Computational Applications in Secondary Metabolite(CAiSMD) 2021 https://caismd.indiayouth.info/

OP03: OsamorSoft: A Tool for Clustering Genomic Data

Victor Chukwudi Osamor[1]*(Professor), Theresa Okediya[1](M.Sc. Student)

[1]Department of Computer and Information Sciences, Covenant University, Ota, Nigeria

vcosamor@gmail.com

Introduction

- Clustering of genomic dataset is often a way of processing through-put data to allow for intuitive unveiling of hidden pattern that connotes some status and functionalities in a system.
- Often, there are challenges of different clustering algorithms giving different results on same dataset due to differences in the techniques in algorithmic development and assumptions.

- Our aim is to describe OsamorSpreadsheet and employ different clustering techniques to cluster molecules from natural products database and
- Investigate the level of cluster quality using our newly developed cluster validation tool called OsamorSoft.

Osamor and Osamor *J Big Data* (2020) 7:48
https://doi.org/10.1186/s40537-020-00325-6

View Edit Select Text

Journal of Big Data

RESEARCH Open Access

OsamorSoft: clustering index for comparison and quality validation in high throughput dataset

Ifeoma Patricia Osamor[1] and Victor Chukwudi Osamor[2*]

*Correspondence: vcosamor@gmail.com; victor.osamor@covenantuniversity.edu.ng
[2] Department of Computer and Information Sciences, College of Science and Technology, Covenant University, Ota, Ogun State, Nigeria

Abstract

The existence of some differences in the results obtained from varying clustering k-means algorithms necessitated the need for a simplified approach in validation of cluster quality obtained. This is partly because of differences in the way the algorithms select their first seed or centroid either randomly, sequentially or some other principles influences which tend to influence the final result outcome. Popular external cluster quality validation and comparison models require the computation of varying cluster-

Chemoinformatics and Natural Products(1)

- Cheminformatics is a field that gives a better understanding of biomolecules (1).

- There has been a substantial increase in both the quality and use of chemoinformatic tools to empower drug research (2).

- Natural products play a significant role in drug discovery and development (3).
- The use of natural products especially plant-derived compounds have brought about an improvement in the treatment of various ailments among humans.
- The current-day pharmaceutical agents are mostly made from natural products, for example, Taxol, camptothecin, and vinblas (4).

- Medicinal plants are a source of novel chemical entities that possess beneficial pharmacological and therapeutics properties (5).

Chemoinformatics and Natural Products (2)

- Though only a fraction of these plants has been analyzed and investigated for their therapeutic potential (6).

- Nature continues to be an abundant source of biologically active and diverse chemotypes, and while relatively few of the actual isolated natural products are developed into clinically effective drugs in their own right (4).

- The cost of chemical trials is high, there is hence a need to pre-sort candidates using computational chemistry and cheminformatics methods.

Clustering and Similarity Measure

- Based on the similarity property principle, drugs with similar molecular structures are likely to have the same properties. This implies that we can identify a novel drug based on its similarity with a known one.

- Clustering is an approach that can be used to identify common characteristics shared by a group of compounds.

Similarity Metric	Mathematical Definition
Tanimoto (Jaccard) coefficient	$S_{AB} = \frac{C}{A + B - C}$
Dice coefficient (Hodgkin index)	$S_{AB} = \frac{2C}{A + B}$
Cosine coefficient (Carbo index)	$S_{AB} = \frac{C}{\sqrt{ab}}$
Soergel distance	$D_{AB} = \frac{a + b - 2c}{a + b - c}$
Euclidean distance	$D_{AB} = \sqrt{a + b - 2c}$

MATERIALS AND METHODS

Materials and Methods

- Software:
  - There are about 25 packages supplied with R (called “standard” and “recommended” packages) and many more are available through the CRAN family of Internet sites (via https://CRAN.R-project.org) and elsewhere [8].
  - Rstudio
  - ChemmineR toolkit is a cheminformatics package for analyzing drug-like small molecule data in R (9).
  - OsamorSoft [10]

- Dataset:
  - The molecules used in this study were gotten from the African Natural Products Database (ANPDB). ANPDB is a database of Natural Products (NPs) that merges NP databases from different African regions [11].
  - A total of 99 molecules from the family of Acanthaceae and Apocyanaceae sourced from NANPDB [12-13] were used and compared with five popular anticancer drugs namely CMP1 (representing Topotecan), CMP2 (representing Docetaxel), CMP3 (representing Irinotecan), CMP4 (representing Camptothecin), CMP5 (representing Cyclophosphamide)

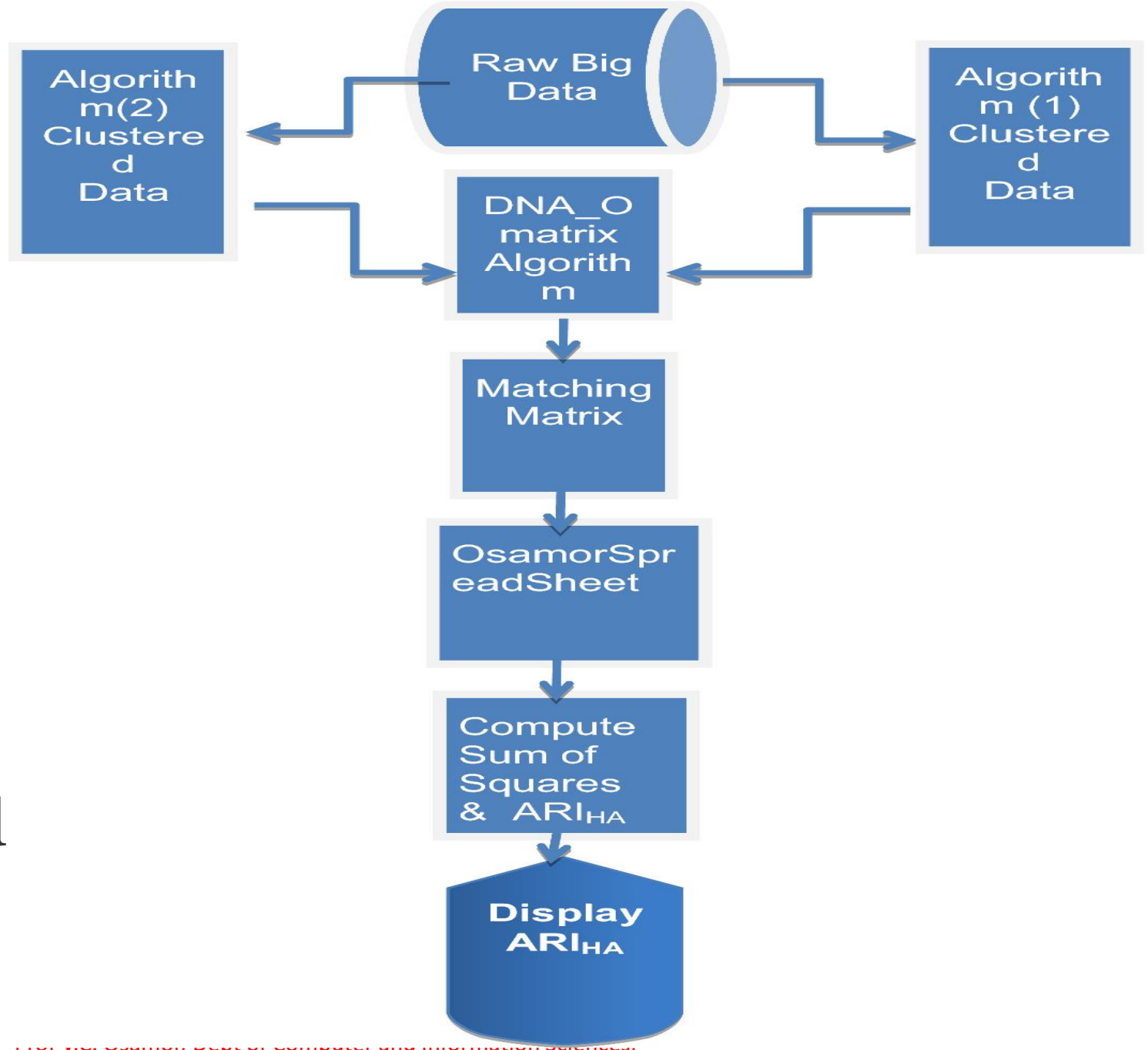

Fig 1: Osamorsoft pipeline allows raw big data access into both Algorithm 1 and 2 (10)

Similarity measurement for molecular structure

- Similarity metrics are used to evaluate how similar two molecules are to each other while dissimilarity metrics are used to test how dissimilar two molecules are to each other (7).

- Many similarity metrics have been proposed and some commonly used metrics in cheminformatics are listed.

To compute the molecular similarities	Atompair fingerprint
Similarity measure	
Clustering of molecules	Kmeans and Hierarchical techniques
Visualization of results	
To Investigate the level of cluster quality	using our newly developed cluster validation tool called OsamorSoft (10).

Table 2: Materials and Methods

RESULTS

Results- Choosing number of clusters

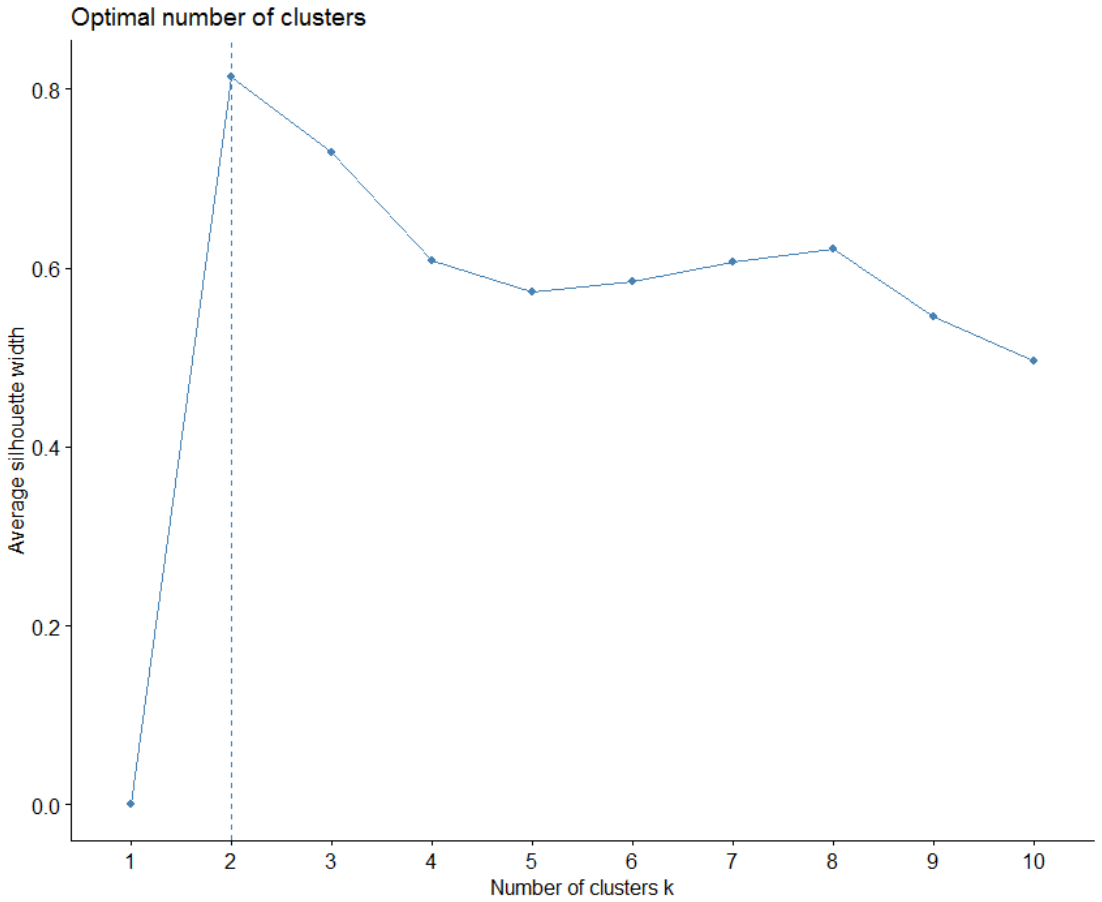

Figure 1: Optimum k

Results- Visualization of molecules

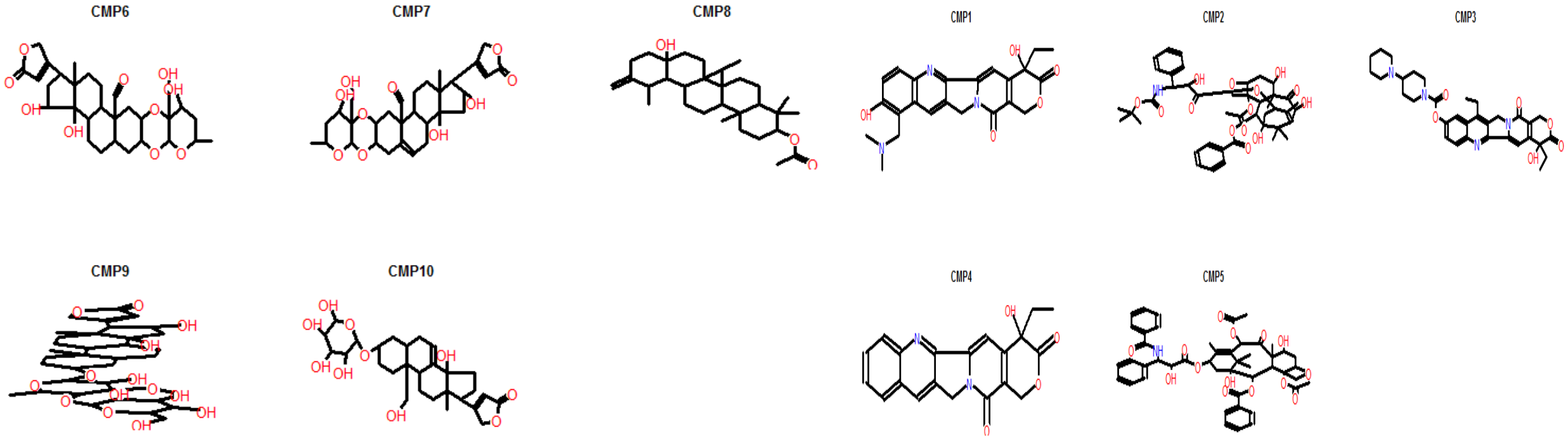

Figure 2: Visualization of compounds

Results- K-means clustering

```
K-means clustering with 2 clusters of sizes 2, 20

Cluster means:
        [,1]
1 0.1451613
2 0.5143972

Clustering vector:
  CMP1   CMP4   CMP3  CMP47  CMP65  CMP66  CMP69  CMP68  CMP95  CMP57  CMP46  CMP90
     1      1      2      2      2      2      2      2      2      2      2      2
 CMP67 CMP100  CMP91  CMP25  CMP49  CMP50  CMP51  CMP88  CMP48  CMP52
     2      2      2      2      2      2      2      2      2      2

within cluster sum of squares by cluster:
[1] 0.04214360 0.02605894
 (between_SS / total_SS =  78.4 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
> |
```

Figure 3: Kmeans result

Results- Hierarchical clustering

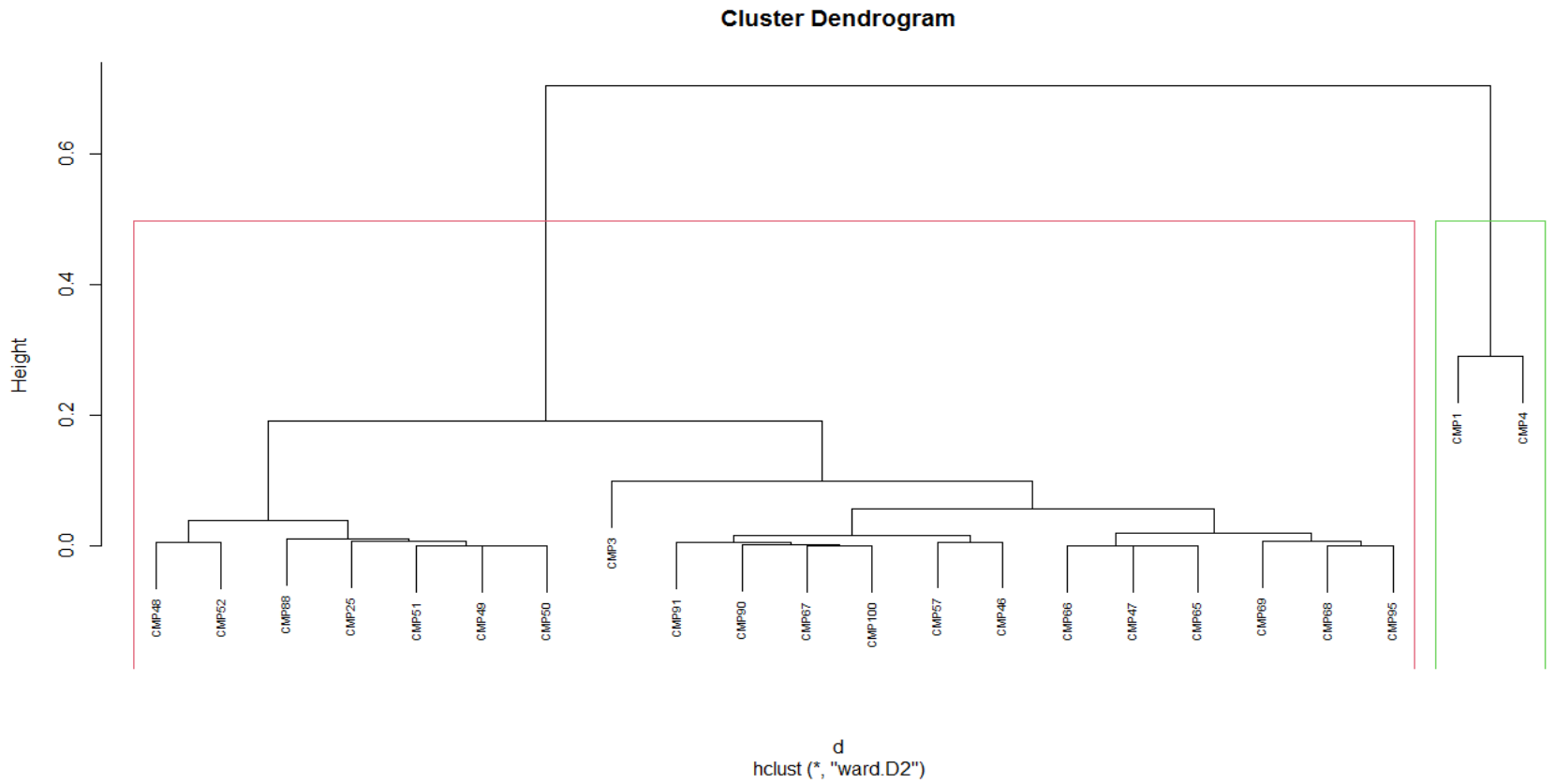

Figure 4: Hierarchical clustering- dendogram

Results

- Based on the calculation of the molecular fingerprints using the Tanimoto method as a similarity measure with a cutoff of 0.4, we were able to reduce the molecules to 22.

- Irinotecan was grouped with 19 natural molecules while Topotecan and Camptothecin were put in the same clusters

Conclusion

- This study aimed to investigate the potential of the studied natural compounds as possible lead molecules.
- As the 99 molecules were reduced to 22 due to cut-of of 0.4, Irinotecan was grouped with 19 natural molecules while Topotecan and Camptothecin were put in the same clusters.
- From the available preliminary result obtained in addition to the similarly property principle, molecules with similar structures are likely to have the same properties.
- This likely portends that they have similar properties and suggests that further test as a potential drug candidate may be necessary.
- The enhancement of OsamorSoft to maximally evaluate the cluster quality of Natural Products is desirable.

References

- [1] Begam F. B, Kumar S.J. A study on cheminformatics and its applications on modern drug discovery. Procedia Eng [Internet]. 2012;38(Icmoc):1264–75. Available from: http://dx.doi.org/10.1016/j.proeng.2012.06.156
- [2] Shinde RS, Deshmukh A. Cheminformatics tools useful for Research Scholar , Research Supervisor , Research and. 2018;5(December):153–6.
- [3] Banerjee P, Erehman J, Gohlke BO, Wilhelm T, Preissner R, Dunkel M. Super Natural II-a database of natural products. Nucleic Acids Res. 2015;43(D1):D935–9.
- [4] Cragg GM, Pezzuto M. Natural Products as a Vital Source for the Discovery of Cancer Chemotherapeutic and Chemopreventive Agents. 2016;25(suppl 2):41–59.
- [5] Ibrahim Khan, Jan SA, Shinwari ZK, Ali M, Khan Y, Kumar T. Ethnobotany and Medicinal Uses of Folklore Medicinal Plants Belonging to Family Acanthaceae: An Updated Review. MOJ Biol Med. 2017;1(2):34–8.
- [6] Rahman HS. Natural Products for Cancer Therapy. iMedPub Journals. 2016;1(2:15).
- [7] Bero SA, Muda AK, Choo YH, Muda NA, Pratama SF. Similarity Measure for Molecular Structure: A Brief Review. J Phys Conf Ser. 2017;892(1).
- [8] Venables WN, Smith DM, Team TRC. An Introduction to R. In: A Programming Environment for Data Analysis and Graphics. 2020.
- [9] Cao Y, Backman T, Horan K, Girke T. ChemmineR : Cheminformatics Toolkit for R. 2021. p. 1–46.
- [10] Osamor IP, Osamor VC. OsamorSoft: clustering index for comparison and quality validation in high throughput dataset. J Big Data [Internet]. 2020;7(1). Available from: https://doi.org/10.1186/s40537-020-00325-6
- [11] Ntie-Kang F, Zofou D, Babiaka SB, Meudom R, Scharfe M, Lifongo LL, et al. AfroDb: a select highly potent and diverse natural product library from African medicinal plants. PLoS One. 2013;8(10):1–15.
- [12] Simoben C.V. , et al. *Pharmacoinformatic investigation of medicinal plants from East Africa, Molecular Informatics* **2020**. DOI: 10.1002/minf.202000163.
- [13] Ntie- Kang F. et al. NANPDB: A Resource for Natural Products from Northern African Sources

  Journal of Natural Products, **2017.** DOI: 10.1021/acs.jnatprod.7b00283