Analyzing and modeling neighboring site dependencies in DNA evolution

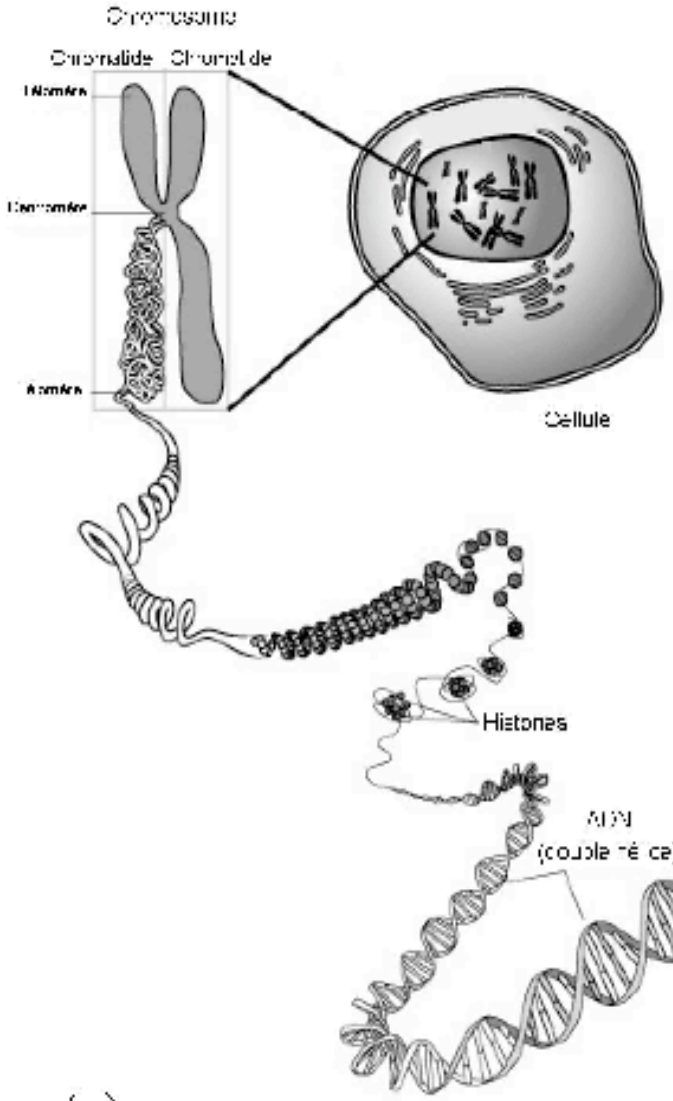
Leonor Palmeira

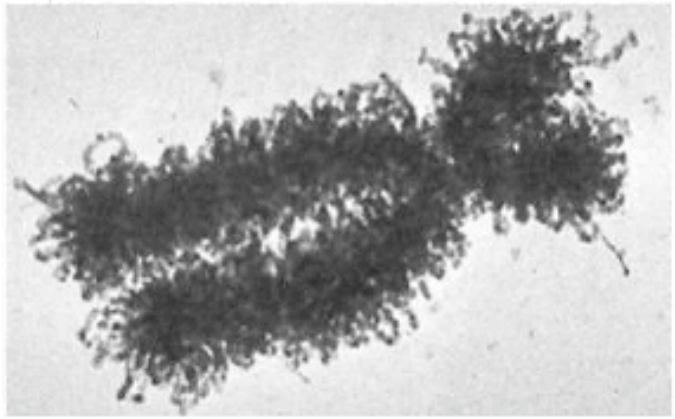
Laboratoire de Biométrie et Biologie Évolutive
UMR CNRS 5558 - INRIA
Université Claude Bernard - Lyon 1

Directeur de recherches: Jean R. Lobry
Co-encadrant: Laurent Guéguen

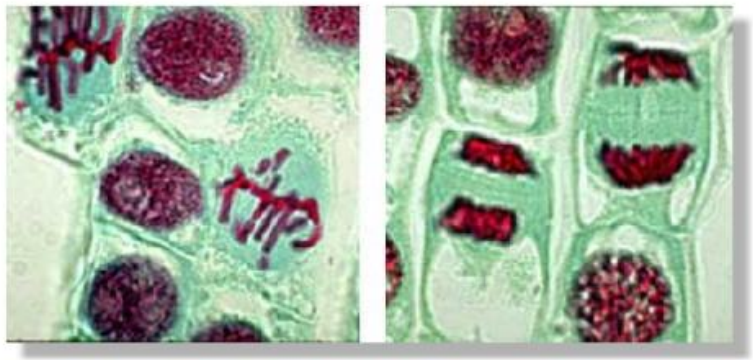
CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

INRIA


Cellule
Histones
(a)


(b)

Sequences change through time

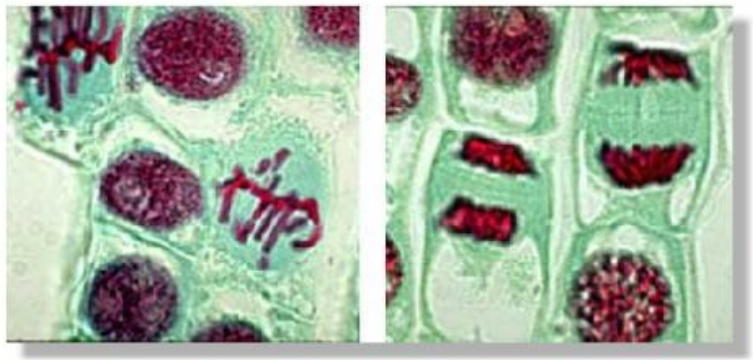


When cells multiply, genomes are transmitted with some errors:

- large scale errors (rearrangements)
- small scale errors
  - mutations
  - insertions-deletions
  - small duplications

$\hookrightarrow$ fixation of these modifications either by selection or by genetic drift

Sequences change through time



When cells mutliply, genomes are transmitted with some errors:

- large scale errors (rearrangements)
- small scale errors
  - mutations
  - insertions-deletions
  - small duplications

$\hookrightarrow$ fixation of these modifications either by selection or by genetic drift

Motivations

Dinucleotides
Statistics
Ultraviolet light

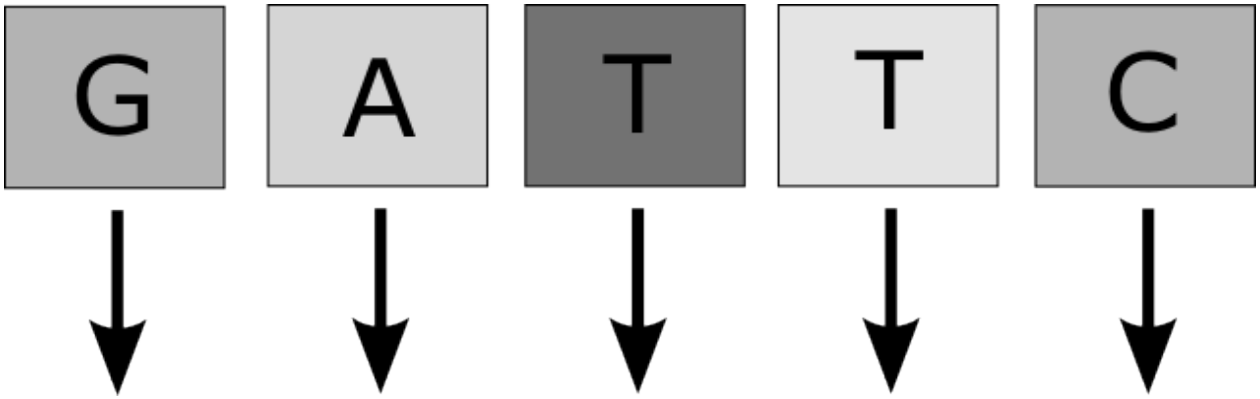
Modeling evolution
Motivations
Modeling
Applications

Conclusions

Perspectives

Motivations

General hypothesis: sites evolve independently



$\hookrightarrow$ mathematical simplification

$\hookrightarrow$ treat sites individually

$\Rightarrow$ strong implications on real data

Motivations

Dinucleotides
Statistics
Ultraviolet light

Modeling evolution
Motivations
Modeling
Applications

Conclusions

Perspectives

Substitutions are influenced by neighboring sites

→ The CpG effect in methylated genomes

- methylated CpG dinucleotides mutate towards TpG(CpA)
- C in a CpG dinucleotide mutates $\sim$ 10 times faster
- strong effect in vertebrate genomes

Motivations

Dinucleotides
Statistics
Ultraviolet light

Modeling evolution
Motivations
Modeling
Applications

Conclusions

Perspectives

Selective pressures on specific dinucleotides

→ Ultraviolet light damages adjacent pyrimidines

- pyrimidine dinucleotides (C and T) are damaged by UV light
- adjacent Cs and Ts are possible targets of negative selective pressure

Motivations

Dinucleotides
Statistics
Ultraviolet light

Modeling evolution
Motivations
Modeling
Applications

Conclusions

Perspectives

Motivations

Sites do not evolve independently

$\Rightarrow$ how to evaluate a sequence's deviation from the "independent sites" hypothesis?

- statistical analysis of dinucleotide representation

$\Rightarrow$ how to explain and describe the preservation of dependencies between sites in evolution?

- model evolution with neighboring site dependencies

Motivations

Dinucleotides
Statistics
Ultraviolet light

Modeling evolution
Motivations
Modeling
Applications

Conclusions

Perspectives

One strong implication of site independence

$$f_{XY} = f_X \times f_Y \iff \frac{f_{XY}}{f_X f_Y} = 1$$

Invalidated in a wide variety of cases:

- TpA widely under-represented in Bacteria, Chloroplasta, Eukaryotes
- correlation between CpG, TpG(CpA) representation in vertebrates

Motivations

Dinucleotides
Statistics
Ultraviolet light

Modeling evolution
Motivations
Modeling
Applications

Conclusions

Perspectives

The ρ statistic

A simple statistic

$$\rho_{XY} = \frac{f_{XY}}{f_X f_Y}$$

- Widely used, as a measure of dinucleotide over- and under-representation (Karlin, 1994; Duret and Galtier, 2000; Oakes et al, 2007)
- No clear boundaries for rejecting the null model.

↪ Abusive use of inaccurate thresholds. limitation # 1

- The null model is invalidated for most sequences.

↪ The rejection of the null model does not bear much information. limitation # 2

Motivations

Dinucleotides
Statistics
Ultraviolet light

Modeling evolution
Motivations
Modeling
Applications

Conclusions

Perspectives

The z-score statistic

General z-score statistic

$$z_{score} = \frac{\omega_{XY} - E(\omega_{XY})}{\sqrt{Var(\omega_{XY})}}$$

where ω_{XY} is a measure associated to dinucleotide XY.

$E(\omega_{XY})$ and $Var(\omega_{XY})$ can either be obtained by simulation or by exact analytical calculation.

When the central limit theorem can be applied:

$$z\text{-score} \sim \mathcal{N}(0,1)$$

$\hookrightarrow$ solves limitation # 1 of the ρ statistic

The z-score statistic

Motivations

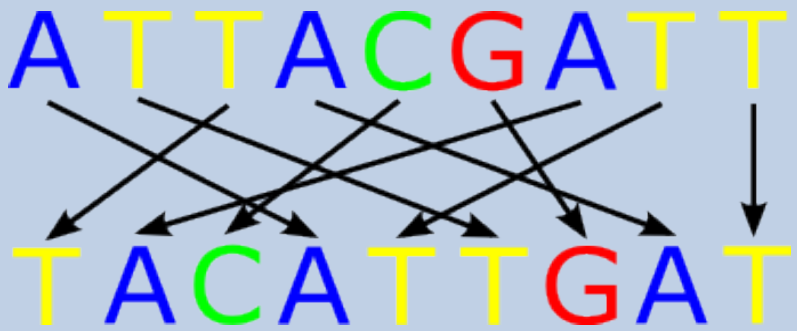
Dinucleotides
Statistics
Ultraviolet light

Modeling evolution
Motivations
Modeling
Applications

Conclusions

Perspectives

z-score with a base permutation model (z-base)



with analytical results available from Schbath, 1995.

- the base composition (*i.e* G+C content) of the studied sequence is accounted for.

Motivations

Dinucleotides
Statistics
Ultraviolet light

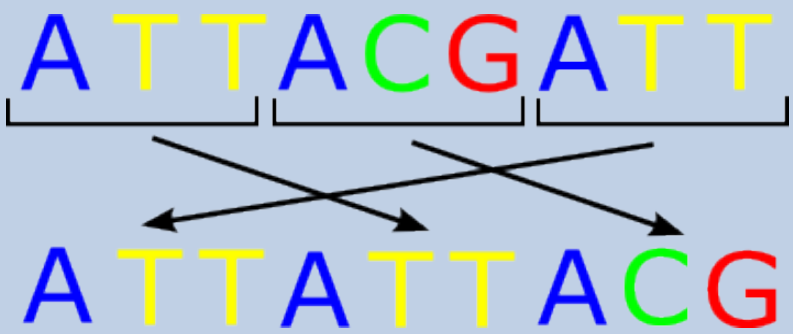
Modeling evolution
Motivations
Modeling
Applications

Conclusions

Perspectives

The z-score statistic

z-score with a codon permutation model (z-codon)



with analytical results available from Gautier *et al.*, 1985.

- the codon usage bias of the studied sequence is accounted for.

$\hookrightarrow$ solves partially limitation # 2 of the ρ statistic

$\Rightarrow$ ρ and z-score statistics presented here were implemented in SEQINR - R library for statistical analysis of sequences.

Motivations

Dinucleotides
Statistics
Ultraviolet light

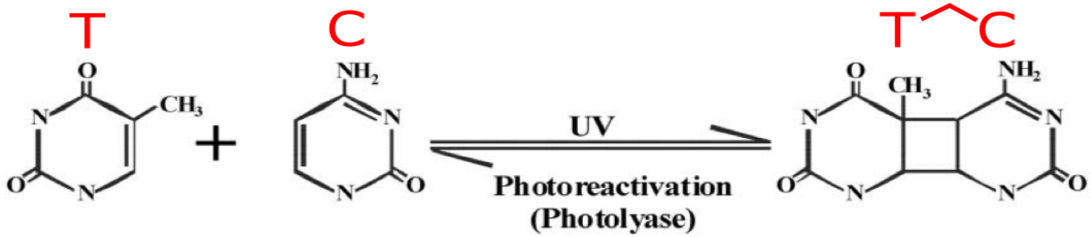
Modeling evolution
Motivations
Modeling
Applications

Conclusions

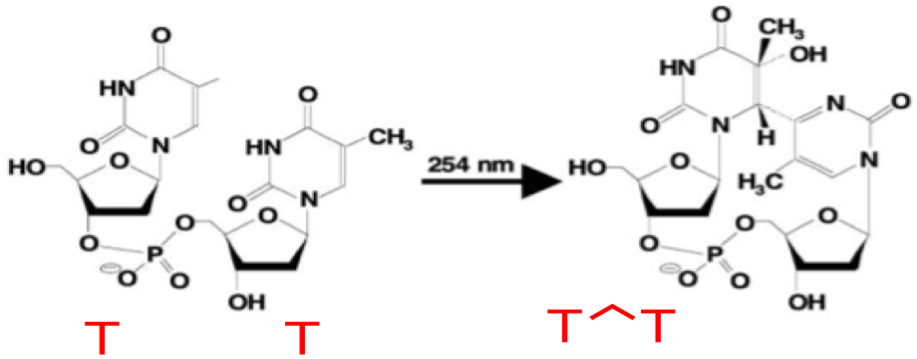
Perspectives

The damages caused by ultraviolet light

- pyrimidine dimers between adjacent pyrimidine bases (C and T) $\sim$ 75 %



- pyrimidine (6-4) pyrimidone photoproducts $\sim$ 25 %



$\hookrightarrow$ lead to a local DNA distortion which blocks transcription and replication if it is not repaired in time.

Motivations

Dinucleotides
Statistics
Ultraviolet light

Modeling evolution
Motivations
Modeling
Applications

Conclusions

Perspectives

The impact of ultraviolet light on genomes

Is there an impact of ultraviolet light on genomic base composition?

- T-rich dinucleotides might be preferentially damaged (Singer and Ames, 1970)
- G+C related to UV exposition (Singer and Ames, 1970)
- subsequent criticism of these results (Bak *et al.*, 1972)

$\hookrightarrow$ other statistics available than G+C content

$\Rightarrow$ z-score on pyrimidine dinucleotides to investigate possible under-representation

General study on all Bacteria

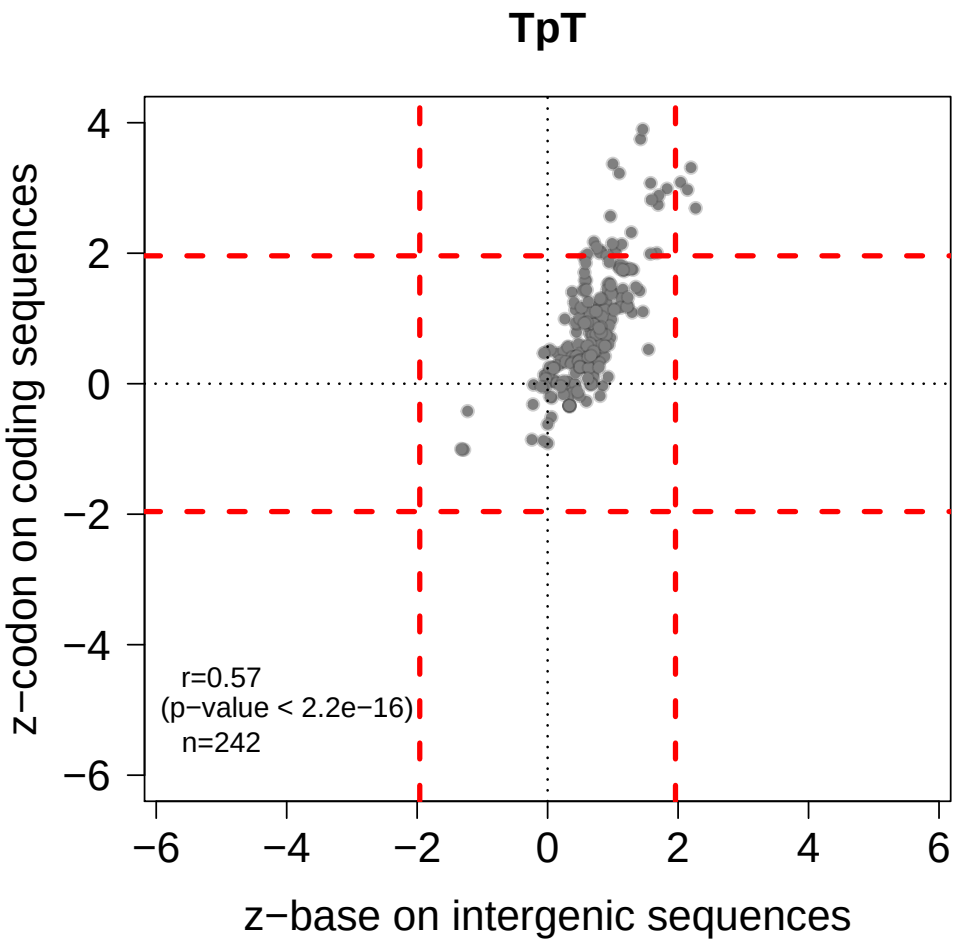
Motivations

Dinucleotides
Statistics
Ultraviolet light

Modeling evolution
Motivations
Modeling
Applications

Conclusions

Perspectives



Motivations

Dinucleotides
Statistics
Ultraviolet light

Modeling evolution
Motivations
Modeling
Applications

Conclusions

Perspectives

A case study on an oceanic Bacteria

Prochlorococcus marinus as an ideal example

- one of the most abundant micro-organisms in oceans, involved in a great part of the oceans primary production
- stratified habitat in the water column
  - three fully sequenced strains of *Prochlorococcus marinus*
  - adapted to different depths in the water column
  - *i.e.* exposed to different UV contents

Photo credit: Genoscope - Centre National de Séquençage.

A case study on an oceanic Bacteria - Results

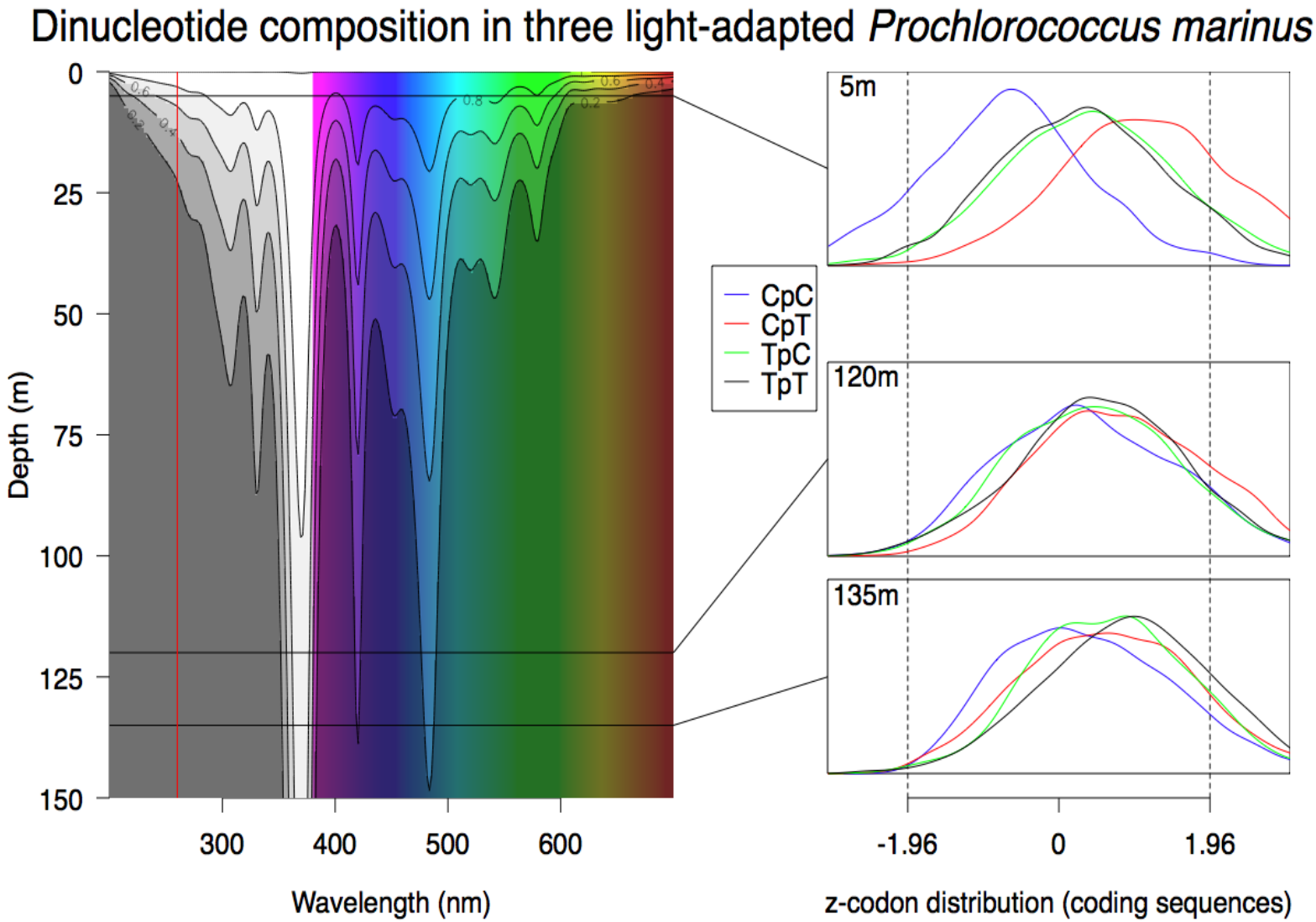
Motivations

Dinucleotides
Statistics
Ultraviolet light

Modeling evolution
Motivations
Modeling
Applications

Conclusions

Perspectives



Motivations

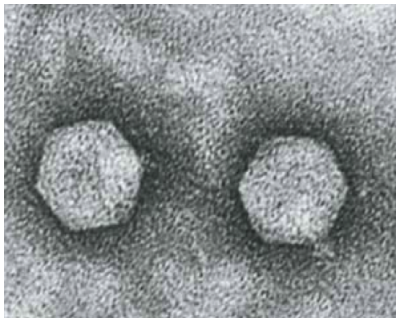
Dinucleotides
Statistics
Ultraviolet light

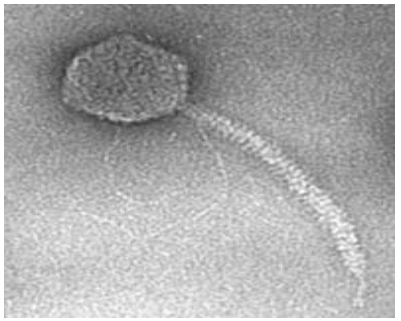
Modeling evolution
Motivations
Modeling
Applications

Conclusions

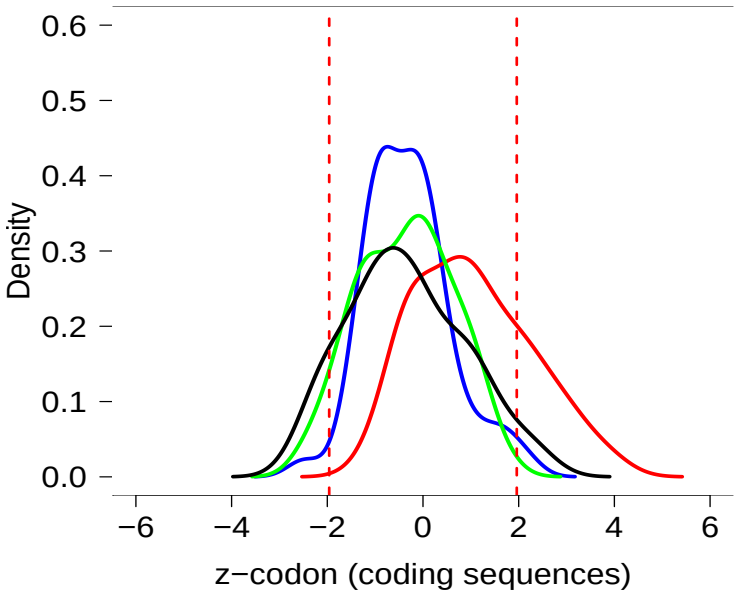
Perspectives

Are Viruses more vulnerable?

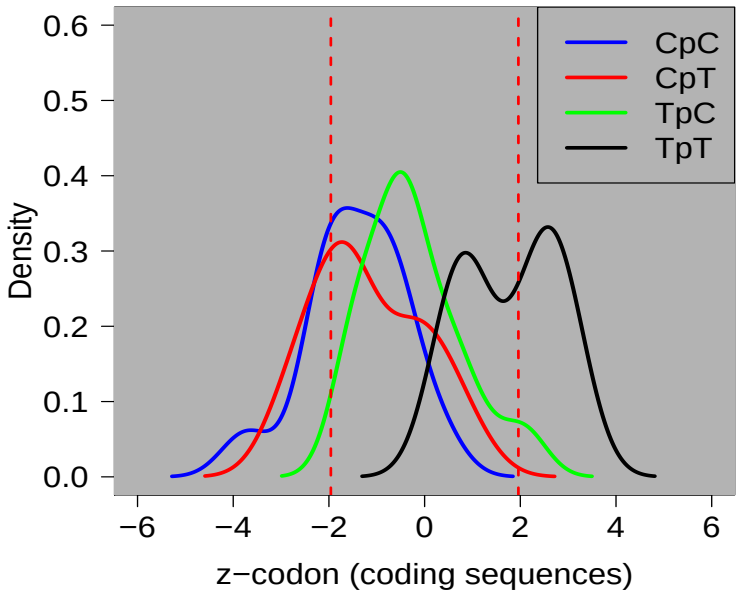




Infects surface P. marinus



Infects Geobacillus at 10900m depth



Motivations

Dinucleotides
Statistics
Ultraviolet light

Modeling evolution
Motivations
Modeling
Applications

Conclusions

Perspectives

Conclusions

⇒ No systematic effect of UV light on dinucleotide content
- ↪ pyrimidine dinucleotides are not avoided.
- ↪ true for all Bacteria and for a set of Virus taken as a good example.

⇒ Protection mechanisms

⇒ Repair mechanisms
- ↪ use the host's repair machinery.

Motivations

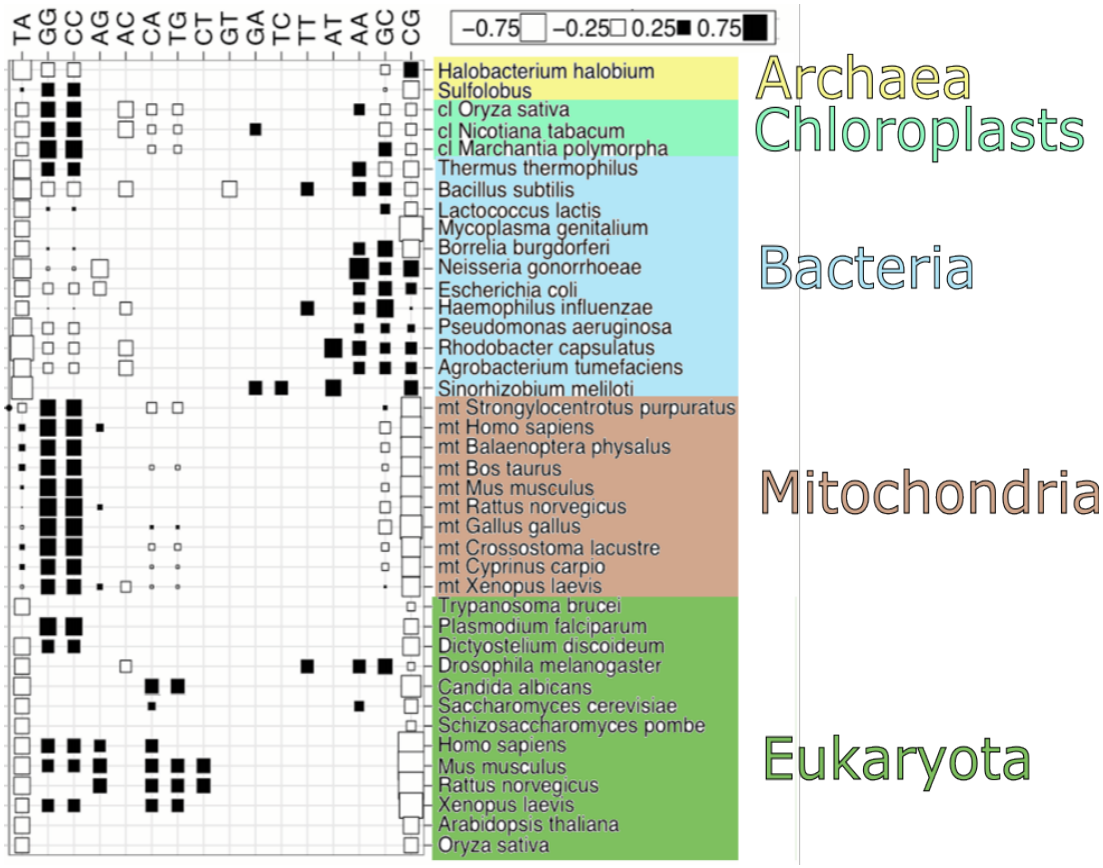
Dinucleotides
Statistics
Ultraviolet light

Modeling evolution
Motivations
Modeling
Applications

Conclusions

Perspectives

Dinucleotides relation to sequence evolution



Data compiled from Burge *et al.* 1992, Brendel *et al.* 1992, Cardon *et al.* 1994, Karlin *et al.* 1994, 1997.

Motivations

Dinucleotides
Statistics
Ultraviolet light

Modeling evolution
Motivations
Modeling
Applications

Conclusions

Perspectives

The importance of being realistic

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)*

Some consequences of using the hypothesis of independent sites

- bias in estimating evolutionary distance between two sequences (von Haeseler and Schöniger, 1998)
- bias in phylogenetic reconstruction (von Haeseler and Schöniger, 1998)

Motivations

Dinucleotides
Statistics
Ultraviolet light

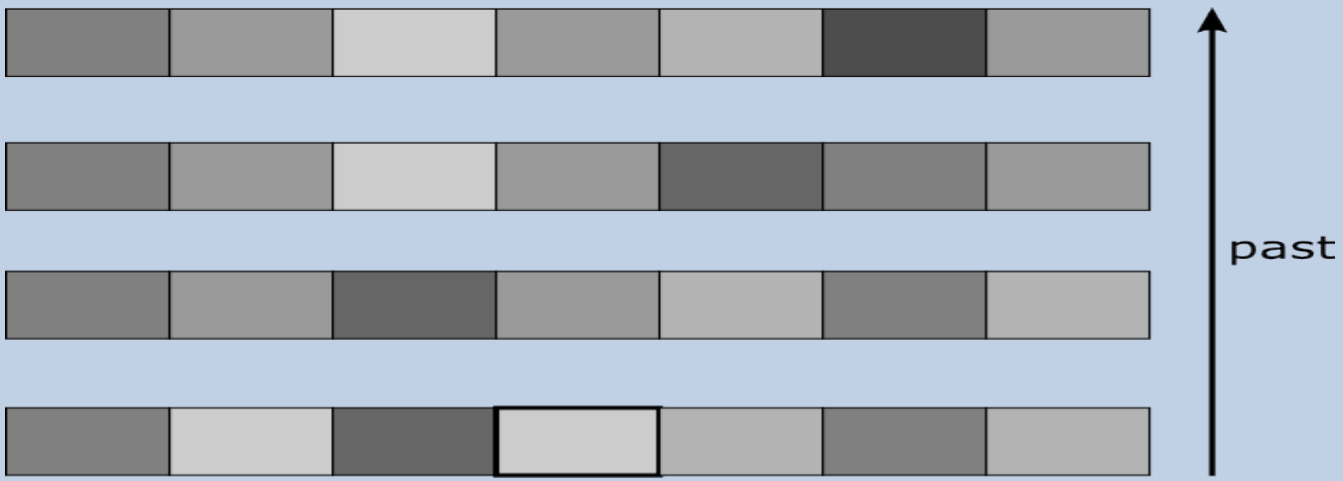
Modeling evolution
Motivations
Modeling
Applications

Conclusions

Perspectives

A general model of neighbor-dependent substitutions

The dependency cone problem



Motivations

Dinucleotides
Statistics
Ultraviolet light

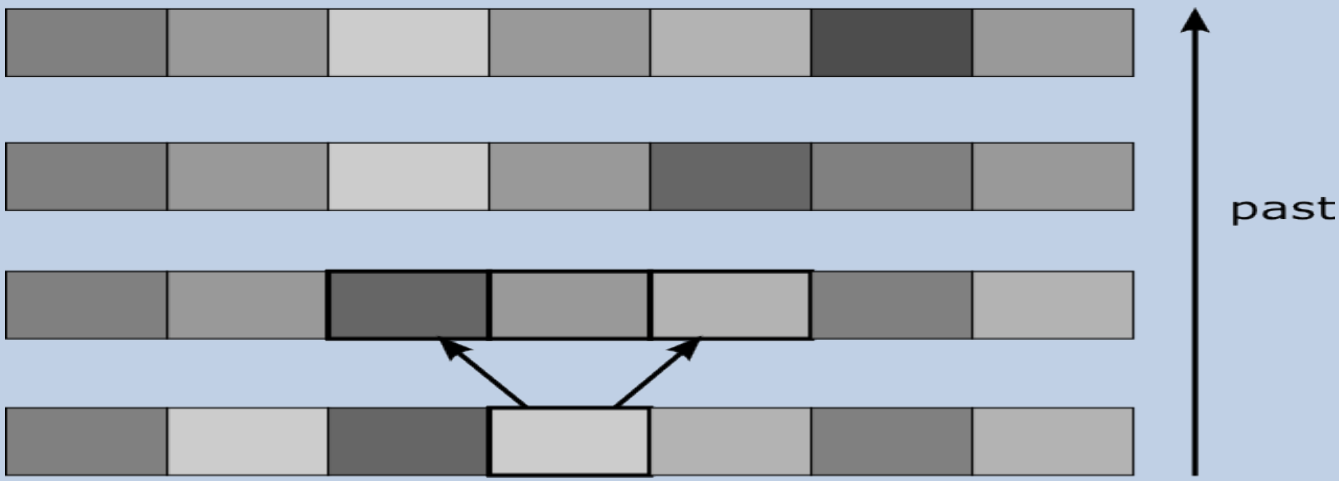
Modeling evolution
Motivations
Modeling
Applications

Conclusions

Perspectives

A general model of neighbor-dependent substitutions

The dependency cone problem



Motivations

Dinucleotides
Statistics
Ultraviolet light

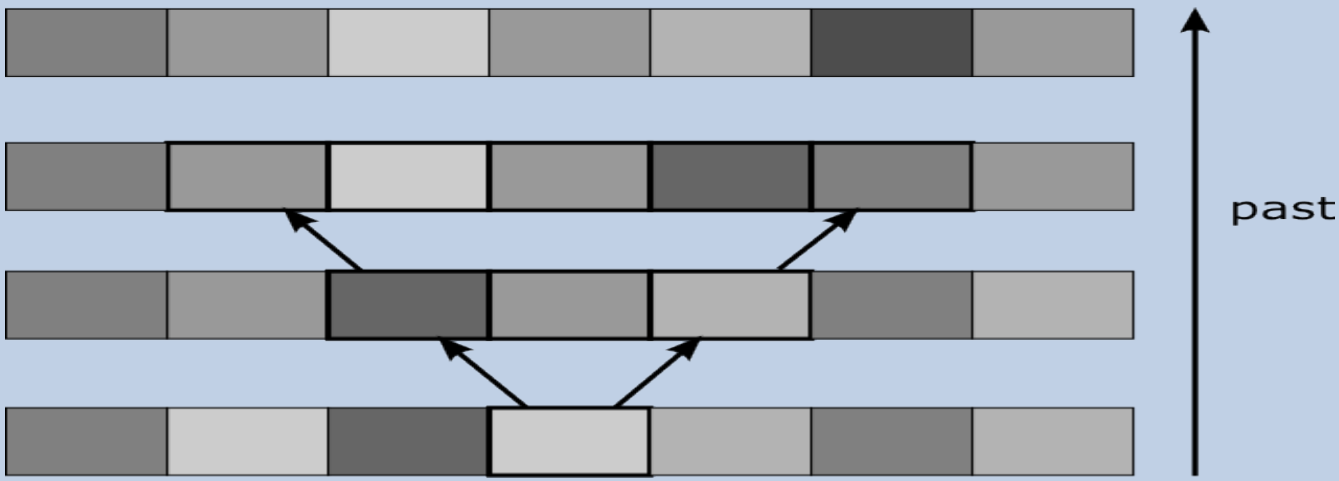
Modeling evolution
Motivations
Modeling
Applications

Conclusions

Perspectives

A general model of neighbor-dependent substitutions

The dependency cone problem



Motivations

Dinucleotides
Statistics
Ultraviolet light

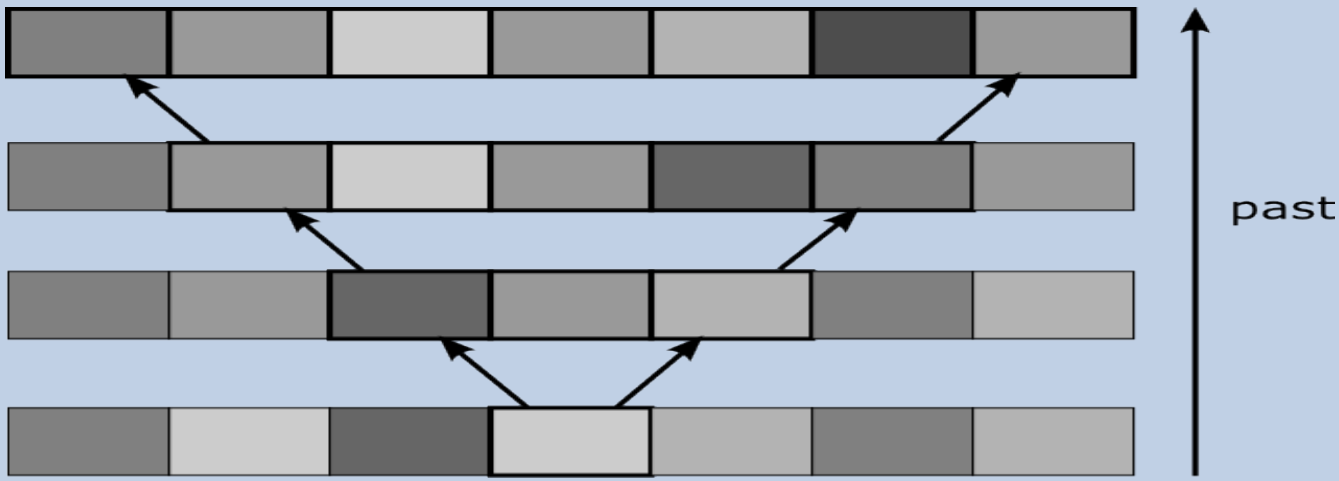
Modeling evolution
Motivations
Modeling
Applications

Conclusions

Perspectives

A general model of neighbor-dependent substitutions

The dependency cone problem



Motivations

Dinucleotides
Statistics
Ultraviolet light

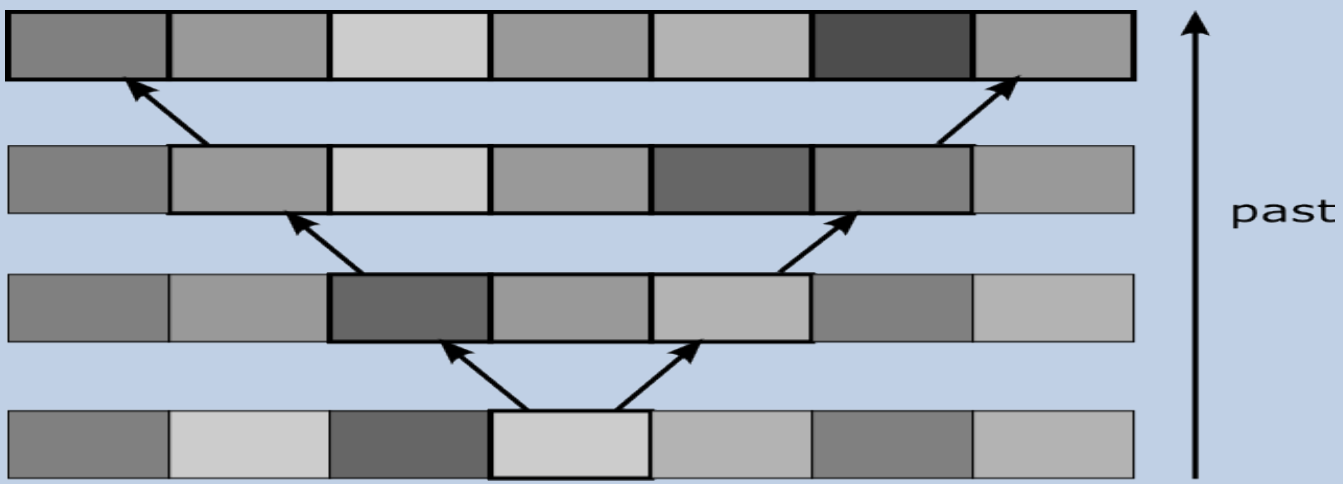
Modeling evolution
Motivations
Modeling
Applications

Conclusions

Perspectives

A general model of neighbor-dependent substitutions

The dependency cone problem



$\Rightarrow$ complex models, no analytical results available

Hyper-mutable CpG dinucleotides

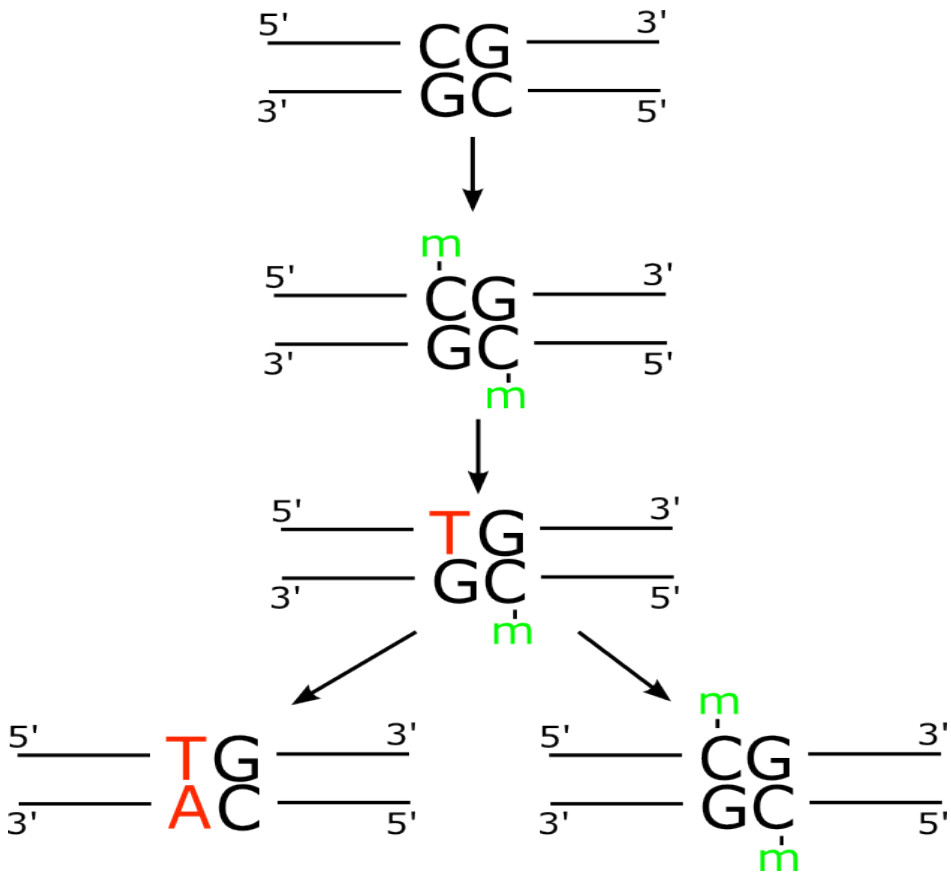
Motivations

Dinucleotides
Statistics
Ultraviolet light

Modeling evolution
Motivations
Modeling
Applications

Conclusions

Perspectives



Motivations

Dinucleotides
Statistics
Ultraviolet light

Modeling evolution
Motivations
Modeling
Applications

Conclusions

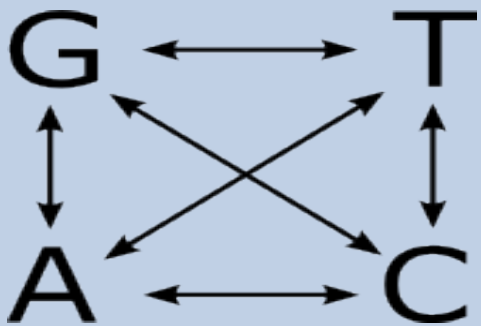
Perspectives

A general solvable model of neighbor-dependent substitutions

(Bérard, Gouéré and Piau, 2005)

Combining:

- a simple nucleotide substitution model of the form



Motivations

Dinucleotides
Statistics
Ultraviolet light

Modeling evolution
Motivations
Modeling
Applications

Conclusions

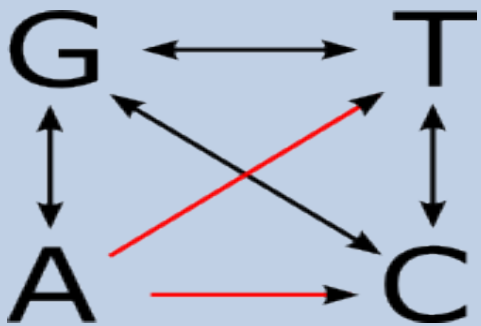
Perspectives

A general solvable model of neighbor-dependent substitutions

(Bérard, Gouéré and Piau, 2005)

Combining:

- a simple nucleotide substitution model of the form



Motivations

Dinucleotides
Statistics
Ultraviolet light

Modeling evolution
Motivations
Modeling
Applications

Conclusions

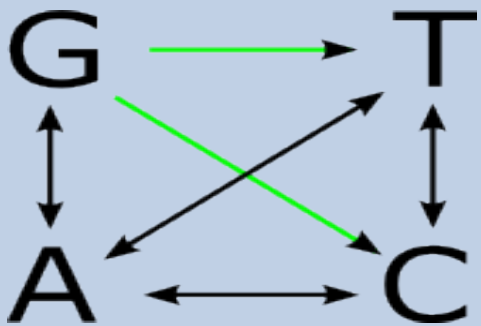
Perspectives

A general solvable model of neighbor-dependent substitutions

(Bérard, Gouéré and Piau, 2005)

Combining:

- a simple nucleotide substitution model of the form



Motivations

Dinucleotides
Statistics
Ultraviolet light

Modeling evolution
Motivations
Modeling
Applications

Conclusions

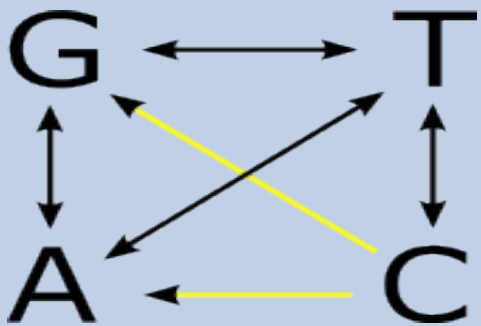
Perspectives

A general solvable model of neighbor-dependent substitutions

(Bérard, Gouéré and Piau, 2005)

Combining:

- a simple nucleotide substitution model of the form



Motivations

Dinucleotides
Statistics
Ultraviolet light

Modeling evolution
Motivations
Modeling
Applications

Conclusions

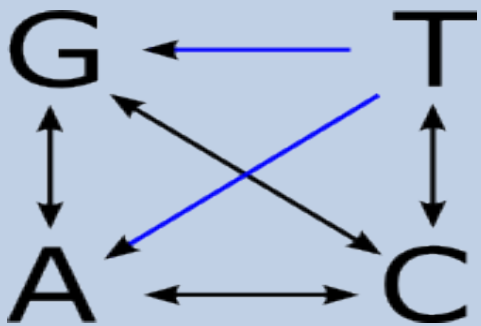
Perspectives

A general solvable model of neighbor-dependent substitutions

(Bérard, Gouéré and Piau, 2005)

Combining:

- a simple nucleotide substitution model of the form



Motivations

Dinucleotides
Statistics
Ultraviolet light

Modeling evolution
Motivations
Modeling
Applications

Conclusions

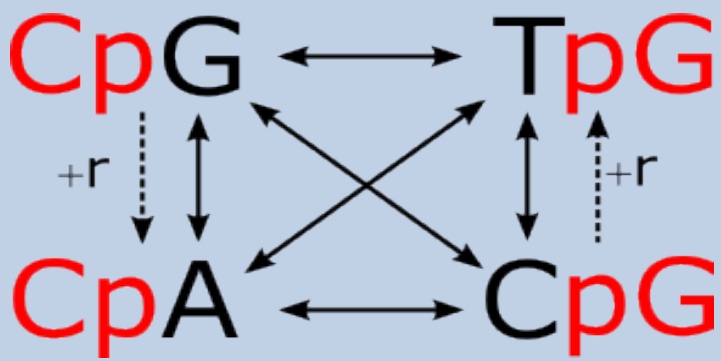
Perspectives

A general solvable model of neighbor-dependent substitutions

(Bérard, Gouéré and Piau, 2005)

Combining:

- a simple nucleotide substitution model
- and all dinucleotide substitution processes of the form YpR $\longrightarrow$ YpR



$\Rightarrow$ stationary distributions become analytically solvable.

Motivations

Dinucleotides
Statistics
Ultraviolet light

Modeling evolution
Motivations
Modeling
Applications

Conclusions

Perspectives

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 - chromosome 21

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

Motivations

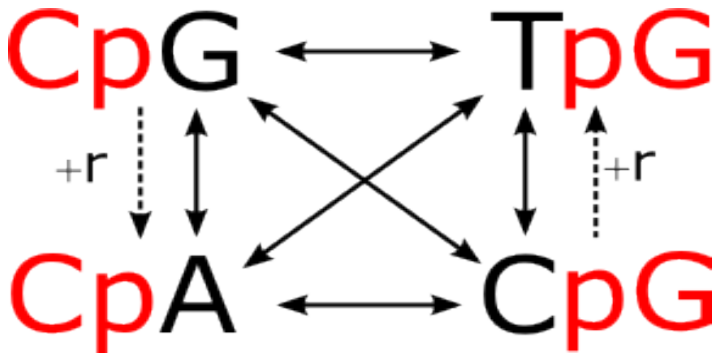
Dinucleotides
Statistics
Ultraviolet light

Modeling evolution
Motivations
Modeling
Applications

Conclusions

Perspectives

Tamura+CpG



α stands for the transitions (between C and T, or A and G)
β stands for the transversions (all other substitutions)

both these rates are multiplied by:

- θ for substitutions towards G or C
- $(1-\theta)$ for substitutions towards A or T

$+$ CpG substitutions (rate r)
$\Rightarrow$ describes G+C variation along the genome

Motivations

Dinucleotides
Statistics
Ultraviolet light

Modeling evolution
Motivations
Modeling
Applications

Conclusions

Perspectives

Stationary distribution

Dinucleotides

$$\pi_{CG} = \frac{\theta^2(3\beta+\alpha)(\alpha+\beta)}{4(3\beta+\alpha)(\alpha+\beta)+4r(3\beta+\alpha+\theta(\alpha+\beta))}$$

$$\pi_{TG} = \pi_{CA} = \frac{\theta(1-\theta)(3\beta+\alpha)(\alpha+\beta)+r\theta(\alpha+(3-\theta)\beta)}{4(3\beta+\alpha)(\alpha+\beta)+4r(3\beta+\alpha+\theta(\alpha+\beta))}$$

$$\pi_{TA} = \frac{1}{4} - \pi_{CG} - 2\pi_{CA}$$

Nucleotides

$$\pi_C = \pi_G = \frac{\theta}{2} - r\frac{\pi_{CG}}{\alpha+\beta}$$

$$\pi_A = \pi_T = \frac{1-\theta}{2} + r\frac{\pi_{CG}}{\alpha+\beta}$$

Motivations

Dinucleotides
Statistics
Ultraviolet light

Modeling evolution
Motivations
Modeling
Applications

Conclusions

Perspectives

Estimating the parameters

The θ parameter

$$\theta = 4\pi_{CA} + 4\pi_{TA} + 2\pi_C - \sqrt{(4\pi_{CA} + 4\pi_{TA} + 2\pi_C - 1)^2 + 4\pi_{TA}}$$

CpG substitution rate

$$\frac{r}{\alpha + \beta} = \frac{4(\pi_{CA} + \pi_{TA}) - \sqrt{(4\pi_{CA} + 4\pi_{TA} + 2\pi_C - 1)^2 + 4\pi_{TA}}}{2\pi_{CG}}$$

- Stationary distributions can be estimated by the observed frequencies
- Estimation from one sequence only.

Motivations

Dinucleotides
Statistics
Ultraviolet light

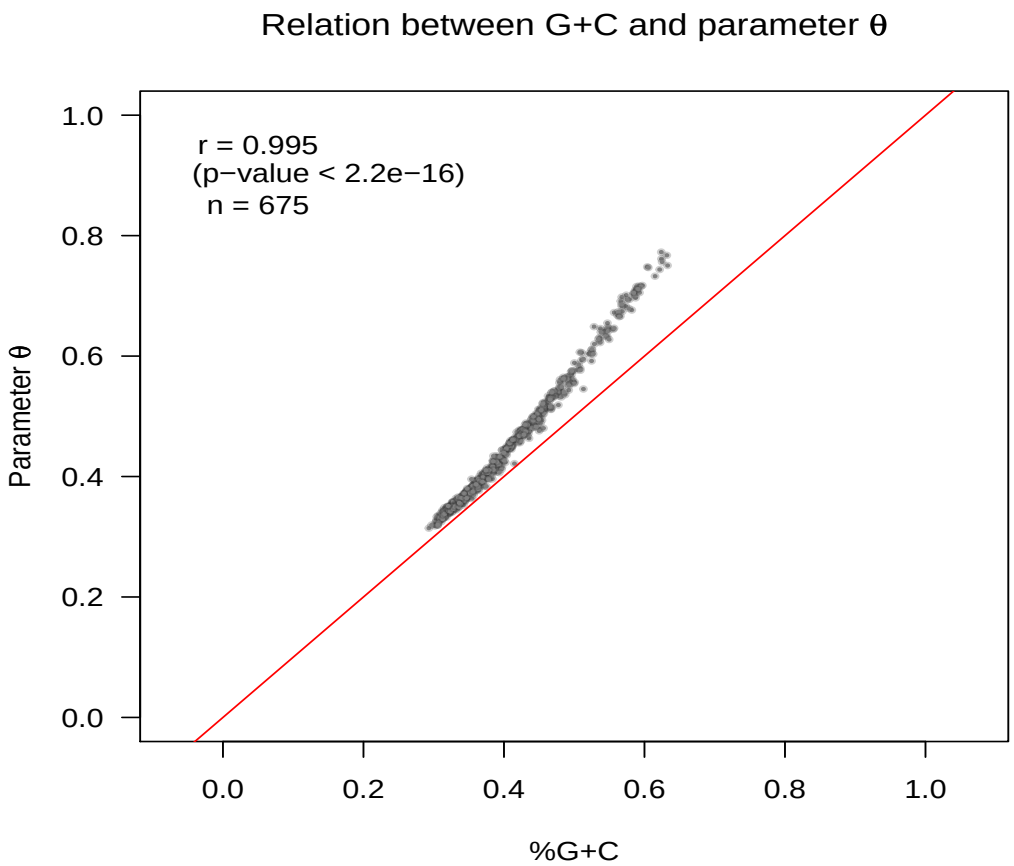
Modeling evolution
Motivations
Modeling
Applications

Conclusions

Perspectives

The θ parameter and its relation to G+C content – Example of chromosome 21

$$\%(G+C) = \pi_C + \pi_G = \theta - 2\frac{r}{\alpha+\beta}\pi_{CG}$$



The r parameter as a measure of CpG substitutions – Example of chromosome 21

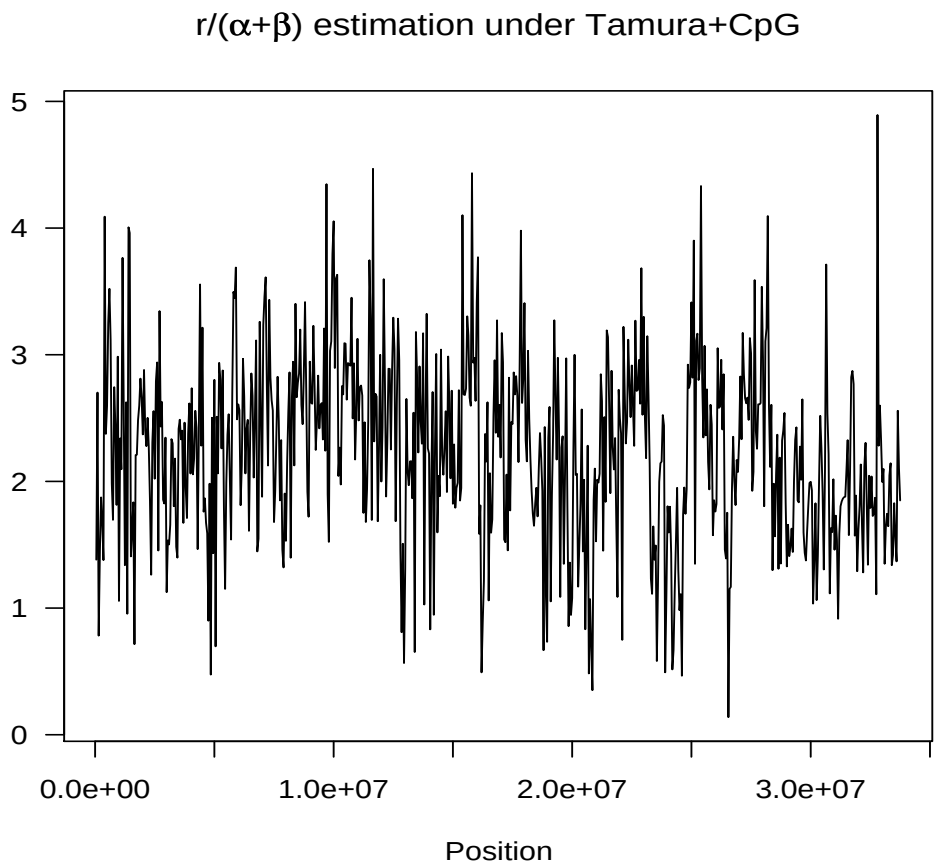
Motivations

Dinucleotides
Statistics
Ultraviolet light

Modeling evolution
Motivations
Modeling
Applications

Conclusions

Perspectives



Motivations

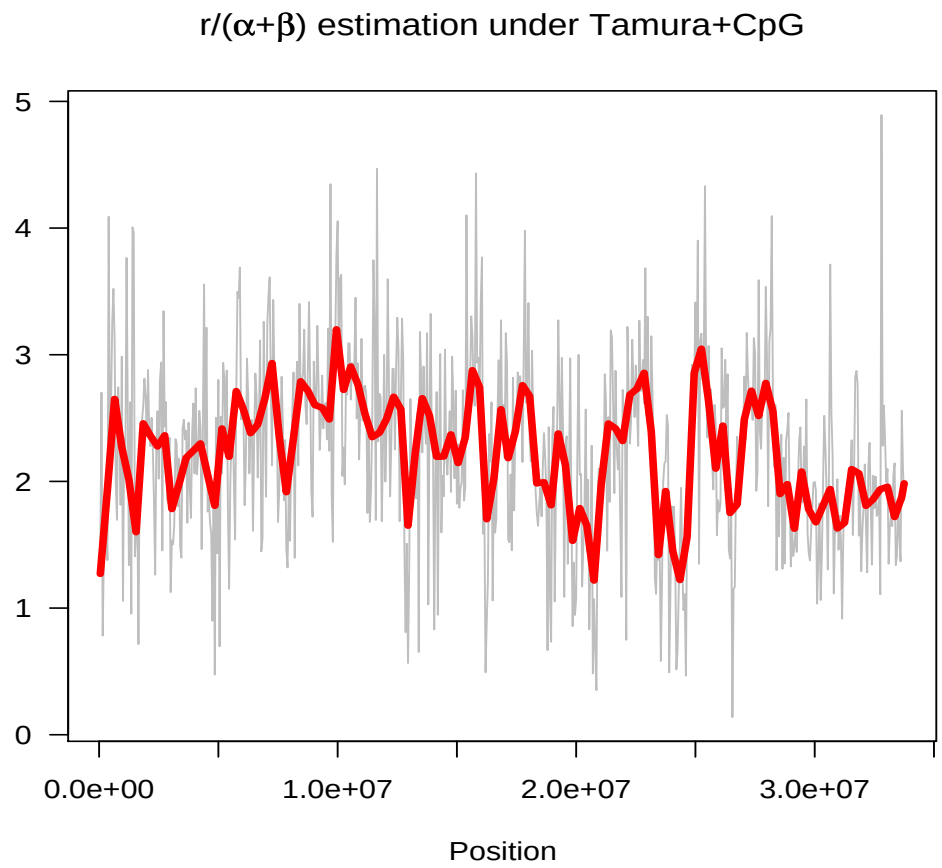
Dinucleotides
Statistics
Ultraviolet light

Modeling evolution
Motivations
Modeling
Applications

Conclusions

Perspectives

The r parameter as a measure of CpG substitutions – Example of chromosome 21



Motivations

Dinucleotides
Statistics
Ultraviolet light

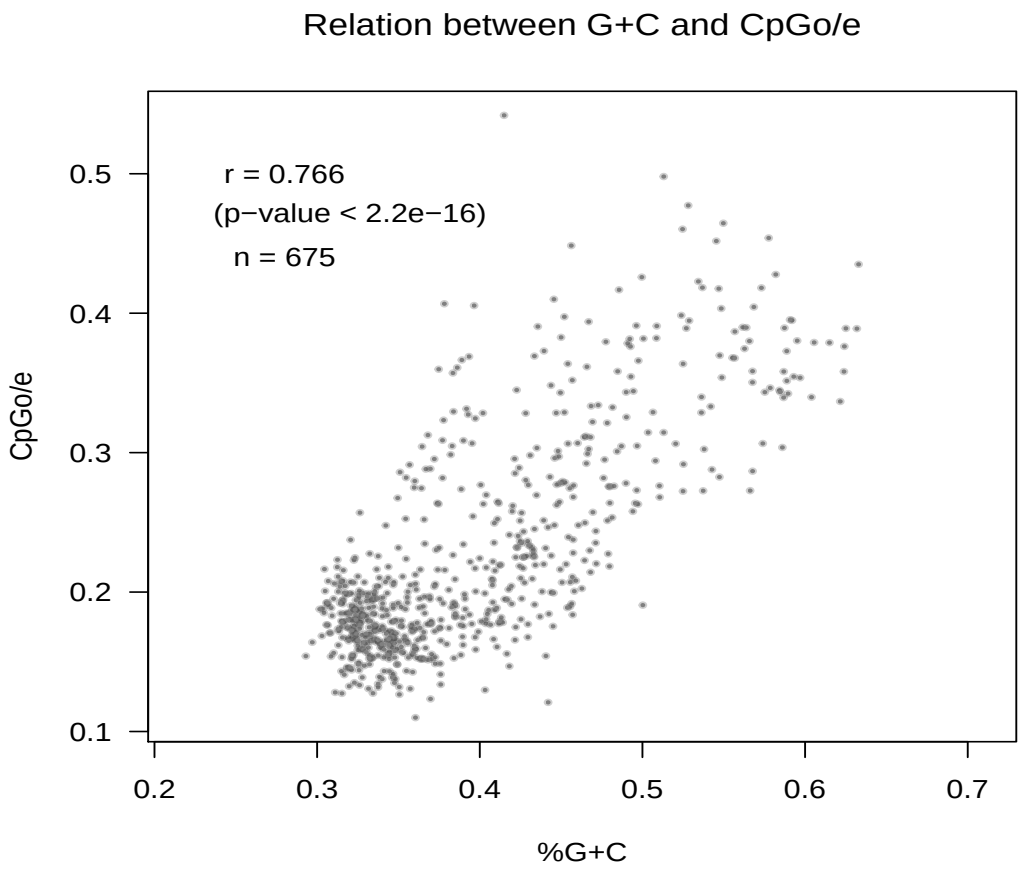
Modeling evolution
Motivations
Modeling
Applications

Conclusions

Perspectives

Relation between G+C and CpGo/e – Example of chromosome 21

$$CpGo/e = f_{CG}/(f_C \times f_G)$$



Motivations

Dinucleotides
Statistics
Ultraviolet light

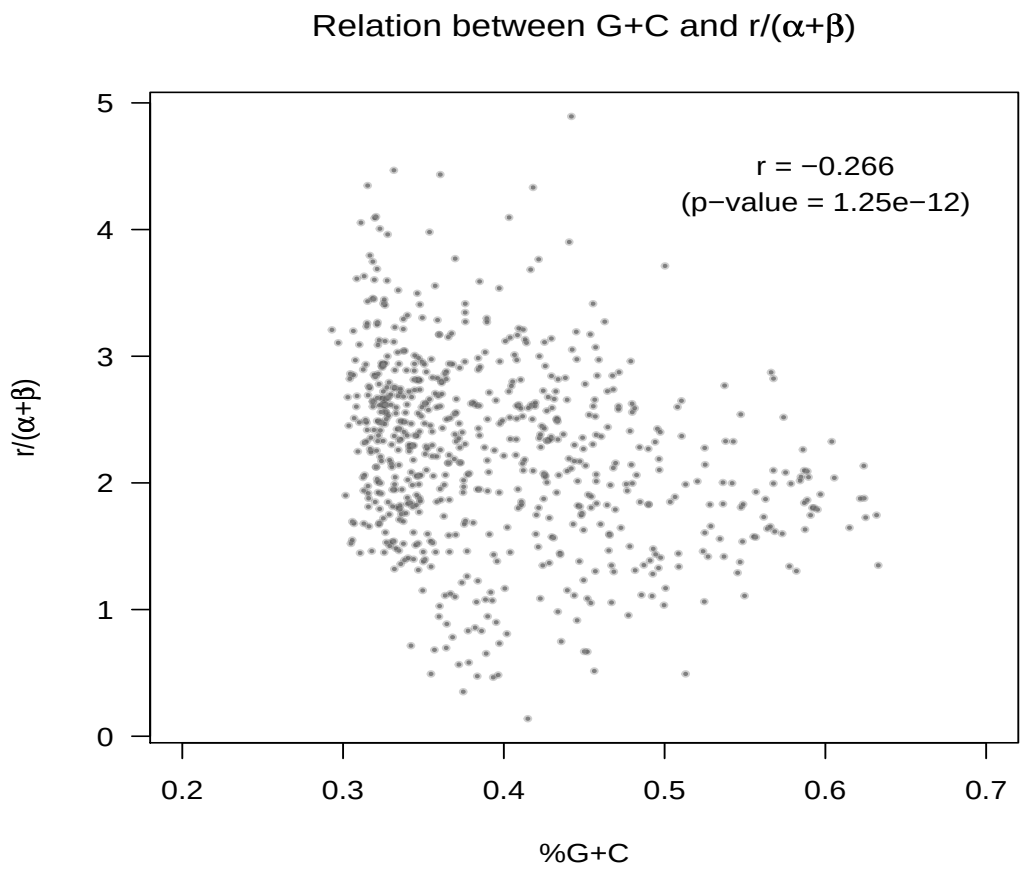
Modeling evolution
Motivations
Modeling
Applications

Conclusions

Perspectives

Relation between G+C and $\frac{r}{\alpha+\beta}$ – Example of

chromosome 21



Motivations

Dinucleotides
Statistics
Ultraviolet light

Modeling evolution
Motivations
Modeling
Applications

Conclusions

Perspectives

Conclusions

- Dinucleotides are far from being formed by chance
- Substitutions are influenced by direct neighbors
- Neighbor-dependent substitutions must be taken into account
- These models are not necessarily intractable:
  - stationary distribution
  - parameter estimation

$\hookrightarrow$ recover CpG substitution rates from the sequence only.

Motivations

Dinucleotides
Statistics
Ultraviolet light

Modeling evolution
Motivations
Modeling
Applications

Conclusions

Perspectives

Perspectives

Short term

- compare $\frac{r}{\alpha + \beta}$ with observed substitution patterns
- write general estimations of the model parameters for this class of models

Long term

- enlarge results to non stationary sequences
- and to other substitutions than YpR $\longrightarrow$ YpR
- incorporate these models in phylogenetic inference methods

Muito obrigada.

Thank you very much.

Vielen Dank.

Merci beaucoup.

Muchas gracias.

Mulţumesc foarte mult.

Mange tak.

Shukran jazilan.

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

INRIA

Muito obrigada.

Thank you very much.

Vielen Dank.

Merci beaucoup.

Muchas gracias.

Mulţumesc foarte mult.

Mange tak.

Shukran jazilan.

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

INRIA

Muito obrigada.

Thank you very much.

Vielen Dank.

Merci beaucoup.

Muchas gracias.

Mulţumesc foarte mult.

Mange tak.

Shukran jazilan.

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

INRIA

Muito obrigada.

Thank you very much.

Vielen Dank.

Merci beaucoup.

Muchas gracias.

Mulţumesc foarte mult.

Mange tak.

Shukran jazilan.

Muito obrigada.

Thank you very much.

Vielen Dank.

Merci beaucoup.

Muchas gracias.

Mulţumesc foarte mult.

Mange tak.

Shukran jazilan.

Muito obrigada.

Thank you very much.

Vielen Dank.

Merci beaucoup.

Muchas gracias.

Mulțumesc foarte mult.

Mange tak.

Shukran jazilan.

Muito obrigada.

Thank you very much.

Vielen Dank.

Merci beaucoup.

Muchas gracias.

Mulțumesc foarte mult.

Mange tak.

Shukran jazilan.

Muito obrigada.

Thank you very much.

Vielen Dank.

Merci beaucoup.

Muchas gracias.

Mulțumesc foarte mult.

Mange tak.

Shukran jazilan.