Split Scores on Phylogenetic Trees and Applications

Laura Kubatko	OSU
Dennis Pearl	OSU
John Rhodes	UAF

Elizabeth S. Allman

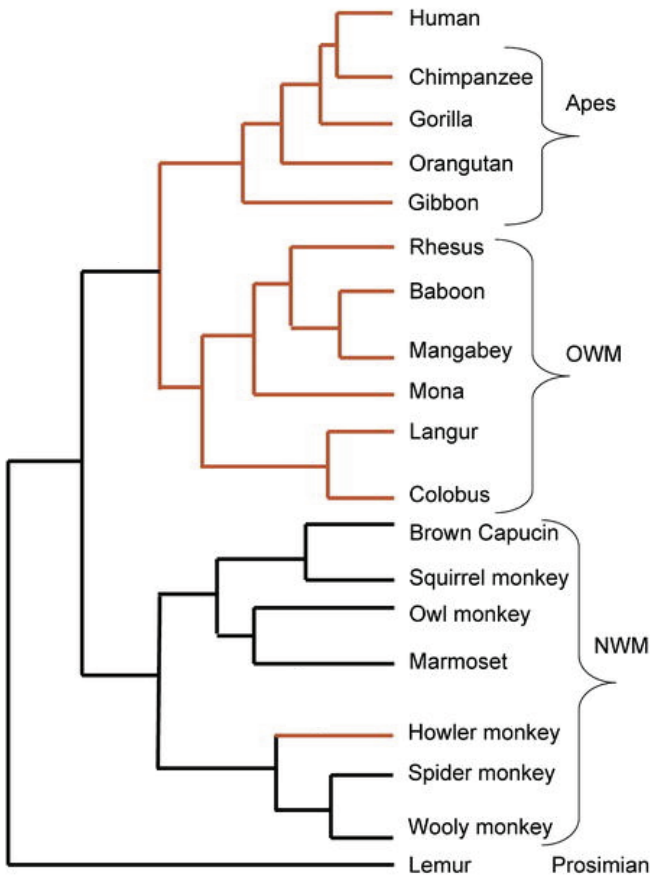
University of Alaska

Fairbanks

SIAM Conference on the Life Sciences

July 12, 2016

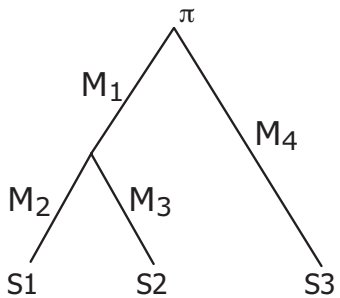
Theory: Split scores



Phylogenetic Tree of Primates

Under the general Markov model on phylogenetic trees, one can associate a score with an edge in the tree.

General Markov model on T



Model parameters =

$$\begin{cases} \text{tree topology} \\ \text{root distribution vector } \boldsymbol{\pi} = (\pi_i) \\ \text{Markov matrix on each edge } M_e \end{cases}$$

Under the iid assumption, one can compute the

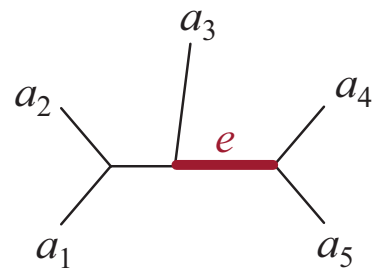
expected site pattern frequencies, e.g. p_{AAAAA}

```
Gorilla      AAGCTTCACCGGCGCAGTTGTTCTTATAATTGCCCACGGACTTACATCAT...
Orangutan    AAGCTTCACCGGCGCAACCACCCTCATGATTGCCCATGGACTCACATCCT...
Human        AAGCTTCACCGGCGCAGTCATTCTCATAATCGCCCACGGGCTTACATCCT...
Chimpanzee   AAGCTTCACCGGCGCAATTATCCTCATAATCGCCCACGGACTTACATCCT...
Gibbon       AAGCTTTACAGGTGCAACCGTCCTCATAATCGCCCACGGACTAACCTCTT...
```

Data: *empirical* $\widehat{p}_{AAAAA}$. Ideally, $\widehat{p}_{AAAAA} \approx p_{AAAAA}$.

Idea: Associate score to edge in tree.

Pattern frequency array has a natural 'flattening' according to splits in the tree:



$\mathbf{a_1a_2a_3 \mid a_4a_5}$.

The flattening for this split is the matrix

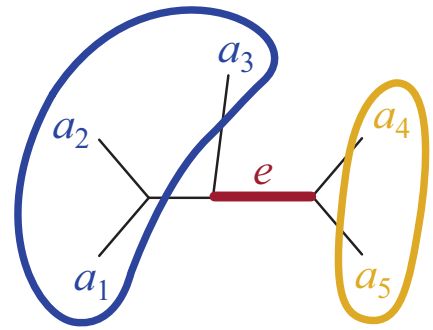
$$\mathbf{AAA} \rightarrow \begin{pmatrix} p_{\mathbf{AAA}AA} & p_{\mathbf{AAA}AC} & p_{\mathbf{AAA}AG} & \cdots & p_{\mathbf{AAA}TT} \\ p_{AACAA} & p_{AACAC} & p_{AACAG} & \cdots & p_{AACTT} \\ p_{AAGAA} & p_{AAGAC} & p_{AAGAG} & \cdots & p_{AAGTT} \\ p_{AATAA} & p_{AATAC} & p_{AATAG} & \cdots & p_{AATTT} \\ p_{ACAAA} & p_{ACAAC} & p_{ACAAG} & \cdots & p_{ACATT} \\ p_{ACCAA} & p_{ACCAC} & p_{ACCAG} & \cdots & p_{ACCTT} \\ p_{ACGAA} & p_{ACGAC} & p_{ACGAG} & \cdots & p_{ACGTT} \\ p_{ACTAA} & p_{ACTAC} & p_{ACTAG} & \cdots & p_{ACTTT} \\ \vdots & & & \ddots & \vdots \end{pmatrix} \quad \mathbf{TT} \downarrow$$

Rows and columns are indexed by states at conglomerate variables $a_1a_2a_3$ and a_4a_5.

A 16×64 matrix is defined analogously for the split $a_1a_2 \mid a_3a_4a_5$.

Idea: Associate score to edge in tree.

Pattern frequency array has a natural 'flattening' according to splits in the tree:



$\mathbf{a_1a_2a_3} \mid a_4a_5$.

The flattening for this split is the matrix

$$\mathbf{AAA} \rightarrow \begin{pmatrix} p_{\mathbf{AAA}AA} & p_{\mathbf{AAA}AC} & p_{\mathbf{AAA}AG} & \cdots & & p_{\mathbf{AAA}\mathbf{TT}} \\ p_{AACAA} & p_{AACAC} & p_{AACAG} & \cdots & & p_{AAC\mathbf{TT}} \\ p_{AAGAA} & p_{AAGAC} & p_{AAGAG} & \cdots & & p_{AAG\mathbf{TT}} \\ p_{AATAA} & p_{AATAC} & p_{AATAG} & \cdots & & p_{AAT\mathbf{TT}} \\ p_{ACAAA} & p_{ACAAC} & p_{ACAAG} & \cdots & & p_{ACA\mathbf{TT}} \\ p_{ACCAA} & p_{ACCAC} & p_{ACCAG} & \cdots & & p_{ACC\mathbf{TT}} \\ p_{ACGAA} & p_{ACGAC} & p_{ACGAG} & \cdots & & p_{ACG\mathbf{TT}} \\ p_{ACTAA} & p_{ACTAC} & p_{ACTAG} & \cdots & & p_{ACT\mathbf{TT}} \\ \vdots & & & & \ddots & \vdots \end{pmatrix} \quad \mathbf{TT} \downarrow$$

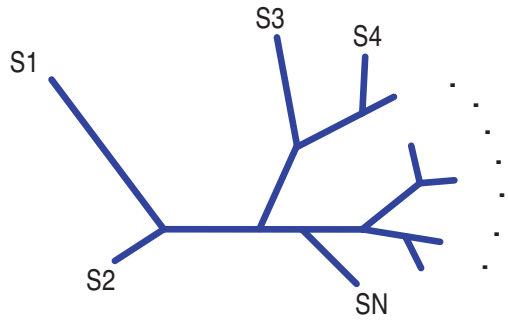
Rows and columns are indexed by states at conglomerate variables $a_1a_2a_3$ and a_4a_5.

A 16×64 matrix is defined analogously for the split $a_1a_2 \mid a_3a_4a_5$.

Theorem:

Given GM on an n-taxon tree, each matrix flattening corresponding to an edge of the tree has rank 4.

Otherwise the rank > 4.



Germ of idea:

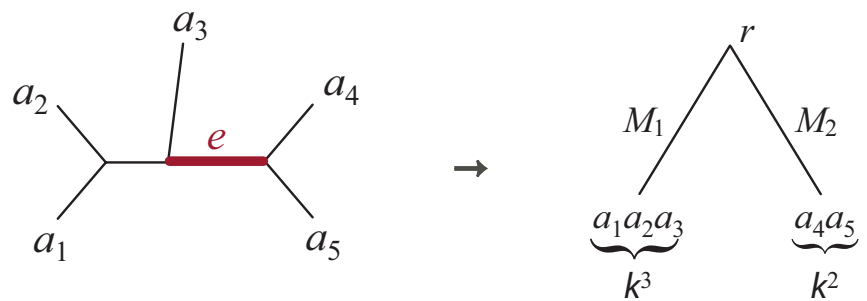
$$\text{matrix flattening rank} = 4 \iff \text{edge in tree}$$

$$\Updownarrow$$

$$\text{singular value decomposition}$$

Theorem: Given GM on an n-taxon tree, each matrix flattening corresponding to an edge of the tree has rank 4. Otherwise the rank > 4.

Focusing on edge e leads to a 'simpler' graphical model:



$$\text{5-dim tensor} \quad \rightsquigarrow \quad 4^3 \times 4^2 \text{ matrix}$$

$$P = \left(p_{AAAAA,\, etc}\right) \text{ flattens to a matrix } P_e$$

with a matrix factorization $P_e = M_1^T \operatorname{diag}(\boldsymbol{\pi}_\mathbf{r}) M_2$.

Connections to tensor rank, latent class models, ...

Other applications of result

- Ideals and Varieties:

  Allman-Rhodes: "Phylogenetic Ideals and Varieties for the general Markov model on trees"

  Draisma-Kuttler: Ideal theoretic version of theorems

- Parameter identifiability results: (Covarion models, mixture models, ...)

- Tree reconstruction algorithms:

  Eriksson 2005: Tree construction algorithm using SVD (GM)

  Kubatko and Chifman 2014: `SVDquartets` (GTR)

  Casanellas and Fernández-Sánchez 2016: `Erik+2` quartet-based (GM)

How can we use these ideas for inference?

- **Software in C:** `SplitSup`
  - Input: aligned sequences, splits to be tested
  - Steps:

    Construct the matrix flattening for current split
    (sparse matrix, binary encoding of patterns)

    Compute four largest singular values
    (svdlibc package)

    Compute **score** and write to file. Repeat.
- $\rightsquigarrow$ *EXTREMELY FAST*
- Our **score** based on the SVD is in the range $[0, 1]$ with 0 good. For the split $A \mid B$ with matrix flattening $P_{A|B}$, the score is

$$\sqrt{\frac{\sum_{i=5}^{k} \sigma_i^2}{\sum_{i=1}^{k} \sigma_i^2}} = \frac{\sqrt{\sum_{i=5}^{k} \sigma_i^2}}{\|P_{A|B}\|_F}$$

How can we use these ideas for inference?

- **Software in C:** `SplitSup`
  - Input: aligned sequences, splits to be tested
  - Steps:

    Construct the matrix flattening for current split
    (sparse matrix, binary encoding of patterns)

    Compute four largest singular values
    (svdlibc package)

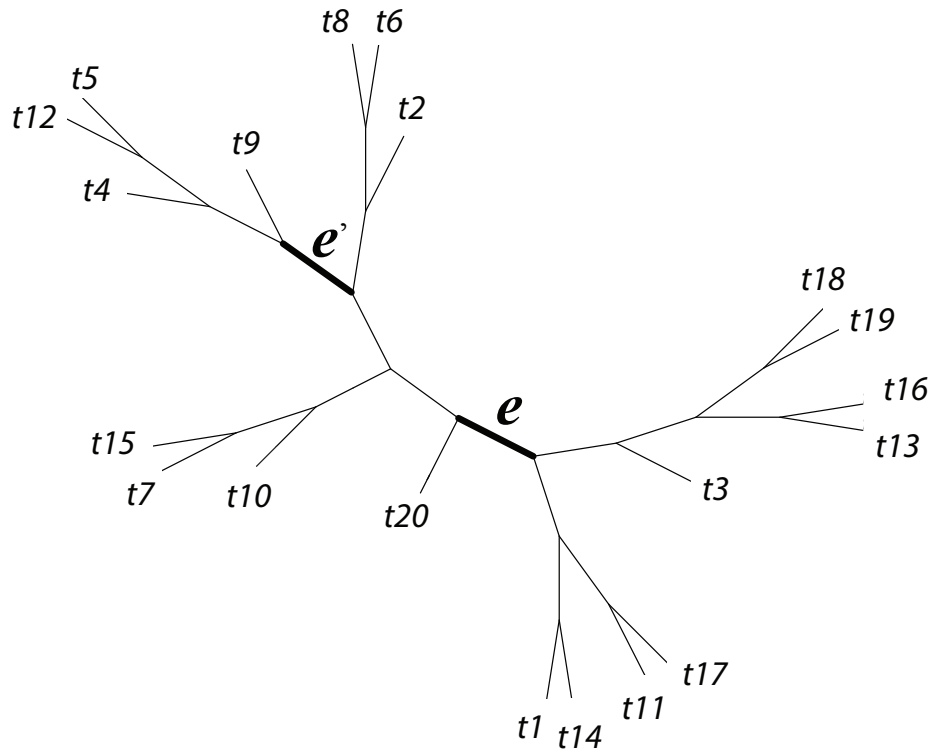
    Compute **score** and write to file. Repeat.

- $\rightsquigarrow$ *EXTREMELY FAST*
- Our **score** based on the SVD is in the range $[0, 1]$ with 0 good. For the split $A \mid B$ with matrix flattening $P_{A|B}$, the score is

$$\sqrt{\frac{\sum_{i=5}^{k} \sigma_i^2}{\sum_{i=1}^{k} \sigma_i^2}} = \frac{\sqrt{\sum_{i=5}^{k} \sigma_i^2}}{\|P_{A|B}\|_F}$$

Properties of splits scores

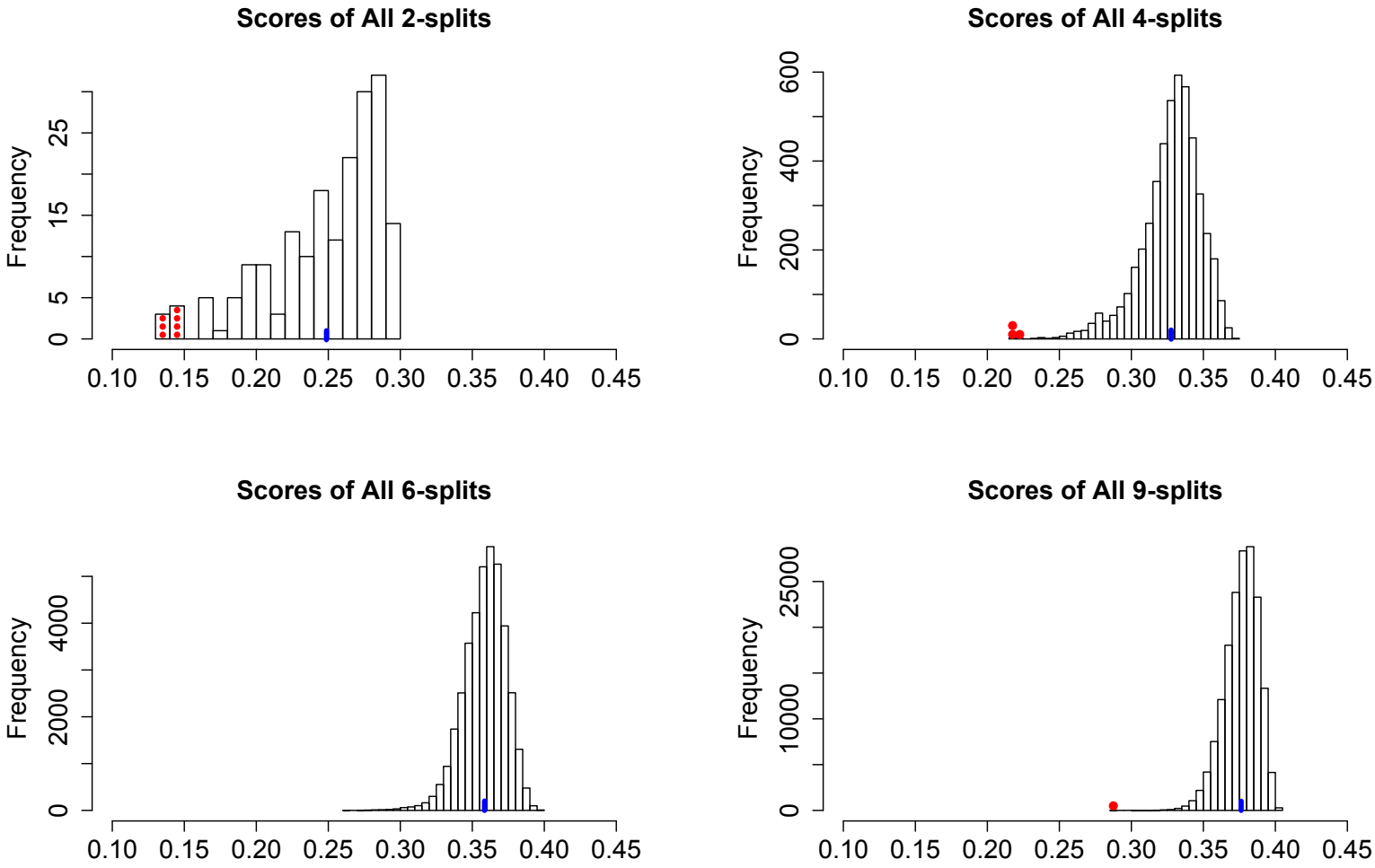
20 taxon tree used for simulations



Simulations: Jukes-Cantor model

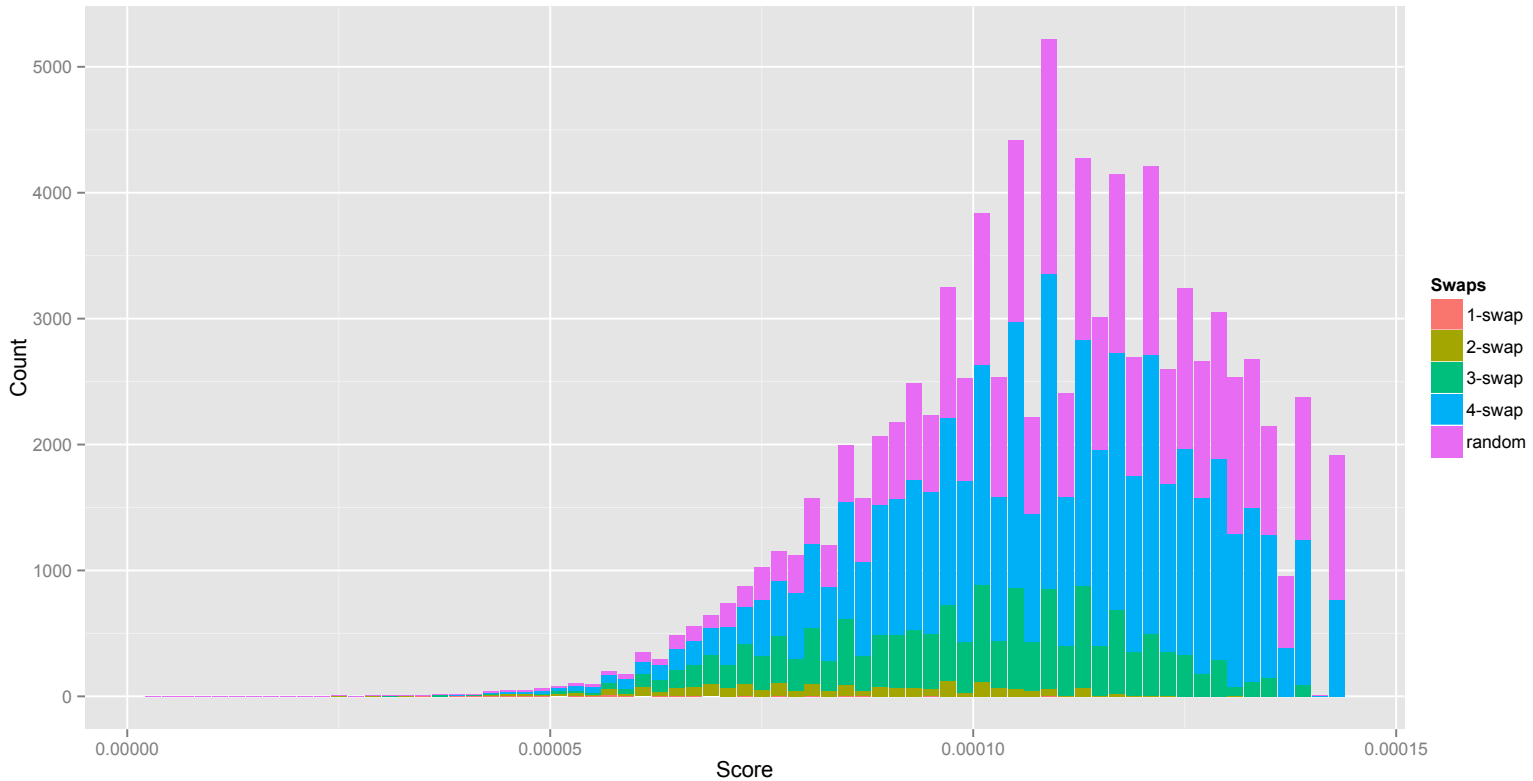
e' = 4-split, e = 9-split, false splits

Application 1: Identify true phylogenetic splits



True splits in tree shown with red dots. Smallest scores.

Application 1: Identify close to true phylogenetic splits

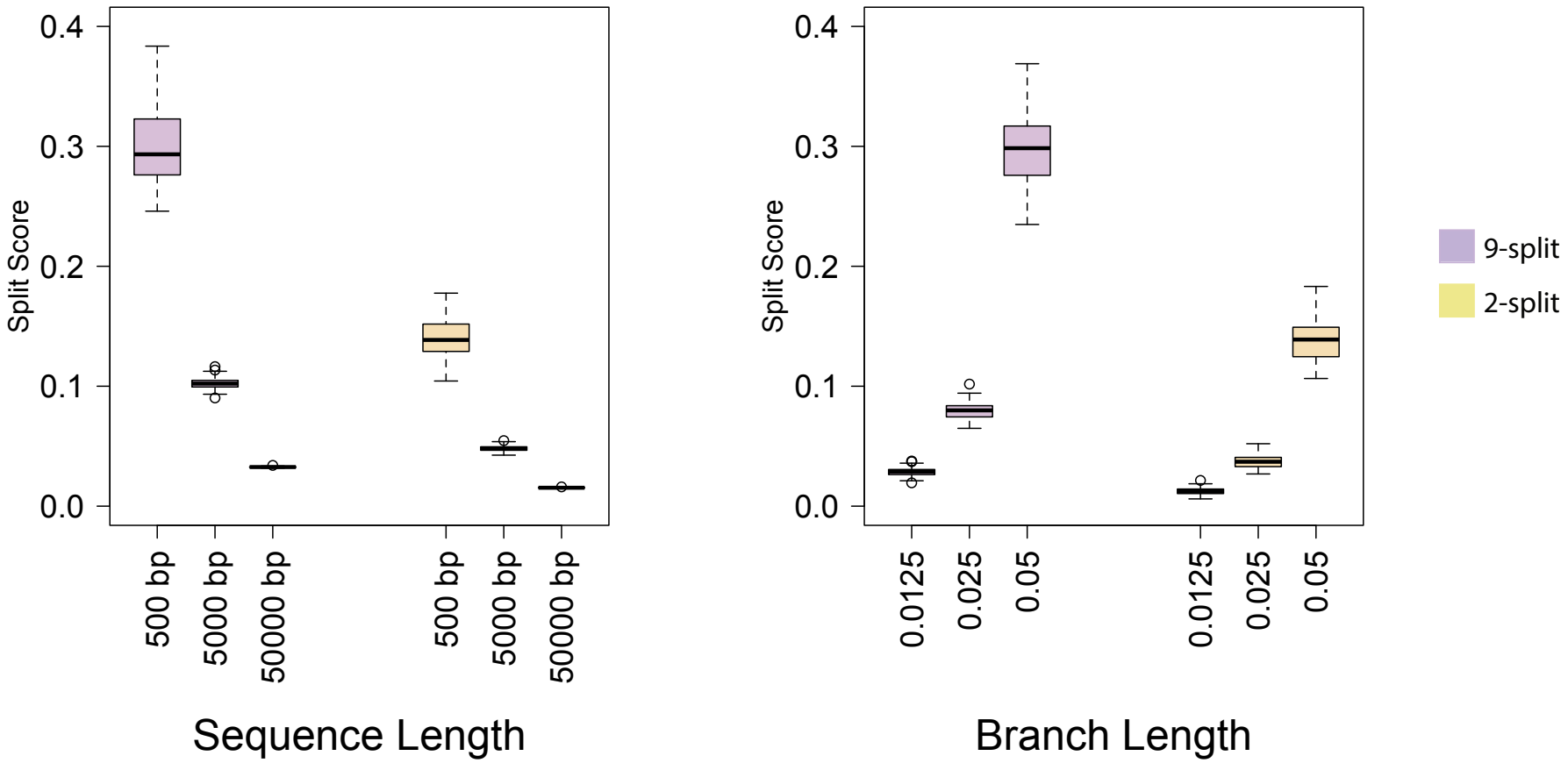


Fix true 10-split and plot scores for 10-splits that differ by 1 taxon, 2 taxa, ...

Properties: Sequence length and branch length effects

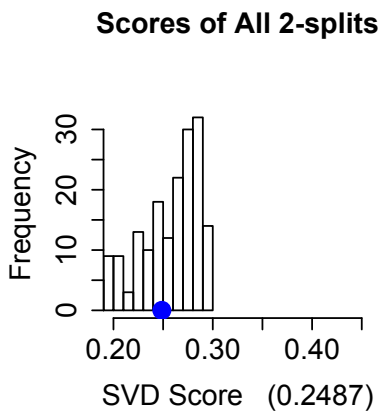
(L) Split scores for true splits decrease with sequence length.

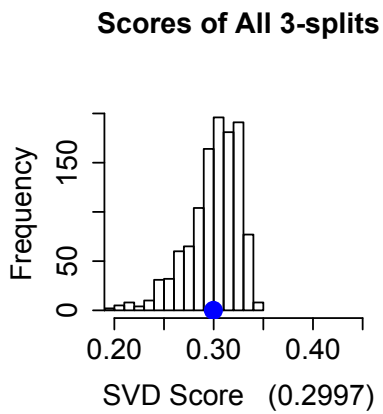
(R) Split scores scale with tree diameter.

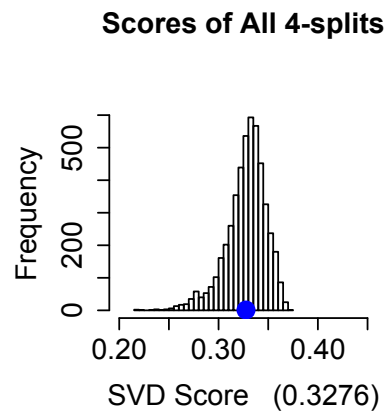


Properties: Split size effect

Split scores increase with split size.





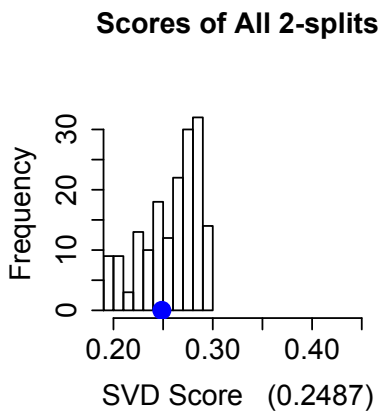


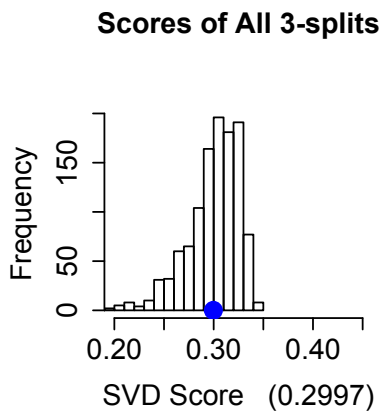
Mean of the observed distribution of split scores for a fixed size $k \,\big|\, (N-k)$ increases with k. Standard deviation decreases.

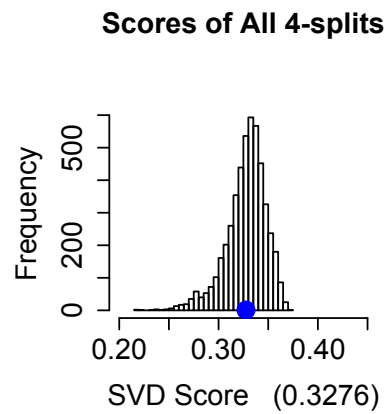
Caution: For use with data, comparisons should be made among splits of same size $k \,\big|\, N-k$.

Properties: Split size effect

Split scores increase with split size.







Dimension of rank varieties $4(4^k + 4^{N-k}) - 16$

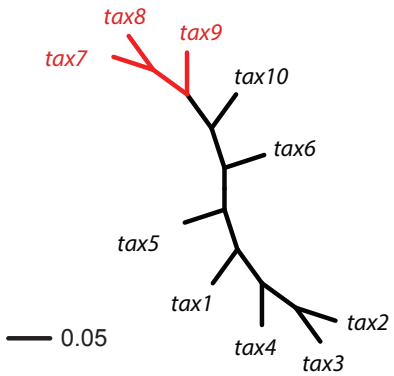
$$\min \text{ at } k = \frac{\lfloor N \rfloor}{2}, \quad \max \text{ at } k = 2$$

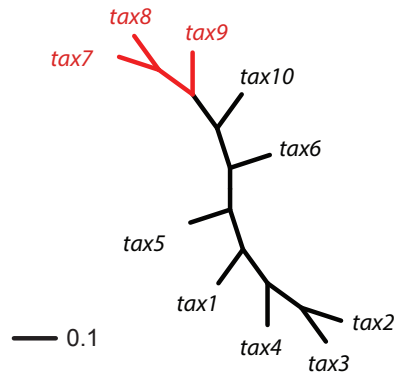
$\rightsquigarrow$ why Eriksson algorithm performs poorly in practice

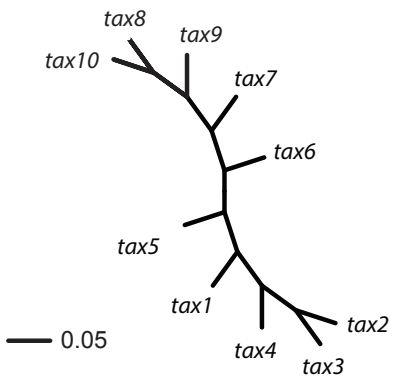
Application 2: Detecting shifts in evolutionary process along a chromosome

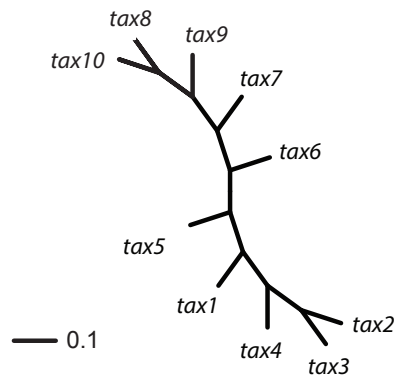
'Sliding Window Analysis'

Four phylogenies along a chromosome







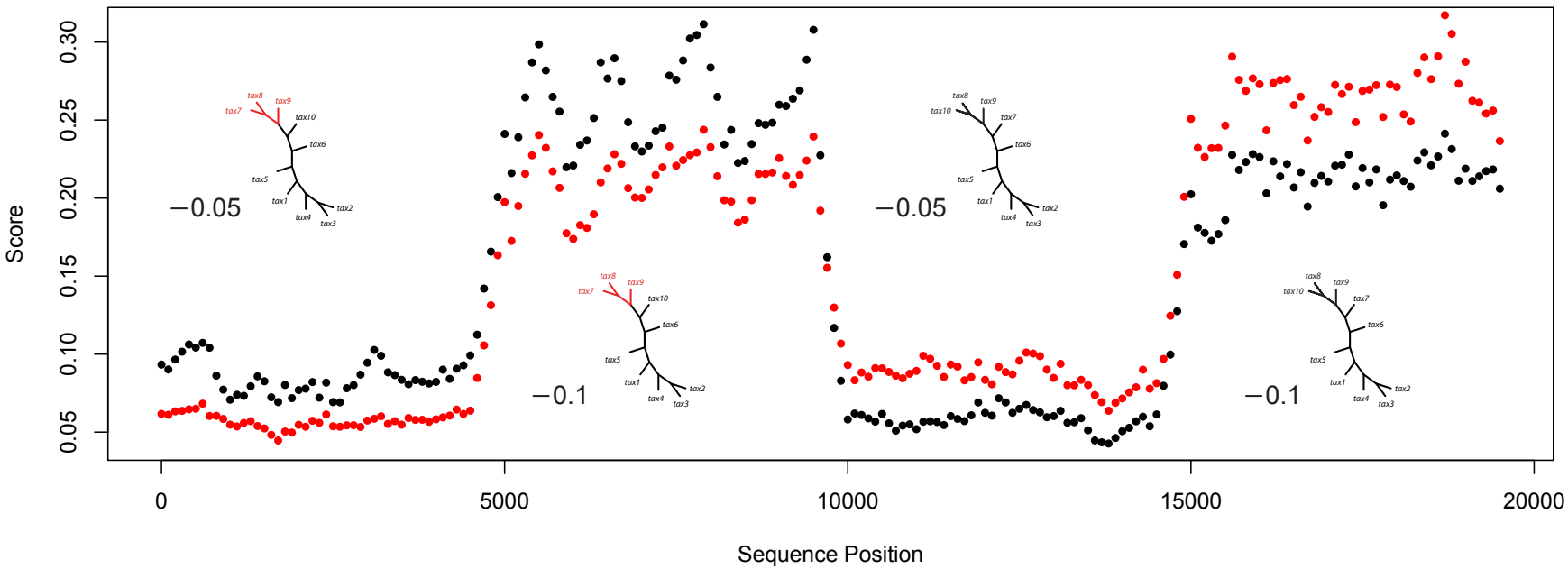


5000 bp were simulated for each model tree, JC model, then concatenated

Application 2: Detecting shifts in evolutionary process along a chromosome

7-8-9 split

8-9-10 split



Lower scores indicate better fit. 'Best' score shifts with model tree.

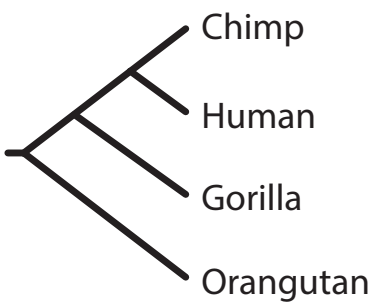
Methods: JC model 5000 bp per tree, concatenated block size: 500 bp, offset: 100 bp

Application 2: Detecting shifts in evolutionary process along a chromosome

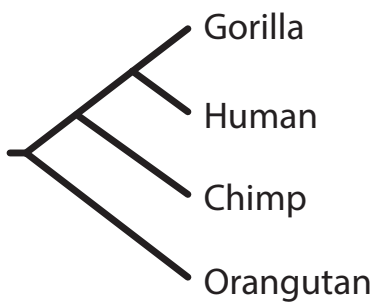
How does this perform for empirical data?

Consider the chromosome 7 data of Patterson et al. *Nature* (2006), used here to examine the relationships among human-chimp-gorilla.

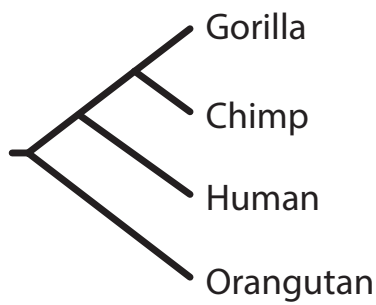
Three possible gene trees:



(HC)G



(GH)C

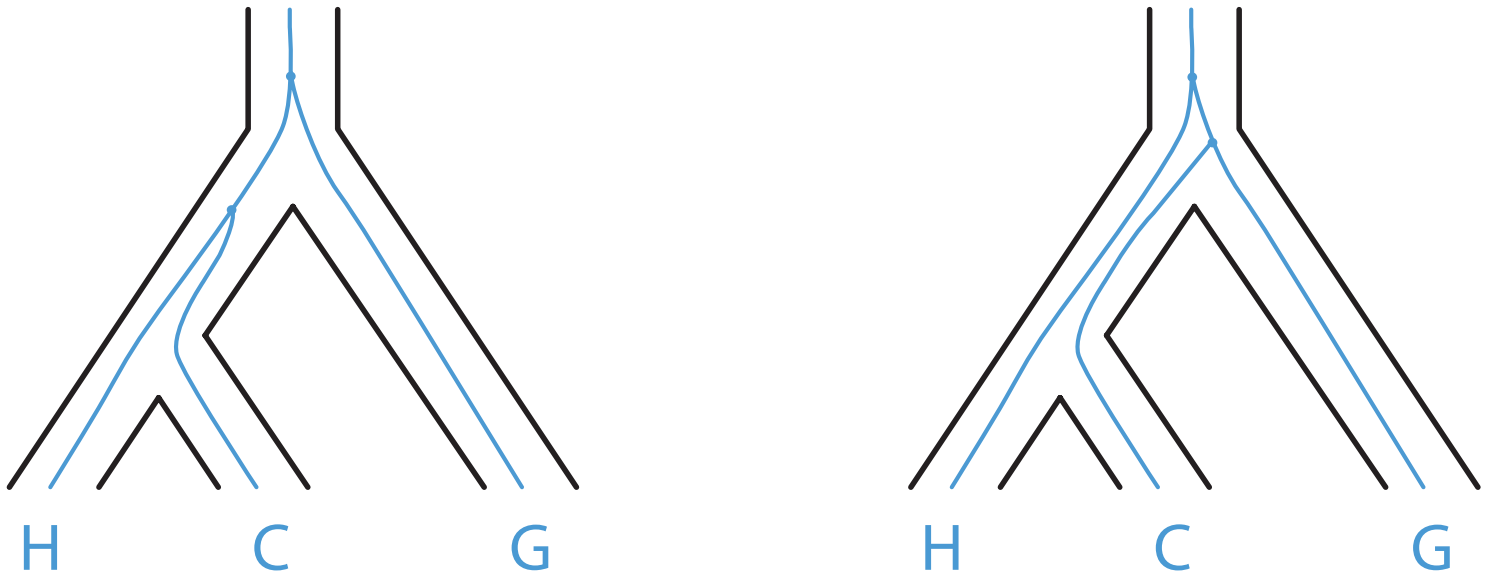


(GC)H

Species trees $\neq$ gene trees.

Application 2: Detecting shifts in evolutionary process along a chromosome

Species trees $\neq$ gene trees.



Gene trees $((H, C), G)$ and $(H, (C, G))$

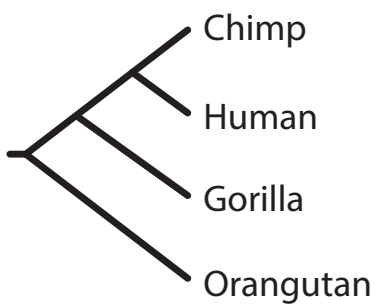
in species tree $((h, c), g)$

Application 2: Detecting shifts in evolutionary process along a chromosome

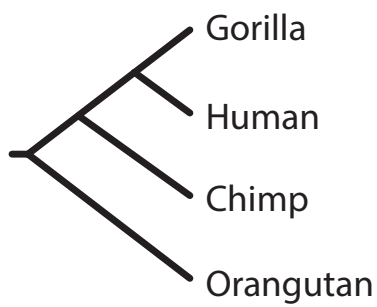
How does this perform for empirical data?

Consider the chromosome 7 data of Patterson et al. *Nature* (2006), used here to examine the relationships among human-chimp-gorilla.

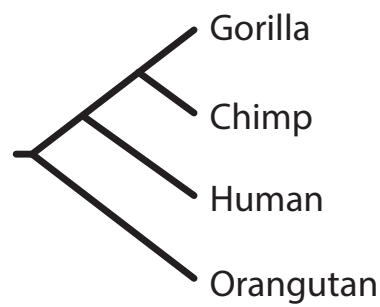
Three possible trees:



(HC)G



(GH)C



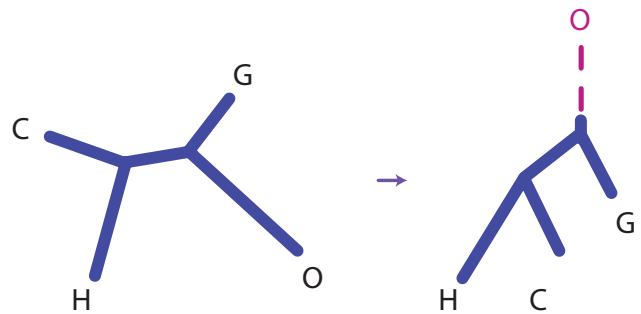
(GC)H

QUICK scan of Chromosome 7 in blocks of size 2000 bp

How does this perform for empirical data?

Consider the chromosome 7 data of Patterson et al. *Nature* (2006), used here to examine the relationships among human-chimp-gorilla.

Three possible trees:

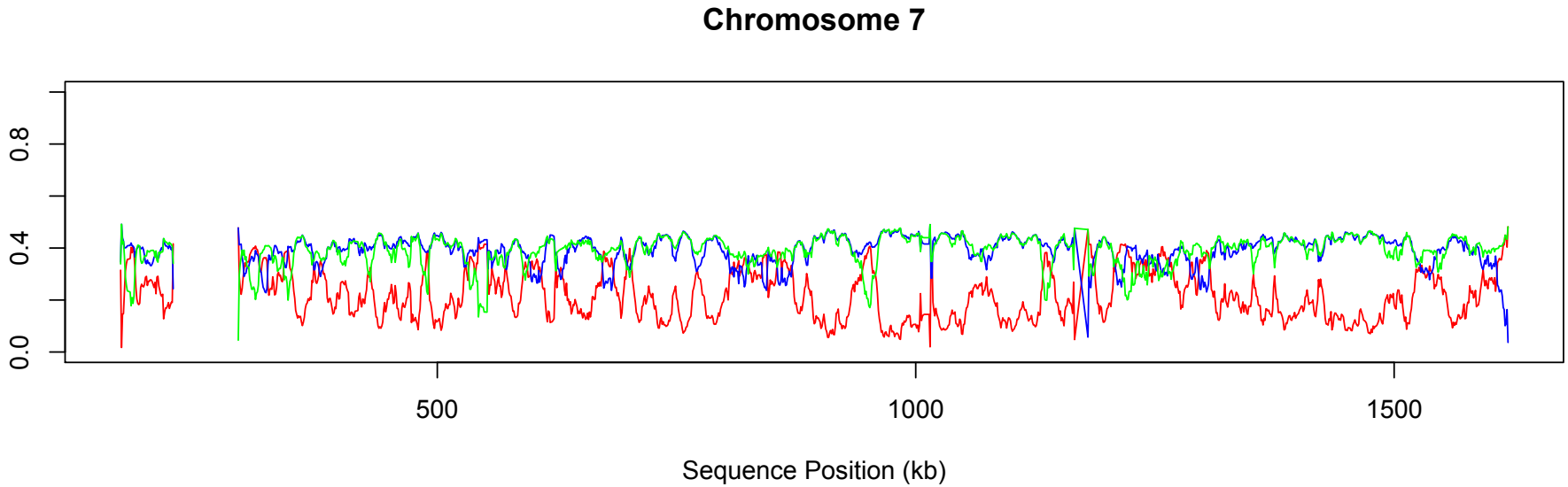


Outgroup is used to root tree, then deleted.

QUICK scan of Chromosome 7 in blocks of size 2000 bp

Application 2: Detecting shifts in evolutionary process along a chromosome

Empirical data of Patterson et al. (2006):



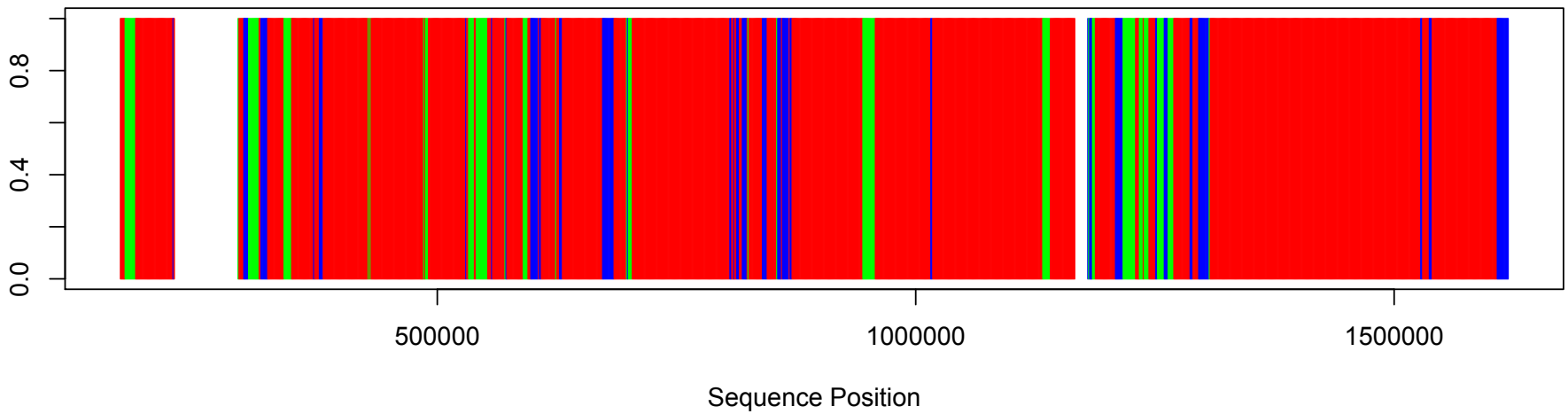
(HC)G (GH)C (GC)H

Chr 7 has
1,903,271 bp

window size:	2000 bp
slide size:	1000 bp
min no of sites required:	500 bp

Application 2: Detecting shifts in evolutionary process along a chromosome

Empirical data of Patterson et al. (2006):



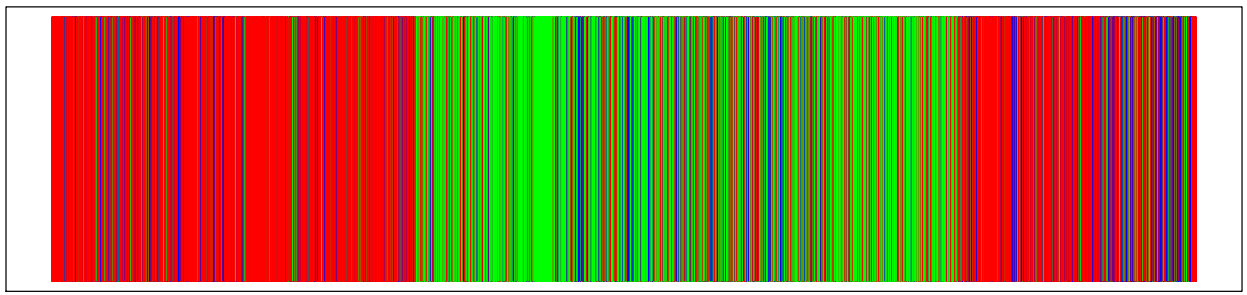
(HC)G (GH)C (GC)H

Chr 7 has
1,903,271 bp

window size:	2000 bp
slide size:	1000 bp
min no of sites required:	500 bp

Application 2: Detecting shifts in evolutionary process along a chromosome

Mosquito Data – Fontaine et al. (2015): phylogenomic analysis of whole genomes from *Anopheles gambiae* species complex.



Sequence Position

Sliding window method picks out rough location of known chromosome 2L inversion as a shift in the phylogeny.

Chr 2L has $\sim$ 37.5-million bp

window size:	10,000 bp
slide size:	1000 bp
min no of sites required:	500 bp

red vertical line = *An. gambiae-An. coluzzii* clade (known sister species)
green vertical line = *An. coluzzii-An. arabiensis* clade

Application 2: Detecting shifts in evolutionary process along a chromosome

Viral Data — Alicai et al. (2016):

Whole-genome alignment of the 10 distinct genes for all 29 individual samples (3 new).

$$\begin{cases} \text{14 samples of Cassava Brown Streak Virus (CBSV)} \\ \text{15 samples of Ugandan Cassava Brown Streak Virus (UCBSV)} \end{cases}$$

Phylogenetic analysis:

Strong support for CBSV | UCBSV split.

CBSV has an accelerated rate of evolution compared to UCBSV, matching field observations of increased virulence for these strains.

Sliding window analysis: Do changes in the score across the genome indicate which genes may be involved in the shift in evolutionary rate of CBSV?

Application 2: Detecting shifts in evolutionary process along a chromosome

Viral Data — Alicai et al. (2016):

Whole-genome alignment of the 10 distinct genes for all 29 individual samples (3 new).

$$\begin{cases} \text{14 samples of Cassava Brown Streak Virus (CBSV)} \\ \text{15 samples of Ugandan Cassava Brown Streak Virus (UCBSV)} \end{cases}$$

Phylogenetic analysis:

Strong support for CBSV | UCBSV split.

CBSV has an accelerated rate of evolution compared to UCBSV, matching field observations of increased virulence for these strains.

Sliding window analysis: Do changes in the score across the genome indicate which genes may be involved in the shift in evolutionary rate of CBSV?

Application 2: Detecting shifts in evolutionary process along a chromosome

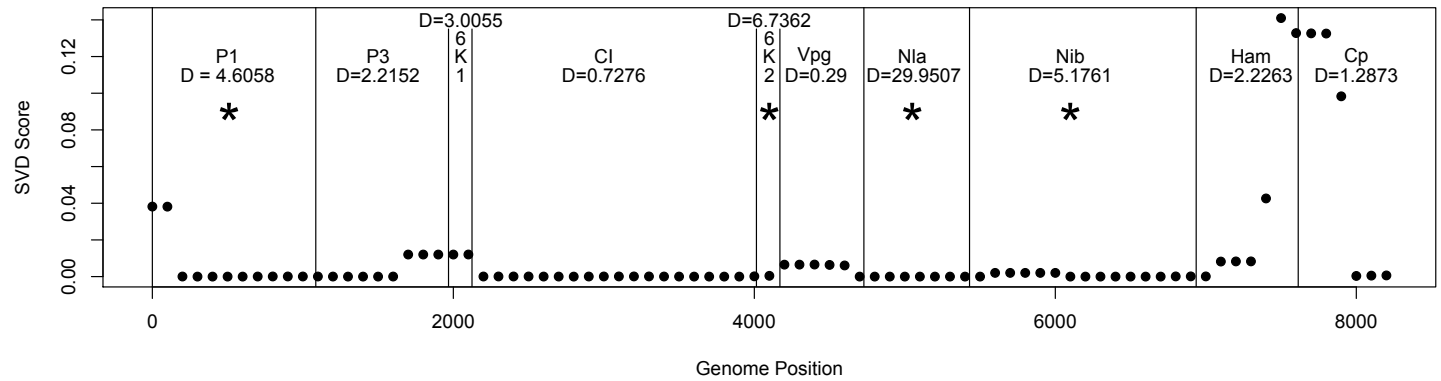
Viral Data — Alicai et al. (2016):

14 samples CBSV
15 samples UCBSV

Consider 14-split CBSV | UCBSV:

alignment is $\sim$ 8.8k bp

window size: 500 bp
slide size: 100 bp
min no of sites required: 100 bp



- indicates score.

$*$ indicates significant difference in evolutionary rate across split.

Gene boundaries are given by black vertical lines.

Thanks to my collaborators

Laura Kubatko, The Ohio State University

NIH National Institutes of Health

Dennis Pearl, The Ohio State University

John Rhodes, University of Alaska Fairbanks