Phyloclustering: A Model-Based Approach for Identifying Microbial Populations

Wei-Chen Chen

pbdR Core Team

2018 Symposium on Data Science and Statistics (SDSS)

Disclaimer

Any opinions, findings, and conclusions or recommendations expressed in this presentation are those of the authors.

Nothing in this content has been formally disseminated by the U.S. Department of Health & Human Services or by U.S. Food and Drug Administration, and should not be construed to represent any determination or policy of University, Agency, Administration, or National Laboratory.

Outline

Motivation

Equine Infectious Anemia Virus (EIAV)

Background

Mixture Multivariate Normal Distribution

Model-based Clustering

Clustering for Nucleotide Sequences

Phyloclustering Approach

Continuous Time Markov Chain (CTMC) Model

Mixture transition probability

EM Algorithm

Simulation Study

Data Analysis

EIAV Result

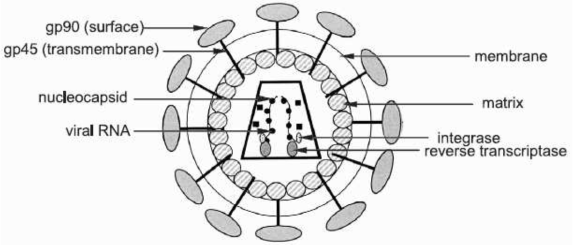
Summary

Motivation I

Equine Infectious Anemia Virus (EIAV)

- Leroux, Cadoré, and Montelaro (2004).
- ”Country cousin” of HIV.
- Lentivirus in the Retrovirus family infect equines.
- A persistent infection characterized by recurring febrile episodes associating with viremia, thrombocytopenia, and wasting symptoms.

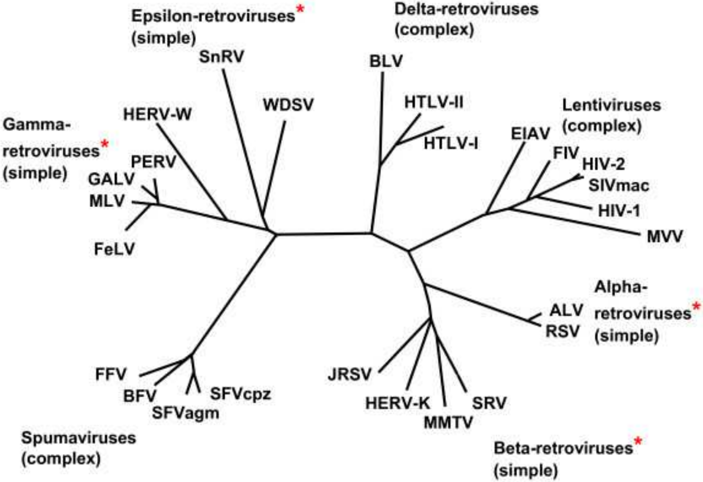
ISU Horse Barn (2006).



Leroux, Cadoré, and Montelaro (2004).

Motivation II

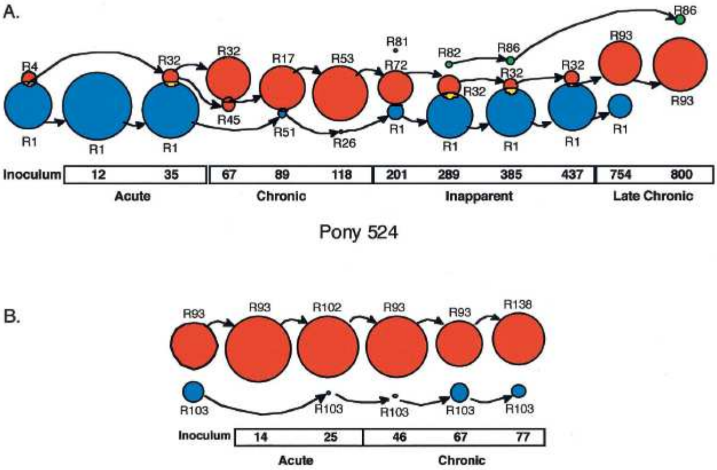
Phylogeny of Retroviruses



Weiss (2006).

Motivation III

PAQ: Partition Analysis of Quasispecies (Baccam et.al. (2001)).



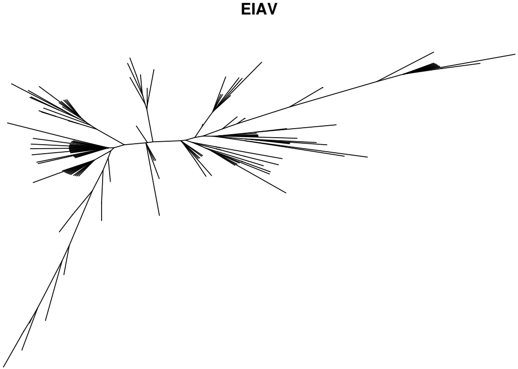
Pony 524

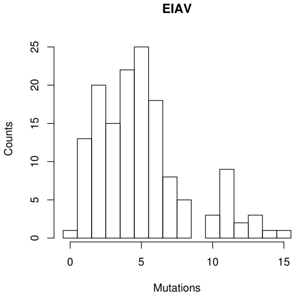
Pony 625

Baccam et al. (2003).

Motivation IV

146 EIAV *rev* sequences of pony 524.





Mutation counts for 146 sequences.

Motivation V

Number of bifurcating unrooted trees N_U for $n \geq 3$ sequences is

$$N_U = \frac{(2n-5)!}{2^{n-3}(n-3)!}.$$

Number of sequences	Number of unrooted trees
2	1
3	1
4	3
5	15
6	105
7	945
⋮	⋮
17	6,190,283,353,629,375
18	191,898,783,962,510,625
19	6,332,659,870,762,850,625
20	221,643,095,476,699,771,875

Felsenstein (1978) or Graur and Li (2000).

Goals of Phyloclustering

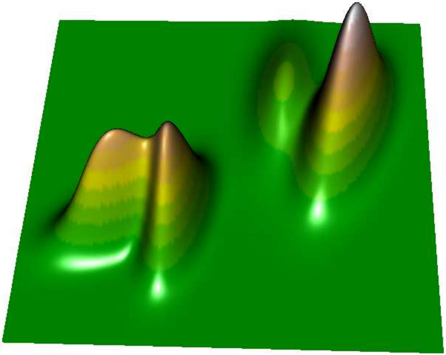
- to identify population centers where sequences may diverge from,
- to establish a model based approach to cluster sequences with phylogenetic meaning,
- to distinguish population structure based on classifications, and
- to aggregate trustworthy sequence information.

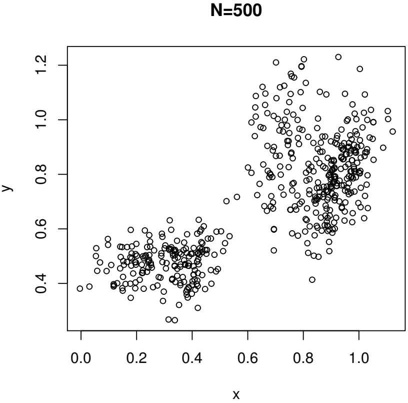
Mixture Multivariate Normal (MVN) Distribution

Mixture MVN with K components in p dimension:

$X_1, \ldots, X_N \overset{iid}{\sim} \phi(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\phi(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{k=1}^{K} \eta_k \phi_k(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k)$
where

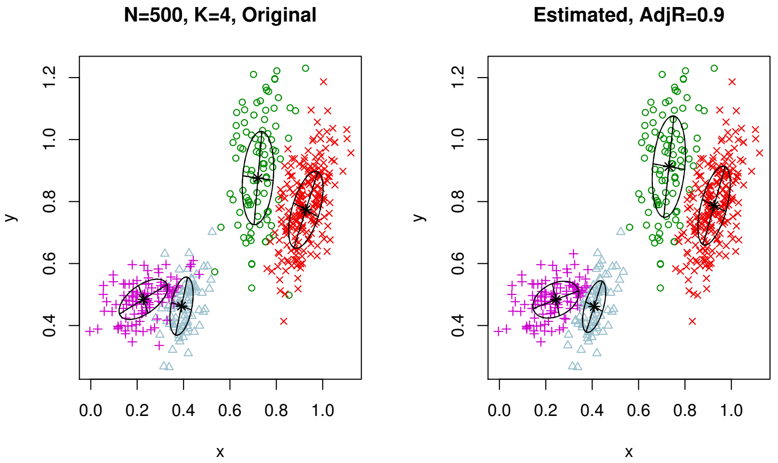
$$\phi_k(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)'\Sigma_k^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k)\right\}$$





Question: Are there four clusters? Where are they?

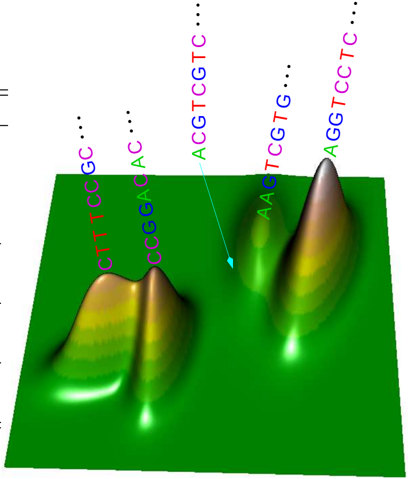
Model-based Clustering



Model-based clustering based on the mixture MVN model.

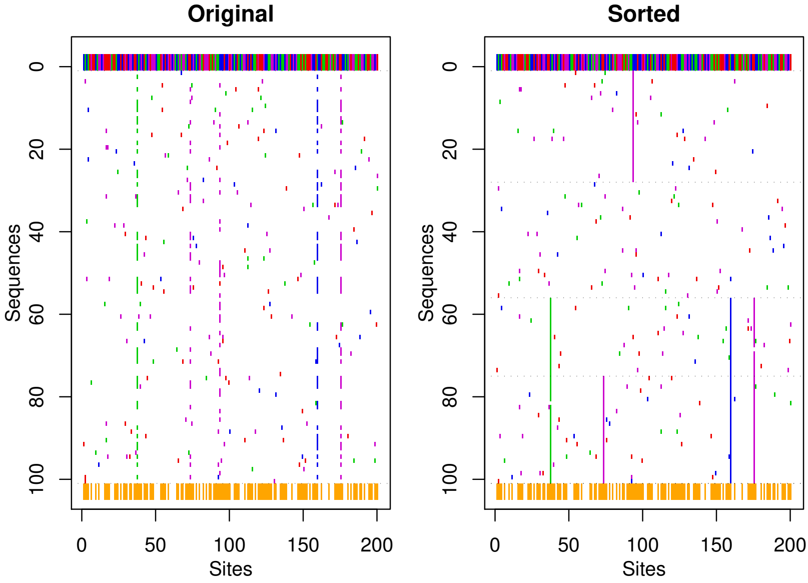
Clustering for Nucleotide Sequences

Id	Sequence	Center
1	ACGTCGTC···	AAGTCGTG···
2	AAGTCGTG···	
3	AAGTCGAG···	
4	AGGTCGCG···	
5	CCGGACAC···	CCGGACAC···
6	CCGGACAC···	
7	CTTGCCGC···	CTTTCCGC···
8	CTTTCCGC···	
9	AGGTCCTC···	AGGTCCTC···
10	AGGTCCTC···	



Question: How do we model/cluster this kind of data?

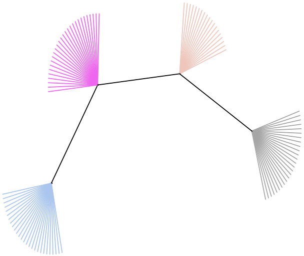
A Toy Dataset



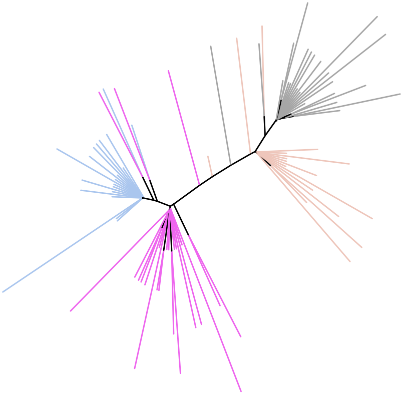
Green: A, Blue: G, Magenta: C, Red: T, Orange: segregating site.

Phylogenetic Approach

True tree for the toy dataset



Neighbor joining tree (K80)

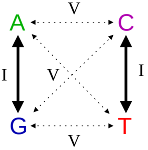


Question: What is the model for mutation process?

Continuous Time Markov Chain (CTMC) Model

Nucleotide substitution model: JC69 (Jukes & Cantor (1969)), K80 (Kimura (1980)), HKY85 (Hasegawa, Kishino & Yano (1985)).

For example, HKY85 defines $\boldsymbol{Q}_{x,y} = (q_{xy})_{4\times 4}$ as



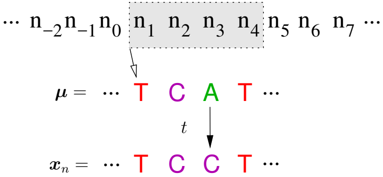
$$q_{xy} = \begin{cases} \pi_y & \text{if } x \text{ and } y \text{ differ by a transversion (V),} \\ \kappa\pi_y & \text{if } x \text{ and } y \text{ differ by a transition (I),} \end{cases}$$

for $y \neq x$, $q_{xx} = -\sum_{y\neq x} q_{xy}$ where $x, y \in \{A, G, C, T\}$.

$$\begin{array}{c c}
 & \begin{array}{cccc} \quad A \quad & \quad G \quad & \quad C \quad & \quad T \quad \end{array} \\
\begin{array}{c} A \\ G \\ C \\ T \end{array} & \begin{pmatrix} 1-\kappa\pi_G-\pi_C-\pi_T & \kappa\pi_G & \pi_C & \pi_T \\ \kappa\pi_A & 1-\kappa\pi_A-\pi_C-\pi_T & \pi_C & \pi_T \\ \pi_A & \pi_G & 1-\pi_A-\pi_G-\kappa\pi_T & \kappa\pi_T \\ \pi_A & \pi_G & \kappa\pi_C & 1-\pi_A-\pi_G-\kappa\pi_C \end{pmatrix}
\end{array}$$

$$\boxed{\text{CTMC: if } \boldsymbol{Q}_{x,y} = \boldsymbol{UDU}^{-1} \Rightarrow \boldsymbol{P}_{x,y}(t) = e^{\boldsymbol{Q}_{x,y}t} = \boldsymbol{U}e^{\boldsymbol{D}t}\boldsymbol{U}^{-1}}$$

Transition Probability



$\cdots$ n$_{-2}$n$_{-1}$n$_0$ n$_1$ n$_2$ n$_3$ n$_4$ n$_5$ n$_6$ n$_7$ $\cdots$
$\boldsymbol{\mu} =$ $\cdots$ T C A T $\cdots$
t
$\boldsymbol{x}_n =$ $\cdots$ T C C T $\cdots$

- $\boldsymbol{x}_n = (x_{n1}, \ldots, x_{nL}) \in \mathcal{S}^L$ where $x_{nl} \in \mathcal{S} = \{\text{A}, \text{G}, \text{C}, \text{T}\}$.
  - Assume mutations among sites are independent.
  - Assume $\boldsymbol{x}_n$ evolves from a population center $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_L) \in \mathcal{S}^L$.
  - Assume an substitution model, $\boldsymbol{Q}_{x,y}$.
  - Assume evolving time t between $\boldsymbol{\mu}$ and $\boldsymbol{x}_n$.

  Transition probability: $p_{\boldsymbol{\mu},\boldsymbol{x}_n}(t) = \prod_{l=1}^{L} P_{\mu_{kl},x_{nl}}(t)$.
- Distribution of mutation process:

$$\boxed{\phi(\boldsymbol{x}_n|\boldsymbol{\mu}, \boldsymbol{Q}, t) = p_{\boldsymbol{\mu},\boldsymbol{x}_n}(t).}$$

Mixture Transition Probability

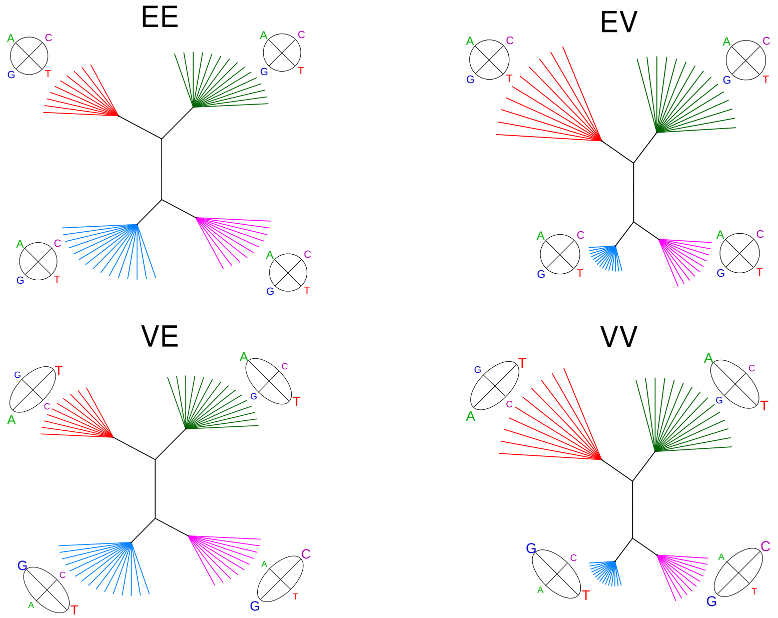
Mixture Transition Probability:

- Mixture proportion: $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)$, $\eta_k > 0$, and $\sum_{k=1}^{K} \eta_k = 1$.
- Dominant sequence (Center): $\boldsymbol{\mu}_k = (\mu_{k1}, \ldots, \mu_{kL}) \in \mathcal{S}^L$ where $\mu_{kl} \in \mathcal{S}$.
- CTMC model (Dispersion): $\boldsymbol{Q}_k$ and t_k.

Possible CTMC models:

- EE: $\boldsymbol{Q}_1 = \boldsymbol{Q}_2 = \cdots = \boldsymbol{Q}_K$ and $t_1 = t_2 = \cdots = t_K$.
- EV: $\boldsymbol{Q}_1 = \boldsymbol{Q}_2 = \cdots = \boldsymbol{Q}_K$ and $t_1 \neq t_2 \neq \cdots \neq t_K$.
- VE: $\boldsymbol{Q}_1 \neq \boldsymbol{Q}_2 \neq \cdots \neq \boldsymbol{Q}_K$ and $t_1 = t_2 = \cdots = t_K$.
- VV: $\boldsymbol{Q}_1 \neq \boldsymbol{Q}_2 \neq \cdots \neq \boldsymbol{Q}_K$ and $t_1 \neq t_2 \neq \cdots \neq t_K$.

Examples of CTMC models



EE

EV

VE

VV

EM Algorithm for Mixture Model

- Log likelihood: let $\mathbf{\Theta} = \{\boldsymbol{\eta}, \boldsymbol{\mu}, \boldsymbol{Q}, t\}$,

$$\log L(\mathbf{\Theta}|\boldsymbol{x}) = \sum_{n=1}^{N} \log \left[\sum_{k=1}^{K} \eta_k \phi_k(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{Q}_k, t_k) \right].$$

- Augment data for missing information:

$$Z_{nk} = I(n \in \mathcal{G}_k) \text{ for } n = 1, \ldots, N \text{ and } k = 1, \ldots, K.$$

- Log complete-data likelihood:

$$\log L_c(\mathbf{\Theta}, \boldsymbol{Z}|\boldsymbol{x}) = \sum_{n=1}^{N} \sum_{k=1}^{K} Z_{nk} \left[\log \eta_k + \log \phi_k(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{Q}_k, t_k) \right].$$

- EM algorithm: (Dempster et.al. 1977)
  1. E-step: $Q(\mathbf{\Theta}|\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{Z}}[\log L_c(\mathbf{\Theta}, \boldsymbol{Z}|\boldsymbol{x})]$.
  2. M-step: $\max_{\mathbf{\Theta}} Q(\mathbf{\Theta}|\boldsymbol{x})$.
  3. Iterate E- and M-steps until convergence which yields

$$\hat{\mathbf{\Theta}} = \underset{\mathbf{\Theta}}{\operatorname{argmax}} \log L(\mathbf{\Theta}|\boldsymbol{x}).$$

EM Algorithm for Phyloclustering with EE Model

- E-step:

$$z_{nk}^{(s)} = \mathbb{E}_{\boldsymbol{Z}}[Z_{nk}|\boldsymbol{x}, \boldsymbol{\Theta}^{(s-1)}] = \frac{\eta_k^{(s-1)}\phi_k(\boldsymbol{x}_n|\boldsymbol{\mu}_k^{(s-1)},\boldsymbol{Q}^{(s-1)},t^{(s-1)})}{\phi(\boldsymbol{x}_n|\boldsymbol{\mu}^{(s-1)},\boldsymbol{Q}^{(s-1)},t^{(s-1)})}$$

  where $n = 1, \ldots, N$ and $k = 1, \ldots, K$.

- M-step:
  - $\eta_k^{(s)} = \sum_{n=1}^N z_{nk}^{(s)}/N$.
  - $\boldsymbol{\mu}_k^{(s)}(\boldsymbol{Q}, \boldsymbol{t})$ obtained by comparing transition probabilities,

$$\begin{aligned} \mu_{kl}^{(s)}(\boldsymbol{Q}, t) &= \operatorname*{argmax}_{\mu \in \mathcal{S}} \textstyle\sum_{n=1}^N z_{nk}^{(s)} \log \phi_k(x_{nl}|\mu(\boldsymbol{Q}, t), \boldsymbol{Q}, t) \\ &= \operatorname*{argmax}_{\mu \in \mathcal{S}} \textstyle\sum_{a \in \mathcal{S}} \left[\left(\sum_{n \ni x_{nl} = a} z_{nk}^{(s)} \right) N_{\{x_l = a\}} \log p_{\mu,s}(t) \right]. \end{aligned}$$

  - $\boldsymbol{Q}^{(s)}$ and $t^{(s)}$ obtained numerically to maximize profile likelihood.

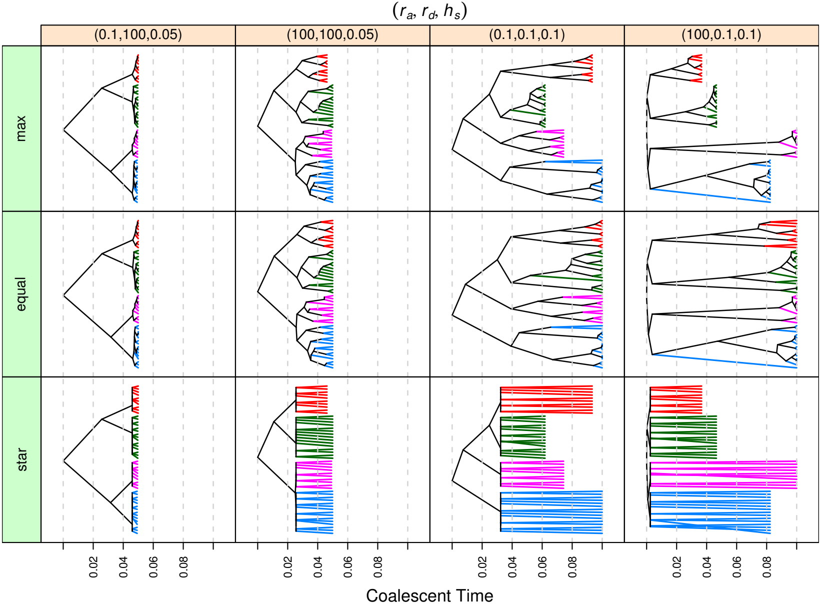
Challenges of EM Algorithm

1. Improve slow confergence of EM algorithm:
   - ECM (Meng & Rubin (1993)).
   - AECM (Meng & van Dyk (1997)).
   - APECM (Chen & Maitra (2011)).
2. Initialization schemes to improve convergent results:

   – Method:
   - Neighbor joining tree (Saitou & Nei (1987))
   - Partition Around Medoids (PAM) (Kaufman & Rousseeuw (1990))
   - K-Medoids (Theodoridis & Koutroumbas (2006))
   - Manually

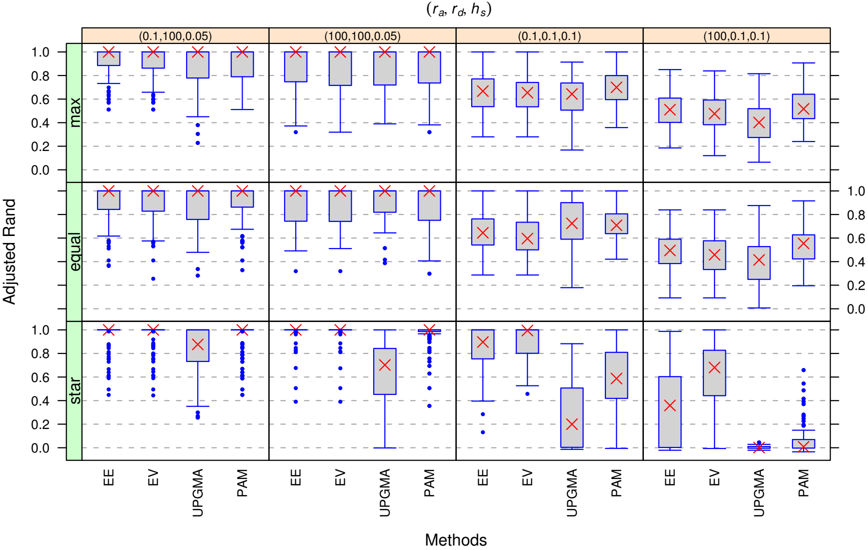
   – Procedure:
   - em-EM (Biernacki, Celeux, & Govaert (2003))
   - Rand-EM (Maitra (2007))
   - Exhausted EM

Simulation Study I



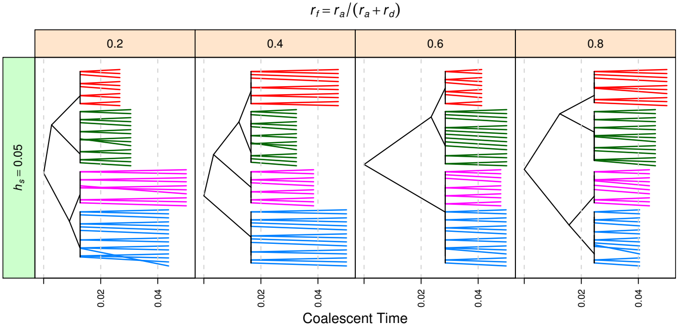
r_a: growth rate of ancestor tree, r_d: growth rate of descendent tree, h_s: total height.

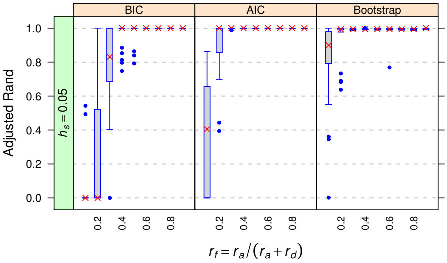
Results of Simulation Study I

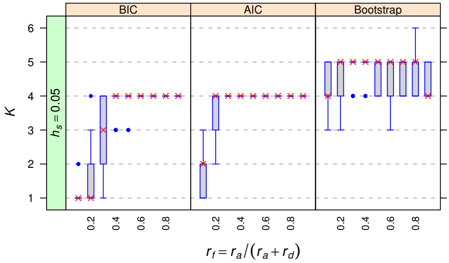


Results of EE (phyclust), EV (phyclust), UPGMA, and PAM.

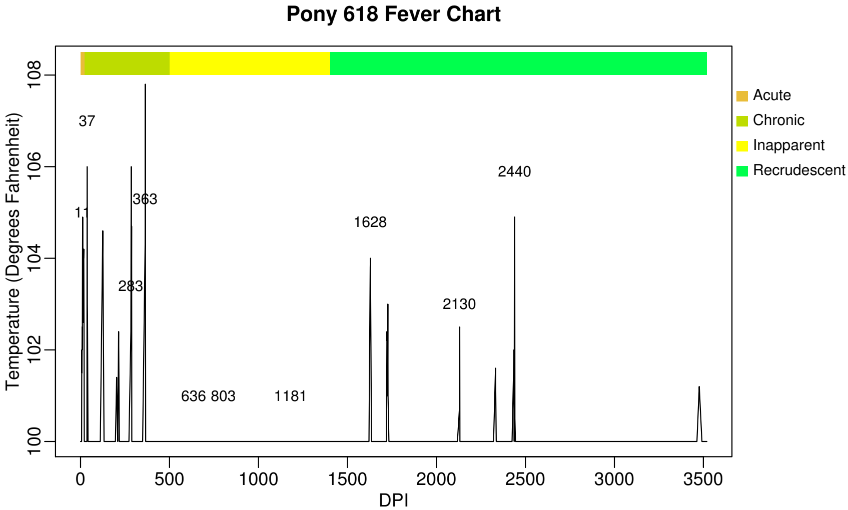
Simulation Study II and Results





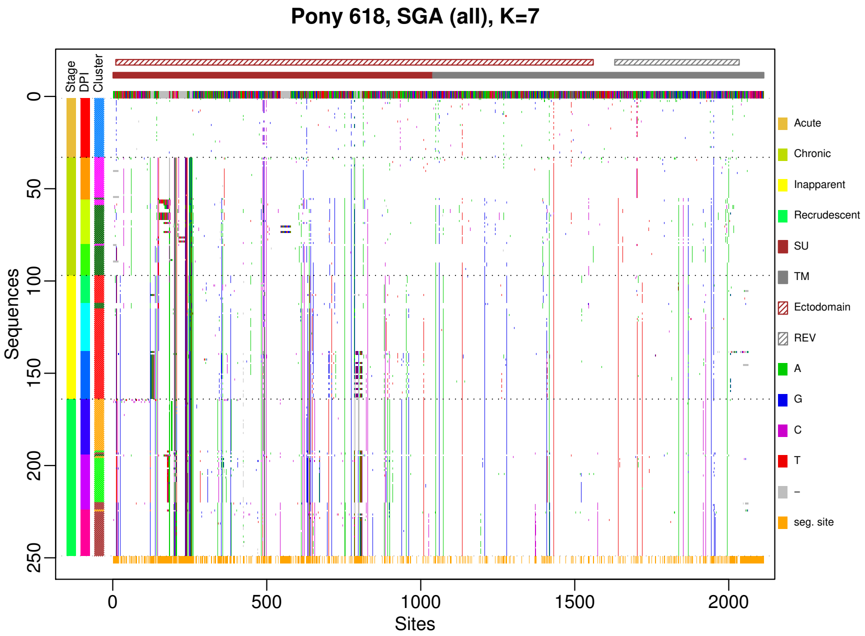


EIA Disease Progress

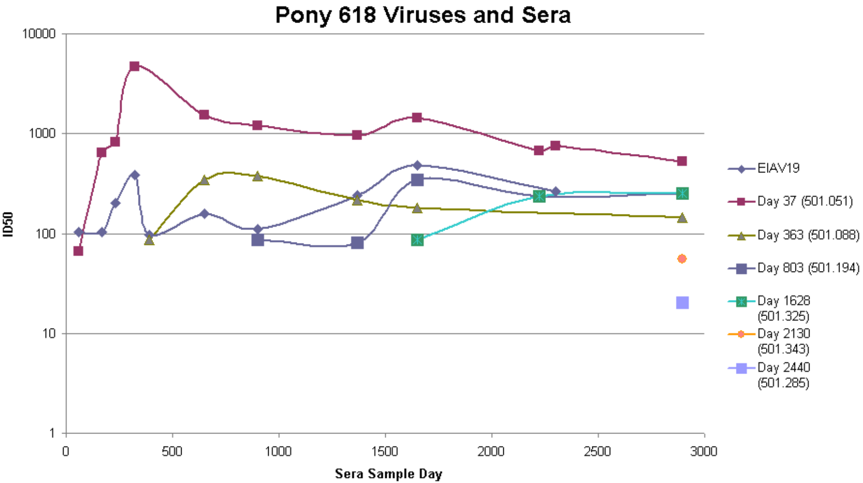


Cierra Pairett (2011), ”Longitudinal analysis of genetic and antigenic variation in EIAV env”, Iowa State University.

EIAV Phyloclustering Results

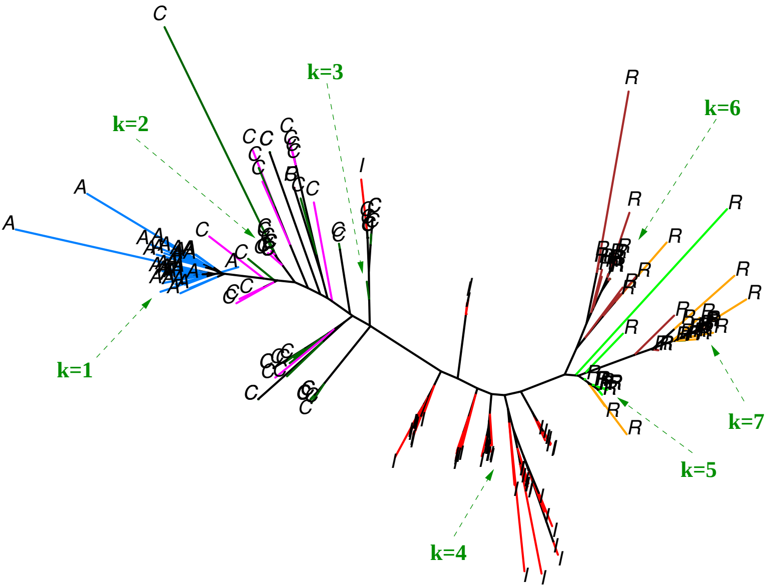


EIAV ID50 Result



Cierra Pairett (2011), ”Longitudinal analysis of genetic and antigenic variation in EIAV env”, Iowa State University.

EIAV Tree



Pony618, SGA (all), K=7

A: Acute, C: Chronic, I: Inapparent, R: Recrudescent.

Summary

- `phyclust`: an R package for Phylogenetic Clustering (`https://cran.r-project.org/package=phyclust`).
- Identify number of clusters.
- Initialization problem for EM algorithm.
- Potential extensions:
  - Reduce number of parameters (Hierarchical model for center sequences.)
  - Dependent structure along sites (Hidden Markov model.)

Acknowledgement

- Dr. Karin Dorman
- Dr. Ranjan Maitra
- Dr. Susan Carpenter
- Cierra Pairett

Thank you!