Biological Models with Combinatorial Spaces

B. Kirkpatrick

University of British Columbia

Feb 7, 2013

Key Idea

Many biological processes can be modeled using a Markov model with a discrete combinatorial state space.

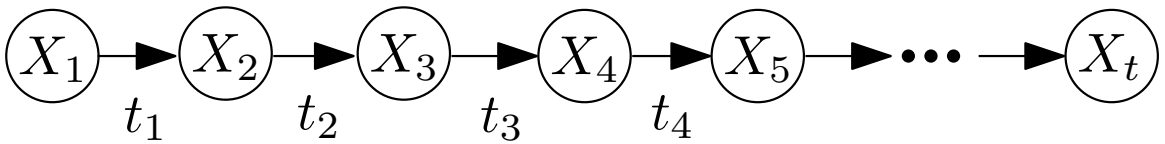
Outline

1. Introduction: models for genetics and nucleic acids
   - Markov models
   - Models on discrete combinatorial spaces
2. Family genetics
   - Discrete-time hidden Markov models (HMMs)
3. RNA/DNA folding pathways
   - Continuous-time Markov chains (CTMCs)
4. Conclusions

Markov Models

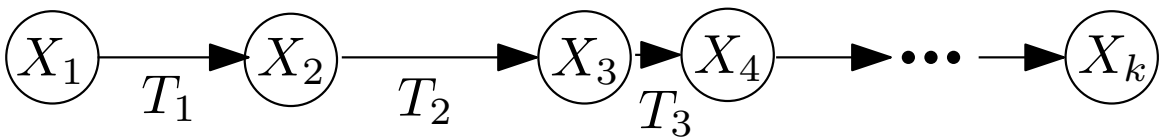
State Space: Ω where $X_i \in \Omega$ for all i
Markov Property: $\mathbb{P}[X_i|X_1, X_2, ..., X_{i-1}] = \mathbb{P}[X_i|X_{i-1}]$

Discrete-Time Markov Chain



$t_i = 1 \; \forall i$
$t = \sum_i t_i \in \mathbb{N}$

Continuous-Time Markov Chain



T_i positive
$T = \sum_i T_i \in \mathbb{R}$

DTMC: Fixed (unit) time interval versus
CTMC: Time drawn from exponential distributions

Computational Goals

To compute statistics of the model, including:

- Transition probabilities $\mathbb{P}[X_i|X_{i-1}, t]$
- Marginal probability distributions $\mathbb{P}[X_i]$
- Joint probability distributions $\mathbb{P}[X_1, X_2, ..., X_k]$
- Likelihoods $\mathbb{P}[X_1 = x_1, X_2 = x_2, ..., X_k = x_k]$
- Maximum likelihood paths
  $\max_{\{x_1,x_2,...,x_k\}} \mathbb{P}[X_1 = x_1, X_2 = x_2, ..., X_k = x_k]$

We might also be interested in functions of these statistics.

Discrete Combinatorial State Spaces

Markov models have discrete combinatorial state spaces if

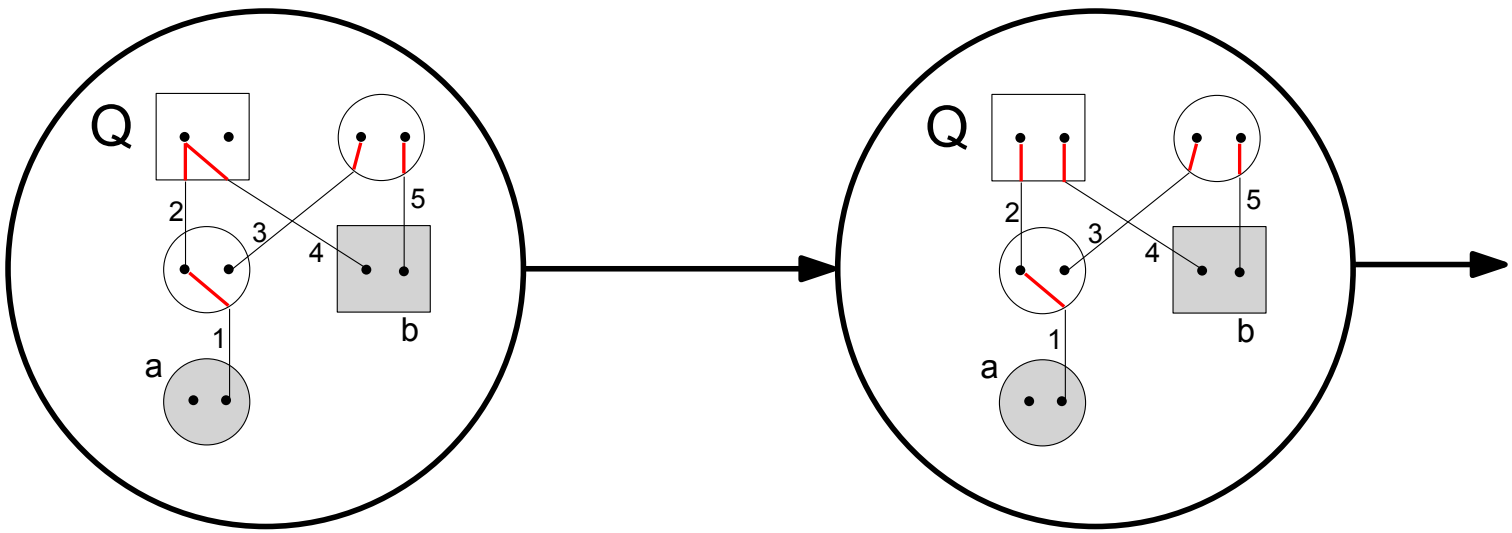
- the state space Ω is countable, and
- typical instances have exponential state spaces.

Representing the marginal probability $\mathbb{P}[X_i = x_i]$ requires a table the size of $|\Omega|$ provided that there is no sparsity to exploit. This representation can be *prohibitively* expensive no matter how good the exact inference algorithm that one uses.

Family Genetics

$|\Omega| = O(2^n)$ for n edges (parent-child pairs)

States: Inheritance possibilities
Transition Probabilities: Function of the recombination probability
Model Time: Represents distance along the genome



Family Genetics - Exact Approach

What can we learn from the model about biology?

- No laboratory method can observe inheritance in humans
- This model allows us to predict the inheritance
- This may aid in understanding inheritance, disease, and genetic relationships

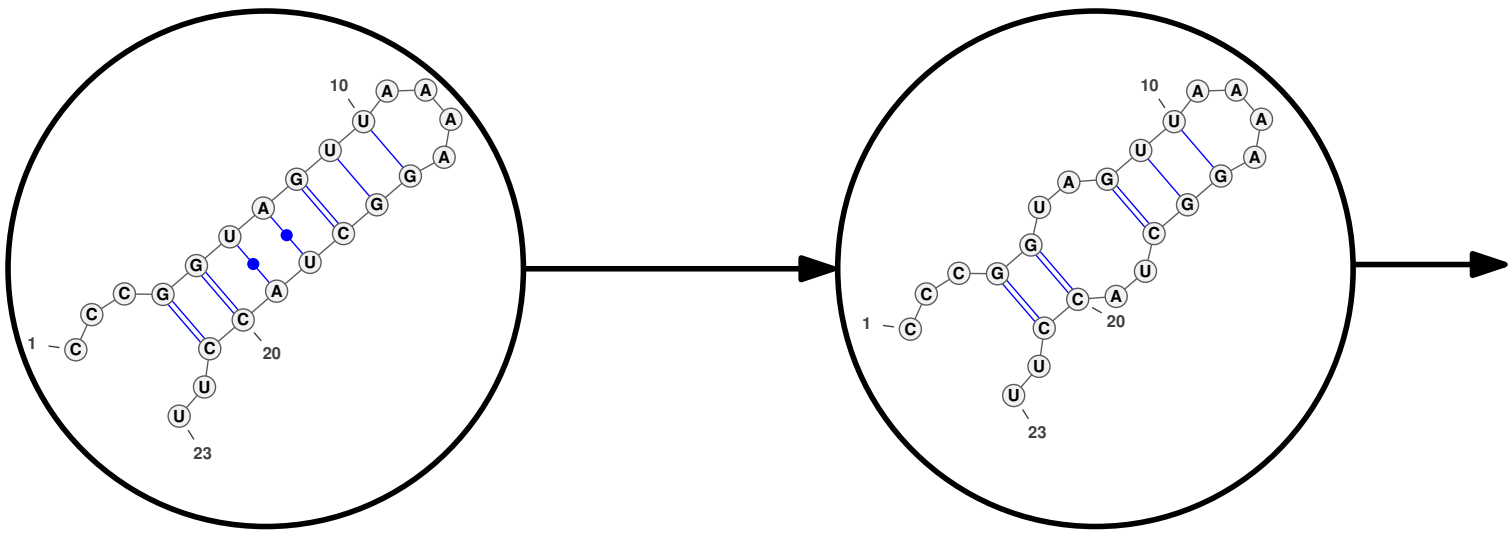
What research question did we pursue?

- **Question:** How can we efficiently compute exact likelihoods for this model?
- **Solution:** Remove the maximal number of repeat calculations.

Nucleic Acids (DNA and RNA)

$|\Omega| = O(3^n)$ for n bases

States: Secondary structures of the DNA/RNA molecules
Transition Probabilities: Function of the free energies of the secondary structures
Model Time: Represents physical time



Nucleic Acids - Approximation Approach

What can we learn from the model about biology?

- Microscopes cannot observe DNA/RNA structures over time
- Simulations allow us to predict this
- This may aid in understanding how DNA/RNA functions in the cell

What research question did we pursue?

- **Question:** How do we approximate the full $O(3^n)$ model with a model of significantly smaller size?
- **Solution:** Judiciously choose a smaller CTMC on secondary structures S that shares properties with the full model.

Key Idea

These two biological processes can be modeled using a Markov model with a discrete combinatorial state space.

Many more biological processes share the features of these two.

Family Genetics

Pedigrees or Family 'Trees'

Directed acyclic graph++

nodes individuals

boxes male

circles female

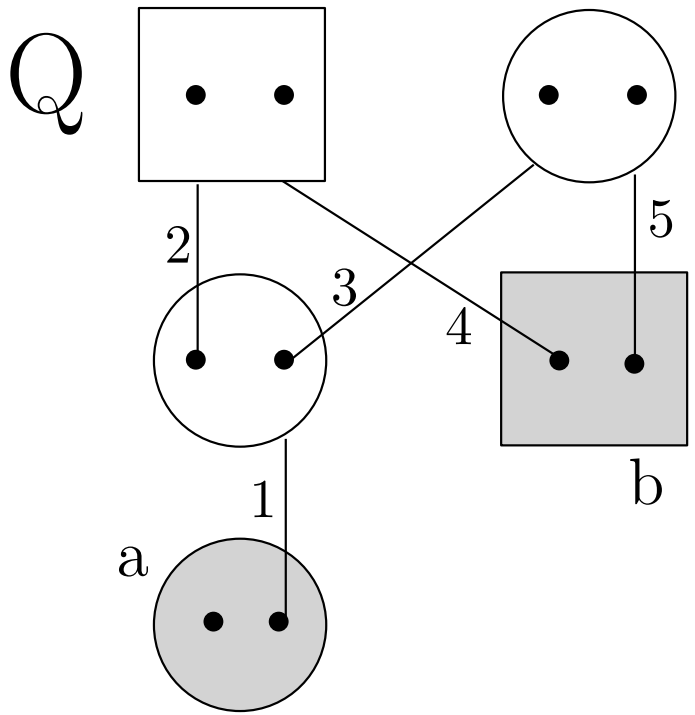
shaded has data

edges from parent to child

One-Site Inheritance

dots diploid alleles

edges inheritance options

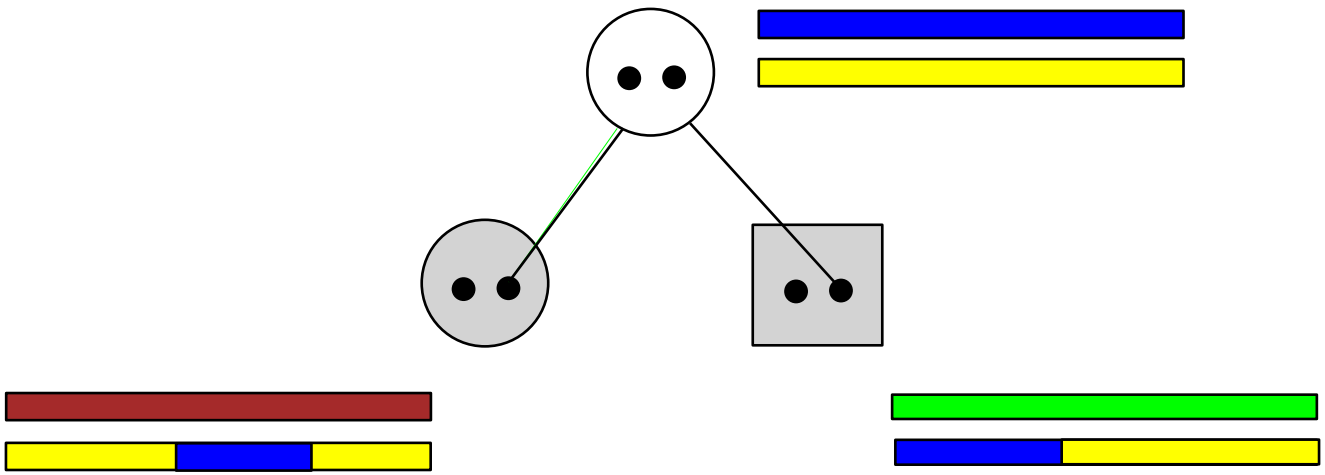


avuncular

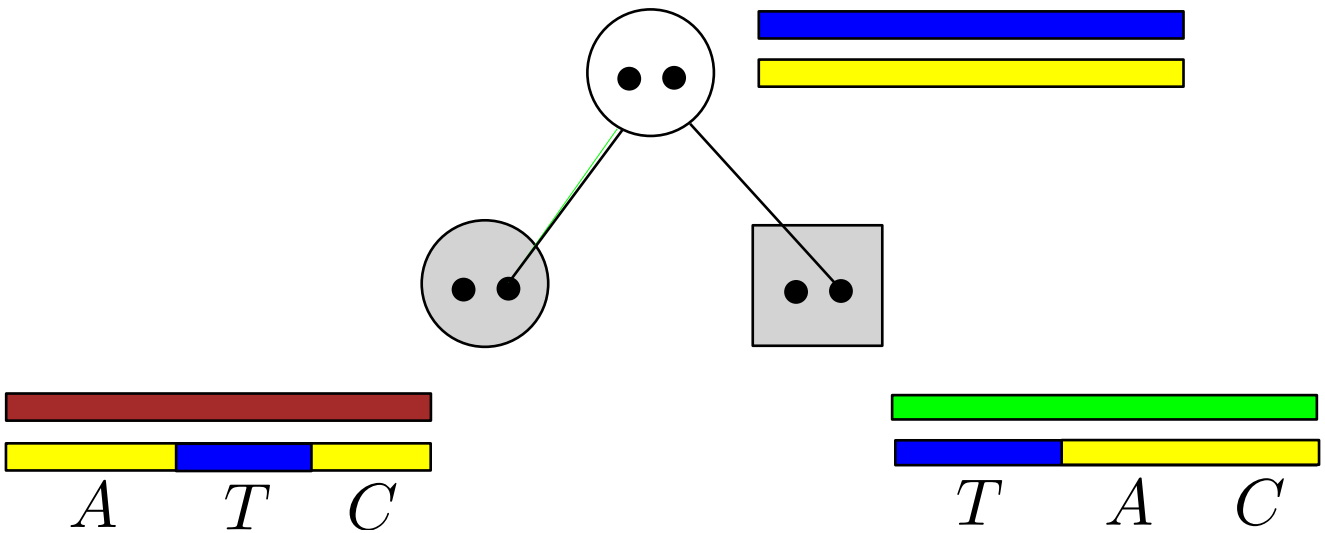
Why use families (pedigrees)?

1. Relationship inference - the pedigree that best generates the observed genomes
2. Pedigree correction - which edges of pedigree are incorrect
3. Forensics - infer a relationship to decide which individuals are suspicious
4. Disease gene-finding - use a pedigree to find genes that correlate to a disease
5. Recombination probability inference - use a pedigree to predict recombinations

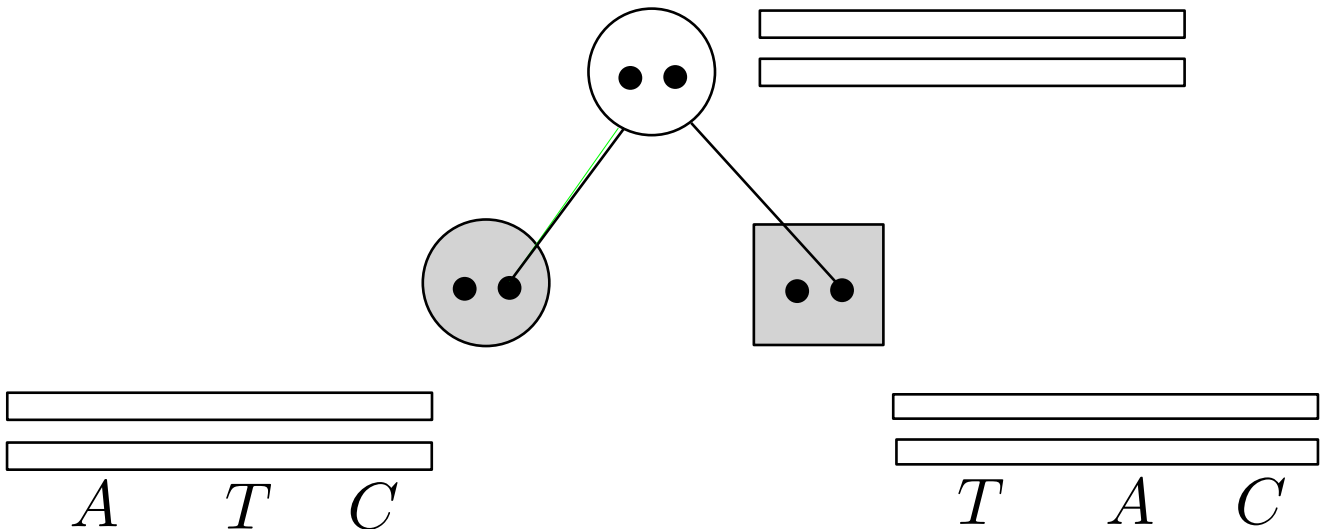
Human Genetics



Human Genetics

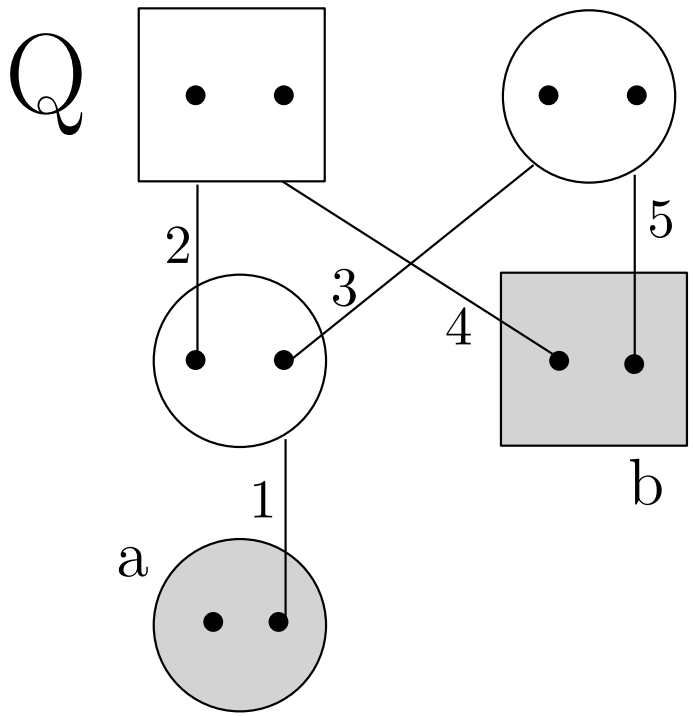


Human Genetics



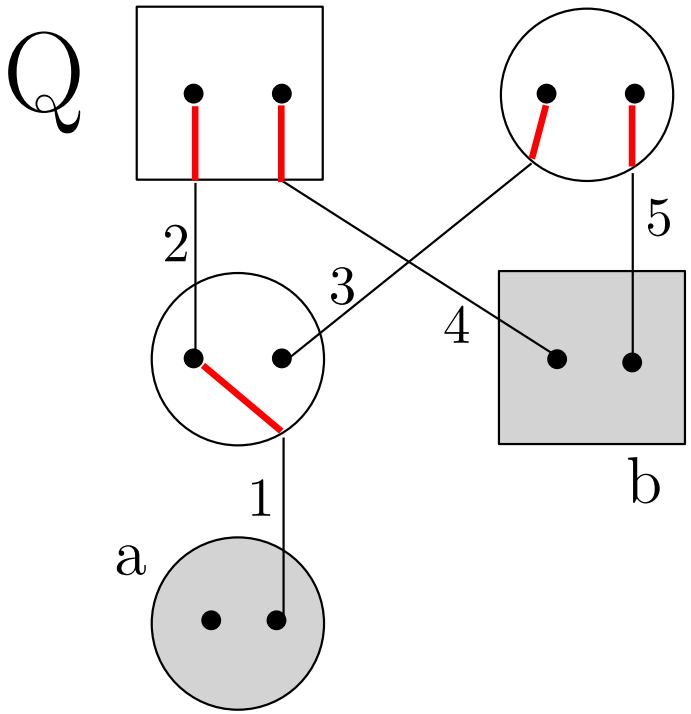
θ_t is the recombination probability per meiosis between a pair of sites t and $t+1$

Inheritance Paths

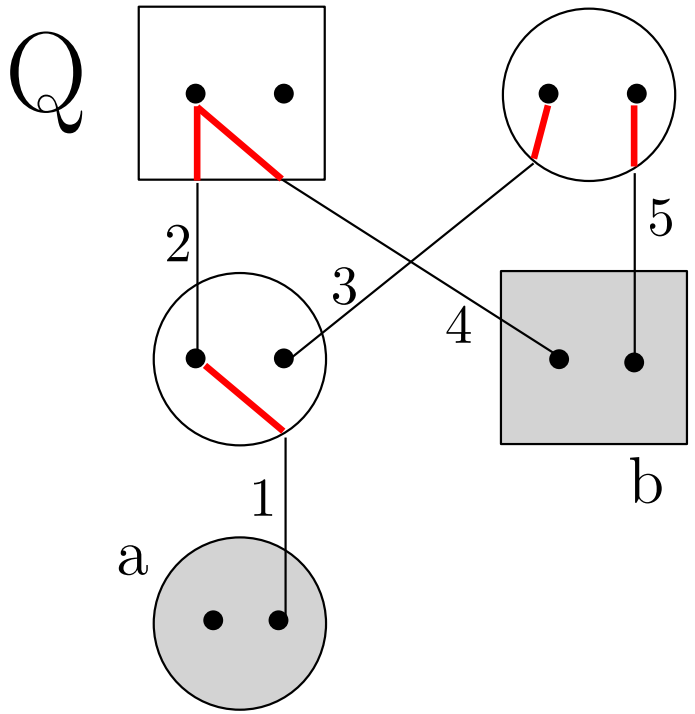


avuncular

Inheritance Paths

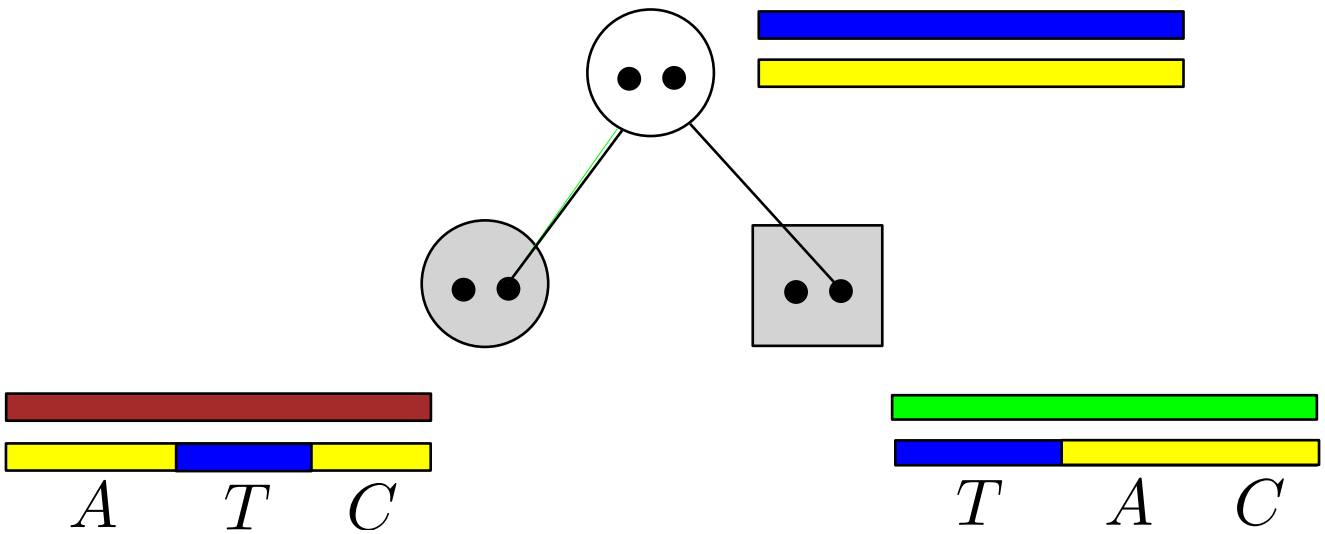


Inheritance Path



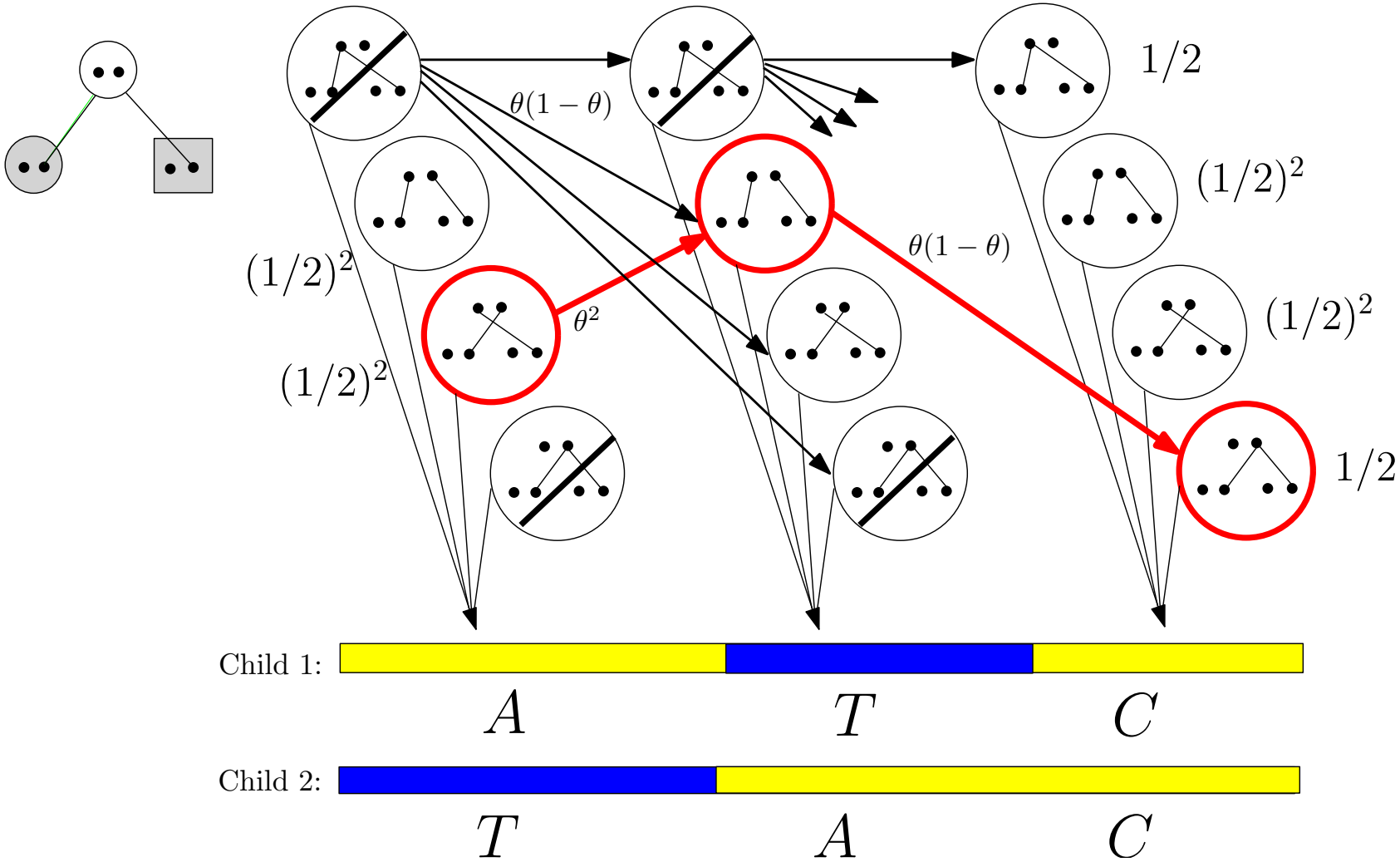
$O(2^n)$ inheritance paths (n edges)

Data

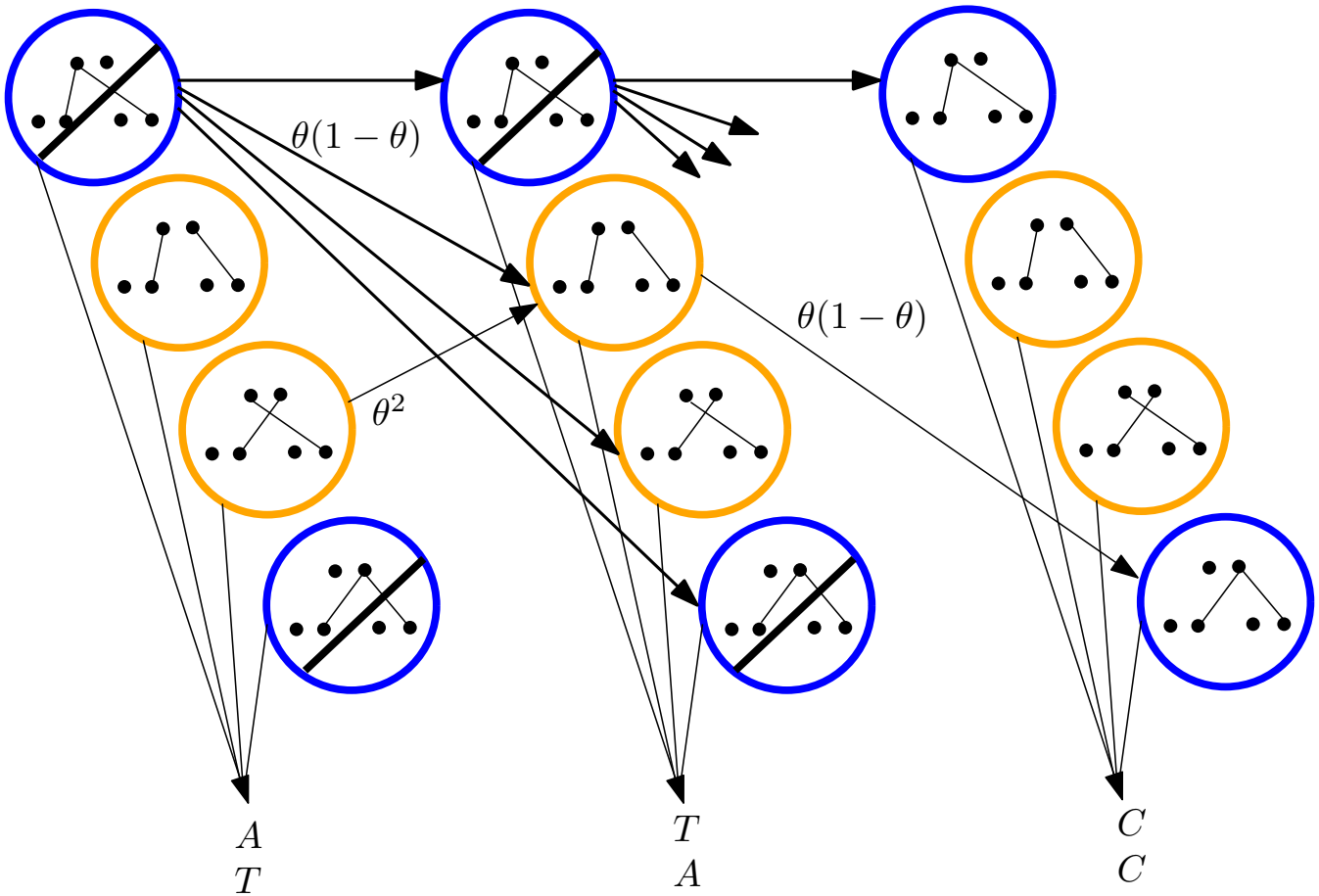


[Kirkpatrick and Kirkpatrick, 2012]

Likelihood


$\theta(1-\theta)$
$1/2$
$(1/2)^2$
θ^2
$(1/2)^2$
$\theta(1-\theta)$
$(1/2)^2$
$(1/2)^2$
$1/2$
Child 1:
A
T
C
Child 2:
T
A
C

Maximum Ensemble Partition



There exists an $O(n2^{2n})$ algorithm for finding the unique *maximum* ensemble state-space $m(P)$ that preserves the likelihood [Kirkpatrick and Kirkpatrick, 2012].

Ensemble Partition

An **ensemble partition** $\{W_1, W_2, ..., W_k\}$ which is a partition of Ω is choose such that the following is true

- $\{X_t\}$ is a Markov chain for an HMM on states Ω having likelihood $P_X[G]$.
- $\{Y_t\}$ is an Markov chain for an HMM on states $\{W_1, ..., W_k\}$ having likelihood $P_Y[G]$.
- $\{W_1, ..., W_k\}$ is chosen such that $P_X[G] = P_Y[G]$.

A **maximum ensemble partition** is the states $\{W_1, ..., W_k\}$ for some $\{Y_t\}$ such that $P_X[G] = P_Y[G]$ which minimizes k.

Why?

Assume we can find the $\{Y_t\}$ with the maximum ensemble partition.

1. The likelihood computation on $\{Y_t\}$ is *faster* than on $\{X_t\}$, because the computation is polynomial in the number of hidden states.

2. If we can show certain properties of an optimal $\{Y_t\}$, such as uniqueness, then we can use it to determine whether two HMMs are identifiable.

Key Contributions

- Algorithm with running-time of $O(nk2^n)$ where $k < 2^n$ improving on the $O(n!2^{2n})$ running-time for best existing algorithm.
- Proof that the optimal partition is *unique* which can be used to prove whether the HMM is *identifiable*.
- Solutions that involved computational group theory.

[Kirkpatrick and Kirkpatrick, 2012] [Kirkpatrick, 2012a]

Nucleic Acids (RNA and DNA)

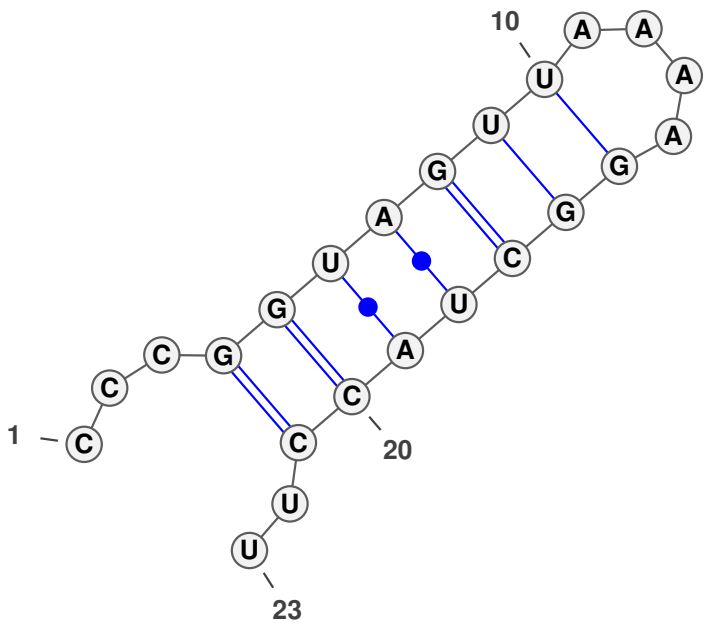
RNA/DNA

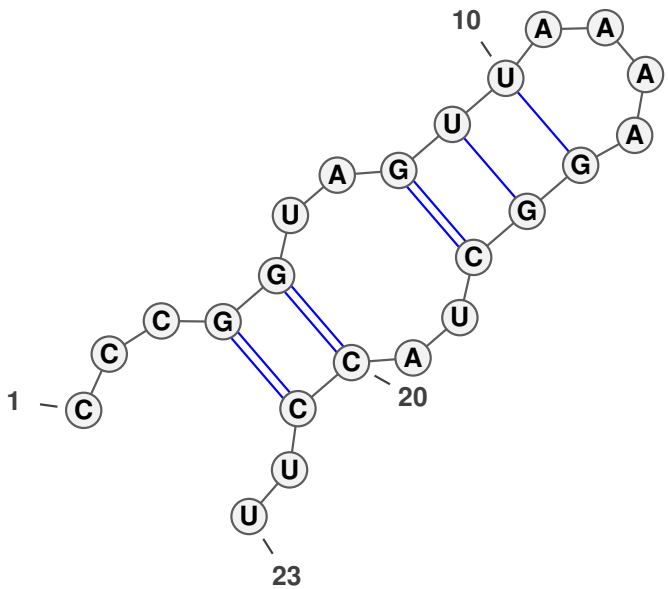
sequence of n nucleotides A,C,G,U/T

base pairs C-G, A-U/T, and G-U/T

secondary structure is the sequence and a set of base pairs

free energy of a secondary structure i is $E(i)$, computed in linear time





Folding Pathways

folding pathway – is a sequence of secondary structures where each successive structure differs by 1 base pair

folding trajectory – is an alternating sequence of secondary structures and holding times where the sequence of structures is a folding pathway

population kinetics – are the fraction of folding trajectories that are in a particular secondary structure at a particular time, given a large number of folding trajectories

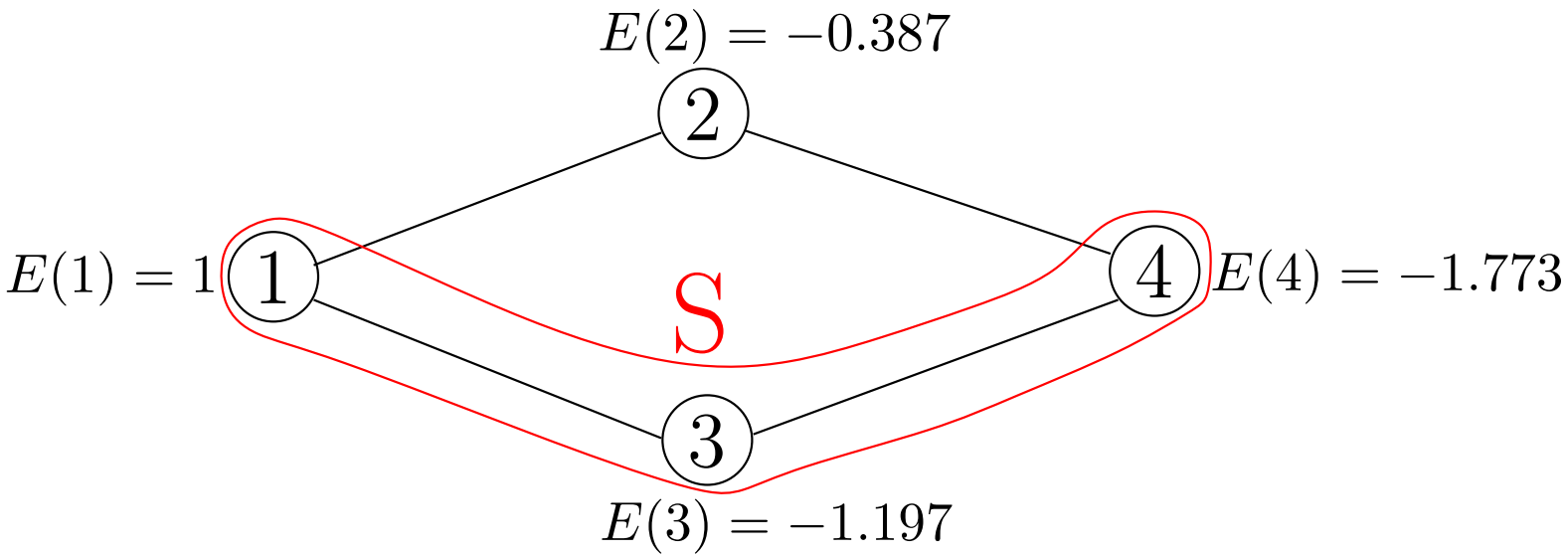
Approach

Problem: The state space of the full process is large: $O(3^n)$.

Question: Is there a process on a subset of the state space S on which well approximates the full process?

We compare the full process to two alternative subset processes for a given subset S.

Example



Node numbers are structures.

Subset S must be connected so that the process is ergodic.

CTMC: Full Model (Kawasaki)

The neighborhood $N(i)$ of i is every structure that differs from i by exactly one base pair. N is symmetric: $j \in N(i)$ iff $i \in N(j)$.

The infinitesimal generator matrix for $i \neq j$ is

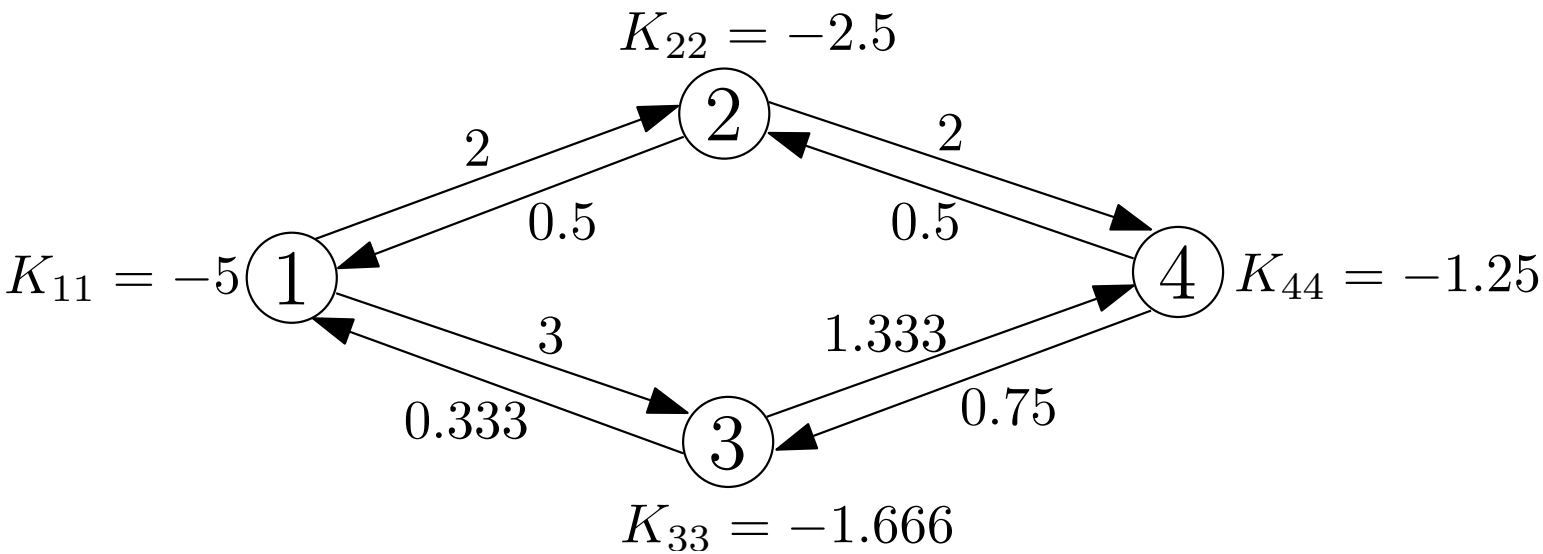
$$K_{ij} = \begin{cases} e^{(E(i)-E(j))/(2\rho\tau)} & \text{if } j \in N(i) \\ 0 & \text{otherwise} \end{cases}$$

and

$$K_{ii} = -\sum_{j \neq i} K_{ij}.$$

Constant ρ is the gas constant, and τ is the temperature.

Example: Rates for K



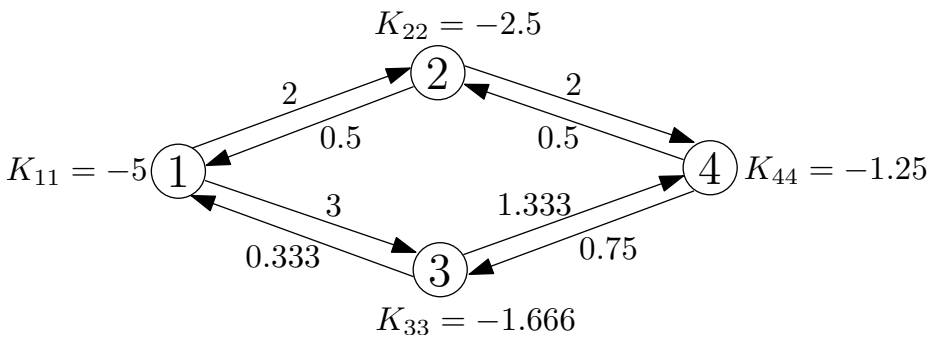
Monte Carlo Sampling of Folding Trajectories

(Doob-Gillespie Algorithm)
For generator K, step δ, for the folding trajectory at state i

1. Simulate holding time t_δ from $exp(-K_{ii})$
2. Choose the next secondary structure x_δ from $N(i)$ with probability

$$H_{ij}^K = \begin{cases} \frac{-K_{ij}}{K_{ii}} & \text{if } i \neq j \\ 0 & \text{otherwise} \end{cases}$$

e.g.
$x_1 = 1, t_1 = 0.02,$
$x_2 = 3, t_2 = 0.8,$
$x_3 = 1, t_3 = 0.03, \ldots$



CTMC Approx #1: The Subset Model

Subset of state-space is $S \subseteq \Omega$

The infinitesimal generator matrix is

$$L_{ij} = \begin{cases} K_{ij} & \text{if } i \in S \text{ and } j \in N(i) \cap S \\ 0 & \text{otherwise} \end{cases}$$

and

$$L_{ii} = -\sum_{j \neq i} L_{ij}.$$

CTMC Approx #2: The Trajectory Subset Model

Again, the subset of state-space is S

The infinitesimal generator matrix is

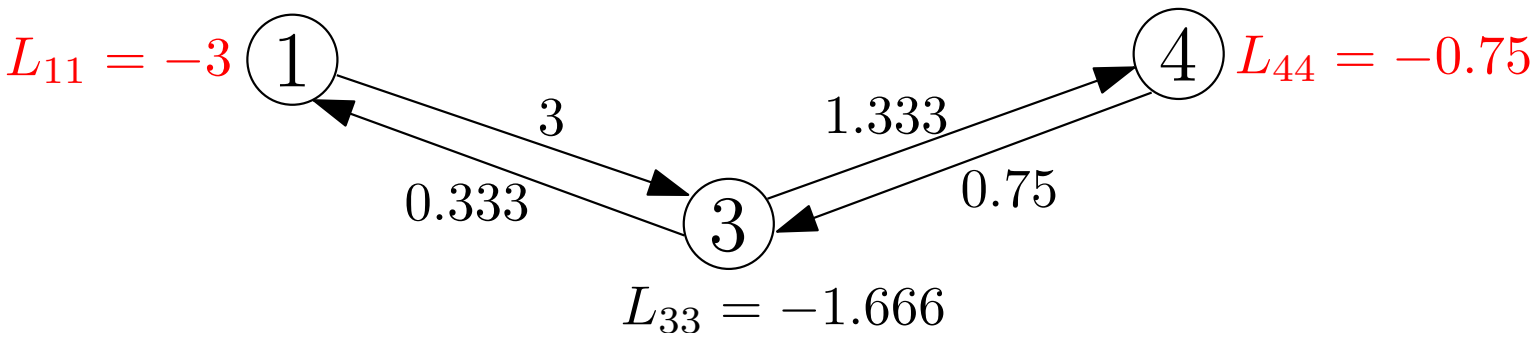
$$\Gamma_{ij} = \begin{cases} \frac{\sum_{k \in \Omega_{-i}} K_{ik}}{\sum_{k \in S_{-i}} K_{ik}} K_{ij} & \text{if } i \in S \text{ and } j \in N(i) \cap S \\ 0 & \textit{otherwise} \end{cases}$$

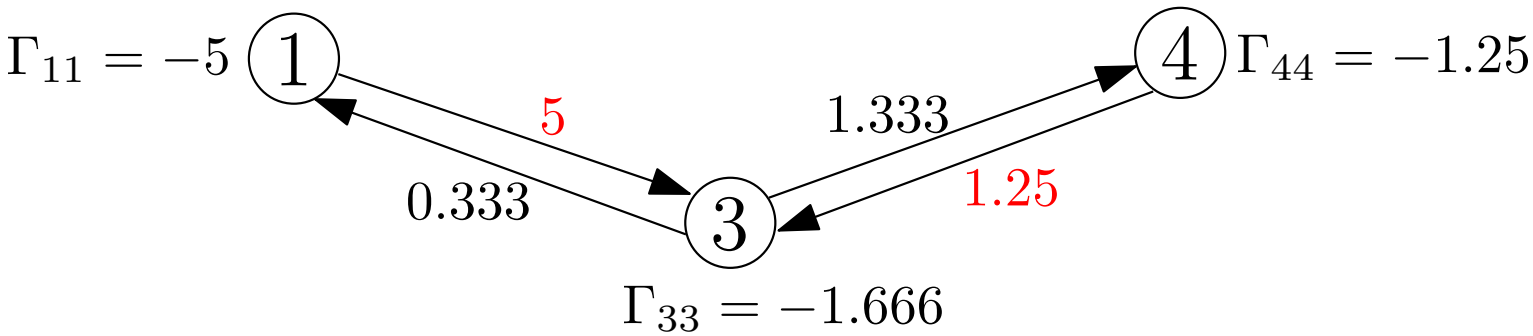
and

$$\Gamma_{ii} = -\sum_{j \neq i} \Gamma_{ij}.$$

making the Γ_{ii} holding times at least as fast as those for L_{ii}. [Kirkpatrick et al., 2012b]

Example: Rates





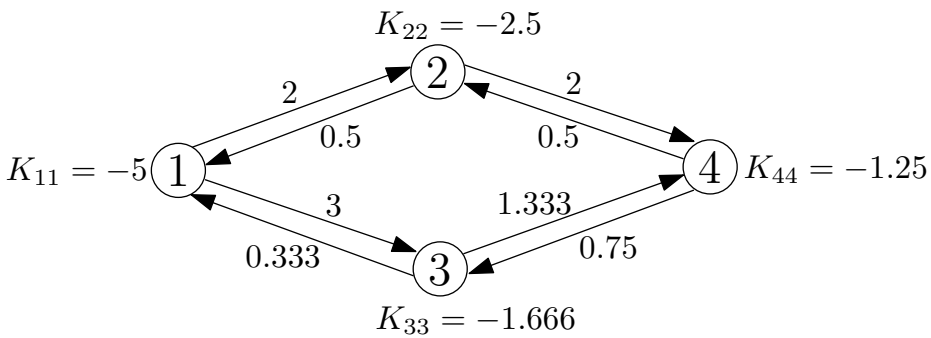
Monte Carlo Sampling of Folding Trajectories

For generator K, step δ, for the folding trajectory at state i

1. Simulate holding time t_δ from $exp(-K_{ii})$
2. Choose the next secondary structure x_δ from $N(i)$ with probability

$$H_{ij}^K = \begin{cases} \frac{-K_{ij}}{K_{ii}} & \text{if } i \neq j \\ 0 & \text{otherwise} \end{cases}$$

e.g.
$x_1 = 1, t_1 = 0.02,$
$x_2 = 3, t_2 = 0.8,$
$x_3 = 1, t_3 = 0.03,$ ...



Why Trajectory Sampling?

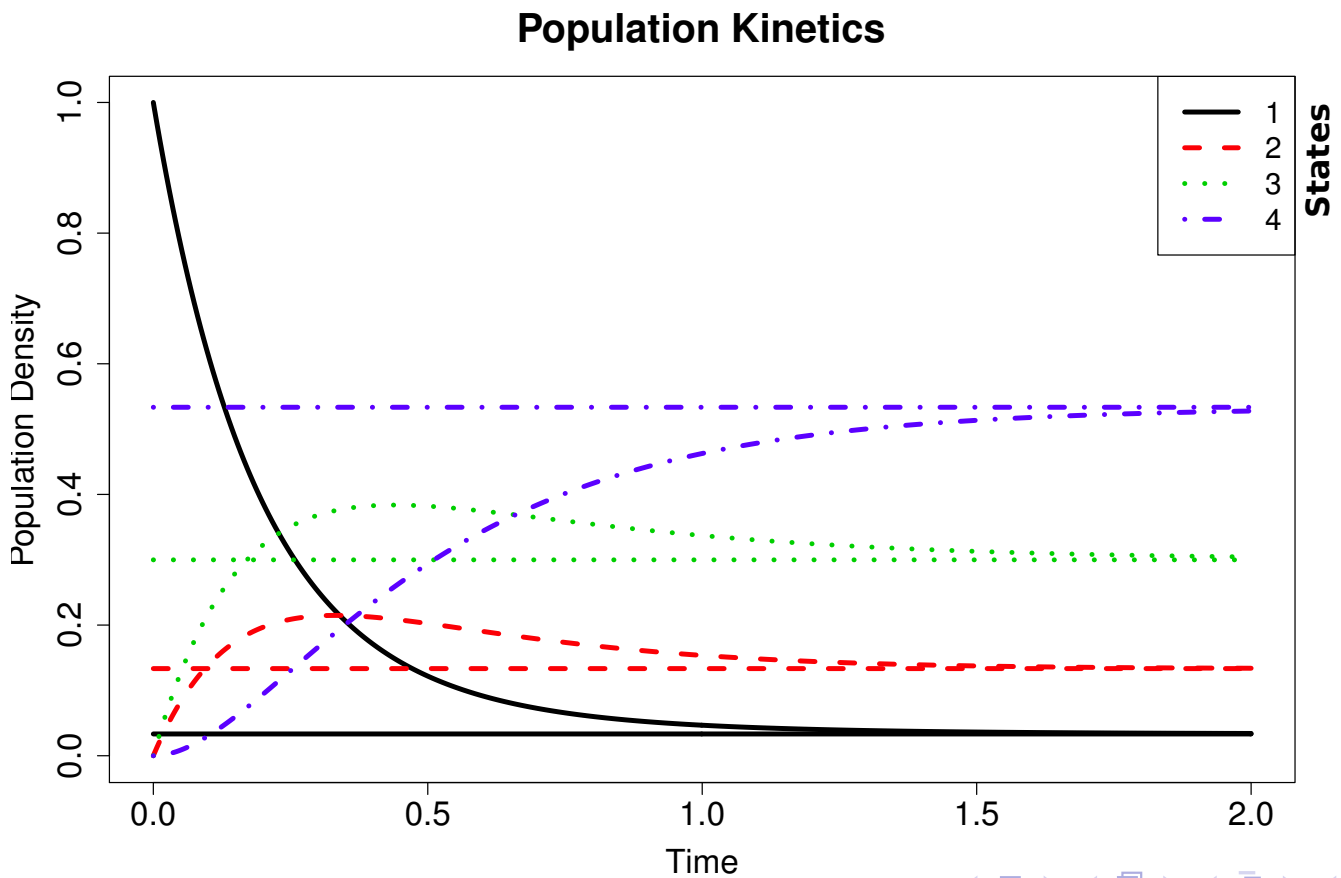
Holding times $\Gamma_{ii} = K_{ii}$ and
Transition probabilities $\frac{-\Gamma_{ij}}{\Gamma_{ii}} \propto H_{ij}^{K}$

We can simulate trajectories in Γ by simulating in K and rejecting those trajectories with nodes not in S. Trajectories from Γ are trajectories in K.

[Kirkpatrick et al., 2012b]

Example: Population Kinetics

$P^K(t) = y_0 e^{tK}$ is the transition matrix for process Q where the i, j entry gives the probability of the process transitioning from state i to state j in t time.



Inaccuracy of R Compared with K

R has state-space S and K has state-space Ω. Let

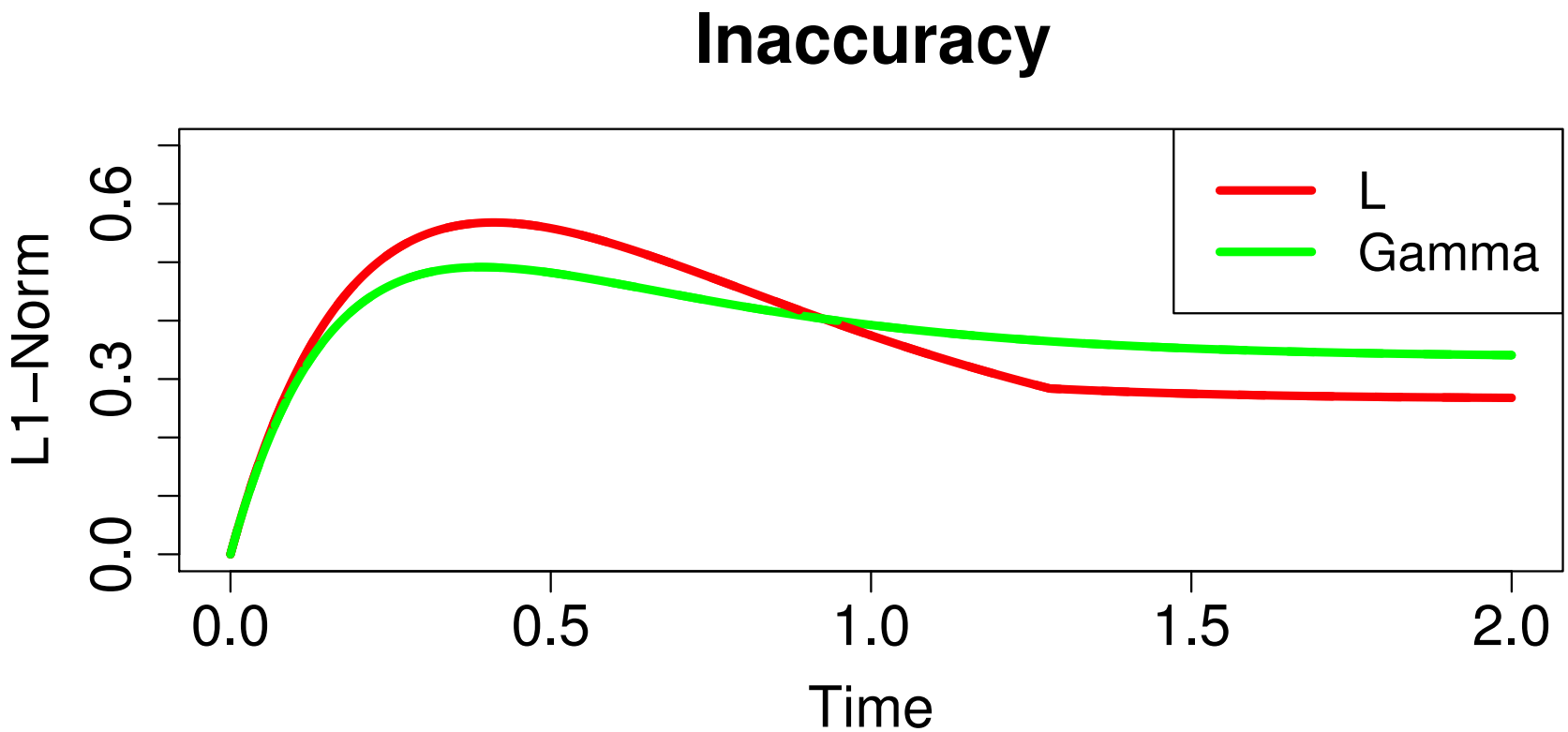
$$A(R,t) := \left\| y_0 e^{Kt} - y_0 e^{Rt} \right\|_1 .$$

Computing $\|a-b\|_1 = \sum_{i\in D} |a_i - b_i|$ for support D:

(Support Ω) sum over Ω after:
defining $(y_0 e^{Rt})_i = 0$ for $i \notin S$,

[Kirkpatrick et al., 2012b]

Inaccuracy Example



Simulation

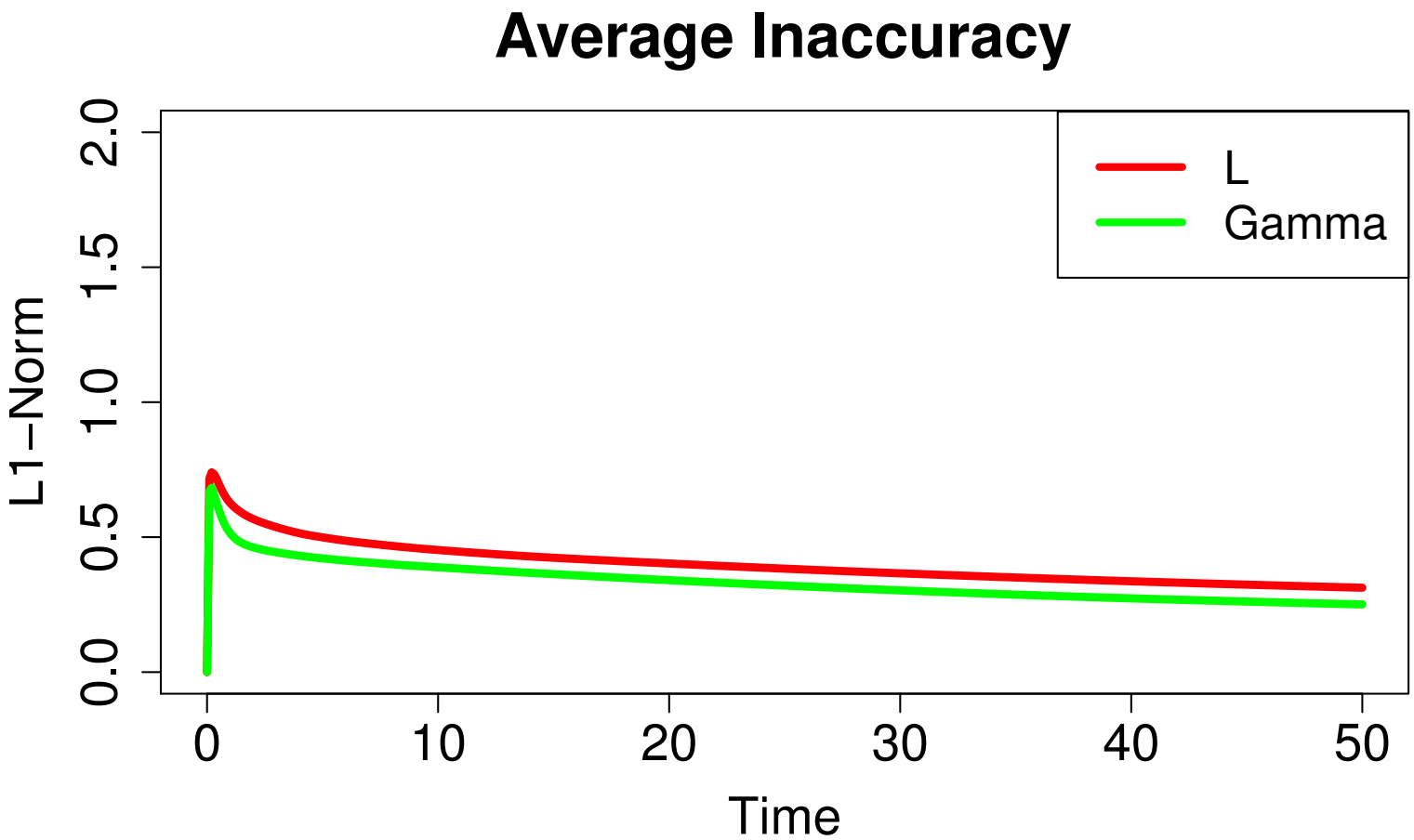
100 replicates with varying S

- real RNA sequences
- Turner 2004 energies
- 10% of secondary structures chosen for S

From this, we can compute

- the infinitesimal generator matrices: K, L, Γ
- the matrix exponentials: e^{Kt}, e^{Lt}, $e^{\Gamma t}$
- the inaccuracies: $A(K,t)$, $A(L,t)$, $A(\Gamma,t)$

Results: Support Ω



Conclusions

Conclusions:

- Γ is a new model with correct holding times.
- On average over various choices of S, our model Γ is superior.
- Simulation was performed on real RNA sequences.

Key Idea

These two biological processes can be modeled using a Markov model with a discrete combinatorial state space. Similar processes include:

CTMC

Phylogenetics with indels
Chemical reaction networks
Recombination
Coalescent

DTMC or HMM

Multiple sequence alignment
Identity by descent
Protein modeling
Biological networks

Thanks to....

- Nucleic Acids: Monir Hajiaghayi and Anne Condon
- Pedigrees: Kay Kirkpatrick
- Everyone who gave me feedback on the slides.


Donnelly, K. P. (1983).
The probability that related individuals share some section of genome identical by descent.
Theoretical Population Biology, 23(1):34 – 63.


Kirkpatrick, B., Monir Hajiaghayi, and Anne Condon (2012).
A New Model for RNA Folding
submitted.


Kirkpatrick, B. (2012).
Non-Identifiable Pedigrees and a Bayesian Solution
ISBRA 2012.


Kirkpatrick, B. and Kirkpatrick, K. (2012).
Optimal state-space reduction for pedigree Hidden Markov Models.
arXiv, Feb. 2012.


McPeek, M.S.. (2002).
Inference on pedigree structure from genome screen data.
Statistica Sinica, 12(1):311–336.


Pinto, N., Silva, P. V., and Amorim, A. (2010).
General derivation of the sets of pedigrees with the same kinship coefficients.
Hum Hered, 70(3):194–204.


Thompson, E. A. (1975).
The estimation of pairwise relationships.
Annals of Human Genetics, 39(2):173–188.

Family Genetics: More Examples

Likelihood

The likelihood of an HMM is a product of the following:
The *transition probabilities*

$$Pr[X_{t+1} = y \mid X_t = x] = \theta_t^{R(x,y)}(1-\theta_t)^{n-R(x,y)} \quad (1)$$

where $R(x, y)$ is the number of recombinations between x and y and n is the number of edges (meioses) in the pedigree.
The *emission probability* at time t is

$$Pr[G_t = g_t \mid X_t = x] \propto h(g_t)$$

where h is a poly-time computable function.

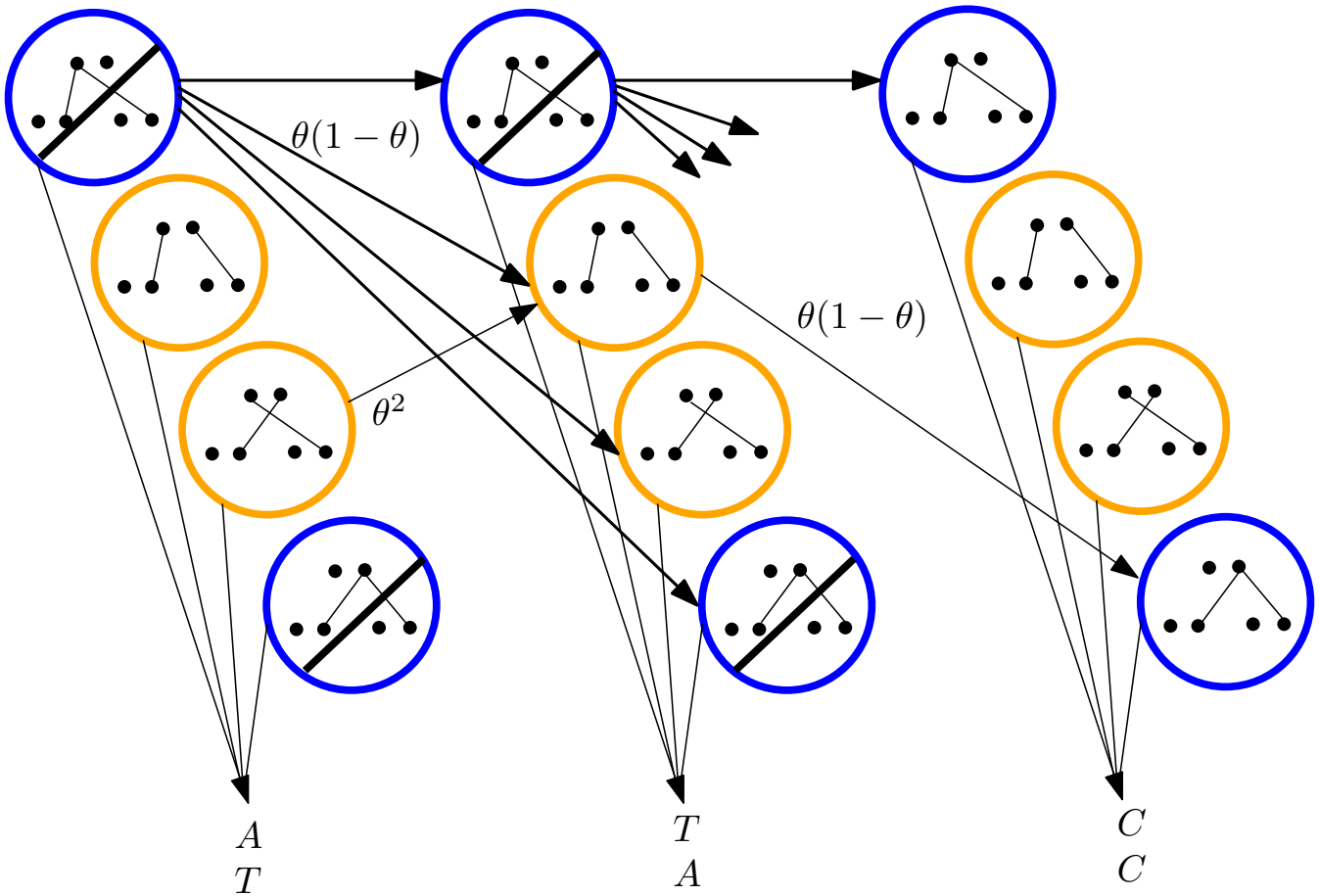
Emission Probability

An *inheritance graph* R_x for inheritance vector x contains two nodes for each individual $i \in I(P)$, called i_0 and i_1. R_x has edges $((p_j(i))_{x_e} i_j)$ for each edge $e = (p_j(i), i) \in E(P)$. Let $CC(R_x)$ be the connected components of R_x. Let $\tilde{g}_t$ is the ordered alleles (g^0_{it}, g^1_{it}). Let $\tilde{G}_t$ be the $\tilde{g}_t$ consistent with x where each connected component in $CC(R_x)$ has exactly one allele. Then for site t, the *emission probability* is

$$Pr[G_t = g_t \mid X_t = x, P] \propto \sum_{\tilde{g}_t \in \tilde{G}_t} \prod_{i \in F(P)} Pr[h(i, \tilde{g}_t)]$$

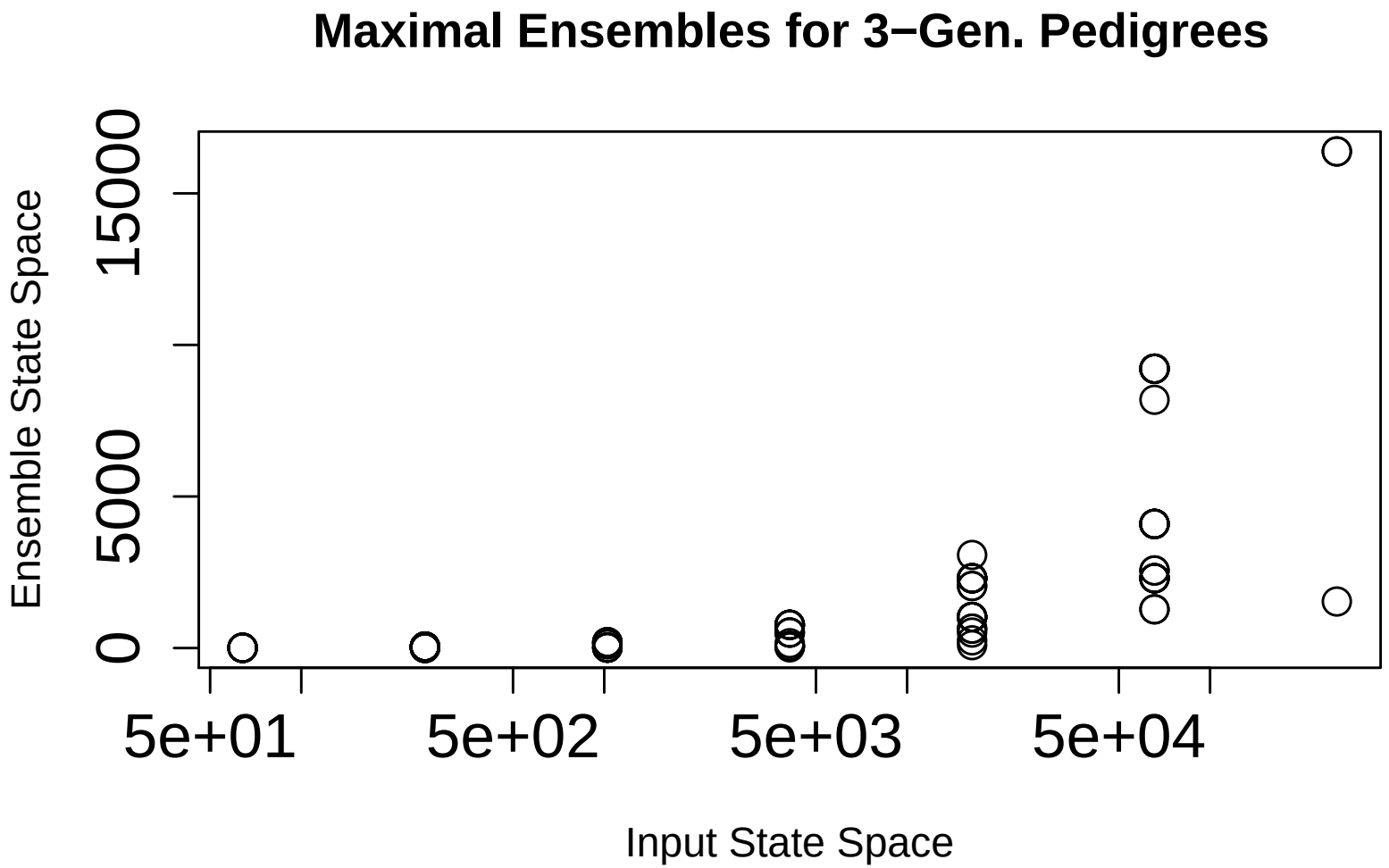
where $h(i, \tilde{g}_t)$ is the allele of $\tilde{g}_t$ that is assigned to founder i.

Maximum Ensemble Partition



There exists an $O(n2^{2n})$ algorithm for finding the unique *maximum* ensemble state-space $m(P)$ that preserves the likelihood [Kirkpatrick and Kirkpatrick, 2012].

Maximum Ensembles



General Identifiability Criteria

Let $m(P)$ be the maximum ensemble states for pedigree P.
A *proper isomorphism*, ψ, is an isomorphism $\psi : m(P) \rightarrow m(Q)$ that preserves the likelihood:

Transition Equality $Pr[Y_{t+1}^P \mid Y_t^P] = Pr[\psi(Y_{t+1}^P) \mid \psi(Y_t^P)] \ \ \forall t$

Emission Equality $Pr[G \mid Y_t^P] = Pr[G \mid \psi(Y_t^P)] \ \ \forall t$

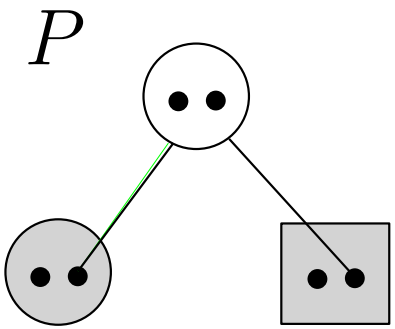
where Y_t^P is the hidden state for pedigree graph P.

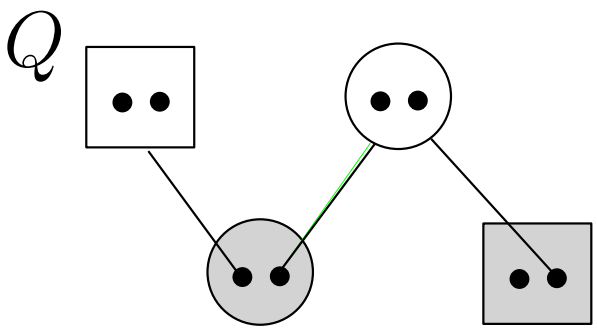
Theorem

There exists proper isomorphism ψ if and only if the likelihoods for $\mathcal{P}$ and $\mathcal{Q}$ are non-identifiable, $Pr[G \mid \theta, P] = Pr[G \mid \theta, Q]$, for all G and $\theta = (\theta_1, ..., \theta_{T-1})$ where T is the number of sites and $T \geq 2$.

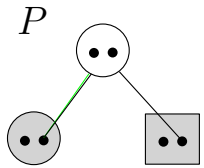
[Kirkpatrick, 2012a]

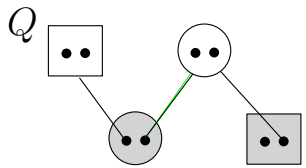
General Theorem Example

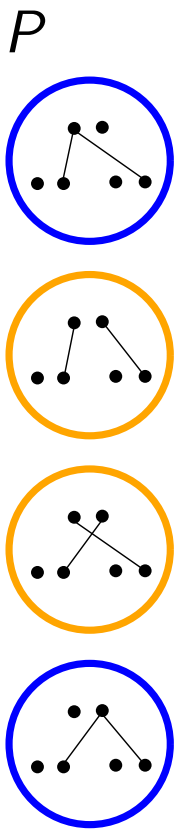


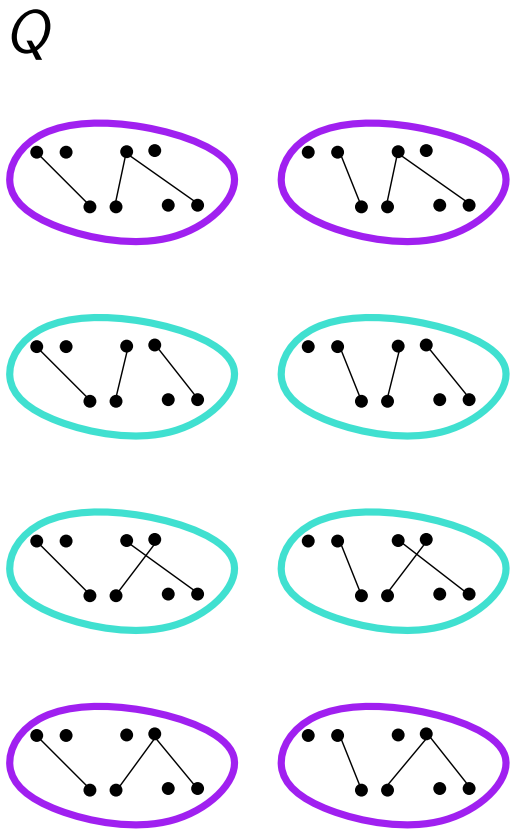


General Theorem Example

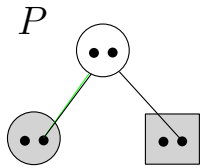


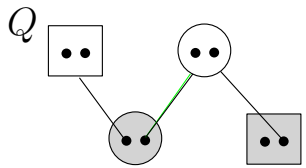


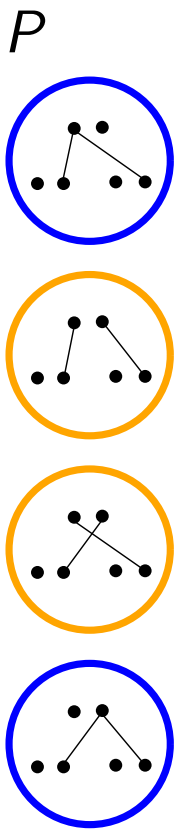


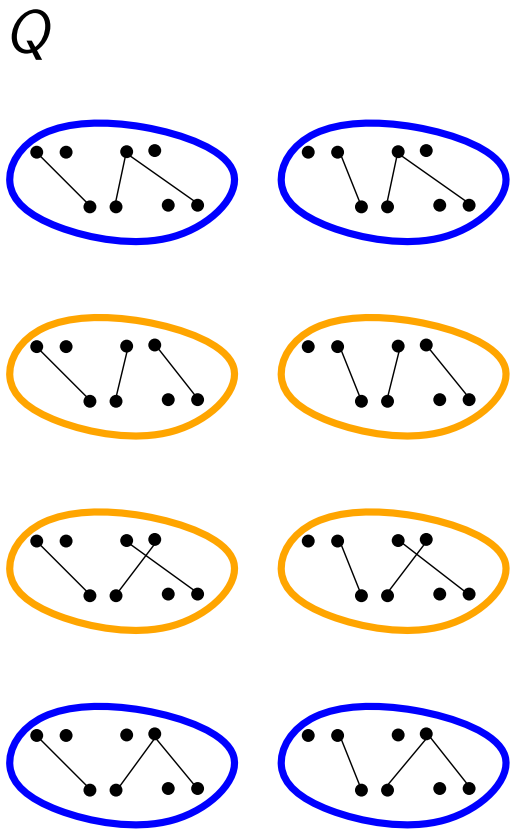


General Theorem Example

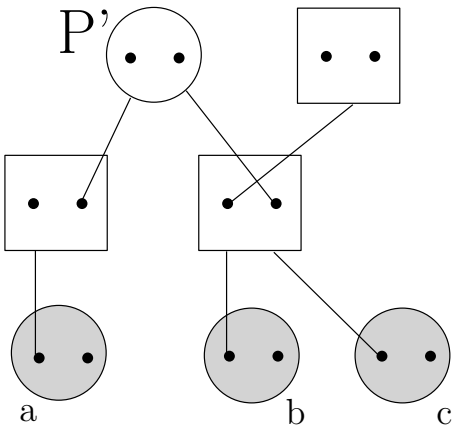




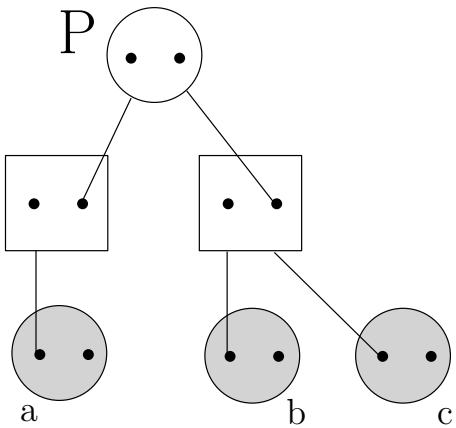




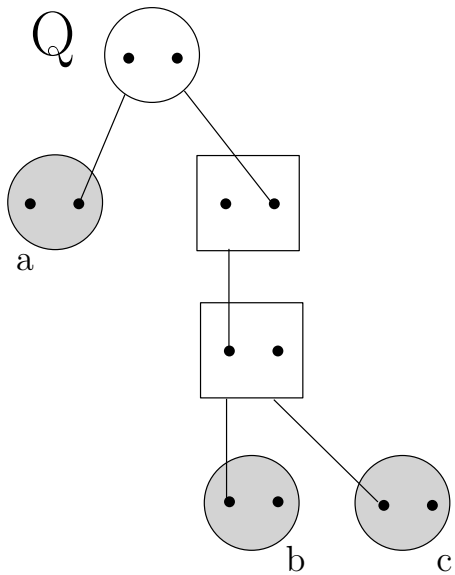
Non-Identifiable Pedigrees (linked and unlinked)



Superfluous Edge



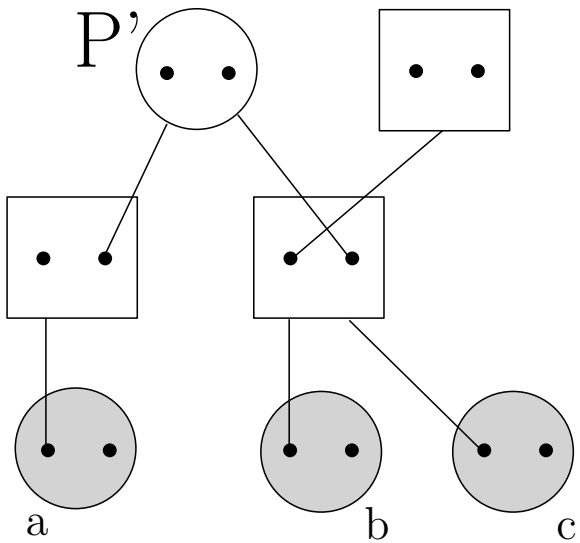
Half-Cousins a and b



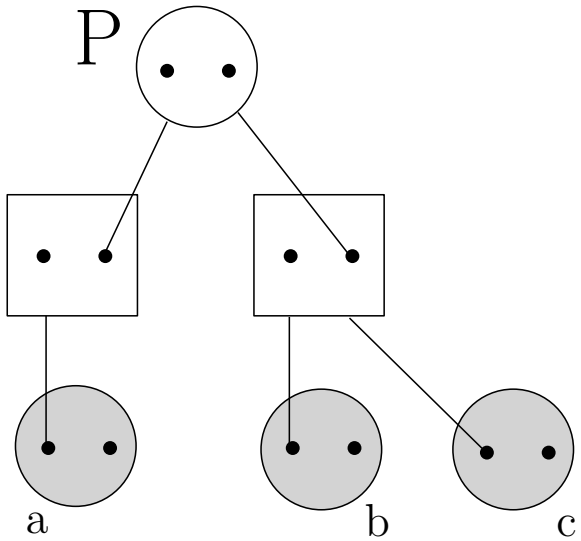
Grand-Half-Avuncular a and b

An example of three individuals with data. Obtained through application of the theorem.

Superfluous Individuals

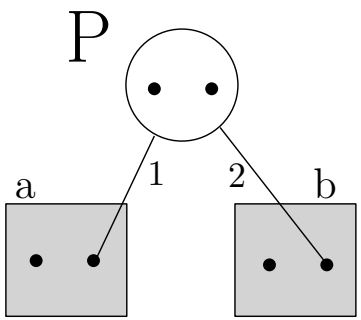


Superfluous Edge

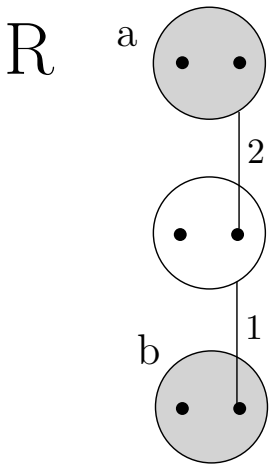


Half-Cousins a and b

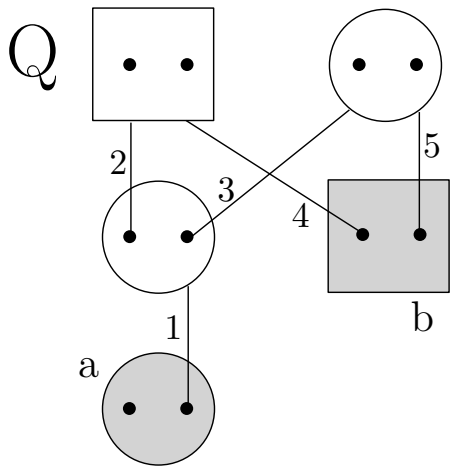
Identical Kinship, Non-identifiable (unlinked), Identifiable (linked)



Half-siblings



Grand-parent-grand-child



Avuncular (i.e. uncle-niece)

Using the kinship result, we can recapitulate the known fact [Pinto et al., 2010] that these three pedigrees have identical kinship.

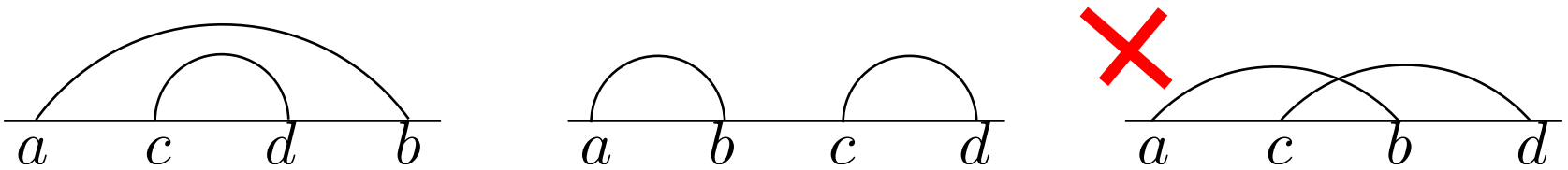
Nucleic Acids: More Examples

RNA

sequence of n nucleotides A,C,G,U

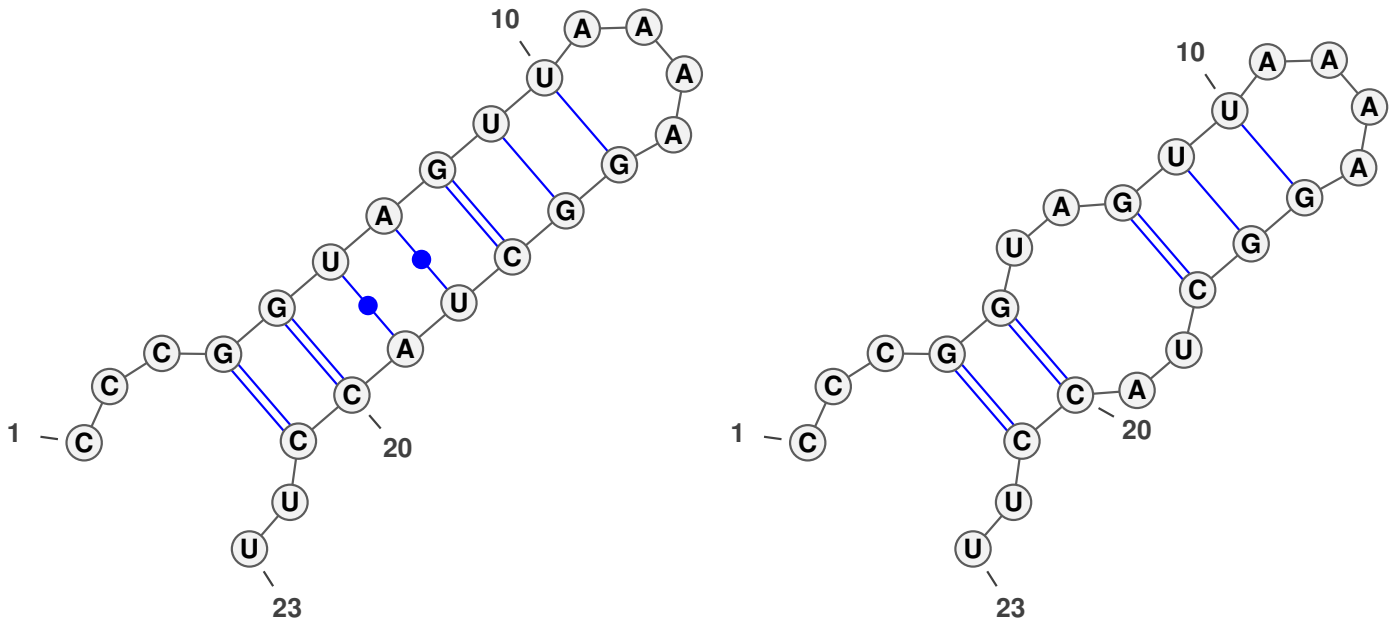
base pairs C-G, A-U, and G-U

secondary (2°) structure is pseudoknot-free, there are no two base pairs, a, b and c, d where $a < b$ and $c < d$ with overlapping intervals such that $a < c < b < d$.



There are $O(3^n)$ pseudoknot-free 2° structures.

Examples



free energy of 2° structure i is $E(i)$, computed using an $O(n)$ algorithm

Matrix Exponential

$P^Q(t)$ is the transition matrix for process Q where the i, j entry gives the probability of the process transitioning from state i to state j in t time.
The transition probabilities satisfy the differential equation

$$dP^Q(t)/dt = QP^Q(t), \quad t \geq 0.$$

The matrix exponential provides a solution to the above ODE

$$P^Q(t) = e^{tQ} := \sum_{n=1}^{\infty} (Qt)^n/(n!).$$

The population kinetics are given by $y_0 e^{tQ}$.