Hierarchical network analysis of co-occurring bioentities in literature

Heejung Yang1,2,4*, Namgil Lee2,3,4, Beomjun Park2, Jinyoung Park1, Jiho Lee1, Hyeon Seok Jang1 & Hojin Yoo2

Biomedical databases grow by more than a thousand new publications every day. The large volume of biomedical literature that is being published at an unprecedented rate hinders the discovery of relevant knowledge from keywords of interest to gather new insights and form hypotheses. A text-mining tool, PubTator, helps to automatically annotate bioentities, such as species, chemicals, genes, and diseases, from PubMed abstracts and full-text articles. However, the manual re-organization and analysis of bioentities is a non-trivial and highly time-consuming task. ChexMix was designed to extract the unique identifiers of bioentities from query results. Herein, ChexMix was used to construct a taxonomic tree with allied species among Korean native plants and to extract the medical subject headings unique identifier of the bioentities, which co-occurred with the keywords in the same literature. ChexMix discovered the allied species related to a keyword of interest and experimentally proved its usefulness for multi-species analysis.

The currently enhanced computing power is boosting the acquisition and processing of scientific data obtained from wet and dry lab experiments. In the fields of biology and chemistry, a huge amount of literature is being published every day and uploaded to the databases in real time. Several databases are currently available applying manual curation or in silico approaches for data management in arbitrary forms1–3. Hence, finding the meaningful information from large databases is almost like 'finding a needle in a haystack'. Therefore, automation techniques, such as text-mining and natural language processing (NLP) methods, were developed to help convert raw scientific texts into well-structured scientific data6,7. Recently, many machine learning technique for NLP have been applied to significantly improve and utilize the text mining performances of models8–13.

PubTator Central (PTC) is a state-of-the-art text-mining service for automated annotation of bioentities including genes/proteins, genetic variants, diseases, chemicals, species, and cell lines in about 30 million abstracts and 3 million full-text articles available in PubMed (https://www.ncbi.nlm.nih.gov/pubmed)14,15. The bioentities co-occurring in the body of literature and extracted by PTC can be re-organized as a complex network and used for literature-based discovery.

Herein, we introduce the bioentity extraction tool, ChexMix, and applied it to extract inter-relationships between medical subject headings (MeSH, https://www.ncbi.nlm.nih.gov/mesh) terms and taxonomy identifiers (TaxIDs) in the National Center for Biotechnology Information (NCBI) Taxonomy from the co-occurrence relationships from PTC16. ChexMix is an open source python module for accessing and processing various forms of data from multiple biomedical databases. It collects the bioentities, such as species, chemicals, and diseases, which were annotated by the PTC, from abstracts and/or free full-text biomedical articles stored in the PubMed database. ChexMix converts and links the bioentities found in the literature with the unique identifiers of the species (TaxID), chemicals (MeSH), and diseases (MeSH). The association between these bioentities can be modulated into bipartite or multipartite networks, or hierarchical trees, aiding to inspect and simply understand the holistic structures of the information associated with targeted topics queried as keywords.

Each bioentity has its own hierarchical organization system according to the bioentity type. For taxonomy, species is located in the lowest rank of the taxonomic hierarchy and is involved with the higher ranks, genus and family. The hierarchical location and distance in the taxonomic tree between the species bioentities provide clues to discover other hidden bioentities. Species under the same genus share more similar features compared with different species under another genus. For chemicals, since similar structures can interact with proteins holding...
Amentoflavone was confirmed by its isotopic peak at 537.4 m/z \([M + H]^-\) detected by liquid chromatography–mass spectrometry. Among them, the leaves of *V. erosum* contained the highest amount of amentoflavone (7.39 mg/g) compared with *Selaginella tamariscina*, which is the representative natural ingredient for anti-wrinkle effect in the cosmetic industry. Overall, the summarization of hierarchical bioentities information using ChexMix is expected to help inspect massive sparse bioentities in databases in future investigations.

The performance of the results from ChexMix was quantitatively evaluated based on the extracted bioentities using a set of keywords which are associated with the original keyword ‘amentoflavone’. 243 networks of taxonomies were obtained using ChexMix from MeSH terms of chemicals co-occurred with ‘amentoflavone’ in the literature, and they were analyzed by the basic network properties and similarity metrics (Supplementary Table S3). The similarity metrics compared each of the 243 networks with the network of ‘amentoflavone’, where the number of true positives was calculated by the number of common nodes in both of the networks (Supplementary Table S3).

Additionally, ChexMix can also integrate the results from multi-keywords. The MeSH identifiers for bioentities co-occuring with the keywords of interests could be used for connecting the results by two different queries (Fig. 4). For instance, two species names, *Taxus cuspidata* and *Podophyllum peltatum*, were queried by ChexMix to avoid duplicated studies and find novel bioactive sources, the analysis was focused on the allied species belonging to the *Viburnum* genus, retrieving 19 samples of different parts of eight species native to Korea that were not previously studied on amentoflavone-related topics (Fig. 3, Supplementary Table S2). Next, the existence of amentoflavone was evaluated in samples of these plants and quantified by HPLC. The presence of amentoflavone was confirmed by its isotopic peak at 537.4 m/z \([M + H]^-\) detected by liquid chromatography–mass spectrometry. Among them, the leaves of *V. erosum* contained the highest amount of amentoflavone (7.39 mg/g) compared with *Selaginella tamariscina*, which is the representative natural ingredient for anti-wrinkle effect and the major source of amentoflavone in the cosmetic industry. Overall, the summarization of hierarchical bioentities information using ChexMix is expected to help inspect massive sparse bioentities in databases in future investigations.

The performance of the results from ChexMix was quantitatively evaluated based on the extracted bioentities using a set of keywords which are associated with the original keyword ‘amentoflavone’. 243 networks of taxonomies were obtained using ChexMix from MeSH terms of chemicals co-occurred with ‘amentoflavone’ in the literature, and they were analyzed by the basic network properties and similarity metrics (Supplementary Table S3). The similarity metrics compared each of the 243 networks with the network of ‘amentoflavone’, where the number of true positives was calculated by the number of common nodes in both of the networks (Supplementary Table S3).
and generated two small networks consisting of bioentities with MeSH identifiers extracted from PubTator. It was possible to inspect the co-occurred bioentities among the MeSH identifiers in the integrated network. The network of each species showed different MeSH identifier profiles and MeSH identifiers related to ‘cancer’, in particular ‘ovarian neoplasms’, co-occurred. This agrees with the fact that paclitaxel of *T. cuspidata* and podophyllotoxin of *P. peltatum* are well-known potent anticancer drugs for ovarian cancer25–27.

Here, a usage scenario of ChexMix to alleviate the complex task of compiling large data by narrowing down the scope of bioentities or grouping similar bioentities using the hierarchical relationships was described. Firstly, to obtain the appearance counts of bioentities in literature queried by keywords of interest, ChexMix collects PubMed and PMC literature followed by fetching annotations within that data from PTC and converting them into unique identifiers according to the respective bioentity class. ChexMix allows Boolean operators (‘AND’, ‘OR’, ‘NOT’), double quotes for phrases, and asterisk for truncated terms for PubMed literature search. Each bioentity extracted from ChexMix is classified within more general categories of bioentity and arranged in a hierarchical structure.

When single or multi keywords of interest are entered in ChexMix, bioentities in all citations that have keywords are retrieved and automatically mapped into unique identifiers. The search results indicate the co-occurrence of bioentities in the available literature, allowing to link them and yielding the co-occurrence network.
ChexMix makes the process straightforward by managing data access from multiple sources and providing functions to manipulate the network data structure.

The analysis is mainly focused on taxonomy terms to inspect the species that biologically affect physiological disorders or diseases within the network. Each taxonomy name in the search results is listed in a hierarchical form. Trivial bioentities are located on the higher ranks of the list. Other near species within the obtained taxonomic tree are expected to have similar biological effects, representing potential alternative biomedical options. ChexMix can also generate the connections between taxonomic terms and MeSH identifiers, which are located under ‘Diseases [C]’ and ‘Chemicals and Drugs [D]’, in the same literature. MeSH identifiers co-occurring with a taxonomic term in the literature are expected to have a close relationship.

In Fig. 4, the intersection set of MeSH terms co-occurring with each taxonomy keyword is highlighted on the whole network resulting from the union set of two networks. Networks generated from a single keyword in ChexMix can be simply reprocessed by the combination of set operations, such as union, difference, and intersection with other networks. The re-organization of complex networks from single or multi keywords provides new insights or clues for bioentities in PubMed, the biggest biomedical database.

In the present study, we have focused on how to use ChexMix to construct a taxonomic tree or a co-occurrence network from multi keywords, and analyze the networks from bioentities identified by PTC. We designed ChexMix for easily adapting the diverse types of bioentities and integrating other existing databases as well as recently introduced state-of-art text mining systems. We hope ChexMix will be utilized for other researchers to integrate other datasets, and manipulate and visualize the relationships between bioentities.

Figure 3. (A) Chemical structure of amentoflavone. (B) Chromatograms of the five samples with the highest amentoflavone content determined as described in the “Methods” section. AMEN, amentoflavone; VCL, leaves of *Viburnum carlesii*; VDF, fruits of *V. furcatum*; VDSt, leaves of *V. dilatatum*; VEL, leaves of *V. erosum*; VEst, stems of *V. erosum*. 
Methods

Data processing. ChexMix currently obtains biomedical data from multiple databases using their web application programming interface (APIs) or bulk data files. For example, Entrez API allows to query Entrez databases, such as PubMed and PubMed Central (PMC), using combinations of keywords. PTC provides a web API to fetch annotations of biomedical concepts, such as taxonomy and MeSH identifiers, in a publication. ChexMix also manages to download and parse bulk data files from biomedical databases. For example, ChexMix loads the data from PTC, including NCBI taxonomy and MeSH that inherently have relationships between entities therein, and transforms it into internal network data structures. ChexMix also grants the possibility to construct, manipulate, and simplify the network data structures.

Bioentities extraction and visualization. The keywords of interest can be input as single words or phrases. The results are output in hierarchical tree format according to their own taxonomic or hierarchically-organized rules for each type of bioentity. In the case of taxonomy information, species names in the literature are encoded into unique identifiers, TaxID, and hierarchically re-organized in the classification rules of NCBI taxonomy. In the present study, hierarchical results were applied to discover relevant species with lower taxonomic ranks (family and genus levels) using the list of the Korean medicinal plants of the Korea Plant Extract Bank (KPEB). The results were visualized in the network format using the Gephi software (ver. 0.9.2).

Sample preparation. To prove the usefulness of ChexMix, 18 Viburnum samples, including V. carlesii, V. dilatatum, V. wrightii, V. opulus, V. furcatum, and V. awubuki, were purchased from the KPEB of the Korea Research Institute of Bioscience and Biotechnology, Korea. V. erosum was collected from the Medicinal Herb Garden of the College of Pharmacy, Seoul National University, Korea and deposited in the Medicinal Herbarium of Kangwon National University with the accession number KNUVE-1. The use of plants in the present study complies with the guidelines of the Medicinal Herbarium of Kangwon National University. Each powdered sample (1 g) was extracted using 80% methanol for 1 h using an ultrasonic apparatus and the filtrate was dried using a vacuum rotary evaporator. Next, the samples were dissolved in 100% methanol at a concentration of 10 mg/mL and filtered through a 0.45 μm polytetrafluoroethylene membrane before analysis.

High-performance liquid chromatography (HPLC) analysis. The samples were analyzed on a 1260 quaternary pump, an autosampler, and a multiple wavelength detector (Agilent Technologies, Santa Clara, CA, USA). Chromatographic separation was performed using a Hector-M C18 column (250 × 4.6 mm I.D.; 5 μm, RSTech, Daejeon, Korea). The ultraviolet detector was set at a wavelength of 260 nm. The mobile phase was a gradient solvent system consisting of solvent A (0.1% formic acid in water) and solvent B (MeOH) as follows: isocratic 95% A (0–10 min), linear gradient 95–80–30% A (10–20–30 min) and isocratic 30% A (30–40 min). The flow rate was 1.0 mL/min, and aliquots of 10 μL were injected using an autosampler.
The in-house codes used to extract the bioentities and classify them, as well as to perform the data visualization, along with some examples, are publicly available at the github repository: https://github.com/bionsight/chexmix.

Received: 22 March 2021; Accepted: 3 May 2022
Published online: 12 May 2022

References
1. Mendez, D. et al. ChEMBL: Towards direct deposition of bioassay data. Nucleic Acids Res. 47, D930–D940 (2019).
2. Gilson, M. K. et al. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic Acids Res. 44, D1045–D1053 (2016).
3. Wassermann, A. M. & Bajorath, J. BindingDB and ChEMBL: Online compound databases for drug discovery. Expert Opin. Drug Discov. 6, 683–687 (2011).
4. Wishart, D. S. et al. DrugBank 5.0: A major update to the DrugBank database for 2018. Nucleic Acids Res. 46, D1074–D1082 (2018).
5. Davis, A. P. et al. The comparative toxicogenomics database. Update 2019. Nucleic Acids Res. 47, D948–D954 (2019).
6. Wilson, S. et al. Automated literature mining and hypothesis generation through a network of Medical Subject Headings. bioRxiv https://doi.org/10.1101/403667 (2018).
7. Himmelstein, D. S. et al. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. Elife 6, e26726 (2017).
8. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv arXiv:1810.04805 Cs (2019).
9. Takeuchi, K. & Collier, N. Bio-medical entity extraction using support vector machines. Artif. Intell. Med. 33, 125–137 (2005).
10. Ohta, T., Tateisi, Y. & Kim, J.-D. The GENIA corpus: an annotated research abstract corpus in molecular biology domain. In Proceedings of the Second International Conference on Human Language Technology Research, 82 (Association for Computational Linguistics, 2002) https://doi.org/10.3115/1289189.1289260.
11. Yadav, S., Bhattacharyya, P. Entity extraction in biomedical corpora: an approach to evaluate word embedding features with PSO based feature selection. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers 1159–1170 (Association for Computational Linguistics, 2017).
12. Perera, N., Dehmer, M. & Emmert-Streib, F. Named entity recognition and relation detection for biomedical information extraction. Front. Cell Dev. Biol. 8, 673 (2020).
13. Sänger, M. & Leser, U. Large-scale entity representation learning for biomedical relationship extraction. Bioinformatics 37, 236–242 (2021).
14. Wei, C.-H., Kao, H.-Y. & Lu, Z. PubTator: A web-based text mining tool for assisting biocuration. Nucleic Acids Res. 41, W518–W522 (2013).
15. Wei, C.-H., Allot, A., Leaman, R. & Lu, Z. PubTator central: Automated concept annotation for biomedical full text articles. Nucleic Acids Res. 47, W587–W593 (2019).
16. Federhen, S. The NCBI taxonomy database. Nucleic Acids Res. 40, D136–D143 (2012).
17. Ertl, P. An algorithm to identify functional groups in organic molecules. J. Cheminform. 9, 1–7 (2017).
18. Ertl, P. & Schulumann, T. Cheminformatics analysis of natural product scaffolds: Comparison of scaffolds produced by animals, plants, fungi and bacteria. bioRxiv https://doi.org/10.1101/2020.01.28.922955 (2020).
19. Djoumbou Feunang, Y. et al. ClassyFire: Automated chemical classification with a comprehensive, computable taxonomy. J. Cheminform. 8, 61 (2016).
20. Uhlen, M. et al. Tissue-based map of the human proteome. Science 347, 1260419–1260419 (2015).
21. Uhlen, M. et al. A pathology atlas of the human cancer transcriptome. Science 357, eaan2507 (2017).
22. Yu, S. et al. A review on the phytochemistry, pharmacology, and pharmacokinetics of amentoflavone, a naturally-occurring biflavonoid. Molecules 22, 299 (2017).
23. Park, N.-H., Lee, C.-W., Bae, J. & Na, Y. J. Protective effects of amentoflavone on Lamin A-dependent UVB-induced nuclear aberration in normal human fibroblasts. Bioorg. Med. Chem. Lett. 21, 6482–6484 (2011).
24. Yuan, C. Simultaneous determination of selagellinines and biflavones in Selaginella tamariscina and S. pulvinata by HPLC. China J. Chin. Mater. Medica https://doi.org/10.4268/cjcm20120918 (2012).
25. Baird, R. D., Tan, D. S. P. & Kaye, S. B. Weekly paclitaxel in the treatment of recurrent ovarian cancer. Nat. Rev. Clin. Oncol. 7, 575–582 (2010).
26. Zhao, W. et al. Challenges and potential for improving the druggability of podophyllotoxin-derived drugs in cancer chemotherapy. Nat. Prod. Rep. https://doi.org/10.1039/D0NP0041H (2021).
27. Mukherjee, A., Basu, S., Sarkar, N. & Ghosh, A. Advances in cancer therapy with plant based natural products. Curr. Med. Chem. 8, 1467–1486 (2001).
28. Lee, N., Yoo, H. & Yang, H. Cluster analysis of medicinal plants and targets based on multipartite network. Biomolecules 11, 546 (2021).
29. Swainston, N. et al. libChEBI: An API for accessing the ChEBI database. J. Cheminform. 8, 11 (2016).
30. Bastian, M., Heymann, S. & Jacomy, M. Gephi: An Open Source Software for Exploring and Manipulating Networks. (2009) https://doi.org/10.13140/2.1.1341.1520.

Acknowledgements
This work was supported by the 2019 Research Grant (PoNT) from Kangwon National University and the Basic Science Research Program through the National Research Foundation of Korea (NRF) grant funded by the Korean government (MEST) (NRF-2021R1C1C1011857).

Author contributions
Heejung Y. and Hojin Y. conceived and coordinated the project. J.P., J.L. and H.S.J. collected Viburnum samples and analyzed the content of amentoflavone. B.P., Heejung Y. and Hojin Y. wrote the in-house python codes. B.P., N.L., Hojin Y. and Heejung Y. contributed to discussions, and Heejung Y. and Hojin Y. wrote the manuscript.

Competing interests
The authors declare no competing interests.
Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-12093-9.

Correspondence and requests for materials should be addressed to H.Y.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2022