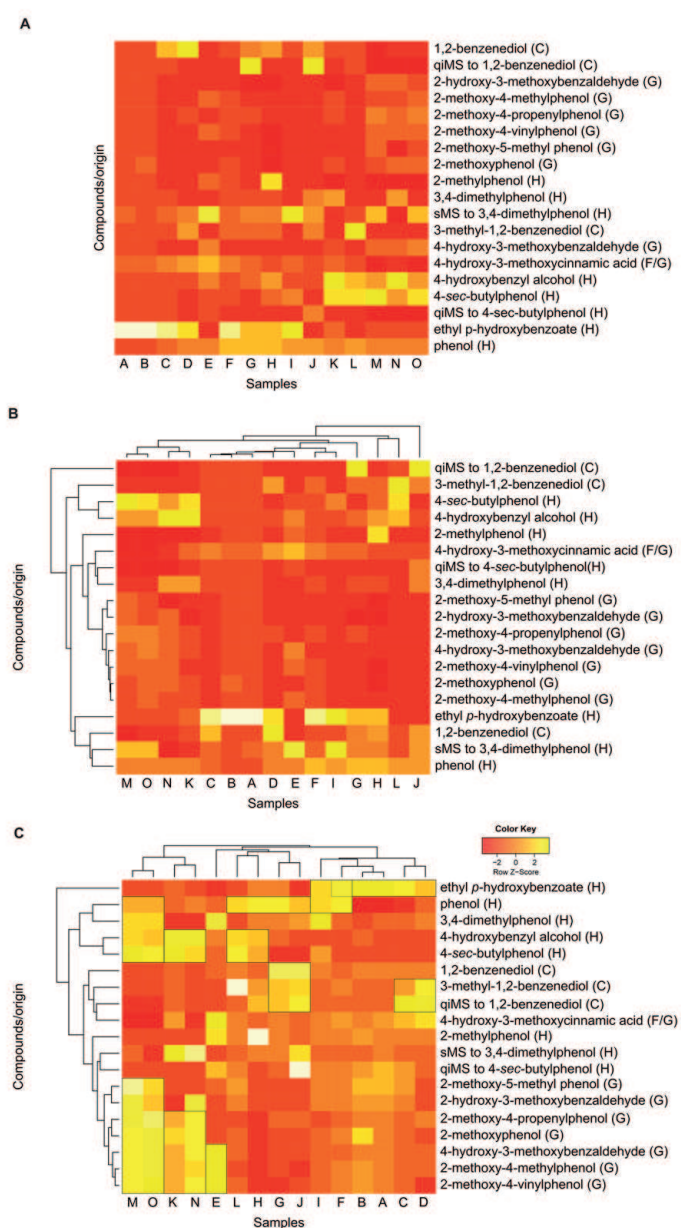


Figure 4.

Representation of the importance of using standardized data for the interpretation of the results. Non-standardized data: A) just ordered alphabetically; it is not possible to identify abundance patterns. B) Data arranged based on the HCA; trace compounds are overshadowed by the most abundant ones. C) Standardized data; compounds with the same origin share patterns of abundance and high similarity.

two different groups of compounds. Another example includes guaiacols, which are derived from guaiacyl (G) units. Under the same conditions of pyrolysis and composition of the samples, their abundance patterns should be the same. In the clustering analysis (CA) of **Figure 3A**, the guaiacols appear together forming a single group. For carbohydrate derivatives, abundance patterns with high similarity can also be identified for related compounds or putative isomers. **Figure 3B** shows the abundance patterns for ethyleneglycol diacetate and compounds with *quasi* identical (qiMS) or similar (sMS) mass spectrum. Another similar example is the independent origin of catechols and guaiacols in some lignocellulosic samples [21]. Catechols can be produced from guaiacols by secondary reactions at high temperatures [12, 21, 36]. However, as seen in **Figure 4**, the catechol abundance patterns across the samples, under the same experimental conditions, are clearly different from those samples with a predominance of G lignin. Therefore, catechols can be considered independently derived from those derivatives from G lignin.

2. Cheminformatics applied to Py-GC/MS

Increased computational capacity, development of powerful deconvolution algorithms and technological advances in analysis equipment have allowed the design of specialized software for chemical analysis. Areas such as omics sciences have particularly benefited from the rise of cheminformatics [26]. However, the application of untargeted analysis is becoming broader and is no longer restricted to the discovery and characterization of compounds in metabolomics. In this sense, it is possible to use the spectral deconvolution software for the processing of the data resulting from Py-GC/MS [21]. Open source software follows the same principle as native GC/MS software for spectra deconvolution and compound identification. However, it allows the use of different input formats for the raw datasets, regardless of the type, resolution and brand of the GC/MS equipment [26, 28]. In addition, different parameters can be adjusted to improve the informative quality of the results; e.g., the parameters used for deconvolution, the use of quality controls and normalization of the relative abundances for a batch of samples, alignment parameters and identification of compounds, use of different reference libraries for mass spectra, retention indices and times of retention. Because Py-GC/MS produces a large number of derived compounds, a lot of information is generated (i.e., mass spectra recorded by the detector in the MS). Omics tools allow deconvolution of all acquired mass spectra for a batch of samples in independent experiments. Basically, the peaks are detected by deconvolution of the mass spectra, smoothing the data points by the least squares method or by linear weighted smoothing average [28, 39]. Afterwards, both the first and second derivatives are considered together with the amplitude of the ions to identify the noise threshold. Based on the noise levels, the initial retention times are calculated for each peak. For the final detection of the peaks, the unsmoothed raw chromatogram is used as a control [28]. The deconvoluted spectra for the batch of samples are aligned based on the similarity of their mass spectra and their RTs. Finally, they are compared with those spectra in the reference MS libraries and the compounds can be identified based on the maximum fit of their RT, RI and mass spectra [26]. Additionally, the deconvoluted datasets for a batch of samples can be normalized and exported in table format. The information contained in the output file is important for comparative analysis: i.e., EI fragmentation pattern, quant mass (m/z of the main ion), averaged RT, InChIKey, total similarity with the reference spectrum and relative abundance of each compound normalized for the entire batch of samples [28]. This information can be used for comparative analysis by multivariate methods. Alternatively, it can

be compared with databases such as the Chemical Entities of Biological Interest (ChEBI) ontology [25], to infer biological characteristics of the original samples based on their derivatives from pyrolysis [21].

The comparative analysis of lignocellulosic samples is highly favored by the normalization process on the data obtained for a batch of samples [21]. The normalization of the deviations of the MS signal intensities is carried out including a series of quality control (QC) samples. The QC samples are one or more samples obtained by combining all samples in the batch. For lignocellulosic materials it is suitable to use alternately one QC sample for every five samples analyzed [21]. The data obtained from the measurement of the QC samples are smoothed by the Lowess of the single-degree least-squares. The coefficients generated on the QC samples are interpolated using the cubic spline and finally all the datasets are aligned based on the spline interpolation result [28].

Additionally, the unknown compounds can be annotated using their elemental formulas and *in silico* mass spectra fragmentation based on public spectral databases, such as MassBank, LipidBlast and GNPS [27, 28]. Currently most open access MS reference libraries are focused on the compounds of interest; i.e., metabolomics and lipidomics. Several of them include precursors or derivatives of lignocellulosic biomass, such as anhydro sugars, furans, pyrans and phenols and their derivatives. Actually, as the areas of application of omics tools diversify (for spectral deconvolution and compound annotation) it can be expected that the diversity and number of compounds incorporated in open access databases will increase.

2.1 Multivariate analysis on exported Py-GC/MS data

Interpretation of the results obtained by Py-GC/MS is a complex process. This is due to the large number of compounds that are generated by pyrolysis and the little information provided by compounds with ambiguous origin, often very numerous (as described above). Multivariate analysis applied to Py-GC/MS data from various materials helps to make data management easier, reduce the information obtained and facilitate the interpretation of the results. It has been used to characterize lignocellulosic samples and other biological samples [40–43].

A common application of Py-GC/MS material analysis is the classification of samples based on the similarities of the compounds they produce. For example, to evaluate different experimental systems [44, 45] or for the optimization of two different methods [46]. It was recently used to characterize and classify lignocellulosic samples applying cheminformatics from a chemotaxonomic approach [21].

Classification of the observations into groups requires the calculation of the distance between each pair of observations. As a result, a distance matrix is obtained, also called a dissimilarity matrix. The distance most commonly used by computational algorithms is the Euclidean distance [47]; i.e., the root sum-of-squares of differences for a set of vectors [48]. As a result, observations with high values of features will be grouped together, likewise, observations with low features values will be grouped together.

Apart from the normalization performed by the spectral deconvolution software on the output datasets, it is highly recommended to standardize the variables before measuring the dissimilarities between observations [49]. This step is considered necessary as it can have a great impact on the results of the analysis on biological data [49, 50]. **Figure 4** represents the differences between non-standardized data and standardized data. In standardization, the values of each variable are weighed by a scale factor in order to give more weight to the small but potentially significant changes in signal intensity [51]. Thus, the standard deviation and the mean usually take values of one and zero, respectively. On the other hand, standardization will help to obtain equivalent similarities regardless of the distance method used (e.g.,

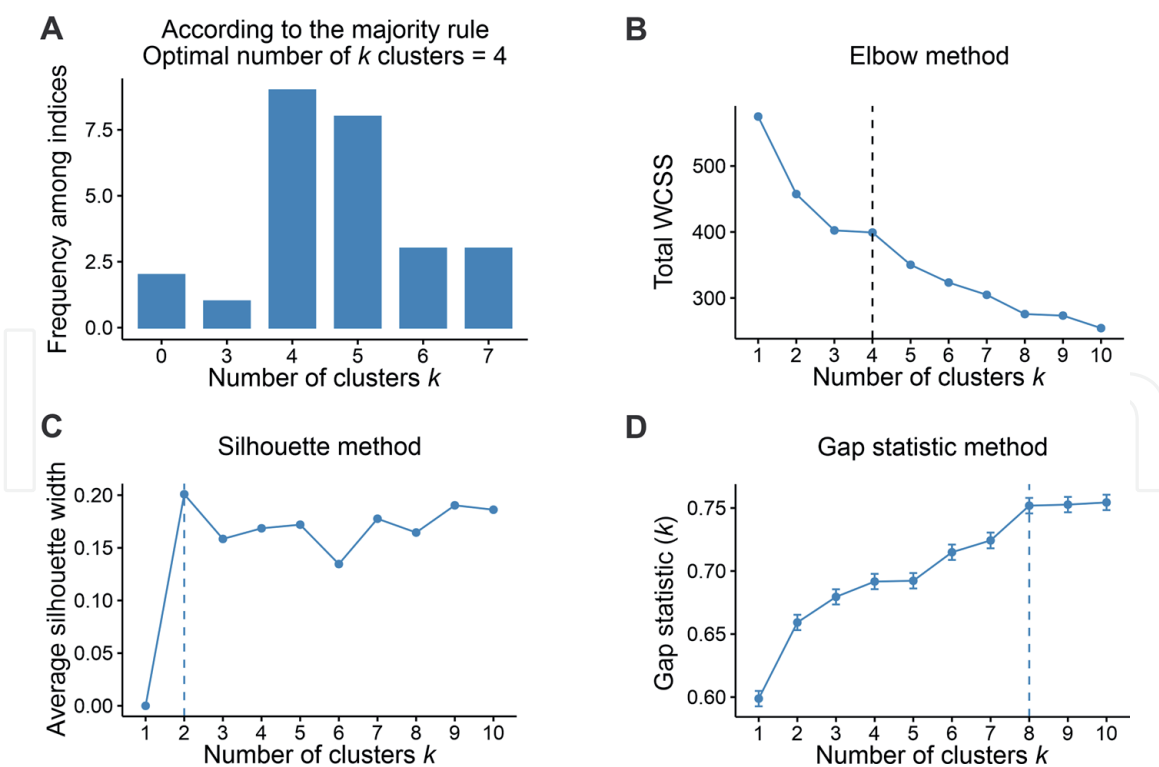
Euclidian, Manhattan, Correlation or Eisen). For example, when using standardized data, there is a functional relationship between Pearson's correlation coefficient and the standardized Euclidean distance, so that both results are comparable [48].

2.2 Groupings by k -means partition

The k -means algorithm is commonly used in the partition of N -dimensional population into k series based on a sample [52, 53]. Where k -series corresponds to the number of clusters to be calculated, arbitrarily specified by the researcher. The algorithm consists of classifying objects forming k clusters, so that for each group the intra-class similarity is minimized, but in turn, each group is as different as possible from the rest [54, 55]. Since the members of each cluster are the most similar to each other, the centre (centroid) of each group is represented by the respective mean. Briefly, the standard procedure for the computational algorithm is as follows: 1) the researcher specifies an arbitrary number of k clusters to be calculate. Alternatively, centroids can also be specified; 2) if the centroids are not specified, they are obtained randomly for each group; 3) by calculating the Euclidean distance, each object is assigned to its closest centroid; 4) the centroids are updated considering the recently incorporated objects; 5) each observation is reviewed with respect to the other clusters to confirm their membership to the respective group. The assignment and update steps are repeated until convergence or the total number of iterations are reached [53]. This method implies advantages when the author has prior knowledge of the analyzed data. For example, in taxonomy, the number of k clusters can refer to the number of data classes to classify [56, 57] or to the taxa that are known or those that want to be tested [21]. In the validation or optimization analysis of methods, it could correspond to the number of systems or criteria that are being considered [58]. An optimal number of k clusters can be more efficient when combined with other multivariate analysis techniques; e.g., in analysis of hierarchical clustering on principal components with partition of k -means (HCPC), which will be explained in the subsequent sections. If there is not enough information to select a specific number of k clusters, the optimal number of k partitions can be inferred using the "elbow" method [49, 59, 60]. The method consists of applying the k -means algorithm to the data, adopting different numbers of k clusters. Then graphically represent the internal variance of the groups, using the number of groups and their respective total within-cluster sum-of-squares (WCSS). The optimal number of k clusters will be indicated by the point where the slope of the WCSS tends to flatten, that is, where the variance is minimized [59, 61, 62]. Due to the randomness with which the initial centroids are selected, it is possible to observe variation in the clusters obtained when replicating the analysis. A suggested solution is to calculate the k -means algorithm several times and select the number of k clusters that generates the lowest WCSS [49]. Furthermore, it is suggested to compare different indices and select an optimal number of k clusters based on the majority rule (**Figure 5**).

2.3 Principal component analysis

Among multivariate analyses, Principal Component Analysis (PCA) is the most common method for extracting information from large datasets generated by analytical pyrolysis [3, 12]. The PCA has different objectives, it is mainly used to reduce the dimensions of the datasets by extracting the most important information. In addition, it is useful to simplify the description of the data series and to analyze the structure of the observations and variables [63–65]. The PCA generates principal components (PC) that result from linear combinations of the original variables (e.g., the identified compounds). The number of these new variables can

**Figure 5.**

Comparison of different methods for calculating the optimal number of k clusters. A) Optimal number of k clusters suggested by the majority rule by analysing all indexes. B) Elbow method. C) Silhouette method. D) Gap Statistic method.

be arbitrarily defined. Commonly, the first component explains the largest possible variance of the dataset and the second, being orthogonal to the first, will be calculated to represent the largest possible variance. The factor scores correspond to the values of these new variables for the original observations (e.g., relative abundances of the compounds). The eigenvalues associated with each component correspond to the sum of the squared factor scores for each component. Thus, the contribution of each observation to a component (i.e., importance of the observation) is represented by the ratio of the square factor score of the observation by the eigenvalue associated with that component. Contributions for a given component can take values from zero to one, so the sum of all contributions for that component is equal to one [65]. Alternatively, the correlation of the two new variables generated by the PCA can be represented by a biplot [66]. Thus, it is possible to know the compounds that contribute the most to the sets obtained in the PCA (Figure 6). As stated, the first two components extracted by the PCA represent the largest variances for the data series. However, to determine the optimal number of components to consider, it is suggested to perform the “scree” test, plotting the eigenvalues as a decreasing function of their size [64]. In the graph, an “elbow” will be observed after the point where the slope of the curve decreases (flattens), therefore the optimal number of components must include all the components before that point (Figure 7A).

2.4 Classification of samples using only the most informative compounds

Multivariate analyses are very useful when working with a large number of data. If lignocellulosic samples are analyzed by Py-GC/MS and the deconvolution method is applied, hundreds of derived compounds can be expected for each sample [21]. The PCA and clustering analysis allows separately to reduce the dimensionality of the datasets, identify relationships between the variables, and quantify the significance of the variables that can explain the resulting clusters [67].

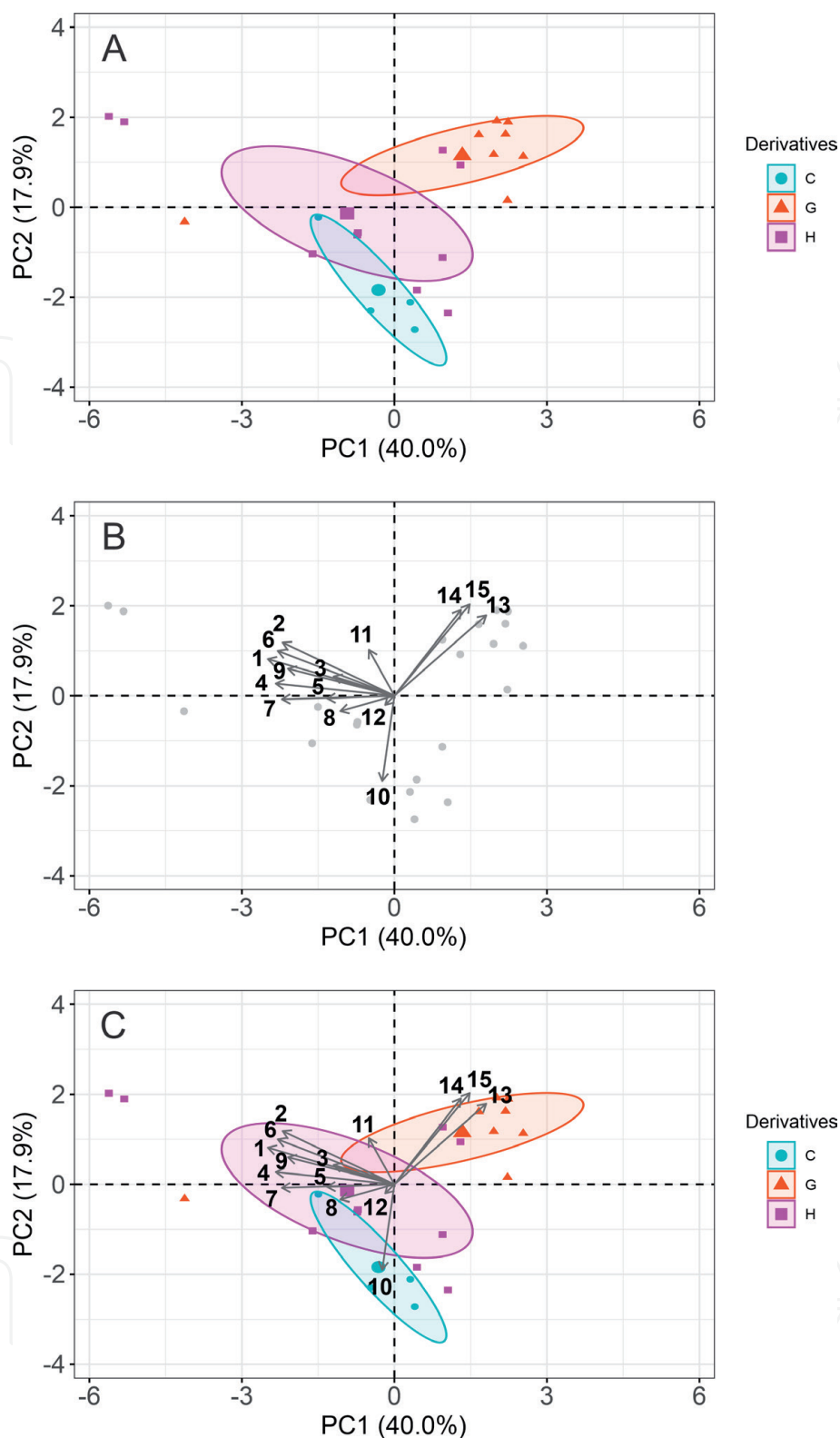
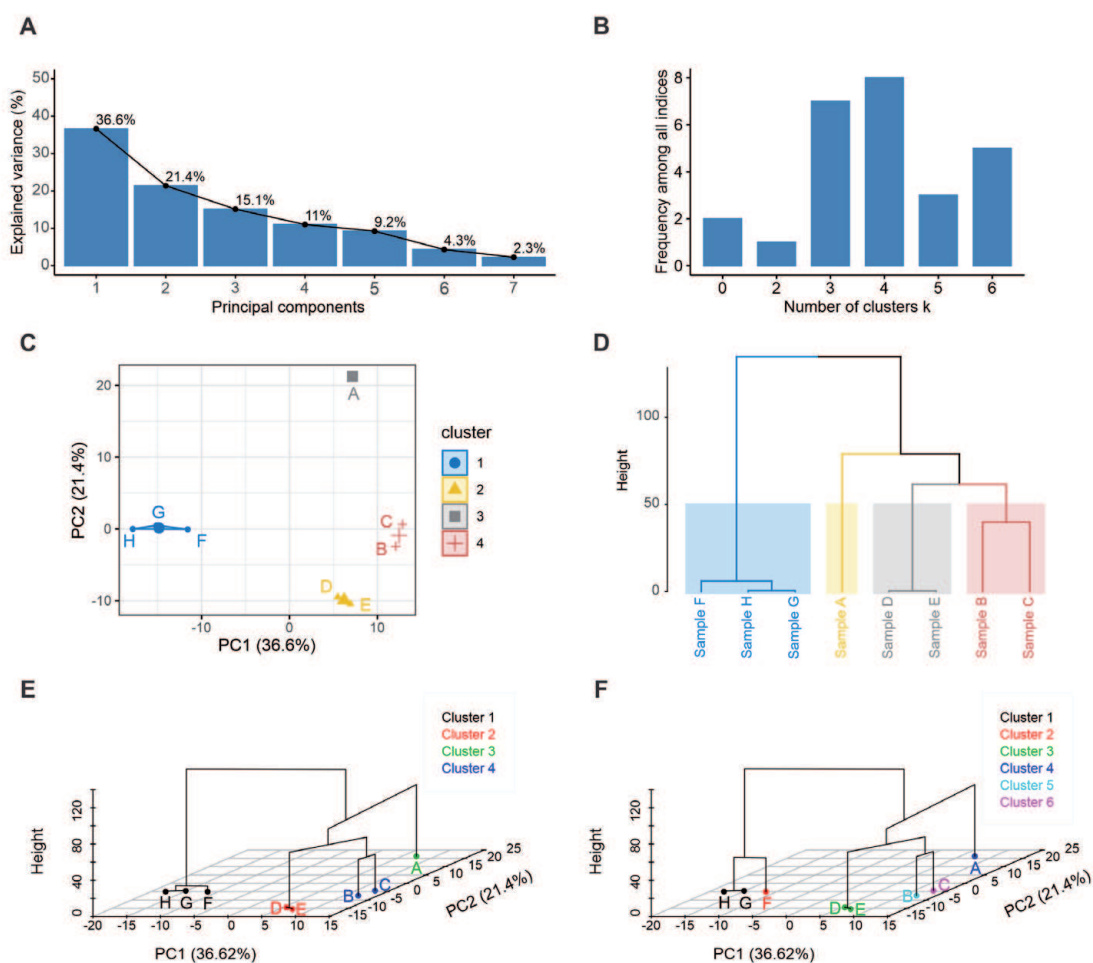


Figure 6.
PCA results: the correlation between the variables generated by the PCA for lignin derivatives is shown. A) Compounds clustered according to their origin: C, catechols; H, phenols; G, guaiacols. B) Biplot that represents the correlation between variables. C) Confidence intervals for the correlation between variables; ellipses represent a significance level of 99%.

The dimensionality of the data directly influences the results; the higher the dimensionality the classifications obtained will be more reliable [68, 69]. For the analysis of chemical compounds in materials the optimal relation of data points to variables is 6:1 or higher, with an absolute minimum of 3:1 [69–71]. However, to achieve these high proportions in the optimal ratio it is necessary to increase the number of experiments. An alternative to achieve the optimal relationship

**Figure 7.**

HCPC analysis for minimizing noise resulting in Py-CG/MS analysis. A) Scree plot, to determine the number of components that explain most of the variance. Number of components used = 5. B) Optimal number of k clusters. Optimal k clusters suggested by the majority rule = 4. C) Factorization of the data series using the PCA. D) Initial hierarchical clustering on the reduced matrix generated by the PCA. E) Clustering obtained using the number of k clusters suggested by the majority rule (the same suggested by the “elbow” method). F) Clusters obtained using a non-optimal number of k clusters.

when it is not possible to increase the number of experiments is by reducing the number of variables [68]. In that sense, the HCPC analysis is a very powerful tool (Figure 7A–C). Compared with PCA and CA, the HCPC analysis increases the objectivity and robustness of the results. That is, the classifications are restricted only to the dimensions that contain the most significant information [67, 72]. In this way, the statistical noise caused by the many uninformative derivatives of pyrolysis is minimized [21]. In addition, it improves the visualization of the data and provides information on the variables (i.e., compounds) that contribute predominantly to the resulting clusters [21, 67]. The HCPC is an exploratory statistical analysis whose computational algorithm can be summarized in three steps: first, the reduction of dimensions can be by any factorial method. PCA for quantitative variables, multiple correspondence analysis for categorical data, or multiple factor analysis to jointly integrate different data blocks [72, 73]. This step allows the determination of the relationships between the concentrations of most abundant compounds and the trace compounds. In addition, it simplifies the dataset by reducing the number of variables to only two principal components that explain most of the variance [74] (Figure 7C). Second, the hierarchical cluster analysis (HCA), by using the Euclidean distance, form clusters of samples according to the similarities in their chemical composition [73, 74] (Figure 7D). Each object is treated as a single cluster

and pairs of groups are successively merged until all clusters merge into one large group [48]. The algorithm uses Ward's method to minimize the total intragroup variance [47, 72, 75]. Finally, the partition with k -means allows to stabilize the groupings obtained by the HCA [67, 73] (**Figure 7E**). In this way, the HCPC applied to the data resulting from Py-GC/MS of lignocellulosic materials allows the samples to be classified based on the abundance patterns of the most informative compounds. That is, statistical noise generated by uninformative, ambiguous, or noisy compounds is suppressed [21].

2.5 Simplified visualization of abundance and similarity patterns from Py-GC/MS data

The heat map method is a simple but highly efficient tool for the graphical representation of large datasets (**Figure 3**). This method is very useful in studies where it is necessary to interpret a large amount of quantitative data; e.g., metabolomics, proteomics, lipidomics, and genomics [76–78]. The quantitative data (i.e., relative abundances of the ions detected by the MS) are represented in different color scales in the format of a two-dimensional matrix [79, 80]. The basic structure of the matrix is given by columns and rows; each column represents a sample and each row represents a compound [76]. The quantitative values correspond to the relative abundance for each compound in each sample. For a certain range of values a particular color is assigned. The highest relative abundances are represented by one end of the color scale and the lowest abundances are represented by the opposite end of that color scale [77]. Additionally, the columns and rows of the matrix are rearranged to recognize significant patterns in the heat map. To do this, rows and columns with similar profiles are arranged so that they are closer to each other, making these profiles easily visible to the eye [79, 80]. The permutation of rows and columns is made based on the result of the CA on the correlation matrix of the variables for each set of variables [77]. Alternatively, the dendrograms resulting from the CA can be represented at the edges of the matrix, both for the samples and for the compounds [77, 79, 81, 82]. This form of representation of the relative abundances is so efficient that after rearranging the rows and columns of the matrices the abundance patterns of the compounds become obvious [76, 83].

The standardization (e.g., Z-transformation) of the variables from each series of variables highly influences the correct representation of the similarity patterns obtained [77, 80]. If raw, non-standardized data are used, the low relative abundances will be obscured by the higher relative abundances (**Figure 4A–C**). When using transformed data it is possible to infer that those compounds with similar abundance patterns imply equal origins [21, 79].

An interactive variant of the heat map method has been referred by several authors in the field of metabolomics [76, 84, 85]. Of course, this can also be applied to Py-GC/MS data. This online variant allows the visualization of important information from the mass spectra on the matrix. Metadata such as mass spectrum, retention time, extracted ion chromatograms (EICs), box and whisker plots as well as matches for each compound can be displayed in real time for each observation [76, 86].

On the other hand, alternative methods for interpreting the data resulting from Py-GC/MS have emerged recently. The Van Krevelen (VK) diagrams have been successfully applied for interpretation of high resolution GC/MS data [3, 87]. These diagrams allow to visualize the chemical composition of complex chemical mixtures by plotting the H:C ratio against the O:C ratio for every compound in the mixture [6]. Thus, the VK diagrams provide information about the classes of compounds present and allow accurately evaluate the number of compounds in

a sample [88]. Furthermore, van Krevelen diagrams play an important role in the deconvolution of high resolution MS spectra for complex lignin samples [6].

3. Potential areas of cheminformatics applied to Py-GC/MS

Due to its versatility, Py-GC/MS has been successfully applied to different areas of knowledge. Among these areas, cheminformatics reviewed in this chapter also has important application opportunities. Environmental, chemical and materials sciences, engineering, energy and biorefinery, biology, biotechnology, and conservation and restoration of cultural heritage are among the most cited in the literature. The fields of application are also varied; for example, in the development and optimization of the properties of new materials and resources, such as synthetic polymers, resins and biofuels [3, 10, 11, 13]. On the other hand, several samples of environmental materials have been characterized by analytical pyrolysis; e.g., organic matter, soil and pollutants in different natural substrates [89–91]. In addition, Kush [13] list a series of applications for polymers, in which the following can be highlighted: 1) identification of polymers through the use of reference libraries, 2) qualitative analysis of copolymers, 3) investigation of thermal stability and kinetics degradation of polymers and copolymers and 4) determination of monomers in polymers and volatile organic compounds.

4. Conclusions

Cheminformatics detailed in this chapter can be applied to the analysis of any type of polymeric materials by Py-GC/MS. The use of open access software to deconvolution of mass spectra streamlines the processing of the resulting data series for a large number of samples. The computational processing capacity of current equipment makes this technique suitable for any laboratory with a Py-GC/MS equipment. In a few minutes a large number of samples can be processed: e.g., deconvolution, alignment and identification of compounds for 30 samples can take about 30 min. On the other hand, the interpretation of the results is greatly aided by the use of the chemometric techniques exemplified here. In addition, cheminformatics makes it possible to compare the mass spectra of the studied compounds, not only with commercial databases, but with other open access databases. Some of the open access databases contain relevant biological information about the compounds (e.g., the ontology of CheBI, MassBank, LipidBlast and GNPS). This is important in studies of materials (e.g., in the case of elements with carcinogenic potential), or of biological interest (e.g., in samples with antibacterial, antibiotic, or medicinal properties). There are currently a significant number of open access MS libraries. Actually, with the diversification of the application field for deconvolution software it is expected that the number of mass spectra in open access libraries will increase. Finally, studies like this leave open the possibility of knowing most of the chemical compounds that take part in the decomposition and secondary reactions during pyrolysis of polymeric materials.

Acknowledgements

JR-R thanks to DGAPA, UNAM by the postdoctoral fellowship [Programa de becas posdoctorales en la UNAM; comunicado 113/2017] and to Facultad de Estudios Superiores Zaragoza, UNAM for supporting this work.

IntechOpen

IntechOpen

Author details

Jorge Reyes-Rivera
Facultad de Estudios Superiores Zaragoza, Universidad Nacional Autónoma de México, Mexico City, Mexico

*Address all correspondence to: jorge.reyes@zaragoza.unam.mx

IntechOpen

© 2021 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

References

- [1] Chen, H. Chemical composition and structure of natural lignocellulose. In *Biotechnology of lignocellulose*. Dordrecht: Springer; 2014. p. 25-71
- [2] Bulushev DA, Ross JR. Catalysis for conversion of biomass to fuels via pyrolysis and gasification: a review. *Catal Today*. 2011;171:1-13
- [3] Grams, J. Chromatographic analysis of bio-oil formed in fast pyrolysis of lignocellulosic biomass. *Reviews in Analytical Chemistry*. 2020;39(1), 65-77
- [4] Briens, C., Piskorz, J., & Berruti, F. Biomass valorization for fuel and chemicals production--A review. *International Journal of Chemical Reactor Engineering*. 2008;6(1)
- [5] Agblevor, F. A., Evans, R. J., & Johnson, K. D. Molecular-beam mass-spectrometric analysis of lignocellulosic materials: I. Herbaceous biomass. *Journal of Analytical and Applied Pyrolysis*. 1994;30(2), 125-144
- [6] Letourneau, D. R., & Volmer, D. A. Mass spectrometry-based methods for the advanced characterization and structural analysis of lignin: A review. *Mass Spectrometry Reviews*. 2021
- [7] Sun, Z., Fridrich, B., de Santi, A., Elangovan, S., & Barta, K. Bright side of lignin depolymerization: toward new platform chemicals. *Chemical reviews*, 2018;118(2), 614-678
- [8] Prothmann, J., Li, K., Hulteberg, C., Spégel, P., Sandahl, M., & Turner, C. Nontargeted Analysis Strategy for the Identification of Phenolic Compounds in Complex Technical Lignin Samples. *ChemSusChem*. 2020;13(17), 4605
- [9] Bridgwater, A. V., & Peacocke, G. V. C. Fast pyrolysis processes for biomass. *Renewable and Sustainable Energy Reviews*. 2000;4(1), 1-73. doi:10.1016/s1364-0321(99)00007-6
- [10] French, R., & Czernik, S. Catalytic pyrolysis of biomass for biofuels production. *Fuel Processing Technology*. 2010;91(1), 25-32
- [11] Isahak, W. N. R. W., Hisham, M. W., Yarmo, M. A., & Hin, T. Y. Y. A review on bio-oil production from biomass by using pyrolysis method. *Renewable and sustainable energy reviews*. 2012;16(8), 5910-5923
- [12] Lourenço, A., Gominho, J., & Pereira, H. Chemical characterization of lignocellulosic materials by analytical pyrolysis. In *Analytical Pyrolysis*. IntechOpen; 2018
- [13] Kusch P. Pyrolysis-Gas Chromatography/Mass Spectrometry of Polymeric Materials, *Advanced Gas Chromatography - Progress in Agricultural, Biomedical and Industrial Applications*. Dr. Mustafa Ali Mohd (Ed.); 2012. ISBN: 978-953-51-0298-4
- [14] Yang H, Yan R, Chen H, Lee DH, Zheng C. Characteristics of hemicellulose, cellulose and lignin pyrolysis. *Fuel*. 2007;86:1781-1788
- [15] Lu, Q., Yang, X. C., Dong, C. Q., Zhang, Z. F., Zhang, X. M., & Zhu, X. F. Influence of pyrolysis temperature and time on the cellulose fast pyrolysis products: Analytical Py-GC/MS study. *Journal of Analytical and Applied Pyrolysis*. 2011;92(2), 430-438
- [16] Faix, O., Meier, D., & Fortmann, I. Thermal degradation products of wood. A collection of electron-impact (EI) mass spectra of monomeric lignin derived products. *Holz als Roh-und Werkstoff*. 1990;48(9), 351-354
- [17] Ralph, J., & Hatfield, R. D. Pyrolysis-GC-MS characterization of forage materials. *Journal of Agricultural and Food Chemistry*. 1991;39(8), 1426-1437

- [18] Reyes-Rivera, J., Soto-Hernández, M., Canché-Escamilla, G., & Terrazas, T. Structural characterization of lignin in four cacti wood: implications of lignification in the growth form and succulence. *Frontiers in plant science*. 2018;9, 1518
- [19] del Río, J. C., Gutiérrez, A., & Martínez, Á. T. Identifying acetylated lignin units in non-wood fibers using pyrolysis-gas chromatography/mass spectrometry. *Rapid communications in mass spectrometry*. 2004;18(11), 1181-1185
- [20] del Río, J. C., Martínez, Á. T., & Gutiérrez, A. Presence of 5-hydroxyguaiacyl units as native lignin constituents in plants as seen by Py-GC/MS. *Journal of analytical and applied pyrolysis*. 2007;79(1-2), 33-38
- [21] Reyes-Rivera, J., Solano, E., Terrazas, T., Soto-Hernández, M., Arias, S., Almanza-Arjona, Y. C., & Polindara-García, L. A. Classification of lignocellulosic matrix of spines in Cactaceae by Py-GC/MS combined with omic tools and multivariate analysis: A chemotaxonomic approach. *Journal of Analytical and Applied Pyrolysis*. 2020;148, 104796
- [22] Meier, D., & Faix, O. Pyrolysis-gas chromatography-mass spectrometry. In *Methods in lignin chemistry*. Berlin, Heidelberg. Springer; 1992. p. 177-199
- [23] Brunow, G., Lundquist, K., & Gellerstedt, G. Lignin. In *Analytical methods in wood chemistry, pulping, and papermaking*. Berlin, Heidelberg: Springer; 1999. p. 77-124
- [24] Wampler, T. P. Analytical pyrolysis: An overview. In: Wampler T.P., editor. *Applied Pyrolysis Handbook*. 2nd ed. New York: Taylor Francis Group; 2007. p. 288
- [25] Degtyarenko, K., De Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., ... & Ashburner, M. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic acids research*. 2007;36(suppl_1), D344-D350
- [26] Tsugawa, H., Cajka, T., Kind, T., Ma, Y., Higgins, B., Ikeda, K., ... & Arita, M. MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nature methods*. 2015;12(6), 523-526
- [27] Tsugawa, H., Kind, T., Nakabayashi, R., Yukihira, D., Tanaka, W., Cajka, T., ... & Arita, M. Hydrogen rearrangement rules: computational MS/MS fragmentation and structure elucidation using MS-FINDER software. *Analytical chemistry*. 2016;88(16), 7946-7958
- [28] Lai, Z., Tsugawa, H., Wohlgemuth, G., Mehta, S., Mueller, M., Zheng, Y., ... & Fiehn, O. Identifying metabolites by integrating metabolome databases with mass spectrometry cheminformatics. *Nature methods*. 2018;15(1), 53-56
- [29] Zhang, B., Zhong, Z., Ding, K., & Song, Z. Production of aromatic hydrocarbons from catalytic co-pyrolysis of biomass and high density polyethylene: analytical Py-GC/MS study. *Fuel*. 2015;139, 622-628
- [30] Lu, Q., Zhou, M. X., Li, W. T., Wang, X., Cui, M. S., & Yang, Y. P. Catalytic fast pyrolysis of biomass with noble metal-like catalysts to produce high-grade bio-oil: analytical Py-GC/MS study. *Catalysis today*. 2018;302, 169-179
- [31] Marques, A. V., & Pereira, H. Aliphatic bio-oils from corks: A Py-GC/MS study. *Journal of Analytical and Applied Pyrolysis*. 2014;109, 29-40
- [32] Faix O, Fortman I, Bremer J, Meier D. Thermal degradation products of wood. Gas chromatographic separation and mass spectrometric characterization of polysaccharide derived products. *Holz Roh Werkst*. 1991;49:213-219

- [33] Luo, Z., Wang, S., Liao, Y., & Cen, K. Mechanism study of cellulose rapid pyrolysis. *Industrial & engineering chemistry research*. 2004;43(18), 5605-5610
- [34] Zhu X, Lu Q. Production of chemicals from selective fast pyrolysis of biomass. In: Momba M, Bux F, editors. *Croatia: Biomass. Sciyo*; 2010. p. 147-16
- [35] Demirbas A. Pyrolysis mechanisms of biomass materials. *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*. 2009;31(13):1186-1193
- [36] Kawamoto, H. Lignin pyrolysis reactions. *Journal of Wood Science*. 2017;63(2), 117-132
- [37] Ponder, G. R., & Richards, G. N. Thermal synthesis and pyrolysis of a xylan. *Carbohydrate Research*. 1991;218, 143-155
- [38] Dobelev, G., Rossinskaja, G., Telysheva, G., Meier, D., & Faix, O. Cellulose dehydration and depolymerization reactions during pyrolysis in the presence of phosphoric acid. *Journal of Analytical and Applied Pyrolysis*. 1999;49(1-2), 307-317
- [39] Savitzky, A., & Golay, M. J. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*. 1964;36(8), 1627-1639
- [40] Zodrow, E. L., & Mastalerz, M. Chemotaxonomy for naturally macerated tree-fern cuticles (Medullosales and Marattiales), Carboniferous Sydney and Mabou sub-basins, Nova Scotia, Canada. *International Journal of Coal Geology*. 2001;47(3-4), 255-275
- [41] Alves, A., Gierlinger, N., Schwanninger, M., & Rodrigues, J. Analytical pyrolysis as a direct method to determine the lignin content in wood: Part 3. Evaluation of species-specific and tissue-specific differences in softwood lignin composition using principal component analysis. *Journal of Analytical and Applied Pyrolysis*. 2009;85(1-2), 30-37
- [42] Mattonai, M., Licursi, D., Antonetti, C., Galletti, A. M. R., & Ribechini, E. Py-GC/MS and HPLC-DAD characterization of hazelnut shell and cuticle: Insights into possible re-evaluation of waste biomass. *Journal of Analytical and Applied Pyrolysis*. 2017;127, 321-328
- [43] Xin, X., Pang, S., de Miguel Mercader, F., & Torr, K. M. The effect of biomass pretreatment on catalytic pyrolysis products of pine wood by Py-GC/MS and principal component analysis. *Journal of Analytical and Applied Pyrolysis*. 2019;138, 145-153
- [44] Gómez, X., Meredith, W., Fernández, C., Sánchez-García, M., Díez-Antolínez, R., Garzón-Santos, J., & Snape, C. E. Evaluating the effect of biochar addition on the anaerobic digestion of swine manure: application of Py-GC/MS. *Environmental Science and Pollution Research*, 2018;25(25), 25600-25611
- [45] Raja Sabaradin, R. Z., & Osman, R. Evaluation of evidence value of car primer using pyrolysis-gas chromatography-mass spectrometry (Py-GC-MS) and chemometrics. *Science Letters (ScL)*. 2021;15(1), 45-37
- [46] Maurer, J., Buffaz, K., Massonnet, G., Roussel, C., & Burnier, C. Optimization of a Py-GC/MS method for silicone-based lubricants analysis. *Journal of Analytical and Applied Pyrolysis*. 2020;149, 104861
- [47] Murtagh, F., & Legendre, P. Ward's hierarchical agglomerative clustering method: which algorithms implement

Ward's criterion?. *Journal of classification*. 2014;31(3), 274-295

[48] Kassambara, A. *Practical guide to cluster analysis in R: Unsupervised machine learning*. Sthda. 2017;1

[49] van den Berg, R. A., Hoefsloot, H. C., Westerhuis, J. A., Smilde, A. K., & van der Werf, M. J. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC genomics*. 2006;7(1), 1-15

[50] Smilde, A. K., van der Werf, M. J., Bijlsma, S., van der Werff-van der Vat, B. J., & Jellema, R. H. Fusion of mass spectrometry-based metabolomics data. *Analytical chemistry*. 2005;77(20), 6729-6736

[51] Bouhleb, J., Bouveresse, D. J. R., Abouelkaram, S., Baéza, E., Jondreville, C., Travel, A., ... & Rutledge, D. N. Comparison of common components analysis with principal components analysis and independent components analysis: Application to SPME-GC-MS volatolomic signatures. *Talanta*. 2018;178, 854-863

[52] MacQueen, J. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. 1967;1(14), 281-297

[53] Kalra, M., Lal, N., & Qamar, S. K-mean clustering algorithm approach for data mining of heterogeneous data. In *Information and Communication Technology for Sustainable Development* Singapore: Springer; 2018. p. 61-70

[54] Arthur, D., & Vassilvitskii, S. How slow is the k-means method?. In *Proceedings of the twenty-second annual symposium on Computational geometry*. 2006; 144-153

[55] Nazeer, K. A., & Sebastian, M. P. Improving the Accuracy and Efficiency of the k-means Clustering Algorithm. In *Proceedings of the world congress on engineering: London: Association of Engineers*. 2009;1, 1-3

[56] Asosheh, A., & Ramezani, N. A comprehensive taxonomy of DDOS attacks and defense mechanism applying in a smart classification. *WSEAS Transactions on Computers*. 2008;7(4), 281-290

[57] Pandey, K. K., & Shukla, D. A study of clustering taxonomy for big data mining with optimized clustering MapReduce model. *decision making*. 2019;1(5), 30

[58] Verma, N., Sharma, V., Kumar, R., Sharma, R., Joshi, M. C., Umapathy, G. R., ... & Chopra, S. On the spectroscopic examination of printed documents by using a field emission scanning electron microscope with energy-dispersive X-ray spectroscopy (FE-SEM-EDS) and chemometric methods: application in forensic science. *Analytical and bioanalytical chemistry*. 2019;411(16), 3477-3495

[59] Kuraria, A., Jharbade, N., & Soni, M. Centroid Selection Process Using WCSS and Elbow Method for K-Mean Clustering Algorithm in Data Mining. *International Journal of Scientific Research in Science, Engineering and Technology*. 2018;190-195

[60] Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (, April). Integration k-means clustering method and elbow method for identification of the best customer profile cluster. In *IOP Conference Series: Materials Science and Engineering*. IOP Publishing. 2018;336(1), 012017

[61] Marutho, D., Handaka, S. H., & Wijaya, E. The determination of cluster number at k-mean using elbow method and purity evaluation on headline news.

In 2018 International Seminar on Application for Technology of Information and Communication. 2018; 533-538

[62] Duong, M. Q., Le Hong Lam, B. T. M., Tu, G. Q. H., & Hieu, N. H. Combination of K-Mean clustering and elbow technique in mitigating losses of distribution network. GMSARN International. 2019;153-158

[63] Jolliffe, I. T. Principal Component Analysis, Encyclopedia of Statistics in Behavioral Science. 2002;30 (3), 487

[64] Kassambara, A. Multivariate Analysis II: Practical Guide to Principal Component Methods in R. 2017

[65] Abdi, H., & Williams, L. J. Principal component analysis. Wiley interdisciplinary reviews: computational statistics. 2010;2(4), 433-459

[66] Gower, J. C., Lubbe, S. G., & Le Roux, N. J. Understanding biplots. John Wiley & Sons. 2011

[67] Strbova, K., Ruzickova, J., & Raclavska, H. Application of multivariate statistical analysis using organic compounds: Source identification at a local scale (Napajedla, Czechia). Journal of environmental management. 2019;238, 434-441

[68] Goodner, K. L., Dreher, J. G., & Rouseff, R. L. The dangers of creating false classifications due to noise in electronic nose and similar multivariate analyses. Sensors and Actuators B: Chemical. 2001;80(3), 261-266

[69] Brenet, S., John-Herpin, A., Gallat, F. X., Musnier, B., Buhot, A., Herrier, C., ... & Hou, Y. Highly-selective optoelectronic nose based on surface plasmon resonance imaging for sensing volatile organic compounds. Analytical chemistry. 2018;90(16), 9879-9887

[70] Wang, B., Cancilla, J. C., Torrecilla, J. S., & Haick, H. Artificial sensing

intelligence with silicon nanowires for ultrasensitive detection in the gas phase. Nano letters. 2014;14(2), 933-938

[71] Shehada, N., Cancilla, J. C., Torrecilla, J. S., Pariente, E. S., Brönstrup, G., Christiansen, S., ... & Haick, H. Silicon nanowire sensors enable diagnosis of patients via exhaled breath. ACS nano. 2016;10(7), 7047-7057

[72] Husson, F., Josse, J., & Pages, J. Principal component methods-hierarchical clustering-partitional clustering: why would we need to choose for visualizing data. Applied Mathematics Department. 2010;1-17

[73] Moyne, O., Castelli, F., Bicout, D. J., Bocard, J., Camara, B., Cournoyer, B., & Le Gouëllec, A. Metatypes of *Pseudomonas aeruginosa* Correlate with Antibiotic Resistance, Virulence and Clinical Outcome in Cystic Fibrosis Chronic Infections. Metabolites. 2021;11(2), 63

[74] Merrot, P., Juillot, F., Noël, V., Lefebvre, P., Brest, J., Menguy, N., ... & Morin, G. Nickel and iron partitioning between clay minerals, Fe-oxides and Fe-sulfides in lagoon sediments from New Caledonia. Science of the Total Environment. 2019;689, 1212-1227

[75] Ward Jr, J. H. Hierarchical grouping to optimize an objective function. Journal of the American statistical association. 1963;58(301), 236-244

[76] Ivanisevic, J., Benton, H. P., Rinehart, D., Epstein, A., Kurczy, M. E., Boska, M. D., ... & Siuzdak, G. An interactive cluster heat map to visualize and explore multidimensional metabolomic data. Metabolomics. 2015;11(4), 1029-1034

[77] Haarman, B. C. B., Riemersma-Van der Lek, R. F., Nolen, W. A., Mendes, R., Drexhage, H. A., & Burger, H. Feature-expression heat maps—A new visual

- method to explore complex associations between two variable sets. *Journal of biomedical informatics*. 2015;53, 156-161
- [78] Gu, Z., Eils, R., & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*. 2016;32(18), 2847-2849
- [79] Gehlenborg, N., & Wong, B. Heat maps. *Nature Methods*. 2012;9(3), 213
- [80] Key, M. A tutorial in displaying mass spectrometry-based proteomic data using heat maps. *BMC bioinformatics*. 2012;13(16), 1-13
- [81] Sneath, P. H. The application of computers to taxonomy. *Microbiology*. 1957;17(1), 201-226
- [82] Ling, R. L. A computer generated aid for cluster analysis. *Communications of the ACM*. 1973;16(6), 355-361
- [83] Vita, F., Taiti, C., Pompeiano, A., Bazihizina, N., Lucarotti, V., Mancuso, S., & Alpi, A. Volatile organic compounds in truffle (*Tuber magnatum* Pico): comparison of samples from different regions of Italy and from different seasons. *Scientific reports*. 2015;5(1), 1-15
- [84] Patti, G. J., Tautenhahn, R., Rinehart, D., Cho, K., Shriver, L. P., Manchester, M., & Siuzdak, G. A view from above: cloud plots to visualize global metabolomic data. *Analytical chemistry*. 2013;85(2), 798-804
- [85] Tautenhahn, R., Cho, K., Uritboonthai, W., Zhu, Z., Patti, G. J., & Siuzdak, G. An accelerated workflow for untargeted metabolomics using the METLIN database. *Nature biotechnology*. 2012;30(9), 826-828
- [86] Gowda, H., Ivanisevic, J., Johnson, C. H., Kurczy, M. E., Benton, H. P., Rinehart, D., ... & Siuzdak, G. Interactive XCMS Online: simplifying advanced metabolomic data processing and subsequent statistical analyses. *Analytical chemistry*. 2014;86(14), 6931-6939
- [87] Kim, J. Y., Park, J., Hwang, H., Kim, J. K., Song, I. K., & Choi, J. W. Catalytic depolymerization of lignin macromolecule to alkylated phenols over various metal catalysts in supercritical tert-butanol. *Journal of Analytical and Applied Pyrolysis*. 2015;113, 99-106
- [88] Brockman, S. A., Roden, E. V., & Hegeman, A. D. Van Krevelen diagram visualization of high resolution-mass spectrometry metabolomics data with OpenVanKrevelen. *Metabolomics*. 2018;14(4), 1-5
- [89] Fabbri, D., Trombini, C., & Vassura, I. Analysis of polystyrene in polluted sediments by pyrolysis—gas chromatography—mass spectrometry. *Journal of chromatographic science*. 1998;36(12), 600-604
- [90] White, D. M., Garland, D. S., Beyer, L., & Yoshikawa, K. Pyrolysis-GC/MS fingerprinting of environmental samples. *Journal of Analytical and Applied Pyrolysis*. 2004;71(1), 107-118
- [91] Campo, J., Nierop, K. G., Cammeraat, E., Andreu, V., & Rubio, J. L. Application of pyrolysis-gas chromatography/mass spectrometry to study changes in the organic matter of macro-and microaggregates of a Mediterranean soil upon heating. *Journal of Chromatography A*. 2011;1218(30), 4817-4827