Received: 1 December 2023 | Accepted: 16 March 2024

DOI: 10.1002/aps3.11604

Applications in Plant Sciences

SOFTWARE NOTE

BAD2matrix: Phylogenomic matrix concatenation, indel coding, and more

Nelson R. Salinas[1] | **Gil Eshel**[2] | **Gloria M. Coruzzi**[2] | **Rob DeSalle**[3] | **Michael Tessler**[1,3,4] | **Damon P. Little**[1]

[1]Lewis B. and Dorothy Cullman Program for Molecular Systematics, The New York Botanical Garden, Bronx, New York, USA

[2]Center for Genomics and Systems Biology, New York University, New York, New York, USA

[3]Institute for Comparative Genomics, American Museum of Natural History, New York, New York, USA

[4]Department of Biology, Medgar Evers College, City University of New York, Brooklyn, New York, USA

Correspondence

Damon P. Little, Lewis B. and Dorothy Cullman Program for Molecular Systematics, The New York Botanical Garden, Bronx, New York, USA.
Email: dlittle@nybg.org

Abstract

Premise: Common steps in phylogenomic matrix production include biological sequence concatenation, morphological data concatenation, insertion/deletion (indel) coding, gene content (presence/absence) coding, removing uninformative characters for parsimony analysis, recording with reduced amino acid alphabets, and occupancy filtering. Existing software does not accomplish these tasks on a phylogenomic scale using a single program.

Methods and Results: BAD2matrix is a Python script that performs the above-mentioned steps in phylogenomic matrix construction for DNA or amino acid sequences as well as morphological data. The script works in UNIX-like environments (e.g., LINUX, MacOS, Windows Subsystem for LINUX).

Conclusions: BAD2matrix helps simplify phylogenomic pipelines and can be downloaded from https://github.com/dpl10/BAD2matrix/tree/master under a GNU General Public License v2.

KEYWORDS

concatenation, gene content, gene presence/absence, indel coding, morphology, occupancy filtering, phylogenomics, reduced amino acid alphabets

Matrix concatenation is an early, critical task in all phylogenomic pipelines. Phylogenomics, in this paper, references phylogenetic and evolutionary analyses using massive amounts of genome and/or transcriptome data. Although there are numerous programs that can concatenate matrices at the multilocus scale (e.g., 2matrix [Salinas and Little, 2014], SequenceMatrix [Vaidya et al., 2011], iPhy [Jones et al., 2011], AIR-appender [Kumar et al., 2009]), input datasets for phylogenomic analyses of eukaryotes are typically one to two orders of magnitude larger than would be used in even the largest traditional multilocus phylogenetic datasets. Accordingly, few publicly available, stand-alone programs (e.g., SCaFoS [Roure et al., 2007], IQ-TREE [Minh et al., 2020], Phyx [Brown et al., 2017]) can accomplish the concatenation task on a genomic scale. Furthermore, most programs omit a substantial amount of genetic data while composing their matrices, namely, insertions/deletions (indels) and gene content (presence/absence); morphology is also often ignored in phylogenomic tools. Additionally, most phylogenomic practitioners remove genes from the concatenated matrix that have not been recovered for a set proportion (e.g., 25–50%) of the terminal organisms (Philippe et al., 2004; Roure et al., 2013; Streicher et al., 2016)—a process we will refer to as occupancy filtering. At a phylogenomic scale, analytic efficiency is also critically important and can be improved by reducing problematic or unnecessary portions of the data. For example, removing loci with missing data typically increases efficiency, and parsimony analyses are made much more efficient by the removal of parsimony uninformative characters during matrix construction.

Previously, we produced 2matrix (Salinas and Little, 2014), which focused on matrix concatenation at a phylogenetic scale and allowed both molecular and morphological data to be concatenated. While 2matrix performs well, it is nonfunctional for large-scale molecular datasets because it stores all sequence data in RAM in order

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.
© 2024 The Author(s). *Applications in Plant Sciences* published by Wiley Periodicals LLC on behalf of Botanical Society of America.

to produce output quickly. For this reason, the ability to function with thousands of genes has been heavily requested by 2matrix users. Furthermore, 2matrix does not compute gene content or conduct occupancy filtering, as these are typically neither necessary nor relevant to studies involving a few loci.

Indels and gene content coding provide additional phylogenomic data that are generally helpful for phylogenomics. Indels can improve resolution and support (Giribet and Wheeler, 1999; Rokas and Holland, 2000; Simmons et al., 2007; Paśko et al., 2011; Yuri et al., 2013; Donath and Stadler, 2018; Suvorov et al., 2020), whereas gene content characters are similarly useful for tree building (Snel et al., 1999; Rosenfeld et al., 2017) and thus should be a useful total-evidence matrix addition to help resolve deep-level nodes in phylogenomic reconstructions. Both indels and gene content characters retain information about gene content that is typically lost in phylogenomic matrices. Furthermore, reduced amino alphabets often lessen issues with saturation and compositional heterogeneity, which can negatively impact tree-building (Susko and Roger, 2007; Feuda et al., 2017). Lastly, morphology is relevant but is often ignored for phylogenomics: We believe incorporating these data should further help to resolve difficult nodes that are of interest in phylogenomics (e.g., morphology for the early-diverging animal lineages [Neumann et al., 2021]).

There are programs currently available for some of the functions detailed here, but none that provide a single path for all the methods. A partial list of publicly available tools that are built to work with phylogenomic-scale data includes: SCaFoS (Roure et al., 2007) for concatenation and occupancy filtering; NGS-Indel Coder (Boutte et al., 2019), GATK (McKenna et al., 2010), and SIDIER (Muñoz-Pajares, 2013) for indel coding; Phyx for concatenation and uninformative position filtering (Brown et al., 2017); and Roary (Page et al., 2015) and BPGA (Chaudhari et al., 2016) for gene content recording.

Here we present Big-Ass Dataset 2matrix (BAD2matrix), which concatenates gene and morphology files, thereby facilitating partitioned analyses. BAD2matrix is the only concatenation program currently available that can also code indels, record gene content, recode amino acid sequences in reduced alphabets, remove parsimony uninformative sequence characters, and/or allow for occupancy filtering. It formats output matrices for IQ-TREE 2 (Minh et al., 2020), RAxML (Stamatakis, 2014; Minh et al., 2020), TNT (Goloboff and Catalano, 2016), and FastTree (Price et al., 2010), providing a smooth transition to tree-building with a wide variety of popular programs.

METHODS AND RESULTS

BAD2matrix is a serial script written in Python 3 and is partially based on 2matrix (Salinas and Little, 2014). The program produces a concatenated matrix from a directory of many individual FASTA (Pearson and Lipman, 1988) files of DNA or amino acid sequences. The OrthologID (Chiu et al., 2006) convention is used for sequence names, i.e., ">species#sequenceID", where "species" is the name of the terminal taxon used (e.g., "Dryopteris_intermedia" or "SARS-CoV-2") and "sequenceID" is a unique identifier for the particular sequence (e.g., "seq. 101"). Therefore, by default, data across alignments will be concatenated by the species name only. This can be overridden by the user with the "-f" flag, in which case the entire accession given in the FASTA file is used (e.g., "Dryopteris_intermedia#seq. 101"), and accession names must be consistent across all FASTA files. When using this option, we recommend completely avoiding sequenceIDs. For both naming conventions, only letters, numbers, periods, and underscores are preserved in sequenceIDs to prevent incompatibility with downstream phylogenetic programs. Filenames in the input directory are used to determine locus names (e.g., for a file named "matK.fasta", the locus will be named "matK"). Both sequence and morphology data can be concatenated.

If the user has specified a cutoff for occupancy coding ("-m x" flag), files that have occupancy below that threshold are skipped and no data are output for them. The occupancy threshold x is the upper percentile of genes in the distribution of missing data to be retained (e.g., if $x = 25$, genes in the upper 25 percentiles of occupancy will be retained, thereby removing genes in the lower 75 percentiles). Authors have suggested different cutoffs for occupancy, but have typically acknowledged that missing data should be limited to some degree (Philippe et al., 2004; Roure et al., 2013; Streicher et al., 2016). Only one file is read into memory (RAM) at a time, which is very RAM efficient, but demanding in regard to disk use (read/write). For parsimony analysis, only informative sequence characters are included in the output. Amino acids can also be recoded using any of the 13 reduced amino acid alphabets available with bins ranging from two to 18 character states; a few recent phylogenomic works focused on six-character binnings of amino acid data (Feuda et al., 2017; Laumer et al., 2018; Neumann et al., 2021).

BAD2matrix offers the option to code and append a binary gene content (absence/presence) matrix to the concatenated sequence data. Absence and presence are recorded based on whether a sequence for a specific taxon is present in a given FASTA file. For example, if only sequence data from taxon 1 is missing for a given locus, taxon 1 is coded "0" while all other taxa in the matrix are coded "1". Similarly, BAD2matrix optionally produces an indel absence/presence matrix using "simple indel coding" (Simmons and Ochoterena, 2000). This type of indel coding records absence ("0")/presence ("1") of unique indels (different 5′ and/or 3′ positions). When an indel is a subset of a longer indel, the taxa with the longer indel are coded as not applicable ("-") for the shorter indel. The resulting matrix can be formatted for several of the most popular likelihood-based phylogenomics programs: RAxML (RAxML-NG version 1.1.0; extended PHYLIP [Stamatakis, 2014; Kozlov et al., 2019]) or IQ-TREE 2 (extended PHYLIP [Minh et al., 2020]). Formatting for FastTree 2 (Price et al., 2010), a shortcut likelihood program, is also available;

TABLE 1 Command-line flags for BAD2matrix.

Option flag[a]	Description	Required
-a *2 | 3 | 4 | 5 | 6|6dso|6kgb| 6sr | 8 | 10 | 11 | 12 | 15 | 18 | 20*	Number of amino acid states (default = 20). Reduction with option "6dso" follows Dayhoff et al. (1978); option "6kgb" follows Kosiol et al. (2004); option "6sr" follows Susko and Roger (2007); option "11" follows Buchfink et al. (2015); and all other options follow Murphy et al. (2000).	No
-d *directory-name*	Specify the input directory of aligned FASTA files. Names should use the following convention: ">species#sequenceID". Characters other than letters, numbers, periods, and underscores will be deleted. Use "-f" for an alternate naming convention.	Yes
-f	Use full FASTA names rather than default settings (see "-d" description for default). Characters other than letters, numbers, periods, and underscores will be deleted.	No
-g	Do not code gene content (absence/presence). If this flag is not set, gene content is coded.	No
-i	Do not code indels. If this flag is not set, indels are coded.	No
-m *x*	Retain the upper *x* percentile of genes in the distribution of missing sequences. By default, $x = 1$ (i.e., include all genes with four or more sequences).	No
-n *root-name*	Specify the *root-name* for output files.	Yes
-r	Folder containing a morphological matrix or a set of ortholog duplication matrices. Datasets should be saved as .tsv tables. Multiple states should be separated by pipes ("|").	No

[a]Italicized text following option flags should be specified by the user.

TABLE 2 Example of an output concatenated matrix, showing gene alignments, simple indel codings, and gene content (presence/absence) codings.

Genes				Indels				Gene content			
G 1	**G 2**	**…**	**G 1 K**	**I 1**	**I 2**	**…**	**I 1 K**	**GC 1**	**GC 2**	**…**	**GC 1 K**
ACTG	C - - A	…	—	0	1	…	*	1	1	…	0
AC -G	C - - A	…	—	1	1	…	*	1	1	…	0
—	—	…	TGAC	?	?	…	*	0	0	…	1
ACTG	CTCA	…	TGCC	0	0	…	*	1	1	…	1
AC -G	CTGA	…	TGCC	1	0	…	*	1	1	…	1

*Loci without indels are not used for indel coding.

however, morphology, indel, and gene content data cannot be included due to limitations inherent in FastTree 2. The matrix can also be formatted for parsimony analysis in TNT version 1.5 using extended XREAD (Goloboff and Catalano, 2016). A detailed explanation of the command-line flags is provided in Table 1, and an example of an output concatenated matrix is provided in Table 2. The code is designed for UNIX-like environments and has been tested on MacOS version 11.6 (Apple, Cupertino, California, USA) and Ubuntu version 22.04 (Canonical, London, United Kingdom). Along with the code, an example dataset (Little, 2006) composed of multiple DNA regions and a morphology matrix are included.

CONCLUSIONS

BAD2matrix is deposited on GitHub (https://github.com/dpl10/BAD2matrix, see Data Availability Statement), where instructions, example data, and further documentation can be accessed. It carries out critical first steps in phylogenomic matrix production and simplifies analytic options that are often ignored. These options—concatenation of sequence and morphological data, indel coding, gene content (presence/absence) coding, removal of parsimony uninformative sequence characters, reduced amino acid alphabets, and occupancy filtering—are often omitted simply because there are no programs that easily implement them. These BAD2matrix analytic options allow users to maximize their data and more fully control matrix composition at phylogenomic scale.

AUTHOR CONTRIBUTIONS

G.E., R.D., G.M.C., M.T., and D.P.L. conceived the project, and N.R.S., G.E., M.T., and D.P.L. designed the project. N.R.S., G.E., and D.P.L. acquired the data. N.R.S., G.E., and D.P.L. performed the coding. N.R.S., R.D., M.T., and D.P.L. wrote the manuscript. G.M.C., R.D., and D.P.L. provided funding. All authors approved the final version of the manuscript.

ACKNOWLEDGMENTS

The authors thank the United States Department of Energy (Biological and Environmental Research [BER]

grant DE-SC0014377), the U.S. National Science Foundation (Division of Environmental Biology [DEB] grants 1655050 and 1758800), and the Zegar Family Foundation for funding this project.

DATA AVAILABILITY STATEMENT

All code has been deposited in GitHub (https://github.com/dpl10/BAD2matrix) and is available under a GNU General Public License v2. Code releases are automatically mirrored in Zenodo (Salinas and Little, 2023; https://zenodo.org/doi/10.5281/zenodo.10028408).

ORCID

Nelson R. Salinas http://orcid.org/0000-0002-4812-8674
Damon P. Little http://orcid.org/0000-0001-9635-6164

REFERENCES

Boutte, J., M. Fishbein, A. Liston, and S. C. K. Straub. 2019. NGS-Indel Coder: A pipeline to code indel characters in phylogenomic data with an example of its application in milkweeds (*Asclepias*). *Molecular Phylogenetics and Evolution* 139: 106534.

Brown, J. W., J. F. Walker, and S. A. Smith. 2017. Phyx: Phylogenetic tools for unix. *Bioinformatics* 33: 1886–1888.

Buchfink, B., C. Xie, and D. H. Huson. 2015. Fast and sensitive protein alignment using DIAMOND. *Nature Methods* 12: 59–60.

Chaudhari, N. M., V. K. Gupta, and C. Dutta. 2016. BPGA: An ultra-fast pan-genome analysis pipeline. *Scientific Reports* 6: 24373.

Chiu, J. C., E. K. Lee, M. G. Egan, I. N. Sarkar, G. M. Coruzzi, and R. DeSalle. 2006. OrthologID: Automation of genome-scale ortholog identification within a parsimony framework. *Bioinformatics* 22: 699–707.

Dayhoff, M. O., R. Schwartz, and B. Orcutt. 1978. A model of evolutionary change in proteins. *In* M. O. Dayhoff [ed.], Atlas of protein sequence and structure, 345–352. National Biomedical Research Foundation, Washington, D.C., USA.

Donath, A., and P. F. Stadler. 2018. Split-inducing indels in phylogenomic analysis. *Algorithms for Molecular Biology* 13: 12.

Feuda, R., M. Dohrmann, W. Pett, H. Philippe, O. Rota-Stabelli, N. Lartillot, G. Wörheide, and D. Pisani. 2017. Improved modeling of compositional heterogeneity supports sponges as sister to all other animals. *Current Biology* 27: 3864–3870.

Giribet, G., and W. C. Wheeler. 1999. On gaps. *Molecular Phylogenetics and Evolution* 13: 132–143.

Goloboff, P. A., and S. A. Catalano. 2016. TNT version 1.5, including a full implementation of phylogenetic morphometrics. *Cladistics* 32: 221–238.

Jones, M. O., G. D. Koutsovoulos, and M. L. Blaxter. 2011. iPhy: An integrated phylogenetic workbench for supermatrix analyses. *BMC Bioinformatics* 12: 30.

Kosiol, C., N. Goldman, and N. H. Buttimore. 2004. A new criterion and method for amino acid classification. *Journal of Theoretical Biology* 228: 97–106.

Kozlov, A. M., D. Darriba, T. Flouri, B. Morel, and A. Stamatakis. 2019. RAxML-NG: A fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* 35: 4453–4455.

Kumar, S., A. Skjaeveland, R. J. S. Orr, P. Enger, T. Ruden, B.-H. Mevik, F. Burki, et al. 2009. AIR: A batch-oriented web program package for construction of supermatrices ready for phylogenomic analyses. *BMC Bioinformatics* 10: 357.

Laumer, C. E., H. Gruber-Vodicka, M. G. Hadfield, V. B. Pearse, A. Riesgo, J. C. Marioni, and G. Giribet. 2018. Support for a clade of Placozoa and Cnidaria in genes with minimal compositional bias. *eLife* 7: e36278.

Little, D. P. 2006. Evolution and circumscription of the true cypresses (Cupressaceae: Cupressus). *Systematic Botany* 31: 461–480.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* 20: 1297–1303.

Minh, B. Q., H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. von Haeseler, and R. Lanfear. 2020. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution* 37: 1530–1534.

Muñoz-Pajares, A. J. 2013. SIDIER: Substitution and indel distances to infer evolutionary relationships. *Methods in Ecology and Evolution* 4: 1195–1200.

Murphy, L. R., A. Wallqvist, and R. M. Levy. 2000. Simplified amino acid alphabets for protein fold recognition and implications for folding. *Protein Engineering* 13: 149–152.

Neumann, J. S., R. Desalle, A. Narechania, B. Schierwater, and M. Tessler. 2021. Morphological characters can strongly influence early animal relationships inferred from phylogenomic data sets. *Systematic Biology* 70: 360–375.

Page, A. J., C. A. Cummins, M. Hunt, V. K. Wong, S. Reuter, M. T. G. Holden, M. Fookes, et al. 2015. Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 31: 3691–3693.

Paśko, Ł., P. G. P. Ericson, and A. Elzanowski. 2011. Phylogenetic utility and evolution of indels: A study in neognathous birds. *Molecular Phylogenetics and Evolution* 61: 760–771.

Pearson, W. R., and D. J. Lipman. 1988. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences, USA* 85: 2444–2448.

Philippe, H., E. A. Snell, E. Bapteste, P. Lopez, P. W. H. Holland, and D. Casane. 2004. Phylogenomics of eukaryotes: Impact of missing data on large alignments. *Molecular Biology and Evolution* 21: 1740–1752.

Price, M. N., P. S. Dehal, and A. P. Arkin. 2010. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5: e9490.

Rokas, A., and P. W. H. Holland. 2000. Rare genomic changes as a tool for phylogenetics. *Trends in Ecology & Evolution* 15: 454–459.

Rosenfeld, J. A., S. Oppenheim, and R. DeSalle. 2017. A whole genome gene content phylogenetic analysis of anopheline mosquitoes. *Molecular Phylogenetics and Evolution* 107: 266–269.

Roure, B., N. Rodriguez-Ezpeleta, and H. Philippe. 2007. SCaFoS: A tool for selection, concatenation and fusion of sequences for phylogenomics. *BMC Evolutionary Biology* 7(Suppl 1): S2.

Roure, B., D. Baurain, and H. Philippe. 2013. Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. *Molecular Biology and Evolution* 30: 197–214.

Salinas, N. R., and D. P. Little. 2014. 2matrix: A utility for indel coding and phylogenetic matrix concatenation. *Applications in Plant Sciences* 2(1): 1300083.

Salinas, N. R., and D. P. Little. 2023. BAD2matrix 1.0. Available at Zenodo repository: https://doi.org/10.5281/zenodo.10028408 [posted 20 October 2023; accessed 19 June 2024].

Simmons, M. P., and H. Ochoterena. 2000. Gaps as characters in sequence-based phylogenetic analyses. *Systematic Biology* 49: 369–381.

Simmons, M. P., K. Müller, and A. P. Norton. 2007. The relative performance of indel-coding methods in simulations. *Molecular Phylogenetics and Evolution* 44: 724–740.

Snel, B., P. Bork, and M. A. Huynen. 1999. Genome phylogeny based on gene content. *Nature Genetics* 21: 108–110.

Stamatakis, A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30: 1312–1313.

Streicher, J. W., J. A. Schulte II, and J. J. Wiens. 2016. How should genes and taxa be sampled for phylogenomic analyses with missing data? An empirical study in iguanian lizards. *Systematic Biology* 65: 128–145.

Susko, E., and A. J. Roger. 2007. On reduced amino acid alphabets for phylogenetic inference. *Molecular Biology and Evolution* 24: 2139–2150.

Suvorov, A., J. Hochuli, and D. R. Schrider. 2020. Accurate inference of tree topologies from multiple sequence alignments using deep learning. *Systematic Biology* 69: 221–233.

Vaidya, G., D. J. Lohman, and R. Meier. 2011. SequenceMatrix: Concatenation software for the fast assembly of multi-gene datasets with character set and codon information. *Cladistics* 27: 171–180.

Yuri, T., R. T. Kimball, J. Harshman, R. C. K. Bowie, M. J. Braun, J. L. Chojnowski, K.-L. Han, et al. 2013. Parsimony and model-based analyses of indels in avian nuclear genes reveal congruent and incongruent phylogenetic signals. *Biology* 2: 419–444.

How to cite this article: Salinas, N. R., G. Eshel, G. M. Coruzzi, R. DeSalle, M. Tessler, and D. P. Little. 2024. BAD2matrix: Phylogenomic matrix concatenation, indel coding, and more. *Applications in Plant Sciences* 12(6): e11604. https://doi.org/10.1002/aps3.11604