Detecting Evolutionary Inter-Gene Heterogeneity in *Borrelia burgdorferi*

ELISA LOZA  
Department of Mathematical Sciences  
University of Bath
Contents

1. What is a *phylogenetic analysis*?
Contents

1. What is a *phylogenetic analysis*?

2. Conventional (homogeneous) model for likelihood-based phylogenetic inference.
1. What is a *phylogenetic analysis*?

2. Conventional (homogeneous) model for likelihood-based phylogenetic inference.

3. Downsides of the homogeneous model.
Contents

1. What is a phylogenetic analysis?

2. Conventional (homogeneous) model for likelihood-based phylogenetic inference.

3. Downsides of the homogeneous model.

4. An improved model that accounts for heterogeneity.
1. What is a phylogenetic analysis?

2. Conventional (homogeneous) model for likelihood-based phylogenetic inference.

3. Downsides of the homogeneous model.

4. An improved model that accounts for heterogeneity.

5. Applications to Borrelia burgdorferi data.
Phylogenetic likelihood methods

• Phylogenetics is the reconstruction and analysis of trees and other parameters to describe and understand the evolution of organisms.
Phylogenetic likelihood methods

- Phylogenetics is the reconstruction and analysis of trees and other parameters to describe and understand the evolution of organisms.

- Likelihood-based phylogenetic analyses start by observing the aligned DNA sequences of $s$ organisms:

  TCAAGCTATACCCGAT...
  TATACCAGCTATAGCT...
  CAAAGCTATACCCGAT...
  CAAAGCTATACCCGAT...
  ...
The homogeneous model

- The homogeneous model for independent observations $y_1 = (T,T,C,C)'$, $y_2 = (C,A,A,A)'$, …, $y_n = (T,A,A,T)'$, is:

$$y_i \sim f( \cdot | \lambda, t, Q) \text{ independently for } i = 1, 2, \ldots, n$$
Model parameters

\[ y_i \sim f(\cdot | \chi, t, Q) \text{ independently for } i = 1, 2, \ldots, n \]
Model parameters

\[ y_i \sim f(\cdot \mid x, t, Q) \] independently for \( i = 1, 2, \ldots, n \)

- A bifurcating tree with \( s \) leaves,
Model parameters

\[ y_i \sim f(\cdot | \chi, t, Q) \quad \text{independently for } i = 1, 2, \ldots, n \]

- A bifurcating tree with \( s \) leaves,

- A set of positive real-valued branch lengths,

\[ t = (t_1, t_2, \ldots, t_5) \]
Model parameters

\[ y_i \sim f( \cdot \mid \mathbf{X}, \mathbf{t}, Q) \quad \text{independently for } i = 1, 2, \ldots, n \]

- A bifurcating tree with \( s \) leaves,

- A set of positive real-valued branch lengths,
  \[ \mathbf{t} = (t_1, t_2, \ldots, t_5) \]

- A rate matrix \( Q \) specifying a Markov process of character substitution along
DNA data may be not homogeneous
DNA data may be not homogeneous
DNA data may be not homogeneous
DNA data may be not homogeneous
Borrelia burgdorferi

• *Borrelia burgdorferi* is one of the bacterial species responsible for Lyme disease.
Borrelia burgdorferi

• *Borrelia burgdorferi* is one of the bacterial species responsible for Lyme disease.

• To fully understand the disease, it is crucial to unveil the evolutionary properties of its genetic variants (strains).
Borrelia burgdorferi

- *Borrelia burgdorferi* is one of the bacterial species responsible for Lyme disease.

- To fully understand the disease, it is crucial to unveil the evolutionary properties of its genetic variants (strains).

- Phylogenetic analysis is an essential tool.
Identification of *B. burgdorferi* strains
Identification of *B. burgdorferi* strains
Identification of *B. burgdorferi* strains
Are the loci congruent in evolution, such that valid inferences can be made under a homogeneous phylogenetic model?
The Q + t mixture model

- Finite mixture models provide a natural way to model heterogeneous data.

\[ f(\cdot | x, t, Q) \]
The $Q + t$ mixture model

- Finite mixture models provide a natural way to model heterogeneous data.

$$f(\cdot | \lambda, t, Q) + f(\cdot | \lambda, t, Q)$$
The Q + t mixture model

- Finite mixture models provide a natural way to model heterogeneous data.

\[ f(\cdot | \lambda, t, Q) + f(\cdot | \lambda, t, Q) + \ldots + f(\cdot | \lambda, t, Q) \]
The Q + t mixture model

• Finite mixture models provide a natural way to model heterogeneous data.

\[ w f( \cdot | \lambda, t, Q) + w f( \cdot | \lambda, t, Q) + \ldots + w f( \cdot | \lambda, t, Q) \]

for \( w + w + \ldots + w = 1 \)
The Q + t mixture model

- Finite mixture models provide a natural way to model heterogeneous data.

\[ y_i \sim w f(\cdot | \lambda, t, Q) + w f(\cdot | \lambda, t, Q) + \ldots + w f(\cdot | \lambda, t, Q) \]

for \( w + w + \ldots + w = 1 \)

and ind. for \( i = 1, 2, \ldots, n \)
A branch-length mixture model

\[ \gamma_i \sim w f(\cdot | \chi, t, Q) + w f(\cdot | \chi, t, Q) + \ldots + w f(\cdot | \chi, t, Q) \]

for \( w + w + \ldots + w = 1 \)

and ind. for \( i = 1, 2, \ldots, n \)
A branch-length mixture model

\[ \gamma_i \sim w f(\cdot | \chi, t, Q) + w f(\cdot | \chi, t, Q) + \ldots + w f(\cdot | \chi, t, Q) \]

for \( w + w + \ldots + w = 1 \)
and ind. for \( i = 1, 2, \ldots, n \)
A branch-length mixture model

\[ \gamma_i \sim w f( \cdot | \chi, t, Q) + w f( \cdot | \chi, t, Q) + \ldots + w f( \cdot | \chi, t, Q) \]

for \( w + w + \ldots + w = 1 \)

and ind. for \( i = 1, 2, \ldots, n \)
A branch-length mixture model

\[ \gamma_i \sim \sum w f(\cdot | \chi, t, Q) \]

for \( w + w + \ldots + w = 1 \)

and ind. for \( i = 1, 2, \ldots, n \)
A branch-length mixture model

\[ \gamma_i \sim w f(\cdot | \lambda, t, Q) + w f(\cdot | \lambda, t, Q) + \ldots + w f(\cdot | \lambda, t, Q) \]

for \( w + w + \ldots + w = 1 \)
and ind. for \( i = 1, 2, \ldots, n \)
The $Q + t$ mixture model

$$y_i \sim w f(\cdot | \chi, t, Q) + w f(\cdot | \chi, t, Q) + \ldots + w f(\cdot | \chi, t, Q)$$

for $w + w + \ldots + w = 1$

and ind. for $i = 1, 2, \ldots, n$
The $Q + t$ mixture model

• A label$_i$ identifies the specific process from which the $i$-th site is generated.
The Q + t mixture model

- A label$_i$ identifies the specific process from which the $i$-th site is generated.

$$p \left( \text{label}_i = \square \right) = \odot$$

independently for $i = 1, 2, ..., n$
The Q + t mixture model

- A label \( i \) identifies the specific process from which the \( i \)-th site is generated.

\[
p(\text{label}_i = \square) = \infty
\]

independently for \( i = 1, 2, \ldots, n \)
The $Q + t$ mixture model

• A $\text{label}_i$ identifies the specific process from which the $i$-th site is generated.

\[ p(\text{label}_i = \square) = \infty \]

independently for $i = 1, 2, \ldots, n$
The $Q + t$ mixture model

- A $\text{label}_i$ identifies the specific process from which the $i$-th site is generated.

$$p \left( \text{label}_i = \square \right) = \varnothing \quad \text{for} \quad \square = \:\square_1, \:\square_2, \: \ldots, \:\square_n$$

independently for $i = 1, 2, \ldots, n$
The $Q + t$ mixture model

- Once the label for site $i$ is known,

$$y_i \mid \square \sim f(\cdot \mid \chi, t, Q)$$

independently for $i = 1, 2, \ldots, n$
The $Q + t$ mixture model

- Once the label $i$ for site $i$ is known,

$$y_i | \mathbf{\cdot} \sim f(\cdot | \chi, t, Q)$$

independently for $i = 1, 2, \ldots, n$
The Q + t mixture model

- Once the label for site $i$ is known,

\[ y_i | \mathbf{\Theta} \sim f( \cdot | \lambda, t, Q ) \]

independently for $i = 1, 2, \ldots, n$
The Q + t mixture model: an example

• Consider a DNA alignment:
The $Q + t$ mixture model: an example

- Consider a DNA alignment:

- Sites are modelled by:

$$
\gamma_i \sim w \ f( \cdot | \chi, t, Q) + w \ f( \cdot | \chi, t, Q)
$$

independently for $i = 1, 2, \ldots, n$
The $Q + t$ mixture model: an example

- Consider a DNA alignment:
The \( Q + t \) mixture model: an example

- Consider a DNA alignment:
The Q + t mixture model: an example

• Consider a DNA alignment:
The Q + t mixture model: an example

- Consider a DNA alignment:
The $Q + t$ mixture model: an example

- Consider a DNA alignment:
The Q + t mixture model: an example

- Consider a DNA alignment:
The Q + t mixture model: an example

- Consider a DNA alignment:
The $Q + t$ mixture model: an example

$$y_i \mid \square \sim f(\cdot \mid \chi, t, Q)$$

independently for $i = 1, 2, \ldots, m$
The $Q + t$ mixture model: an example

$y_i | \pi \sim f(\cdot | \lambda, t, Q)$

independently for $i = m+1, \ldots, n$
Analysis of B. burgdorferi: the ‘housekeeping genes’ alignment
Analysis of *B. burgdorferi*: the ‘housekeeping genes’ alignment
Analysis of *B. burgdorferi*: the ‘housekeeping genes’ alignment
Analysis of *B. burgdorferi*: the ‘housekeeping genes’ alignment
Analysis of B. burgdorferi: the ‘housekeeping genes’ alignment
Analysis of *B. burgdorferi*: the ‘housekeeping genes’ alignment
Analysis of *B. burgdorferi*: the ‘housekeeping genes’ alignment
Analysis of *B. burgdorferi*: the ‘housekeeping genes’ alignment

Site classification probabilities

Site number
Analysis of *B. burgdorferi*: the ‘housekeeping genes’ alignment

Posterior densities of stationary frequencies
Analysis of *B. burgdorferi*: the ‘housekeeping genes’ alignment

Posterior densities of substitution rates

- $r_{AC}$
- $r_{AG}$
- $r_{AT}$
- $r_{CG}$
- $r_{CT}$
- $r_{GT}$
Analysis of *B. burgdorferi*: the ‘housekeeping genes’ alignment

Posterior densities of branch lengths
Analysis of *B. burgdorferi*: the ‘housekeeping g. |ospC’ alignment
Analysis of *B. burgdorferi*: the ‘housekeeping g. | ospC’ alignment
Analysis of *B. burgdorferi*: the ‘housekeeping g. ospC’ alignment
Analysis of B. burgdorferi: the ‘housekeeping g. ospC’ alignment
Analysis of *B. burgdorferi*: the ‘housekeeping g. ospC’ alignment
Analysis of *B. burgdorferi*: the ‘housekeeping g.|ospC’ alignment

Site classification probabilities

Site number
Analysis of *B. burgdorferi*: the ‘housekeeping g. ospC’ alignment

Posterior densities of stationary frequencies

- $\pi_A$
- $\pi_C$
- $\pi_G$
- $\pi_T$
Analysis of B. burgdorferi: the ‘housekeeping g. | ospC’ alignment

Posterior densities of substitution rates
Analysis of *B. burgdorferi*: the ‘housekeeping g.|ospC’ alignment

Posterior densities of branch lengths

---

**Histograms of Branch Lengths**

- **Interior Branch Lengths**
  - The histogram shows the distribution of interior branch lengths. The mode is around 1.5, with a peak density of 3.

- **Terminal Branch Lengths**
  - The histogram illustrates the distribution of terminal branch lengths. The mode is slightly lower than the interior branch lengths, around 1.0, with a peak density of 2.
Conclusions

- A more realistic phylogenetic model that accommodates heterogeneity.
Conclusions

• A more realistic phylogenetic model that accommodates heterogeneity.

• The $Q+t$ mixture model automatically recovers the evolutionary identity of a site.
Conclusions

• A more realistic phylogenetic model that accommodates heterogeneity.

• The $Q+t$ mixture model automatically recovers the evolutionary identity of a site.

• It is a suitable indicator of evolutionary homogeneity or heterogeneity among large-scale concatenations of genes.
Conclusions

• It is relevant testing for homogeneity as a concatenation of genes will produce valid inferences only when there is evolutionary congruence.
Conclusions

• It is relevant testing for homogeneity as a concatenation of genes will produce valid inferences only when there is evolutionary congruence.

• *B. burgdorferi* data is just one application of many other possibilities.
Acknowledgements

- Merrilee Hurn, Mathematical Sciences
- Tony Robinson, Mathematical Sciences
- Gabi Margos, Biology and Biochemistry
- Klaus Kurtenbach, Biology and Biochemistry