
Karen Hogenboom
Beth Sheehan
Heidi Imker
Hello
my name is
Data Management Workshop Series

• Introduction to Data Management
  – 10-11 am, Thursday, October 1, 2015

• Documentation and Organization for Data and Processes
  – 10-11 am, Thursday, October 8, 2015

• Making Research Data Public: Why, What, and How
  – 10-11 am, Thursday, November 5, 2015
Objectives

1. Learn the benefits of making data available in order to identify how sharing can advance your research.

2. Assess the legal and ethical considerations that may apply to your data in order to make informed decisions about what data to share and how to share it.

3. Use tools to identify and evaluate data sharing options in order to guide your strategy for publishing data.
Why Share Your Data?
Benefits

Statistical binning enables an accurate coalescent-based estimation of the avian tree
Sivash Mirarab, M. Shamsuzzoha Bayzid, Bastien Boussau, Tandy Warnow

Gene tree incongruence arising from incomplete lineage sorting (ILS) can reduce the accuracy of concatenation-based estimations of species trees. Although coalescent-based species tree estimation methods can have good accuracy in the presence of ILS, they are sensitive to gene tree estimation error. We propose a pipeline that uses bootstrapping to evaluate whether two genes are likely to have the same tree, then it groups genes into sets using a graph-theoretic optimization and estimates a tree on each subset using concatenation, and finally produces an estimated species tree from these trees using the preferred coalescent-based method. Statistical binning improves the accuracy of MP-EST, a popular coalescent-based method, and we use it to produce the first genome-scale coalescent-based avian tree of life.

Species trees provide a basis for understanding how life evolved on earth, as well as having applications to comparative genomics, orthology detection, protein function inference, and biodiversity analysis. Estimations of species trees are typically built with multiple loci (sometimes complete genomes—but not always); in some cases, them throughout the genome. One advantage of such a phylogenomic approach is that it enables more data to be used in tree estimation (11). However, there is phylogenetic analysis may be complete genes, e.g., some may contain only the exons or only the introns of some gene, and some may not be based on genes at all. However, to be consistent with other literature on the subject (2, 10), we refer to phylogenetic trees on genomic loci as gene trees. Some of these summary methods are created on the basis of the multispecies coalescent model (30) and are statistically consistent under that model (31, 32), which means they will reconstruct the true species tree with poor phylogenetic signal (e.g., because of short sequence length) in individual genes, which is a potential problem for coalescent-based summary methods (21, 29). Moreover, many realistic biological conditions (including short branches in gene trees) make completely accurate gene tree estimation from limited sequence data highly unlikely (30).

Phylogenetic analyses can utilize very large numbers of genomic loci to estimate the species tree, but genome-scale data sets contain loci that have reduced phylogenetic signal so that their estimated gene trees have reduced bootstrap support (28) (31). Although it is not known how summary methods are affected when only some of the loci have low signal, coalescent-based summary methods have reduced accuracy on data sets where all the gene sequence alignments are short (39).

This challenge confronted the avian phylogenomics project (33), where a species tree estimated with a concatenated maximum likelihood analysis on 14,446 loci had a succession of short branches suggesting of a radiation, and the tree also conflicted with estimated gene trees. Furthermore, most loci had low phylogenetic signal, which resulted in average BS of only 32% for the bifurcating maximum-likelihood trees estimated on those loci (Fig. 3). Although much of the distance between estimated gene trees and the estimated species tree was related to the low support branches, even after collapsing low support branches, there was still substantial conflict among the gene trees (Fig. 3B and supplementary text), suggestive of ILS. Thus, not only is there gene


Citations as of Jan 12th: 16
Benefits

IDEALS Home → College of Engineering → Dept. of Computer Science → Warnow Research Group → Statistical binning improves species tree estimation in the presence of gene tree incongruence

Statistical binning improves species tree estimation in the presence of gene tree incongruence

Abstract:
Gene tree incongruence arising from incomplete lineage sorting (ILS) can reduce accuracy of concatenation-based estimations of species trees. While coalescent-based species tree estimation methods can have good accuracy in the presence of ILS, they are sensitive to gene tree estimation error. We propose a pipeline that uses bootstrapping to evaluate whether two genes are likely to have the same tree, then groups genes into sets using a graph-theoretic optimization and estimates a tree on each subset using concatenation, and finally produces an estimated species tree from these trees using the preferred coalescent-based method. Statistical binning improves the accuracy of MP-EST, a popular coalescent-based method, and we use it to produce the first genome-scale coalescent-based avian tree of life.

Article Citation:
Siavash Mirarab, Md Shamsuzzoha Bayzid, Bastien Boussau, and Tandy Warnow Science 33 December 2014: 546 (8215), 3230468 [DOI: 10.1126/science.1250468]

Datasets:
Thirteen datasets are available for download and use. Use the browse menu at right to access all of the datasets. Please refer to this README file for a description of the files included in this collection.

Mirarab, Siavash; Bayzid, Md Shamsuzzoha; Boussau, Bastien; Warnow, Tandy (2014) IDEALS http://dx.doi.org/10.13012/C5MW2F2P.

Downloads as of Sept 25th: 692
Starting with a Data Sharing Plan Can...

- **SAVE TIME**: Following a repository’s guidelines from the beginning of your project helps avoid extra processing and data clean-up later.
Common Deposit Options

- IDEALS: https://www.ideals.illinois.edu/
- Figshare: http://figshare.com/
- Hint: Use re3data (http://www.re3data.org/) to search or browse for repositories accepting data in your discipline
- Examples:
  - ICPSR: https://www.icpsr.umich.edu/icpsrweb/
  - tDAR: http://www.tdar.org/
Factors to Consider
File Types, Formats and Sizes
Disciplines or Areas of Research
Deposit Fees
Intellectual Property, Copyright, and Licensing
This seems like a combination of license concerns when you are using licensed data and license concerns when you deposit data. Should we separate them?

Hogenboom, Karen L, 2/18/2015
Copyright and Intellectual Property

• Copyright law does not apply to all data, but an incarnation of data (e.g., data in a table) is copyrightable

• Question is often who **owns** and **controls** the data
  – What are the PI’s restrictions on data? Is it yours to share?
  – Are you using data from a third party?
  – Are you working with an external party, particularly a commercial entity or industry partner?
Other Legal Considerations

- Student data
- Health information
- Patent

- Funding agency requirements
- Contract with funder of research
Privacy Concerns and Sensitive Data
Ethical Considerations

• Ensuring privacy of research subjects and confidentiality of the information they provide
• Informed consent
• Role of Institutional Research Board (IRB)
  – Protection of human subjects
  – Review of research projects and instruments
Ethical Considerations: Designing Studies for Confidentiality

• What data is necessary in order to answer the research question?
• Can the data be combined with other data to identify subjects?
• Can a sample be used rather than a population?
Ethical Considerations: Preparing Data for Sharing

- Strip identifiable information
- Decrease geographic granularity
- Generalize categories
- Top code
Data Access Restrictions Required?
Desired Data Use/Reuse Tracking and Statistics
How do I know whether it is a “good” repository?
How to Share/Publish Data

• **Assess** your project’s data sharing needs, requirements, restrictions.

• **Identify and evaluate** suitable data sharing/publication options.

• **Compare** individual project needs with features of data sharing/publication options, and select “best” option.
Objectives

1. Learn the benefits of making data available in order to identify how sharing can advance your research.

2. Assess the legal and ethical considerations that may apply to your data in order to make informed decisions about what data to share and how to share it.

3. Use tools to identify and evaluate data sharing options in order to guide your strategy for publishing data.
Contact

Karen Hogenboom
Numeric and Spatial Data Librarian
hogenboo@illinois.edu

Beth Sheehan
Social Work/Anthropology/Sociology Librarian
edivince@illinois.edu

Heidi Imker
Director, Research Data Service
imker@illinois.edu
Data Management Workshop Series

- **Introduction to Data Management**
  - 10-11 am, Thursday, October 1, 2015

- **Documentation and Organization for Data and Processes**
  - 10-11 am, Thursday, October 8, 2015

- **Making Research Data Public: Why, What, and How**
  - 10-11 am, Thursday, November 5, 2015
Scholarly Commons
http://www.library.illinois.edu/sc/index.html

Library Home Page
http://www.library.illinois.edu/

LibGuides
Built by librarians to help you through the research process. Search by subject, find your subject specialist librarian, browse for new guides, or use the Ask a Librarian chat service.
http://uiuc.libguides.com/index.php
Useful?
Tell your friends and colleagues!

Not Useful?
Tell us!