The Multispecies Network Coalescent and Phylogenetic Network Inference

16 June 2016
Phylogenomics Symposium and Software School
Austin Convention Center

Luay Nakhleh
Department of Computer Science
Rice university
All talks before noon today focused on species tree inference under the multispecies coalescent, where incomplete lineage sorting (ILS) is assumed to be the main cause of gene tree incongruence.
The multispecies coalescent (MSC)
The multispecies coalescent (MSC)

\[ P[((HC)G)] = 1 - \frac{2}{3} e^{-(T_2 - T_1)/N} \]
\[ P[((HG)C)] = \frac{1}{3} e^{-(T_2 - T_1)/N} \]
\[ P[((CG)H)] = \frac{1}{3} e^{-(T_2 - T_1)/N} \]
But, what if hybridization occurred and the phylogeny had short branches?
But, what if hybridization occurred and the phylogeny had short branches?

Gene tree incongruence due to hybridization and ILS!
The rest of this talk

- The multispecies network coalescent (MSNC)
- Phylogenetic network inference under MSNC
- Summary
The multispecies network coalescent (MSNC)
Phylogenetic networks

A leaf-labeled, rooted, directed, acyclic graph (rDAG)

Psi

tree node

reticulation node

branch lengths (coalescent units)
Multispecies network coalescent (MSNC)

Phylogenetic network with branch lengths (in coalescent units)

Inheritance probabilities, one per locus per reticulation node
Multispecies network coalescent (MSNC)
When gene tree $G_i$ is given by its topology alone:

$$P(G_i|\Psi, \Gamma) = \sum_{h \in H_\Psi(G_i)} P(h|\Psi, \Gamma)$$

$$P(h|\Psi, \Gamma) = \frac{w(h)}{d(h)} \prod_{b \in E(\Psi)} \frac{w_b(h)}{d_b(h)} \Gamma[b, j]^{u_b(h)} p_{u_b(h)v_b(h)}(\lambda_b)$$

[Yu, Degnan, Nakhleh, PLoS Genetics, 2012.]
Multispecies network coalescent (MSNC)

- When $G_i$ is given by its topology and branch lengths:

$$p(G_i | \Psi, \Gamma) = \sum_{h \in H_{\Psi}(G_i)} p(h | \Psi, \Gamma)$$

$$p(h | \Psi, \Gamma) = \prod_{b \in E(\Psi)} \left[ \prod_{i=1}^{|T_b(h)|-1} e^{-\frac{(u_b(h)-i+1)}{2}(T_b(h)_{i+1}-T_b(h)_i)} \right] \times e^{-\frac{u_b(h)}{2}(\tau_{\Psi}(b)-T_b(h)_{|T_b(h)|})} \times \Gamma[b, j]^{u_b(h)}$$

[12]

[Yu, Dong, Liu, Nakhleh, PNAS, 2014.]
Improving computation time

Fast algorithms and heuristics for phylogenomics under ILS and hybridization

Yun Yu*, Nikola Ristic, Luay Nakhleh*

From Eleventh Annual Research in Computational Molecular Biology (RECOMB) Satellite Workshop on Comparative Genomics
Lyon, France. 17-19 October 2013

A maximum pseudo-likelihood approach for phylogenetic networks

Yun Yu1 and Luay Nakhleh1,2*

*Correspondence:
nakhleh@rice.edu
1Department of Computer Science, Rice University, Houston, 77005 Texas, USA
Full list of author information is available at the end of the article

Abstract

Background: Several phylogenomic analyses have recently demonstrated the need to account simultaneously for incomplete lineage sorting (ILS) and hybridization when inferring a species phylogeny. A maximum likelihood approach is impractical for large data sets due to the combinatorially explosive number of possible reticulate evolutionary histories.

Results: Under the assumption of ILS, computing the probability of a gene tree given a species tree is a very hard problem. We present two exact algorithms for these two problems that speed up existing techniques significantly and enable analyses of much larger data sets.

Conclusion: The contributions of this paper are threefold. First, we improved computation time to the point where one can now infer phylogenetic networks for large data sets using our method. Second, we introduce a novel concept of weighted ancestral configurations on phylogenetic networks and parameterize our method to speed up the computation even further. Third, we introduce a novel concept of weighted ancestral configurations on phylogenetic networks and parameterize our method to speed up the computation even further.
Inference under MSNC
Inference under MSNC

Input: Sequence alignments for \( m \) (independent) loci

\[ S = \{ S_1, S_2, \ldots, S_m \} \]

Output: Phylogenetic network and inheritance probabilities

\((\Psi, \Gamma)\)
Inference under MSNC

The gene tree shortcut

Input: Gene trees for \( m \) (independent) loci

\[ G = \{g_1, g_2, \ldots, g_m\} \]

Output: Phylogenetic network and inheritance probabilities

\((\Psi, \Gamma)\)
An ML approach

Seek \((\Psi, \Gamma)\) that maximizes

\[
L(\Psi, \Gamma | \mathcal{G}) = \prod_{i=1}^{m} p(G_i | \Psi, \Gamma)
\]

Maximum likelihood inference of reticulate evolutionary histories

Yun Yu\textsuperscript{a,1}, Jianrong Dong\textsuperscript{a}, Kevin J. Liu\textsuperscript{a,b}, and Luay Nakhleh\textsuperscript{a,b,1}

Departments of \textsuperscript{a}Computer Science and \textsuperscript{b}Ecology and Evolutionary Biology, Rice University, Houston, TX 77005

Edited by David M. Hillis, The University of Texas at Austin, Austin, TX, and approved October 7, 2014 (received for review April 30, 2014)

Hybridization plays an important role in the evolution of certain groups of organisms, adaptation to their environments, and diversification of their genomes. The evolutionary histories of such groups are reticulate, and methods for reconstructing them are still in their infancy and have limited applicability. We present a maximum likelihood method for inferring reticulate evolutionary histories while accounting simultaneously for incomplete lineage

to the best of our knowledge, the first method to conduct a search of the phylogenetic network space in search of optimal phylogenies is described in a study by our group (18). However, this method is based on the maximum parsimony criterion: It seeks a phylogenetic network that minimizes the number of “extra lineages” resulting from embedding the set of gene tree topologies within its branches.
Searching the network space

endpoints
(potentially convergent)

search within
a layer

descending
a layer

ascending
a layer

multiple
starting points

[Yu, Dong, Liu, Nakhleh, PNAS, 2014.]
Illustration on a mosquito data set
Extensive introgression in a malaria vector species complex revealed by phylogenomics

Michael C. Fontaine,1,2,* James B. Pease,3* Aaron Steele,4 Robert M. Waterhouse,5,6,7,8 Daniel E. Neafsey,6 Igor V. Sharakhov,9,10 Xiaofang Jiang,10 Andrew B. Hall,10 Flaminia Catteruccia,11,12 Evodia Kakani,11,12 Sara N. Mitchell,12 Yi-Chieh Wu,5 Hilary A. Smith,1,2 R. Rebecca Love,1,2 Mara K. Lawniczak,13† Michel A. Slotman,14 Scott J. Emrich,2,4 Matthew W. Hahn,5,15§ Nora J. Besansky1,2§

*Present address: Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK.
†Corresponding author.
‡§Corresponding author.

In recent years, the importance of introgression in species evolution has increased. Introgression is the transfer of beneficial alleles between closely related species, which can lead to the creation of new species or the modification of existing ones. However, the extent and importance of introgression in the evolution of malaria vectors have not been well understood.

**Introduction**

The malaria vector Anopheles gambiae complex is a group of closely related species that are responsible for the transmission of Plasmodium falciparum, the most deadly form of malaria. Introgression in this complex has been suggested to play a role in the evolution of traits associated with vector competence and adaptation to the environment. However, the extent of introgression within the complex has not been well characterized.

**Methods**

To investigate introgression in the Anopheles gambiae complex, we sequenced 415 datasets from geographic locations across the range of the complex. We then used these datasets to infer the phylogenetic relationships within the complex and to identify regions of introgression.

**Results**

Our results indicate that extensive introgression has occurred within the Anopheles gambiae complex. Introgression has been observed in all regions of the genome, with the greatest extent of introgression occurring in the X chromosome. This suggests that introgression may have played a key role in the evolution of this species complex.

**Discussion**

The extent of introgression observed in the Anopheles gambiae complex is unprecedented. This suggests that introgression may have been a widespread phenomenon in the evolution of this group of species. However, the role of introgression in the evolution of specific traits associated with vector competence remains to be determined.

**Conclusion**

In conclusion, our study provides evidence for extensive introgression in the Anopheles gambiae complex. This suggests that introgression may have played a key role in the evolution of this species complex and should be considered in future studies of vector evolution.
Extensive introgression in a malaria vector species complex revealed by phylogenomics


Breeding in brackish coastal waters of eastern and western Africa, respectively, are minor vectors. *Anopheles quadrimaculatus* plays no role in malaria transmission despite vector competence for *Plasmodium falciparum*, as it tends to bite animals other than humans. On the basis of ecology and behavior, one might predict phylogenetic clustering of the highly anthropophilic vector species. Yet such clustering has not been supported by chromosomal inversion polymorphisms (10, 14). The apparent phylogenetic affinity between *An. arabiensis* and *An. gambiae* supported by molecular markers and shared chromosomal inversion polymorphisms was instead attributed to introgression (15, 16). Introgression is plausible between any geographically overlapping pair of species in the complex, as reproductive isolation is incomplete: Adult female F1 hybrids—although uncommon in nature—are fertile and vigorous (only F1 hybrid males are sterile) (10). Nonethe-
Extensive introgression in a malaria vector species complex revealed by phylogenomics

Michael C. Fontaine,1,2,4 James B. Pease,3 Aaron Steele,6 Robert M. Waterhouse,5,6,7,8 Daniel E. Neafsey,6 Igor V. Sharakhov,9,10 Xiaofang Jiang,10 Andrew B. Hall,10 Flaminia Catteruccia,11,12 Evdoxia Kakani,11,12 Sara N. Mitchell,13 Yi-Chieh Wu,5 Hilary A. Smith,1,2 Rebecca Love,1,2 Mara K. Lawniczak,13,14 Michel A. Slotman,14 Scott J. Emrich,2,4 Matthew W. Hahn,5,13,15 Nora J. Besansky1,4

**MOSQUITO GENOMICS**

breed in brackish coastal waters of eastern and western Africa, respectively, are minor vectors. *Anopheles quadrimaculatus* plays no role in malaria transmission despite vector competence for *Plasmodium falciparum*, as it tends to bite animals other than humans. On the basis of ecology and behavior, one might predict phylogenetic clustering of the highly anthropophilic vector species. Yet such clustering has not been supported by chromosomal inversion phylogenies (10, 14). The apparent phylogenetic affinity between *An. arabiensis* and *An. gambiae* supported by molecular markers and shared chromosomal inversion polymorphisms was instead attributed to introgression (15, 16). Introgression is plausible between any geographically overlapping pair of species in the complex, as reproductive isolation is incomplete: Adult female F1 hybrids—although uncommon in nature—are fertile and vigorous (only F1 hybrid males are sterile) (10). Nonethe-
There are two possible reasons that the overall pattern of divergence of genomic segments may differ from the overall species tree, which has one of the more rarely supported topologies genome-wide. First, there may be introgression when crosses between species produce fertile female hybrids that result in gene flow between species. Second, there may be incomplete lineage sorting, whereby an ancestral population splits into two daughter species and each becomes fixed for a different allele that was segregating at a previously polymorphic locus in the ancestral species. The local phylogeny at such a locus inferred in the two daughter species may then not necessarily be consistent with the branching pattern of the species tree.
There are two possible reasons that the overall pattern of divergence of genomic segments may differ from the overall species tree, which has one of the more rarely supported topologies genome-wide. First, there may be introgression when crosses between species produce fertile female hybrids that result in gene flow between species. Second, there may be incomplete lineage sorting, whereby an ancestral population splits into two daughter species and each becomes fixed for a different allele that was segregating at a previously polymorphic locus in the ancestral species. The local phylogeny at such a locus inferred in the two daughter species may then not necessarily be consistent with the branching pattern of the species tree.
Abstract

The role of hybridization and subsequent introgression has been demonstrated in an increasing number of species. Recently, Fontaine et al. (Science, 347, 2015, 1258524) conducted a phylogenomic analysis of six members of the Anopheles gambiae species complex. Their analysis revealed a reticulate evolutionary history and pointed to extensive introgression on all four autosomal arms. The study further highlighted the complex evolutionary signals that the co-occurrence of incomplete lineage sorting (ILS) and introgression can give rise to in phylogenomic analyses. While tree-based methodologies were used in the study, phylogenetic networks provide a more natural model to capture reticulate evolutionary histories. In this work, we reanalyze the Anopheles data using a recently devised framework that combines the multispecies coalescent with phylogenetic networks. This framework allows us to capture ILS and introgression simultaneously, and forms the basis for statistical methods for inferring reticulate evolutionary histories. The new analysis reveals a phylogenetic network with multiple hybridization events, some of which differ from those reported in the original study. To elucidate the extent and patterns of introgression across the genome, we devise a new method that quantifies the use of reticulation branches in the phylogenetic network by each genomic region. Applying the method to the mosquito data set reveals the evolutionary history of all the chromosomes. This study highlights the utility of ‘network thinking’ and the new insights it can uncover, in particular in phylogenomic analyses of large data sets with extensive gene tree incongruence.

Keywords: Anopheles gambiae, hybridization, incomplete lineage sorting, introgression, phylogenetic networks

Introduction

In a recent study, Fontaine et al. (2015) conducted phylogenomic analyses of the species complex including the malaria vector, Anopheles gambiae. The authors reported a reticulate evolutionary history of this group, including extensive introgression patterns across all four autosomal chromosome arms. They inferred a species tree based on the X chromosome and used information on sequence divergence from the autosomes to hypothesize three hybridization events. This study of recently diverged species highlighted two processes that can be at play during evolution and must be accounted for in phylogenomic analyses. On the one hand, the low levels of divergence mean that species can hybridize and that their genomes may carry introgressed genetic material. On the other hand, the short times between speciation events mean that incomplete lineage sorting (ILS) is likely to occur. Phylogenomic
You can compute the likelihood of a phylogenetic network hypothesis...

\[ \ln L = -13,738.60 \]
You can search for the best augmentation of a tree into a phylogenetic network...

\[ \ln L = -13,681.91 \]
You can search for the network from scratch while controlling the number of reticulations...

\[
\ln L = -13,996.64
\]

\[
\ln L = -13,778.52
\]

\[
\ln L = -13,706.11
\]

\[
\ln L = -13,659.70
\]
You can infer the evolutionary paths within the network of individual genomic regions...
We are interested in obtaining a sample of networks (rather than a single point estimate) penalizing model complexity (to avoid overfitting) specifying knowledge from the biology of hybridization
Bayesian Inference of Reticulate Phylogenies under the Multispecies Network Coalescent

Dingqiao Wen¹*, Yun Yu¹, Luay Nakhleh¹,²*

¹ Computer Science, Rice University, Houston, Texas, United States of America, ² BioSciences, Rice University, Houston, Texas, United States of America

* dw20@rice.edu (DW); nakhleh@rice.edu (LN)

Abstract

The multispecies coalescent (MSC) is a statistical framework that models how gene genealogies grow within the branches of a species tree. The field of computational phylogenetics has witnessed an explosion in the development of methods for species tree inference under...
Summary

- Reticulations create evolutionary histories that are best modeled by phylogenetic networks.
- Other discord processes, such as incomplete lineage sorting (ILS), compound the problem of detecting true reticulations.
- The multispecies network coalescent (MSNC) allows for modeling both and teasing them apart.
- We have developed parsimony, likelihood, and Bayesian approaches to phylogenetic network inference.
PhyloNet

❖ All the Methods are implemented in PhyloNet (open source, JAVA):
  ❖ http://bioinfo.cs.rice.edu/phylonet

Tutorial by Yun Yu tomorrow!
Acknowledgments


❖ **Funding:** NIH, NSF, Guggenheim and Sloan Foundations
Thank you

http://bioinfo.cs.rice.edu/