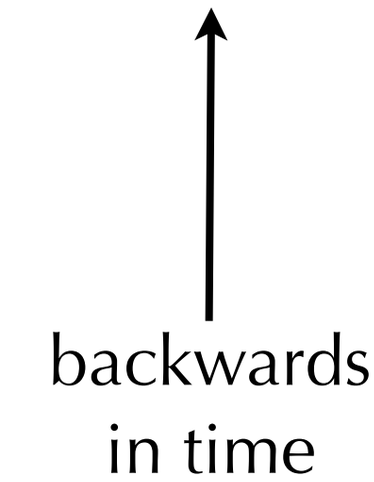
Modelling Genetic Variations with Fragmentation-Coagulation Processes

Yee Whye Teh, Charles Blundell, Lloyd Elliott
Gatsby Computational Neuroscience Unit, UCL

Genetic Variations in Populations

- Inferring histories of human populations.
- Understanding fundamental genetic processes.
- Associating genetic with phenotypic variations.
- Discovering genetic causes of diseases.

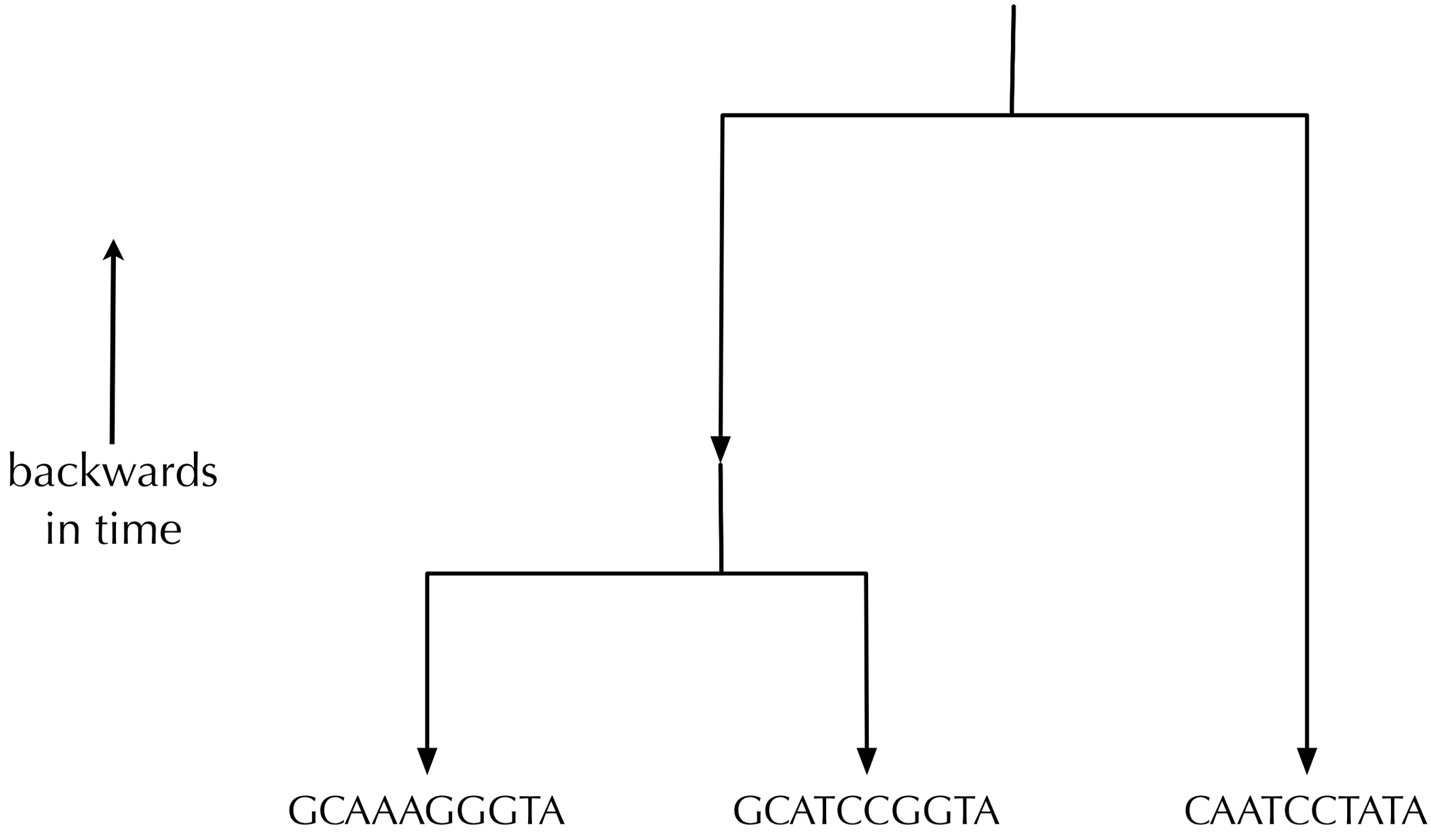
Ancestral Tree



GCAAAGGGTA GCATCCGGTA CAATCCTATA

[Kingman 1982]

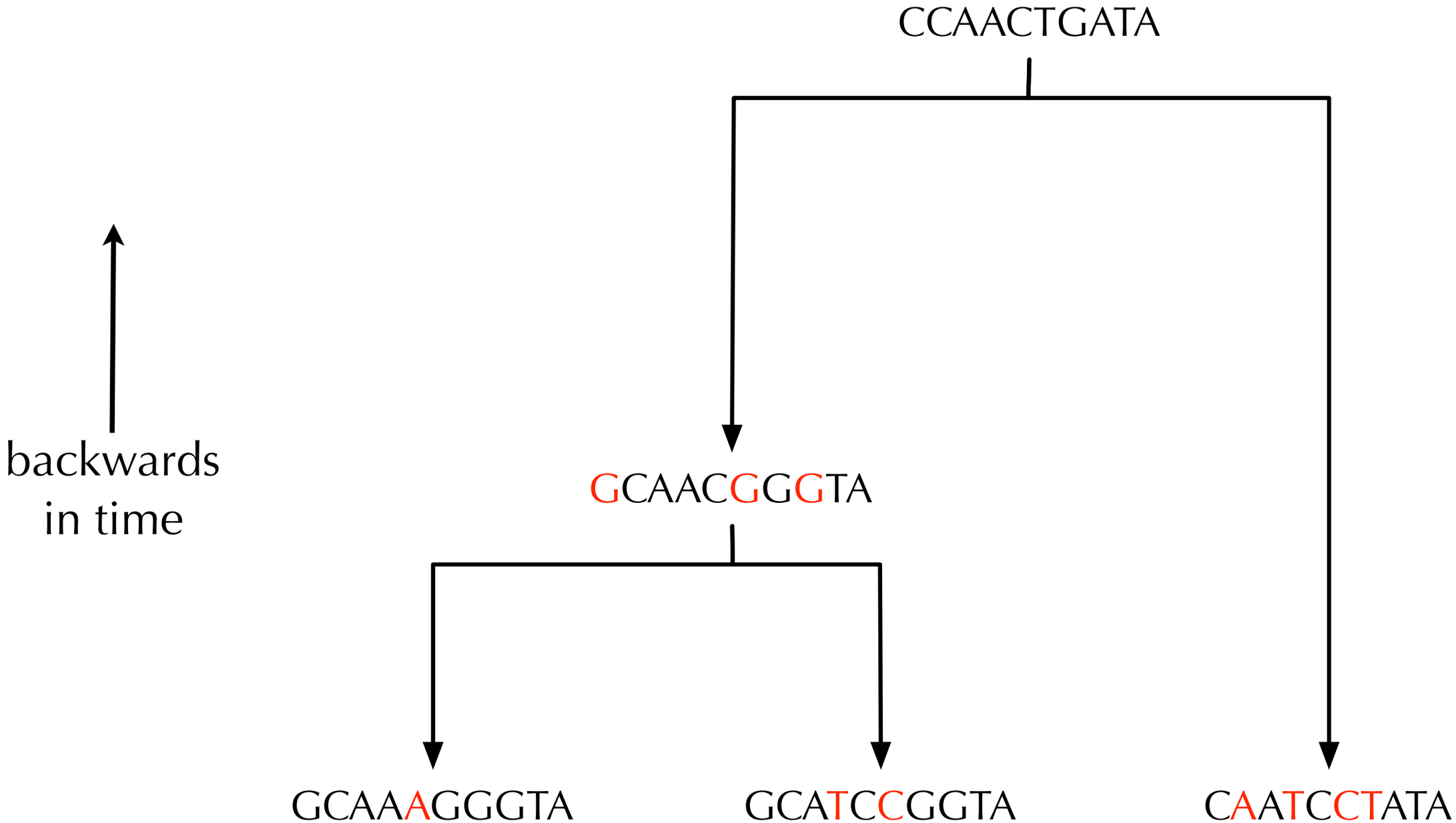
Ancestral Tree



Teh, Blundell, Elliott

[Kingman 1982]

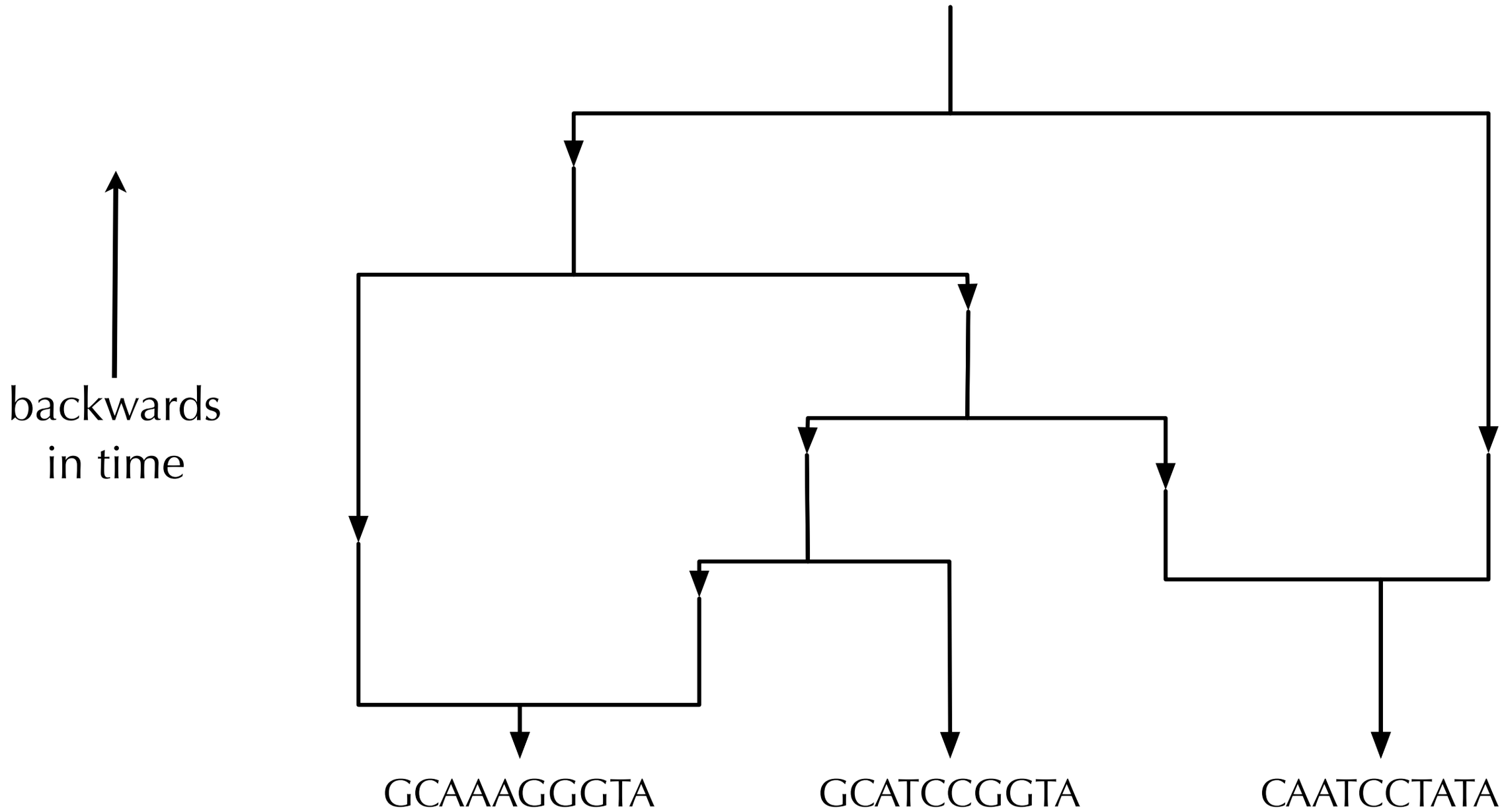
Ancestral Tree



Teh, Blundell, Elliott

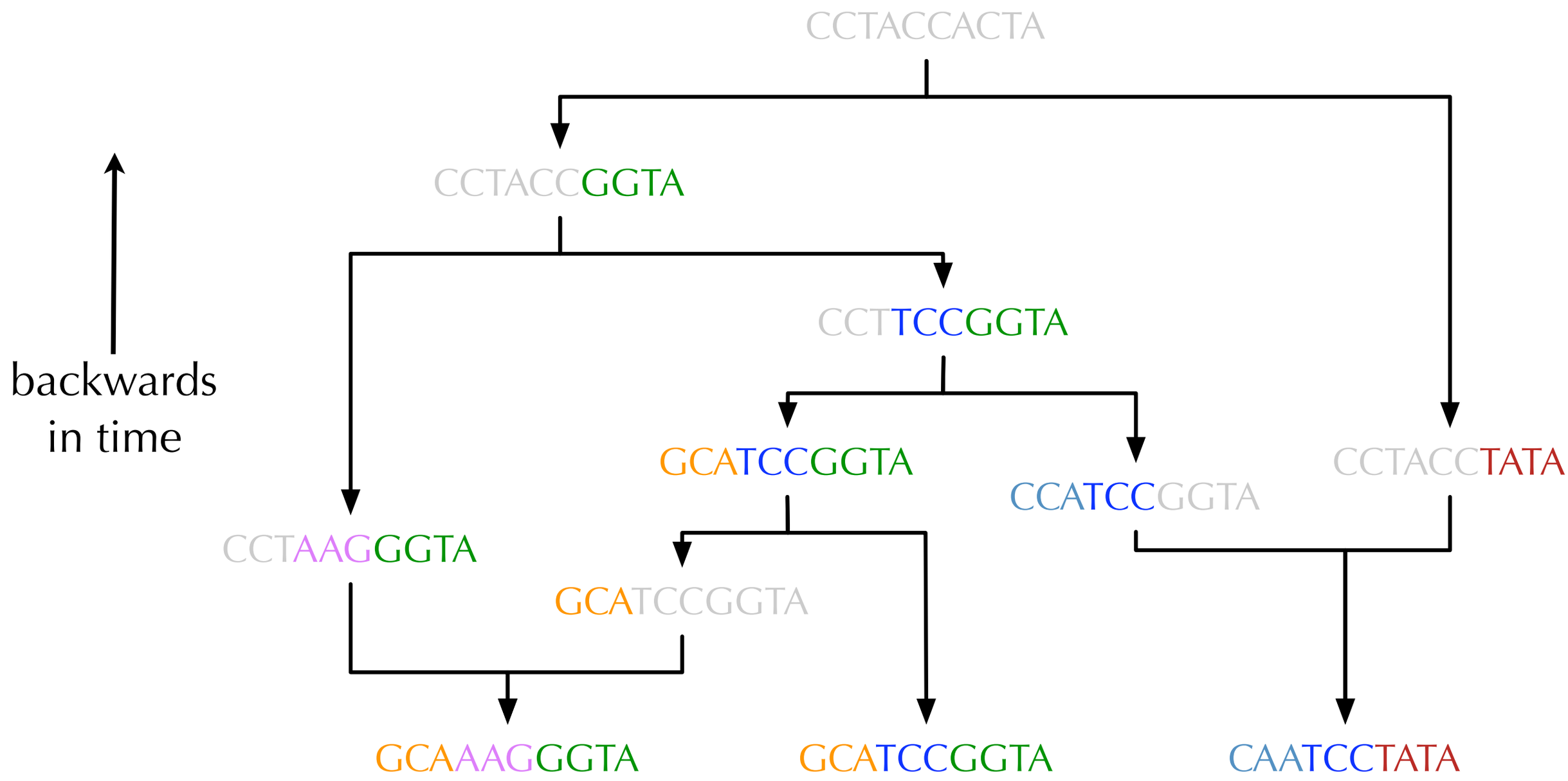
[Kingman 1982]

Ancestral Recombination Graph



[Hudson 1983]

Ancestral Recombination Graph



Teh, Blundell, Elliott

[Hudson 1983]

Mosaic Structure

- Simplification:
  - Blocks of recurring segments;
  - Each DNA sequence composed of multiple blocks.

➡ Hidden Markov models.

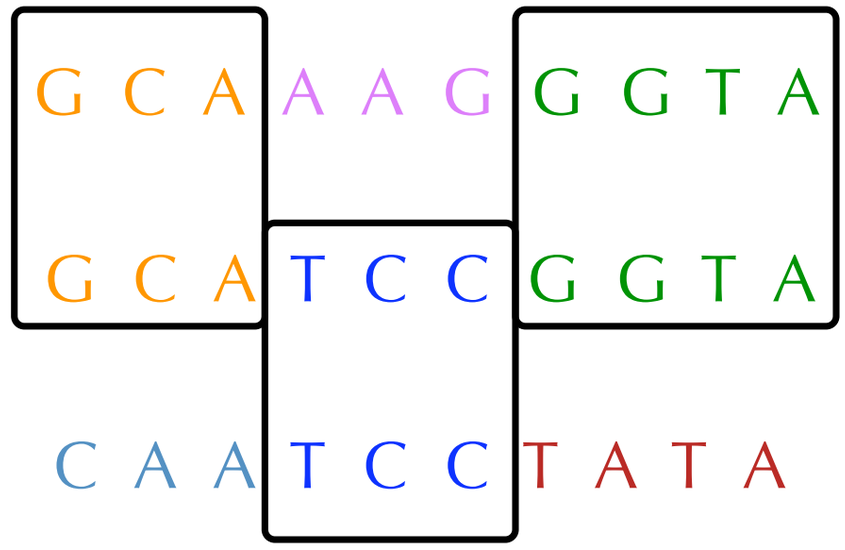


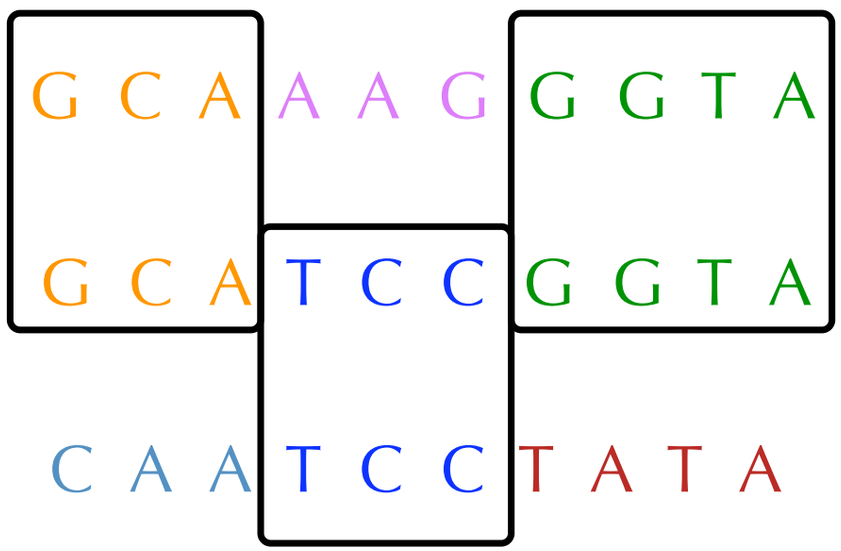
Figure from [Daly et al 2001]

Fragmentation-Coagulation Processes

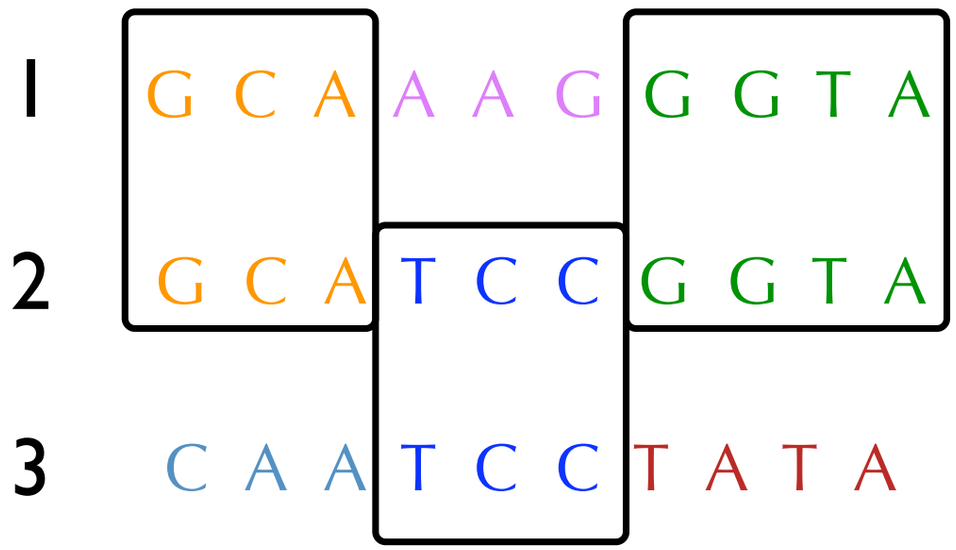
- No need for model selection---Bayesian nonparametric.
- No label switching problem---no labels.

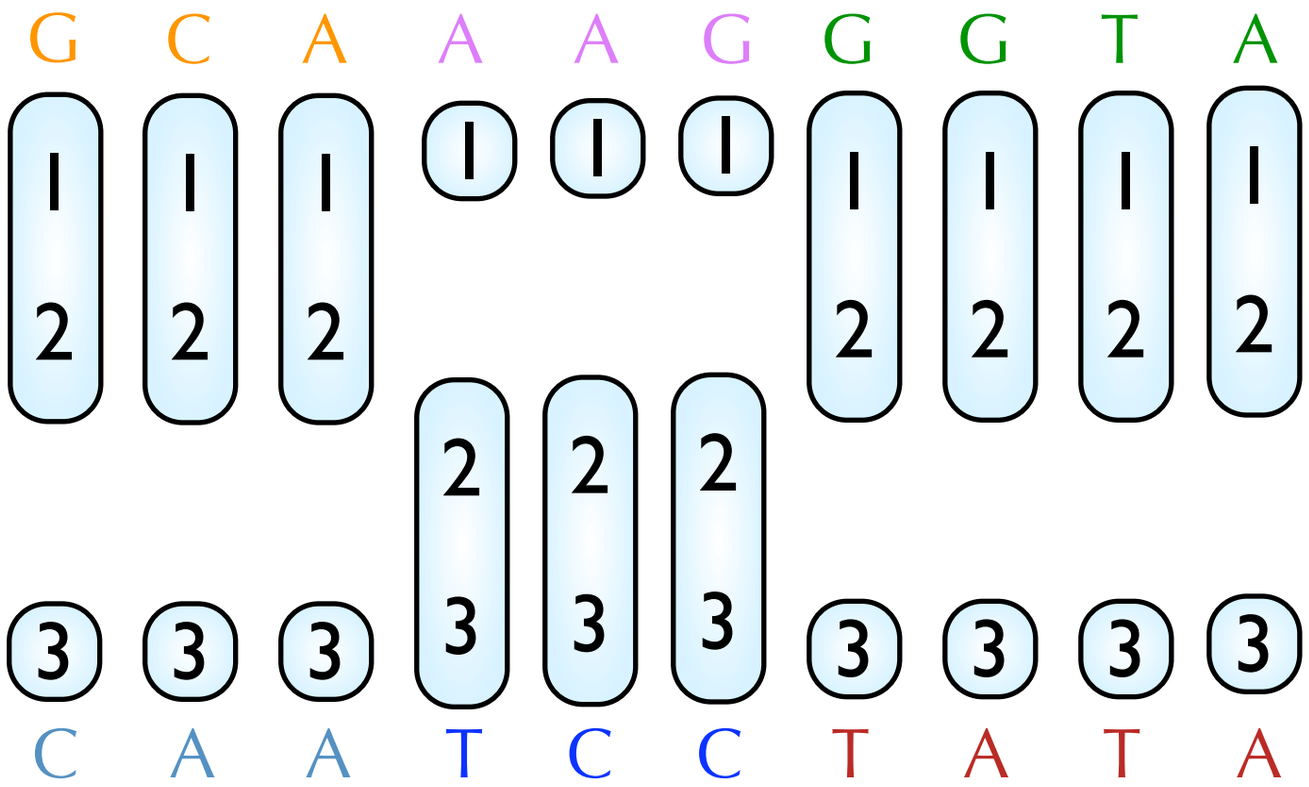
- Idea:
  - Use *unlabelled* partitions of sequences as basic representation.
  - Use a Markov process over partitions to model changing partition structure.

- Partition: set of clusters, e.g. {{1,2},{3}}
  - disjoint, non-empty, contains all sequences.
  - unlabelled



Markov Process over Partitions





$$\{\Pi_t : t = 1, 2, \ldots, T\}$$

$$\{\Pi_t : t \in [0, T]\}$$

Fragmentation-Coagulation Processes

Teh, Blundell, Elliott

Fragmentation-Coagulation Processes

1

3

6

2

7

4

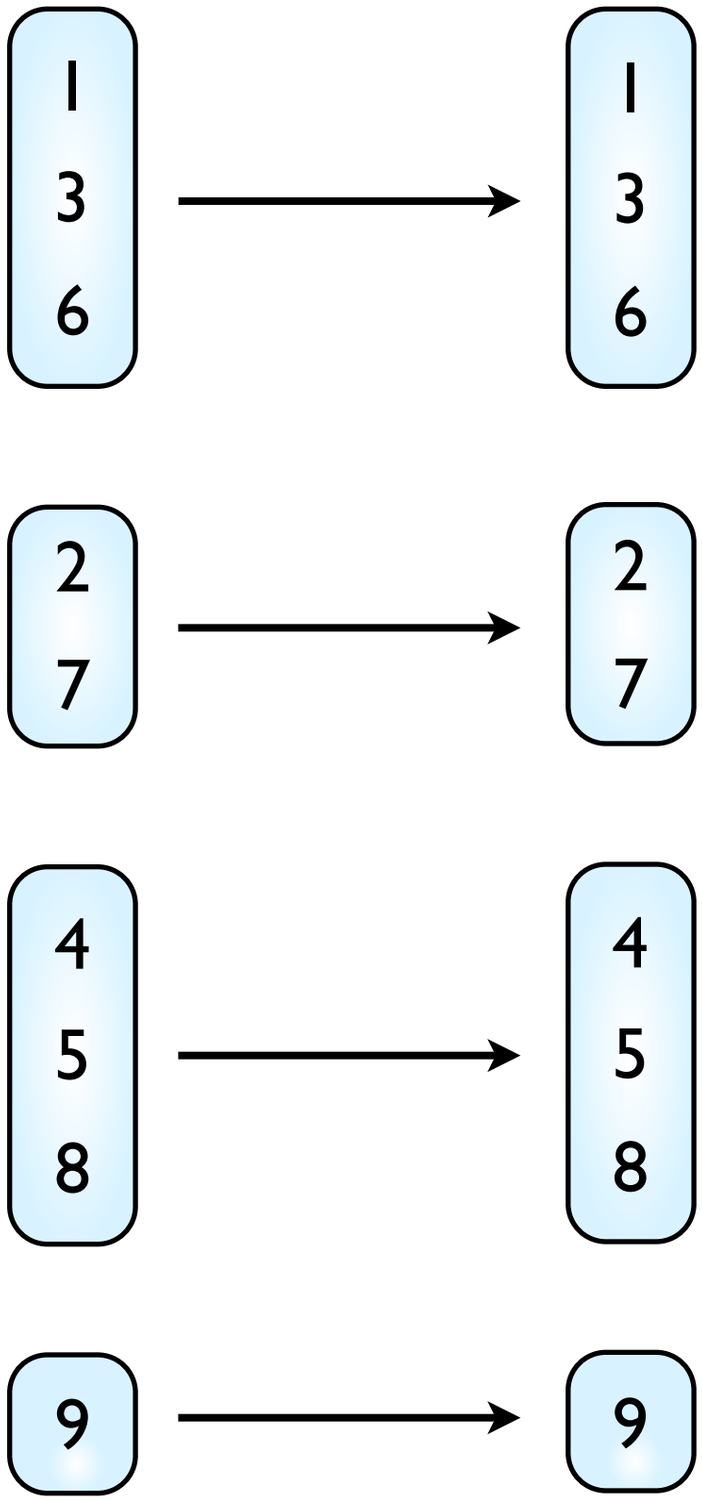
5

8

9

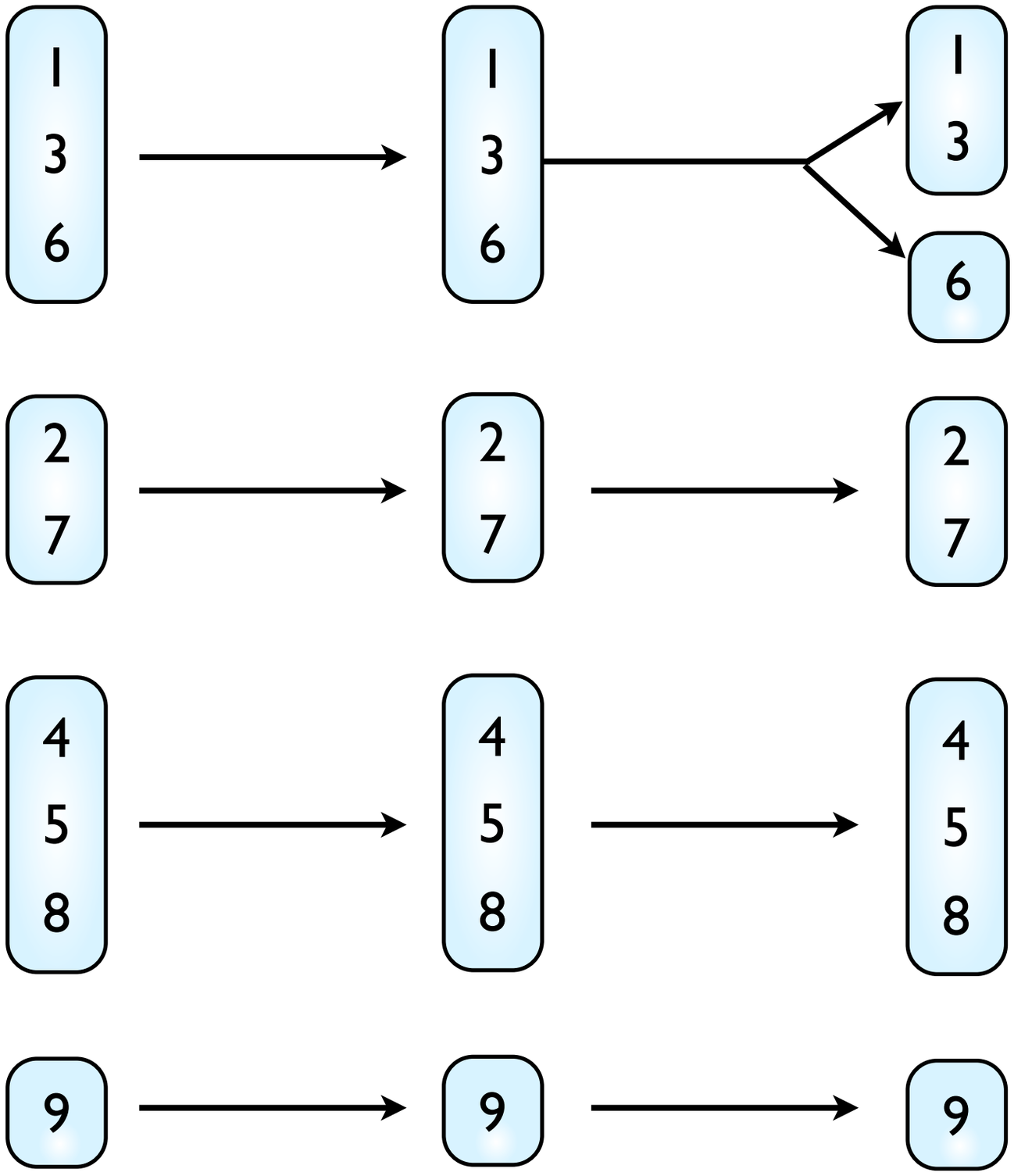
Teh, Blundell, Elliott

Fragmentation-Coagulation Processes



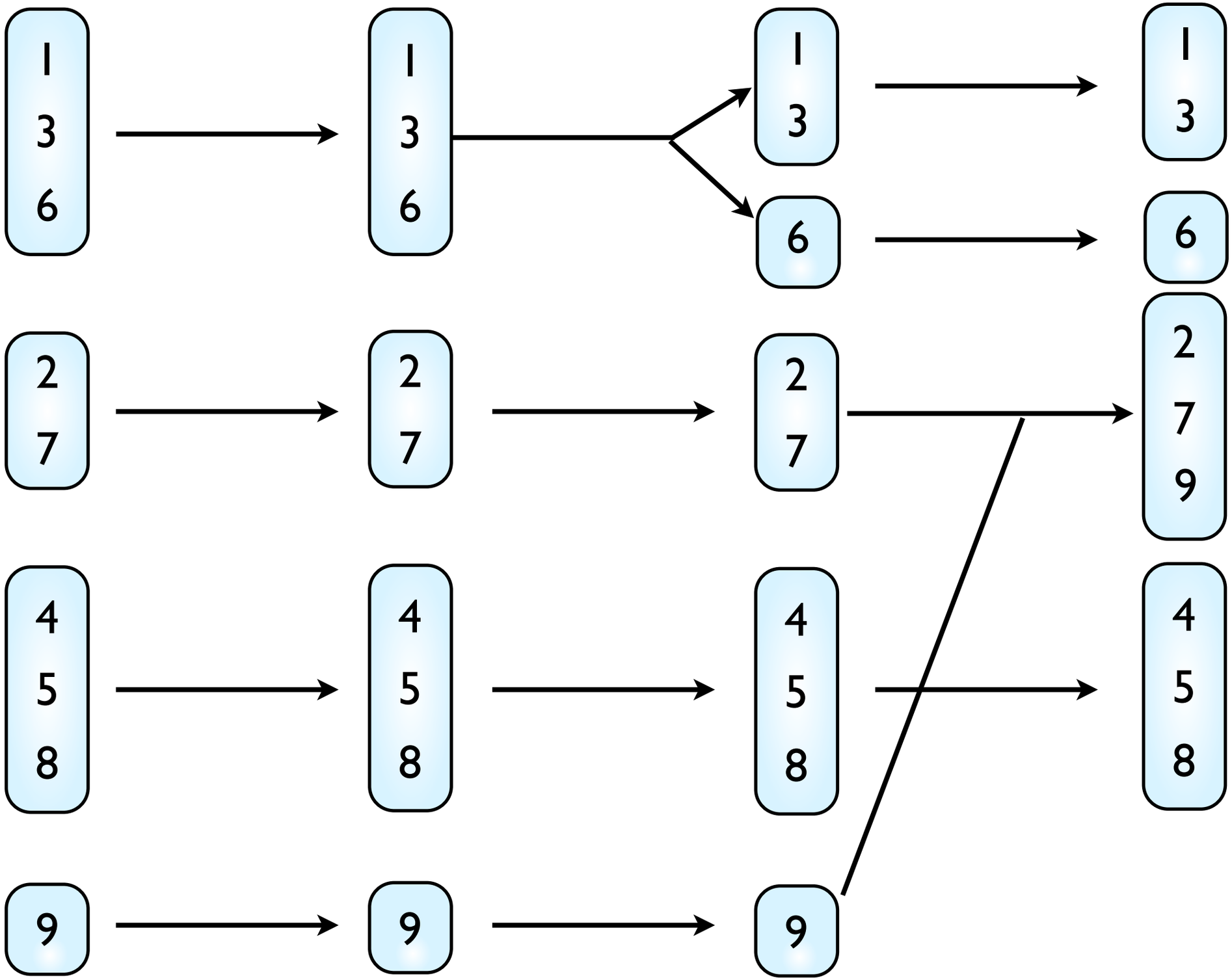
Teh, Blundell, Elliott

Fragmentation-Coagulation Processes



Teh, Blundell, Elliott

Fragmentation-Coagulation Processes

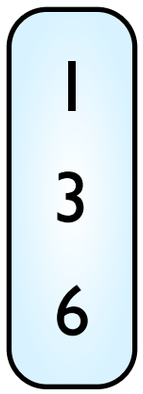


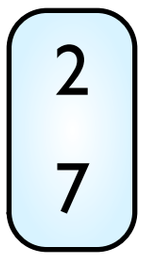
Teh, Blundell, Elliott

Fragmentation-Coagulation Processes

Teh, Blundell, Elliott

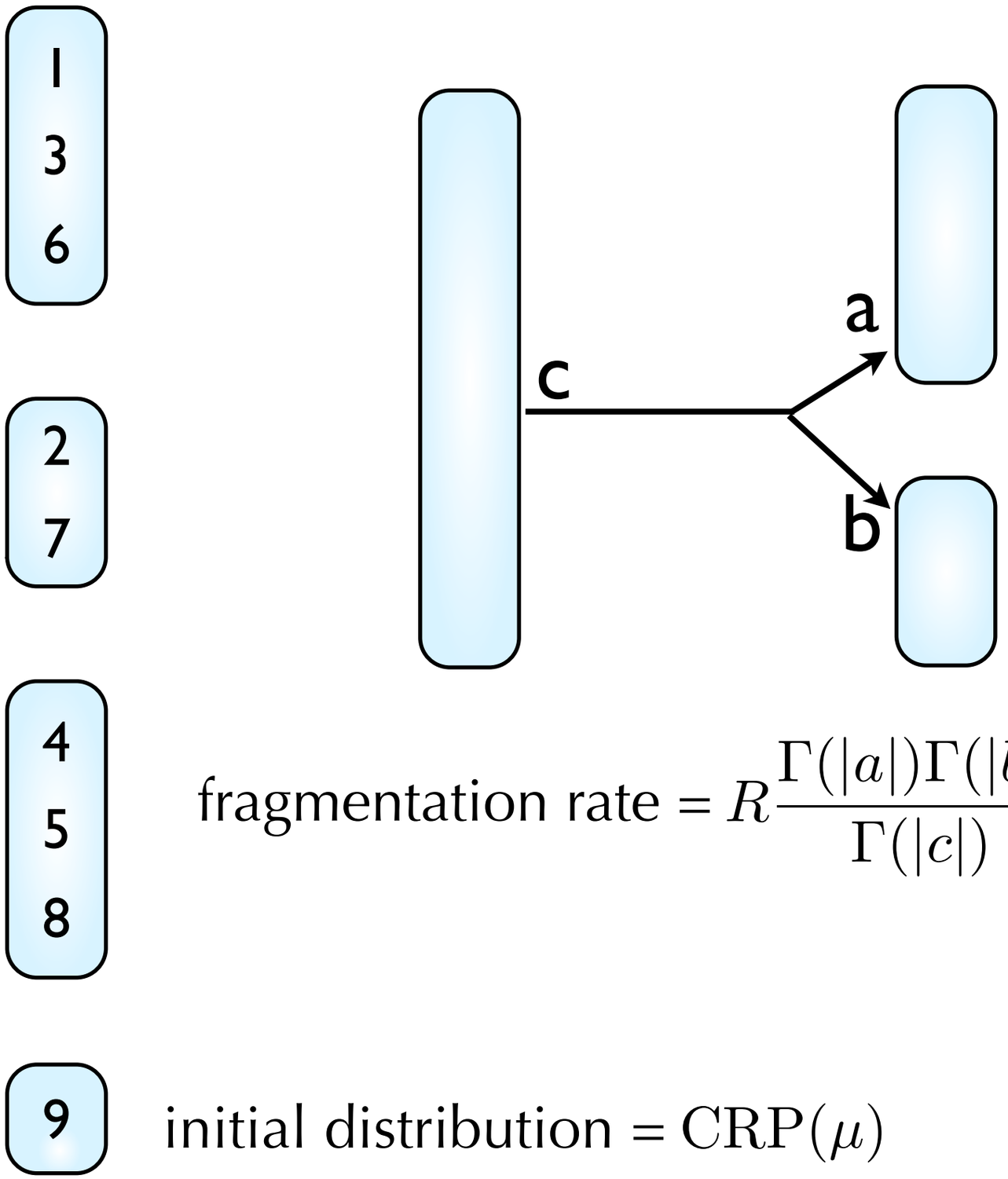
Fragmentation-Coagulation Processes





initial distribution = $\mathrm{CRP}(\mu)$

Fragmentation-Coagulation Processes

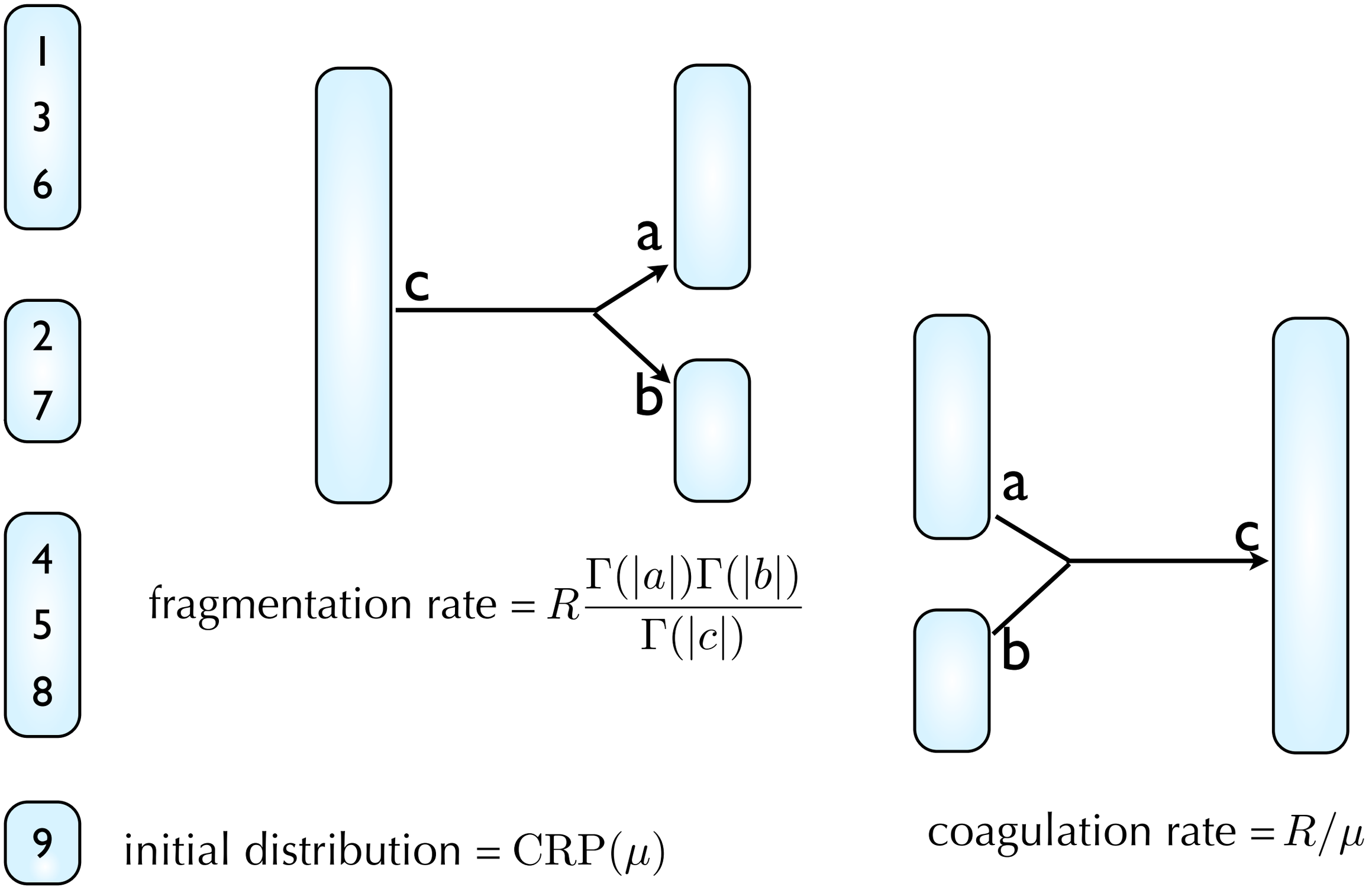


$$\text{fragmentation rate} = R\frac{\Gamma(|a|)\Gamma(|b|)}{\Gamma(|c|)}$$

$$\text{initial distribution} = \mathrm{CRP}(\mu)$$

Teh, Blundell, Elliott

Fragmentation-Coagulation Processes



$$\text{fragmentation rate} = R\frac{\Gamma(|a|)\Gamma(|b|)}{\Gamma(|c|)}$$

$$\text{initial distribution} = \mathrm{CRP}(\mu)$$

$$\text{coagulation rate} = R/\mu$$

Teh, Blundell, Elliott

Fragmentation-Coagulation Processes

- Markov.
- Stationary, with CRP(μ) as equilibrium distribution.
- Reversible.

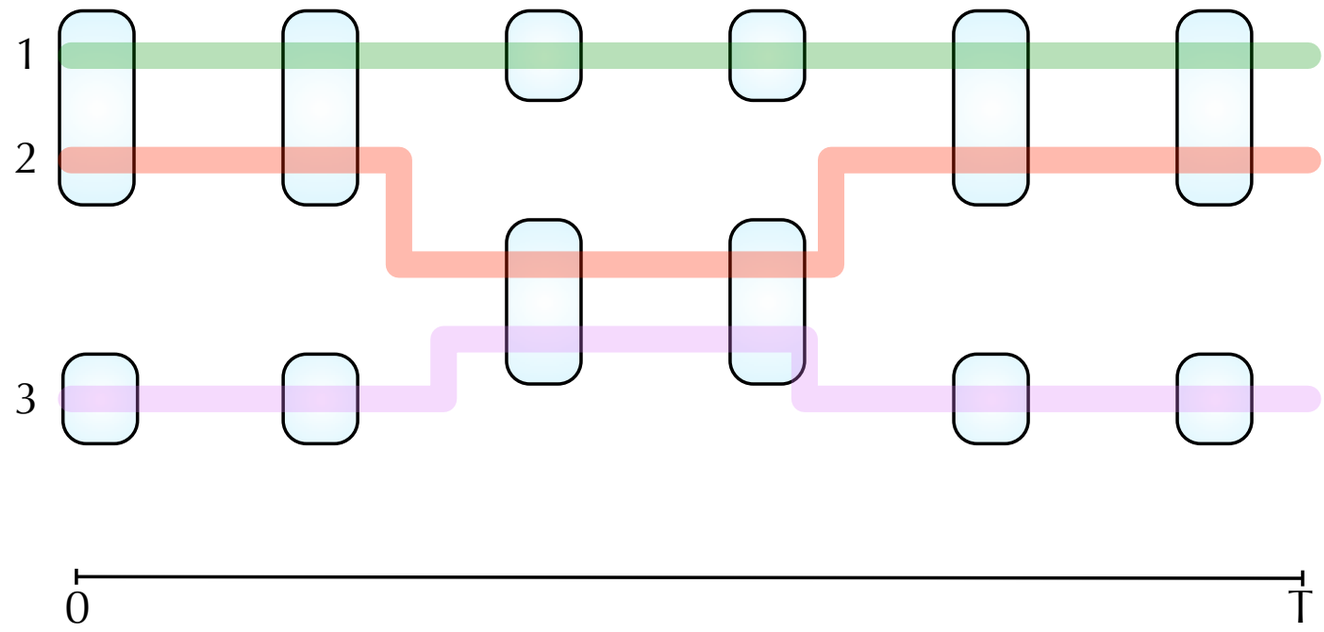
- Exchangeable.

- Dirichlet diffusion tree [Neal 2003] and Kingman's coalescent.

- Simplest example of exchangeable fragmentation-coalescence processes [Berestycki 2004].

Inference

- Gibbs sampling:
  - Resample trajectory of one sequence at each iteration.



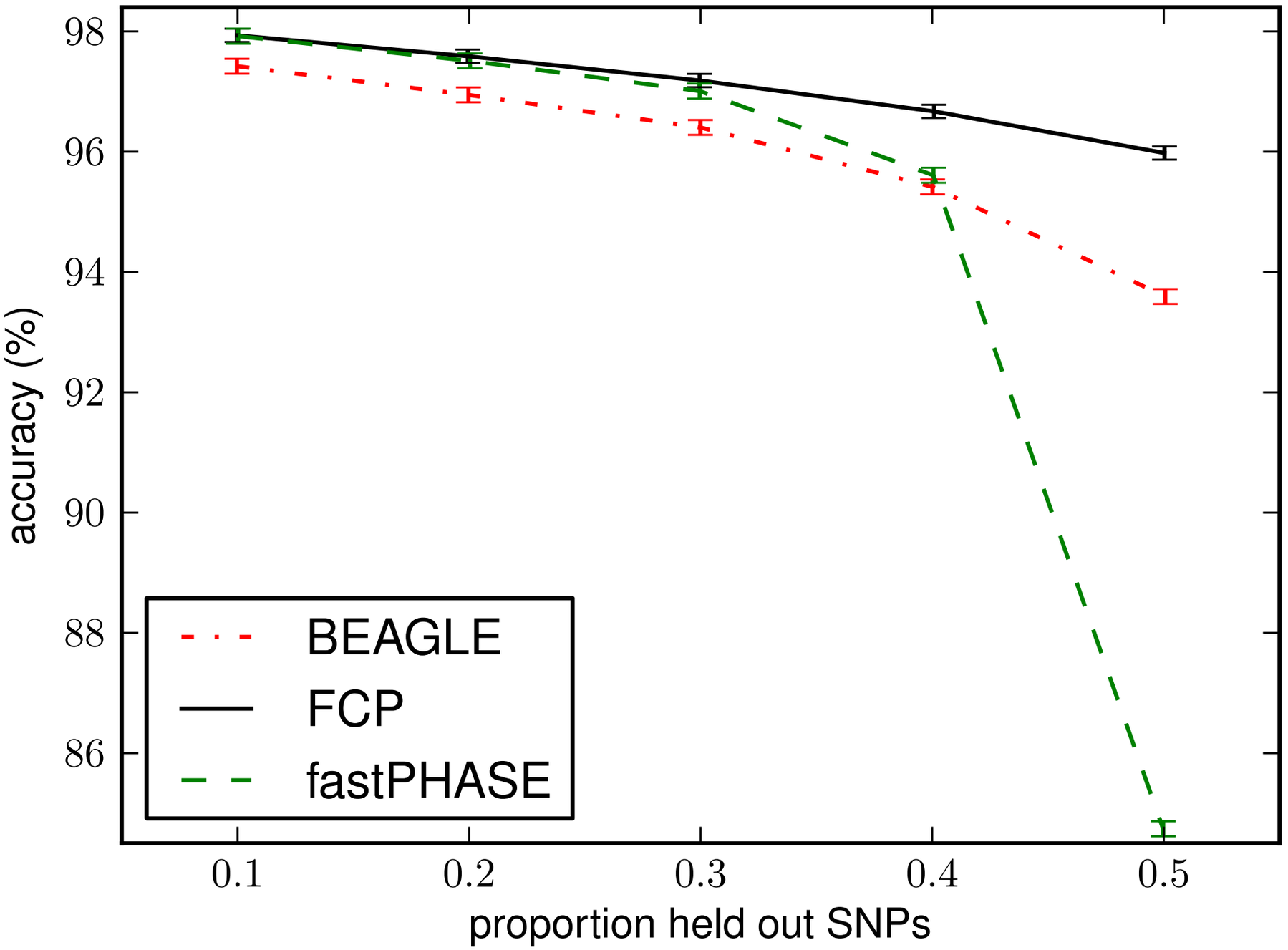
- Dealing with continuous time dynamics:
  - Uniformisation based auxiliary variable Gibbs [Rao & Teh UAI 2011].
  - Forward filtering-backward sampling.

Inference

- Gibbs sampling:
  - Resample trajectory of one sequence at each iteration.

- Dealing with continuous time dynamics:
  - Uniformisation based auxiliary variable Gibbs [Rao & Teh UAI 2011].
  - Forward filtering-backward sampling.

Imputation Results



Teh, Blundell, Elliott

Summary

Poster **T092**

- Modelling the mosaic structure of genetic variations.
- Fragmentation-coagulation processes.
- Bayesian nonparametrics.
- Label switching problem.
- State-of-the-art results.

- Future work: scaling up, and other statistical genetics applications.

Teh, Blundell, Elliott

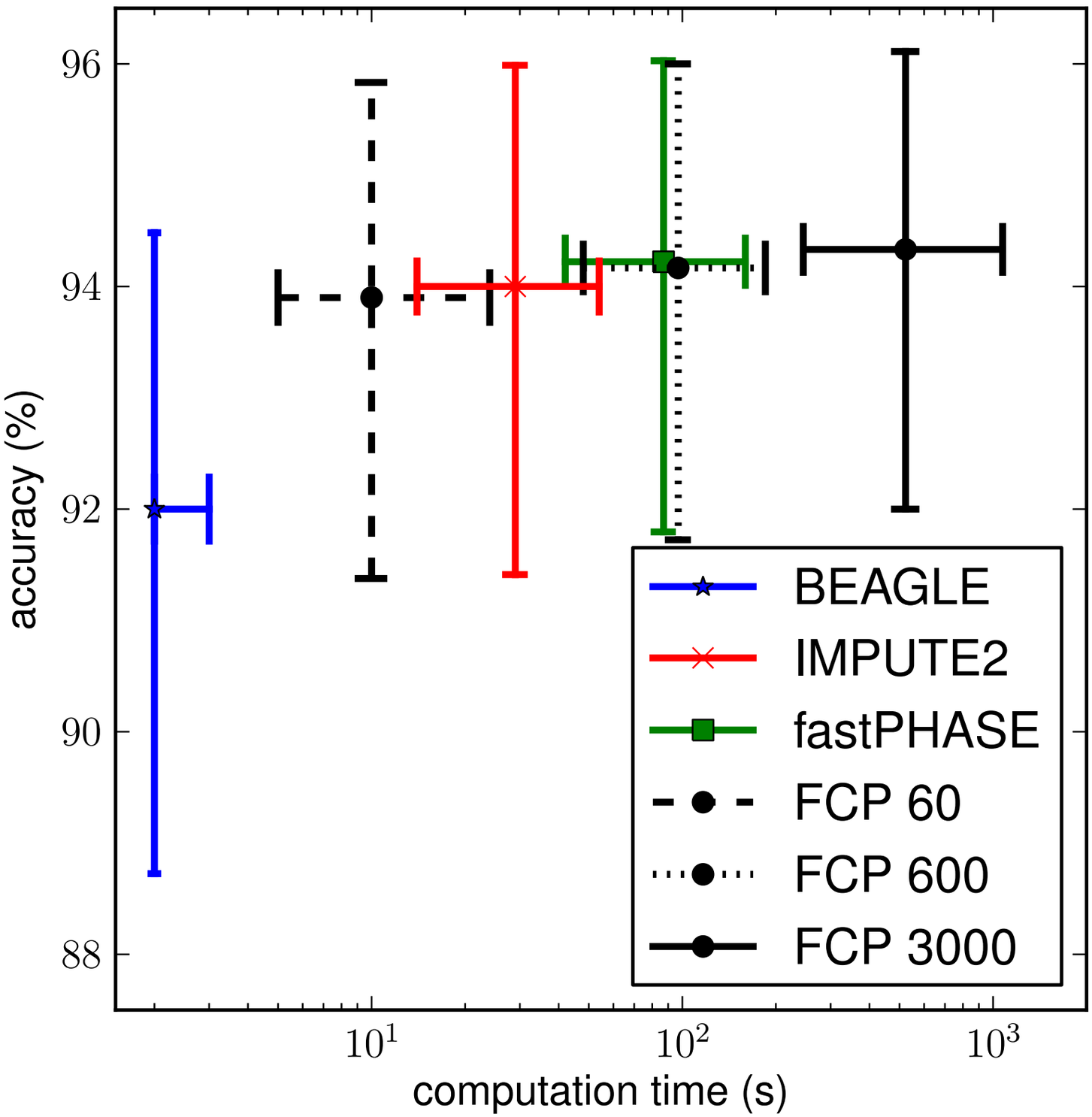
Poster **T092**

Thank You!

- Vinayak Rao and Andriy Mnih
- Chris Holmes, Gil McVean, Lancelot James
- NIPS organisers and audience

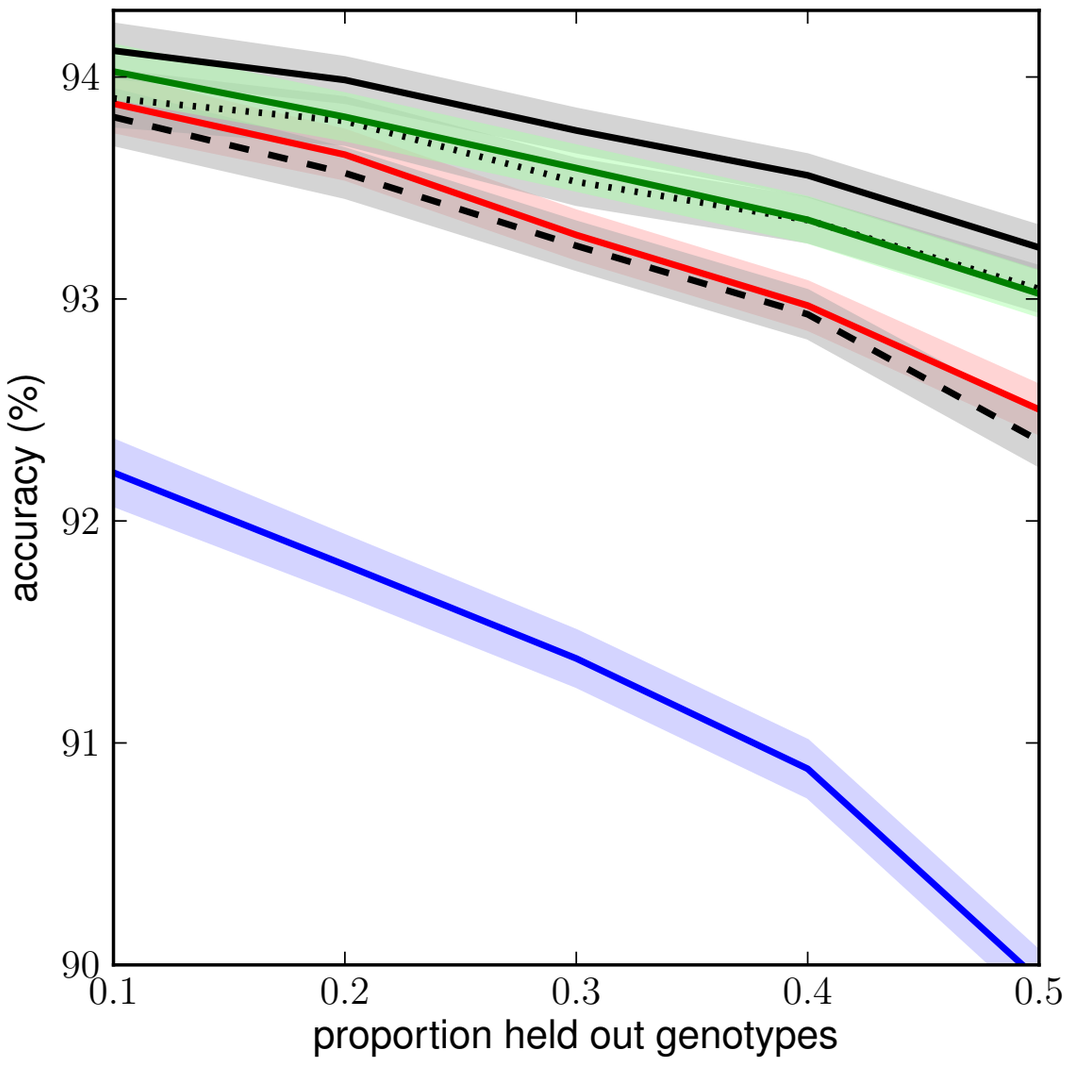
Appendix

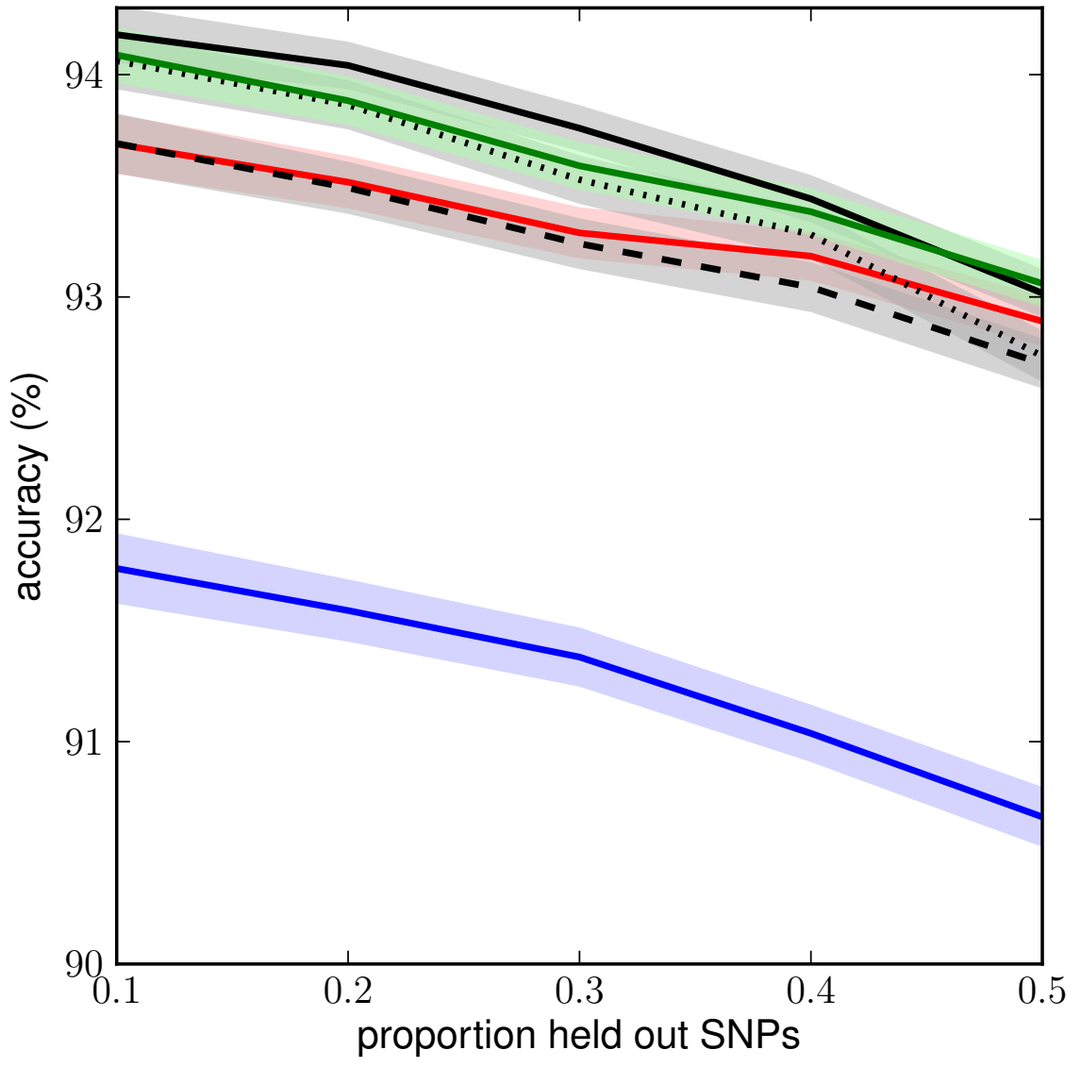
Imputation Experiments: Unphased Data



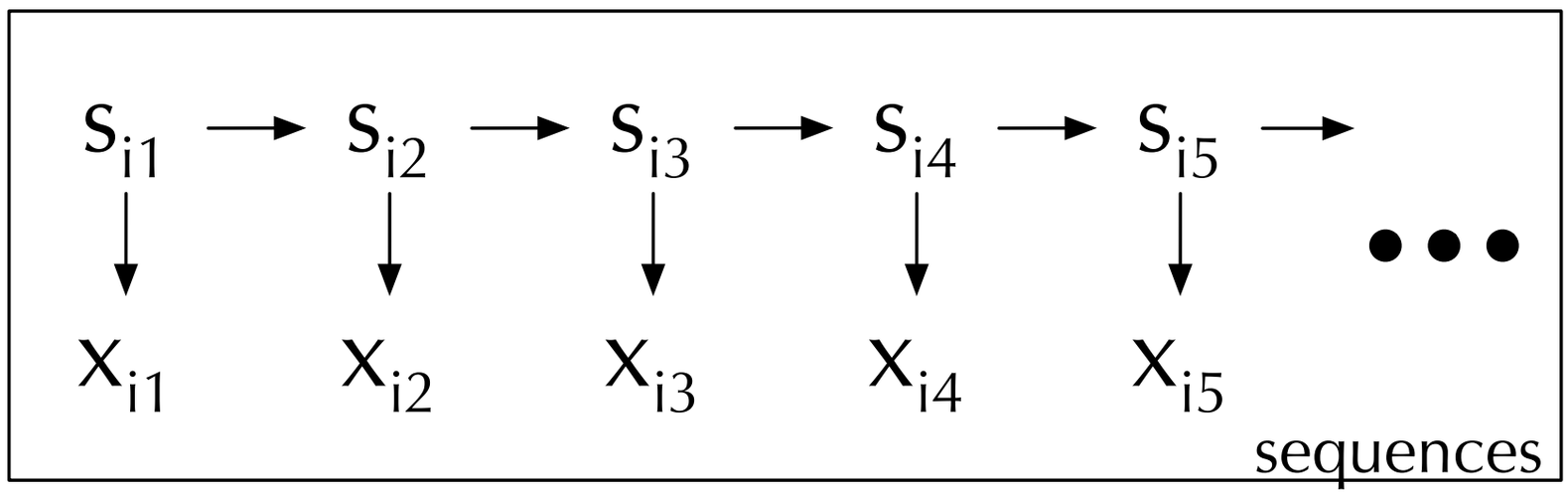
Teh, Blundell, Elliott

Imputation Experiments: Unphased Data



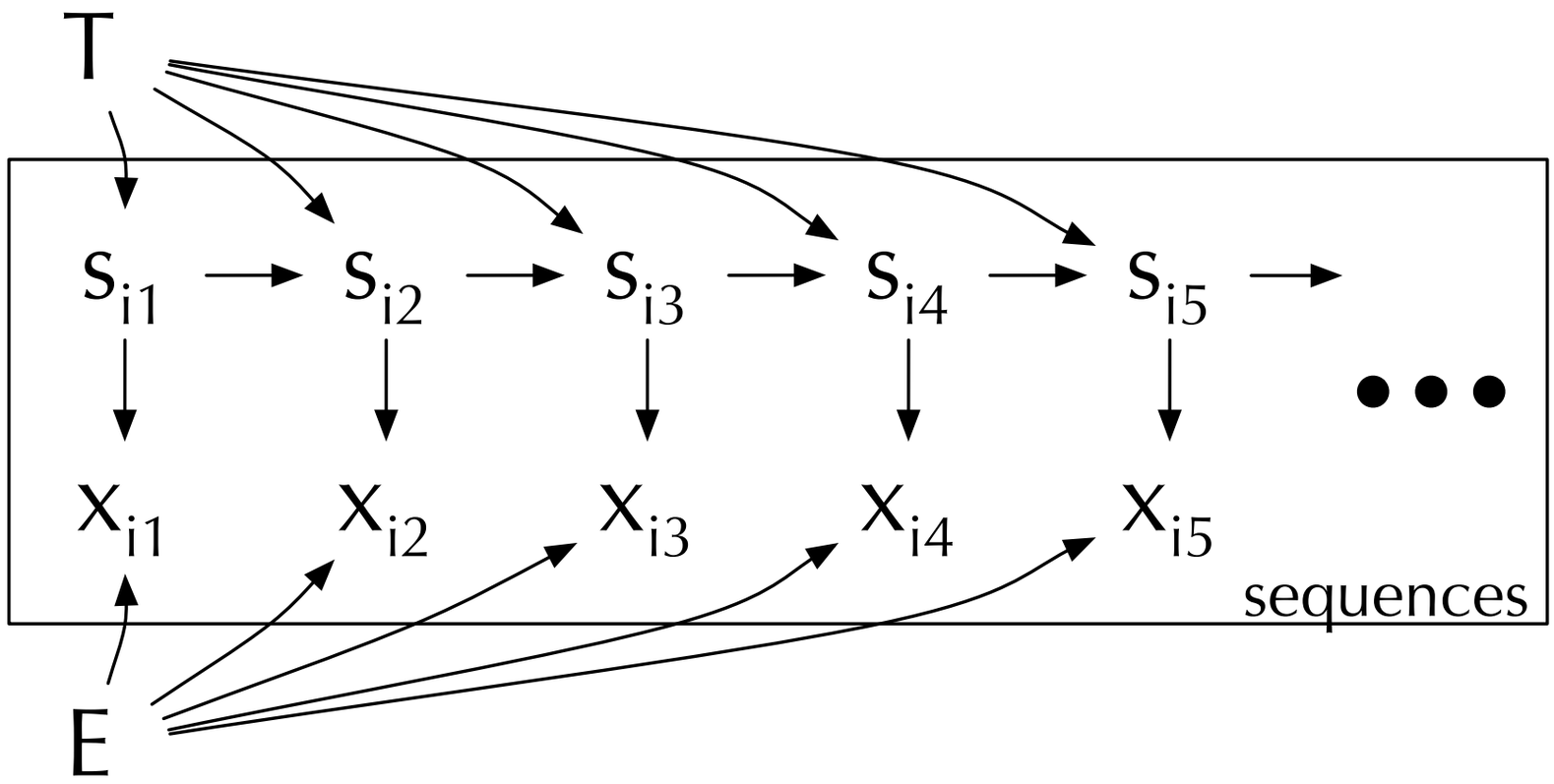


Hidden Markov Models



[Daly et al 2001], [Scheet & Stephens 2006]

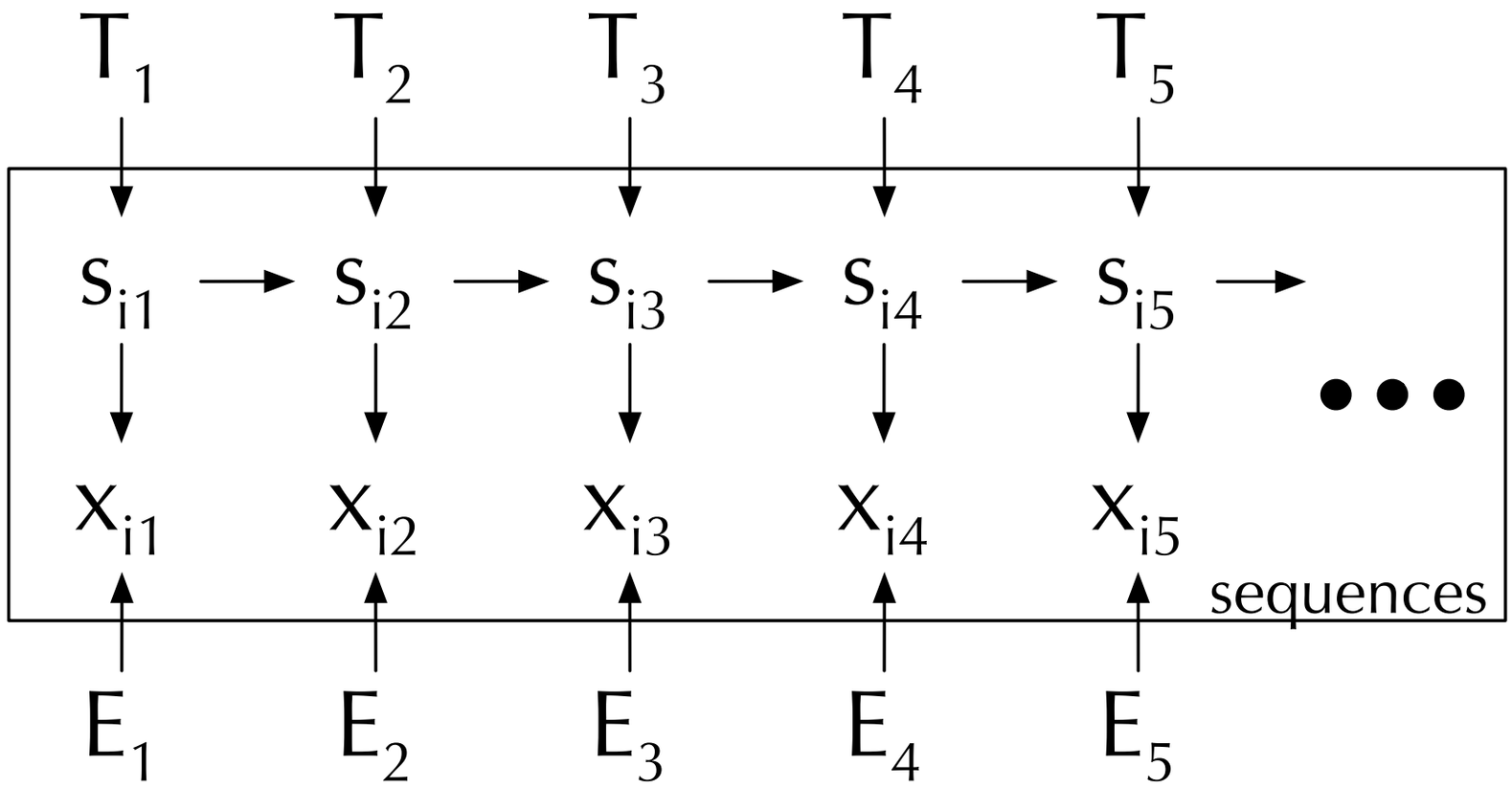
Hidden Markov Models



- Typical: stationary with shared transition and emission probabilities.

[Daly et al 2001], [Scheet & Stephens 2006]

Hidden Markov Models



- Typical: stationary with shared transition and emission probabilities.
- Here: non-stationary with location specific transition and emissions.

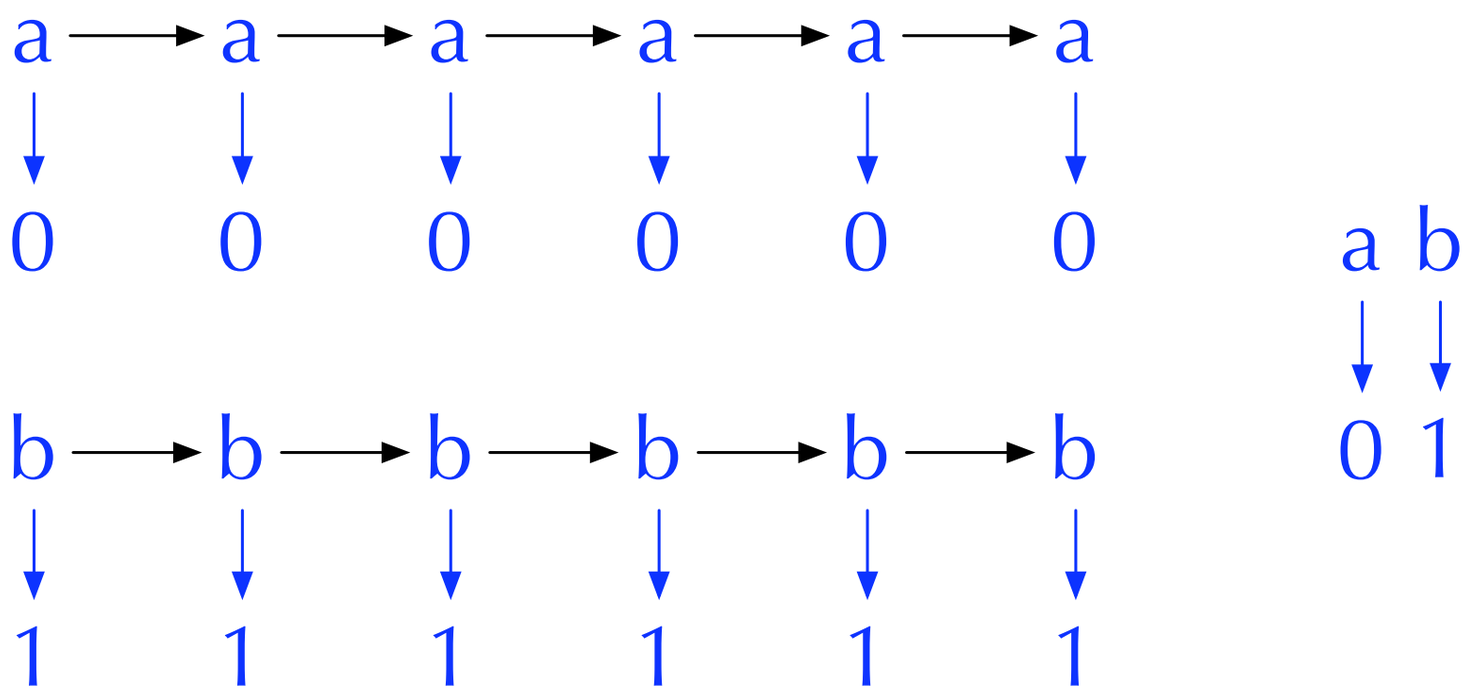
[Daly et al 2001], [Scheet & Stephens 2006]

HMM Label Switching Problem

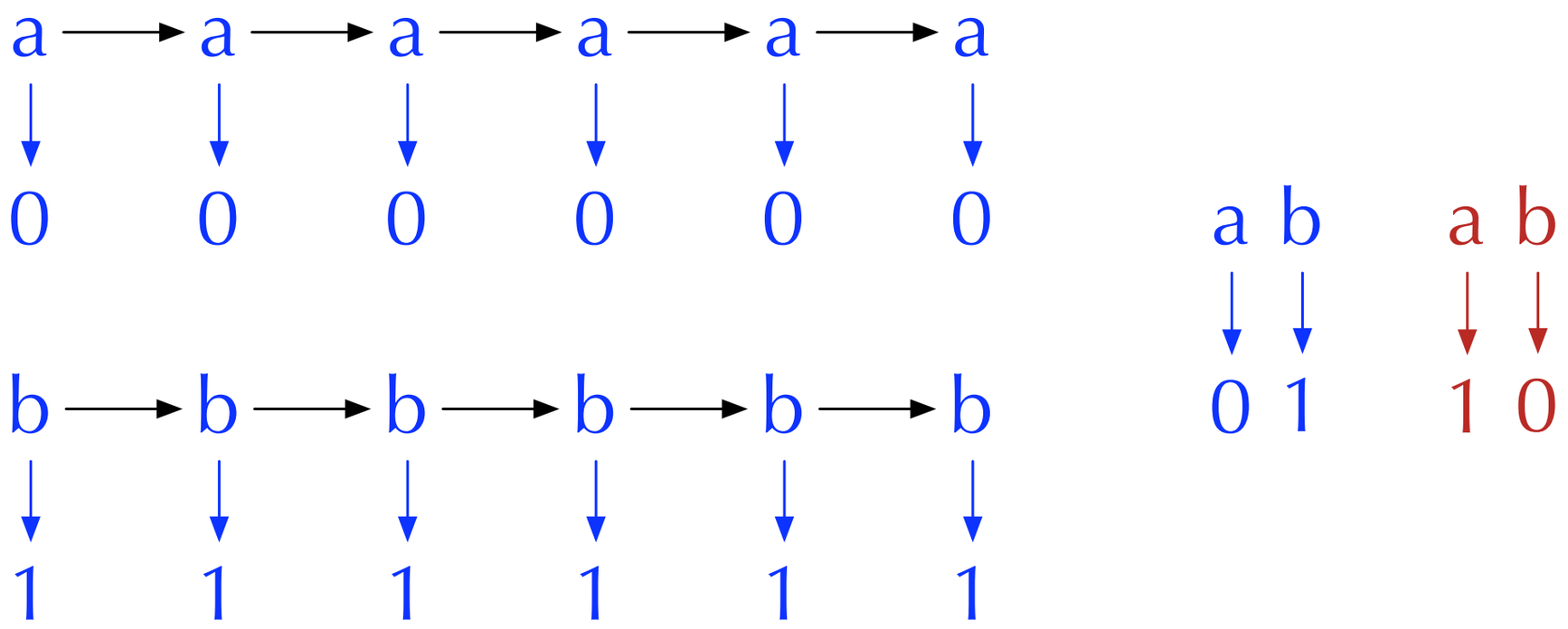
0 0 0 0 0 0

1 1 1 1 1 1

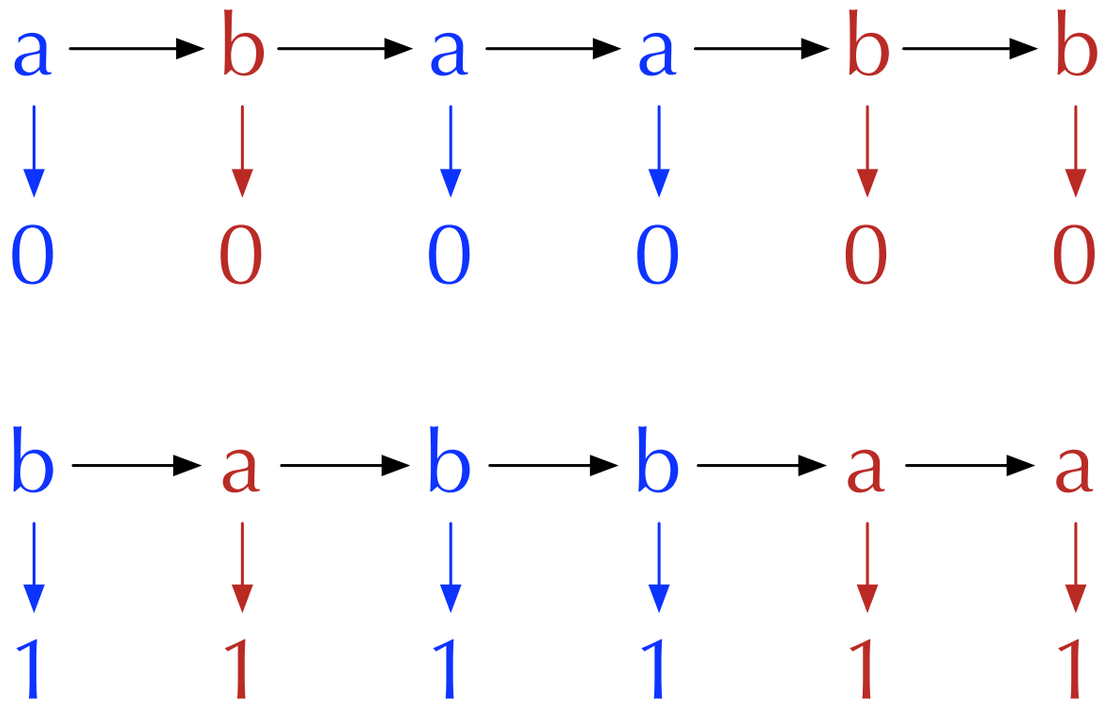
HMM Label Switching Problem

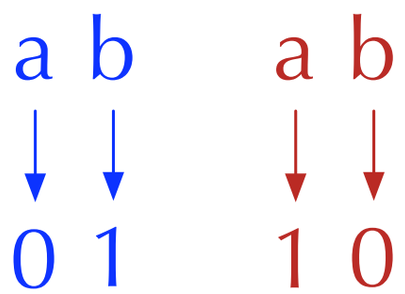


HMM Label Switching Problem

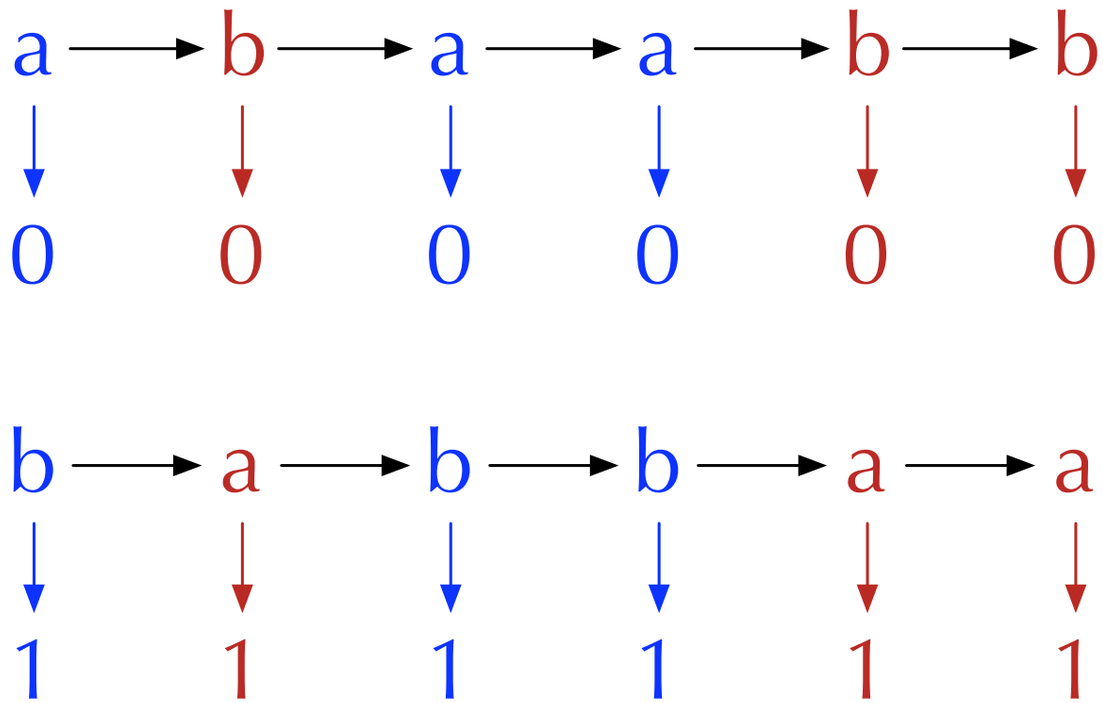


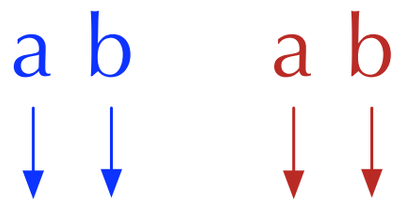
HMM Label Switching Problem

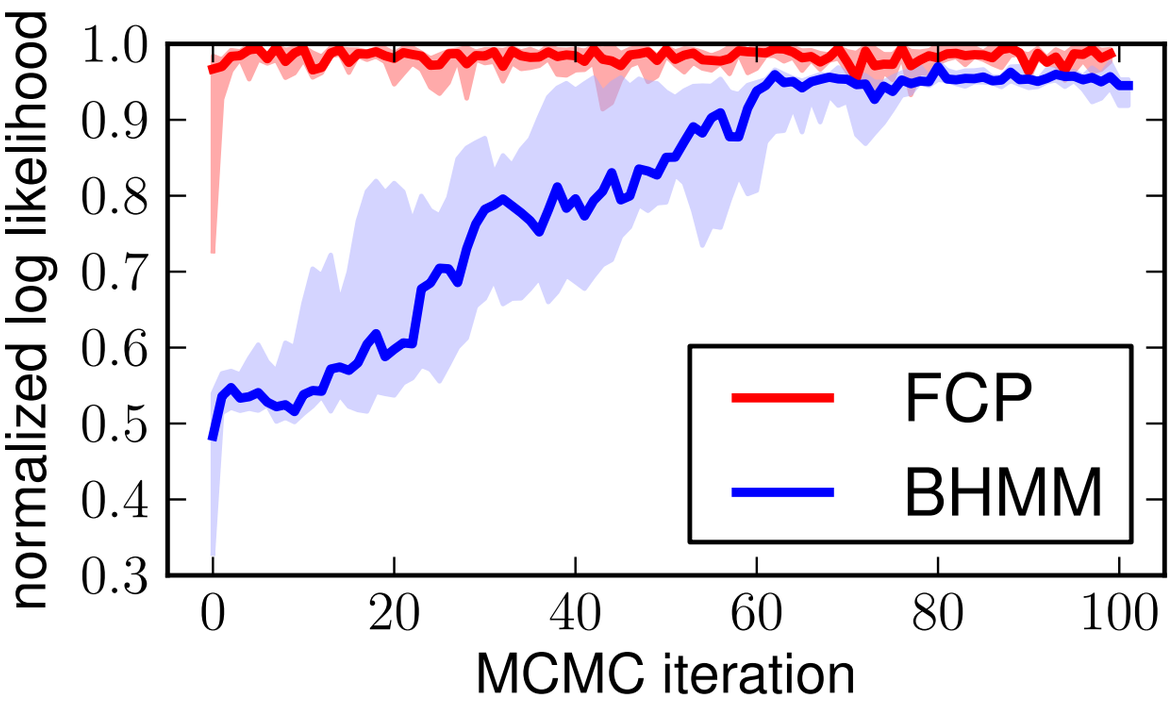


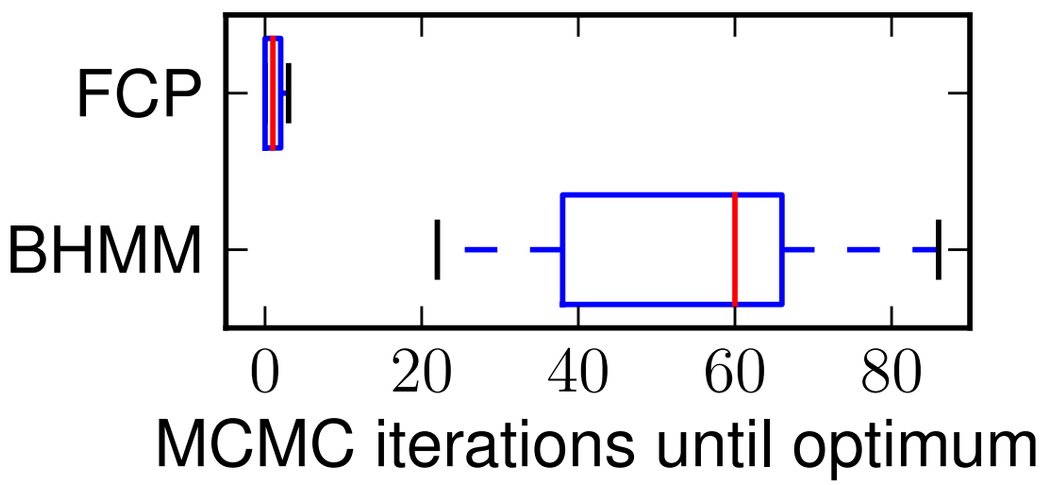


HMM Label Switching Problem



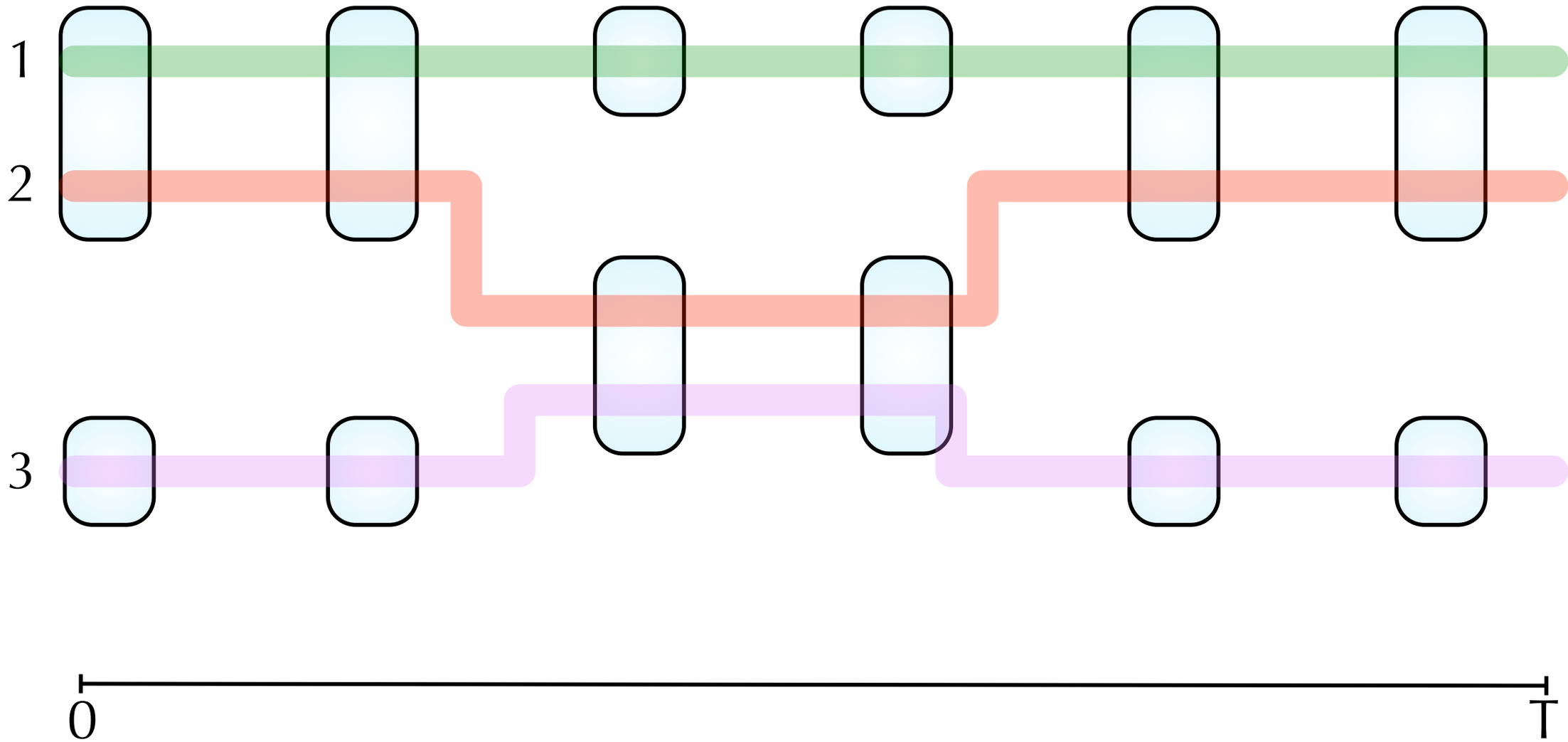




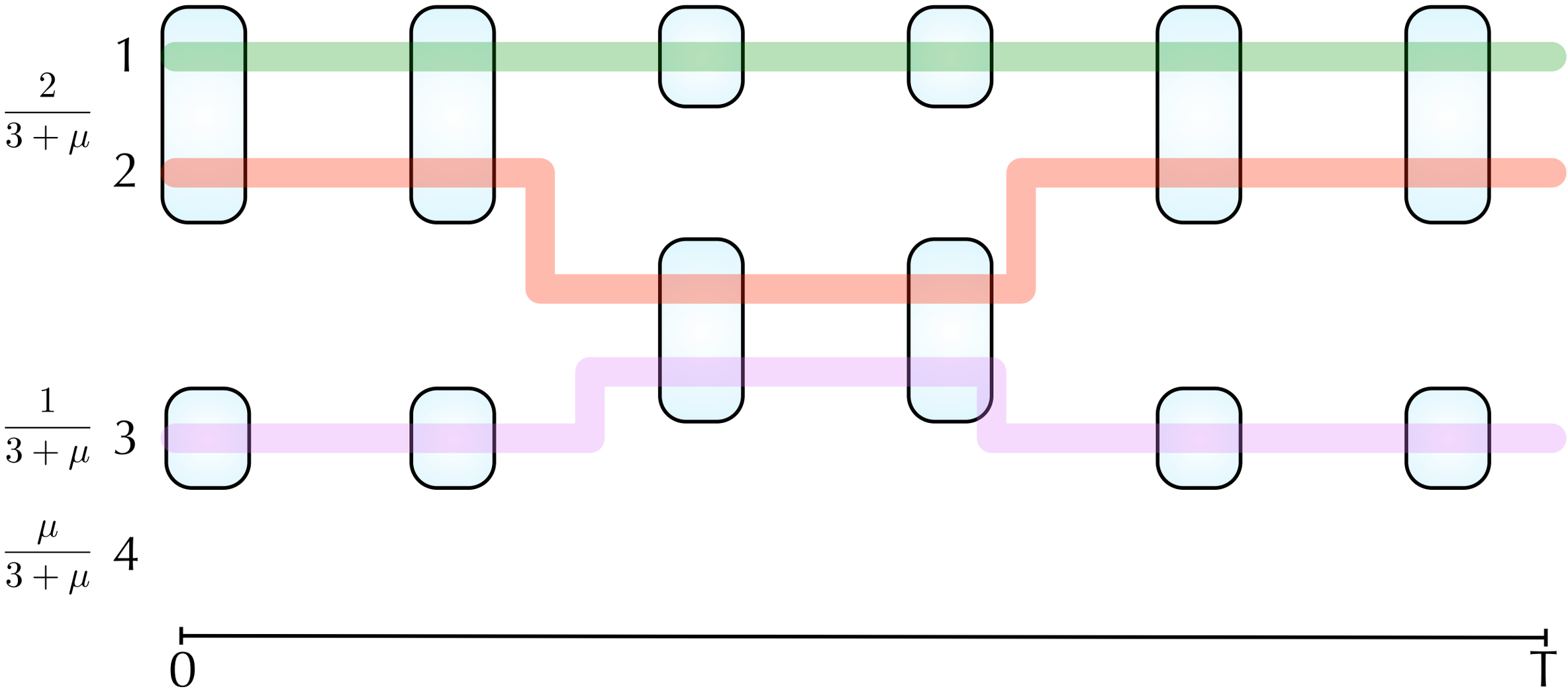


Teh, Blundell, Elliott

Chinese Restaurant Process Through Time

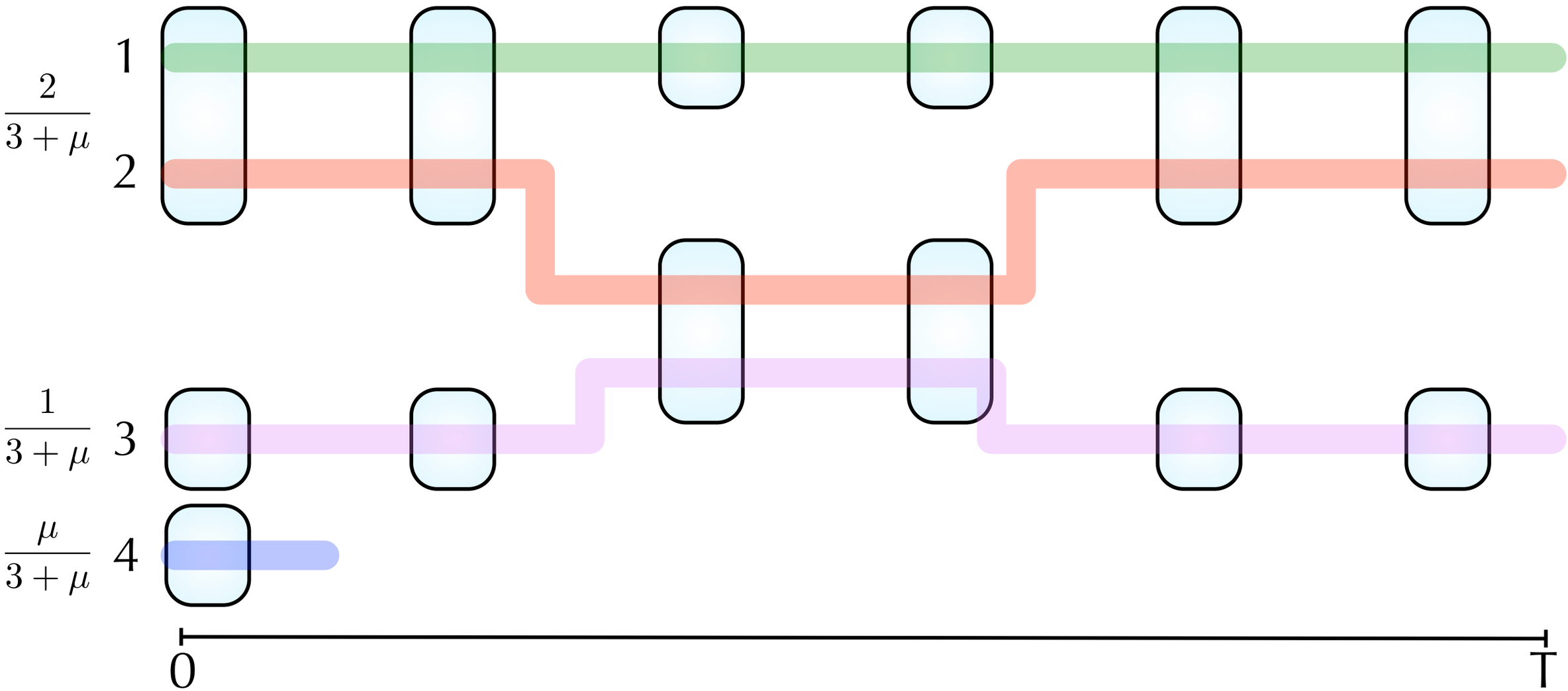


Chinese Restaurant Process Through Time



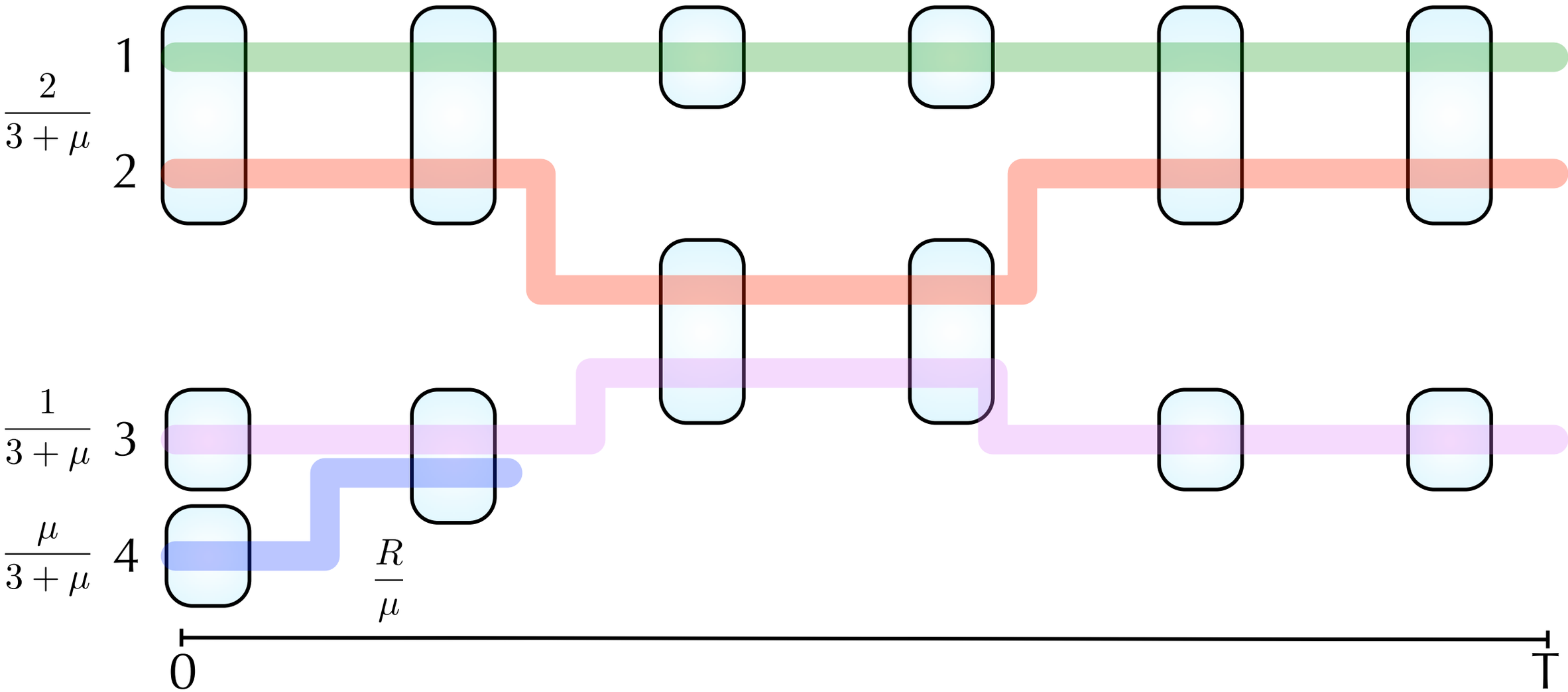
Teh, Blundell, Elliott

Chinese Restaurant Process Through Time



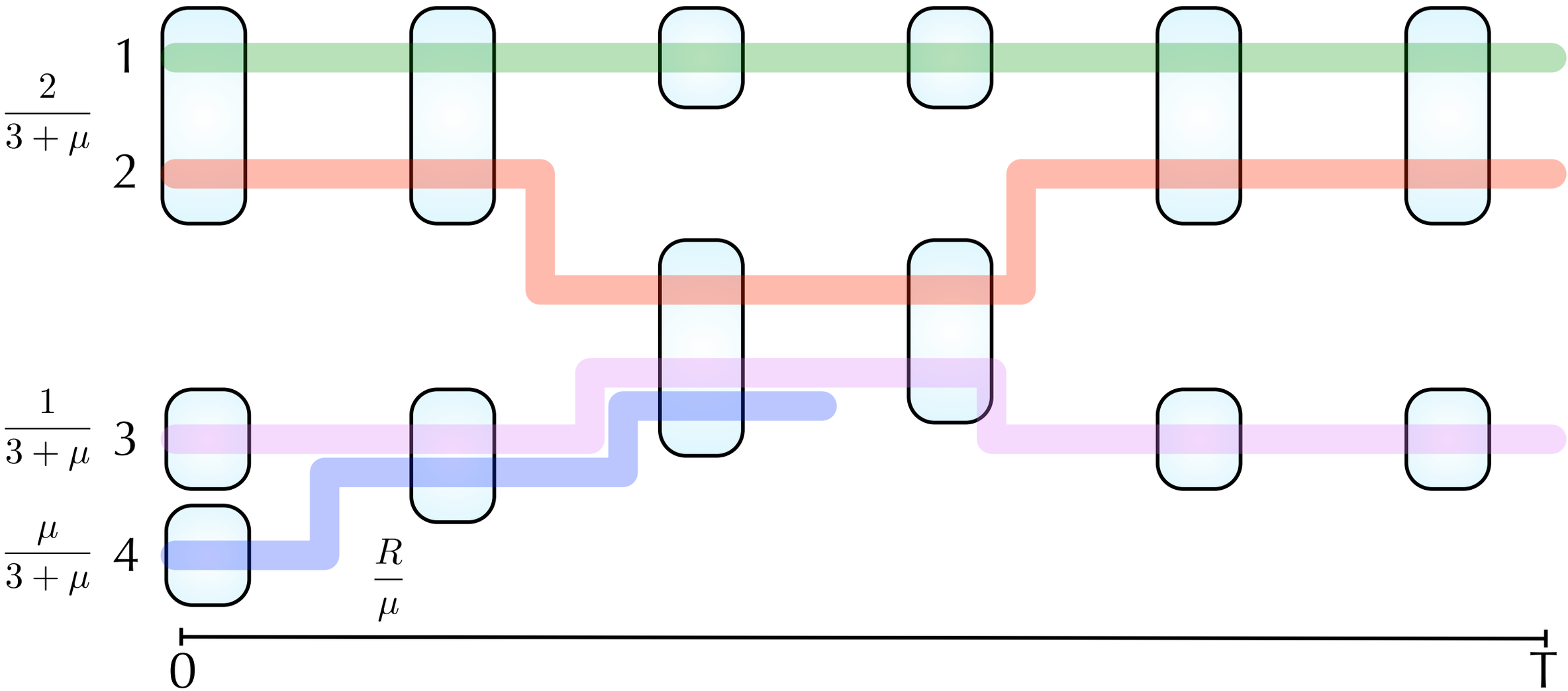
Teh, Blundell, Elliott

Chinese Restaurant Process Through Time



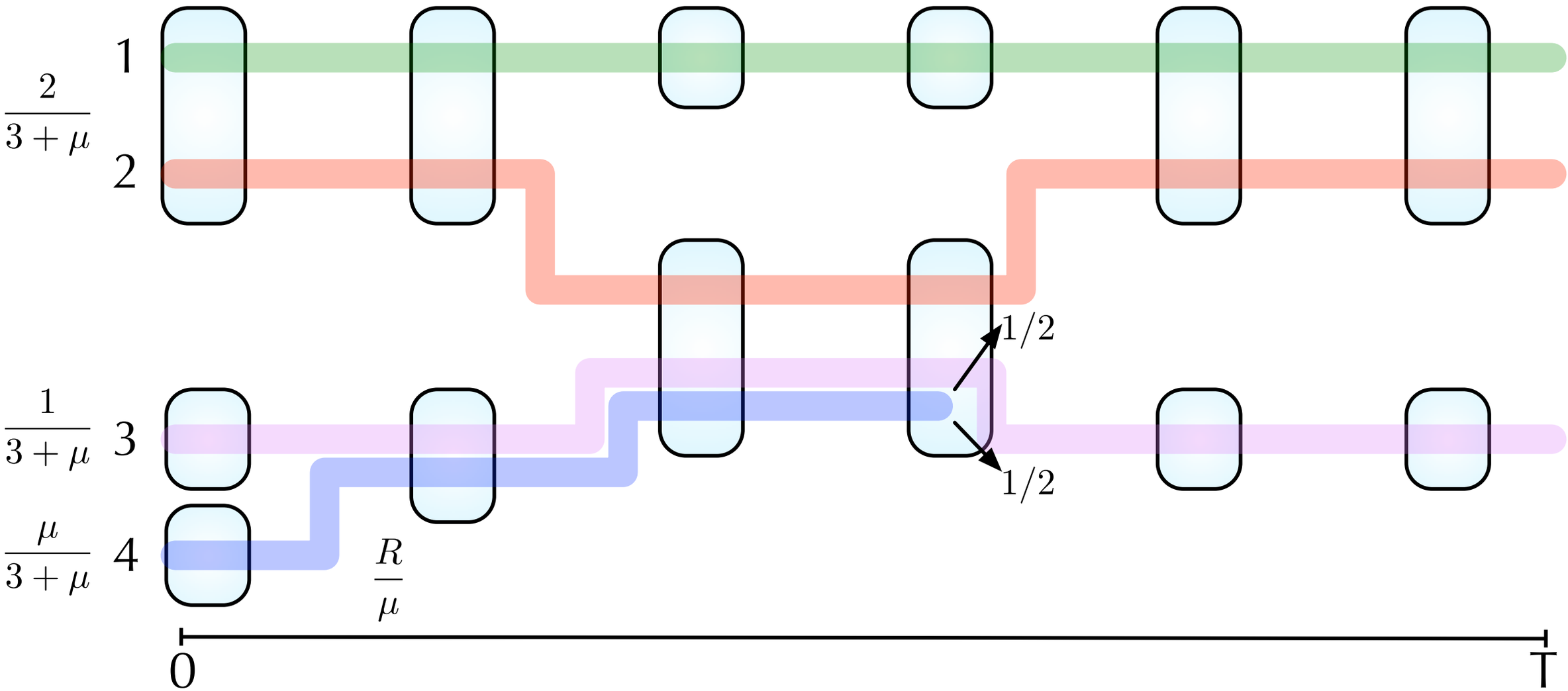
Teh, Blundell, Elliott

Chinese Restaurant Process Through Time



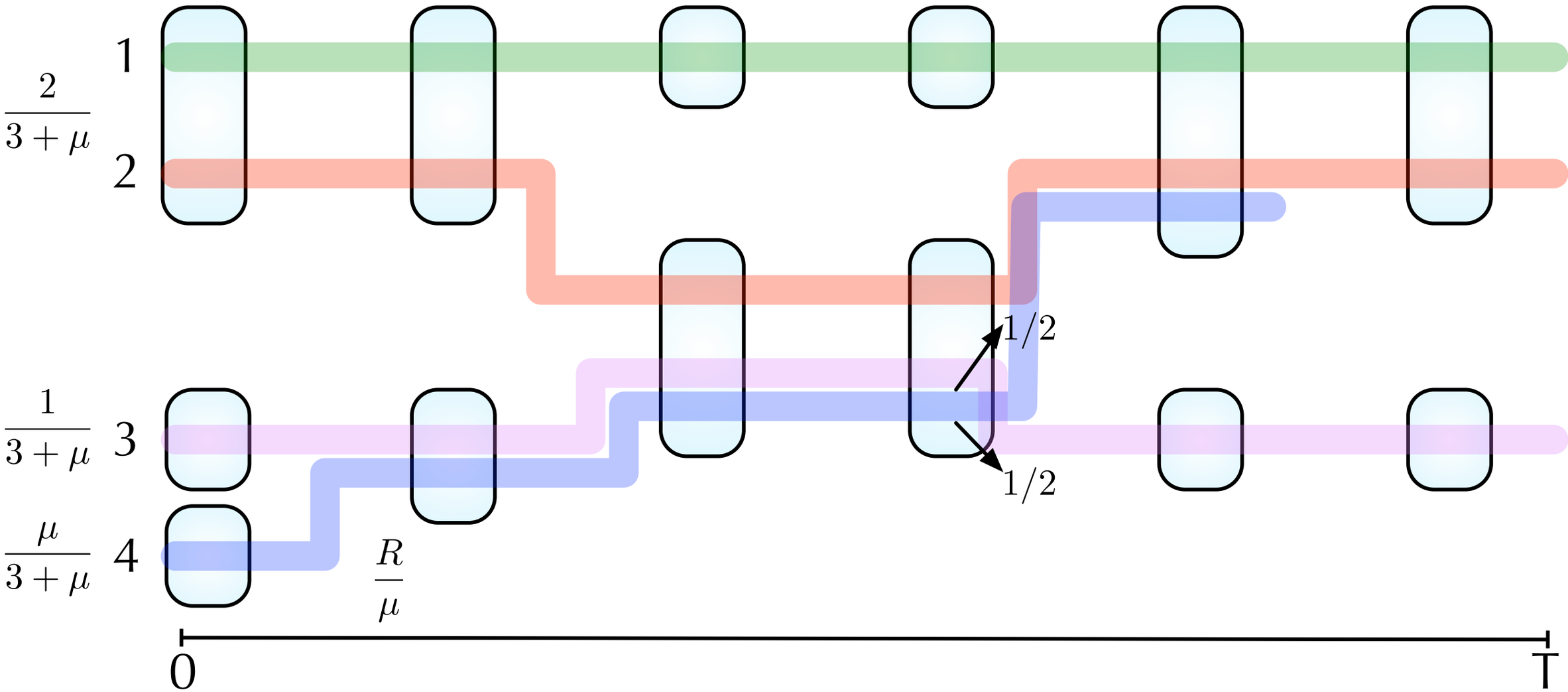
Teh, Blundell, Elliott

Chinese Restaurant Process Through Time



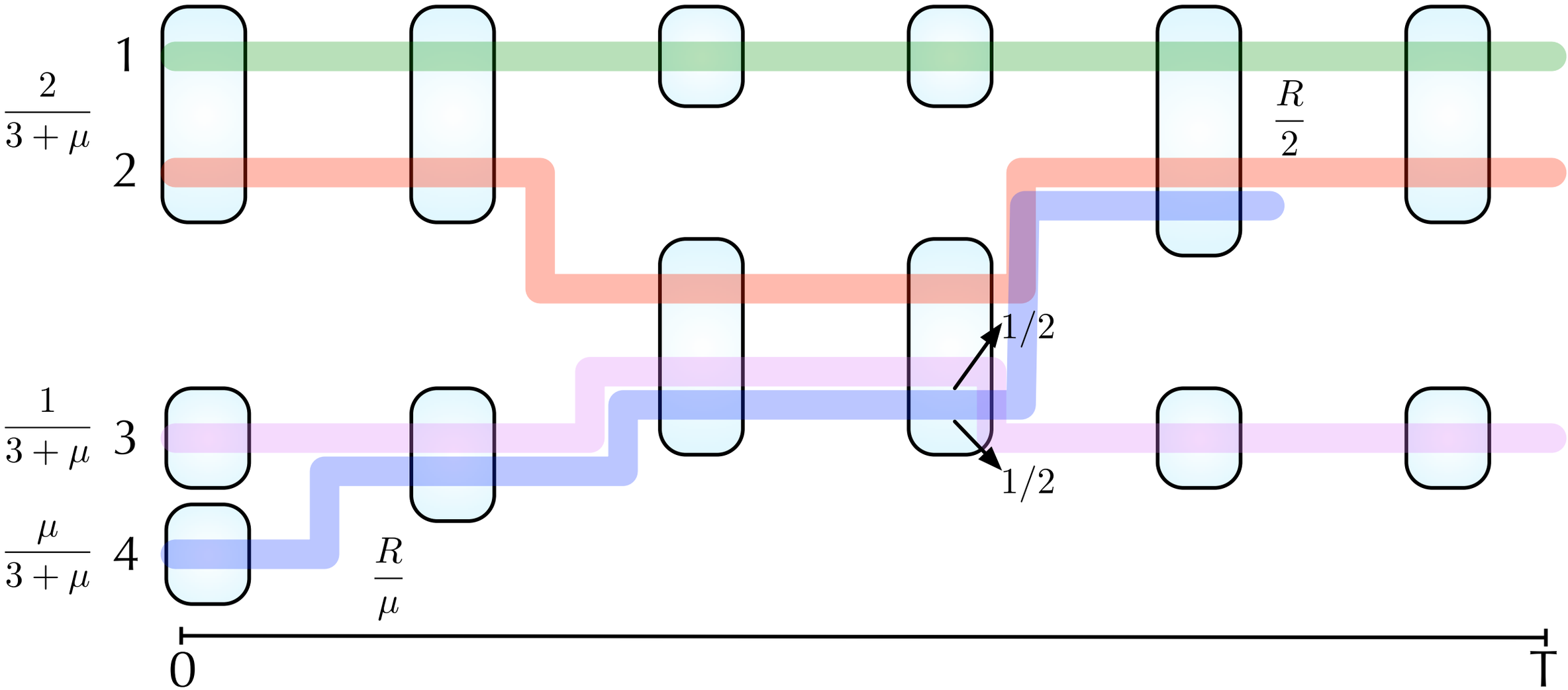
Teh, Blundell, Elliott

Chinese Restaurant Process Through Time



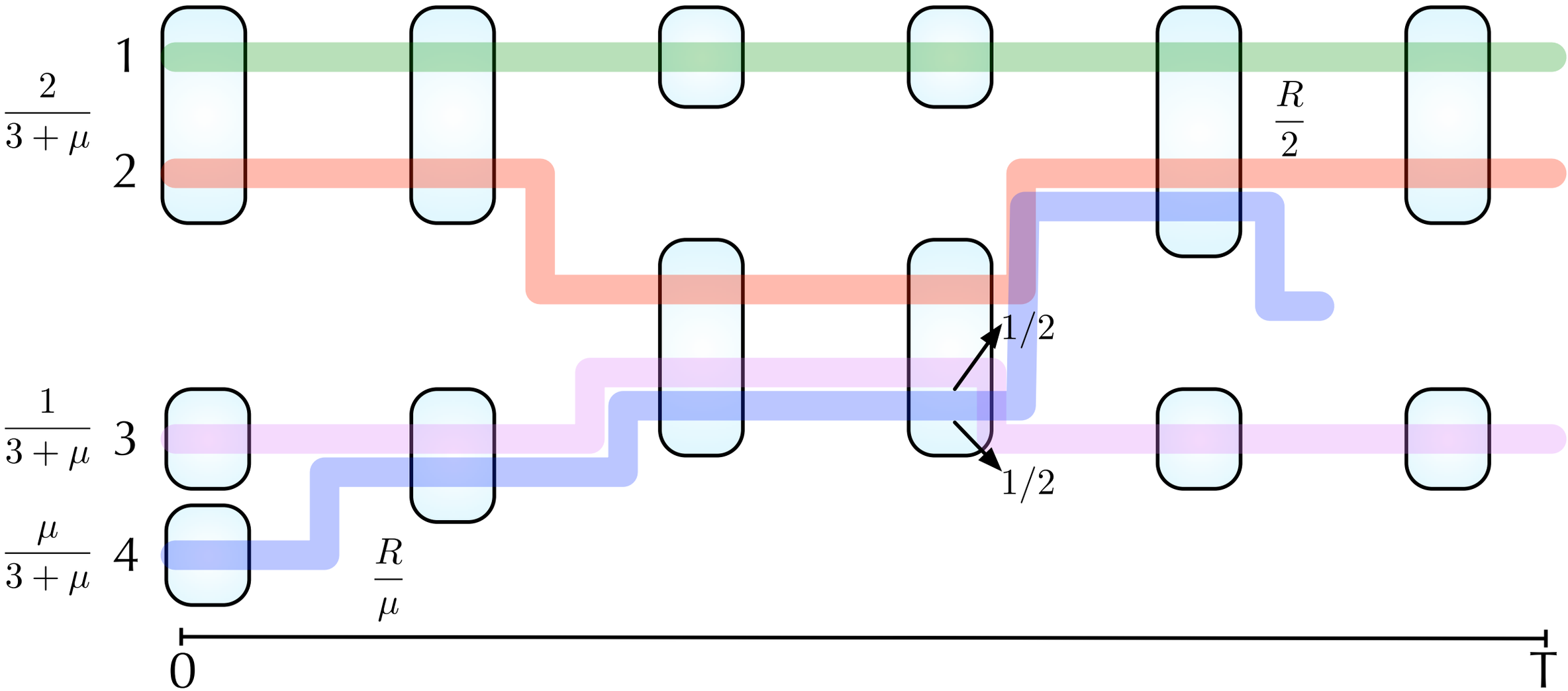
Teh, Blundell, Elliott

Chinese Restaurant Process Through Time

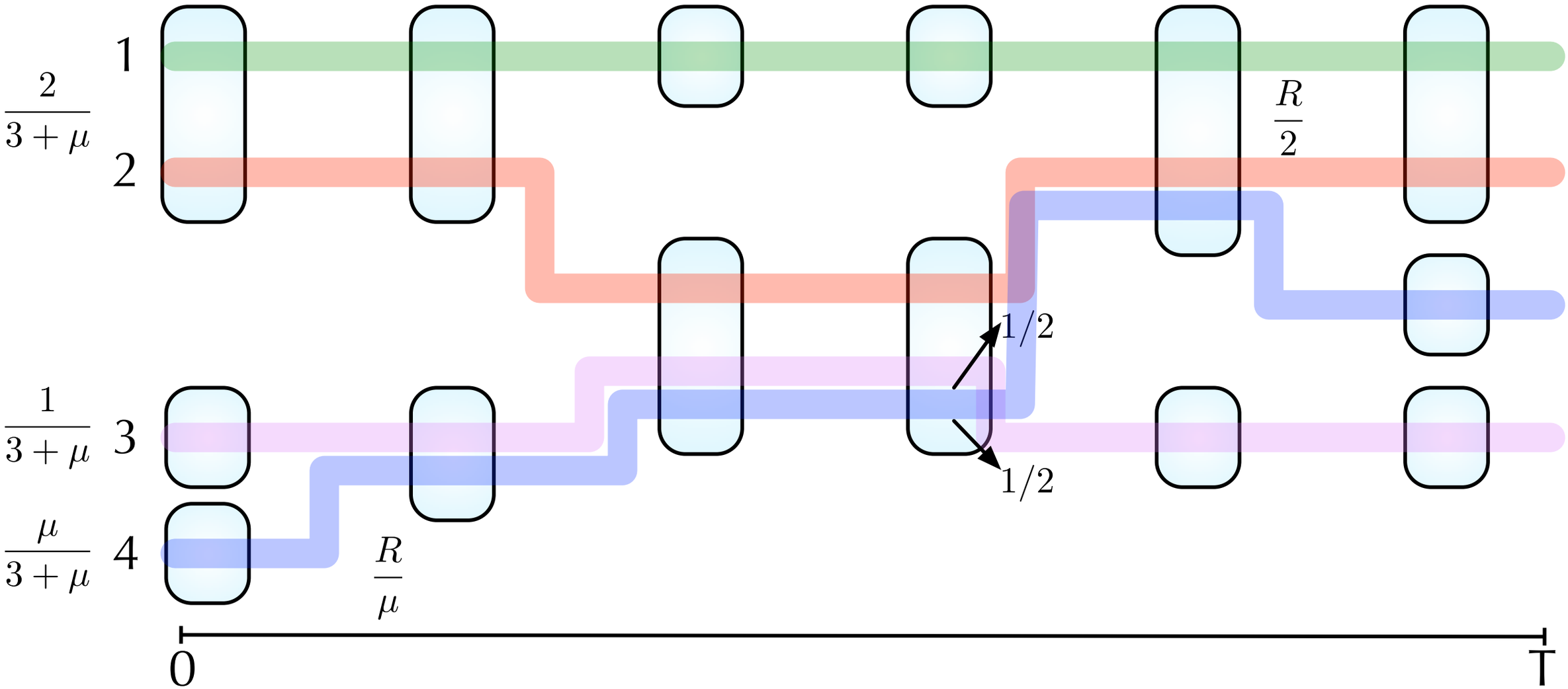


Teh, Blundell, Elliott

Chinese Restaurant Process Through Time



Chinese Restaurant Process Through Time



Teh, Blundell, Elliott

Partitions

- Set [n] = {1,2,...,n} indexing n sequences.
- Partition of [n], e.g.:

$$\{\{1,3,6\},\{2,7\},\{4,5,8\},\{9\}\}$$

  - Non-empty;
  - Disjoint;
  - Union is [n]; and
  - Unlabelled.