ICIC Express Letters
Volume **16**, Number **1**, January **2022**

ICIC International ©2022 ISSN 1881-803X
pp. **9–16**

ANTICANCER COMPOUND IDENTIFICATION MODEL OF RODENT TUBER'S LIQUID CHROMATOGRAPHY-MASS SPECTROMETRY DATA

Iwan Binanto[1,2,*], Harco Leslie Hendric Spits Warnars[1]
Nesti Fronika Sianipar[3,4] and Widodo Budiharto[1]

[1]Computer Science Department, BINUS Graduate Program – Doctor of Computer Science
[3]Food Technology Department, Faculty of Engineering
[4]Research Interest Group Biotechnology
Bina Nusantara University
Jl. K. H. Syahdan No. 9, Kemanggisan, Palmerah, Jakarta 11480, Indonesia
spits.hendric@binus.ac.id; { nsianipar; wbudiharto }@binus.edu

[2]Informatics Department
Sanata Dharma University
Kampus 3, Jl. Paingan, Krodan, Maguwoharjo, Kec. Depok, Kabupaten Sleman
Daerah Istimewa Yogyakarta 55281, Indonesia
*Corresponding author: iwan@usd.ac.id

Received April 2021; accepted July 2021

Abstract. *Chemical compounds identification of Liquid Chromatography-Mass Spectrometry data is challenging, especially anticancer compound identification from the Rodent Tuber sample. This paper proposed a novel model for anticancer compound identification of a sample of Liquid Chromatography-Mass Spectrometry Rodent Tuber data. Data sampling was done by linear systematic sampling. The model begins with data labeling which was giving a name to the compound followed by data labeling of the status anticancer compound or not utilizing webscrapping technique. After the data was labeled, the anticancer compound was identified from the data. This model was successfully identifying the most dominant anticancer compounds contained in the sample of Liquid Chromatography-Mass Spectrometry Rodent Tuber data.*

Keywords: Chemical compound identification, Liquid Chromatography-Mass Spectrometer data, Rodent Tuber, Anticancer compound identification, Linear systematic sampling

1. **Introduction.** Research in the field of chemical compounds interpretation and identification from the Liquid Chromatography-Mass Spectrometry data is challenging. Mass Spectrometry (MS) is used to determine the mass of particles, to determine the elemental composition of a sample, and to describe the chemical structure of molecules, such as peptides and other chemical compounds [1]. When combined with Liquid Chromatography (LC), more comprehensive information can be obtained. It will consist of hundreds of thousands of mass per charge (m/z), retention time, and intensity [2]. This combination called Liquid Chromatography-Mass Spectrometry (LC-MS). It is also widely utilized to examine extracts of medicinal plants.

This paper will present and explain our model to identify anticancer compounds from LC-MS data of Rodent Tuber. It contributed to the development of a method of chemical compound identification in LC-MS. As a case study, we use one of the LC-MS data of Rodent Tuber from the studies of Sianipar et al. [3-8]. Their studies resulted in 10 datasets which are proprietary raw data. This means, only the instrument that produced them can read. The dataset should be converted to a format that is human-readable and easier

DOI: 10.24507/icicel.16.01.9

to analyze. There are two steps, which are 1) convert the raw data to .mzXML which is an open format by an open-source software developed by Chambers et al. [9], namely Proteowizard version 3 for Windows, and 2) convert .mzXML to .xlsx, by our software, developed by Python programming. Only one dataset of ten is used in this paper and it is randomly selected.

Using huge native data would be costly. So, data sampling will be carried out. LC-MS data of Rodent Tuber in this case is time-series data [10], and it is necessary to get samples per period time to obtain a representative sample that represents the actual dataset. We use a sampling technique called Linear Systematic Sampling (LSS) as Algorithm 1, because it is simple and produces a sample that represents the actual dataset [11,12] compared to Kim and Hwang [13]. In this case, the number of samples which is about 18,000 and 18,275 was generated. The sample data from the result of Algorithm 1 will be used in this paper.

Algorithm 1. LSS

```
Get data from the dataset.
Create a list that contains the “Retention Time” column without duplication with
    the pivot feature.
Find the skip ($s$) value with a sample size ($y$), while the number population of
    elements ($X$) is the actual data list:
    $s = \text{len(list)}/18000$
Create a list of indexes with the first index as a random integer Rd in range 1 to $s$,
    which determines the sample to be the units numbered Rd, Rd + $s$, Rd + $2s$, and
    so on.
This index is used to retrieve data in the actual data list and put it on the temporary
    list.
```

This paper is categorized as follows: Section 2 describes the related works, Section 3 describes the research method, Section 4 describes the proposed model, Section 5 describes the result and discussion, and Section 6 focuses on conclusions and future work.

2. Related Works.

The back-to-nature trend by using natural ingredients including plants for treatment is increasing. This plant needs to be examined for its chemical content so that it is used more optimally. One of the plants studied intensively in its ingredients is Rodent Tuber [4,14-19]. Apart from the study of chemical compound content, breeding and its effect on chemical compound content were also investigated [3-8], but they still manually identify the chemical compounds that exist through Gas Chromatography-Mass Spectrometry and Liquid Chromatography-Mass Spectrometry data. This identification is assisted by an online database. It takes a lot of time because of done manually.

The more advanced technology, the greater data generated, and the greater challenges. Done manually becomes impossible. There are already many software tools for this, such as XCMS [20-22] which aim to facilitate interpretation of metabolomic data by scientists with limited background in bioinformatics or statistics; MAVEN [23,24] which provide interactive processing of LC-MS-based metabolomics data; CAMERA [25] which is designed to postprocess XCMS feature lists and to collect all features related to a compound into a compound spectrum; MetaboAnalystR [26]; and the one which is also summarized by Binanto et al. [27]. Almost all of them do not have the feature for identifying anticancer chemical compounds; they discuss more about preprocessing, signal processing, or whole identification of chemical compounds.

3. Research Method.

Our research method directly uses the proposed model as shown in Figure 1. The first stage of this method is an initial stage by implementing Algorithm 1. This algorithm produces a representative sample of the whole dataset which is 18,275

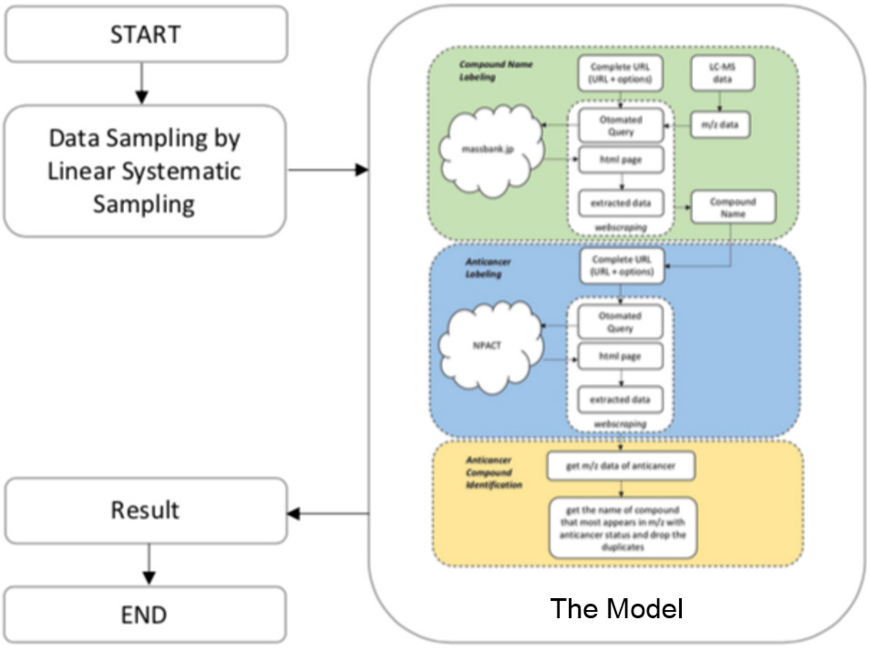

START
Data Sampling by Linear Systematic Sampling
Compound Name Labeling
Complete URL (URL + options)
LC-MS data
massbank.jp
Otomated Query
m/z data
html page
extracted data
Compound Name
webscraping
Anticancer Labeling
Complete URL (URL + options)
NPACT
Otomated Query
html page
extracted data
webscraping
Anticancer Compound Identification
get m/z data of anticancer
get the name of compound that most appears in m/z with anticancer status and drop the duplicates
Result
END
The Model

Figure 1. Research method

records out of 985,925 records. Then this sample was processed using the proposed model depicted in Figure 2.

4. **Proposed Model.** The proposed model has three main stages, which are 1) labeling the name of the compound and its real mass and formula; 2) labeling of compound status: anticancer or not; 3) identification of anticancer compounds. This proposed model is illustrated in Figure 2. Data sampling is not included in this model because for actual conditions, it does not use sample data, but original data.

Labeling is the main thing that must be done in this model because it will be used to retrieve other data from the online database. We utilize an existing online database, namely massbank.jp [28] and NPACT (Naturally occurring Plant-based Anticancerous Compound-Activity-Target database) at http://crdd.osdd.net/raghava/npact/ [29] for labeling. Unfortunately, there is not any piece of information about the Application Programming Interface (API) on those websites. Hence, we use webscraping technique.

Webscraping technique was originally developed for business purposes in the private sector. It is used to get content from websites to analyze certain structured or unstructured data. However, it offers great benefits for those looking for specific information [30-32]. In this case, it will be used for getting compound names from massbank.jp and anticancer status from NPACT.

4.1. **Compound name labeling.** There are input forms for m/z and formulas that can be filled in one of them on massbank site, to get the compound name. So, m/z value is entered into the form to be queried to massbank. There are a lot of options, but we just used the default options.

The first step to do webscraping is to get a complete URL. To get the name of the compound in masbank.jp, the m/z value is used. Manually, m/z value entered into the "Exact Mass" filling form's massbank.jp and the search button clicked; since then, there emerged the complete URL with its options in the browser's address bar as shown in Figure 3. This complete URL will be used for webscraping.

To get the preferred data, we must understand the structure of the HTML page generated by the query. Webscraping technique can run well and get the preferred data by

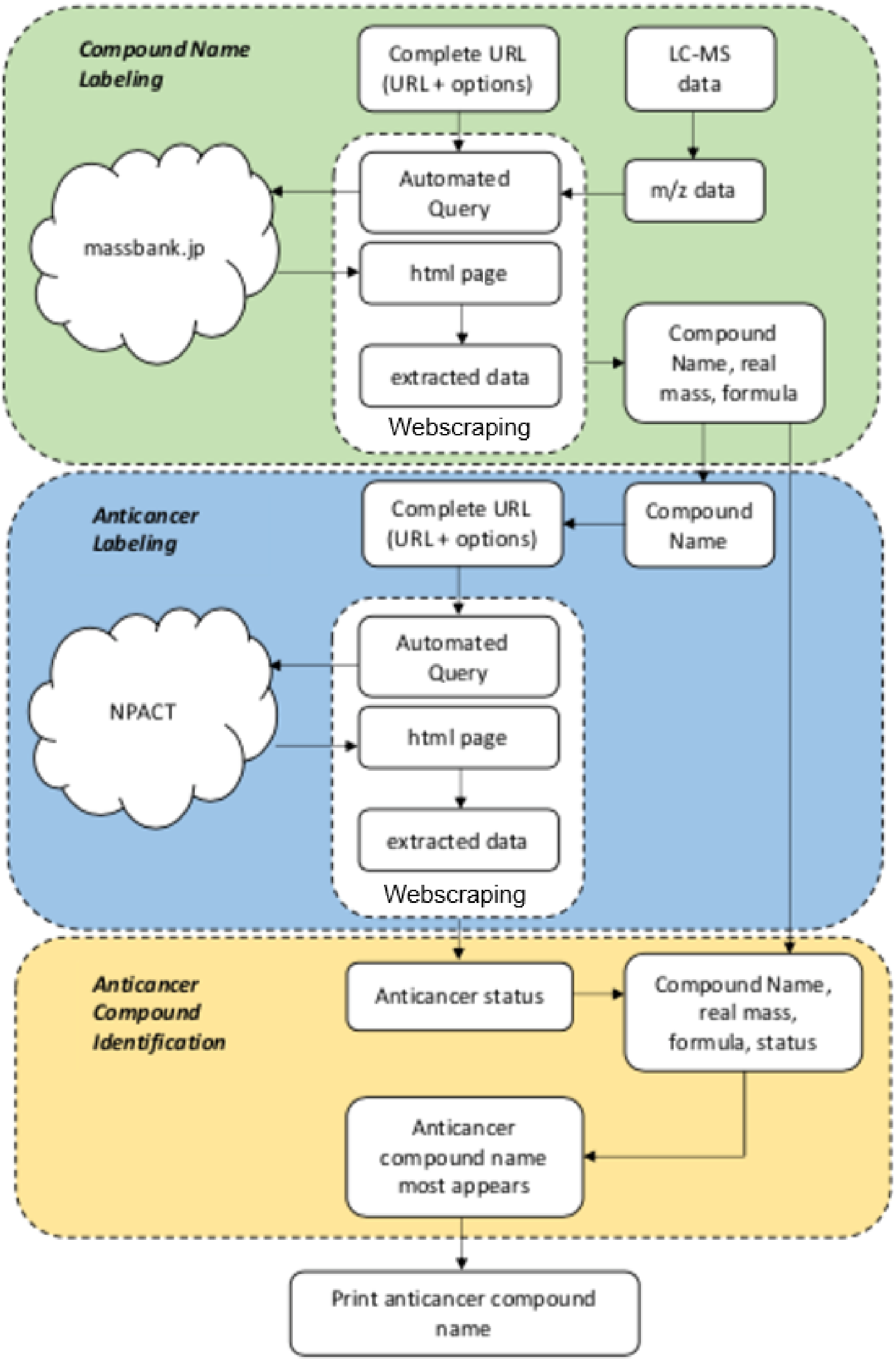
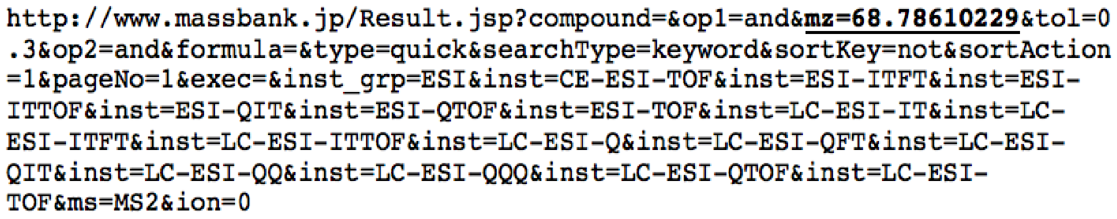

Compound Name Labeling
Complete URL (URL + options)
LC-MS data
massbank.jp
Automated Query
m/z data
html page
extracted data
Compound Name, real mass, formula
Webscraping
Anticancer Labeling
Complete URL (URL + options)
Compound Name
NPACT
Automated Query
html page
extracted data
Webscraping
Anticancer Compound Identification
Anticancer status
Compound Name, real mass, formula, status
Anticancer compound name most appears
Print anticancer compound name

FIGURE 2. The model



```
http://www.massbank.jp/Result.jsp?compound=&op1=and&mz=68.78610229&tol=0
.3&op2=and&formula=&type=quick&searchType=keyword&sortKey=not&sortAction
=1&pageNo=1&exec=&inst_grp=ESI&inst=CE-ESI-TOF&inst=ESI-ITFT&inst=ESI-
ITTOF&inst=ESI-QIT&inst=ESI-QTOF&inst=ESI-TOF&inst=LC-ESI-IT&inst=LC-
ESI-ITFT&inst=LC-ESI-ITTOF&inst=LC-ESI-Q&inst=LC-ESI-QFT&inst=LC-ESI-
QIT&inst=LC-ESI-QQ&inst=LC-ESI-QQQ&inst=LC-ESI-QTOF&inst=LC-ESI-
TOF&ms=MS2&ion=0
```

FIGURE 3. Massbank's complete URL

python programming with its library. The preferred data are compound name, compound formula, and real m/z in massbank.jp.

The value of real m/z is the value closest to the input, because, during the experiment, no m/z values were retrieved the same. In our opinion, it was because signal processing is not carried out and the mass itself is a float number not an integer; upon this, we call it a brute force process.

4.2. **Compound status labeling.** This stage aims to label chemical compound whether it is an anticancer compound or not. To do this, the Naturally occurring Plant-based Anticancer Compound-Activity-Target (NPACT) database is used [29]. It does not provide an application programming interface, so webscraping technique is utilized.

This stage is similar to the compound name labeling stage, and the difference is the detailed steps, such as the complete URL and how to extract data from the obtained HTML page. It is simpler than the compound name labeling stage above.

Using a compound name to get the status of the compound in NPACT. Manually, the compound name entered into the filling form's field provided in NPACT and the search button clicked; since then, there emerged the complete URL in the browser's address bar as shown in Figure 4. This complete URL will be used for webscraping.

`https://webs.iiitd.edu.in/raghava/npact/search_cmpnd.php?cmpnd=`**`sanguinarine`**

Figure 4. NPACT's complete URL

Bold text in Figure 4 is the actual variable that will be replaced with a compound name from the previous stage.

The result of this stage is to obtain the status of the compound name as an anticancer compound (labeled with 1) or not anticancer compound (labeled with 0). Based on this label, the most dominant anticancer compounds will be identified.

4.3. **Identification.** This stage aims to identify the dominant anticancer chemical compounds in the sample of LC-MS Rodent Tuber data. The way it works is to find the most detected anticancer compounds in the sample of LC-MS Rodent Tuber data over the retention time. The algorithm is shown in Algorithm 2.

Algorithm 2. Identification Anticancer Compound

```
Get data from the dataset.
Get the compound labeled "1" (anticancer) and place it in a new array.
Find and count the occurrence of each anticancer compound already in the array.
Print the finding which is the most occurrence anticancer compound.
```

The detected anticancer compound is the most appearing anticancer compound and could be said it is dominant. It caused this compound have strong chemical bonds so from the early time it is detected to the last time.

5. **Result and Discussion.** The use of webscraping for labeling really speeds up and simplifies the task, both for labeling name compounds and for the status of anticancer compounds or not. The data sample contains anticancer and non-anticancer chemical compounds, detected from the early retention time. All the anticancer compounds appear as it is shown in Figure 5. It can be seen that many anticancer compounds were detected at the early retention time. However, not many were detected at the end of the retention time.

Algorithm 2 is used to identify anticancer compounds and successfully identify anticancer compounds contained in the data sample. After the algorithm was run, the results

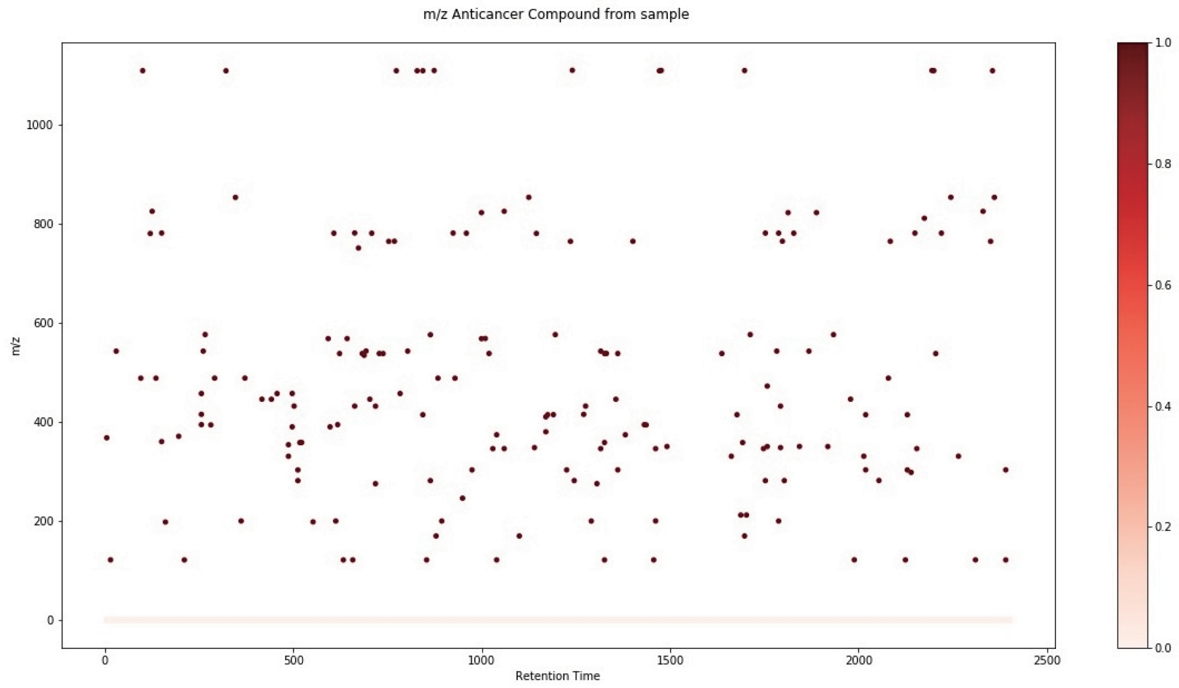


Figure 5. (color online) Anticancer compounds from sample based on m/z

Table 1. The ten most anticancer compound detected

No.	Compound name	Real m/z	First detection in retention time in second	Last detection in retention time in second	Identified count
1.	Benzoic acid	122.03678	15	2390	12
2.	Saikosaponin a	780.99298	150	2220	11
3.	Amentoflavone	538.09003	622	2205	10
4.	Ginsenoside Rb1	1108.60291	100	2355	8
5.	Doxorubicin	543.17407	30	1868	7
6.	Asiatic acid	488.35019	95	2079	7
7.	Digitoxin	764.43469	753	2350	7
8.	Harmalol	200.241	361	1788	6
9.	Pseudobaptigenin	282.0528	512	2054	6
10.	Diosgenin	414.634	843	2129	6

showed the 10 most identified anticancer compound in Table 1. It has been sorted from the most identified compound. “Identified count” represents the number of the same compounds identified.

Table 1 shows benzoic acid with m/z is 122.03678 identified earliest at the 15th second of the retention time and longest at the 2390th second and detected 12 times. This shows benzoic acid has a strong bond so it remains intact from the beginning to the end of the retention time.

6. **Conclusion.** The proposed model has succeeded in identifying existing anticancer compounds and providing information on the dominant anticancer compounds which is benzoic acid with m/z 122.03678. Followed by “Saikosaponin a” as the second dominant, and so forth.

This study uses sample data, so it is still a reflection of the actual data. For future work, the real data will be used and will be compared with the processing results with sample data. In addition, the model will be tested on other LCMS datasets.

REFERENCES

[1] P. R. Kumar, S. R. Dinesh and R. Rini, LCMS – A review and a recent update, *World Journal of Pharmacy and Pharmaceutical Sciences*, vol.5, no.5, pp.377-391, DOI: 10.20959/wjpps20165-6656, 2016.

[2] F. Fernández-Albert, *Machine Learning Methods for the Analysis of Liquid Chromatography-Mass Spectrometry Datasets in Metabolomics*, Ph.D. Thesis, Universitat Politècnica de Catalunya, 2014.

[3] N. F. Sianipar, D. Laurent, R. Purnamaningsih and I. Darwati, Genetic variation of the first generation of rodent tuber (*Typhonium flagelliforme* Lodd.) mutants based on RAPD molecular markers, *HAYATI Journal of Biosciences*, vol.22, no.2, pp.98-104, DOI: 10.4308/hjb.22.2.98, 2015.

[4] N. F. Sianipar, R. Purnamaningsih, D. L. Gumanti, Rosaria and M. Vidianti, Analysis of gamma irradiated-third generation mutants of rodent tuber (*Typhonium flagelliforme* Lodd.) based on morphology, RAPD, and GC-MS markers, *Pertanika J. Trop. Agric. Sci.*, vol.40, no.1, pp.185-202, 2017.

[5] N. F. Sianipar, R. Purnamaningsih, I. Darwati, D. Laurent and Chelen, The effects of gamma irradiation and somaclonal variation on morphology variation of mutant rodent tuber (*Typhonium flagelliforme* Lodd.) LINES, *The 3rd International Conference on Biological Science*, pp.637-645, 2015.

[6] N. F. Sianipar and R. Purnamaningsih, Enhancement of the contents of anticancer bioactive compounds in mutant clones of rodent tuber (*Typhonium flagelliforme* Lodd.) based on GC-MS analysis, *Pertanika J. Trop. Agric. Sci.*, vol.41, no.1, pp.305-320, 2018.

[7] N. F. Sianipar, R. Purnamaningsih and Chelen, Effect of gamma irradiation on protein profile of rodent tuber (*Typhonium flagelliforme* Lodd.) in vitro mutant based on 1D and 2D page analyses, *Jurnal Teknologi*, vol.78, no.10-4, pp.35-40, 2016.

[8] N. F. Sianipar, R. Purnamaningsih and Rosaria, Bioactive compounds of fourth generation gamma-irradiated *Typhonium flagelliforme* Lodd. mutants based on gas chromatography-mass spectrometry, *The 2nd International Conference on Agricultural and Biological Sciences*, DOI: 10.1088/1755-1315/41/1/012025, 2016.

[9] M. C. Chambers, B. Maclean, R. Burke, D. Amodei, D. L. Ruderman, S. Neumann, L. Gatto, B. Fischer, B. Pratt, J. Egertson, K. Hoff, D. Kessner, N. Tasman, N. Shulman, B. Frewen, T. A. Baker, M.-Y. Brusniak, C. Paulse, D. Creasy, L. Flashner, K. Kani, C. Moulding, S. L. Seymour, L. M. Nuwaysir, B. Lefebvre, F. Kuhlmann, J. Roark, P. Rainer, S. Detlev et al., A cross-platform toolkit for mass spectrometry and proteomics, *Nature Biotechnology*, vol.30, no.10, pp.918-920, DOI: 10.1038/nbt.2377, 2012.

[10] I. Binanto, H. Leslie, H. Spits, N. F. Sianipar and W. Budiharto, Understanding LCMS data for identification of chemical compounds contained in rodent tuber: Timeseries or not, *Systematic Reviews in Pharmacy*, vol.12, no.1, pp.648-654, 2021.

[11] S. L. Lohr, *Sampling: Design and Analysis*, 2nd Edition, Brooks/Cole, Cengage Learning, Boston, 2010.

[12] R. Arnab, Systematic sampling, in *Survey Sampling Theory and Applications*, Academic Press, DOI: 10.1016/B978-0-12-811848-1.00004-2, 2017.

[13] Y. Kim and H. Hwang, Approximate consistent weighted sampling for efficient top-k search, *International Journal of Innovative Computing, Information and Control*, vol.16, no.3, pp.1125-1132, 2020.

[14] C. Y. Choo, K. L. Chan, K. Takeya and H. Itokawa, Cytotoxic activity of *Typhonium flagelliforme* (Araceae), *Phytotherapy Research*, vol.15, no.3, pp.260-262, DOI: 10.1002/ptr.717, 2001.

[15] C. Y. Chee, L. C. Kit, W. S. Teng, Y. Hitotsuyanagi and K. Takeya, The cytotoxicity and chemical constituents of the hexane fraction of *Typhonium flagelliforme* (Araceace), *Journal of Ethnopharmacology*, vol.77, no.1, pp.129-131, DOI: 10.1016/S0378-8741(01)00274-4, 2001.

[16] S. Mohan, A. Bustaman, S. Ibrahim, A. S. Al-Zubairi and M. Aspollah, Anticancerous effect of *Typhonium flagelliforme* on human T4-lymphoblastoid cell line CEM-ss, *Journal of Pharmacology and Toxicology*, vol.3, no.6, pp.449-456, DOI: 10.3923/jpt.2008.449.456, 2008.

[17] A. Bustamam, S. Mohan, S. Ibrahim, A. S. Al-Zubairi, M. Aspollah, R. Abdullah and M. M. Elhassan, In vitro ultramorphological assessment of apoptosis on CEMss induced by linoleic acid-rich fraction from *Typhonium flagelliforme* tuber, *Evidence-Based Complementary and Alternative Medicine*, vol.2011, DOI: 10.1093/ecam/neq010, 2011.

[18] E. Purwaningsih, Y. Suciati and E. Widayanti, Anticancer effect of a *Typhonium flagelliforme* L. in raji cells through telomerase expression, *Indonesian Journal of Cancer Chemoprevention*, vol.8, no.1, pp.15-20, 2017.

[19] R. Purnamaningsih and N. F. Sianipar, Analysis of bioactive compounds and morphological traits in Indonesian rodent tuber mutant clones of pekalongan accession using GC-MS, *Jurnal Teknologi*, vol.80, no.2, pp.131-136, 2018.

[20] C. A. Smith, E. J. Want, G. O. Maille, R. Abagyan and G. Siuzdak, XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification, *Analytical Chemistry*, vol.78, no.3, pp.779-787, DOI: 10.1021/ac051437y, 2006.

[21] H. P. Benton, D. M. Wong, S. A. Trauger and G. Siuzdak, XCMS2: Processing tandem mass spectrometry data for metabolite identification and structural characterization, *Analytical Chemistry*, vol.80, no.16, pp.6382-6389, DOI: 10.1021/ac800795f.XCMS, 2009.

[22] H. Gowda, J. Ivanisevic, C. H. Johnson, M. E. Kurczy, H. P. Benton, D. Rinehart, T. Nguyen, J. Ray, J. Kuehl, B. Arevalo, P. D. Westenskow, J. Wang, A. P. Arkin, A. M. Deutschbauer, G. J. Patti and G. Siuzdak, Interactive XCMS online: Simplifying advanced metabolomic data processing and subsequent statistical analyses, *Analytical Chemistry*, vol.86, no.14, pp.6931-6939, DOI: 10.1021/ac500734c, 2014.

[23] E. Melamud, L. Vastag and J. D. Rabinowitz, Metabolomic analysis and visualization engine for LC-MS data, *Analytical Chemistry*, vol.82, no.23, pp.9818-9826, DOI: 10.1021/ac1021166, 2010.

[24] M. F. Clasquin, E. Melamud and J. D. Rabinowitz, LC-MS data processing with MAVEN: A metabolomic analysis and visualization engine, *Current Protocols in Bioinformatics*, DOI: 10.1002/0471250953.bi1411s37.LC-MS, 2012.

[25] C. Kuhl, R. Tautenhahn and S. Neumann, LC-MS peak annotation and identification with CAMERA, *CAMERA Documentation*, 2010.

[26] J. Chong, M. Yamamoto and J. Xia, MetaboAnalystR 2.0: From raw spectra to biological insights, *Metabolites*, vol.9, no.3, DOI: 10.3390/metabo9030057, 2019.

[27] I. Binanto, H. L. H. S. Warnars, N. F. Sianipar and B. S. Abbas, LC-MS analysis: Mini review frequety used open source, *2019 6th International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE)*, Semarang, Indonesia, pp.1-5, 2019.

[28] H. Horai, M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, Y. Ojima, K. Tanaka, S. Tanaka, K. Aoshima, Y. Oda, Y. Kakazu, M. Kusano, T. Tohge, F. Matsuda, Y. Sawada, M. Y. Hirai, H. Nakanishi, K. Ikeda, N. Akimoto, T. Maoka, H. Takahashi, T. Ara, N. Sakurai, H. Suzuki, D. Shibata, S. Neumann, T. Iida, K. Tanaka et al., MassBank: A public repository for sharing mass spectral data for life sciences, *Journal of Mass Spectrometry*, vol.45, no.7, pp.703-714, DOI: 10.1002/jms.1777, 2010.

[29] M. Mangal, P. Sagar, H. Singh, G. P. S. Raghava and S. M. Agarwal, NPACT: Naturally occurring plant-based anti-cancer compound-activity-target database, *Nucleic Acids Research*, vol.41, no.D1, pp.1124-1129, DOI: 10.1093/nar/gks1047, 2013.

[30] M. Herrmann and L. Hoyden, Applied webscraping in market research, *The 1st International Conference on Advanced Research Methods and Analytics*, Valencia, DOI: 10.4995/carma2016.2016.3131, 2016.

[31] M. Shreesha, S. B. Srikara and R. Manjesh, A novel approach for news extraction using webscraping technique, *The 3rd National Conference on Image Processing, Computing, Communication, Networking and Data Analytics (NCICCNDA2018)*, pp.359-362, DOI: 10.21467/proceedings.1.56, 2018.

[32] R. McAlister, Webscraping as an investigation tool to identify potential human trafficking operations in Romania, *Proc. of the 2015 ACM Web Science Conference*, DOI: 10.1145/2786451.2786510, 2015.