Theoretical approaches for the study of dinucleotide content in genomes

Leonor Palmeira, Laurent Guéguen, Jean R. Lobry

Laboratoire de Biométrie et Biologie Évolutive
UMR CNRS 5558
Université Claude Bernard - Lyon 1

TAG – LAPTH, Annecy – 9.11.2006

Modelling of sequence evolution

Usual assumptions :

- a sequence is a set of independently evolving sites
- substitution rates are constant along a sequence
- substitution rates are constant through time
- ...

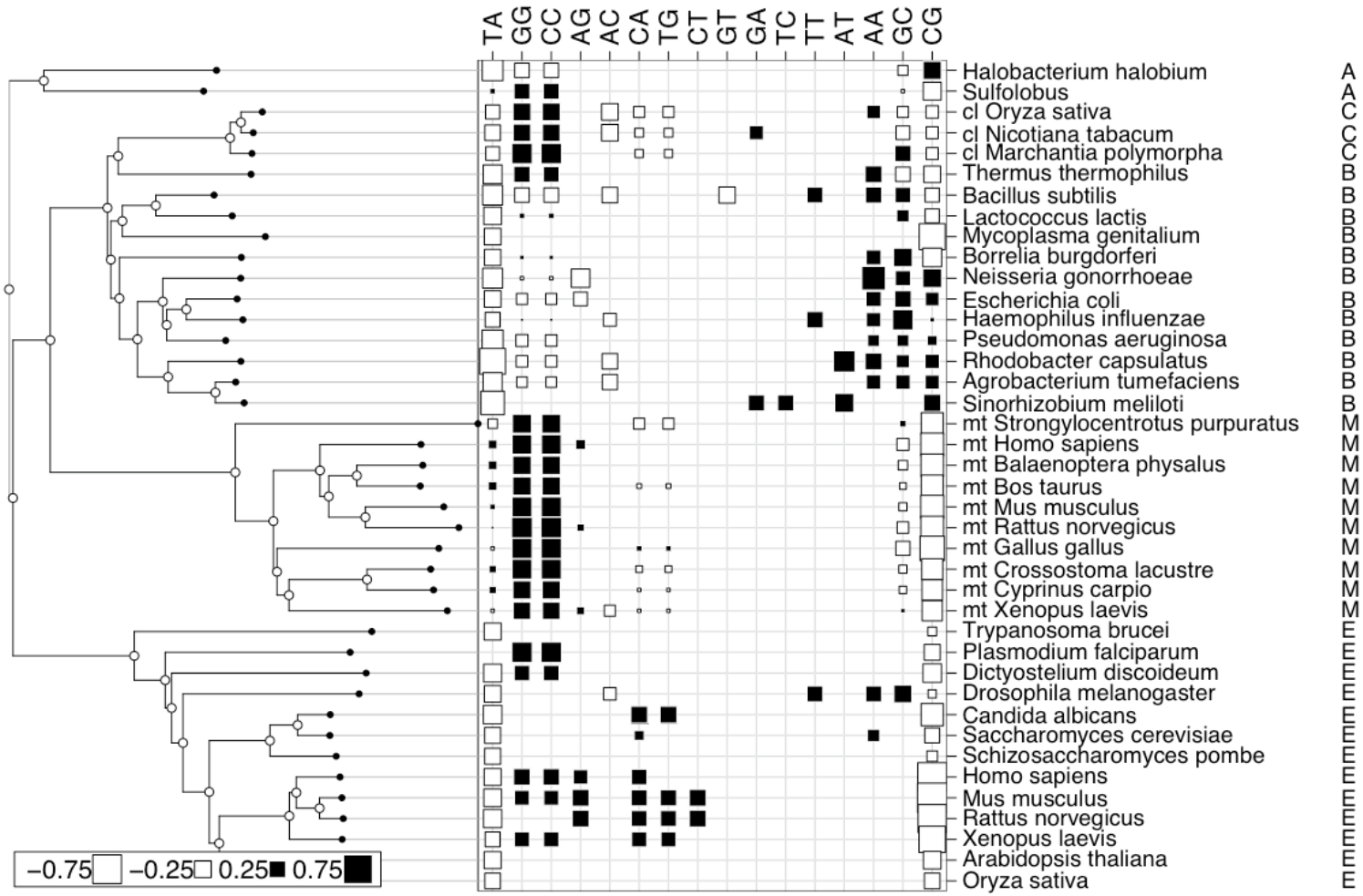
How can we have more realistic models of DNA sequence evolution ?

Sites do not evolve independently

- ▷ biologically unfounded assumption :
  - biochemical evidence (Bird, 1980)
  - statistical evidence (Karlin and Burge, 1995)
- ▷ assumption needed for mathematical purposes

↪ ... a widely used assumption

Sites do not evolve independently



A : Archæa, C : Chloroplasts, B : Bacteria, M : Mitochondria, E :Eukaryota.
Data compiled from Burge *et al.* 1992, Brendel *et al.* 1992, Cardon *et al.* 1994, Karlin *et al.* 1994, 1997.

Measuring dinucleotide composition

Problem #1

Simple statistic estimation of dinucleotide over- and under-representation.

Modelling neighboring-sites dependency

Problem #2

Write a mathematically tractable model to describe neighbor-dependent substitutions in DNA.

Measuring dinucleotide composition : method

Statistics based on the comparison between :

- the observed count $\rho(XY) = \frac{f_{XY}}{f_X f_Y}$ on the studied sequence (Karlin, 1992)
- the estimated count $\rho(XY)$ according to a certain model

– by simulation or by analytical computation –

$$z_{score} = \frac{\rho(XY) - E(\rho(XY))}{\sqrt{Var(\rho(XY))}} \sim \mathcal{N}(0,1)$$

Measuring dinucleotide composition : method

This model preserves :

- the base composition (*i.e* G+C content) of the studied sequence in each permuted sequence.

The calculated statistic allows to answer the question :

- is there a statistically high/low dinucleotide content given the base composition of my sequence ?

Asymptotic results are available (Prum *et al.*, 1995) :

$$\mathrm{E}(\rho_{\mathrm{XY}}) = 1$$

$$\sqrt{\mathrm{V}(\rho_{\mathrm{XY}})} \approx \sqrt{\frac{(1-f_X)(1-f_Y)}{nf_Xf_Y}}$$

How does one deal with codon usage bias ?

- Preferential usage of certain codons for each amino acid.
  - positive correlation between preferred codons and tRNA abundance (*E. coli*, *S. cerevisiae*, *D. melanogaster*).
  - selective pressure for translation efficiency.

$\Rightarrow$ artificial over-representation of dinucleotides contained in preferred codons.

Measuring dinucleotide composition : method

This model preserves :

- the base composition (*i.e* G+C content) of the studied sequence in each permutated sequence.
- the codon usage bias of the studied sequence in each permutated sequence.

The calculated statistic allows to answer the question :

- is there a statistically high/low dinucleotide content given the base composition and the codon usage bias of my sequence ?

Asymptotic results are available (Gautier *et al.*, 1985) :

$$z_{score} = \frac{XY_{3-1} - E(XY_{3-1})}{\sqrt{Var(XY_{3-1})}}$$

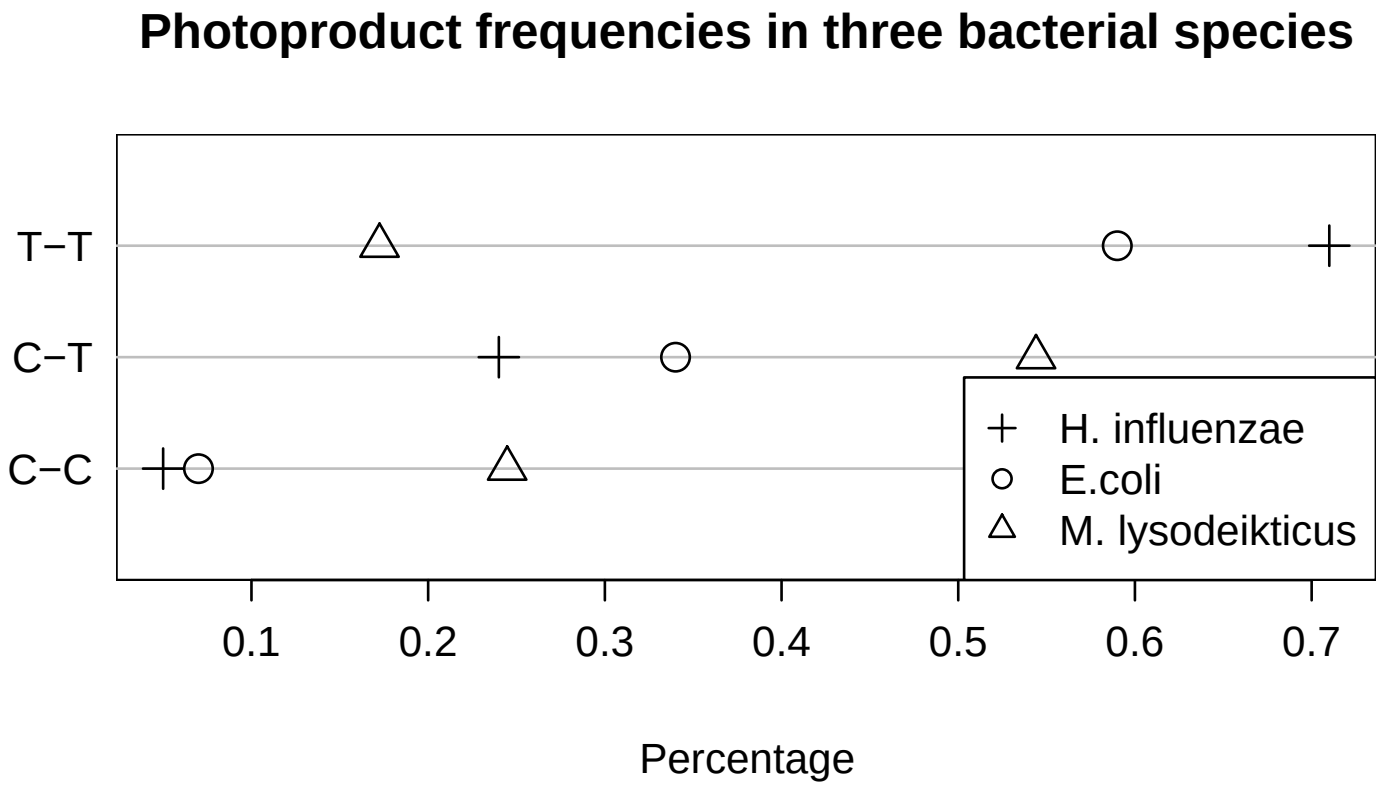
$$E(XY_{3-1}) = \frac{n_1.n_2 - n_3}{n}$$

$$Var(XY_{3-1}) = E(XY_{3-1}) - (E(XY_{3-1}))^2 + \frac{1}{n(n-1)}[(2n_3(n_1 + n_2 - n_1.n_2 - 1) + n_1.n_2(n_1 - 1)(n_2 - 1))]$$

UV impact on genomes

Most frequent photoproducts are formed by covalent links between adjacent pyrimidines (C and T) :

⇒ Blocks **replication** and **transcription** by local DNA distorsion



Data from Setlow, R. B., *Science*, 1966

A controversial study (Singer and Ames, 1970)

- is $G + C$ content a good measure of selective pressure due to UVs ? (Setlow, 1966)
- UV light exposure in the bacterial habitat is very difficult to determine (Bak, *et al.*, 1972)
- Laboratory studies do no support these results (Joux *et al.*, 1999)

Does a high G+C content reduce the number of phototargets ?

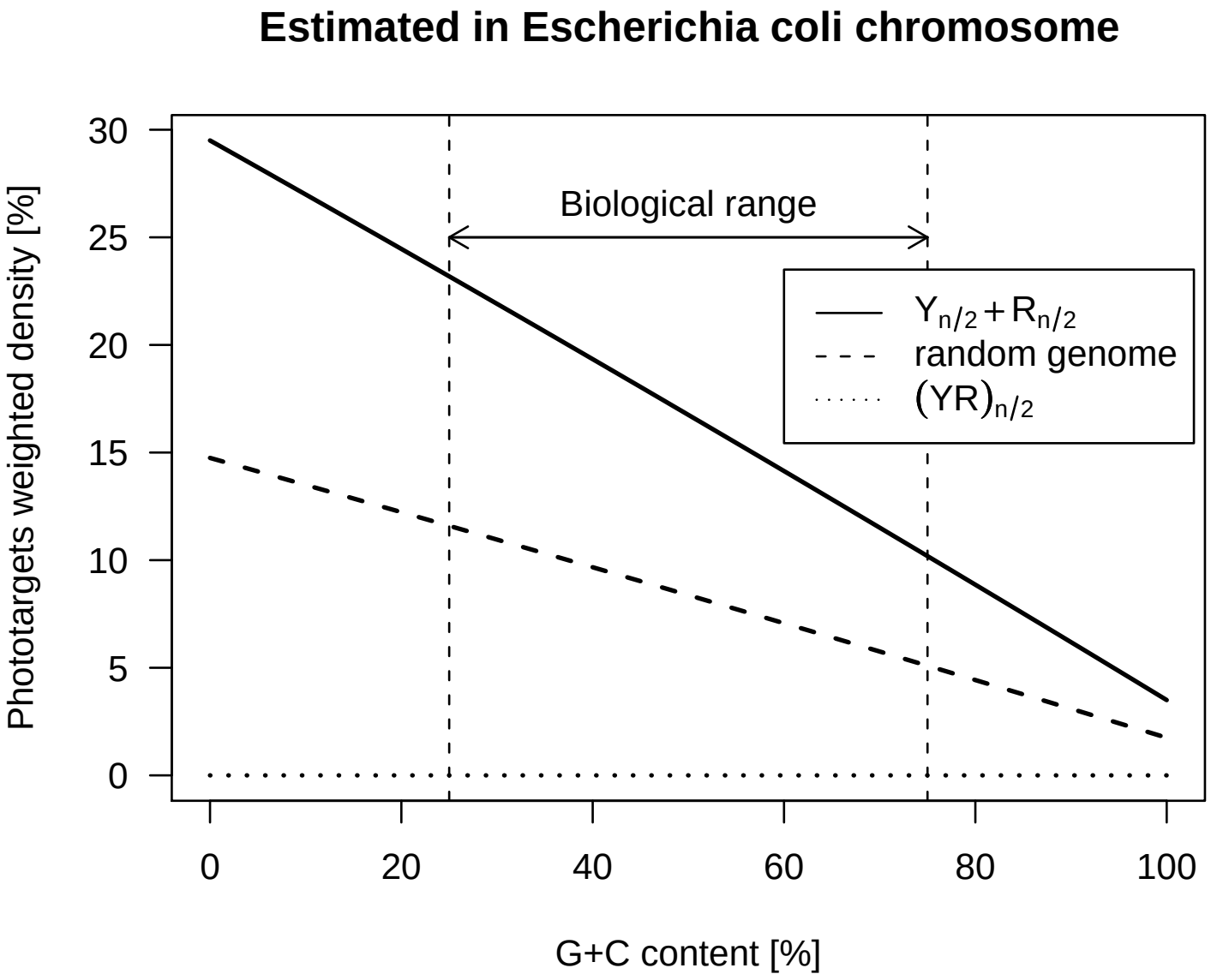
Let us measure phototargets weighted density for a given G+C content :

$$f_c = (G + C)/2 \text{ and } f_t = (1 - (G + C))/2$$

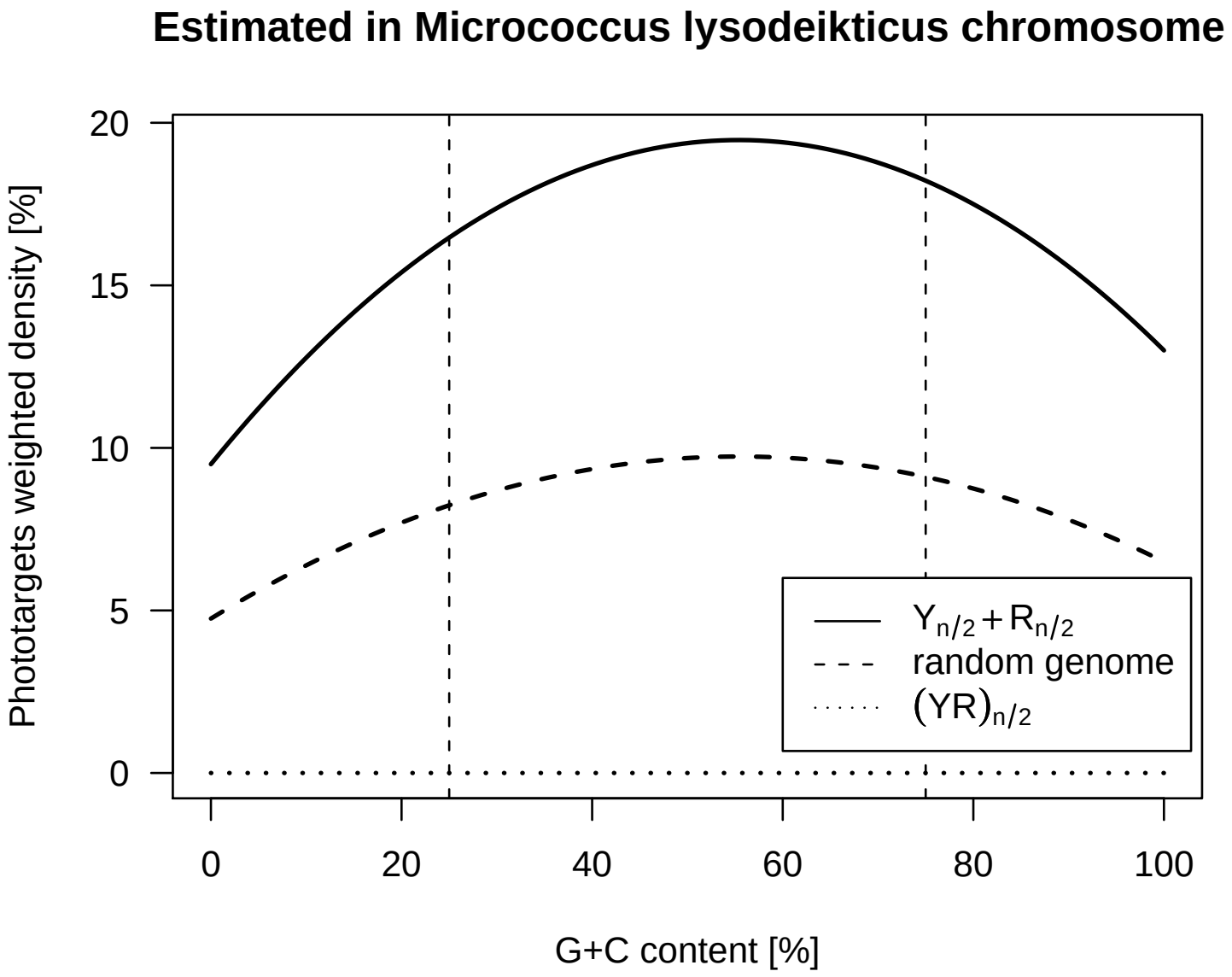
therefore : $f_c + f_t = 1/2$

- interspersed genome YRYRYR... : $(Y + R)_{n/2}$
  $\hookrightarrow$ $N = 0$
- 'random' aggregation of Y and R
  $\hookrightarrow$ $N = s_{tt} f_t^2 + s_{tc}.2(f_t f_c) + s_{cc} f_c^2$
- fully aggregated genome : $Y_{n/2} + R_{n/2}$
  $\hookrightarrow$ $N = 2(s_{tt} f_t^2 + s_{tc}.2(f_t f_c) + s_{cc} f_c^2)$

G+C content : a bad measure for this analysis



G+C content : a bad measure for this analysis



Fully sequenced genomes...

⇒ Complete analysis of the **frequency of pyrimidine dimers** (TT, CT, TC and CC).

- computing resources at the *IN2P3 Computing Center* CC-IN2P3
  - parallelized computing grid
  - Linux platform : 654 bi-processor machines
- analysis and computing done with R's `seqinR` and `ade4` packages.

... a systematic view and a beautiful example

A systematic view

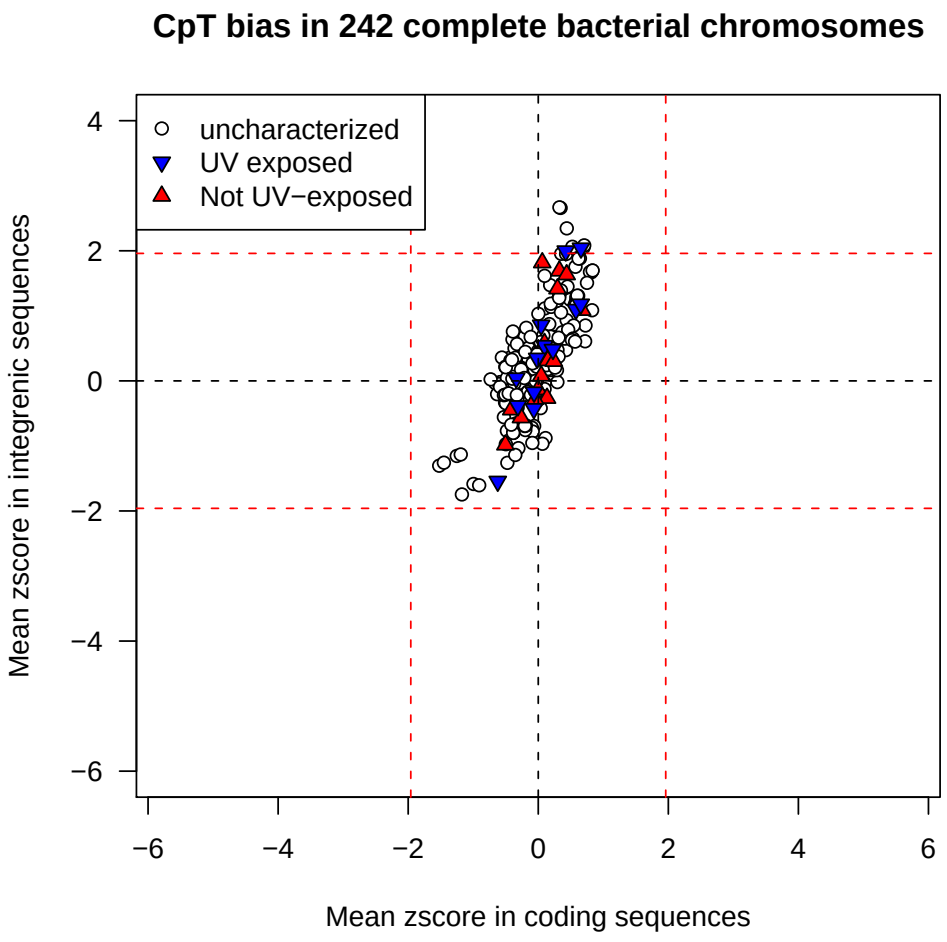
- complete genomes of Bacteria and Archaea

242 completely sequenced Bacteria and Archae chromosomes downloaded from EBI Genome Reviews

A beautiful example

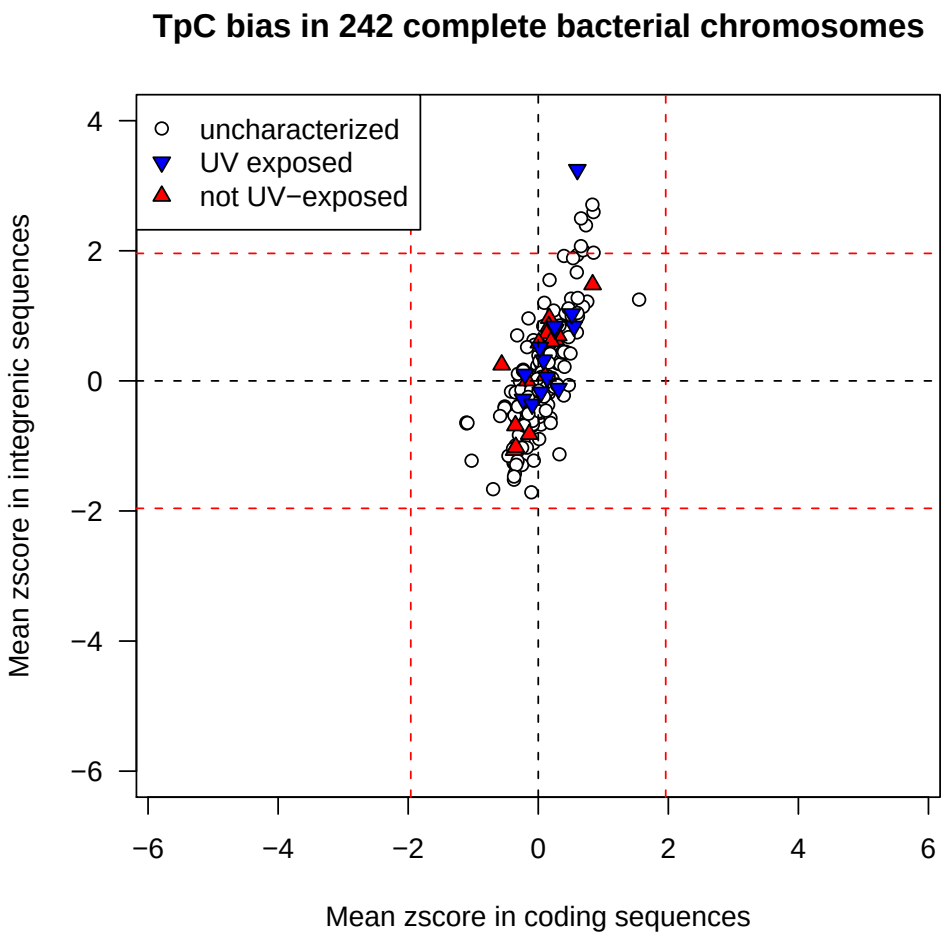
- three fully sequenced strains of *Prochlorococcus marinus*
- adapted to different depths in the water column
- exposed to different UV contents

Coding sequences vs. non-coding sequences



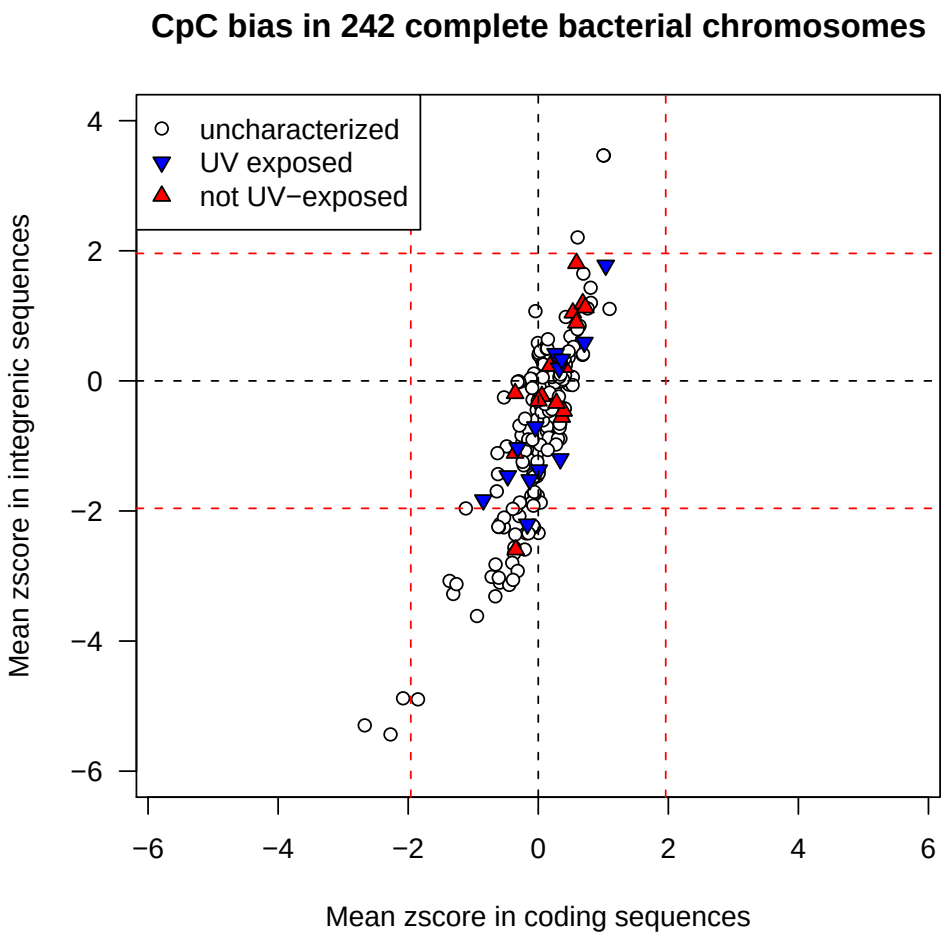
L. Palmeira, L. Guéguen, J. R. Lobry, 2006.

Coding sequences vs. non-coding sequences



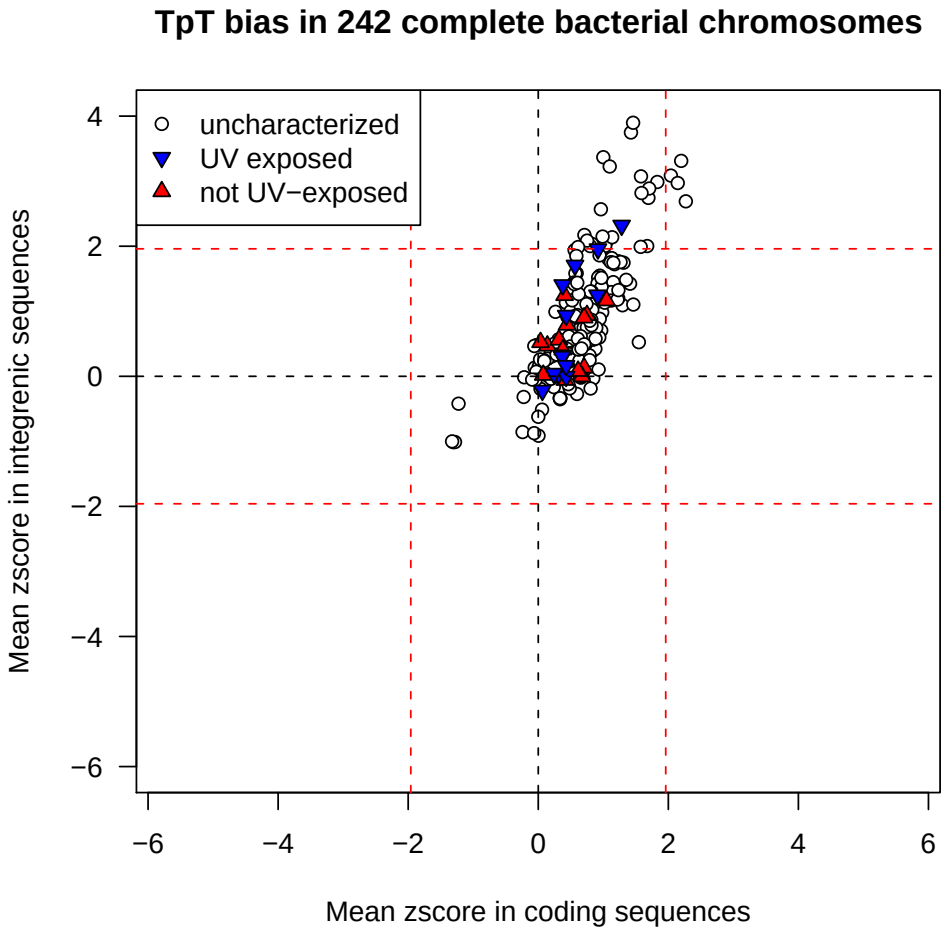
L. Palmeira, L. Guéguen, J. R. Lobry, 2006.

Coding sequences vs. non-coding sequences



L. Palmeira, L. Guéguen, J. R. Lobry, 2006.

Coding sequences vs. non-coding sequences



L. Palmeira, L. Guéguen, J. R. Lobry, 2006.

⇒ **No systematic link** between pyrimidine dimers frequencies and UV exposure in Prokaryotes.

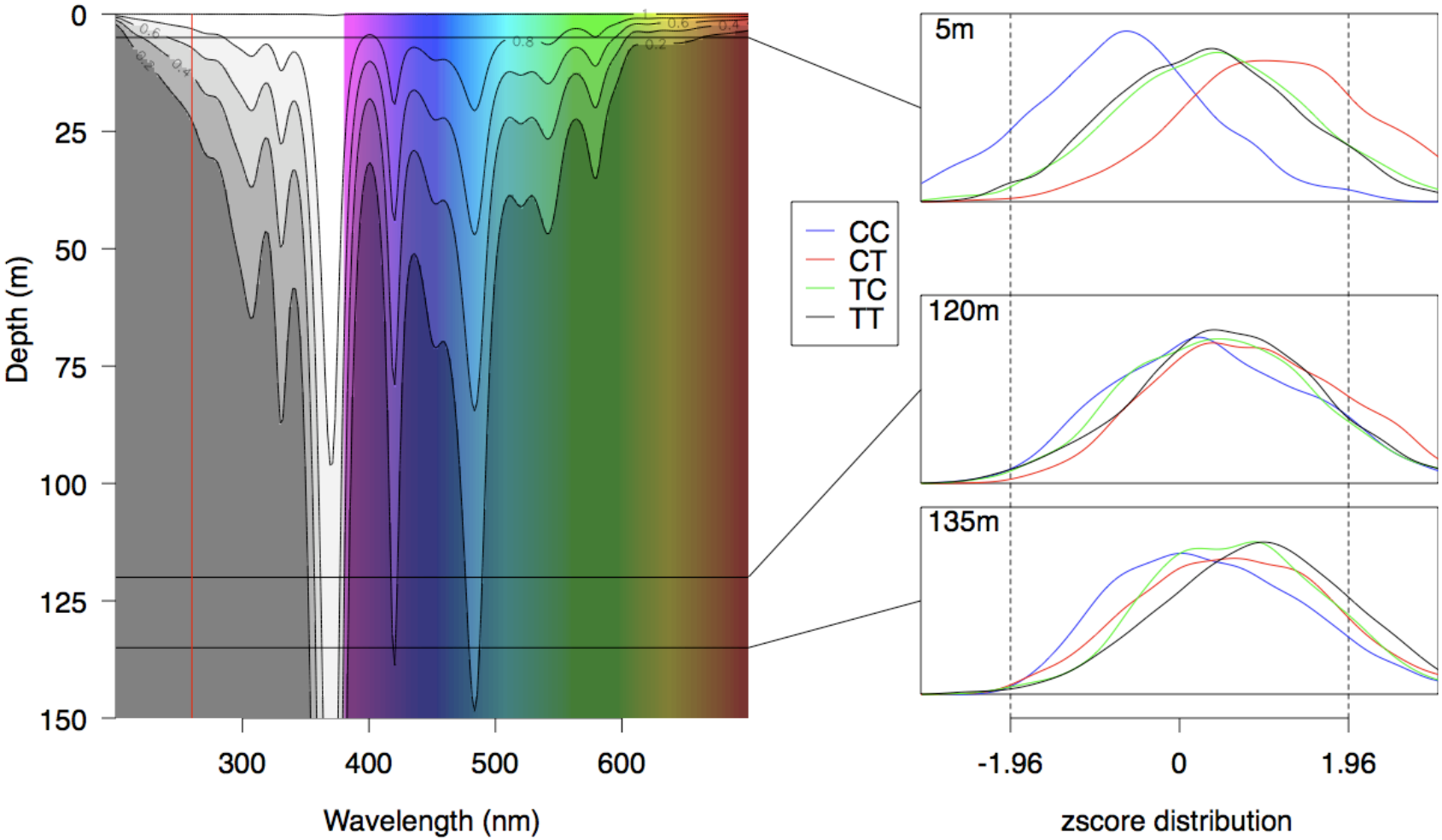
A perfect model organism : *P. marinus*

- one of the most abundant micro-organisms in oceans
- involved in a great part of the primary production
- stratified habitat in the water column
- compact genomes → 11 eleven ecotypes being sequenced (5 completed)

Photo credit : Genoscope - Centre National de Séquençage.

No effect of UV exposure

L. Palmeira, L. Guéguen, J. R. Lobry, to be submitted.
Dinucleotide composition in three light-adapted P. marinus ecotypes



References

This work has led to :

L. Palmeira, L. Guéguen, J. R. Lobry
UV-targeted dinucleotides are not depleted in light-exposed Prokaryotic genomes.
Molecular Biology and Evolution, 2006, 23(11) :2214-2219.

References

This work has led to :

L. Palmeira, L. Guéguen, J. R. Lobry
UV-targeted dinucleotides are not depleted in light-exposed Prokaryotic genomes.
Molecular Biology and Evolution, 2006, 23(11) :2214-2219.

L. Palmeira, L. Guéguen, J. R. Lobry
Genomes under the influence : impact of environmental UV
To be submitted.

References

This work has led to :

L. Palmeira, L. Guéguen, J. R. Lobry
UV-targeted dinucleotides are not depleted in light-exposed Prokaryotic genomes.
Molecular Biology and Evolution, 2006, 23(11) :2214-2219.

L. Palmeira, L. Guéguen, J. R. Lobry
Genomes under the influence : impact of environmental UV
To be submitted.

D. Charif, J. R. Lobry, A. Necșulea and L. Palmeira
SeqinR 1.0-6 : a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis.
http ://cran.r-project.org/

A need for realistic models of nucleotide substitution

Some applications of evolutionary models :

- estimate substitution rates (evolutionary speed)
- estimate the evolutionary distance between two sequences
- constructing a phylogenetic tree from n sequences *(distance methods, maximum likelihood)*

When sites do not evolve independently

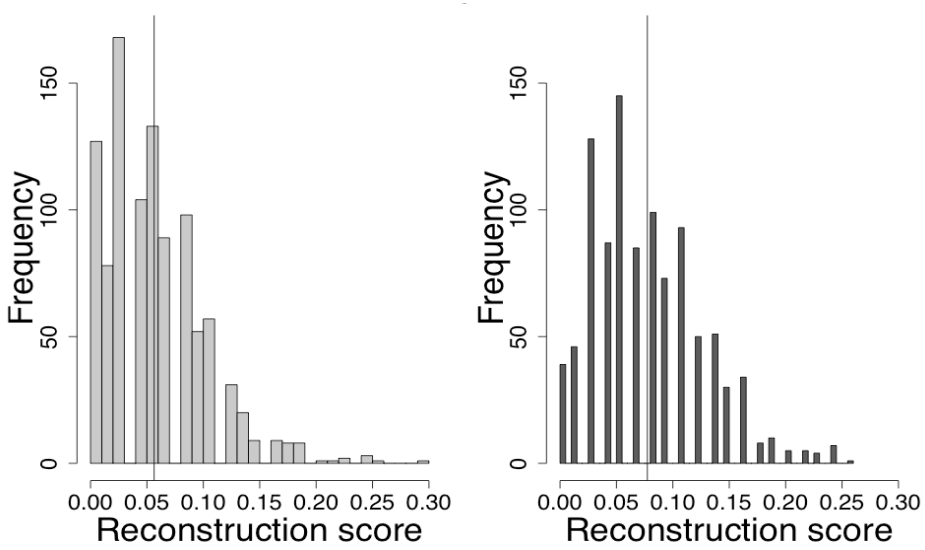
Consequences

- bias in estimating evolutionary distance between two sequences (von Haeseler and Schöniger, 1998)
- bias in phylogenetic reconstruction (von Haeseler and Schöniger, 1998 ; Palmeira, Lobry and Guéguen, *in prep.*)

When sites do not evolve independently

Bias in phylogenetic reconstruction *(Palmeira, Lobry and Guéguen, in prep.)*

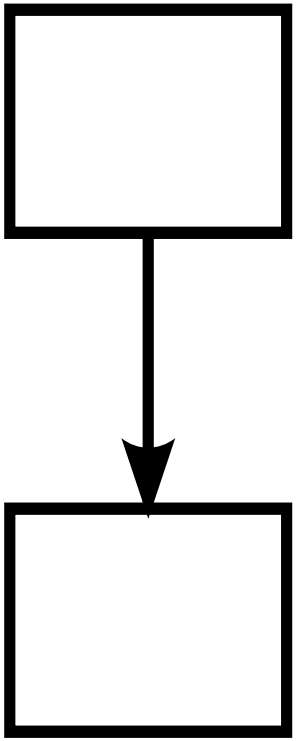
- evolution with either **Kimura** or **Kimura+neighbor-dependent substitution** models
- all phylogenetic reconstructions with Kimura model (*maximum likelihood*)
- tree comparison score (Robinson and Foulds, 1981)



The problem of the dependency cone

Example of the stationary distribution

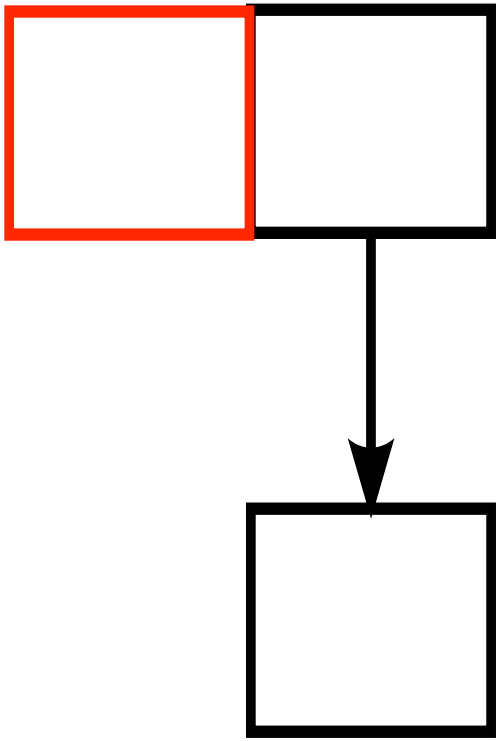
- nucleotide frequencies



The problem of the dependency cone

Example of the stationary distribution

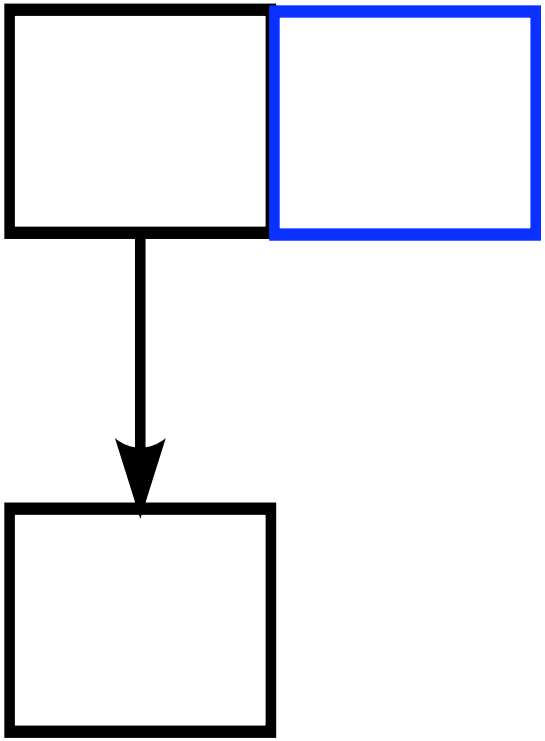
- nucleotide frequencies



The problem of the dependency cone

Example of the stationary distribution

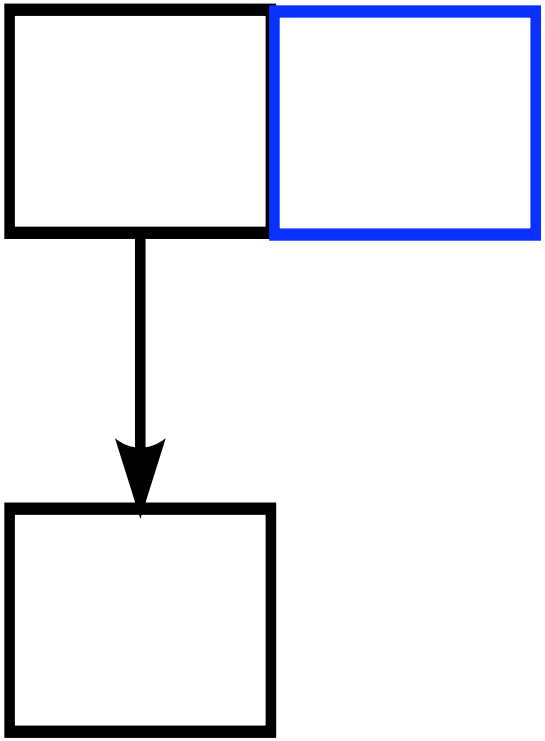
- nucleotide frequencies



The problem of the dependency cone

Example of the stationary distribution

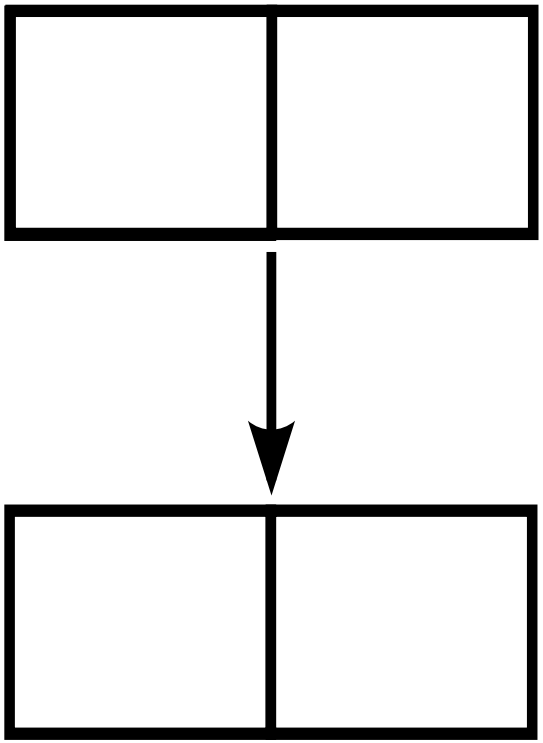
- nucleotide frequencies call for dinucleotide frequencies



The problem of the dependency cone

Example of the stationary distribution

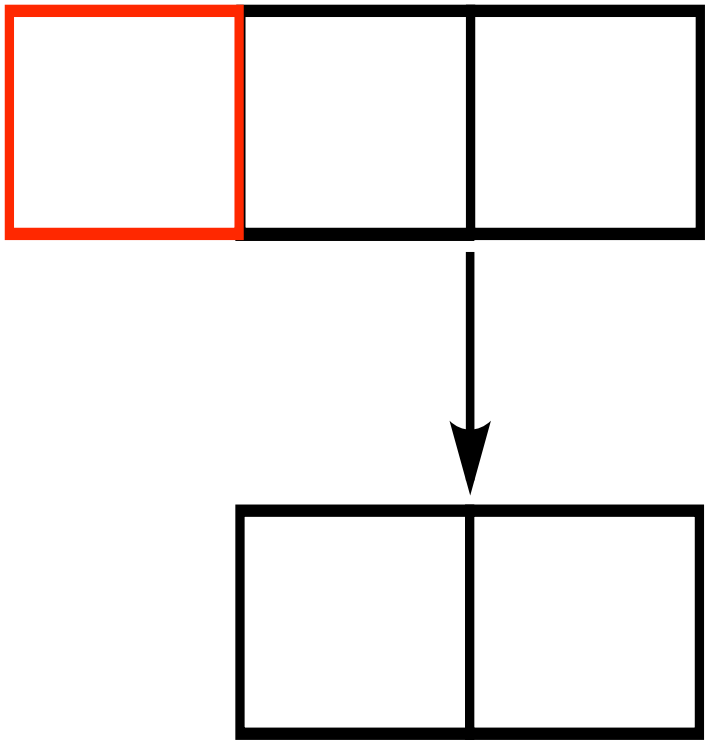
- dinucleotide frequencies



The problem of the dependency cone

Example of the stationary distribution

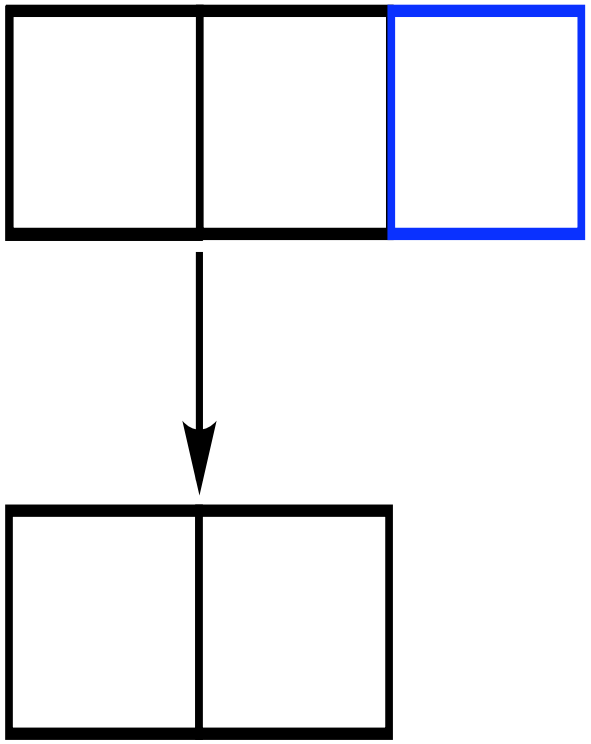
- dinucleotide frequencies



The problem of the dependency cone

Example of the stationary distribution

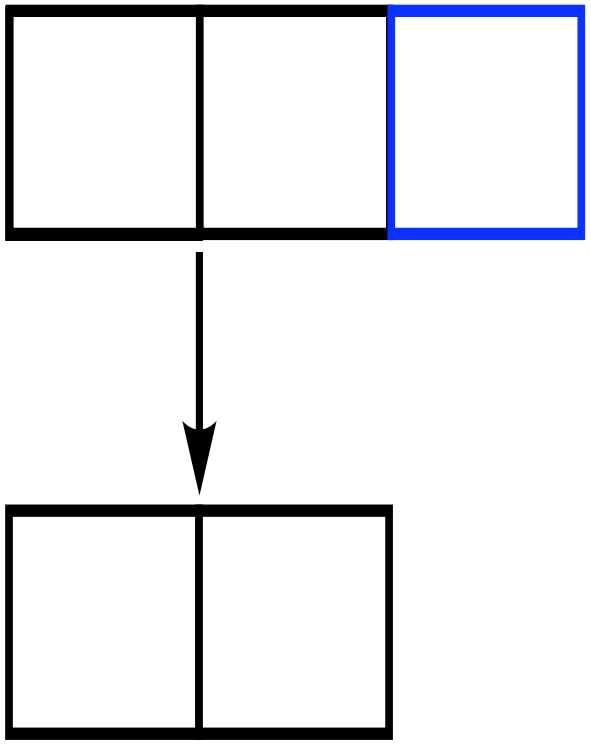
- dinucleotide frequencies



The problem of the dependency cone

Example of the stationary distribution

- dinucleotide frequencies call for trinucleotide frequencies



Modelling neighboring-sites dependencies : a summary

Using the K-cluster approximation

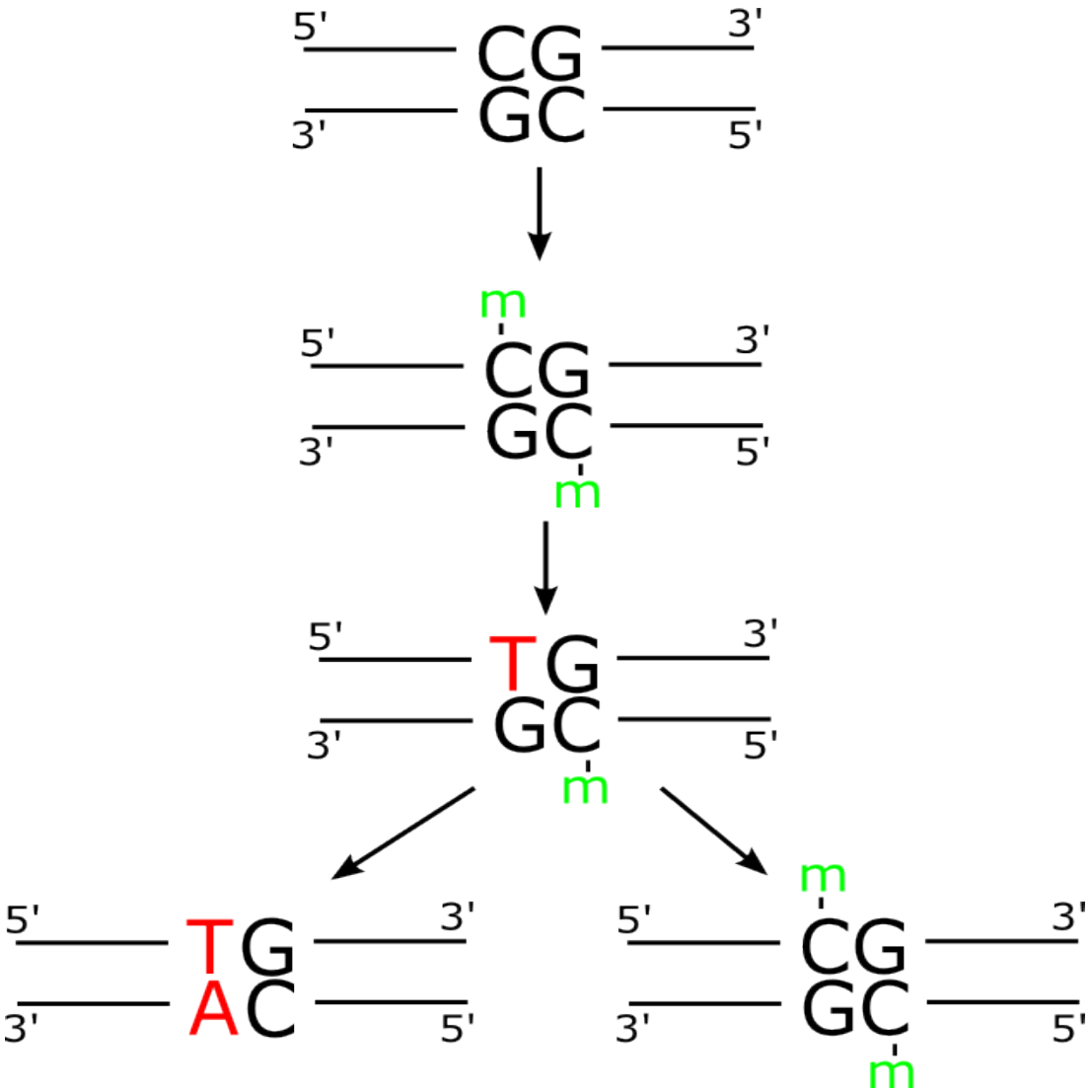
$$f_{xyz} = \frac{f_{xy} f_{yz}}{f_y}$$

- stationary distribution estimation (Arndt, Burge and Hwa, 2002, Lunter and Hein, 2004)
- substitution rates estimation (Arndt, Burge and Hwa, 2002)

$\hookrightarrow$ exact analytical formulas for stationary distribution ?

$\hookrightarrow$ exact analytical formulas for substitution rates ?

CpG methylation-deamination process

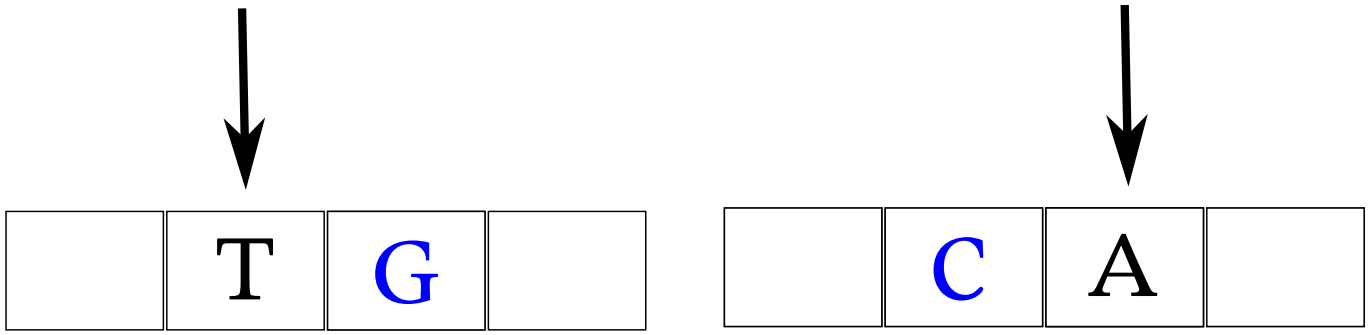


Breaking the dependency cone

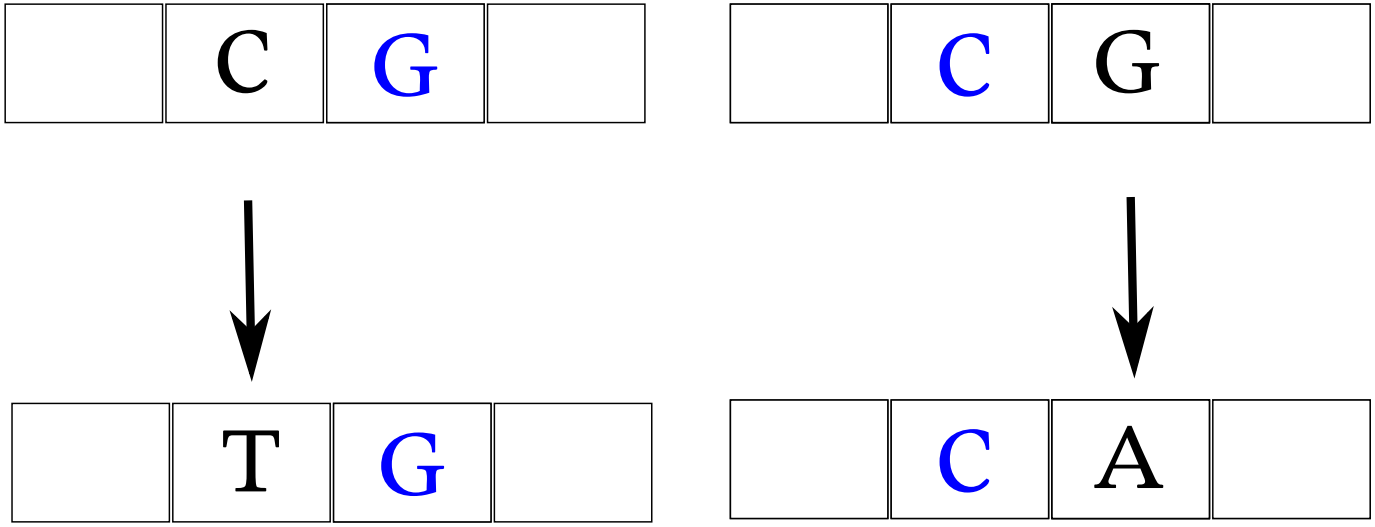
	T		

		A	

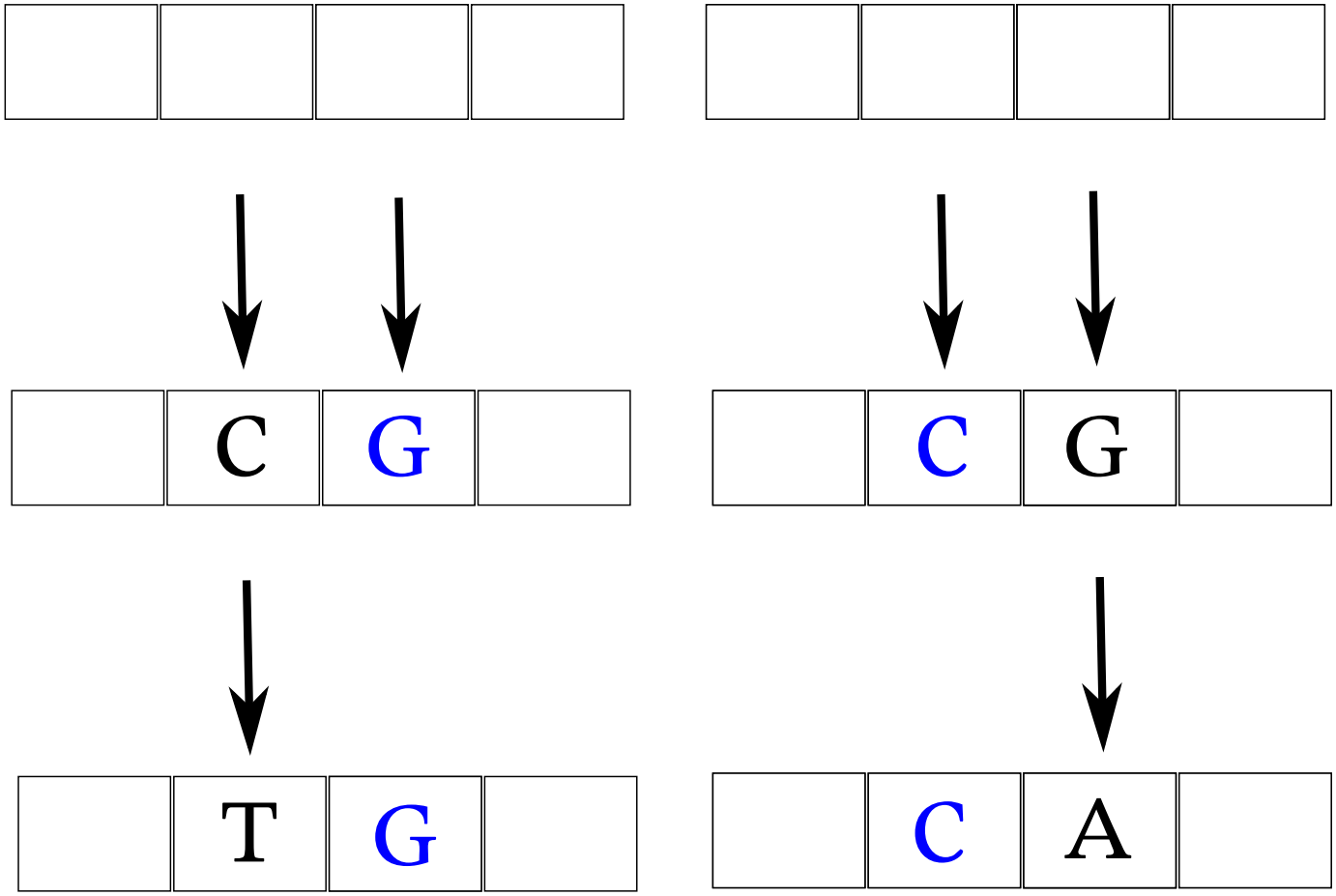
Breaking the dependency cone



Breaking the dependency cone



Breaking the dependency cone



A general solvable model of neighbor-dependent substitutions

(Bérard, Gouéré and Piau, 2005)

Combining :

- a simple nucleotide substitution model of the form :

$$\begin{pmatrix} - & w_T & w_C & v_G \\ w_A & - & v_C & w_G \\ w_A & v_T & - & w_G \\ v_A & w_T & w_C & - \end{pmatrix}$$

(Rzhetsky and Nei, 1995)

v transition rate ; w transversion rate.

- and all dinucleotide substitution process of the form YpR.

$\Rightarrow$ stationary distributions become analytically solvable.

A general solvable model of neighbor-dependent substitutions

Two biologically interesting models analysed :

- Kimura+CpG and Tamura+CpG model
- writing stationary distributions
- deriving substitution rates estimators
- analysis on human data

$\hookrightarrow$A program for simulating evolution is also available.

L. Palmeira, J. R. Lobry and L. Guéguen, in prep.

YpR stationary distributions

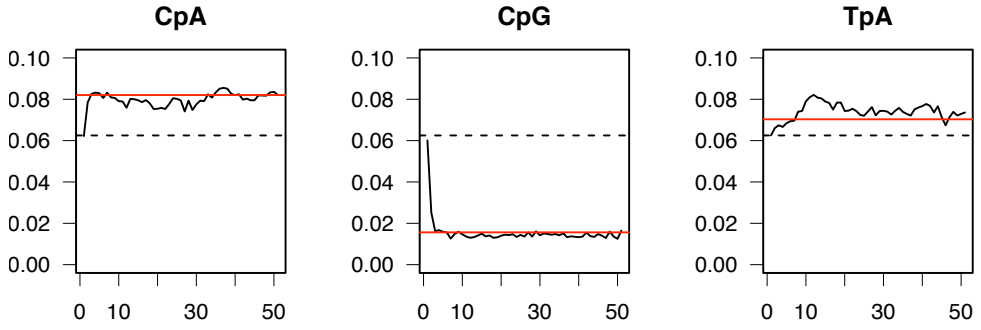
Analytical results

$$F(CA) = F(TG) = \frac{2(w+v)(3w+v) + r(5w+2v)}{32(w+v)(3w+v) + 8r(7w+3v)}$$

$$F(CG) = \frac{(w+v)(3w+v)}{16(w+v)(3w+v) + 4r(7w+3v)}$$

$$F(TA) = \frac{(w+v)(3w+v) + r(2w+v)}{16(w+v)(3w+v) + 4r(7w+3v)}$$

Simulations results



Substitution rates estimation

Analytical results

$$\frac{v}{w} = \frac{6g + 14a - \frac{5}{4}}{-(2g + 6a - \frac{1}{2})}$$

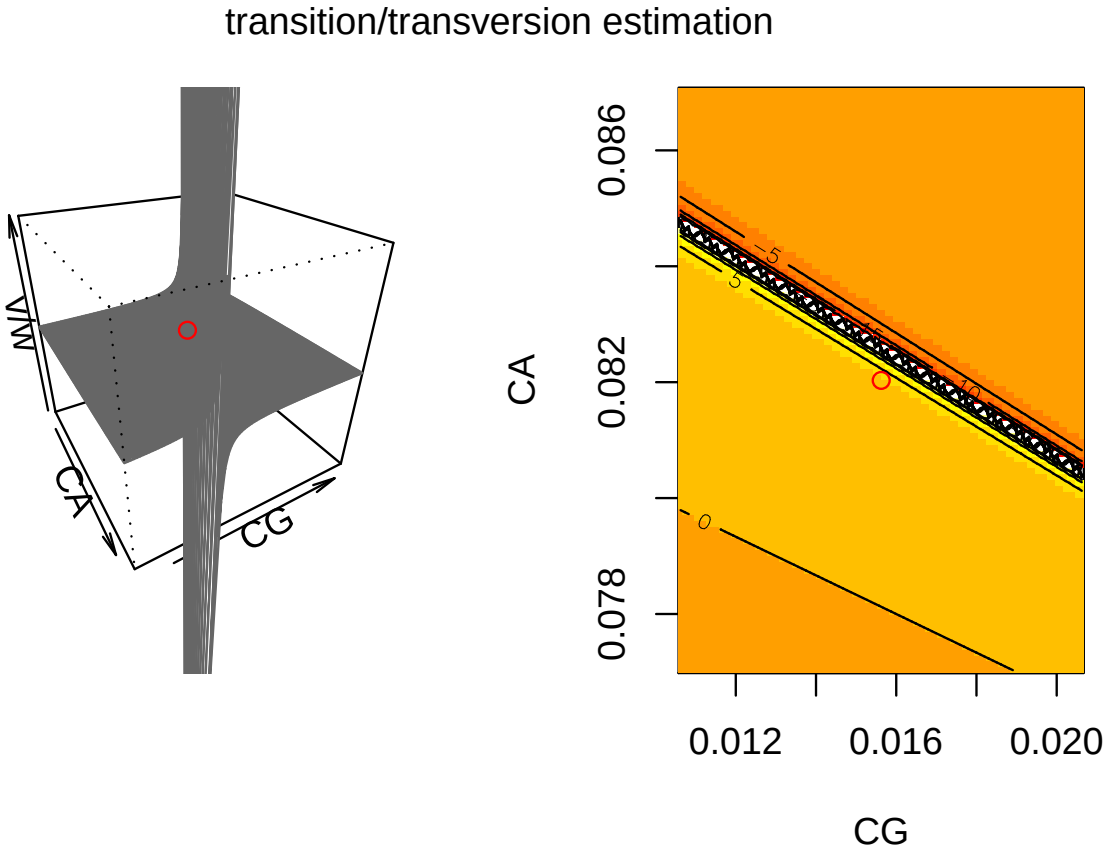
$$\frac{r}{w} = \frac{16a - 1}{4g} * (\frac{v}{w} + 1)$$

where $g = F(CG)$ and $a = F(CA)$

Substitution rates estimation

Sensitivity analysis

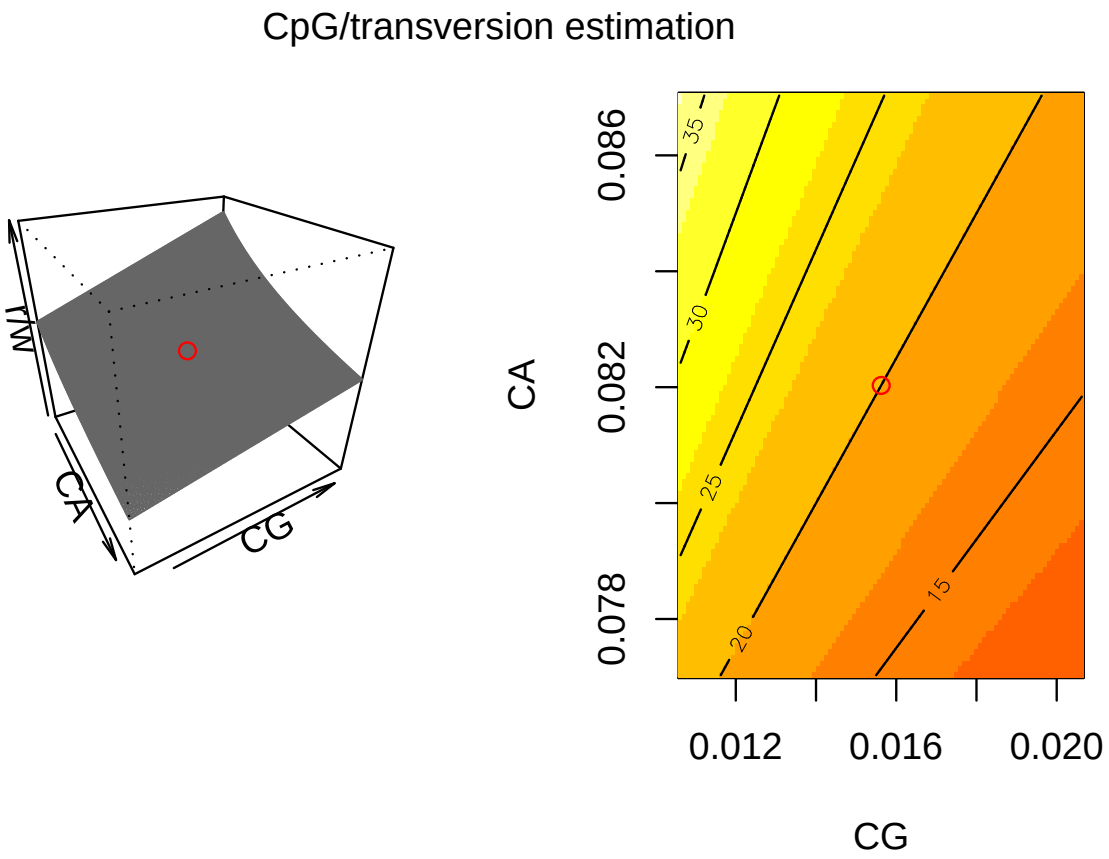
- introduce some noise in equilibrium frequencies
- investigate substitution rates estimation ($\frac{v}{w}$)



Substitution rates estimation

Sensitivity analysis

- introduce some noise in equilibrium frequencies
- investigate substitution rates estimation ($\frac{r}{w}$)



References

This work leads to :

L. Palmeira, J. R. Lobry, L. Guéguen
Models of DNA evolution with neighbor-dependent substitutions.
Work in progress.

References

This work leads to :

- L. Palmeira, J. R. Lobry, L. Guéguen
  Models of DNA evolution with neighbor-dependent substitutions.
  Work in progress.
- L. Palmeira, J. R. Lobry, L. Guéguen
  Neighboring-sites dependencies in evolution affect phylogenetic reconstruction.
  Work in progress.

Perspectives

- analyze over- and under-representation of all dinucleotides in bacteria
- correlations related to neighbor-dependent substitution processes ?
- analyze CpG substitution rates in vertebrate lineages
- analyze CpG substitution rates variation across a genome

Thank you !