Algorithms for Analyzing Intraspecific Sequence Variation

Srinath Sridhar

Computer Science Department
Carnegie Mellon University

March 2, 2009
Outline

1 Motivation

2 Phylogeny Reconstruction
   - Definitions
   - Imperfect Phylogeny Reconstruction
   - Extensions
   - Empirical Results

3 Population Substructure
   - Pure Populations
   - Admixture
How can we characterize and use genomic variation that exists within a single species to understand its recent history?
Significance

- Fundamental to understanding of genome variation
- Disease association tests: ensure association of SNPs to cases/controls not underlying population substructure
- Direct to consumer genotyping: ancestry and life-time risks
Motivation
Phylogeny Reconstruction
Population Substructure

Analysis of Genetic Variation

- Finding genetic variation
  - What forms of variation does the genome exhibit?
- Analyzing evolution of the genome
  - How does one genome transform to another?
- Analyzing genetic distribution in populations
  - How do the variants characterize sub-populations?
Analysis of Genetic Variation

- Finding genetic variation
  - What forms of variation does the genome exhibit?
- Analyzing evolution of the genome
  - How does one genome transform to another?
- Analyzing genetic distribution in populations
  - How do the variants characterize sub-populations?
Finding genetic variation
  - What forms of variation does the genome exhibit?
Analyzing evolution of the genome
  - How does one genome transform to another?
Analyzing genetic distribution in populations
  - How do the variants characterize sub-populations?
Motivation

Phylogeny Reconstruction
Population Substructure

Finding Genetic Variation

- Large segments of mouse genome missing or duplicated
- Newer form of large-scale variation
- Joint work with Cold Spring Harbor Labs; *Nature Genetics* 2007

Citation

‘Breakthrough of the year 2007’ – *Science magazine*
First Part of Talk

Phylogeny reconstruction
Vertex: an individual’s Chromosome 2

Brown Hair
Black Hair
Genetic Distribution in Populations

Second part of Talk

Substructure in populations

European
- 10% black hair
- 90% brown hair

Migration

Asian
- 90% black hair
- 10% brown hair

Migration

- 99% black hair
- 1% brown hair

- 10% black hair
- 90% brown hair
Single Nucleotide Polymorphisms (SNPs)

- Variation due to single base change (SNPs)
- Only two bases per site
- Data-set represented by binary $n \times m$ matrix

Example

<table>
<thead>
<tr>
<th>ACGT</th>
<th>0000</th>
</tr>
</thead>
<tbody>
<tr>
<td>AACT</td>
<td>0110</td>
</tr>
<tr>
<td>TCGA</td>
<td>1001</td>
</tr>
</tbody>
</table>
Outline

1 Motivation

2 Phylogeny Reconstruction
   - Definitions
     - Imperfect Phylogeny Reconstruction
     - Extensions
     - Empirical Results

3 Population Substructure
   - Pure Populations
   - Admixture
Phylogeny Reconstruction

- Input matrix $I: n \times m$ binary
- Rows: taxa (chromosomes of individuals)
- Columns: sites (SNPs)
- Assume all sites contain both 0, 1
Motivation
Phylogeny Reconstruction
Population Substructure

Phylogeny Reconstruction

Definition
A phylogeny is an unrooted tree $T(V, E)$ where each vertex $v \in \{0, 1\}^m$ represents a taxon and an edge represents a single mutation (Hamming distance 1). Then $\text{length}(T) = |E|$.

Definition
A vertex $v$ that represents an input taxon is called a terminal vertex. Every other vertex is a Steiner vertex.
Example

<table>
<thead>
<tr>
<th>Individual</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Steiner: 1100

1 2 3 4

Ind 1: 0000
Ind 2: 1010
Ind 3: 1000
Ind 4: 1101
Ind 5: 0101
Imperfection of Phylogeny

Any phylogeny has length \textit{at least} \( m \)

**Definition**

Phylogeny \( T \) is called \( q \)-imperfect if \( \text{length}(T) = m + q \).

Phylogeny \( T \) is \textit{perfect} if \( \text{length}(T) = m \).

Imperfection \( q \Leftrightarrow q \) \textit{recurrent} mutations
Example

Individual 1: 0 0 0 0
Individual 2: 1 0 1 0
Individual 3: 1 0 0 0
Individual 4: 1 1 0 1
Individual 5: 0 1 0 1

Steiner: 1100

1–imperfect

Srinath Sridhar
Algorithms for Analyzing Intraspecific Sequence Variation
Outline

1 Motivation

2 Phylogeny Reconstruction
   - Definitions
   - Imperfect Phylogeny Reconstruction
   - Extensions
   - Empirical Results

3 Population Substructure
   - Pure Populations
   - Admixture
Problem Definition

- Input: \( n \times m \{0, 1\}\)-matrix \( I \)
- Output: phylogeny \( T \) connecting all \( n \) taxa of \( I \)
- Objective: minimize \( \text{length}(T) \)
- NP-complete, Steiner Minimum Tree over hypercubes
- Traditional approaches: Hill-climbing heuristics, brute-force
Problem Definition

- **Input:** \( n \times m \{0,1\}-\text{matrix } I, \text{ parameter } q \)
- **Output:** phylogeny \( T \) connecting all \( n \) taxa of \( I \)
- **Objective:** minimize \( \text{length}(T) \)
- **Assumption:** \( \text{length}(T^*) \leq m + q \) where \( T^* \) is the optimal tree
Results

<table>
<thead>
<tr>
<th>State</th>
<th>Imperf ((q))</th>
<th>Time (O(nm))</th>
<th>Work</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0</td>
<td>(O(nm))</td>
<td>Gusfield 92</td>
</tr>
<tr>
<td>(k)</td>
<td>(q)</td>
<td>(m^{O(q)}2^{O(q^2k^2)})</td>
<td>Fernandez-Baca and Lagergren 03</td>
</tr>
<tr>
<td>2</td>
<td>(q)</td>
<td>(O(21^q + 8^qnm^2))</td>
<td>ICALP 06, TCBB 07</td>
</tr>
</tbody>
</table>

**Fixed Parameter Tractability**

Other: many heuristics Nearest-neighbor, Tree bisection and reconnection etc
Imperfection

- $\text{imperfect}(l) \overset{\text{def}}{=} \text{imperfect}(T^*)$
- where $T^*$ is the optimal tree
- imperfection: number of duplicate edge labels

Example

```
           0000
          /  \
         2    1
       /   /  \
      4   2 3
     /   /   /
    010 1100 1010
   /     /     /
  0101 1100 1010
```

Srinath Sridhar

Algorithms for Analyzing Intraspecific Sequence Variation
Algorithm Overview

Example

Algorithm

function buildTree(matrix $M$)

1. If imperfect($M$) = 0 return $T_M^*$
2. ‘Guess’ site $j$ that mutates exactly once
3. ‘Guess’ adjacent vertices $u, v$
4. Partition $M$ into $M_0, M_1$ using $j$
5. Return buildTree($M_0$) $\cup$ buildTree($M_1$) $\cup \{(u, v)\}$
Motivation
Phylogeny Reconstruction
Population Substructure

Definitions
Imperfect Phylogeny Reconstruction
Extensions
Empirical Results

Algorithm Overview

Example

2-imperfect

Algorithm

function buildTree(matrix $M$)

1. If imperfect($M$) = 0 return $T_M^*$
2. ‘Guess’ site $j$ that mutates exactly once
3. ‘Guess’ adjacent vertices $u, v$
4. Partition $M$ into $M_0, M_1$ using $j$
5. Return buildTree($M_0$) $\cup$ buildTree($M_1$) $\cup$ \{(u, v)\}
Algorithm Overview

**Example**

2-imperfect

**Algorithm**

```plaintext
function buildTree(matrix M)
  1. If imperfect(M) = 0 return $T_M^*$
  2. ‘Guess’ site $j$ that mutates exactly once
  3. ‘Guess’ adjacent vertices $u, v$
  4. Partition $M$ into $M_0, M_1$ using $j$
  5. Return buildTree($M_0$) $\cup$ buildTree($M_1$) $\cup$ {$(u, v)$}
```
Algorithm Overview

**Example**

```
2-imperfect
```

**Algorithm**

```python
function buildTree(matrix M):
    1. If imperfect(M) = 0 return T^*_M
    2. ‘Guess’ site j that mutates exactly once
    3. ‘Guess’ adjacent vertices u, v
    4. Partition M into M0, M1 using j
    5. Return buildTree(M0) ∪
       buildTree(M1) ∪ {(u, v)}
```
Algorithm Overview

Example

Algorithm

```
function buildTree(matrix M)
    1. If imperfect(M) = 0 return $T^*_M$
    2. ‘Guess’ site $j$ that mutates exactly once
    3. ‘Guess’ adjacent vertices $u, v$
    4. Partition $M$ into $M_0, M_1$ using $j$
    5. Return $\text{buildTree}(M_0) \cup \text{buildTree}(M_1) \cup \{(u, v)\}$
```
Projections: If $\text{imperfect}(M) = 0$ return $T_M^*$

- Let $P(i,j)$ be projection of $I$ on sites $i,j$
- $\text{imperfect}(I) > 0$ iff $\exists i,j$ st $|P(i,j)| = 4$
- Implication: Easy to check if Gusfield’s algorithm

Example

<table>
<thead>
<tr>
<th></th>
<th>P(1, 2) = {(0, 0), (0, 1), (1, 0), (1, 1)}</th>
</tr>
</thead>
<tbody>
<tr>
<td>0000</td>
<td></td>
</tr>
<tr>
<td>0101</td>
<td></td>
</tr>
<tr>
<td>1100</td>
<td></td>
</tr>
<tr>
<td>1010</td>
<td></td>
</tr>
</tbody>
</table>
Projections: If imperfect($M$) = 0 return $T^*_M$

- Sites $i, j$ conflict if $|P(i, j)| = 4$
- Idea: if $i, j$ conflict then $T^*$ contains $i \rightarrow j \rightarrow i$ or $j \rightarrow i \rightarrow j$ path

**Example**

```
\begin{array}{c|cccc}
\text{ij} & 00 & j & 01 & \\
00 & 0000 & 0101 & 1100 & 1010 \\
01 & 0000 & 0101 & 1100 & 1010 \\
\end{array}
```

```
\begin{array}{c|ccc}
i & j & i & \\
0100 & 0101 & 1100 & 1010 \\
0000 & 0101 & 1100 & 1010 \\
1000 & 0101 & 1100 & 1010 \\
\end{array}
```

$i \rightarrow j \rightarrow i$ path or $j \rightarrow i \rightarrow j$ path
‘Guess’ site $j$ that mutates exactly once

- $K$: set of sites that conflict
- If $|K| \geq 2q$ then guess $j \leftarrow_{u.a.r} K$
- $\Pr[j \text{ occurs exactly once in } T^*] \geq 0.5$ (correct guess)

Example

```
K = \{1, 2\}, j = 1
```

![Diagram showing a phylogenetic tree with sites and conflicts]
‘Guess’ adjacent vertices $u, v$

If all vertices in $M_0$ contain state $s$ on site $k$ then $u[k] = s$
therefore $v[k] = s$

Example

All terminals have state $s$. Hence, $u[k]=s$
‘Guess’ adjacent vertices $u, v$

- If both $M_0$ and $M_1$ contain both states on site $k$ then guess $u[k] \leftarrow_{u.a.r.} \{0, 1\}$ (Pr[correct guess] = 0.5)
- If $t$ guesses performed then $\text{imperfect}(M_0) + \text{imperfect}(M_1) \leq \text{imperfect}(M) - t$

Example

$M_0$ contains terminals $v_1, v_2$
St $v_1[k] \neq v_2[k]$

$M_1$ contains terminals $v_3, v_4$
St $v_3[k] \neq v_4[k]$
Each guess has success probability 0.5
Each guess reduces imperfection by at least 1
imperfect(I) = q
Pr[algorithm finds \( T_i^* \)] \( \geq \) 0.25^q
Recap: Running time: exponential in q polynomial in \( n, m \)
Can be derandomized by enumeration
Outline

1. Motivation

2. Phylogeny Reconstruction
   - Definitions
   - Imperfect Phylogeny Reconstruction
   - Extensions
   - Empirical Results

3. Population Substructure
   - Pure Populations
   - Admixture
## Results

**Genotypes:** Conflated combinations of \( \{0, 1\}^m \) sequences

<table>
<thead>
<tr>
<th>Imperf ((q))</th>
<th>Time</th>
<th>Work</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>( O(nm^\alpha(n, m)) )</td>
<td>Gusfield 2003</td>
</tr>
<tr>
<td>0</td>
<td>( O(nm^2) )</td>
<td>Eskin, Halperin and Karp 2004</td>
</tr>
<tr>
<td>0</td>
<td>( O(nm) )</td>
<td>Ding, Filkov and Gusfield 2005</td>
</tr>
<tr>
<td>1</td>
<td>( O(nm^3) )</td>
<td>Song, Wu and Gusfield 2005</td>
</tr>
<tr>
<td>( q, 1 \text{ site} )</td>
<td>( O(nm^{q+2}) )</td>
<td>Satya et al. 2006</td>
</tr>
<tr>
<td>( q )</td>
<td>( nm^{O(q)} )</td>
<td>Sridhar, Blelloch, Ravi, Schwartz 2006</td>
</tr>
</tbody>
</table>
Outline

1 Motivation

2 Phylogeny Reconstruction
   - Definitions
   - Imperfect Phylogeny Reconstruction
   - Extensions
   - Empirical Results

3 Population Substructure
   - Pure Populations
   - Admixture
Practical ILP based algorithm (S, Lam, Blelloch, Ravi, Schwartz 07)

<table>
<thead>
<tr>
<th>Data Set</th>
<th>input</th>
<th>q</th>
<th>time(secs)</th>
<th>pars</th>
<th>penny</th>
</tr>
</thead>
<tbody>
<tr>
<td>human Y</td>
<td>150 × 49</td>
<td>1</td>
<td>0.02</td>
<td>0.02</td>
<td>2.55</td>
</tr>
<tr>
<td>bacterial</td>
<td>17 × 1510</td>
<td>7</td>
<td>4.61</td>
<td>0.08</td>
<td>0.06</td>
</tr>
<tr>
<td>chimp mtDNA</td>
<td>24 × 1041</td>
<td>2</td>
<td>0.14</td>
<td>0.08</td>
<td>2.63</td>
</tr>
<tr>
<td>chimp Y</td>
<td>15 × 98</td>
<td>1</td>
<td>0.02</td>
<td>0.02</td>
<td>0.03</td>
</tr>
<tr>
<td>human mtDNA</td>
<td>40 × 52</td>
<td>21</td>
<td>—</td>
<td>13.39</td>
<td>11.24</td>
</tr>
<tr>
<td>human mtDNA</td>
<td>395 × 830</td>
<td>14</td>
<td>—</td>
<td>53.4</td>
<td>712.95</td>
</tr>
<tr>
<td>human mtDNA</td>
<td>13 × 390</td>
<td>6</td>
<td>9.75</td>
<td>0.02</td>
<td>0.41</td>
</tr>
<tr>
<td>human mtDNA</td>
<td>33 × 405</td>
<td>4</td>
<td>1.36</td>
<td>0.09</td>
<td>0.59</td>
</tr>
</tbody>
</table>
Buddhists and Muslims of Ladakh: 52 mtDNA SNPs
Genome-Wide Scan (Sridhar and Schwartz 2008)

- Sliding window across whole genome
- Construct phylogeny for each window
- Chromosome 2 imperfection on Central Europeans (top) and Africans (bottom)

*x*-axis: genomic position, *y*-axis: imperfection
Recent Work

- Tsai et al. used our method to cluster sub-populations
- CEU: Central Europeans, YRI: Yoruba Africans, CHB: Han Chinese, JPT: Japanese from Tokyo
Empirical Results

- Solved millions of problem instances spanning whole genome
- Provided fine-scale mutation rates across genome
- Software used hundreds of times online
- Exciting new avenues
  - Find sub-populations
  - Find rapidly evolving regions of the genome
Outline

1 Motivation

2 Phylogeny Reconstruction
   - Definitions
   - Imperfect Phylogeny Reconstruction
   - Extensions
   - Empirical Results

3 Population Substructure
   - Pure Populations
   - Admixture
Outline

1. Motivation

2. Phylogeny Reconstruction
   - Definitions
   - Imperfect Phylogeny Reconstruction
   - Extensions
   - Empirical Results

3. Population Substructure
   - Pure Populations
   - Admixture
Problem Overview

Migration

European
10% black hair
90% brown hair

Migration

Asian
90% black hair
10% brown hair

Randomly Sampled

99% black hair
1% brown hair

10% brown hair
90% black hair

Srinath Sridhar
Algorithms for Analyzing Intraspecific Sequence Variation
Example

- Two populations: ‘Asians’ \((p)\) and ‘Europeans’ \((q)\)
- For simplicity, consider two SNPs with state 1 probabilities:
  - \((p_1, p_2) = (0.4, 0.1)\) (Asians)
  - \((q_1, q_2) = (0.3, 0.5)\) (Europeans)
- Randomly sampled European, SNP 2 has state 1: 0.5
Problem Definition

- Input: \( n \times m \)-matrix \( G \)
- Output: classification \( \hat{\theta} : \{1, \ldots, n\} \rightarrow \{0, 1\} \)
- Errors: \( \min \sum_{i=1}^{n} |\theta(i) - \hat{\theta}(i)| \)
  \( \theta \) is the correct classification
- Want to minimize errors (no training data)
Graph Based (RECOMB 2007)

- Graph $G(V, E)$
  - Each vertex represents an individual
  - Edge distance captures genomic distance
- Perform max-cut on $G$

Example

![Graph Example](image.png)
Mathematical Properties

**Distance function properties**
- Expected intra-distance = 0
- Expected inter-distance = $2d^2$, where $d$ is the $L_2$ distance between the two populations

**Convergence**
- When $m = \Omega\left(\frac{\log n}{\gamma^2}\right)$ where
  - $\gamma$: Expected (over SNPs) $L_2^2$ distance between populations
  - $n$: number of individuals
  - $m$: number of SNPs.
- max-cut is the correct partition
- max-cut can be found efficiently (polynomial time)
Accuracy in practice (RECOMB 2007)

89 individuals: 45 Chinese, 44 Japanese
structure: Markov Chain Monte Carlo based (cited 1000+ times)
Outline

1. Motivation

2. Phylogeny Reconstruction
   - Definitions
   - Imperfect Phylogeny Reconstruction
   - Extensions
   - Empirical Results

3. Population Substructure
   - Pure Populations
   - Admixture
Admixture Example

- European: 10% black hair, 90% brown hair
- Asian: 90% black hair, 10% brown hair
- Admixture: 90% brown hair, 10% black hair

Motivation
- Phylogeny Reconstruction
- Population Substructure
- Pure Populations
- Admixture

Srinath Sridhar
Algorithms for Analyzing Intraspecific Sequence Variation
Problem Definition

- **Input:** $n \times m$ matrix $G$
- **Output:** classification
  \[ \hat{\theta} : \{1, \ldots, n\} \times \{1, \ldots, m\} \rightarrow \{0, 0.5, 1\} \]
- **Errors:** $\theta(i, j) \neq \hat{\theta}(i, j)$
  - $\theta$ is the correct classification
- **Ancestry of every locus of every individual**
High Level Idea

- Sliding window of length $w$
- Predict ancestry $\hat{\theta} : \{0, 0.5, 1\}$ for local region
- Combine local predictions
- Software downloaded and used by hundreds of labs including Cornell, UCSF, Scripps, Harvard Medical School etc.

American Journal of Human Genetics 2008
Recap of Contributions

- Finding polymorphisms: copy number variation in mouse
- Phylogeny Reconstruction
  - Fixed parameter tractability for haplotypes
  - Polynomial time (when $q$ is fixed) for genotypes
  - Integer Linear Programming for general problem
  - Genome-wide analysis of phylogenies
- Population Substructure
  - Pure populations: Poly-time, provably correct; outperforms other methods in accuracy (closely related populations) and run-time
  - Admixed populations: outperforms other methods in accuracy (well-separated ancestral populations) and significantly faster
Conclusions and Future Work

- Finding variation
  - Finding copy number changes, reversals, deletions
- Analysis of Variation
  - Phylogenies over sub-populations
  - Richer population models
  - Selection
- Disease Association Tests
- Direct to consumer genotyping
  - No longer controlled studies
  - Identifying relationships: cousins, ancestry