Context dependent evolution and noninformative biological sequences

Marijus Radavičius and Tomas Rekašius

Institute of Mathematics and informatics,
Vilnius Gediminas Technical University, LITHUANIA

1. Introduction:
nucleotide sequences, problems

2. Models:
independent and context-dependent evolution

3. Noninformative nucleotide sequences:
problem statement,
assumptions of a simple evolution model,
definition of "genetic noise"

4. Computer simulations:
long-range dependence,
comparison with real nucleotide sequences

5. Statistical analysis:
data, loglinear models,
Markov property and reversibility

1. Introduction

1.1. Nucleotide sequences

4 nucleotides: A,C,G,T
Nucleotide sequence ...GCAATACGCCTA...

Coding and **non-coding** regions.
Gene is a protein coding nucleotide sequence, and DNA sequences located between genes are non-coding genome sequences.

Evolution: mutation, insertion, deletion.

Properties of nucleotide sequences:
Bacterial genomes: total $\sim (0.5 - 10) \cdot 10^6$ nucleotides, genes $\sim 10^2 - 10^3$.
Human genome: total $3.12 \cdot 10^7$ nucleotides, genes ~ 30000,
non-coding part $\approx 97\%$.

Correlations: long-range dependence in nucleotide sequences indicates a complexity of a system; **Li et al.** (1992), **Peng et al.** (1992), **Buldyrev et al.** (1995), **Karlin et al.** (1993).

1.2. Problems

- **Practical:** prediction and identification of biological function(s) of genes or groups of genes, identification of functionally related genes (regions),

(re)construction of phylogenetic trees...

- **Decoding:** Searching for informative and biologically important regions, extraction of genetic information (mining)

- **Coding:** Modeling and generation of nucleotide sequences,
simulation of DNA evolution

- We are interested in a **opposite problem:** what are **noninformative nucleotide sequences**?

2. Models

2.1. Notation

Let $X(t), t \in T$, be a (discrete time) finite homogeneous Markov chain,

$$X(t) = \{x_l(t) \in \mathcal{A},\ l = 1, \ldots, n,\ t \in T\},$$

Here $T = \{0, 1, \ldots\}$, $\mathcal{A}$ is a finite set;
for DNA sequence $\mathcal{A} = \{A, C, G, T\}$.

Two directions of evolution:

Evolution in time

$$X(t) \longrightarrow X(t+1)$$

In the **stationary** case distribution of $X(t)$ is independent of t.

Thus, it defines probability distribution of a random sequence X on the set of sequences $\mathcal{A}^n$ and we can consider its

"Evolution in space"

$$x_l \longrightarrow x_{l+1}$$

2.2. Independent evolution

The first stochastic models of DNA evolution assumed that the nucleotide along the DNA sequence evolved independently of one another according to the same rule.

Jukes and Cantor (1969), **Kimura** (1980), **HKY model** (Hasegawa et al. (1985))

Consequently, X is a sequence of **i.i.d. variables** i.e. independent and identically distributed nucleotides.
It is not realistic: results of some statistical analysis are presented below.

A natural **generalization:** models with independent **codons** (for coding sequences).

Muse and Gaut (1994), **Goldman and Yang** (1994), **Pedersen et al.** (1998), **Schat and Lange** (2002).

2.3. Context-dependent evolution

In context-dependent evolution model it is assumed that mutations in each site (nucleotide or codon) depend on its nearest neighbours. Usually *continuous time* Markov chains are considered.

Time reversibility implies the Markov property. **Arndt et al. (2003)**

The Markov property in space of the evolution is supposed. **Hwang and Green (2004)** consider a general nonreversible context-dependent nucleotide model, **Siepel and Haussler (2004)**, **Christensen at al. (2004)** context dependent codon model.

Jensen and Petersen (2000,2001) and **Jensen (2005)** a discussion of the relation between time reversibility and the Markov property of the stationary measure of context-dependent evolution is given.

3. Noninformative nucleotide sequences

Problem:

- What does it mean "noninformative nucleotide sequences"?
How to define "genetic noise", i.e. sequence which has no genetically important information?

Direct application:

- Informativeness of a segment in DNA is measured as its distance to noninformative one usually obtained as a random permutation of the initial segment.

Assumptions:

1. **Non-coding regions** of DNA has not direct impact on survival of biological species and thus is not (so) genetically important.

2. Evolution of non-coding regions has **simple structure** and are controlled by local factors.
For instance, here we ignore insertions and deletions and assume that probability of mutation in any site depends exclusively on its nearest neighbours.

3. The **stationary distribution** of non-coding sequence evolution can be treated as "noninformative", i.e. as **"genetic noise"**.

Definition:

Let the evolution $X(t), t \in T$, of nucleotide sequences $x \in \mathcal{A}^n$ in time be a (discrete time) homogeneous Markov chain with a given transition probabilities Π of a simple structure. If there exists its stationary distribution p on $\mathcal{A}^n$, a random sequence X with the distribution p is called **noninformative or genetic noise**.

Assume for simplicity that the site state set $\mathcal{A} = \{0, 1\}$ and consider the Glauber dynamics in time of a random sequence

$$\{X(t), t \in T\} \quad (X(t) \in \mathcal{A}^n)$$

of the length n . Suppose that this dynamics is Markov and homogeneous in both time and space but in each site depends on its nearest neighbours.

Namely,
1. a nucleotide from sequence is selected with probability $1/n$;
2. the selected nucleotide mutates with probability

$$\pi_{uzv} := \mathbf{P}\{x_l(t+1) = \bar{z} | x_{[l-1,l+1]}(t) = uzv\} \tag{1}$$

$$l = 2, \ldots, n-1, \quad u, z, v \in \mathcal{A}, \quad t \in T.$$

Here $\quad \bar{z} = 1 - z$,

$$x_{[l-1,l+1]} = x_{l-1}|x_l|x_{l+1}$$

(we omit the argument t).

Let X denote the "noise" obtained by this evolution, i.e. X is a random sequence of 0's and 1's with the stationary (invariant) distribution of $\{X(t), t \in T\}$. It is completely determined by 8 scalar parameters $\pi := \{\pi_{uzv}\}$.

- What can we say about the properties of the genetic noise X ?

Proposition 1. *If X is homogeneous Markov chain (in space) of order $k < n/4$ then*
(a) $k = 1$, X is reversible and depends on 2 parameters,
(b) the probabilities π are symmetric, $\pi_{uzv} = \pi_{vzu}$ and only the ratios π_{u0v}/π_{u1v} are identifiable.

Proposition 2. *There exists π such that X is not a nonhomogeneous Markov chain (of order 1).*

The PROOF's are rather straightforward and are based on
Hamersley-Clifford theorem for finite random fields.

4. Computer simulations

Let $\mathcal{A} = \{0, 1\}$. The sequence

$$X(t) = \{x_l(t), l = 1, \ldots, n\}, \ t \in T,$$

evolves under the context-dependent mutation model. Probability of nucleotide mutation in the sequence depends on two neighbouring nucleotides (the same Glauber dynamic model).

Several different sets of transition probabilities are considered,

for example:
symmetric where $\pi_{uzv} \equiv \pi_{vzu}$,
non-symmetric where $\pi_{uzv} \not\equiv \pi_{vzu}$.

Simulation of the sequence evolution starts from a random binary sequence and 10^7 **iterations** (mutations) are performed. It is assumed that sequence obtained has (approximately) stationary distribution.

Autocorrelation function

Definition.

Let $X(t)$ be a stationary process and there exists a real number $H \in (\frac{1}{2}, 1)$ and a constant $c_p > 0$ such that autocorrelation function

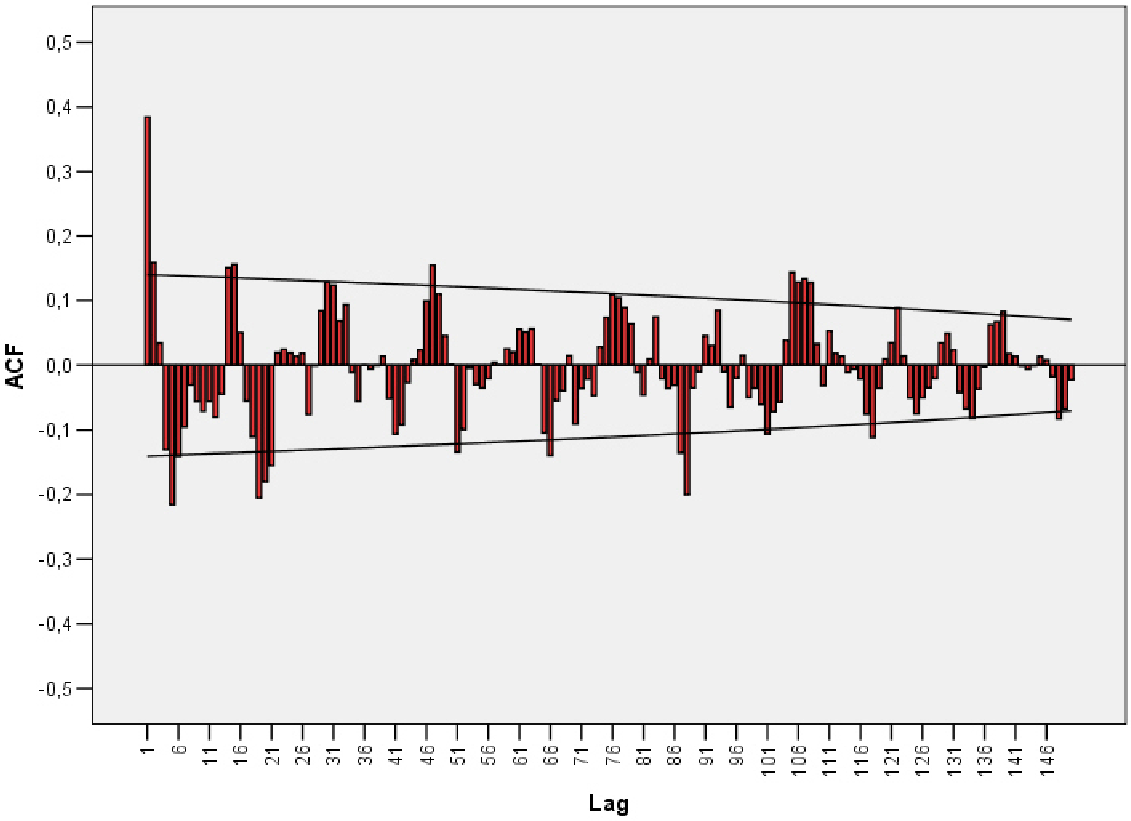
$$\rho(k) \sim c_p |k|^{2H-2}, \quad k \to \infty.$$

Then $X(t)$ is called a stationary process with **long-range dependence**.

The exponent H is called **Hurst parameter**.

For $H = 1/2$ the observations are uncorrelated ($c_p = 0$), and for $H \in (0, \frac{1}{2})$ the process has short-range dependence.

Long memory is characterized by a slow decay of the correlations proportional to k^{2H-2}. A plot of the sequence autocorrelation function should therefore exhibit this decay.



Autocorrelation function of simulated binary nucleotide sequence of length $n = 200$, non-symmetric transition probabilities.

Estimating of parameter H, R/S analysis

Binary nucleotide sequence $x_l, l = 1, \ldots, n$ is subdivided into m non-overlapping blocks. We compute the *rescaled adjusted range* $R(t_i, d)/S(t_i, d)$ for a number of values d where t_i are the starting points of the blocks.

$$R(t_i, d) = max\{0, W(t_i, 1), \ldots, W(t_i, d)\} -$$

$$min\{0, W(t_i, 1), \ldots, W(t_i, d)\},$$

where

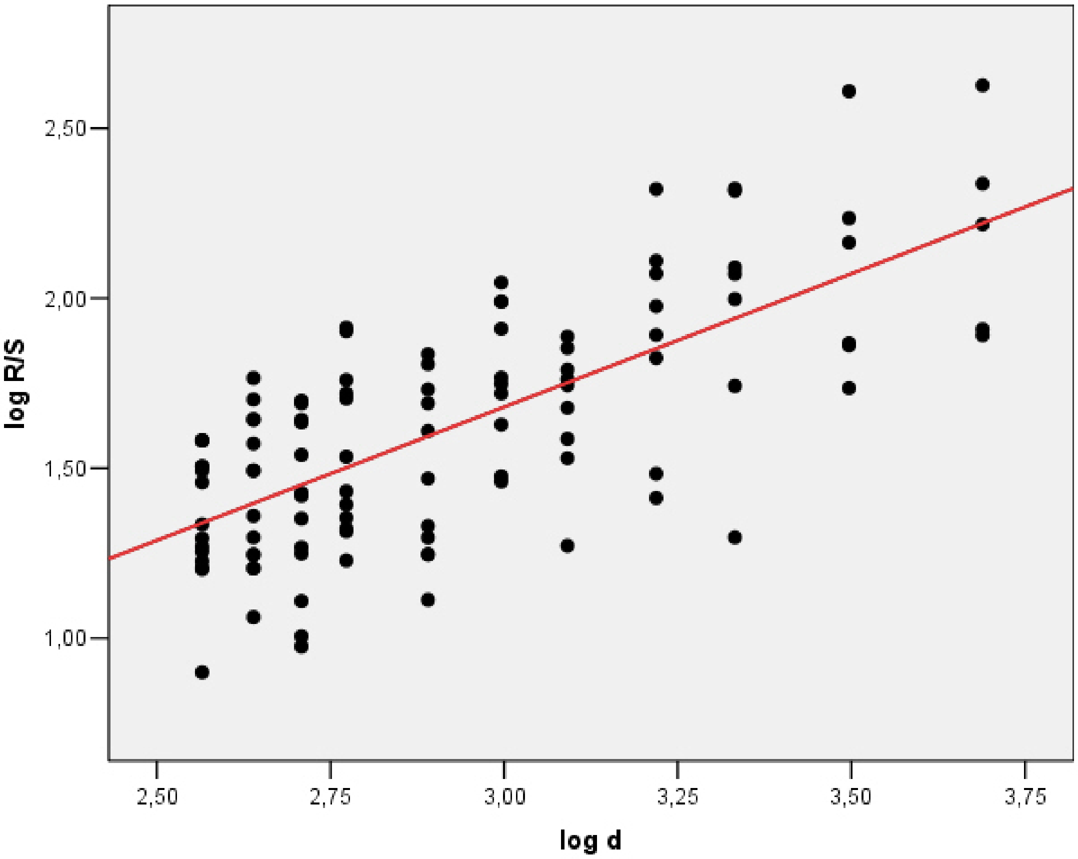
$$W(t_i, k) = \sum_{j=1}^{k} x_{t_i+j-1} - \frac{k}{d} \sum_{j=1}^{d} x_{t_i+j-1},$$

$$k = 1, \ldots, d.$$

$S^2(t_i, d)$ is a sample variance of $x_{t_i}, \ldots, x_{t_i+d-1}$.

For each values of m and d we obtain a number of R/S samples and plot $\log(R/S)$ vs. $\log d$.

A least squares line is fitted to the points of the R/S plot. The slope of the regression line for these R/S samples is an estimate for the Hurst parameter H.

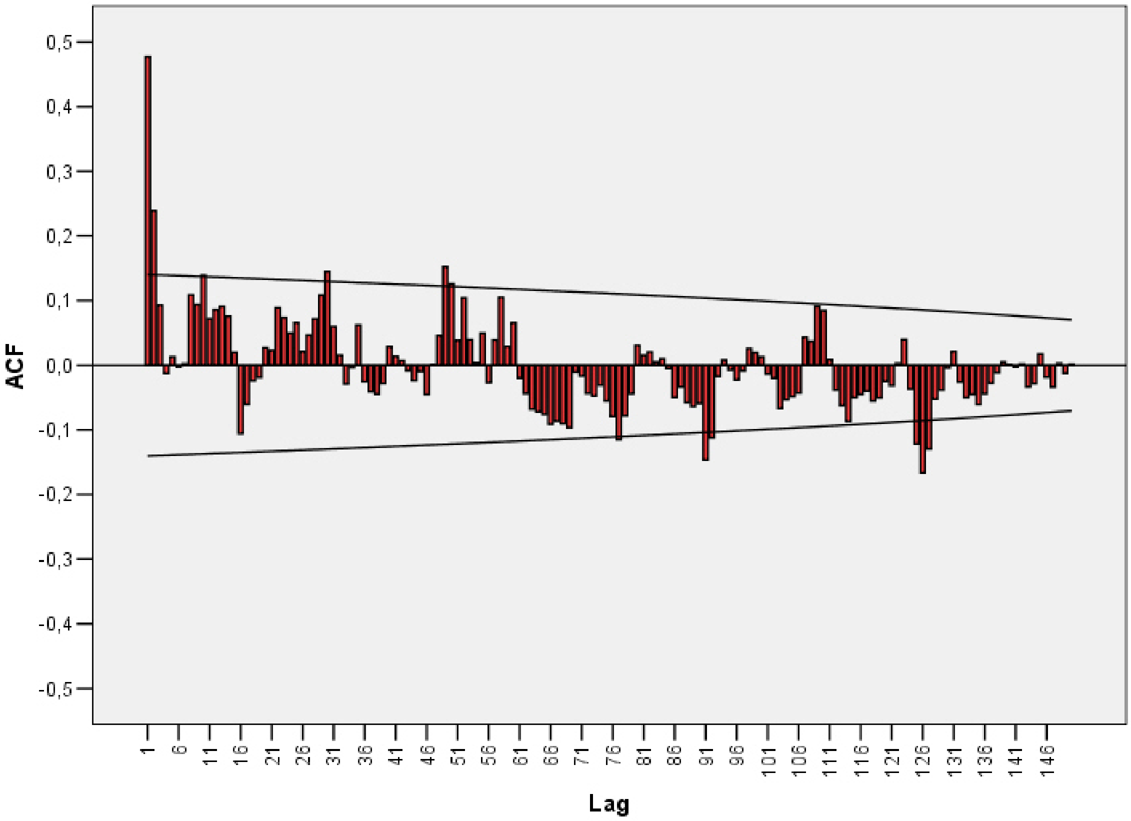


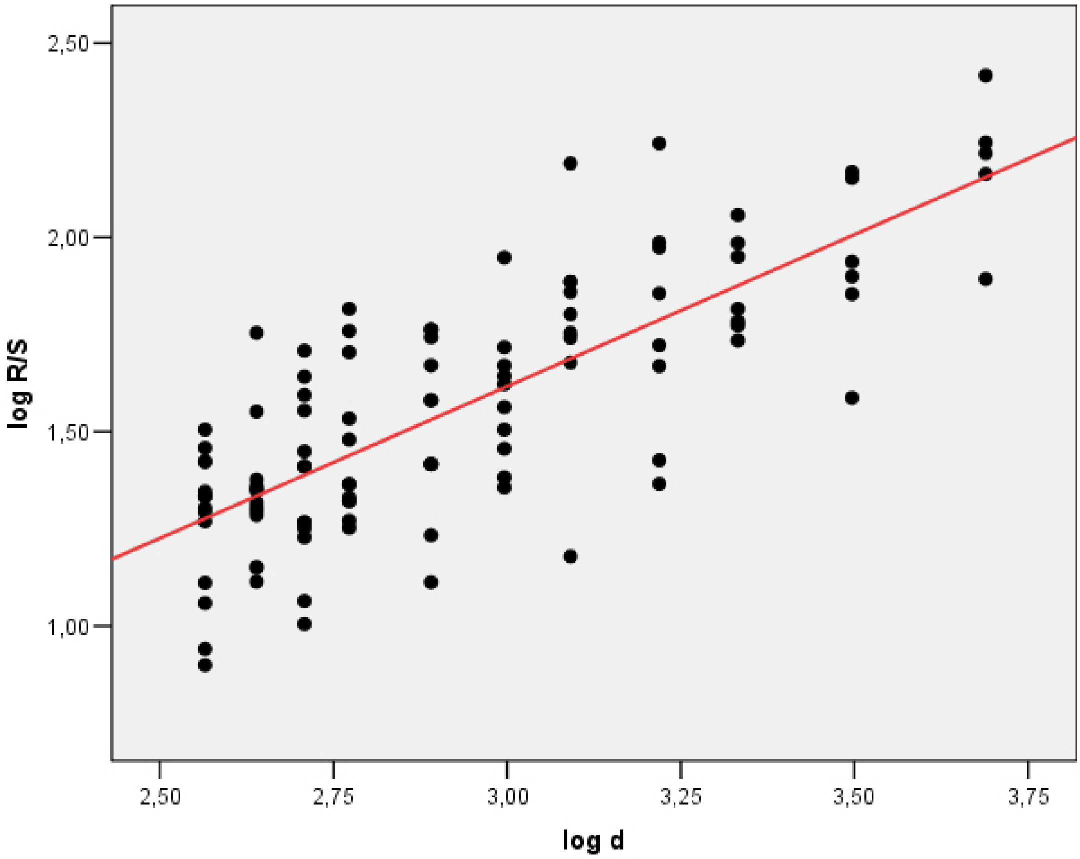
R/S plot for simulated binary nucleotide sequence of length $n = 200$,
non-symmetric transition probabilities.
The Hurst parameter estimate $\widehat{H} = 0.785$

Real non-coding sequence

Below we present autocorrelation function and R/S plots of bacteria *Escherichia coli* non-coding sequence of length $n \approx 200$.
Nucleotide recoding rule:

$\{C, G\} \to \{1\}$, $\{A, T\} \to \{0\}$.
Relative frequency of nucleotides C+G is 50.79%.
Note, that in the generated sequences $\mathbf{P}(0) = \mathbf{P}(1) = 1/2$.





The Hurst parameter estimate for this sequence $\widehat{H} = 0.782$

5. Statistical analysis:

Data: bacterial genomes (GenBank)

We use full genome sequences and sets of their non-coding regions. The data we deal with is of the following form:

$$\{(y_l, z_l), l = 1, \ldots, N\},$$

where

$$y_t = x_{2l}, \quad z_l = (x_{2l-1}, x_{2l+1}), \quad l = 1, \ldots, N.$$

Assumption:

$\{y_l,\ l = 1, \ldots, N\}$ are conditionally independent given $\{z_l,\ l = 1, \ldots, N\}$, and impact of z's on y's is homogeneous (does not depend on sites l).

This assumption is valid, in particular, if X is a homogeneous Markov chain.

- Thus, standard assumptions of regression models hold and we can apply standard statistical software to perform statistical analysis.

We use SAS (proc CATMOD) to fit loglinear model to the data.

Let the state space be $\mathcal{A} = \{A, C, G, T\}$.

Saturated logit model with the reference state 'T':

$$\log\left(\frac{\mathbf{P}\{x_{[2l-1,2l-1]} = uzv\}}{\mathbf{P}\{(x_{[2l-1,2l-1]} = u'T'v\}}\right) = \lambda_z + \lambda^L_{uz} + \lambda^R_{zv} + \lambda^{L\&R}_{uzv},$$

$$u, z, v \in \mathcal{A}, \quad u, z, v \neq' T'.$$

For the **Markov** chains the interaction term $\lambda^{L\&R}$ should be zero.

Markov hypothesis

$$H_0 : \quad \lambda^{L\&R} \equiv 0$$

The **reversibility** to hold the logit model should be symmetric in u and v.

Reversibility hypothesis

$$H_0 : \quad \lambda^L_{uz} \equiv \lambda^R_{zu}$$

We introduced a special **variable 'as'** to indicate the asymmetry. For the symmetric models both the main effect of 'as' and all its interactions should be zero.

We tested this for the full genome of **bacteria** *Escherichia coli* and for its set of non-coding regions.

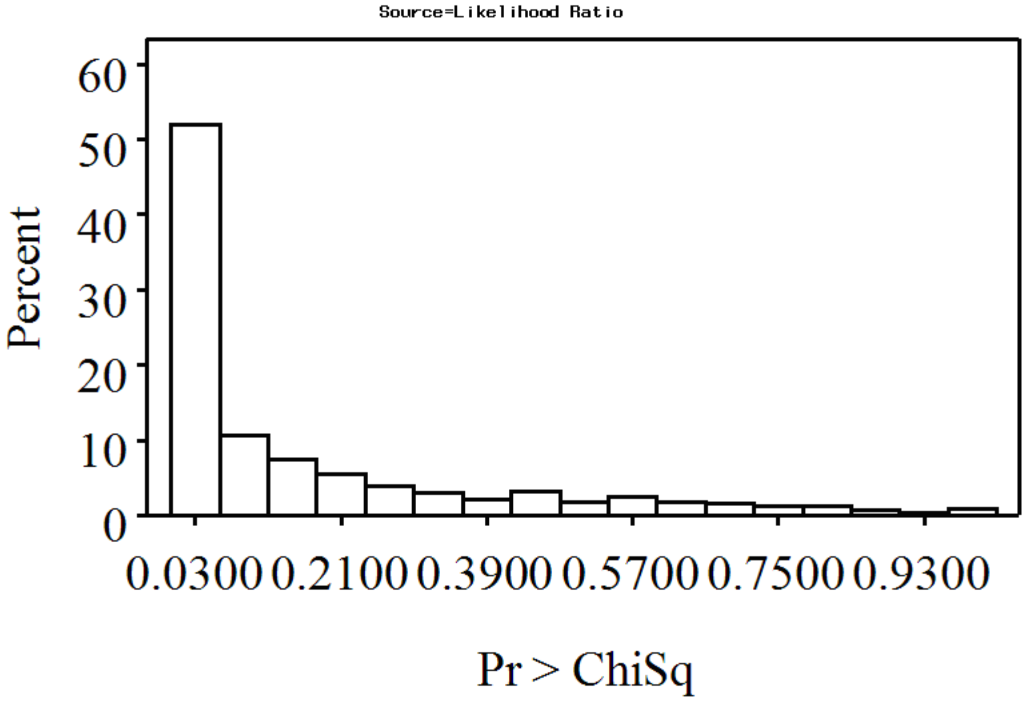
Testing the Markov hypothesis

$$H_0: \quad \lambda^{L\&R} \equiv 0$$

Likelihood Ratio (LR) statistic for full genome:

$DF = 27, \quad LR = 18411.49, \qquad$ p-value $< .00017$

Distribution of p-values of LR statistic for testing the Markov hypothesis for the set of the non-coding regions of the genome.



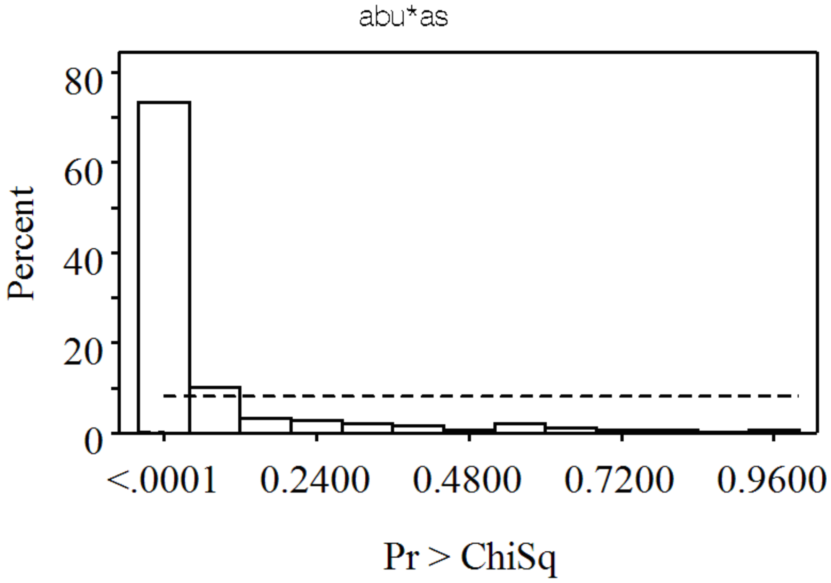
p-values of LR statistic

Testing the reversibility for the full genome

$$H_0 : \quad \lambda^L_{uz} \equiv \lambda^R_{zu}$$

We introduced a special **variable 'as'** to indicate the asymmetry. For the symmetric models both the main effect of 'as' and all its interactions should be zero.

Testing the reversibility:
distribution of p-values for the non-coding regions



Testing for interactions of 'as'