

Genome Rearrangement in Mitochondria and Its Computational Biology

István Miklós¹ and Jotun Hein²

¹Theoretical Biology and Ecology Group Hungarian Academic of Science

²Oxford Centre for Gene Function, University of Oxford

**2nd RECOMB Comparative Genomics Satellite Workshop
16-19 October 2004 Bertinoro, Italy**

Funded by EPSRC, MRC and Hungarian Academy of Science

Outline

- Inferring closely related mitochondrial genomes
- Metropolised partial importance sampler for inferring inversions and transpositions
 - The idea of using partial importance sampling in MCMC
 - Mixing properties
 - Algorithmic properties
 - Results
- Open problems
 - Increasing the computational time needed for one step
 - Increasing the speed of mixing
 - More complicated models
- Conclusions

Inferring closely related mitochondrial genomes

- Based on closely related genomes (breakpoint distance is small)
→ parsimony works well
- Inversions, transpositions, inverted transpositions do happen
- Short mutations are more frequent
- Mutations are correlated, very likely due to the control region
- There might be more complicated mutations causing big breakpoint distance in one step

Biologically more realistic models would be desired

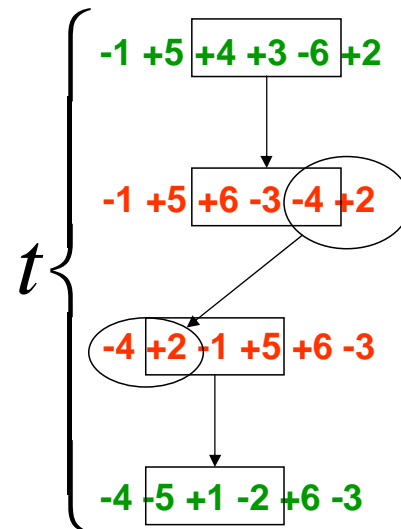
Stochastic modelling genome rearrangements

- Time continuous Markov model
- Each type of mutation has the same rate,
 - α : inversions β : transpositions and inverted transpositions
- The size of the state space is $2^n n!$, where n is the number of genes
- Analytical solution is unknown for any models

What we can calculate is the likelihood of a trajectory

$$\frac{e^{-\binom{n+1}{2}\alpha} (\alpha t)^l e^{-3\binom{n+1}{3}\beta t} (\beta t)^m}{(l+m)!}$$

$$\frac{e^{-21\alpha} (\alpha t)^2 e^{-105\beta t} \beta t}{6}$$



Bayesian approach

We can use the likelihood of trajectories to calculate the likelihood of observing two genomes:

$$P(G_2, G_1 | \alpha, \beta) = P(G_2 | G_1, \alpha, \beta) P_\infty(G_1, \alpha, \beta)$$

$$P(G_2 | G_1, \alpha, \beta) = \sum_{tr \in \text{Traj}(G_1, G_2)} P(tr | \alpha, \beta)$$

Due to symmetry:

$$P_\infty(G_1 | \alpha, \beta) = \frac{1}{2^n n!}$$

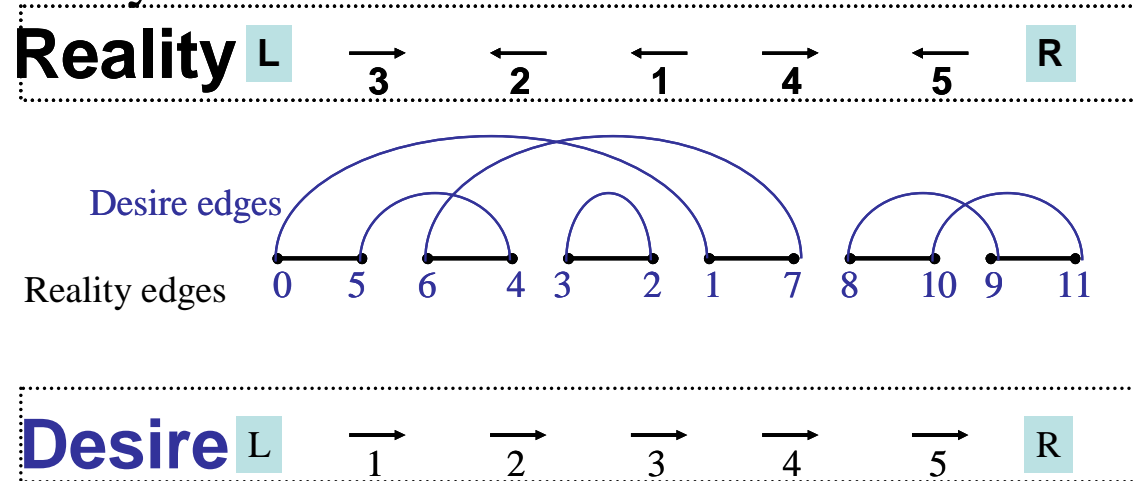
MCMC co-sampling parameters and trajectories.

Flat prior of parameters, Gibbs sampling is easy:

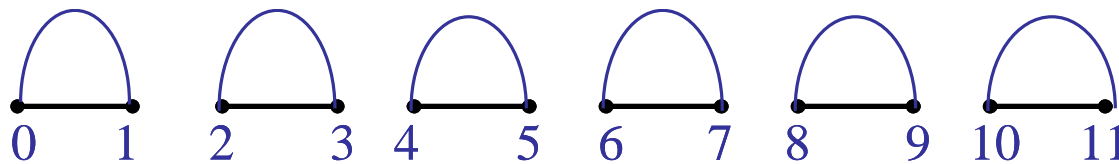
$$\begin{aligned} P(\alpha, \beta | tr) &\propto e^{-\binom{n+1}{2} \alpha t} (\alpha t)^l e^{-3 \binom{m+1}{3} \beta t} (\beta t)^m = \\ &= f(\alpha) g(\beta) \end{aligned}$$

Representation of signed permutations

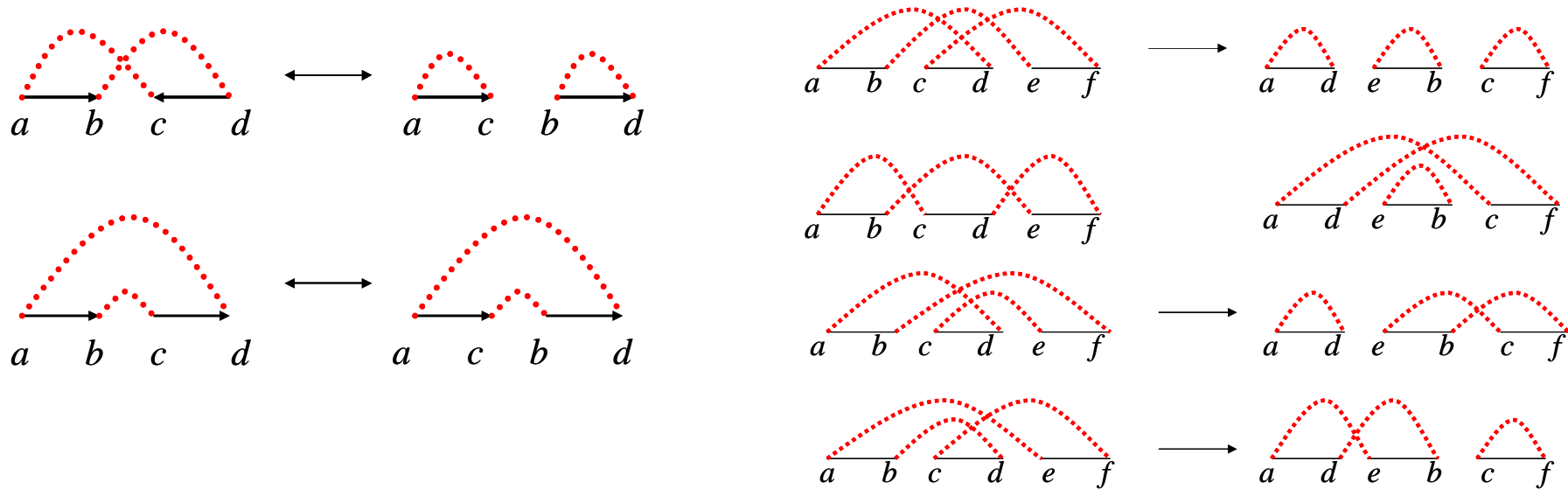
Graph of desire and reality



Based on basic group theory, transforming π_1 to π_2 is equivalent with transforming $\pi_2^{-1}\pi_1$ to the identical permutation. The graph of desire and reality of the identical permutation is $n+1$ cycle, all other permutations have less cycles:



Classification of mutations



Sampling distribution:

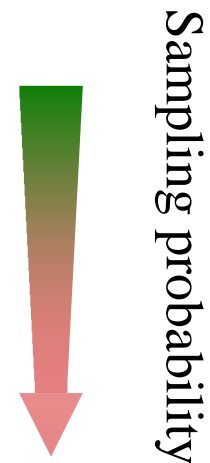
+2 transpositions

+1 transpositions and inversions

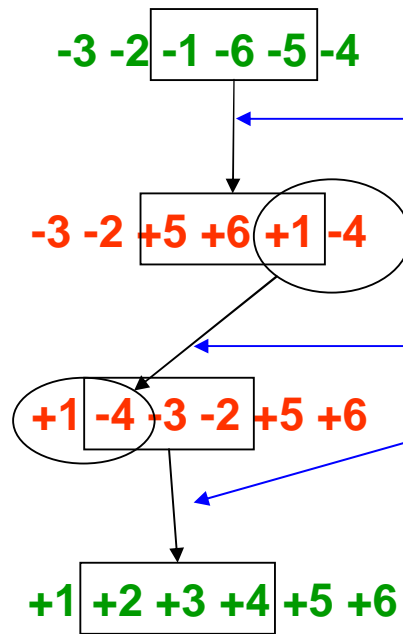
0 inversions

-1 inversions

0, -1, -2 transpositions



Sampling a trajectory

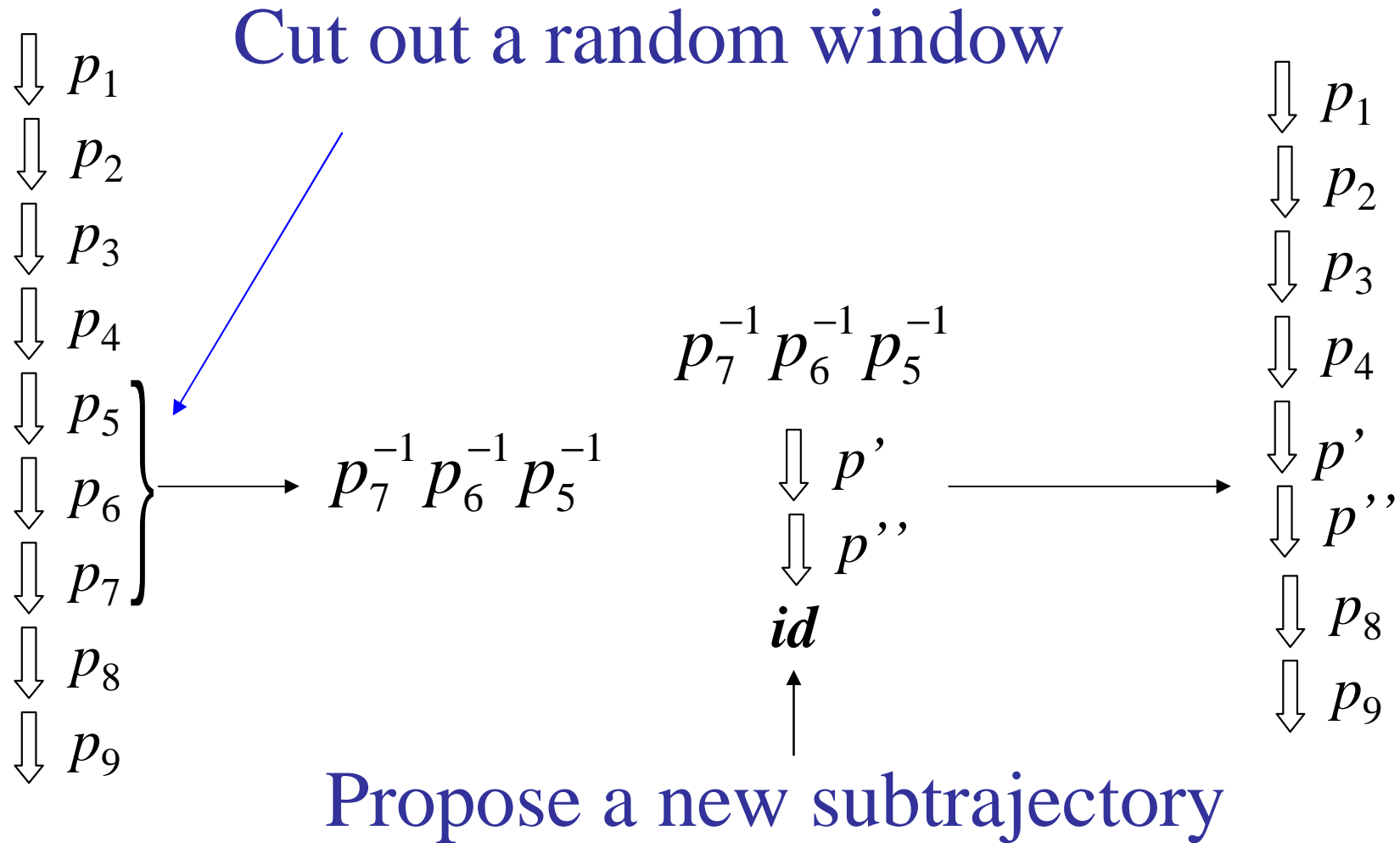


Propose a step in a way described above

Continue till sorted

Make a random decision whether or not to continue

Metropolised Partial Importance Sampling



Metropolis-Hastings ratio

Instead of calculating $P(Y|X)$, we calculate $P(Y, \text{window}|X)$ and use it in the Hastings ratio. This is proper, since the detailed balance still holds.

$$\pi(X)T(Y | X) =$$

$$\pi(X) \sum_i P(Y, w_i | X) \min \left(1, \frac{\pi(Y)P(X, w_i | Y)}{\pi(X)P(Y, w_i | X)} \right) =$$
$$\sum_i \min(\pi(X)P(Y, w_i | X), \pi(Y)P(X, w_i | Y))$$

which is a symmetric function

Window size & mixing

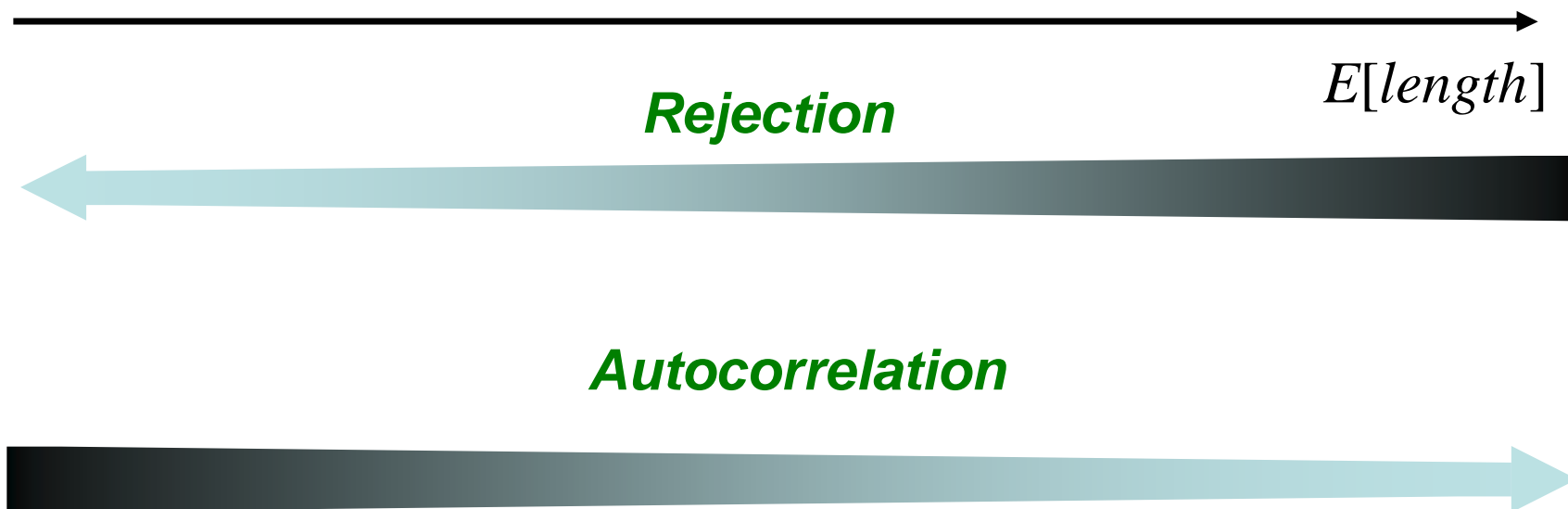
For reversible proposals, we need all possible window sizes

Miklós, (Ittész, Hein): geometric distribution

York, Durrett, Nielsen: $q(l) \propto 1 - \tanh\left(\xi\left(\frac{1}{\alpha N} - 1\right)\right)$

Local movements

Importance sampler

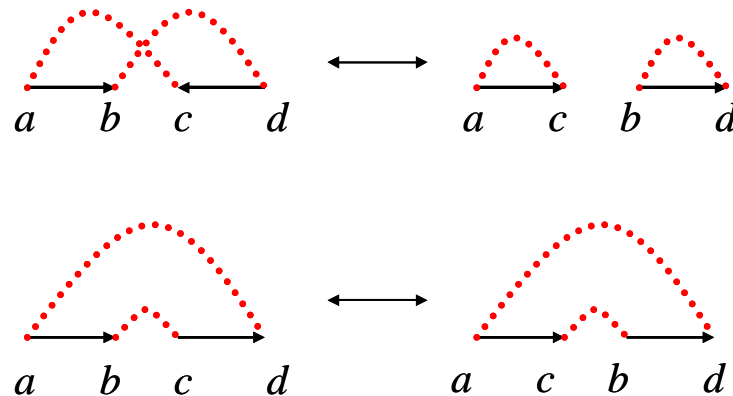


Proposing a step & mixing

How much time does it take to propose a step?

For inversions:

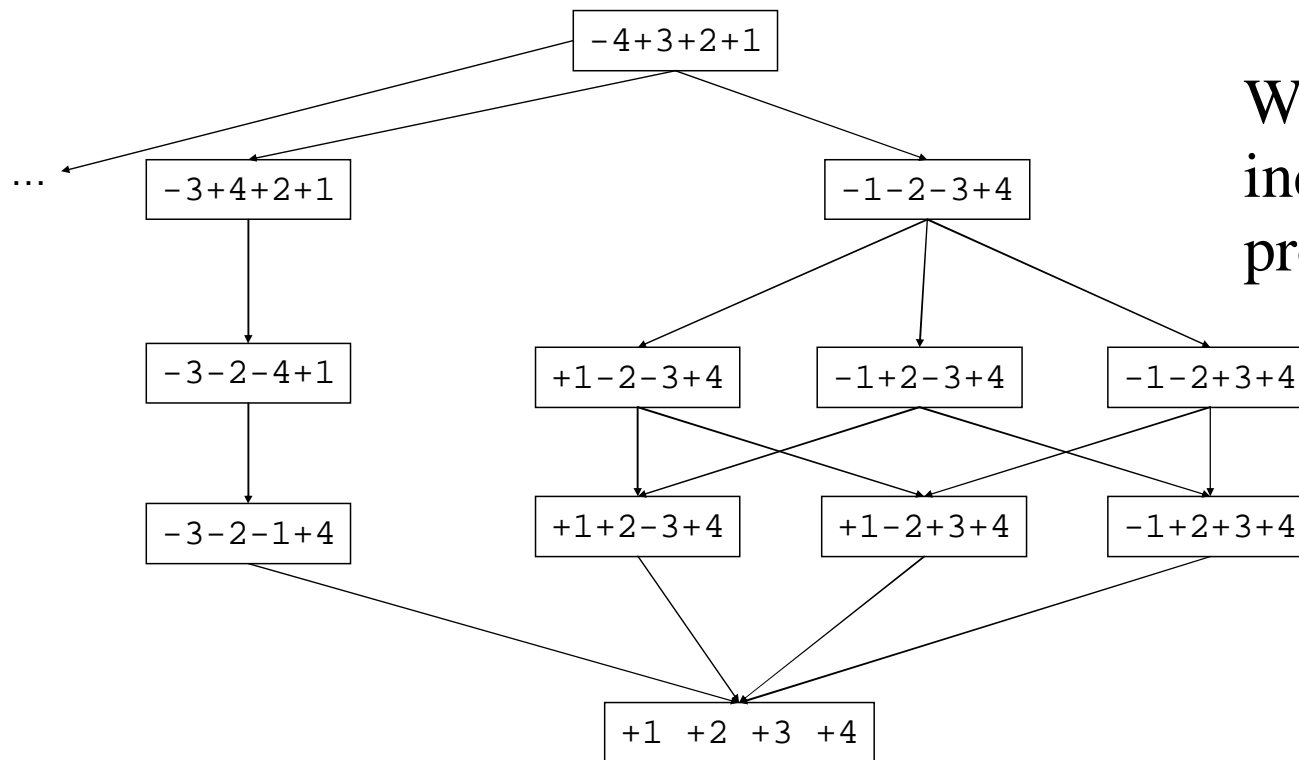
- Based on the number of cycles: it needs $O(n)$ time, where n is the number of genes:



- Based on the Hannenhalli-Pevzner theorem: it needs $O(n^2)$ time!

The later approximates the target distribution better, however, the overall performance is worse!

Why Hannenhalli-Pevzner does not yield an extremely good proposal?



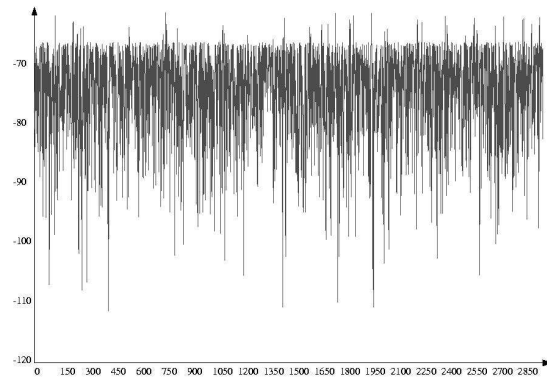
We choose the first steps indicated with the same probability ...

...hence we do not sample paths uniformly!!!

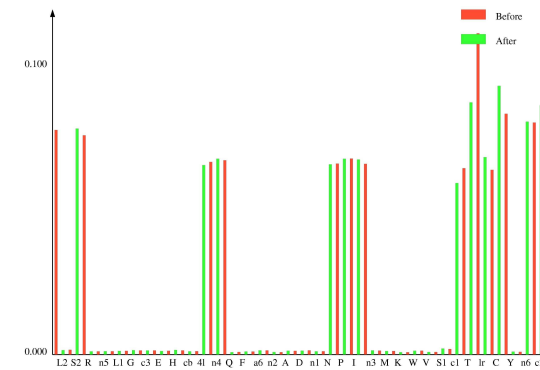
The longer the trajectory, the greater the bias is.

We would like to sample suboptimal paths, as well, according to their probabilities...

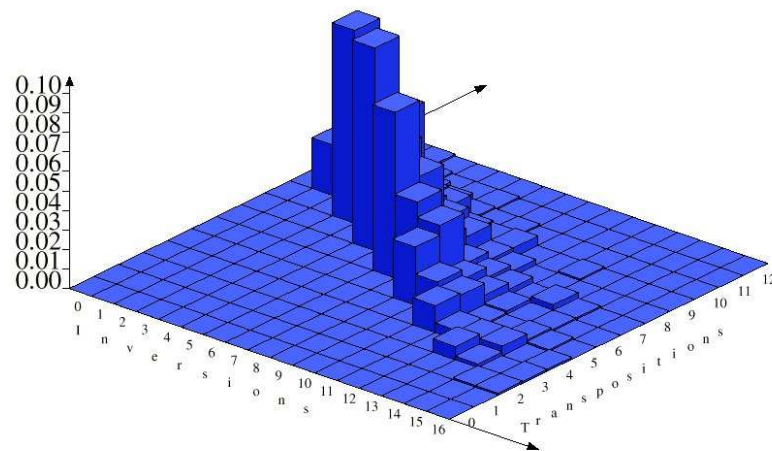
Results



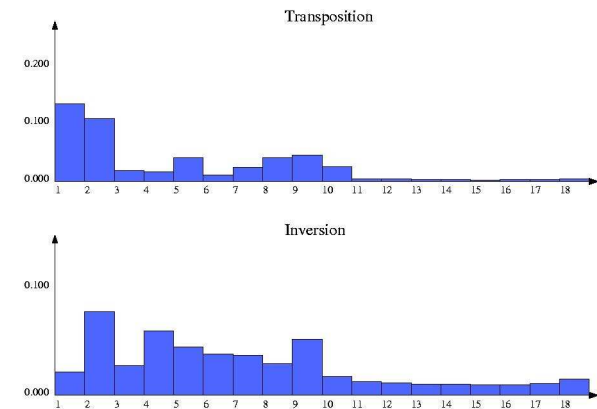
Log likelihood trace



Expected number of breakpoints



Posterior distribution of mutations



Length distribution

Miklós, Ittész, Hein (2004) Paris Genome Rearrangement server. *Bioinformatics*, adv. pub.

<http://www.stats.ox.ac.uk/~miklos/>

Open problems

- **Proposing transpositions in a faster way.**

State-of-the-art: Enumerating & listing +2 and +1 transpositions and inverted transpositions, sampling uniformly of them using this list, sampling other transpositions and inverted transpositions with rejection

Aim: Do it faster (In sub-cubic time)

- **Reusing the old graph of desire and reality after proposing a mutation**

State-of-the-art: Sampling a trajectory consisting of inversions takes $O(NL)$ time, where N is the number of genes, L is the length of trajectory.

Aim: Do it in $O(N+L)$ time

- **Other measurements of goodness of a mutation?**

State-of-the-art: Number of cycles it changes.

Aim: Breakpoints? Something other?

- **Preferring short mutations**

State-of-the-art: Surprisingly it is enough to change the likelihood calculations !!!

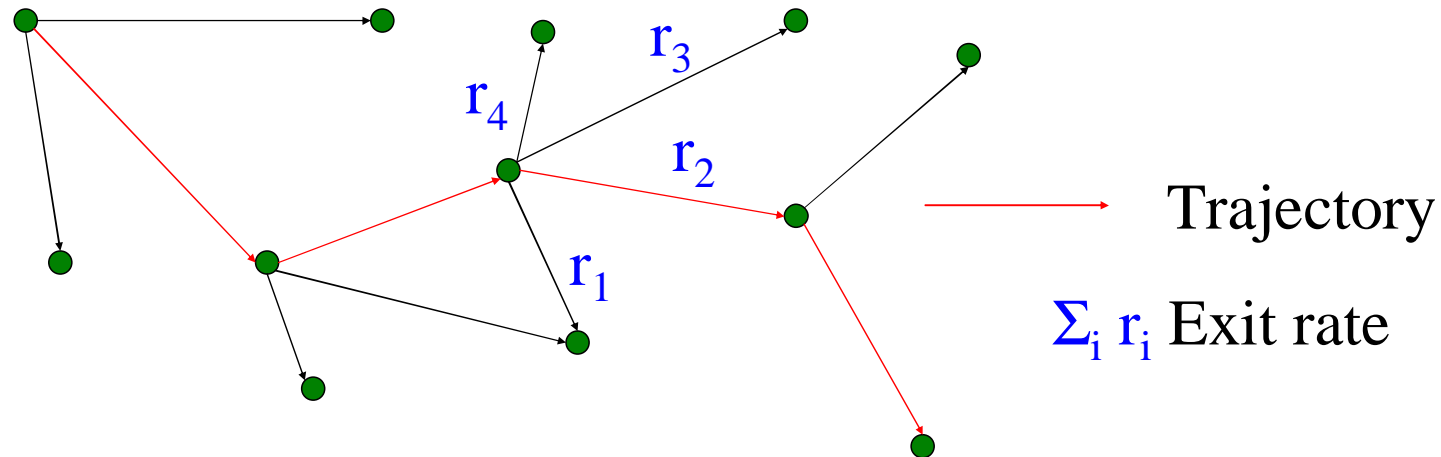
Aim: Other sampling strategy might mix faster

More complicated models I.

Exit rates might differ

Examples:

- Multi-chromosomal genomes
- Gene gain and loss
- Mutation rates depend on the previous mutation



State-of-the-art: Likelihood of trajectory still can be calculated in $O(L^2)$ time

Miklós, Lunter, Holmes (2004) *Mol. Biol. Evol.* 21(3):529-540.

Aim: Do it in time linear with the length of the path

More complicated models II.

- **Duplication-loss model**

State-of-the-art: Only optimisation methods exist (to our best of knowledge)

Aim: Stochastic modelling. Have no idea what a good proposal strategy should be...

- **Duplicated genes, different gene content**

State-of-the-art: Only optimisation methods exist (to our best of knowledge)

Aim: Stochastic modelling. Have no idea what a good proposal strategy should be...

Conclusions

- MCMC can handle more complicated models
- Extension to tree is easy (For example: Larget et al. (2002))
- Posterior distribution of several statistics
- Most likely is not always the more frequent!

Receipt of success

- Choose a mathematically complicated, but biologically interesting problem
- Such that:
 - Defining a stochastic model is easy
 - The problem is decomposable to well defined small parts
 - At least approximation algorithms exists for finding “the best” small part
 - This small part can be further cut out small dimensions. The approximation should be still good is these small dimensions
- Perturb the approximation method such that you can define an ergodic Markov chain
- Use this chain for MCMC
- Enjoy!

Already successfully applied

- Inversions, two genomes
York, Durrett, Nielsen (2002) *J. Comp. Biol.* **9**:808-818
- Inversions on trees
Larget, Simon, Kadane (2002) *J. Roy. Stat. Soc. B.* **64(4)**:681-695
- Inversions and translocations
Durrett, Nielsen, York (2004) *Genetics* **166**:621-629
- Inversions, transpositions, inverted transpositions
Miklós (2003) *Bioinformatics* **19**:ii130-137
Miklós, Ittész, Hein (2004) *Bioinformatics* Adv. pub.
- Cosampling alignments and trees
Miklós, Lunter, Drummond, Jensen, Hein (2004) submitted

In preparation

- Inversions, transpositions, inverted transpositions on trees
- Cosampling alignments, trees and RNA structures including pseudoknots