Implementasi *Hierarchical Clustering* dalam *DNA Barcoding* untuk Menentukan Taksonomi Tumbuhan

SKRIPSI

Diajukan untuk Memenuhi sebagian dari
Syarat Memperoleh Gelar Sarjana Komputer
Program Studi Ilmu Komputer

Oleh

Muhammad Iqbal Zain

1901423

PROGRAM STUDI ILMU KOMPUTER
FAKULTAS PENDIDIKAN MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS PENDIDIKAN INDONESIA
BANDUNG

2023

Implementasi *Hierarchical Clustering* dalam *DNA Barcoding* untuk Menentukan Taksonomi Tumbuhan

Oleh

Muhammad Iqbal Zain

NIM 1901423

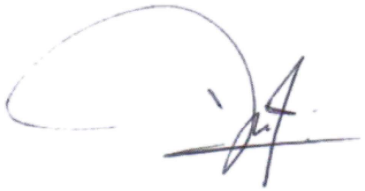
Sebuah Skripsi yang Diajukan untuk Memenuhi Salah Satu Syarat Memperoleh Gelar Sarjana Komputer di Fakultas Pendidikan Matematika dan Ilmu Pengetahuan Alam

MUHAMMAD IQBAL ZAIN

1901423

Implementasi *Hierarchical Clustering* dalam *DNA Barcoding* untuk Menentukan Taksonomi Tumbuhan

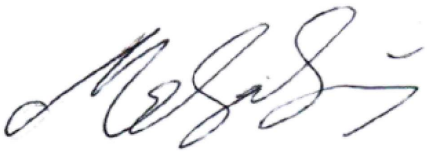
DISETUJUI DAN DISAHKAN OLEH PEMBIMBING:
Pembimbing I,



Prof. Dr. Lala Septem Riza, M.T.

NIP. 197809262008121001

Pembimbing II,



Dr. Rani Megasari, M.T.

NIP. 198705242014042002

Mengetahui,

Ketua Program Studi Ilmu Komputer,

Dr.Muhamad Nursalman, S.Si, M.T.

NIP. 197909292006041002

PERNYATAAN

Dengan ini penulis menyatakan bahwa skripsi dengan judul "Implementasi *Hierarchical Clustering* dalam *DNA Barcoding* untuk Menentukan Taksonomi Tumbuhan" ini beserta seluruh isinya adalah benar-benar karya penulis sendiri. Penulis tidak melakukan penjiplakan atau pengutipan dengan cara-cara yang tidak sesuai dengan etika ilmu yang berlaku dalam masyarakat keilmuan. Atas pernyataan ini, penulis siap menanggung risiko/sanksi apabila di kemudian hari ditemukan adanya pelanggaran etika keilmuan atau ada klaim dari pihak lain terhadap keaslian karya penulis ini.

Bandung, Agustus 2023

Yang Membuat Pernyataan,

Muhammad Iqbal Zain

NIM 1901423

Implementasi *Hierarchical Clustering* dalam *DNA Barcoding* untuk Menentukan Taksonomi Tumbuhan

Oleh

Muhammad Iqbal Zain —- iqbalzain99@gmail.com

1901423

ABSTRAK

Pendekatan *DNA Barcoding* telah digunakan secara luas dalam taksonomi dan filogenetik. Perbedaan dalam urutan DNA tertentu dapat membedakan dan membantu mengklasifikasikan organisme ke dalam taksa. Hal ini telah digunakan dalam kasus-kasus perselisihan taksonomi dimana pendekatan morfologi saja tidak cukup. Penelitian ini bertujuan untuk memanfaatkan *hierarchical clustering*, sebuah metode *unsupervised machine learning*, untuk menentukan dan menyelesaikan sengketa dalam taksonomi famili tumbuhan. Studi kasus *Leguminosae* secara historis ada yang mengklasifikasikan ke dalam tiga famili (*Fabaceae, Caesalpiniaceae, dan Mimosaceae*) tetapi ada juga yang mengklasifikasikan ke dalam satu famili (*Leguminosae*). Penelitian ini dibagi menjadi beberapa tahap, yaitu: (i) *data collection*, (ii) *data preprocessing*, (iii) *finding the best distance method*, and (iv) *determining disputed family*. Data yang digunakan dalam penelitian ini dikumpulkan dari beberapa sumber, termasuk *National Center for Biotechnology Information* (NCBI), jurnal, dan website. Data untuk validasi metode dikumpulkan dari NCBI dan digunakan untuk menentukan *distance method* terbaik untuk membedakan famili atau genera. Data untuk studi kasus pada kelompok *Leguminosae* dikumpulkan dari Berbagai jurnal dan *website*. Percobaan bertujuan untuk mendapatkan *distance method* terbaik yang kemudian digunakan untuk menentukan famili yang disengketakan. Ditemukan bahwa studi kasus *Leguminosae* harus dikelompokkan ke dalam satu famili berdasarkan penelitian ini.

Kata Kunci: *DNA Barcoding, Unsupervised Learning,* Bioinformatika*, Hierarchical Clustering, Machine Learning, Taxonomy, R Programming Language*

Hierarchical Clustering Implementation in DNA Barcoding for Determining Plant Taxonomy

Arranged by

Muhammad Iqbal Zain —- iqbalzain99@gmail.com

1901423

ABSTRACT

The DNA barcoding approach has been used extensively in taxonomy and phylogenetics. The differences in certain DNA sequences are able to differentiate and help classify organisms into taxa. This tool has been used in cases of taxonomic disputes where morphology by itself is insufficient. This research aimed to utilize hierarchical clustering, an unsupervised machine learning method, to determine and resolve disputes in plant family taxonomy. We take a case study of Leguminosae that historically some classify into three families (Fabaceae, Caesalpiniaceae, and Mimosaceae) but others classify into one family (Leguminosae). This study is divided into several phases, which are: (i) data collection, (ii) data preprocessing, (iii) finding the best distance method, and (iv) determining disputed family. The data used in this study are collected from several sources, including National Center for Biotechnology Information (NCBI), journals, and websites. The data for validation of the methods were collected from NCBI and used to determine the best distance method for differentiating families or genera. The data for the case study in the Leguminosae group was collected from journals and a website. The experiment aimed to identify the best distance method, which was then used to determine the disputed family. It was found that the Leguminosae case study should be grouped into one family based on this research.

Keywords: DNA Barcoding, Unsupervised Learning, Bioinformatics, Hierarchical Clustering, Machine Learning, Taxonomy, R Programming Language

KATA PENGANTAR

Puji dan syukur penulis panjatkan kehadirat Allah swt. karena hanya dengan kehendak, berkat, serta karunia-Nya lah penulis dapat menyelesaikan skripsi yang berjudul "Implementasi *Hierarchical Clustering* dalam *DNA Barcoding* untuk Menentukan Taksonomi Tumbuhan" ini dapat terselesaikan.

Penyusunan skripsi ini diajukan untuk memenuhi dan melengkapi salah satu syarat untuk penyusunan skripsi yang merupakan syarat untuk mendapatkan gelar sarjana komputer atas jenjang studi S1 pada Program Studi Ilmu Komputer Fakultas Pendidikan Matematika dan Ilmu Pengetahuan Alam Universitas Pendidikan Indonesia.

Penulis menyadari bahwa dalam penyusunan skripsi ini masih terdapat banyak kekurangan dan keterbatasan yang perlu disempurnakan. Oleh karena itu, penulis sangat mengharapkan saran maupun kritik yang membangun agar tidak terjadi kesalahan yang sama di kemudian hari dan dapat meningkatkan kualitas ke tahap lebih baik.

Bandung, Agustus 2023

Penyusun

UCAPAN TERIMAKASIH

Alhamdulillahirobbilalamin, puji dan syukur kehadirat Allah SWT Yang telah melimpahkan rahmat dan hidayah-Nya sehingga penulis diberikan kelancaran dalam menyelesaikan penulisan skripsi ini. Dalam proses menyelesaikan penelitian dan penyusunan skripsi ini, peneliti banyak mendapat bimbingan, dorongan, serta bantuan dari berbagai pihak. Oleh karena itu, pada kesempatan ini penulis mengucapkan terimakasih serta penghargaan yang setinggi-tingginya, kepada:

1. Diri saya sendiri hingga mencapai titik ini atas segala usaha dan tenaga yang dicurahkan dan terus melakukan yang terbaik kepada diri sendiri.
2. Kedua orang tua yaitu Zainun Mu'tadin dan Seftiana yang selalu memberi dukungan, motivasi, serta doa serta selalu menjadi penyemangat utama dalam menempuh pendidikan tinggi sehingga penulis dapat menyelesaikan skripsi ini.
3. Bapak Prof. Dr. Lala Septem Riza, M.T. selaku pembimbing I atas segala waktu dan tenaga yang dicurahkan untuk membimbing penulis demi terselesaikannya skripsi ini.
4. Ibu Dr. Rani Megasari, M.T. selaku selaku pembimbing II atas saran dan masukan yang diberikan kepada penulis selama proses menyelesaikan penelitian dan penulisan skripsi.
5. Bapak Dr.Muhamad Nursalman, S.Si, M.T. selaku Ketua Program Studi Ilmu Komputer.
6. Bapak Yudi Ahmad Hambali, M.T. selaku Dosen Wali saya.
7. Prof. Topik Hidayat, M.Si., Ph.D. yang telah bekerja sama dalam proses pembuatan jurnal ini dan bimbingan yang telah diberikan.
8. Ibu Rosa Ariani Sukamto, M.T. selaku dosen Program Studi Ilmu Komputer yang sudah banyak memberikan saya kesempatan untuk belajar dan

mengembangkan diri sebagai asisten dosen pada mata kuliah Algoritma dan Pemrograman, Struktur Data, dan Pemrograman Berorientasi Objek.

9. Bapak dan Ibu Dosen Prodi Pendidikan Ilmu Komputer dan Ilmu Komputer yang telah berbagi ilmu yang bermanfaat kepada penulis dan mahasiswa lainnya.
10. Tim penelitian pak Lala yaitu Alif, Yudi, Zulfikar, M Fajar, Fajar Z, Izzudin, Ira, Melvyn, Naufal, Pak Aria, dan Arfiansyah yang senantiasa memberikan dukungan, semangat, dan kegiatan positif lainnya selama proses penelitian dan pengerjaan skripsi.
11. Teman teman dan rekan dari BEM KEMAKOM yang telah memberi pengalaman berharga.
12. Orang orang lainnya yang tidak dapat disebutkan satu persatu.

Semoga semua amal baik yang diberikan kepada penulis mendapatkan balasan yang berlipat ganda dari Allah SWT. Aamiin.

Bandung, Agustus 2023

Muhammad Iqbal Zain

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR PUSTAKA

Abdel-Basset, M., Mohamed, R., Sallam, K. M., Chakrabortty, R. K., & Ryan, M. J. (2020). An Efficient-Assembler Whale Optimization Algorithm for DNA Fragment Assembly Problem: Analysis and Validations. *IEEE Access*, *8*, 222144–222167. https://doi.org/10.1109/ACCESS.2020.3044857

An, Q., Chen, J., Tan, G., Ren, Y., Zhou, J., Liao, H., & Tan, R. (2022). Predicting medicinal resources in Ranunculaceae family by a combined approach using DNA barcodes and chemical metabolites. *Phytochemistry Letters*, *50*, 67–76. https://doi.org/10.1016/j.phytol.2022.04.009

Ani Brown Mary, N., & Dharma, D. (2017). Coral reef image classification employing Improved LDP for feature extraction. *Journal of Visual Communication and Image Representation*, *49*, 225–242. https://doi.org/10.1016/j.jvcir.2017.09.008

Avise, J. C., & Liu, J.-X. (2011). On the temporal inconsistencies of Linnean taxonomic ranks. *Biological Journal of the Linnean Society*, *102*(4), 707–714. https://doi.org/10.1111/j.1095-8312.2011.01624.x

Bates, S. A. (2022, September 14). *Deoxyribonucleic Acid (DNA)*. Genome.Gov. https://www.genome.gov/genetics-glossary/Deoxyribonucleic-Acid

Bentham, G. (1842). Notes on Mimoseae, with a synopsis of species. *London Journal of Botany*, *1*, 318–528.

Berman, J. J. (2016). Understanding Your Data. In *Data Simplification* (pp. 135–187). Elsevier. https://doi.org/10.1016/B978-0-12-803781-2.00004-7

Borges, L., Bruneau, A., Cardoso, D., Crisp, M., Delgado-Salinas, A., Doyle, J. J., Egan, A., Herendeen, P. S., Hughes, C., Kenicer, G., Klitgaard, B., Koenen, E., Lavin, M., Lewis, G., Luckow, M., Mackinder, B., Malécot, V., Miller, J. T., Pennington, R. T., … Wink, M. (2013). Towards a new classification system for legumes: Progress report from the 6th International Legume Conference. *South African Journal of Botany*, *89*, 3–9. https://doi.org/10.1016/j.sajb.2013.07.022

Chambers, J. M. (2014). Object-Oriented Programming, Functional Programming and R. *Statistical Science*, *29*(2). https://doi.org/10.1214/13-STS452

China Plant BOL Group, Li, D.-Z., Gao, L.-M., Li, H.-T., Wang, H., Ge, X.-J., Liu, J.-Q., Chen, Z.-D., Zhou, S.-L., Chen, S.-L., Yang, J.-B., Fu, C.-X., Zeng, C.-X., Yan, H.-F., Zhu, Y.-J., Sun, Y.-S., Chen, S.-Y., Zhao, L., Wang, K., … Duan, G.-W. (2011). Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. *Proceedings of the National Academy of Sciences*, *108*(49), 19641–19646. https://doi.org/10.1073/pnas.1104551108

Cronquist, A. (1981). *An integrated system of classification of flowering plants*. Columbia University Press.

Dahlgren, R. (1983). General aspects of angiosperm evolution and macrosystematics. *Nordic Journal of Botany*, *3*(1), 119–149. https://doi.org/10.1111/j.1756-1051.1983.tb01448.x

Das, S. (2012). Domestication, phylogeny and taxonomic delimitation in underutilized grain *Amaranthus* (Amaranthaceae) – a status review. *Feddes Repertorium*, *123*(4), 273–282. https://doi.org/10.1002/fedr.201200017

Deoxyribonucleic Acid (DNA) Fact Sheet. (2022, September 14). Genome.Gov. https://www.genome.gov/about-genomics/fact-sheets/Deoxyribonucleic-Acid-Fact-Sheet

DeSalle, R., & Goldstein, P. (2019). Review and Interpretation of Trends in DNA Barcoding. *Frontiers in Ecology and Evolution*, *7*, 302. https://doi.org/10.3389/fevo.2019.00302

Dhakad, A. K. (2018). *Molecular phylogeny of selected tree species of families Fabaceae Caesalpiniaceae and Mimosaceae of Uttarakhand* [Forest Research Institute University]. http://hdl.handle.net/10603/203120

Docker, M. F. (Ed.). (2015). *Lampreys: Biology, Conservation and Control: Volume 1* (1st ed. 2015). Springer Netherlands : Imprint: Springer. https://doi.org/10.1007/978-94-017-9306-3

Doyle, J. J., & Luckow, M. A. (2003). The Rest of the Iceberg. Legume Diversity and Evolution in a Phylogenetic Context. *Plant Physiology*, *131*(3), 900–910. https://doi.org/10.1104/pp.102.018150

Edgar, R. C., & Batzoglou, S. (2006). Multiple sequence alignment. *Current Opinion in Structural Biology*, *16*(3), 368–373. https://doi.org/10.1016/j.sbi.2006.04.004

Enrico Bonatesta, C. H.-K. (2017). *Msa* [Computer software]. Bioconductor. https://doi.org/10.18129/B9.BIOC.MSA

Faisal, M., Zamzami, E. M., & Sutarman. (2020). Comparative Analysis of Inter-Centroid K-Means Performance using Euclidean Distance, Canberra Distance and Manhattan Distance. *Journal of Physics: Conference Series*, *1566*(1), 012112. https://doi.org/10.1088/1742-6596/1566/1/012112

Fazekas, A. J., Burgess, K. S., Kesanakurti, P. R., Graham, S. W., Newmaster, S. G., Husband, B. C., Percy, D. M., Hajibabaei, M., & Barrett, S. C. H. (2008). Multiple Multilocus DNA Barcodes from the Plastid Genome Discriminate Plant Species Equally Well. *PLoS ONE*, *3*(7), e2802. https://doi.org/10.1371/journal.pone.0002802

Fazekas, A. J., Kesanakurti, P. R., Burgess, K. S., Percy, D. M., Graham, S. W., Barrett, S. C. H., Newmaster, S. G., Hajibabaei, M., & Husband, B. C. (2009). Are plant species inherently harder to discriminate than animal species using DNA barcoding markers? *Molecular Ecology Resources*, *9*, 130–139. https://doi.org/10.1111/j.1755-0998.2009.02652.x

Federhen, S. (2012). The NCBI Taxonomy database. *Nucleic Acids Research*, *40*(D1), D136–D143. https://doi.org/10.1093/nar/gkr1178

Fišer Pečnikar, Ž., & Buzan, E. V. (2014). 20 years since the introduction of DNA barcoding: From theory to application. *Journal of Applied Genetics*, *55*(1), 43–52. https://doi.org/10.1007/s13353-013-0180-y

Gielly, L., & Taberlet, P. (1994). The use of chloroplast DNA to resolve plant phylogenies: Noncoding versus rbcL sequences. *Molecular Biology and Evolution*, *11*(5), 769–777. https://doi.org/10.1093/oxfordjournals.molbev.a040157

Gonçalves, P. F. M., Oliveira-Marques, A. R., Matsumoto, T. E., & Miyaki, C. Y. (2015). DNA Barcoding Identifies Illegal Parrot Trade. *Journal of Heredity*, *106*(S1), 560–564. https://doi.org/10.1093/jhered/esv035

Greuter, W., McNeill, J., Barrie, F. R., Burdet, H. M., Demoulin, V., Filgueiras, T. S., Nicolson, D. H., Silva, P. C., Skog, J. E., Trehane, P., Turland, N. J., &

Hawksworth, D. L. (Eds.). (2000). *International code of botanical nomenclature (Saint Louis code): Adopted by the sixteenth International botanical congress, St Louis, Missouri, july-august 1999*. Koeltz scientific books.

H. Pagès, P. A. (2017). *Biostrings* [Computer software]. Bioconductor. https://doi.org/10.18129/B9.BIOC.BIOSTRINGS

Hahsler, M., Piekenbrock, M., & Doran, D. (2019). **dbscan**: Fast Density-Based Clustering with *R*. *Journal of Statistical Software*, *91*(1). https://doi.org/10.18637/jss.v091.i01

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Applied Statistics*, *28*(1), 100. https://doi.org/10.2307/2346830

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Vol. 2). Springer.

He, T., Jiao, L., Wiedenhoeft, A. C., & Yin, Y. (2019). Machine learning approaches outperform distance- and tree-based methods for DNA barcoding of Pterocarpus wood. *Planta*, *249*(5), 1617–1625. https://doi.org/10.1007/s00425-019-03116-3

Hebert, P. D. N., Cywinska, A., Ball, S. L., & deWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *270*(1512), 313–321. https://doi.org/10.1098/rspb.2002.2218

Hebert, P. D. N., Ratnasingham, S., & De Waard, J. R. (2003). Barcoding animal life: Cytochrome *c* oxidase subunit 1 divergences among closely related

species. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *270*(suppl_1). https://doi.org/10.1098/rsbl.2003.0025

Heibl, C. (2008). *PHYLOCH: R language tree plotting tools and interfaces to diverse phylogenetic software packages.* [Computer software]. http://www.christophheibl.de/Rpackages.html

Hilu, K. W., & Liang, gping. (1997). The matK gene: Sequence variation and application in plant systematics. *American Journal of Botany*, *84*(6), 830–839. https://doi.org/10.2307/2445819

Hou, D., Larsen, K., & Larsen, S. S. (1996). Caesalpiniaceae (Leguminosae-Caesalpinioideae). *Flora Malesiana*, *12*(2), 409–730.

Hsuan, K. (1983). *Orders and families of Malayan seed plants*. Singapore University Press.

Huang, K.-W., Chen, J.-L., Yang, C.-S., & Tsai, C.-W. (2015). A memetic particle swarm optimization algorithm for solving the DNA fragment assembly problem. *Neural Computing and Applications*, *26*(3), 495–506. https://doi.org/10.1007/s00521-014-1659-0

Ji, Y., Liu, C., Yang, J., Jin, L., Yang, Z., & Yang, J.-B. (2020). Ultra-Barcoding Discovers a Cryptic Species in Paris yunnanensis (Melanthiaceae), a Medicinally Important Plant. *Frontiers in Plant Science*, *11*, 411. https://doi.org/10.3389/fpls.2020.00411

Jiao, L., Lu, Y., He, T., Guo, J., & Yin, Y. (2020). DNA barcoding for wood identification: Global review of the last decade and future perspective. *IAWA Journal*, *41*(4), 620–643. https://doi.org/10.1163/22941932-bja10041

Kassambara, A., & Mundt, F. (2020). *factoextra: Extract and Visualize the Results of Multivariate Data Analyses* (1.0.7) [Computer software]. https://CRAN.R-project.org/package=factoextra

Kotu, V., & Deshpande, B. (2019). Classification. In *Data Science* (pp. 65–163). Elsevier. https://doi.org/10.1016/B978-0-12-814761-0.00004-6

Kumar, S., & Filipski, A. (2007). Multiple sequence alignment: In pursuit of homologous DNA positions. *Genome Research*, *17*(2), 127–135. https://doi.org/10.1101/gr.5232407

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lee, P., Yang, S. T., West, J. D., & Howe, B. (2017). PhyloParser: A Hybrid Algorithm for Extracting Phylogenies from Dendrograms. *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 1087–1094. https://doi.org/10.1109/ICDAR.2017.180

Lewis, G. P. (Ed.). (2005). *Legumes of the world*. Royal Botanic Gardens, Kew.

Little, D. P., & Stevenson, D. Wm. (2007). A comparison of algorithms for the identification of specimens using DNA barcodes: Examples from gymnosperms. *Cladistics*, *23*(1), 1–21. https://doi.org/10.1111/j.1096-0031.2006.00126.x

Lovie, A. D. (1995). Who discovered Spearman's rank correlation? *British Journal of Mathematical and Statistical Psychology*, *48*(2), 255–269.

Lucas, C., Thangaradjou, T., & Papenbrock, J. (2012). Development of a DNA Barcoding System for Seagrasses: Successful but Not Simple. *PLoS ONE*, *7*(1), e29987. https://doi.org/10.1371/journal.pone.0029987

Lukhtanov, V. A., Sourakov, A., & Zakharov, E. V. (2016). DNA barcodes as a tool in biodiversity research: Testing pre-existing taxonomic hypotheses in Delphic Apollo butterflies (Lepidoptera, Papilionidae). *Systematics and Biodiversity*, *14*(6), 599–613. https://doi.org/10.1080/14772000.2016.1203371

Madden, M. J. L., Young, R. G., Brown, J. W., Miller, S. E., Frewin, A. J., & Hanner, R. H. (2019). Using DNA barcoding to improve invasive pest identification at U.S. ports-of-entry. *PLOS ONE*, *14*(9), e0222291. https://doi.org/10.1371/journal.pone.0222291

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, *9*(1), 381–386.

Malkauthekar, M. D. (2013). Analysis of euclidean distance and manhattan distance measure in face recognition. *Third International Conference on Computational Intelligence and Information Technology (CIIT 2013)*, 503–507. https://doi.org/10.1049/cp.2013.2636

Maxim, L. G., Rodriguez, J. I., & Wang, B. (2019). *Defect of Euclidean distance degree* (arXiv:1905.06758). arXiv. http://arxiv.org/abs/1905.06758

Meher, P. K., Sahu, T. K., & Rao, A. R. (2016). Identification of species based on DNA barcode using k-mer feature vector and Random forest classifier. *Gene*, *592*(2), 316–324. https://doi.org/10.1016/j.gene.2016.07.010

Merigó, J. M., & Casanovas, M. (2011). A New Minkowski Distance Based on Induced Aggregation Operators. *International Journal of Computational Intelligence Systems*, *4*(2), 123–133. https://doi.org/10.1080/18756891.2011.9727769

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.

Mondal, A. K., & Mondal, S. (2011). Circumscription of the families within Leguminales as determined by cladistic analysis based on seed protein. *African Journal of Biotechnology*, *10*(15), 2850–2856. https://doi.org/10.5897/AJB10.206

Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: An overview. *WIREs Data Mining and Knowledge Discovery*, *2*(1), 86–97. https://doi.org/10.1002/widm.53

Murtagh, F., & Contreras, P. (2017). Algorithms for hierarchical clustering: An overview, II. *WIREs Data Mining and Knowledge Discovery*, *7*(6). https://doi.org/10.1002/widm.1219

Nielsen, F. (2016). *Introduction to HPC with MPI for Data Science*. Springer International Publishing. https://doi.org/10.1007/978-3-319-21903-5

Nielsen, I. C. (1992). Mimosaceae (Leguminosae-Mimosoideae). *Flora Malesiana*, *11*(1), 1–226.

Nikitina, E. V., Karimov, F. I., Savina, N. V., Kubrak, S. V., & Kilchevsky, A. V. (2021). Inventory of some Tulipa species from Uzbekistan using DNA barcoding. *BIO Web of Conferences*, *38*, 00086. https://doi.org/10.1051/bioconf/20213800086

Nikitina, E. V., Yu. Beshko, N., & Omarov, S. A. (2022). Assessment of plant species diversity (Lamiaceae Lindle.) in Uzbekistan based on DNA barcoding. *IOP Conference Series: Earth and Environmental Science*, *1068*(1), 012042. https://doi.org/10.1088/1755-1315/1068/1/012042

Nishimaki, T., & Sato, K. (2019). An Extension of the Kimura Two-Parameter Model to the Natural Evolutionary Process. *Journal of Molecular Evolution*, *87*(1), 60–67. https://doi.org/10.1007/s00239-018-9885-1

Olson, C. F. (1995). Parallel algorithms for hierarchical clustering. *Parallel Computing*, *21*(8), 1313–1325.

Ouyang, M., Welsh, W. J., & Georgopoulos, P. (2004). Gaussian mixture clustering and imputation of microarray data. *Bioinformatics*, *20*(6), 917–923. https://doi.org/10.1093/bioinformatics/bth007

Papa, Y., Le Bail, P., & Covain, R. (2021). Genetic landscape clustering of a large DNA barcoding data set reveals shared patterns of genetic divergence among freshwater fishes of the Maroni Basin. *Molecular Ecology Resources*, *21*(6), 2109–2124. https://doi.org/10.1111/1755-0998.13402

Parker, C. T., Tindall, B. J., & Garrity, G. M. (2019). International Code of Nomenclature of Prokaryotes: Prokaryotic Code (2008 Revision). *International Journal of Systematic and Evolutionary Microbiology*, *69*(1A), S1–S111. https://doi.org/10.1099/ijsem.0.000778

Patel, S., & Panchal, H. (2014). Evolutionary studies of few species belonging to Leguminosae family based on RBCL gene. *Discovery*, *9*(22), 38–50.

Polhill, R. M., & Raven, P. H. (1981). *Advances in legume systematics*. Royal botanic gardens.

Purty, R., & Chatterjee, S. (2016). DNA barcoding: An effective technique in molecular taxonomy. *Austin J Biotechnol Bioeng*, *3*(1), 1059.

Ristaino, J. B. (2020). The Importance of Mycological and Plant Herbaria in Tracking Plant Killers. *Frontiers in Ecology and Evolution*, *7*. https://doi.org/10.3389/fevo.2019.00521

Rossini, B. C., Oliveira, C. A. M., Melo, F. A. G. de, Bertaco, V. de A., Astarloa, J. M. D. de, Rosso, J. J., Foresti, F., & Oliveira, C. (2016). Highlighting Astyanax Species Diversity through DNA Barcoding. *PLOS ONE*, *11*(12), e0167203. https://doi.org/10.1371/journal.pone.0167203

Saddhe, A. A., & Kumar, K. (2017). DNA barcoding of plants: Selection of core markers for taxonomic groups. *Plant Science Today*, *5*(1), 9–13. https://doi.org/10.14719/pst.2018.5.1.356

Sandionigi, A., Galimberti, A., Labra, M., Ferri, E., Panunzi, E., De Mattia, F., & Casiraghi, M. (2012). Analytical approaches for DNA barcoding data – how to find a way for plants? *Plant Biosystems - An International Journal Dealing with All Aspects of Plant Biology*, *146*(4), 805–813. https://doi.org/10.1080/11263504.2012.740084

Schoch, C. L., Ciufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'Neill, K., Robbertse, B., Sharma, S., Soussov, V., Sullivan, J. P., Sun, L., Turner, S., & Karsch-Mizrachi, I. (2020). NCBI Taxonomy: A comprehensive update on curation, resources and tools. *Database*, *2020*, baaa062. https://doi.org/10.1093/database/baaa062

Schoch, C. L., Seifert, K. A., Huhndorf, S., Robert, V., Spouge, J. L., Levesque, C. A., Chen, W., Fungal Barcoding Consortium, Fungal Barcoding Consortium Author List, Bolchacova, E., Voigt, K., Crous, P. W., Miller, A. N., Wingfield, M. J., Aime, M. C., An, K.-D., Bai, F.-Y., Barreto, R. W.,

Begerow, D., … Schindel, D. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for *Fungi*. *Proceedings of the National Academy of Sciences*, *109*(16), 6241–6246. https://doi.org/10.1073/pnas.1117018109

Sigwart, J. D., Sutton, M. D., & Bennett, K. D. (2018). How big is a genus? Towards a nomothetic systematics. *Zoological Journal of the Linnean Society*, *183*(2), 237–252. https://doi.org/10.1093/zoolinnean/zlx059

Solo, V. (2019). *Pearson Distance is not a Distance* (arXiv:1908.06029). arXiv. http://arxiv.org/abs/1908.06029

Sommerville, I. (2016). *Software engineering* (10. ed., global ed). Pearson.

Soueidan, H., & Nikolski, M. (2015). *Machine learning for metagenomics: Methods and tools*. https://doi.org/10.48550/ARXIV.1510.06621

Spearman, C. (1906). Footrule for measuring correlation. *British Journal of Psychology*, *2*(1), 89.

Spearman, C. (1961). *The proof and measurement of association between two things.*

Stevens, P. F. (2003). History of Taxonomy. In John Wiley & Sons, Ltd (Ed.), *ELS* (1st ed.). Wiley. https://doi.org/10.1038/npg.els.0003093

Stoeckle, M. (2003). Taxonomy, DNA, and the bar code of life. *BioScience*, *53*(9), 796–797.

Takhtajan, A. L. (1980). Outline of the classification of flowering plants (magnoliophyta). *The Botanical Review*, *46*(3), 225–359. https://doi.org/10.1007/BF02861558

Tänzler, R., Sagata, K., Surbakti, S., Balke, M., & Riedel, A. (2012). DNA Barcoding for Community Ecology—How to Tackle a Hyperdiverse, Mostly Undescribed Melanesian Fauna. *PLoS ONE*, *7*(1), e28832. https://doi.org/10.1371/journal.pone.0028832

Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, *22*(22), 4673–4680. https://doi.org/10.1093/nar/22.22.4673

Thulin, M. (1983). *Leguminosae of ethiopia*.

Tripathi, S., & Mondal, A. K. (2012). Taxonomic diversity in epidermal cells (stomata) of some selected Anthophyta under the order Leguminales (Caeselpiniaceae, Mimosaceae and Fabaceae) based on numerical analysis: A systematic approach. *International Journal of Science and Nature*, *3*(4), 778–798.

Usama, M., Qadir, J., Raza, A., Arif, H., Yau, K. A., Elkhatib, Y., Hussain, A., & Al-Fuqaha, A. (2019). Unsupervised Machine Learning for Networking: Techniques, Applications and Research Challenges. *IEEE Access*, *7*, 65579–65615. https://doi.org/10.1109/ACCESS.2019.2916648

Venables, W. N., Smith, D. M., & R Core Team. (2023). An Introduction to R. *CRAN (The Comprehensive R Archive Network)*, 99.

Walker, P. J., Siddell, S. G., Lefkowitz, E. J., Mushegian, A. R., Dempsey, D. M., Dutilh, B. E., Harrach, B., Harrison, R. L., Hendrickson, R. C., Junglen, S., Knowles, N. J., Kropinski, A. M., Krupovic, M., Kuhn, J. H., Nibert, M., Rubino, L., Sabanadzovic, S., Simmonds, P., Varsani, A., … Davison, A. J.

(2019). Changes to virus taxonomy and the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2019). *Archives of Virology*, *164*(9), 2417–2429. https://doi.org/10.1007/s00705-019-04306-w

Wardill, T. J., Graham, G. C., Zalucki, M., Palmer, W. A., Playford, J., & Scott, K. D. (2005). The importance of species identity in the biocontrol process: Identifying the subspecies of Acacia nilotica (Leguminosae: Mimosoideae) by genetic distance and the implications for biological control. *Journal of Biogeography*, *32*(12), 2145–2159. https://doi.org/10.1111/j.1365-2699.2005.01348.x

Watson, J. D., & Crick, F. H. C. (1953). THE STRUCTURE OF DNA. *Cold Spring Harbor Symposia on Quantitative Biology*, *18*(0), 123–131. https://doi.org/10.1101/SQB.1953.018.01.020

Weber, J. H., Schouhamer Immink, K. A., & Blackburn, S. R. (2016). Pearson Codes. *IEEE Transactions on Information Theory*, *62*(1), 131–135. https://doi.org/10.1109/TIT.2015.2490219

Weitschek, E., Fiscon, G., & Felici, G. (2014). Supervised DNA Barcodes species classification: Analysis, comparisons and results. *BioData Mining*, *7*(1), 4. https://doi.org/10.1186/1756-0381-7-4

Williams, B., Halloin, C., Löbel, W., Finklea, F., Lipke, E., Zweigerdt, R., & Cremaschi, S. (2020). Data-Driven Model Development for Cardiomyocyte Production Experimental Failure Prediction. In *Computer Aided Chemical Engineering* (Vol. 48, pp. 1639–1644). Elsevier. https://doi.org/10.1016/B978-0-12-823377-1.50274-3

Winter, D. J. (2017). *rentrez: An R package for the NCBI eUtils API* [Preprint]. PeerJ Preprints. https://doi.org/10.7287/peerj.preprints.3179v2

Winter, D., J. (2017). rentrez: An R package for the NCBI eUtils API. *The R Journal*, *9*(2), 520. https://doi.org/10.32614/RJ-2017-058

Wong, W. H., Tay, Y. C., Puniamoorthy, J., Balke, M., Cranston, P. S., & Meier, R. (2014). 'Direct PCR' optimization yields a rapid, cost-effective, nondestructive and efficient method for obtaining DNA barcodes without DNA extraction. *Molecular Ecology Resources*, *14*(6), 1271–1280. https://doi.org/10.1111/1755-0998.12275

Wyatt, G., & Cohen, S. (1952). A new pyrimidine base from bacteriophage nucleic acids. *Nature*, *170*(4338), 1072–1073.

Xie, Y., Wang, Y., Nallanathan, A., & Wang, L. (2016). An Improved K-Nearest-Neighbor Indoor Localization Method Based on Spearman Distance. *IEEE Signal Processing Letters*, *23*(3), 351–355. https://doi.org/10.1109/LSP.2016.2519607

Xu, H., Li, P., Ren, G., Wang, Y., Jiang, D., & Liu, C. (2021). Authentication of Three Source Spices of Arnebiae Radix Using DNA Barcoding and HPLC. *Frontiers in Pharmacology*, *12*, 677014. https://doi.org/10.3389/fphar.2021.677014

Yang, C.-H., Wu, K.-C., Chuang, L.-Y., & Chang, H.-W. (2022). DeepBarcoding: Deep Learning for Species Classification Using DNA Barcoding. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *19*(4), 2158–2165. https://doi.org/10.1109/TCBB.2021.3056570

Zar, J. H. (2014). Spearman Rank Correlation: Overview. In N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri, & J. L. Teugels (Eds.), *Wiley StatsRef: Statistics Reference Online* (1st ed.). Wiley. https://doi.org/10.1002/9781118445112.stat05964

Zhao, L., Yu, X., Shen, J., & Xu, X. (2018). Identification of three kinds of Plumeria flowers by DNA barcoding and HPLC specific chromatogram. *Journal of Pharmaceutical Analysis*, *8*(3), 176–180. https://doi.org/10.1016/j.jpha.2018.02.002