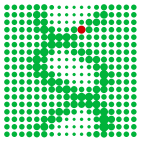
Quantifying the Equilibrium and Irreversibility Properties of the Nucleotide Substitution Process

Federico Squartini and Peter. F. Arndt

Max Planck Institute for Molecular Genetics

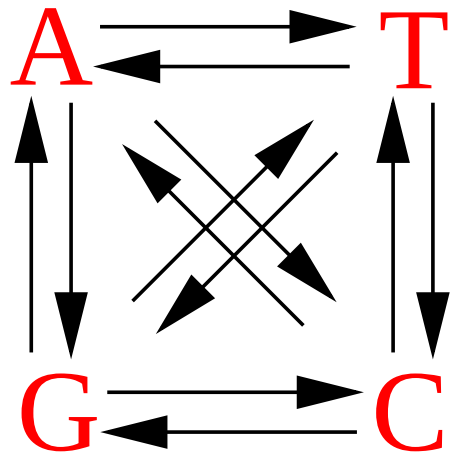


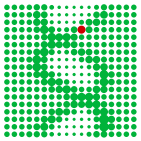
We will talk about disequilibrium and irreversibility. . .

Markovian Sequence Evolution

Nucleotide substitution models: i.i.d Markov models of evolution, i.e. a master equation:

$$\frac{\partial}{\partial t}\rho_\beta(t) = \sum_\alpha Q_{\beta\alpha}\rho_\alpha(t) \qquad \alpha, \beta \in \{\mathtt{A}, \mathtt{G}, \mathtt{C}, \mathtt{T}\}$$



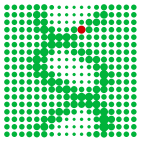


Markovian Sequence Evolution

Nucleotide substitution models: i.i.d Markov models of evolution, i.e. a master equation:

$$\frac{\partial}{\partial t}\rho_\beta(t) = \sum_\alpha Q_{\beta\alpha}\rho_\alpha(t) \qquad \alpha,\beta \in \{\mathtt{A},\mathtt{G},\mathtt{C},\mathtt{T}\}$$

$$Q = \begin{array}{c} \\ \mathtt{A} \\ \mathtt{C} \\ \mathtt{G} \\ \mathtt{T} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{A} & \mathtt{C} & \mathtt{G} & \mathtt{T} \end{array} \\ \begin{pmatrix} \cdot & Q_{\mathtt{AC}} & Q_{\mathtt{AG}} & Q_{\mathtt{AT}} \\ Q_{\mathtt{CA}} & \cdot & Q_{\mathtt{CG}} & Q_{\mathtt{CT}} \\ Q_{\mathtt{GA}} & Q_{\mathtt{GC}} & \cdot & Q_{\mathtt{GT}} \\ Q_{\mathtt{TA}} & Q_{\mathtt{TC}} & Q_{\mathtt{TG}} & \cdot \end{pmatrix} \end{array}.$$



Markovian Sequence Evolution

Nucleotide substitution models: i.i.d Markov models of evolution, i.e. a master equation:

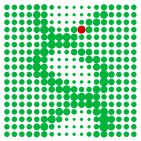
$$\frac{\partial}{\partial t}\rho_\beta(t) = \sum_\alpha Q_{\beta\alpha}\rho_\alpha(t) \qquad \alpha, \beta \in \{\mathtt{A}, \mathtt{G}, \mathtt{C}, \mathtt{T}\}$$

The solution to this equation, with initial condition ρ_0, is:

$$\rho_\beta(t) = \left[e^{Qt}\rho_0\right]_\beta$$

$$P(t) = e^{Qt}$$

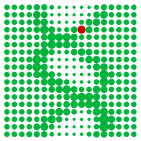
Such a model is not complete. . .



Choosing Parameters

Specifying an evolutionary mode $\Rightarrow$ postulating a form for the rate matrix:

$$Q = \begin{array}{c} \\ \mathtt{A} \\ \mathtt{C} \\ \mathtt{G} \\ \mathtt{T} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{A} & \mathtt{C} & \mathtt{G} & \mathtt{T} \end{array} \\ \begin{pmatrix} \cdot & Q_{\mathtt{AC}} & Q_{\mathtt{AG}} & Q_{\mathtt{AT}} \\ Q_{\mathtt{CA}} & \cdot & Q_{\mathtt{CG}} & Q_{\mathtt{CT}} \\ Q_{\mathtt{GA}} & Q_{\mathtt{GC}} & \cdot & Q_{\mathtt{GT}} \\ Q_{\mathtt{TA}} & Q_{\mathtt{TC}} & Q_{\mathtt{TG}} & \cdot \end{pmatrix} \end{array}.$$



Choosing Parameters

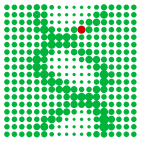
Specifying an evolutionary mode $\Rightarrow$ postulating a form for the rate matrix:

$$Q = \begin{array}{c} \\ \mathtt{A} \\ \mathtt{C} \\ \mathtt{G} \\ \mathtt{T} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{A} & \mathtt{C} & \mathtt{G} & \mathtt{T} \end{array} \\ \begin{pmatrix} \cdot & Q_{\mathtt{AC}} & Q_{\mathtt{AG}} & Q_{\mathtt{AT}} \\ Q_{\mathtt{CA}} & \cdot & Q_{\mathtt{CG}} & Q_{\mathtt{CT}} \\ Q_{\mathtt{GA}} & Q_{\mathtt{GC}} & \cdot & Q_{\mathtt{GT}} \\ Q_{\mathtt{TA}} & Q_{\mathtt{TC}} & Q_{\mathtt{TG}} & \cdot \end{pmatrix} \end{array}.$$

Widely used models:

$$Q = \begin{array}{c} \\ \mathtt{T} \\ \mathtt{C} \\ \mathtt{A} \\ \mathtt{G} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{T} & \mathtt{C} & \mathtt{A} & \mathtt{G} \end{array} \\ \begin{pmatrix} \cdot & \mu & \mu & \mu \\ \mu & \cdot & \mu & \mu \\ \mu & \mu & \cdot & \mu \\ \mu & \mu & \mu & \cdot \end{pmatrix} \end{array}.$$

Jukes-Cantor



Choosing Parameters

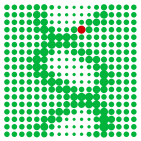
Specifying an evolutionary mode $\Rightarrow$ postulating a form for the rate matrix:

$$Q = \begin{array}{c} \\ \mathtt{A} \\ \mathtt{C} \\ \mathtt{G} \\ \mathtt{T} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{A} & \mathtt{C} & \mathtt{G} & \mathtt{T} \end{array} \\ \begin{pmatrix} \cdot & Q_{\mathtt{AC}} & Q_{\mathtt{AG}} & Q_{\mathtt{AT}} \\ Q_{\mathtt{CA}} & \cdot & Q_{\mathtt{CG}} & Q_{\mathtt{CT}} \\ Q_{\mathtt{GA}} & Q_{\mathtt{GC}} & \cdot & Q_{\mathtt{GT}} \\ Q_{\mathtt{TA}} & Q_{\mathtt{TC}} & Q_{\mathtt{TG}} & \cdot \end{pmatrix} \end{array}.$$

Widely used models:

$$Q = \begin{array}{c} \\ \mathtt{T} \\ \mathtt{C} \\ \mathtt{A} \\ \mathtt{G} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{T} & \mathtt{C} & \mathtt{A} & \mathtt{G} \end{array} \\ \begin{pmatrix} \cdot & \alpha & \beta & \beta \\ \alpha & \cdot & \beta & \beta \\ \beta & \beta & \cdot & \alpha \\ \beta & \beta & \alpha & \cdot \end{pmatrix} \end{array}.$$

Kimura



Choosing Parameters

Specifying an evolutionary mode $\Rightarrow$ postulating a form for the rate matrix:

$$Q = \begin{array}{c} \\ \mathtt{A} \\ \mathtt{C} \\ \mathtt{G} \\ \mathtt{T} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{A} & \mathtt{C} & \mathtt{G} & \mathtt{T} \end{array} \\ \begin{pmatrix} \cdot & Q_{\mathtt{AC}} & Q_{\mathtt{AG}} & Q_{\mathtt{AT}} \\ Q_{\mathtt{CA}} & \cdot & Q_{\mathtt{CG}} & Q_{\mathtt{CT}} \\ Q_{\mathtt{GA}} & Q_{\mathtt{GC}} & \cdot & Q_{\mathtt{GT}} \\ Q_{\mathtt{TA}} & Q_{\mathtt{TC}} & Q_{\mathtt{TG}} & \cdot \end{pmatrix} \end{array}.$$

Widely used models:

$$Q = \begin{array}{c} \\ \mathtt{T} \\ \mathtt{C} \\ \mathtt{A} \\ \mathtt{G} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{T} & \mathtt{C} & \mathtt{A} & \mathtt{G} \end{array} \\ \begin{pmatrix} \cdot & \pi_{\mathtt{T}} & \pi_{\mathtt{T}} & \pi_{\mathtt{T}} \\ \pi_{\mathtt{C}} & \cdot & \pi_{\mathtt{C}} & \pi_{\mathtt{C}} \\ \pi_{\mathtt{A}} & \pi_{\mathtt{A}} & \cdot & \pi_{\mathtt{A}} \\ \pi_{\mathtt{G}} & \pi_{\mathtt{G}} & \pi_{\mathtt{G}} & \cdot \end{pmatrix} \end{array}.$$

Felsenstein

Choosing Parameters

Specifying an evolutionary mode $\Rightarrow$ postulating a form for the rate matrix:

$$Q = \begin{array}{c} \\ \mathtt{A} \\ \mathtt{C} \\ \mathtt{G} \\ \mathtt{T} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{A} & \mathtt{C} & \mathtt{G} & \mathtt{T} \end{array} \\ \begin{pmatrix} \cdot & Q_{\mathtt{AC}} & Q_{\mathtt{AG}} & Q_{\mathtt{AT}} \\ Q_{\mathtt{CA}} & \cdot & Q_{\mathtt{CG}} & Q_{\mathtt{CT}} \\ Q_{\mathtt{GA}} & Q_{\mathtt{GC}} & \cdot & Q_{\mathtt{GT}} \\ Q_{\mathtt{TA}} & Q_{\mathtt{TC}} & Q_{\mathtt{TG}} & \cdot \end{pmatrix} \end{array}.$$

Widely used models:

$$Q = \begin{array}{c} \\ \mathtt{T} \\ \mathtt{C} \\ \mathtt{A} \\ \mathtt{G} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{T} & \mathtt{C} & \mathtt{A} & \mathtt{G} \end{array} \\ \begin{pmatrix} \cdot & k\pi_{\mathtt{T}} & \pi_{\mathtt{T}} & \pi_{\mathtt{T}} \\ k\pi_{\mathtt{C}} & \cdot & \pi_{\mathtt{C}} & \pi_{\mathtt{C}} \\ \pi_{\mathtt{A}} & \pi_{\mathtt{A}} & \cdot & k\pi_{\mathtt{A}} \\ \pi_{\mathtt{G}} & \pi_{\mathtt{G}} & k\pi_{\mathtt{G}} & \cdot \end{pmatrix} \end{array}.$$

Hasegawa-Kishino-Yano

Choosing Parameters

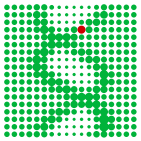
Specifying an evolutionary mode $\Rightarrow$ postulating a form for the rate matrix:

$$Q = \begin{array}{c} \\ \mathtt{A} \\ \mathtt{C} \\ \mathtt{G} \\ \mathtt{T} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{A} & \mathtt{C} & \mathtt{G} & \mathtt{T} \end{array} \\ \begin{pmatrix} \cdot & Q_{\mathtt{AC}} & Q_{\mathtt{AG}} & Q_{\mathtt{AT}} \\ Q_{\mathtt{CA}} & \cdot & Q_{\mathtt{CG}} & Q_{\mathtt{CT}} \\ Q_{\mathtt{GA}} & Q_{\mathtt{GC}} & \cdot & Q_{\mathtt{GT}} \\ Q_{\mathtt{TA}} & Q_{\mathtt{TC}} & Q_{\mathtt{TG}} & \cdot \end{pmatrix} \end{array}.$$

Widely used models:

$$Q = \begin{array}{c} \\ \mathtt{T} \\ \mathtt{C} \\ \mathtt{A} \\ \mathtt{G} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{T} & \mathtt{C} & \mathtt{A} & \mathtt{G} \end{array} \\ \begin{pmatrix} \cdot & k_1\pi_{\mathtt{T}} & \pi_{\mathtt{T}} & \pi_{\mathtt{T}} \\ k_1\pi_{\mathtt{C}} & \cdot & \pi_{\mathtt{C}} & \pi_{\mathtt{C}} \\ \pi_{\mathtt{A}} & \pi_{\mathtt{A}} & \cdot & k_2\pi_{\mathtt{A}} \\ \pi_{\mathtt{G}} & \pi_{\mathtt{G}} & k_2\pi_{\mathtt{G}} & \cdot \end{pmatrix} \end{array}.$$

Tamura-Nei



Two Evolutionary Models

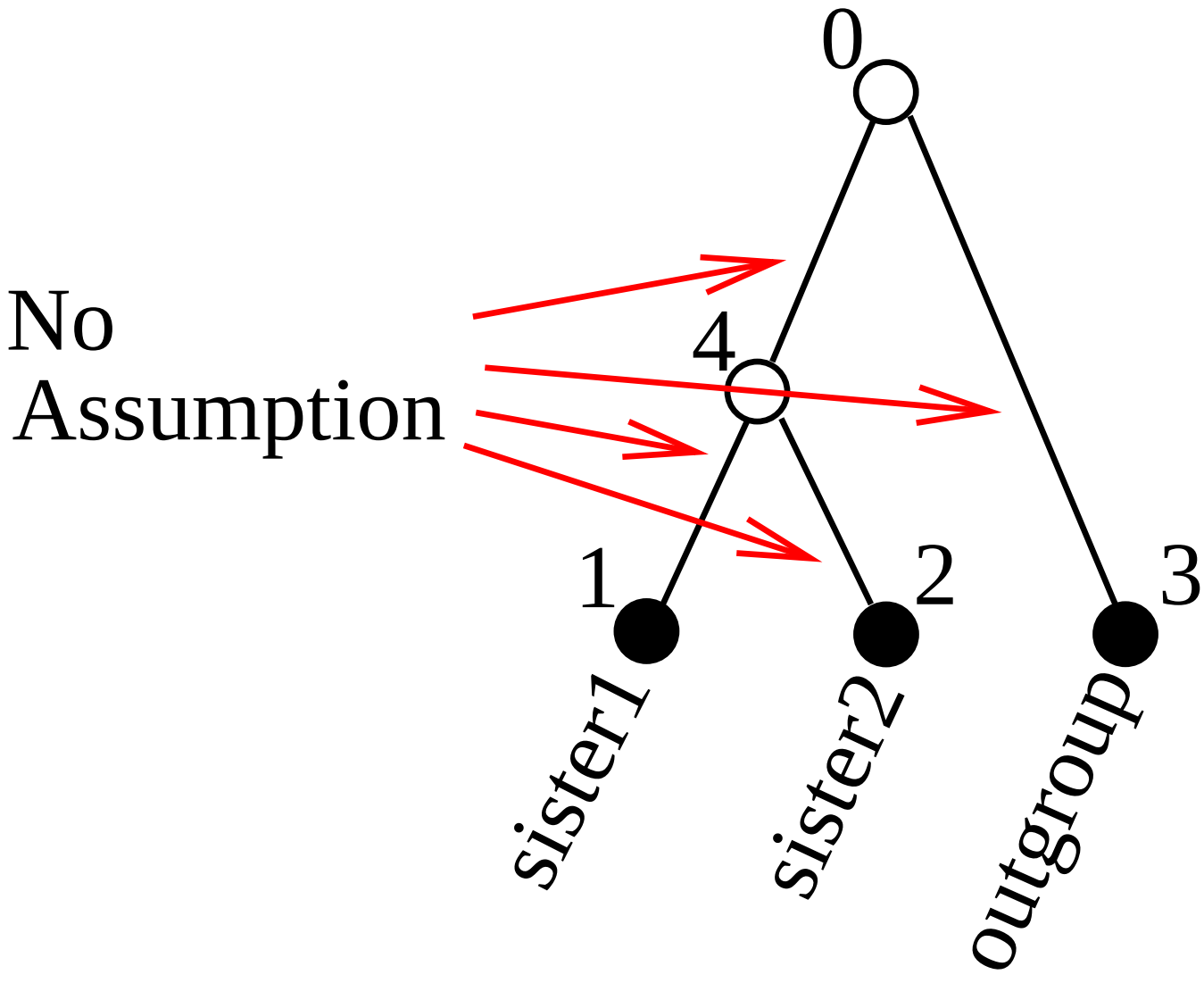
All preceding models are nested into the following:

$$Q_{\text{GTR}} = \begin{array}{c} \\ \mathtt{A} \\ \mathtt{G} \\ \mathtt{T} \\ \mathtt{C} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{A} & \mathtt{G} & \mathtt{T} & \mathtt{C} \end{array} \\ \begin{pmatrix} \cdot & a\pi_{\mathtt{A}} & b\pi_{\mathtt{A}} & c\pi_{\mathtt{A}} \\ a\pi_{\mathtt{G}} & \cdot & d\pi_{\mathtt{G}} & e\pi_{\mathtt{G}} \\ b\pi_{\mathtt{T}} & d\pi_{\mathtt{T}} & \cdot & f\pi_{\mathtt{T}} \\ c\pi_{\mathtt{C}} & e\pi_{\mathtt{C}} & f\pi_{\mathtt{C}} & \cdot \end{pmatrix} \end{array}.$$

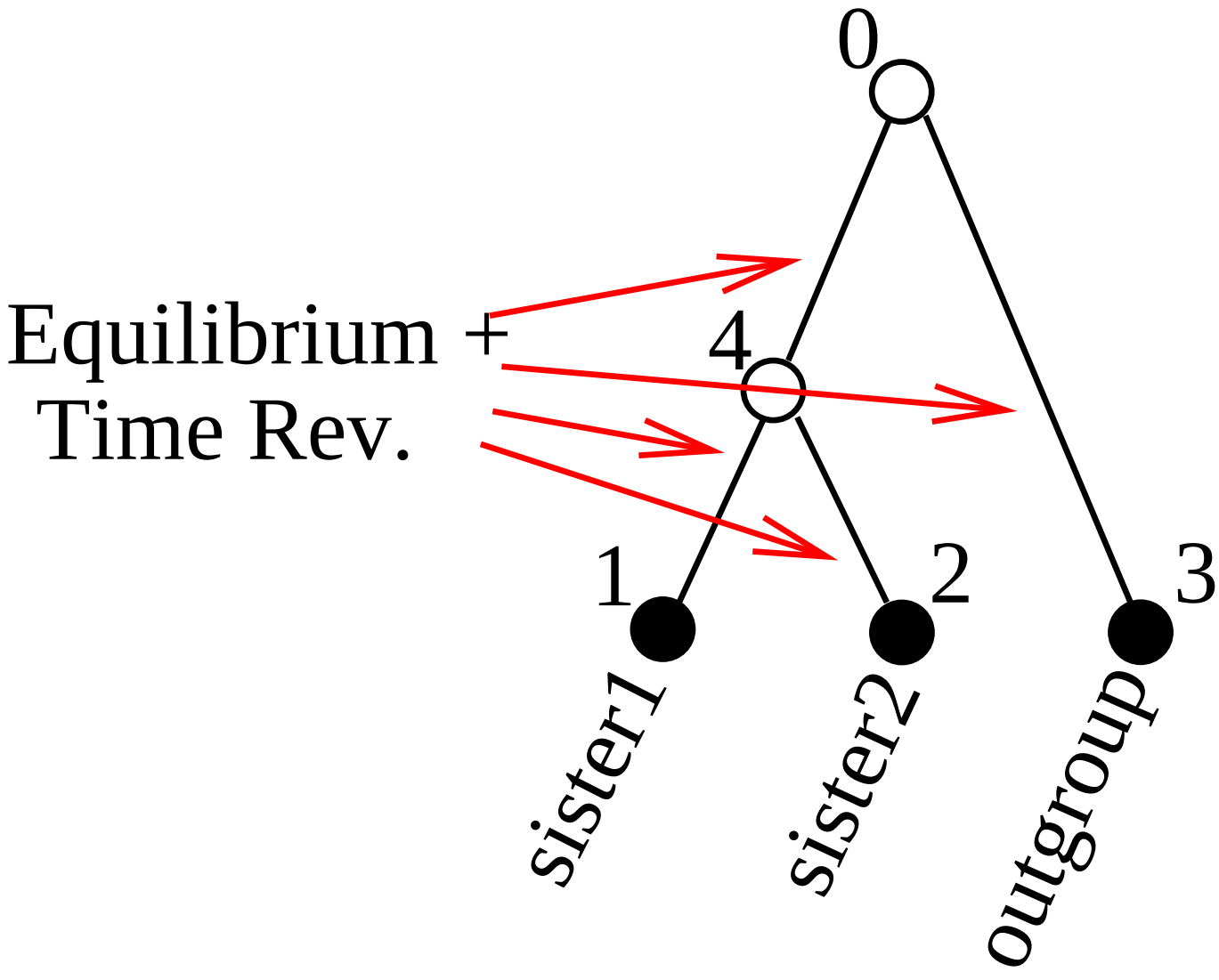
A possible alternative:

$$Q_{\text{RCS}} = \begin{array}{c} \\ \mathtt{A} \\ \mathtt{C} \\ \mathtt{G} \\ \mathtt{T} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{A} & \mathtt{C} & \mathtt{G} & \mathtt{T} \end{array} \\ \begin{pmatrix} \cdot & r_{\mathtt{AC}} & r_{\mathtt{AG}} & r_{\mathtt{AT}} \\ r_{\mathtt{GT}} & \cdot & r_{\mathtt{CG}} & r_{\mathtt{CT}} \\ r_{\mathtt{CT}} & r_{\mathtt{CG}} & \cdot & r_{\mathtt{GT}} \\ r_{\mathtt{AT}} & r_{\mathtt{AG}} & r_{\mathtt{AC}} & \cdot \end{pmatrix} \end{array}.$$

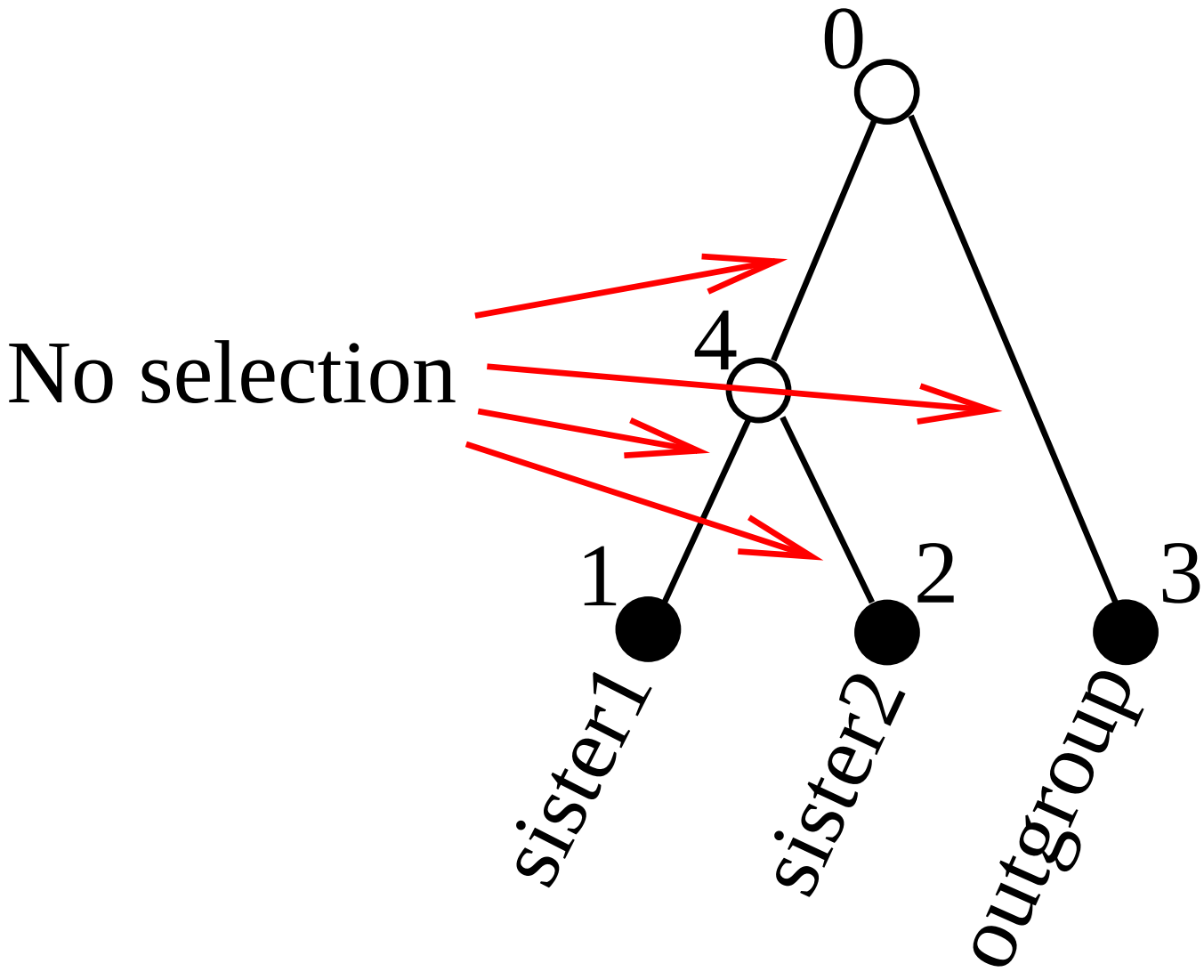
Two Evolutionary Models - 2



Two Evolutionary Models - 2



Two Evolutionary Models - 2

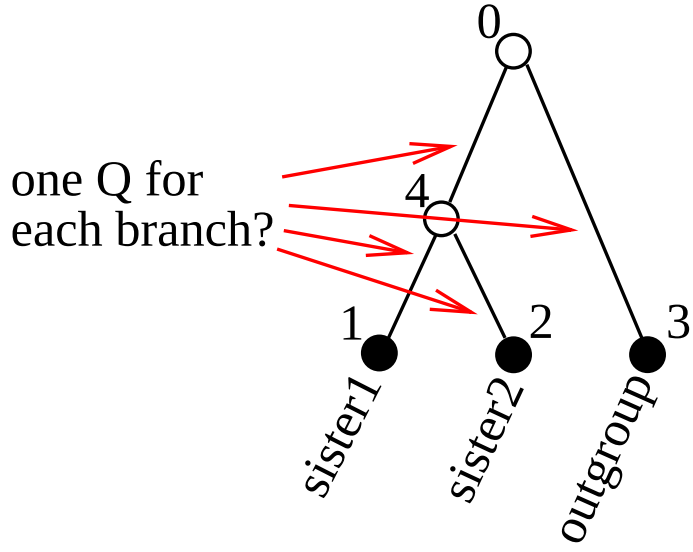


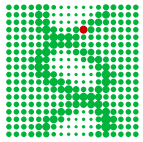
Estimating Parameters

For a given triple alignment $\vec{\alpha}^i$ of nucleotide sequences from 3 species, the likelihood of the alignment is:

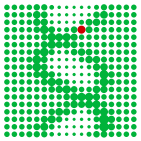
$$L = \prod_{k=1}^{N} \sum_{\alpha^0,\alpha^4 \in \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}} \rho^0_{\alpha^0} \left[P^{30}\right]_{\alpha^3_k \alpha^0} \left[P^{40}\right]_{\alpha^4 \alpha^0} \left[P^{24}\right]_{\alpha^2_k \alpha^4} \left[P^{14}\right]_{\alpha^1_k \alpha^4}$$

The vector ρ^0 represents the ancestral nucleotide distribution at the root node.





Equilibrium



The stationarity index

The equilibrium distribution of a Markov process is defined by:

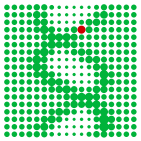
$$Q\pi = 0$$

Just taking the difference between present and stationary distribution:

$$\Delta_\alpha = \rho_\alpha - \pi_\alpha$$

And rearrange the terms:

$$\begin{aligned} \mathrm{STI}_1 &= \Delta_{\mathtt{C}} + \Delta_{\mathtt{G}} = \rho_{\mathtt{GC}} - \pi_{\mathtt{GC}} \\ \mathrm{STI}_2 &= \Delta_{\mathtt{A}} - \Delta_{\mathtt{T}} \\ \mathrm{STI}_3 &= \Delta_{\mathtt{C}} - \Delta_{\mathtt{G}}, \end{aligned}$$



The STI - Reverse complement symmetry

Substituting the equilibrium distribution:

$$(1 - \pi_{CG}, \pi_{CG}, \pi_{CG}, 1 - \pi_{CG})$$

Where:

$$\pi_{\mathtt{CG}} = \frac{r_{\mathtt{GT}} + r_{\mathtt{CT}}}{r_{\mathtt{AC}} + r_{\mathtt{AG}} + r_{\mathtt{GT}} + r_{\mathtt{CT}}}$$

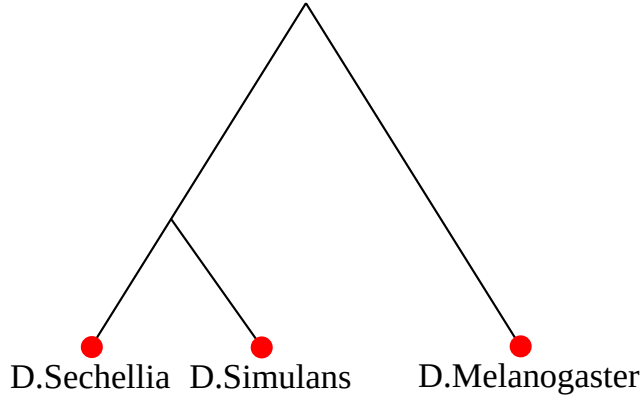
For the reverse complement symmetric model the STI has a simple form:

$$\begin{aligned} \mathsf{STI}_1 &= \rho_{\mathtt{GC}} - \pi_{\mathtt{GC}} \\ \mathsf{STI}_2 &= (\rho_{\mathtt{A}} - \rho_{\mathtt{T}}) \\ \mathsf{STI}_3 &= (\rho_{\mathtt{C}} - \rho_{\mathtt{G}}). \end{aligned}$$

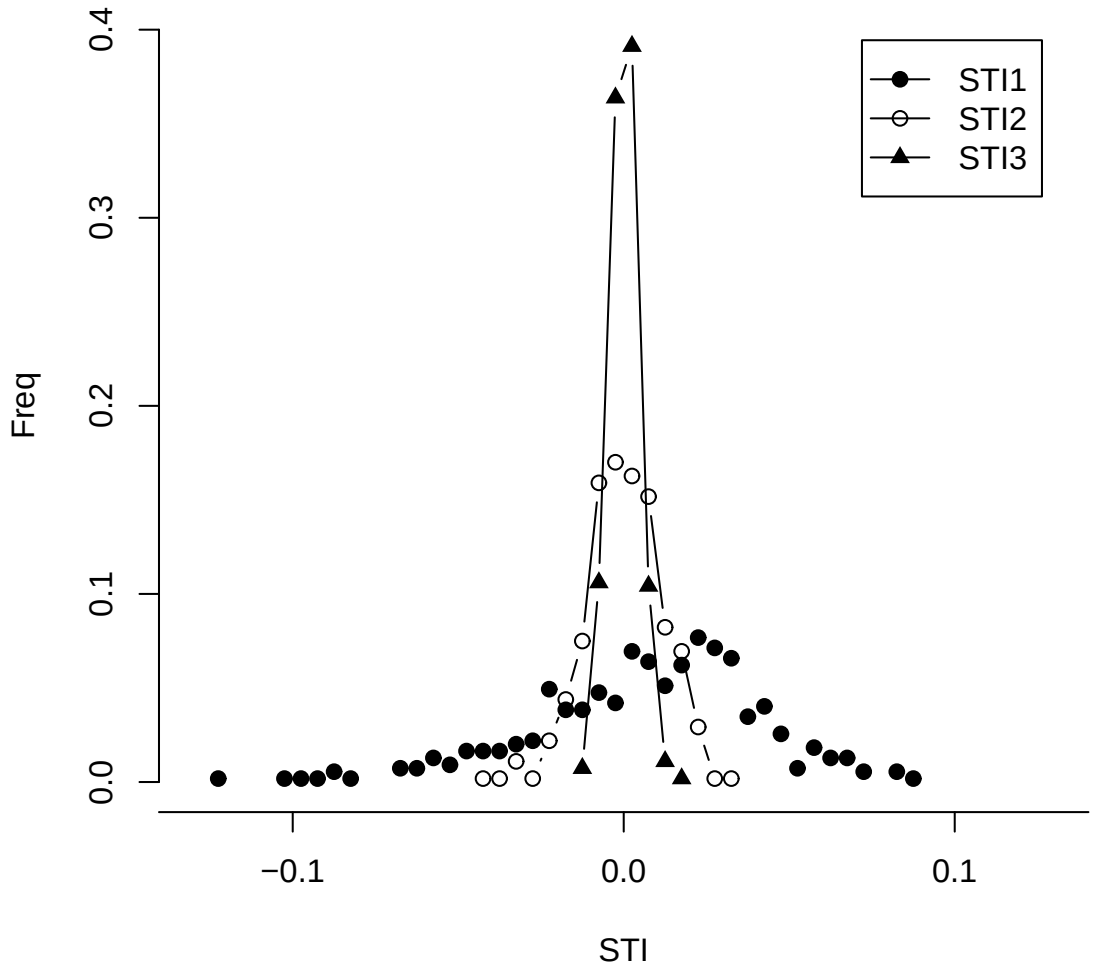
Analysis of the Fly Genome

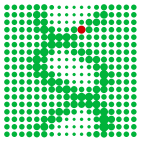
Results about the time reversal properties for the evolution of the fly genome:

- Alignment of 3 Drosophilas: sechellia, simulans and melanogaster
- Removed annotated coding regions
- Rates have been estimated using a maximum likelihood algorithm
- Sliding window analysis, 50kbp length
- For each window we have calculated the stationarity index in the simulans lineage

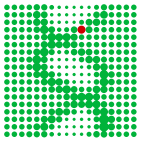


Analysis of the Fly Genome - Stationarity





Reversibility



Time Reversibility: the Detailed Balance

Time reversibility is usually defined in terms of the **detailed balance conditions**:

$$Q_{ji}\pi_i = Q_{ij}\pi_j$$

From which one can derive the General Time Reversible (GTR) Parameterization:

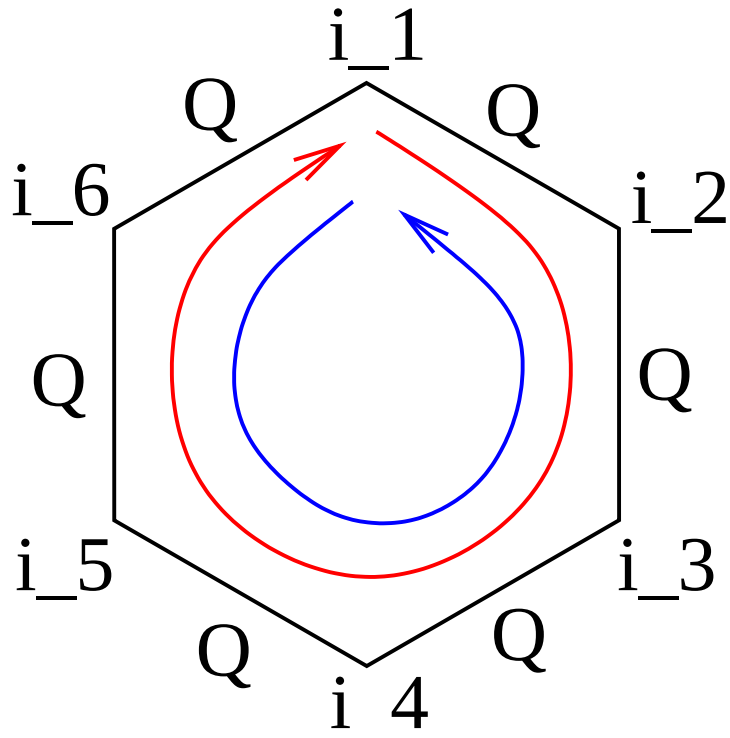
$$Q_{\text{GTR}} = \begin{array}{c} \\ \mathtt{A} \\ \mathtt{G} \\ \mathtt{T} \\ \mathtt{C} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{A} & \mathtt{G} & \mathtt{T} & \mathtt{C} \end{array} \\ \begin{pmatrix} \cdot & a\pi_{\mathtt{A}} & b\pi_{\mathtt{A}} & c\pi_{\mathtt{A}} \\ a\pi_{\mathtt{G}} & \cdot & d\pi_{\mathtt{G}} & e\pi_{\mathtt{G}} \\ b\pi_{\mathtt{T}} & d\pi_{\mathtt{T}} & \cdot & f\pi_{\mathtt{T}} \\ c\pi_{\mathtt{C}} & e\pi_{\mathtt{C}} & f\pi_{\mathtt{C}} & \cdot \end{pmatrix} \end{array}$$

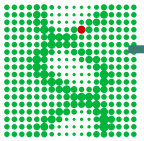
Time reversibility: Kolmogorov Cycle Conditions

A lesser known formulation of time reversibility:

Definition. *A Markov process is said to satisfy the Kolmogorov cycle conditions if the following equality on generators holds:*

$$Q_{i_1 i_n} Q_{i_n i_{n-1}} \dots Q_{i_2 i_1} = Q_{i_1 i_2} \dots Q_{i_{n-1} i_n} Q_{i_n i_1} \qquad \forall i_1, \dots, i_n \in \mathcal{C} \tag{-2}$$





Time reversibility: Kolmogorov Cycle Conditions - 2

Moreover the following proposition (relevant when analyzing biological sequences) holds:

Proposition. *If the coefficients of the rate matrix are strictly positive and if Kolmogorov conditions hold for three cycles then they hold for cycles of arbitrary length.*

Proposition. *Given a four states Markov process with strictly positive rate matrix coefficients, if the conditions:*

$$Q_{\alpha\delta}Q_{\delta\gamma}Q_{\gamma\beta}Q_{\beta\alpha} = Q_{\alpha\beta}Q_{\beta\gamma}Q_{\gamma\delta}Q_{\delta\alpha}, \qquad \text{(-2)}$$

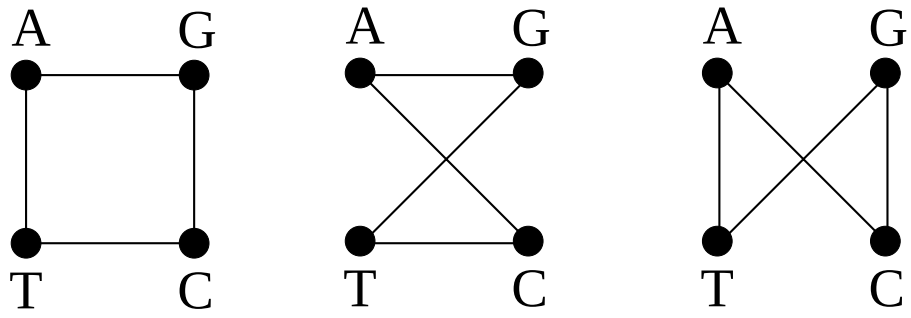
hold for $(\alpha, \beta, \gamma, \delta)$ *equal to* $(\mathtt{A}, \mathtt{G}, \mathtt{C}, \mathtt{T})$, $(\mathtt{A}, \mathtt{G}, \mathtt{T}, \mathtt{C})$ *and* $(\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T})$ *then Kolmogorov conditions hold for* 3*-cycles.*

Ans lastly:

Proposition. *If the coefficients of the rate matrix are strictly positive and if Kolmogorov conditions hold for four cycles then they hold for cycles of arbitrary length.*

IRI - The general iid case

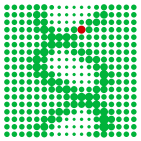
To check reversibility for nucleotide sequences we need to check the following conditions on four cycles:



$$\mathsf{IRI}_1 := \frac{Q_{\mathtt{AG}}Q_{\mathtt{GC}}Q_{\mathtt{CT}}Q_{\mathtt{TA}} - Q_{\mathtt{AT}}Q_{\mathtt{TC}}Q_{\mathtt{CG}}Q_{\mathtt{GA}}}{Q_{\mathtt{AG}}Q_{\mathtt{GC}}Q_{\mathtt{CT}}Q_{\mathtt{TA}} + Q_{\mathtt{AT}}Q_{\mathtt{TC}}Q_{\mathtt{CG}}Q_{\mathtt{GA}}}$$

$$\mathsf{IRI}_2 := \frac{Q_{\mathtt{AC}}Q_{\mathtt{CT}}Q_{\mathtt{TG}}Q_{\mathtt{GA}} - Q_{\mathtt{AG}}Q_{\mathtt{GT}}Q_{\mathtt{TC}}Q_{\mathtt{CA}}}{Q_{\mathtt{AC}}Q_{\mathtt{CT}}Q_{\mathtt{TG}}Q_{\mathtt{GA}} + Q_{\mathtt{AG}}Q_{\mathtt{GT}}Q_{\mathtt{TC}}Q_{\mathtt{CA}}}$$

$$\mathsf{IRI}_3 := \frac{Q_{\mathtt{AC}}Q_{\mathtt{CG}}Q_{\mathtt{GT}}Q_{\mathtt{TA}} - Q_{\mathtt{AT}}Q_{\mathtt{TG}}Q_{\mathtt{GC}}Q_{\mathtt{CA}}}{Q_{\mathtt{AC}}Q_{\mathtt{CG}}Q_{\mathtt{GT}}Q_{\mathtt{TA}} + Q_{\mathtt{AT}}Q_{\mathtt{TG}}Q_{\mathtt{GC}}Q_{\mathtt{CA}}}$$



Iri for the Reverse Complement Symmetric Model

Out of the previous indices we get a specialized version of the IRI:

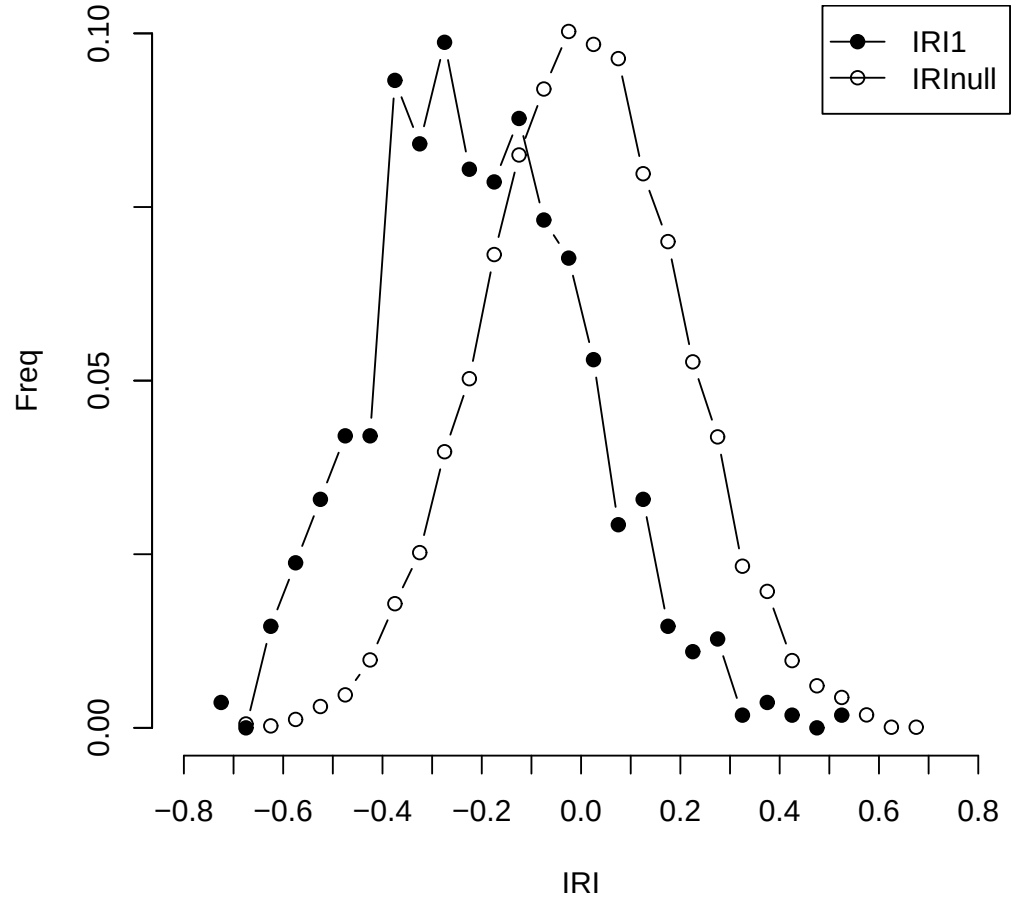
$$\begin{aligned} \mathsf{IRI}_1 &= \frac{r_{\mathtt{AG}}^2 r_{\mathtt{GT}}^2 - r_{\mathtt{AC}}^2 r_{\mathtt{CT}}^2}{r_{\mathtt{AG}}^2 r_{\mathtt{GT}}^2 + r_{\mathtt{AC}}^2 r_{\mathtt{CT}}^2} \\ \mathsf{IRI}_2 &= 0 \\ \mathsf{IRI}_3 &= 0 \end{aligned}$$

The IRI_1 will thus be comprised in the interval $[-1, 1]$ and if the system under study evolves time symmetrically:

$$\mathsf{IRI}_1 = 0$$

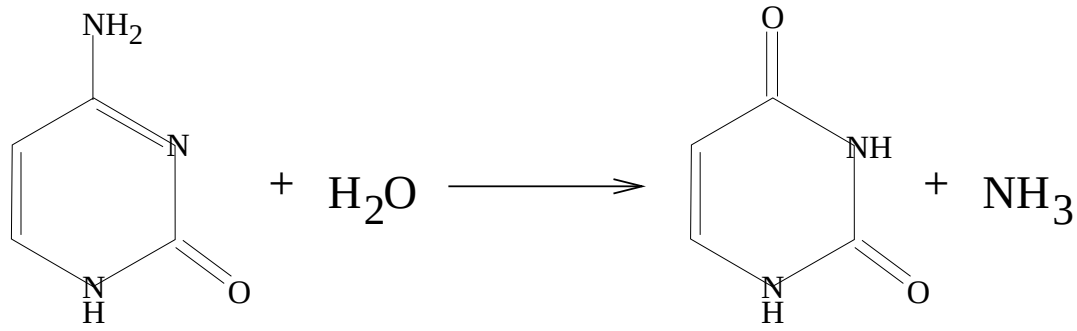
Irreversibility in the Fly Genome

Plots of the IRI for the Drosophila simulans genome and for the null model:



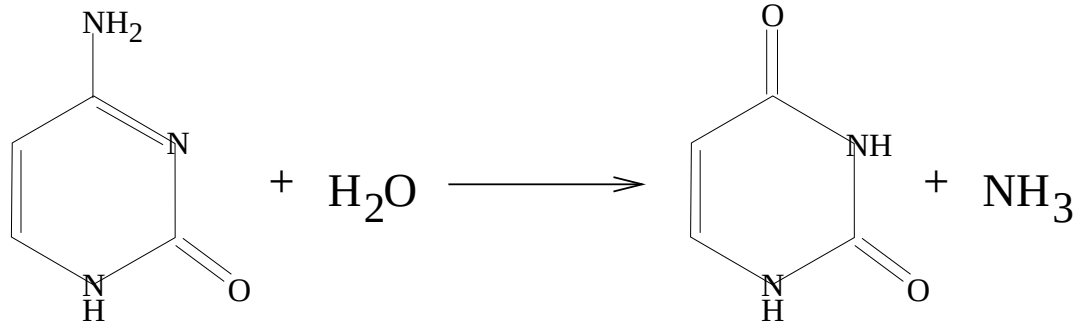
If water is around. . .

Cytosine can easily decay into Uracil:

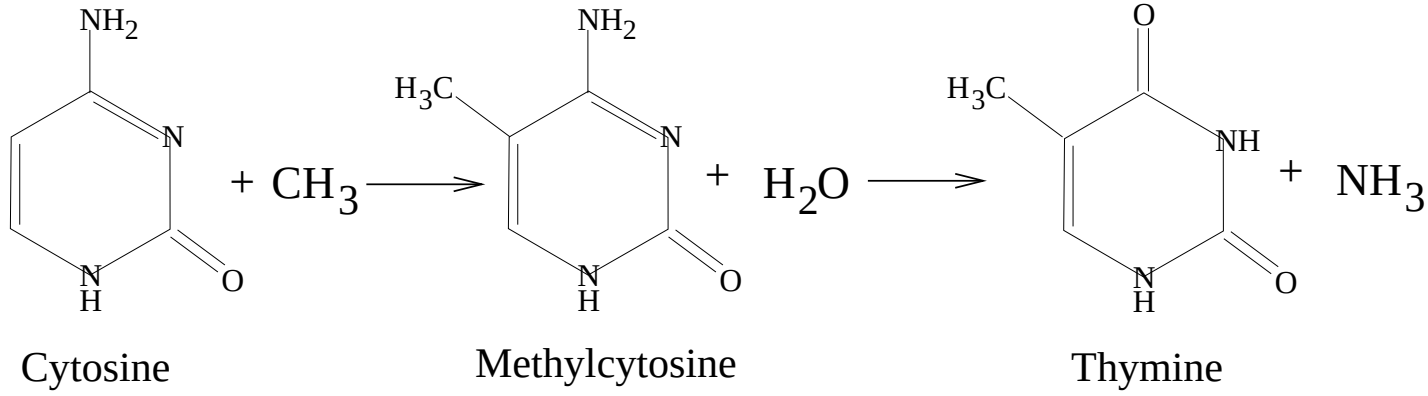


If water is around...

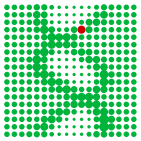
Cytosine can easily decay into Uracil:



On the other hand GpC pairs often occur in a methylated form:



The net effect is the decay of CpG pairs into TpG and CpA pairs.



A Nucleotide Substitution Model with CpG Decay

We need to extend the configuration space:

$$\mathcal{C} = s_1 \times \ldots \times s_N \qquad s_i \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}.$$

We assume the following form for the generator:

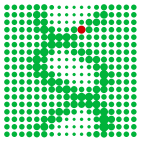
$$\mathcal{Q} = \sum_{i=1}^{N} \mathcal{Q}_i + \sum_{i=1}^{N-1} \mathcal{Q}_{i,i+1}^{\mathtt{CpG}}.$$

Where:

$$\mathcal{Q}_i = \underbrace{\mathbb{I} \otimes \ldots \otimes \mathbb{I}}_{i-1} \otimes Q \otimes \underbrace{\mathbb{I} \otimes \ldots \otimes \mathbb{I}}_{N-i}.$$

And:

$$\mathcal{Q}_{i,i+1}^{\mathtt{CpG}} = \underbrace{\mathbb{I} \otimes \ldots \otimes \mathbb{I}}_{i-1} \otimes Q^{\mathtt{CpG}} \otimes \underbrace{\mathbb{I} \otimes \ldots \otimes \mathbb{I}}_{N-i-1}.$$



The IRI of a Process with CpG Decay

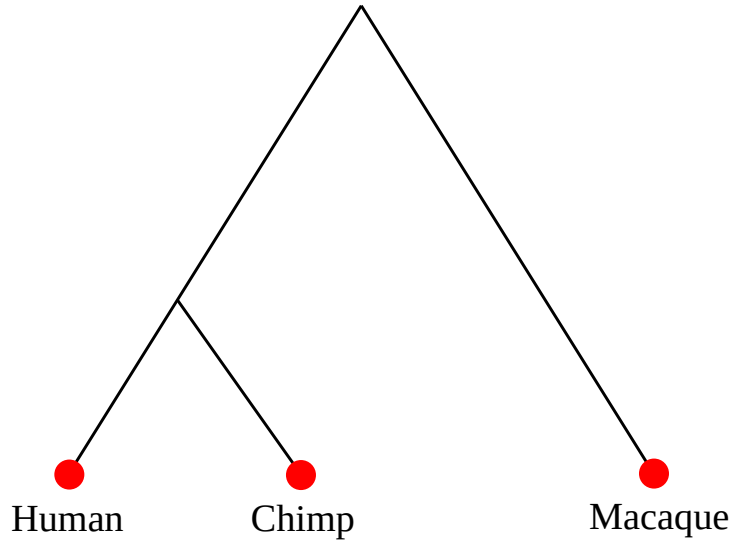
We get two IRI's in this case:

$$\mathsf{IRI}_1 := \frac{r_{\mathtt{AG}}^2 r_{\mathtt{GT}}^2 - r_{\mathtt{AC}}^2 r_{\mathtt{CT}}^2}{r_{\mathtt{AG}}^2 r_{\mathtt{GT}}^2 + r_{\mathtt{AC}}^2 r_{\mathtt{CT}}^2}$$

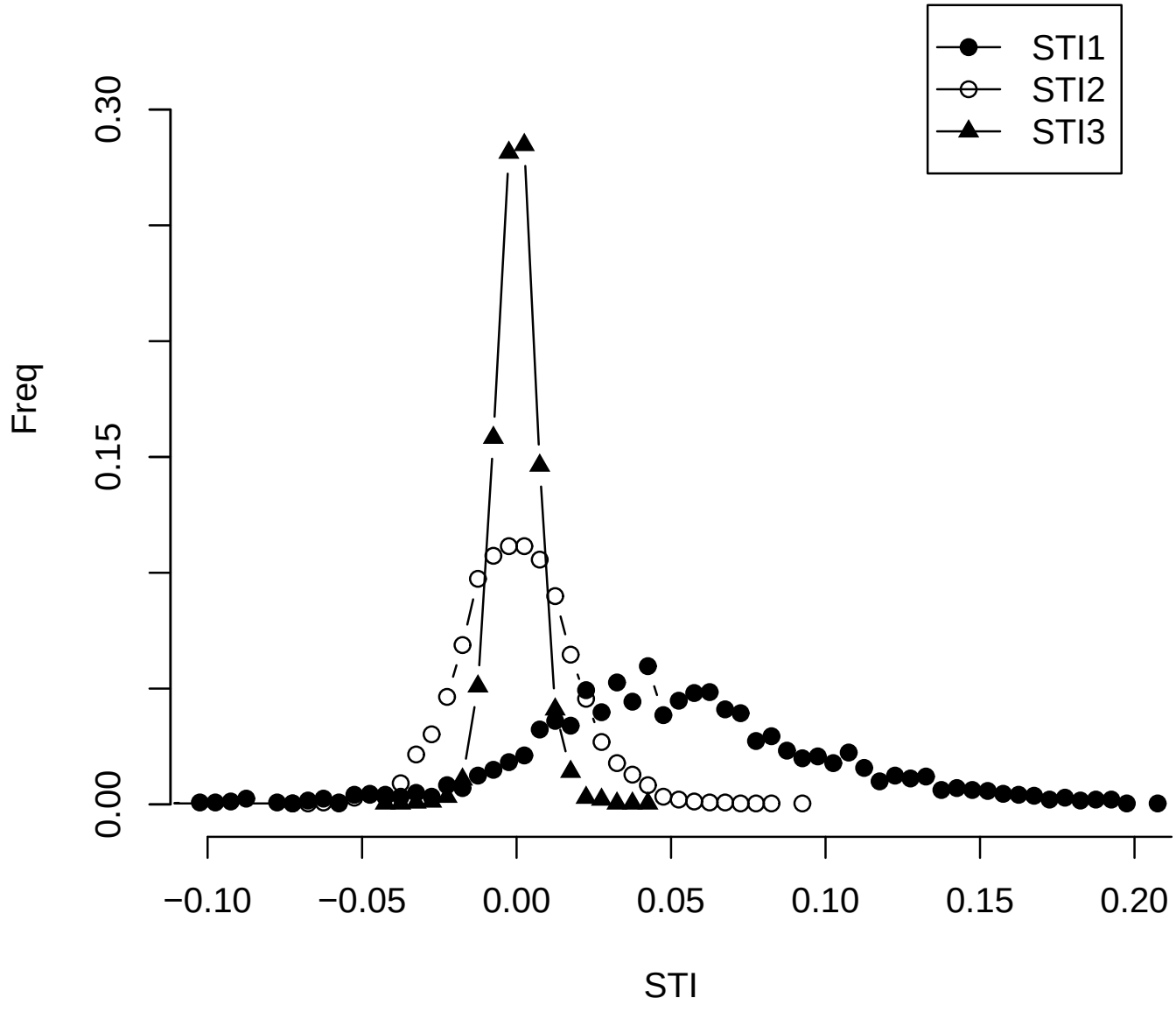
$$\mathsf{IRI}_{\mathtt{CpG}} := \frac{r_{\mathtt{GT}}^2 (r_{\mathtt{AG}} + r_{\mathtt{CpG}})^2 - (r_{\mathtt{CT}} + r_{\mathtt{CpG}}^{\mathtt{rev}})^2 r_{\mathtt{AC}}^2}{r_{\mathtt{GT}}^2 (r_{\mathtt{AG}} + r_{\mathtt{CpG}})^2 + (r_{\mathtt{CT}} + r_{\mathtt{CpG}}^{\mathtt{rev}})^2 r_{\mathtt{AC}}^2}$$

Analysis of the Human Genome

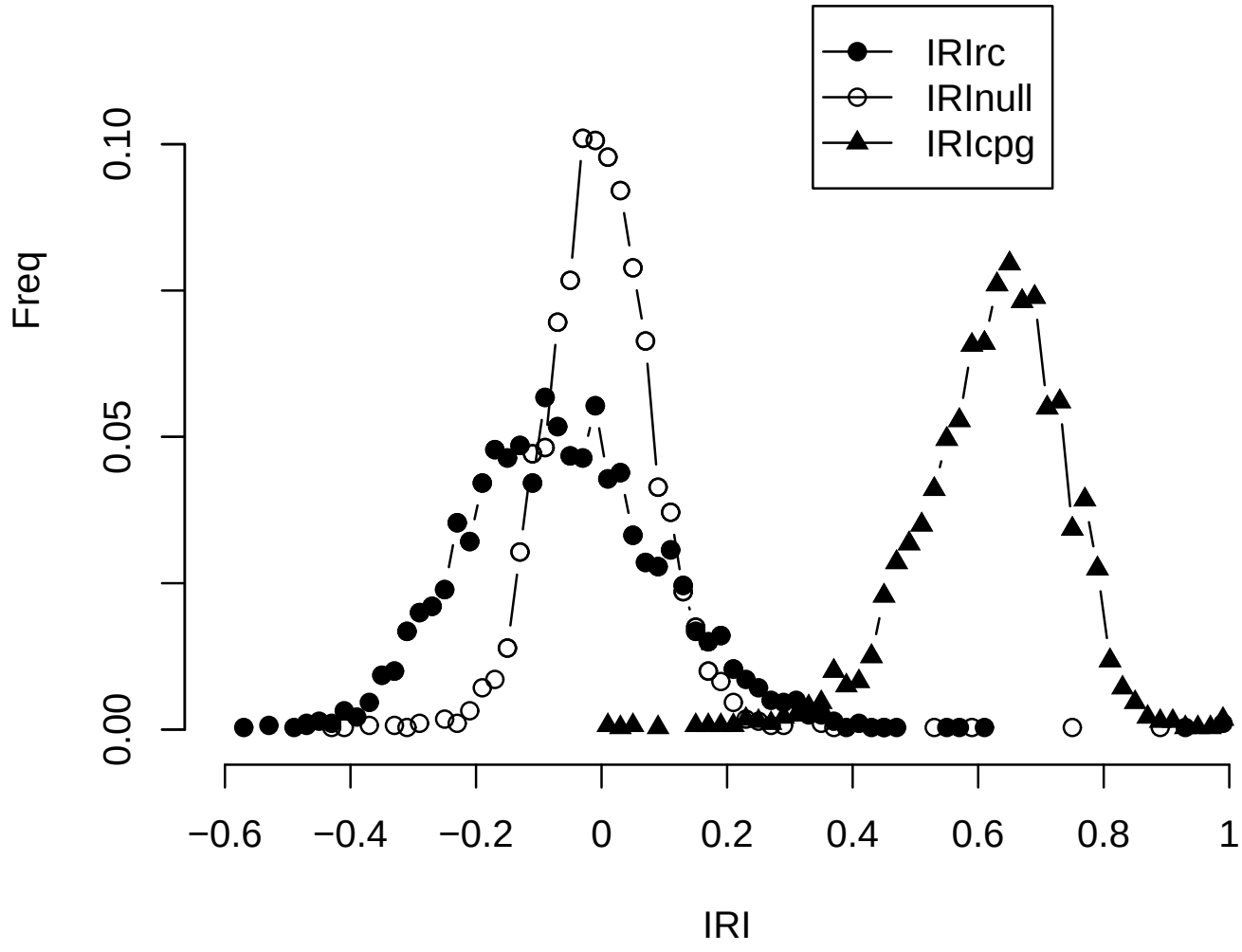
- Alignment of Human, Chimp and Rhesus Macaque genomes
- Rates have been estimated using a maximum likelihood algorithm
- Sliding window analysis, 1 Mbp length
- For each window we have calculated the STIs, IRI_{RC} and IRI_{CpG} in the human lineage

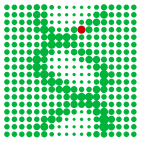


STI Human



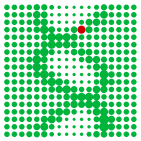
IRI Human


IRIrc
IRInull
IRIcpg
Freq
0.00
0.05
0.10
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1
IRI



Summary

- Commonly used evolutionary models assume equilibrium and reversibility
- We have introduced indices to test for equilibrium (STI) and reversibility(IRI) on each single branch of a given phylogeny
- Analysis in Drosophila and Human show clear violation of the equilibrium/reversibility.
- Further work has to be done to asses how this violations affect specific bioinformatic algorithms.



It’s Evolution Baby. . .

Thank you!