From Genes to Genomes and Beyond: a Computational Approach to Evolutionary Analysis

Kevin J. Liu, Ph.D.
Rice University Dept. of Computer Science
Evolution: Unifying Theme

#1

• “Nothing In Biology Makes Sense Except in the Light of Evolution” – 1973 essay by T. Dobzhansky, a famous biologist

• Overarching goal: use evolutionary principles to:
  - Create computational methodology to analyze heterogeneous large-scale biological data,
  - Then apply findings to obtain new biological and biomedical discoveries
Big Data: Unifying Theme #2

Number of sequences

1000+

Less than 1000

Sequence length

Single gene or a few genes

Entire genome

Sequence length
Big Data: Unifying Theme #2

Number of sequences

1000+

Less than 1000

Graduate work: SATé, SATé-II, DACTAL, etc.

Big Data: Unifying Theme #2

Number of sequences

1000+
Less than 1000

Graduate work:
SATé, SATé-II, DACTAL, etc.

Single gene or a few genes

Postgraduate work:
PhyloNet-HMM, etc.

Entire genome

Sequence length
Big Data: Unifying Theme #2

- **Number of sequences**
  - 1000+
  - Less than 1000

- **Sequence length**
  - Single gene or a few genes
  - Entire genome

**Graduate work**:
- SATé,
- SATé-II,
- DACTAL,
- etc.

**Postgraduate work**:
- PhyloNet-HMM,
- etc.
The Spread of Warfarin Resistance Between Two Mouse Species

Report

Adaptive Introgression of Anticoagulant Rodent Poison Resistance by Hybridization between Old World Mice

Ying Song, Stefan Endepols, Nicole Klemann, Dania Richter, Franz-Rainer Matuschka, Ching-Hua Shih, Michael W. Nachman, and Michael H. Kohn

1Department of Ecology and Evolution Biology, Rice University, Houston, TX 77005, USA
2Environmental Science, Bayer CropScience AG, D-40789 Monheim, Germany
3D-48231 Warendorf, Germany
4Division of Pathology, Department of Parasitology, Charité–Universitätsmedizin, D-10117 Berlin, Germany
5Department of Ecology and Evolution, University of Arizona, Tucson, AZ 85721, USA

The emergence of warfarin resistance in humans and rodents in response to exposure to anticoagulants [2]; additional SNPs in vkc1 await such experimental proof. A mere ~10 years after the inception of warfarin as a rodenticide in the 1950s, reports of resistant Norway rats (Rattus norvegicus) emerged between 1960 and 1969, followed by reports of resistant house mice (Mus musculus spp.) in 1964, roof rats (R. rattus) in 1972, and other rat species (e.g., R. tiomanicus, R. r. diardii, and R. lusea) [3, 8–10]. Resistant rodent colonies have been discovered in Europe, the Americas, Asia, and Australia [8]. In response to such warfarin-resistant colonies, other anticoagulant rodenticides were developed that target VKOR, including coumatetralyl,
Warfarin and Adverse Events

- Warfarin is the most widely prescribed blood thinner
- Treatment is complicated because every patient is different
  - Gene mutations confer resistance or susceptibility
- Annually,
  - 85,000 serious bleeding events
  - 17,000 strokes
  - Cost: $1.1 billion

McWilliam et al. AEI-Brookings Joint Center 2006.
The \textit{Vkorcl} Gene and Personalized Warfarin Therapy

- Mutant \textit{Vkorcl} gene contributes to warfarin resistance
- Warfarin resistant individuals require larger-than-normal dose to prevent clotting complications (like stroke)

Warfarin is Really Glorified Rodent Poison

Reproduced from UTMB.
Recasting the Study of Introgression as a Computational Question

• Humans inadvertently started a gigantic drug trial by giving warfarin to mice in the wild

• Mice shared genes (including one that confers warfarin resistance) to survive (Song et al. 2011)
  - Gene sharing occurred between two different species (introgression)

• To find out results from the drug trial, we just need to analyze the genomes of introgressed mice and locate the introgressed genes
Related Applications

• Similar computational approaches can be used to study gene flow between species in other contexts
  - Constitutes basic research of interest to the NSF

• Wide range of applications of interest to different funding agencies, including:
  - The role of horizontal gene transfer in the spread of antibiotic resistance in bacteria (NIH)
  - Metabolism of hybrid yeast species, with applications in metabolic engineering (DOE)
  - Disease resistance of hybrid plant species (USDA)
**Problem:** Computational Introgression Detection

**Input:**

<table>
<thead>
<tr>
<th>Species</th>
<th>Genome ID</th>
<th>Introgressed?</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>x</td>
<td>Unknown</td>
</tr>
<tr>
<td>A</td>
<td>a₁</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>aₖ</td>
<td>No</td>
</tr>
<tr>
<td>B</td>
<td>b₁</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>bₐ</td>
<td>No</td>
</tr>
</tbody>
</table>
### Problem: Computational Introgression Detection

#### Input:

<table>
<thead>
<tr>
<th>Species</th>
<th>Genome ID</th>
<th>Introgressed?</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>x</td>
<td>Unknown</td>
</tr>
<tr>
<td>A</td>
<td>a₁</td>
<td>No</td>
</tr>
<tr>
<td>A</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>A</td>
<td>aₖ</td>
<td>No</td>
</tr>
<tr>
<td>B</td>
<td>b₁</td>
<td>No</td>
</tr>
<tr>
<td>B</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>B</td>
<td>bₙ</td>
<td>No</td>
</tr>
</tbody>
</table>

#### Output:

- Probability that x contains introgressed material from species B
Naïve Sliding Windows

1. Break the genome into segments using a sliding-window (or other approaches)

2. Estimate a local tree in between every pair of breakpoints
Sliding Windows (Example)
Sliding Windows (Example)
Sliding Windows (Example)
Sliding Windows (Example)

Gene tree incongruence!
“Horizontal” Gene Tree Incongruence (Example)
“Horizontal” Gene Tree Incongruence (Example)
Sliding Windows: Results

Vkorc1 gene

Hybrid origin?

Yes

No

Position along chromosome 7 (megabases)

Percentage of hybrid origin

Chromosome
Sliding Windows Approach Is Too Simplistic

- Gene tree incongruence can occur for reasons other than introgression
- The organisms in our study included “vertical” gene tree incongruence due to:
  - Incomplete lineage sorting
  - Recombination
“Vertical” Gene Tree Incongruence (Example)
How to Disentangle “Horizontal” and “Vertical” Gene Tree Incongruence?
Insight from Meng and Kubatko (2009)
Insight from Meng and Kubatko (2009)

“Pull apart” species network into two “parental trees”
Disentangling “Horizontal” and “Vertical” Gene Tree Incongruence
Disentangling “Horizontal” and “Vertical” Gene Tree Incongruence
Insight #1

• “Horizontal” and “vertical” incongruence between neighboring gene trees represent two different types of dependence

• Model the two dependence types using two classes of transitions in a graphical model
Insight #2

• DNA sequences are observed, not gene trees

• Under traditional models of DNA sequence evolution, the probability $P[s|g]$ of observing DNA sequences $s$ given a gene tree $g$ can be efficiently calculated using dynamic programming (Felsenstein’s pruning algorithm)
Insight #1 + Insight #2 = Use a Hidden Markov Model (HMM)
PhyloNet-HMM: Hidden States

- **s₀**
- **q₁**
- **q₂**
- **q₃**
- **r₁**
- **r₂**
- **r₃**

- **Non-introgressed**
- **Introgressed**
PhyloNet-HMM

• Each hidden state $s_i$ is associated with a gene tree $g(s_i)$ contained within a “parental” tree $f(s_i)$

• The set of HMM parameters $\lambda$ consists of
  
  ‣ The initial state distribution $\pi$
  ‣ Transition probabilities $a_{ij} = \begin{cases} P[g(s_i)|f(s_i)] \cdot \gamma & \text{if } s_i \text{ and } s_j \text{ in different rows} \\ P[g(s_i)|f(s_i)] \cdot (1 - \gamma) & \text{if } s_i \text{ and } s_j \text{ in same row} \end{cases}$

  where $\gamma$ is the “vertical” parental tree switching frequency and

  $\Pr[g(s_i)|f(s_i)]$ is calculated using formula of Degnan and Salter (2005)

  ‣ The emission probabilities $b_i = \Pr[O_i|g(s_i)]$

• Use a model of nucleotide substitution like Jukes-Cantor (1969)
Three HMM-related Problems

1. What is the likelihood of the model given the observed DNA sequences?

2. Which sequence of hidden states best explains the observed DNA sequences?

3. How do we choose parameter values that maximize the model likelihood?
First HMM-related Problem

• Let $q_t$ be PhyloNet-HMM’s hidden state at time $t$, where $1 \leq t \leq k$ and $k$ is the length of the input observation sequence $O$.

• What is the likelihood of the model given the observed DNA sequences $O$?
  
  - Forward algorithm calculates “prefix” probability $\alpha_t(i)$
  
  - Backward algorithm calculates “suffix” probability $\beta_t(i)$

  - Model likelihood is $P[O|\lambda] = \sum_{i=1}^{N} \alpha_k(i)$. 
Second HMM-related Problem

- Which sequence of states best explains the observation sequence?

- Posterior decoding probability $\gamma_t(i)$ is the probability that the HMM is in state $s_i$ at time $t$, which can be computed as:

$$\gamma_t(i) = \frac{\alpha_t(i) \beta_t(i)}{P[O|\lambda]}.$$
Third HMM-related Problem

• How do we choose parameter values that maximize the model likelihood?
  - Perform local search to optimize the criterion
    \[
    \arg\max_\lambda P[O|\lambda]
    \]
Related HMM-based Approaches

• CoalHMM (Mailund et al. 2012)
  - Models introgression + incomplete lineage sorting + recombination (with a simplifying assumption)
  - Currently supports two sequences only
  - Assumes that time is discretized

• Other approaches that don’t account for introgression (e.g., Hobolth et al. 2007)
PhyloNet-HMM Scan of Chromosome 7

PhyloNet-HMM Scan of Whole Genome

(a) Percentage with introgressed origin (%)

(b) Total length with introgressed origin (Mb)
Scaling PhyloNet-HMM

• Previous analyses (at most five genomes and a single introgression event) required more than a CPU-month on a large cluster

• Problem is combinatorial in both the number of genomes and the number of introgression events

• Challenge: efficient and accurate introgression detection from hundreds of genomes or more
Big Data: Unifying Theme #2

Number of sequences

- Less than 1000
- 1000+

Sequence length

- Single gene or a few genes
- Entire genome

Graduate work:
- SATé, SATé-II, DACTAL, etc.

Postgraduate work:
- PhyloNet-HMM, etc.
SATé: Simultaneous Alignment and Tree estimation (Liu et al. Science 2009)

- Standard methods for alignment and tree estimation have unacceptably high error and/or cannot analyze large datasets
- SATé is more accurate than all existing methods on datasets with up to thousands of taxa
- 24 hour analyses using standard desktop computer
- SATé-II (Liu et al. Systematic Biology 2011) is more accurate and faster than SATé on datasets with up to tens of thousands of taxa using a standard desktop computer
DNA Sequence Evolution (Example)

Substitutions

AAGG CTT

AAGACTT

-3 mil yrs

-2 mil yrs

-1 mil yrs
today
The true alignment is:

...ACGGTG\textcolor{magenta}{CAGTT}ACCA...  

...AC\textcolor{red}{CAGT}CACC\textcolor{blue}{CATAGA}...
DNA Sequence Evolution (Example)

Substitutions

- ATCGGGCAT
- TAGCCCA
- TAGACTT
- AGCA
- AGCG

Insertions

- AAGACTTT

Deletions

- 3 mil yrs
- 2 mil yrs
- 1 mil yrs
- today
DNA Sequence Evolution (Example)

-3 mil yrs

-2 mil yrs

-1 mil yrs

today

ATCGGGCAT

TAGCCCA

TAGACTT

AGCA

AGCG

AAGGCTT

AAGACTT

TGGACTT

AGCG
Tree and Alignment Estimation Problem (Example)

\[
\begin{align*}
u &= \text{ATCTGGGCAT} \\
v &= \text{TAGCCCA} \\
w &= \text{TAGACTT} \\
x &= \text{AGCA} \\
y &= \text{AGCG}
\end{align*}
\]
Many Trees

- Number of trees $|T|$ grows exponentially in the number of species $n$
  \[ |T| = (2n - 5)! \]
- NP-hard optimization problems
Two-phase Methods

\[
\begin{align*}
  u &= \text{AGGCTATCACCTGACCTCCA} \\
  v &= \text{TAGCTATCAGCCCGC} \\
  w &= \text{TAGCTGACCAGC} \\
  x &= \text{TCACGACCAGAC}
\end{align*}
\]

Phase 1: Align

\[
\begin{align*}
  u &= \text{-AGGCTATCACCTGACCTCCA} \\
  v &= \text{TAG-CTATCAGCCCGC} \\
  w &= \text{TAG-CTGACCAGC} \\
  x &= \text{-TCACGACCAGAC}
\end{align*}
\]

Phase 2: Estimate Tree
Many Methods

Alignment method

- ClustalW
- MAFFT
- Muscle
- Prank
- Opal
- Probcons (and Probtree)
- Di-align
- T-Coffee
- Etc.
Many Methods

Alignment method
- ClustalW
- MAFFT
- Muscle
- Prank
- Opal
- Probcons (and Probtree)
- Di-align
- T-Coffee
- Etc.

Phylogeny method
- Maximum likelihood (ML)
  - RAxML
- Bayesian MCMC
- Maximum parsimony
- Neighbor joining
- UPGMA
- Quartet puzzling
- Etc.
Simulation Study
(Liu et al. Science 2009)

Simulation using ROSE

- Model trees with 1000 taxa
- Biologically realistic model with:
  - Varied rates of substitutions
  - Varied rates of insertions and deletions
  - Varied gap length distribution
    - Long
    - Medium
    - Short
False Negative (FN): an edge in the true tree that is missing from the estimated tree

Missing branch rate: the percentage of edges present in the true tree but missing from the estimated tree
Alignment Error

- **False Negative (FN):** pair of nucleotides present in true alignment but missing from estimated alignment

- **Alignment SP-FN error:** percentage of paired nucleotides present in true alignment but missing from estimated alignment
Results

1000 taxon models ranked by difficulty

- ML(ClustalW)
- ML(Muscle)
- ML(Prank+GT)
- ML(MAFFT)
- ML(TrueAln)
Problem with Two-phase Approach

• **Problem**: two-phase methods fail to return reasonable alignments and accurate trees on large and divergent datasets
  - manual alignment
  - unreliable alignments excluded from phylogenetic analysis
Simultaneous Estimation of a Tree and Alignment

\[ u = \text{AGGCTATCACCTGACCTCCA} \]
\[ v = \text{TAGCTATCACGACCGC} \]
\[ w = \text{TAGCTGACCGC} \]
\[ x = \text{TCACGACCGACA} \]
Simultaneous Estimation Methods

• Methods based on statistical models
  - Limited to datasets with a few hundred taxa
  - Unknown accuracy on larger datasets
• Parsimony-based methods
  - Slower than two-phase methods
  - No more accurate than two-phase methods
Results

1000 taxon models ranked by difficulty
Problem with Two-phase Approach

- **Problem**: two-phase methods fail to return reasonable alignments and accurate trees on large and divergent datasets
- **Insight**: divide-and-conquer to constrain dataset divergence and size
Obtain initial alignment and estimated ML tree
SATé Algorithm

Obtain initial alignment and estimated ML tree

Tree

Alignment

Insight:
Use tree to perform divide-and-conquer alignment
SATé Algorithm

Obtain initial alignment and estimated ML tree

Estimate ML tree on new alignment

Tree

Alignment

Insight:
Use tree to perform divide-and-conquer alignment
**SATé Algorithm**

1. Obtain initial alignment and estimated ML tree.
2. Estimate ML tree on new alignment.
3. **Insight:** Use tree to perform divide-and-conquer alignment.

**Insight:** Iterate - use a moderately accurate tree to obtain a more accurate tree.

If new alignment/tree pair has worse ML score, realign using a different decomposition.

Repeat until termination condition (typically, 24 hours).
SATé iteration
(Actual decomposition produces 32 subproblems)
SATé iteration
(Actual decomposition produces 32 subproblems)

Decompose based on input tree

A
B
C
D

A
B
C
D
SATé iteration
(Actual decomposition produces 32 subproblems)

Decompose based on input tree

Align subproblems

(Actual decomposition produces 32 subproblems)
SATé iteration
(Actual decomposition produces 32 subproblems)

Decompose based on input tree

Align subproblems

Merge subproblems
SATé iteration
(Actual decomposition produces 32 subproblems)

Decompose based on input tree

Align subproblems

Estimate tree on merged alignment

Merge subproblems

(Actual decomposition produces 32 subproblems)
SATé iteration
(Actual decomposition produces 32 subproblems)

Decompose based on input tree

Align subproblems

Estimate tree on merged alignment

Merge subproblems
Results

1000 taxon models ranked by difficulty
Results

1000 taxon models ranked by difficulty
Selected Current Contributions

Number of sequences

- Single gene or a few genes
  - Less than 1000
  - 1000+

Graduate work:
- SATé, SATé-II, DACTAL, etc.

Postgraduate work:
- PhyloNet-HMM, etc.

Sequence length

- Entire genome
Future Work: Topic #1

Number of sequences

1000+

Less than 1000

Graduate work:
SATé, SATé-II, DACTAL, etc.

Postgraduate work:
PhyloNet-HMM, etc.

Future work:
Divide and conquer on species networks

Sequence length

Single gene or a few genes

Entire genome
Future Work: Topic #2

Number of sequences:
- Single gene or a few genes
- Entire genome or 1000 sequences
- Less than 1000 sequences
- 1000+ sequences

Sequence length:

Graduate work:
- SATé
- SATé-II
- DACTAL, etc.

Postgraduate work:
- PhyloNet-HMM, etc.

Future work:
- Divide and conquer on species networks
Future Work: Topic #2

- **Number of sequences**
  - Less than 1000
  - 1000
  - 1000+

- **Sequence length**
  - Single gene
  - Entire genome or a few genes

**Graduate work:**
- SATé,
- SATé-II,
- DACTAL,
- etc.

**Postgraduate work:**
- PhyloNet-HMM,
- etc.

**Future work:**
- Divide and conquer on species networks

Biological and Biomedical Insight
Future Work: Topic #2

- **Graduate work:** SATé, SATé-II, DACTAL, etc.
- **Future work:** Divide and conquer on species networks
- **Postgraduate work:** PhyloNet-HMM, etc.

**Number of sequences:**
- Less than 1000
- 1000+

**Sequence length:**
- Single gene
- Entire genome or a few genes

**Biological and Biomedical Insight:**
- Interactomic Data
- Gene Expression Data
- Phenotypic (Trait) Data
Acknowledgments

• Thanks to my postdoctoral mentors (Luay Nakhleh and Michael H. Kohn), my graduate adviser and co-adviser (Tandy Warnow and C. Randal Linder), and their labs.

• Supported in part by:
  – A training fellowship from the Keck Center of the Gulf Coast Consortia, on Rice University’s NLM Training Program in Biomedical Informatics (Grant No. T15LM007093).
  – NLM (Grant No. R01LM00949405 to Luay Nakhleh)
  – NHLBI (Grant No. R01HL09100704 to Michael Kohn)
Questions?

- My website can be found at http://www.cs.rice.edu/~kl23
- Nakhleh lab website can be found at http://bioinfo.cs.rice.edu/