Article **Journal Applied Computing**
June, 2023 Vol.7 No.20 6-11

Computational algorithm for search and detection of metabolites in CDF(GCxGCxMS) files

Algoritmo computacional para la búsqueda y detección de metabolitos en archivos CDF (GCxGCxMS)

AMARO-RODRÍGUEZ, Isidro†*

Tecnológico Nacional de México/Instituto Tecnológico de Durango, México.

ID 1st Author: *Isidro, Amaro-Rodríguez* / **ORC ID:** 0000-0001-7502-2894, **CVU CONAHCYT ID:** 398361

DOI: 10.35429/JCA.2023.20.7.6.11 Received January 15, 2023; Accepted June 30, 2023

Abstract

Metabolomics is the most novel of the omic sciences, which is responsible for the study and comparison of metabolites. In this study, gas chromatography mass spectrometry (GC-MS) and two-dimensional gas chromatography time-of-flight mass spectrometry (GCxGC-TOFMS) were used for the identification of metabolites found in the blood, which, throw their results in CDF files for further analysis. Metabolic profiles of healthy patients were analyzed using multivariate data analysis techniques, which showed clear differences between the metabolites found, and compared with the universal Golm Metabolome Database (GMD), allowing them to be identified with a high percentage of coincidence. The results found reinforce the advantages of GCxGC-TOFMS and the role of metabolomics in the characterization of metabolites useful for different areas of biology.

Chromatography gas, Mass spectrometry, Metabolomics

Resumen

La metabolómica es considerada la más novedosa de las ciencias –ómicas, la cual se encarga del estudio y comparación de los metabolitos. En este estudio, la espectrometría de masas de cromatografía de gases (GC-MS) y la espectrometría de masas bidimensional de tiempo de vuelo de cromatografía de gases (GCxGC-TOFMS) se emplearon para la identificación de metabolitos que se encuentran en la sangre, las cuales arrojan sus resultados en archivos con formato de documento computable (CDF), para su posterior análisis. Se analizaron perfiles metabólicos de pacientes sanos mediante técnicas de análisis de datos multivariantes, los cuales mostraron diferencias claras entre los metabolitos encontrados, y comparados con la Base de datos universal Golm Metabolome (GMD), lo que permitió identificarlos con un alto porcentaje de coincidencia. Los resultados encontrados refuerzan las ventajas de GCxGC-TOFMS y el papel de la metabolómica en la caracterización de los metabolitos útiles para diferentes áreas de la biología.

Cromatografía de gases, Espectrometría de masas, Metabolómica

Citation: AMARO-RODRÍGUEZ, Isidro. Computational algorithm for search and detection of metabolites in CDF(GCxGCxMS) files. Journal Applied Computing. 2023. 7-20:6-11.

* Correspondence to the Author (E-mail: iamaro@itdurango.edu.mx)
† Researcher contributing as first author.

Introduction

Metabolomics is the study of chemical processes involving all the endogenous and exogenous metabolites in a cell or body fluid. Conventionally speaking, metabolomics studies all the metabolites with molecular mass below 5000 Da, which represents and reflects the functional activities of biological processes. A metabolomic study should, in theory, be able to detect, identify and quantify all the metabolites present in each sample at a given moment; the metabolomic map that is obtained is the representation of biological processes which, in turn, are influenced by individual tissue genetic features, regulation of gene expression, protein abundance and environmental influences (Troisi et al., 2020).

The use of GCxGC-TOFMS offers some advantages compared to the classic one-dimensional GC-MS technique, including improved chromatographic resolution (no increase in analytical execution time), increased sensitivity, improved metabolite identification and separation of reactive artifacts from metabolite peaks (Pasikanti et al., 2010).

By studying metabolic profiles, diseases, risk factors and/or biomarkers (Villavicencio et al.,
2012) can bediscovered,and for this purpose, one of the ways to separatemetabolites from fluids is by using the two-dimensional gas chromatography method coupled to flight-time mass spectrometry (GCxGC-TOFMS), and then switch to detecting them; for which an application was developed, which is designed to process CDF files, from which the characteristics of the profiles are extracted in order to detect their peaks and compare them with the GOLM database (Golm Metabolome Database), and identify the possible metabolites found.

Bioinformatics has brought with it a new mindset and way of carrying out research in biological processes, this has generated new concerns and new sciences that have been developed in order to cover as much as possible the biological elements that surround each process generated in the organisms.

Thus, when trying to analyze large volumes of data generated by complex methods that yield raw signals such as the use of gas chromatograph or mass spectrometry make bioinformatics gain strength. One of the sciences that has been put to work with these elements is so-called cosmic science, which incorporates disciplines such as genomics, proteomics, transcriptomics and metabolomics (Patti et al., 2012).

What is considered to be the most novel is the metabolomic which is responsible for the study and comparison of metabolites, which are molecules of low molecular weight present in each cell, tissue or organism, which are involved in biological processes; it generally studies the set of metabolites present in a biological system, particularly in biofluids such as urine, blood, cerebrospinal fluid, saliva or even in cellular tissues or cultures.

Metabolites are low- and medium molecular weight molecules, less than 1,500 on the Dalton scale, which are involved in cellular processes and reveal how metabolism is working in a given organ. The absence or presence of some metabolites, as well as the relative concentration between them, may be an indicator of disease states or predisposition factors (Díaz, 2016).

Many of today's approaches to metabolomics are characterized by an ongoing transition from qualitative to quantitative methods, similar to the prior development of genomics, transcriptomics and proteomics. Applying the state of the most advanced high-performance technologies as a result of significant growth in the size and complexity of the data generated that leads to increased demand for computational methods of visualization, annotation, and data mining.

Two-dimensional gas chromatography coupled with flight time mass spectrometry (GCxGC- TOFMS) is a maturation technique that stands out for its ability to analyze complex mixtures and has been successfully applied in metabolic research (Gutiérrez, 2002).

ISSN-2531-2952

Objective and goals

This research makes use of an image processing-based algorithm, used for peak detection, which was adjusted with conventional statistics for the purpose of identification.

In addition, the information produced by the CDF files, which extracts the main components of mass, intensity and time, is interpreted and graphs are made for optional representation; detects the peaks found in the signals and compares them with the GOLM database, of the German Research Foundation, with the mixtures obtained in the blood (serum or plasma as treated), which will allow to detect the certain components and thus be able to detect the metabolites, in certain projects, locate anomalies for further analysis and classification, which is the main objective of the research.

Within this research, too, the suitability of different computational programs of processing biomedical signals was analyzed, to obtain relevant information from different metabolic samples, the purpose of which was the development of an application that allowed the reading of files (NCDF or CDF) generated by a gas chromatograph and mass spectrometer.

Methods and materials

Processing algorithms seek to extract characteristics by converting fixed-size signals into vectors, which helps pattern detection, and in this case the peaks that characterize each signal.

When using an image recognition algorithm, applied to signal recognition, it requires an adequacy in the form of the application of the same.

According to Hidalgo (2015) image recognition as a scientific field is considered part of computer vision studies, which in turn is part of a larger field such as artificial intelligence. Its main objective is to identify specific contours, shapes and objects that are performed on tasks such as identifying and classifying images and localizing areas in an image.

A. Algorithm

The developed system presents analysis, search and navigation functionalities through cdf files generated by the chromatograph; this system makes use of signal characteristics such as intensity, and spatial location of signals in the matrix.

The algorithm consists of searching for the maximum elements, in relation to the intensity of the signals, within the array of coefficients,

by doing this, we detect a possible peak, and a search is performed towards the adjoining corners, to determine whether the data belongs to the peak in question or to an independent peak; ana adaptation to the Binarization algorithm, so that it seeks to find behaviors in the sequence ofsignals, to be classified as elements of the same peak or not.

B. Standardization

An improvement in algorithm usage optimization is normalization, transforming the studied variable into a similar variable that has the same proportions, but on a standard scale. The most common form used was to obtain the mean of the signal intensities and divide them by their respective standard deviation.

Another way wasto convert the signals to a unit scale, giving the highest signal strength the value of 1 and converting the other data to its equivalent ratio.

C. Materials

In the development of the application, computer equipment was used, with great information processing capabilities; using basic software such as Excel, and specialized such as R and Matlab, where different algorithms and applications were programmed and tested for the realization of research.

Development

The research was descriptive transsectional, where ten files were available, which were obtained through the blood treatment, obtained by clinically healthy patients and passed by a GCxGC-TOFMS chromatograph.

ISSN-2531-2952

Files containing information from the different compounds, of which it was required to identify the metabolic footprint, which allowed for different specific tests for pattern detection. This opted for an algorithm based on digital image processing, developingan application in the R Language for reading the generated CDF files.

The application developed is divided into four phases: 1) Reading and interpreting the CDF files, coming from the blood treatise by means of the gas chromatograph. 2) Generation of graphical representations; coupled with the loading of the GOLM database, which is the Knowledge Database. 3) Detection of spikes of uploaded files. 4) Comparison between signals obtained in the study against GOLM, to determine their similarities. 5) Concentrate of the results and their export for interpretation.

From its reading and interpretation, graphs are generated that can help the detection of peaks, visually, and thus look for similarities with the metabolites of GOLM. See figures1 and 2 (a) and (b).

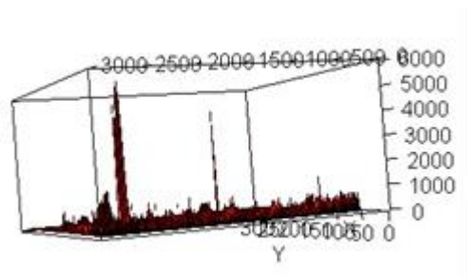
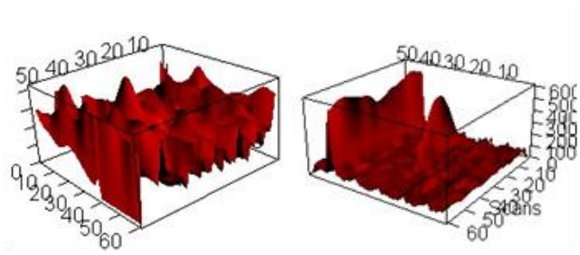


Figure 1 Plot a sample file with all its variables
Source: (Own elaboration, with the use of the R-Studio program)



(a) (b)

Figure 2 Graph of totals and maximums per observed cycle
Source: (Own elaboration, with the use of the R-Studio program)

The application also carries out an updated GOLM Metaboloma database load, which facilitates the search and dissemination of reference mass spectra of biologically active metabolites quantified by gas chromatography (GC) coupled to mass spectrometry (MS), which is a public BD supported by a Very Reliable German Research Foundation (Deutsche Forschungsgemeinschaft, DFG), and comprises mass spectra and retention time rates of pure reference substances and frequently observed mass spectrum labels. This GMD spectral mass library with built-in decision trees is freely accessible for non-commercial use in http://gmd.mpimp-golm.mpg.de/ (Hummel, 2008)

With the help of this database it is possible to compare the signals it has and those obtained in the study, and thus look for similarities between the different metabolites found. see Figure 3.

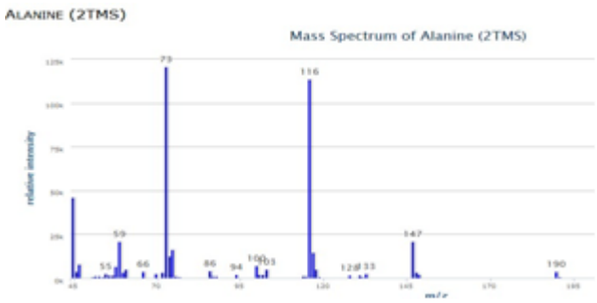


Figure 3 Example graph shown by GOLM on the intensity and m/z ratio of the Alanine compound (2TMS). *Source: http://gmd.mpimp-golm.mpg.de/Spectrums/fca40e62-fe2f-47b7-b20b-f66766c1343d.aspx*

With all the above information, loaded in the application, we proceed to the detection of peaks of the desired files or files, which need a minimum height factor for the peaks, which is given according to the standard deviation. At this point we make use of the developed algorithm, which emphasizes image processingalgorithmof binarization for imageprocessing es, which helps in the detectionof peaks in the file and performs an approximation of the computational time that the machine will need to be able to find the metabolites.

ISSN-2531-2952

The peaks found are generated in a list, shown in a csv file, where the peaks with their similarity percentage and the most similar metabolite of GOLM are found along with their name for quick identification, this for each file analyzed, thus achieving the identification of the metabolites contained in the sample.

Results

Using the developed software, the following results were obtained, ascan be seen in Table 1, the percentage of similarities of the peaks found, as well astheir identification, isobserved, thus finding the metabolites existing in the samples studied.

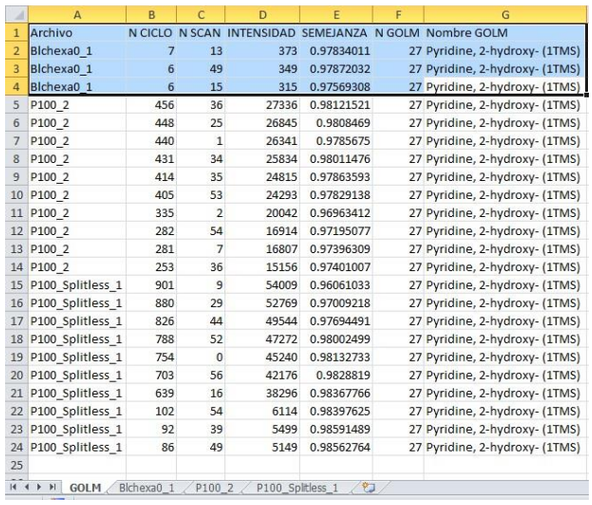
N CICLO	N SCAN	PICO	SEMEJANZA	N GOLM	Nombre GOLM
7	13	373	98%	27	Pyridine, 2-hydroxy- (1TMS)
N CICLO	N SCAN	PICO	SEMEJANZA	N GOLM	Nombre GOLM
6	49	349	98%	27	Pyridine, 2-hydroxy- (1TMS)
N CICLO	N SCAN	PICO	SEMEJANZA	N GOLM	Nombre GOLM
6	15	315	98%	27	Pyridine, 2-hydroxy- (1TMS)

Table 1 Results generated on peaks in a given sample
Source: (Own elaboration)

In the end, the results are concentrated on the similarities with the highest percentage with the metabolites recorded in goLM, for each of the files, which are presented in a csv file, which shows the percentage of similarity of the metabolites found, as well as their position and name, as seen in table 1.

With the help of normalization, it is possible to improve the results obtained by the system. The results show each of the analyzed files, with their respective peaks found and their percentage of similarity to the GOLM database, allowing to determine the name of the metabolite in question.

With these results, it will be possible to identify the different types of metabolites found in blood samples, and thus be able to determine characteristic patterns that help in the diagnosis of diseases, for early detection or improvement of treatment.



	A	B	C	D	E	F	G
1	Archivo	N CICLO	N SCAN	INTENSIDAD	SEMEJANZA	N GOLM	Nombre GOLM
2	Blchexa0_1	7	13	373	0.97834011	27	Pyridine, 2-hydroxy- (1TMS)
3	Blchexa0_1	6	49	349	0.97872032	27	Pyridine, 2-hydroxy- (1TMS)
4	Blchexa0_1	6	15	315	0.97569308	27	Pyridine, 2-hydroxy- (1TMS)
5	P100_2	456	36	27336	0.98121521	27	Pyridine, 2-hydroxy- (1TMS)
6	P100_2	448	25	26845	0.9808469	27	Pyridine, 2-hydroxy- (1TMS)
7	P100_2	440	1	26341	0.9785675	27	Pyridine, 2-hydroxy- (1TMS)
8	P100_2	431	34	25834	0.98011476	27	Pyridine, 2-hydroxy- (1TMS)
9	P100_2	414	35	24815	0.97863593	27	Pyridine, 2-hydroxy- (1TMS)
10	P100_2	405	53	24293	0.97829138	27	Pyridine, 2-hydroxy- (1TMS)
11	P100_2	335	2	20042	0.96963412	27	Pyridine, 2-hydroxy- (1TMS)
12	P100_2	282	54	16914	0.97195077	27	Pyridine, 2-hydroxy- (1TMS)
13	P100_2	281	7	16807	0.97396309	27	Pyridine, 2-hydroxy- (1TMS)
14	P100_2	253	36	15156	0.97401007	27	Pyridine, 2-hydroxy- (1TMS)
15	P100_Splitless_1	901	9	54009	0.96061033	27	Pyridine, 2-hydroxy- (1TMS)
16	P100_Splitless_1	880	29	52769	0.97009218	27	Pyridine, 2-hydroxy- (1TMS)
17	P100_Splitless_1	826	44	49544	0.97694491	27	Pyridine, 2-hydroxy- (1TMS)
18	P100_Splitless_1	788	52	47272	0.98002499	27	Pyridine, 2-hydroxy- (1TMS)
19	P100_Splitless_1	754	0	45240	0.98132733	27	Pyridine, 2-hydroxy- (1TMS)
20	P100_Splitless_1	703	56	42176	0.9828819	27	Pyridine, 2-hydroxy- (1TMS)
21	P100_Splitless_1	639	16	38296	0.98367766	27	Pyridine, 2-hydroxy- (1TMS)
22	P100_Splitless_1	102	54	6114	0.98397625	27	Pyridine, 2-hydroxy- (1TMS)
23	P100_Splitless_1	92	39	5499	0.98591489	27	Pyridine, 2-hydroxy- (1TMS)
24	P100_Splitless_1	86	49	5149	0.98562764	27	Pyridine, 2-hydroxy- (1TMS)
25							

GOLM | Blchexa0_1 | P100_2 | P100_Splitless_1

Figure 5 Concentrated results on metabolites found in the samples analyzed
Source: (Own elaboration)

Conclusions

In this study, the advantages of using the GCxGC-TOFMS tool for metabolic fingerprint detection were demonstrated. In addition, with the development of the application, it was possible to characterize metabolites in a biological sample, the results of which when compared to the GOLM database yield useful results in comparing the different metabolites found in a sample, which will help to detect patterns and thus have the ability to make decisions according to the type of research. It should be noted the use of these tools have numerous applications such as the detection of biomarkers in oils (Silva, 2011) detection of tissue biomarkers (Welthangen, 2005), detection and confirmationof drugs, among others.

In addition, the possibility for the application of different mathematical, computational and bioinformatics tools for the exploration and integration of large heterogeneous atomic data sets is left open, with the intention of improving what has been achieved to this day, and thus providing a reliable tool to be able to diagnose or detect predisposition factors on a disease specific to a human being. Since by knowing more accurately each of the metabolites found in a sample, it will be possible to make comparisons between the population of control patients, those who are already sick or have any symptoms, and thus generate a model capable of making a reliable diagnosis.

ISSN-2531-2952

Future jobs

There are some points to be modified to maximize the optimal operation of the application, one of which is the improvement of the algorithm, which could be done by removing the baseline from the signals from the sample, which is equivalent to the elimination of the "noise" generated by the procedure itself; another would be the parallelization of the routine to find similarities more quickly, since depending on the number of peaks found, it is the amount of time to find its similarity in comparison with the metabolites recorded in GOLM.

Another important point is the being able to generate a package for the R statistical language, useful in the detection of spikes in files generated by the GCxGC-TOFMS process, which is available on the CRAN (http://cran.r- project.org) platform, for use and modification by researchers who require it.

References

Díaz Fernández, Usnavia, & Rodríguez Ferreiro, Annarli Olivia. (2016).

“Biotechnology applications in the development of personalizedmedicine". MEDISAN, 20(5), 678-687.

Gutierrez Bouzán, Ma Carmen Droguet, Marta. (2002). "Gas chromatography and mass spectrometry: identification of odor-causing compounds". Institute of Textile Research and Industrial Cooperation, num. 122, 35-41.

Hidalgo, I., and Sanchez, R. (2015). "Character recognition using images on gas meters in real-world environments (End-of-Grade Work)".Complutense University of Madrid, Spain.

Hummel, J., Selbig, J., Walther, D., & Kopka, J. (2008). "The Golm Metabolome Database: A database for GC-MS based metabolite profiling".In J. Nielsen & M. Jewett (Eds.), Metabolomics a powerful tool in systems biology. Topics in current genetics Vol. 18 (pp. 75–96). Berlin, Heidelberg, New York: Springer.

Kumar Pasikanti, Kishore & Rahmat, Juwita & Cai, Shirong & Mahendran, Ratha & Esuvaranathan, Kesavan & C Ho, Paul & Chun Yong Chan, Eric. (2010). "Metabolic footprinting of tumorigenic and nontumorigenic uroepithelial cells using two-dimensional gas chromatography time-of-flight mass spectrometry".Analytical and bioanalytical chemistry. 398. 1285-93. 10.1007/s00216-010-4055-3.

Patti GJ, Yanes O, Siuzdak G. (2012). "Innovation: Metabolomics: The Apogee of the Omic Trilogy".Nature Reviews Molecular Cell Biolog. 263-269. doi: 10.1038/nrm3314.

Schmidt, C. W. (2004). "Metabolomics: what's happening downstream ofDNA." Environmental Health Perspectives, 112(7), A410–A415.

Silva, Raphael & G. M. Aguiar, Helen & D. Rangel, Mário& A. Azevedo, Débora &Radler de Aquino Neto, Francisco. (2011). "Comprehensive two-dimensional gas chromatography with time of flight mass spectrometry applied to biomarker analysis of oils fromColombia"..Fuel. 90. 2694-2699. 10.1016/j.fuel.2011.04.026. JC Lindon, JK Nicholson, E Holmes. (2007). Handbook of Metabonomics and Metabolomics, Elsevier.

Troisi, J., Cavallo, P., Colucci, A., Pierri, L., Scala, G., Symes, S., Jones, C., & Richards, S. (2020). Metabolomics in genetic testing. Advances in clinical chemistry, 94, 85–153. https://doi.org/10.1016/bs.acc.2019.07.009

Villavicencio, Btglaudia&Colmenarez, Luis& Rivero, Yliana. (2012). "Genomic, Metagenomic, Metabolomic and GenomicS SCIENCES". Lascienciaosmicas.blogspot.mx. Recovered from http://lascienciaosmicas.blogspot.mx/

Welthagen, W., Shellie, R. A., Spranger, J., Ristow, M., Zimmermann, R., &Fiehn, O. (2005). "Comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry (GC- GC-TOF) for high resolution metabolomics: Biomarker discovery on spleen tissue extracts of obese NZO compared to lean C57BL/6 mice". Metabolomics, 1(1), 65-73. DOI: 10.1007/s11306-005-1108-2

ISSN-2531-2952