An evolution model for Limited Insertion Independent of Substitution (LIIS)

S. Lèbre, C. Michel

Université de Strasbourg
-
ICube - CNRS UMR 7005
Équipe de Bioinformatique théorique,
Fouille de données et Optimisation stochastique

Evolution models for Substitution-Insertion-Deletion

Over the last 20 years: 3 main classes of stochastic evolution models for Substitution-Insertion-Deletion (SID)

- **Two classes of reversible models** (Thorne et al. (1991) and Mc Guire et al. (2001))
  - Useful property for unrooted phylogenetic trees inference
  - Classical for substitution models
  - Strong theoretical constraints for the insertion-deletion process $\rightsquigarrow$ e.g. for the alignment of 2 sequences: reversibility $\Leftrightarrow$ equality of the insertion and deletion frequencies
- **One class of non-reversible model** (Rivas, 2005)
  - Development of an alignment algorithm DNA sequence "with gap"
  - Problem: a uniform deletion rate μ should not affect residue occurence probability $P(t)$ at time t

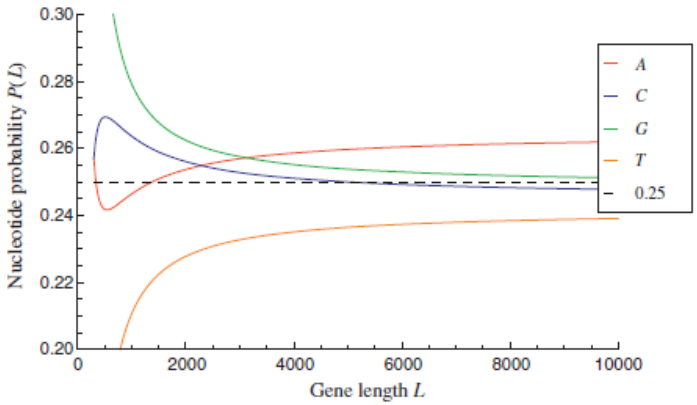
Evolution models for Substitution-Insertion-Deletion

- So far, evolution models are mostly designed for alignment, phylogeny,...

$$\begin{array}{cccccccc} - & T & G & T & - & C & - \\ G & - & C & - & A & C & A \end{array}$$

- Our motivations for developing the **LIIS model**

1. Focus on **sequence content** evolution (e.g. GC content $P_{G+C}(t)$, nucleotide $P_A(t)$, codon $P_{ATG}(t)$, etc ...)
2. Insertion, Deletion **independent of the Substitution** equilibrium distribution



Outline

IDIS model: Insertion Deletion Independent of Substitution

- Let $N_i(t)$ be the random variable for the **number of occurrences of nucleotide** i in the sequence, $1 \leq i \leq K$ ($K = 4$ for $\{A, C, G, T\}$).
- Random vector $(N_i(t))_{1 \leq i \leq 4}$ follows a **Birth-Death process** with instantaneous rates (at time t)

Birth	Death
$r_i n(t)$	$dn_i(t)$

- The sequence size $N(t) = \sum_{1 \leq i \leq 4} N_i(t)$ follows a **Birth-Death process** with linear growth
  - birth rate $\lambda(t) = \sum_i r_i \ n(t)$
  - death rate $\mu(t) = \sum_i dn_i(t) = dn(t)$

Then $E(N(t)) = n_0 e^{(\sum_i r_i - d)t}$ $\rightsquigarrow$ population dynamics (Malthus)

Average behaviour description of the sequence content

Assumption (1)

Insertion of nucleotide i: $n_i'(t) = r_i n(t) - d n_i(t)$

- Then $n'(t) = \left(\sum_i r_i - d\right) n(t)$ and $n(t) = n_0 e^{\left(\sum_i r_i - d\right)t}$

$$
\begin{aligned}
p_i'(t) &= \frac{\partial}{\partial t}\left(\frac{n_i(t)}{n(t)}\right) \\
&= \frac{1}{[n(t)]^2}(n(t)n_i'(t) - n'(t)n_i(t)) \\
&= r_i - \left(\sum_j r_j\right) p_i(t) \quad \text{(independent of deletion rate } d\text{)}
\end{aligned}
$$

- The vector $P(t) = (p_i(t))_{1 \leq i \leq 4}$ of the average nucleotide ratio satisfies the differential equation :

$$P'(t) = R - rP(t) \qquad \text{with } r = \sum_i r_i$$

Average behaviour description of the sequence content

Assumption (2)

Classical residue substitution model with stochastic matrix M
$\rightsquigarrow$ *site evolution i.i.d.*

- Then the sequence ratio satisfies the differential equation

$$P'(t) = (M - I) \cdot P(t)$$

where M is stochastic in column.

Average behaviour description of the sequence content

Assumption (3)

Insertion-deletion independent of substitution

- Then
$$\begin{aligned} P'(t) &= \underbrace{(M-I)\cdot P(t)}_{\text{Substitution}} + \underbrace{(-rP(t)+R)}_{\text{Insertion-Deletion}} \\ &= A\cdot P(t) + R \end{aligned}$$

  with $A = M - (1+r)I$, $r = \sum_i r_i$.

- Analytical solution?
  - Non-homogeneous matrix differential equation
  - Coefficients A and R independent of time t

$\Rightarrow$ General solution by the **method of variation parameters**:

$$P(t) = e^{A(t-t_0)}P(t_0) + e^{At}\left(\int_{t_0}^{t} e^{-Au}du\right)R$$

If the substitution matrix M is diagonalizable

Proposition

Whenever M **can be diagonalized with eigenvalues** $(\lambda_k)_{1\leq k\leq K}$*, then the residue average ratio $P(t)$ at time t defined by the IDIS model satisfies*

$$P(t) = Q \cdot D_1(t) \cdot Q^{-1} \cdot P(t_0) + Q \cdot D_2(t) \cdot Q^{-1} \cdot R$$

where

- *Q an associated eigenvector matrix of M, such that the k^{th} column of Q is an eigenvector for eigenvalue λ_k*
- $R = (r_i)_{1\leq i\leq K}$ *(tel que* $r = \sum_{1\leq i\leq K} r_i > 0$*)*
- $D_1(t) = Diag\left(\left(e^{-(r+1-\lambda_k)(t-t_0)}\right)_{1\leq k\leq K}\right)$
- $D_2(t) = Diag\left(\left(\frac{1}{r+1-\lambda_k}(1 - e^{-(r+1-\lambda_k)(t-t_0)})\right)_{1\leq k\leq K}\right)$
- $P(t_0) = (p_i(t_0))_{1\leq i\leq K}$

IDIS model

Proposition

- *Whenever* M **can be diagonalized with eigenvalues** $(\lambda_k)_{1\leq k\leq K}$, *let* Q **be an associated eigenvector matrix of** M
  (the k*th column of* Q *being an eigenvector for eigenvalue* λ_k*)*
- *For all* $1 \leq k \leq K$, *we define matrix* O_k *of size* $K \times K$ *by using the eigenvector matrix* Q *of* M,

$$\forall 1 \leq i,j \leq K,\ O_k[i,j] = Q[i,k] \cdot Q^{-1}[k,j]$$

$$P(t)=\left(\sum_{k=1}^{K} \frac{1}{r+1-\lambda_k} O_k\right) R + \sum_{k=1}^{K} O_k \cdot \left(P(t_0) - \frac{1}{r+1-\lambda_k} R\right) e^{-(r+1-\lambda_k)(t-t_0)}$$

with

- $R = (r_i)_{1\leq i\leq K}$
- $r = \sum_{1\leq i\leq K} r_i$ *is the total insertion rate*
- $P(t_0) = (p_i(t_0))_{1\leq i\leq K}$

IDIS model as a function of sequence length $l = n(t)$

- IDISL model

$$P(l) = \left(\sum_{k=1}^{K} \frac{1}{r+1-\lambda_k} O_k \right) R + \sum_{k=1}^{K} O_k \cdot \left(P(l_0) - \frac{1}{r+1-\lambda_k} R \right) \left(\frac{l}{l_0} \right)^{-\frac{r+1-\lambda_k}{r-d}}$$

with

  - $R = (r_i)_{1 \leq i \leq K}$
  - $r = \sum_{1 \leq i \leq K} r_i$ is the total insertion rate
  - d is the deletion rate
  - $P(t_0) = (p_i(t_0))_{1 \leq i \leq K}$
  - $(\lambda_k)_{1 \leq k \leq K}$ are the eigenvalues of M

- Indeed: from the Insertion-Deletion assumption:

$$n_i'(t) = r_i \times n(t) - d \times n_i(t)$$

Then $n(t) = n(t_0) e^{(r-d)(t-t_0)}$

and $e^{-(t-t_0)} = \left(\frac{l}{l_0} \right)^{-\frac{1}{r-d}}$ with $l_0 = n(t_0)$.

Properties

Proposition (Parameter scale)

When multiplying all the substitution–insertion–deletion parameters, i.e.

- *the non-diagonal elements* $[m_{ij}]_{i \neq j}$ *of the substitution matrix* M,
- *the insertion rates* $[r_i]_{1 \leq i \leq K}$
- *and the deletion rate* d,

by a scalar α, *then*

- *the average residue ratio* $P(t)$ *at time* t *satisfies*

$$P\left(\frac{t}{\alpha}; [\alpha m_{ij}], [\alpha r_i], \alpha d\right) = P(t; [m_{ij}], [r_i], d).$$

- *the average residue ratio* $P(l)$ *as a function of the sequence length* l *remains unchanged*

$$P(l; [\alpha m_{ij}]_{i \neq j}, [\alpha r_i], \alpha d) = P(l; [m_{ij}]_{i \neq j}, [r_i], d).$$

Properties

Proposition (Time step - Time inversion)

The residue average ratio $P(t)$ defined by

$$P(t)=\left(\sum_{k=1}^{K}\frac{1}{r+1-\lambda_k}O_k\right)R+\sum_{k=1}^{K}O_k\cdot\left(P(t_0)-\frac{1}{r+1-\lambda_k}R\right)e^{-(r+1-\lambda_k)(t-t_0)}$$

is obtained for all t_0, including $t_0 > t$.

- Similar result for $P(l)$ as a function of initial length l_0
- Allows us to go back in time
- Only constraint: $p_i(t) > 0$ i.e. all ratios $p_i(t)$ must be positive
  $\Rightarrow$ Depending on the evolution parameters, we cannot go further than the disparition of one residue $(\exists i, 1 \leq i \leq K, p_i(t) = 0)$

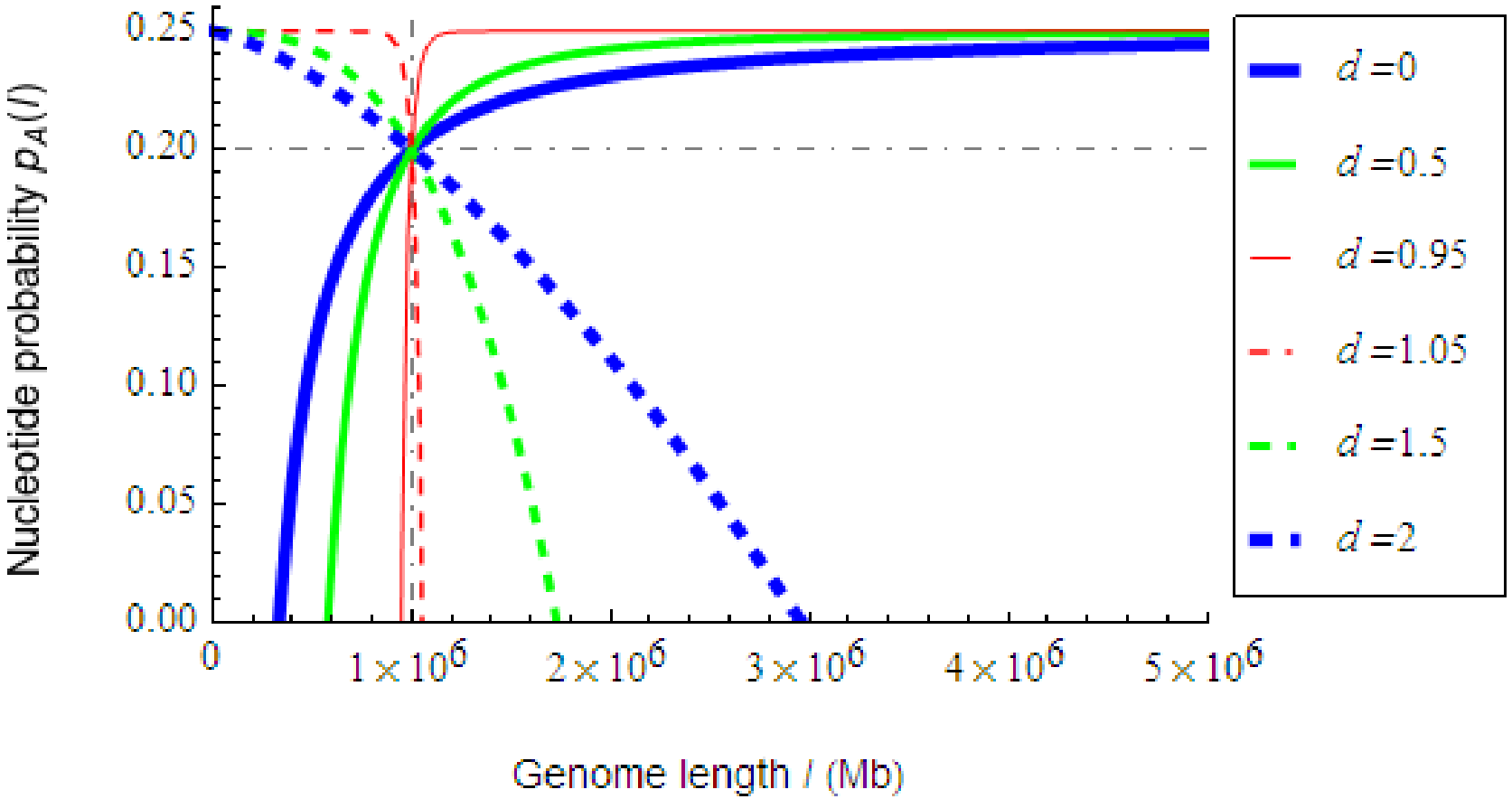
Properties

Proposition (Fixed point)

The LIIS model admits a fixed point which is reached by the sequence content vector after an infinite amount of time,

$$\begin{aligned}\lim_{t\to\infty} P(t) &= \lim_{l\to\infty, r-d>0} P(l) \\ &= \left(\sum_{k=1}^{K} \frac{1}{r+1-\lambda_k} O_k\right) R\end{aligned}$$

Inverse evolution and limits



- Initial conditions:
  $l_0 = 1$ Mb (vertical dot dashed line)
  $p_A(l_0) = 0.2$ (horizontal dot dashed line).
- HKY substitution matrix M with parameters: $\alpha = 0.2$, $\beta = 0.1$, $\pi_A = \pi_C = \pi_G = \pi_T = 0.25$

IDISL-SYM3 model

- SYM3 stochastic substitution matrix (Kimura, 1982)

$$M_{SYM3} = \begin{pmatrix} n & c & a & b \\ c & n & b & a \\ a & b & n & c \\ b & a & c & n \end{pmatrix}$$

where $n = 1 - (a + b + c)$.

- Eigenvalues:

$$\{\lambda_1 = 1 - 2(a + b), \lambda_2 = 1 - 2(a + c), \lambda_3 = 1 - 2(b + c), \lambda_4 = 1\}$$

- Eigenvectors:

$$\{v_1 = \{-1, -1, 1, 1\}, v_2 = \{1, -1, -1, 1\}, \\ v_3 = \{-1, 1, -1, 1\}, v_4 = \{1, 1, 1, 1\}\}.$$

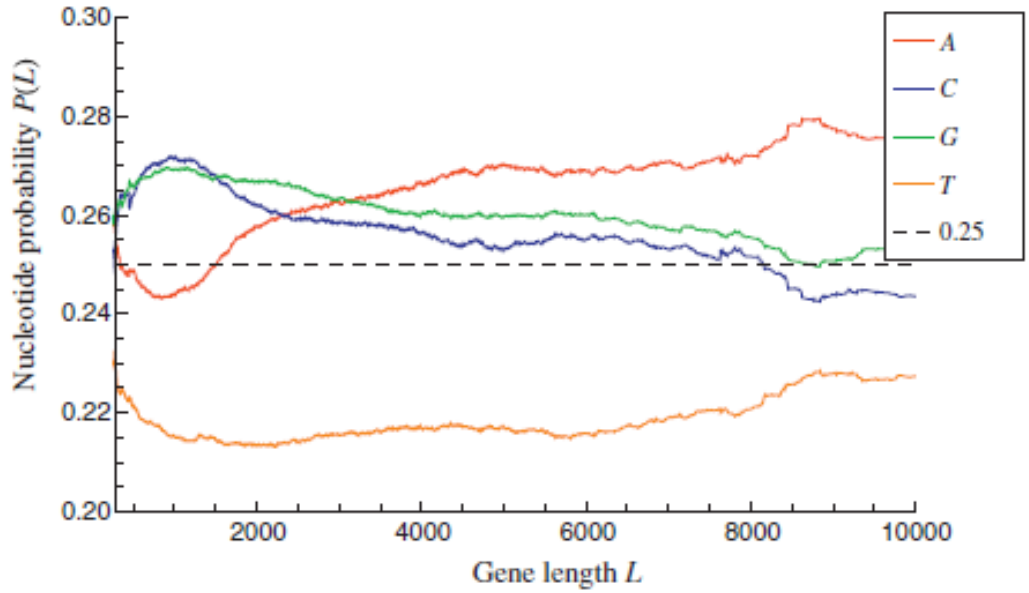
IDISL-SYM3 model

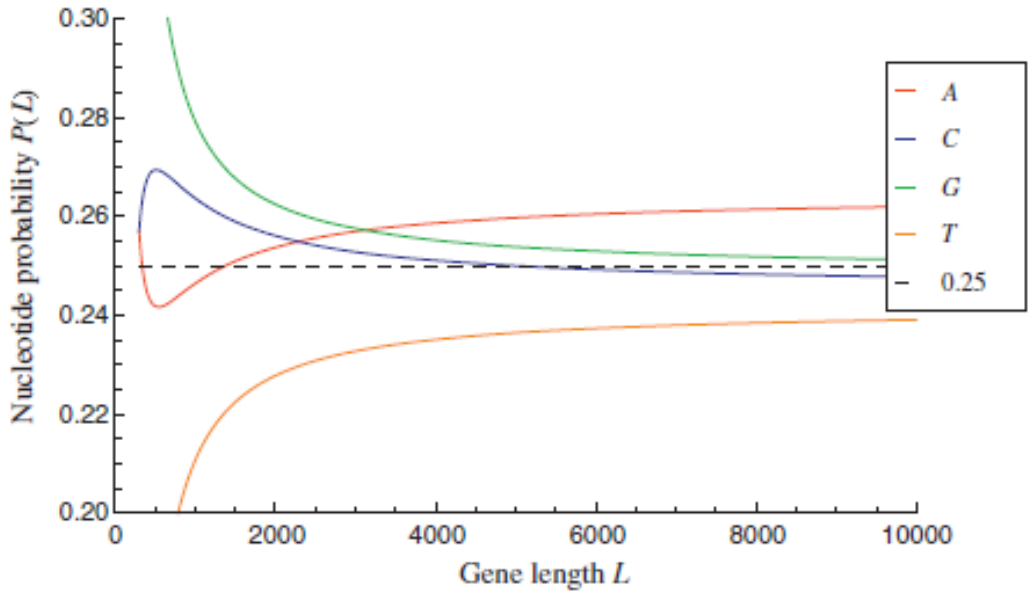
$$P(t) = \frac{1}{4} \begin{pmatrix} 1 + \frac{r_1}{z_1} + \frac{r_2}{z_2} + \frac{r_3}{z_3} + \left(p_1 - \frac{r_1}{z_1}\right) e^{-z_1 t} + \left(p_2 - \frac{r_2}{z_2}\right) e^{-z_2 t} + \left(p_3 - \frac{r_3}{z_3}\right) e^{-z_3 t} \\ 1 + \frac{r_1}{z_1} - \frac{r_2}{z_2} - \frac{r_3}{z_3} + \left(p_1 - \frac{r_1}{z_1}\right) e^{-z_1 t} - \left(p_2 - \frac{r_2}{z_2}\right) e^{-z_2 t} - \left(p_3 - \frac{r_3}{z_3}\right) e^{-z_3 t} \\ 1 - \frac{r_1}{z_1} - \frac{r_2}{z_2} + \frac{r_3}{z_3} - \left(p_1 - \frac{r_1}{z_1}\right) e^{-z_1 t} - \left(p_2 - \frac{r_2}{z_2}\right) e^{-z_2 t} + \left(p_3 - \frac{r_3}{z_3}\right) e^{-z_3 t} \\ 1 - \frac{r_1}{z_1} + \frac{r_2}{z_2} - \frac{r_3}{z_3} - \left(p_1 - \frac{r_1}{z_1}\right) e^{-z_1 t} + \left(p_2 - \frac{r_2}{z_2}\right) e^{-z_2 t} - \left(p_3 - \frac{r_3}{z_3}\right) e^{-z_3 t} \end{pmatrix}$$

where

- $z_k = r + 1 - \lambda_k$ for all $k = 1, 2, 3$
- $r = r_A + r_C + r_G + r_T$;
- $r_1 = r_A + r_C - r_G - r_T$, $r_2 = r_A - r_C - r_G + r_T$, $r_3 = r_A - r_C + r_G - r_T$;
- $p_1 = P_A(0) + P_C(0) - P_G(0) - P_T(0)$, $p_2 = P_A(0) - P_C(0) - P_G(0) + P_T(0)$, $p_3 = P_A(0) - P_C(0) + P_G(0) - P_T(0)$.

Illustration on a human gene dataset (NCBI, built 36 version 3)





IDISL-HKY model

- HKY stochastic substitution matrix (Hasegawa et al., 1985)

$$M_{HKY} = \begin{pmatrix} n_A & \beta\pi_A & \alpha\pi_A & \beta\pi_A \\ \beta\pi_C & n_C & \beta\pi_C & \alpha\pi_C \\ \alpha\pi_G & \beta\pi_G & n_G & \beta\pi_G \\ \beta\pi_T & \alpha\pi_T & \beta\pi_T & n_T \end{pmatrix}$$

(here stochastic in column)

- Eigenvalues:

$$\{\lambda_1 = 1-\beta, \lambda_2 = 1-\alpha\pi_R-\beta\pi_Y, \lambda_3 = 1-\alpha\pi_Y-\beta\pi_R, \lambda_4 = 1\}$$

- Eigenvectors:

$$\left\{ v_1 = \left\{ -\frac{\pi_Y\pi_A}{\pi_R\pi_T}, \frac{\pi_C}{\pi_T}, -\frac{\pi_Y\pi_G}{\pi_R\pi_T}, 1 \right\}, v_2 = \{-1,0,1,0\}, \right.$$
$$\left. v_3 = \{0,-1,0,1\}, v_4 = \left\{ \frac{\pi_A}{\pi_T}, \frac{\pi_C}{\pi_T}, \frac{\pi_G}{\pi_T}, 1 \right\} \right\}.$$

IDISL-HKY model

- For all $l = n(t)$ and $l_0 = n(t_0)$

$$P(l) = P_K + k_{1,R,Y} \begin{pmatrix} \frac{\pi_A}{\pi_R} \\ -\frac{\pi_C}{\pi_Y} \\ \frac{\pi_G}{\pi_R} \\ -\frac{\pi_T}{\pi_Y} \end{pmatrix} \left(\frac{l}{l_0}\right)^{-\frac{\mu_1}{r-d}} + \begin{pmatrix} k_{2,A,G} \left(\frac{l}{l_0}\right)^{-\frac{\mu_2}{r-d}} \\ k_{3,C,T} \left(\frac{l}{l_0}\right)^{-\frac{\mu_3}{r-d}} \\ -k_{2,A,G} \left(\frac{l}{l_0}\right)^{-\frac{\mu_2}{r-d}} \\ -k_{3,C,T} \left(\frac{l}{l_0}\right)^{-\frac{\mu_3}{r-d}} \end{pmatrix}$$

where P_K is a constant

and for all $1 \leq i \leq 3$, $j \in \{A, C, R\}$, $k \in \{G, T, Y\}$,

$$k_{i,j,k} = \frac{\pi_j(r_k - \mu_i P_k(l_0)) - \pi_k(r_j - \mu_i P_j(l_0))}{(\pi_j + \pi_k)\mu_i},$$

$$\mu_i = r + 1 - \lambda_i$$

IDISL-HKY model

P_K is the limit when $t \to \infty$ and insertion is predominant $(r - d > 0)$

$$P_K = \lim_{I \to \infty, r-d>0} P(I) = \begin{pmatrix} \pi_A \left(1 - \frac{l_{1,R,Y}}{\pi_R}\right) - l_{2,A,G} \\ \pi_C \left(1 + \frac{l_{1,R,Y}}{\pi_Y}\right) - l_{3,C,T} \\ \pi_G \left(1 - \frac{l_{1,R,Y}}{\pi_R}\right) + l_{2,A,G} \\ \pi_T \left(1 + \frac{l_{1,R,Y}}{\pi_Y}\right) + l_{3,C,T} \end{pmatrix}$$

with, for all $1 \leq i \leq 3$, $j \in \{A, C, R\}$, $k \in \{G, T, Y\}$,

$$l_{i,j,k} = \frac{\pi_j r_k - \pi_k r_j}{(\pi_j + \pi_k)\,\mu_i}.$$

Outline

1. Insertion Deletion Independent of Substitution (IDIS) model
2. Application to GC content
3. Limited Insertion Independent of Substitution (LIIS) model

CG content modeling

- Data
  - The length (number of nucleotides) and the GC content of bacterial genome were extracted from the **NCBI website** `www.ncbi.nlm.nih.govgenomeslproks.cgi`.
  - Bacterial genomes are classified according to their taxonomic group and their anaerobic/non-aerobic property
  - Groups with more than 30 genomes are included.

- Assumptions:

(H1) The residue frequencies in bacterial genomes are still **transient**.

(H2) The species in a taxonomic group are subject to a **common evolution process** governed by the mutation parameters $([m_{ij}]_{i \neq j}, [r_i], d)$ up to a multiplicative constant, each species having a specific evolution speed.

IDISL-HKY model for GC content

- Assumptions (G=C and A=T)

$$\begin{cases} p_C(l_0) = p_G(l_0),\ p_A(l_0) = p_T(l_0), \\ \pi_C = \pi_G,\ \pi_A = \pi_T, \\ r_C = r_G,\ r_A = r_T. \end{cases}$$

- With $\kappa = \frac{\alpha+\beta}{2r}$, the GC content $p_{G+C}(l) = p_C(l) + p_G(l)$ reads:

$$p_{G+C}(l) = 2\left(\frac{\frac{r_C}{r} + \kappa\pi_C}{1+\kappa}\right) + 2\left(p_C(l_0) - \frac{\frac{r_C}{r} + \kappa\pi_C}{1+\kappa}\right)\left(\frac{l}{l_0}\right)^{-\frac{1+\kappa}{1-\frac{d}{r}}}$$

- Then $p_{G+C}(l)$ is a polynom:

$$p_{G+C}(l) = a + b\left(\frac{l}{l_0}\right)^{-c}$$

Domain of definition

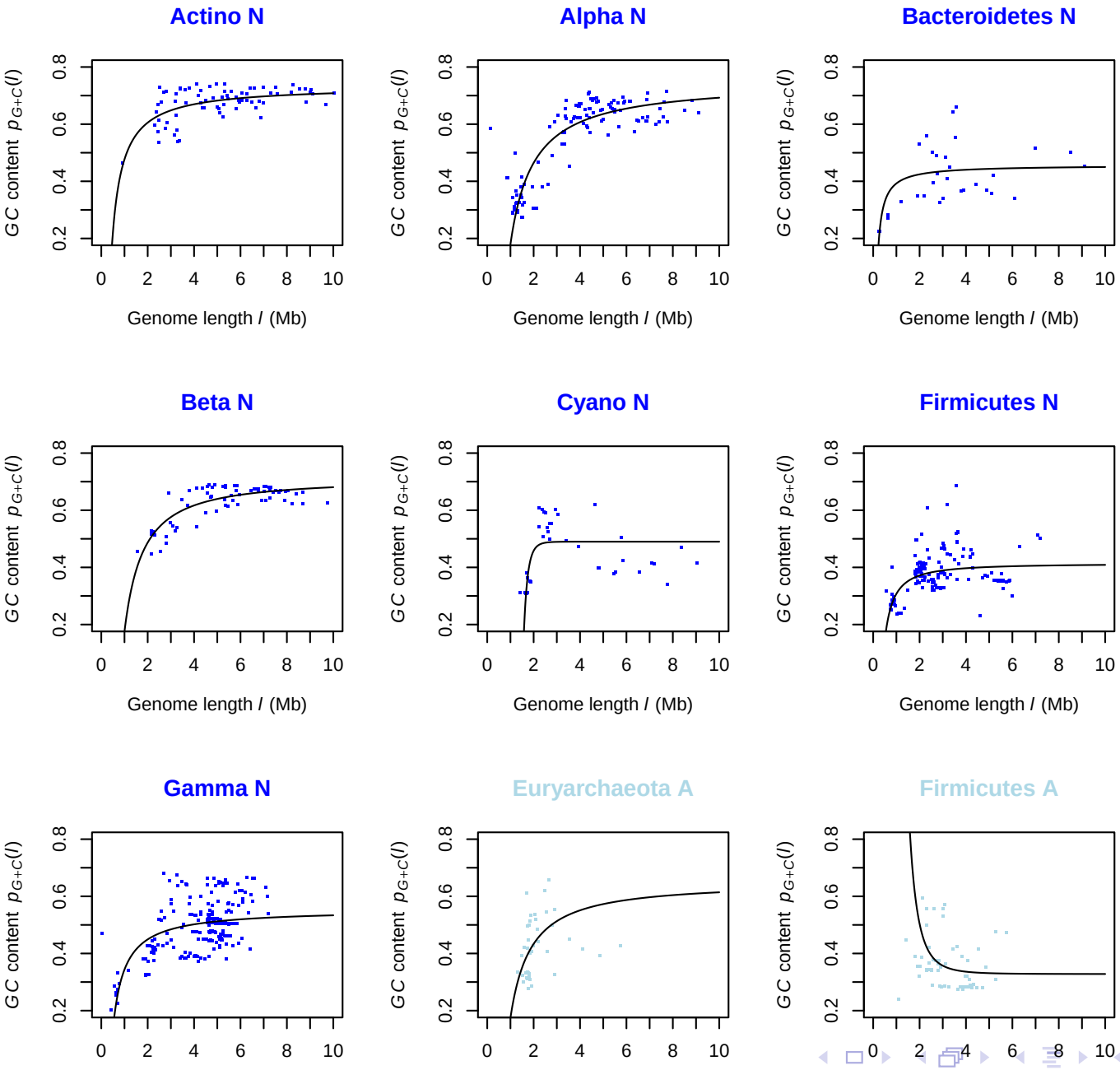
- The GC content model parameters $(r_c, d, \alpha, \beta, p_C(l_o))$ satisfy

$$\begin{cases} d = -\frac{2r(1-c)+\alpha+\beta}{2c} \\ r_C = r_G = \frac{a}{2}r + \frac{\alpha+\beta}{2}\left(\frac{a}{2} - \pi_C\right) \\ p_C(l_0) = p_G(l_0) = \frac{1}{2}(a+b) \end{cases} .$$

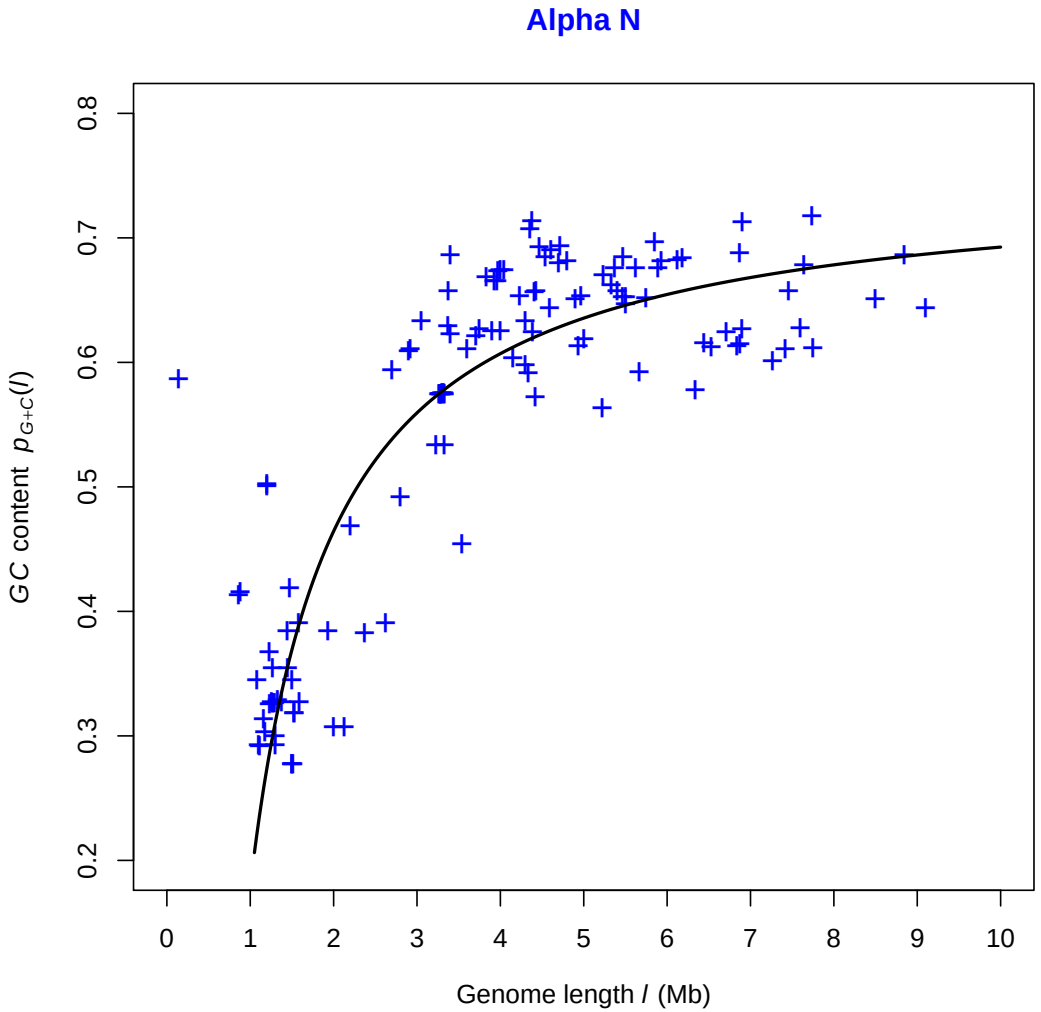
- Leading to the following domain of definition

$$\begin{cases} 0 \leq a \leq 1 \\ -a \leq b \leq 1-a \\ c \geq 1 \text{ if } r > d \\ c < 0 \text{ if } r < d \end{cases} .$$

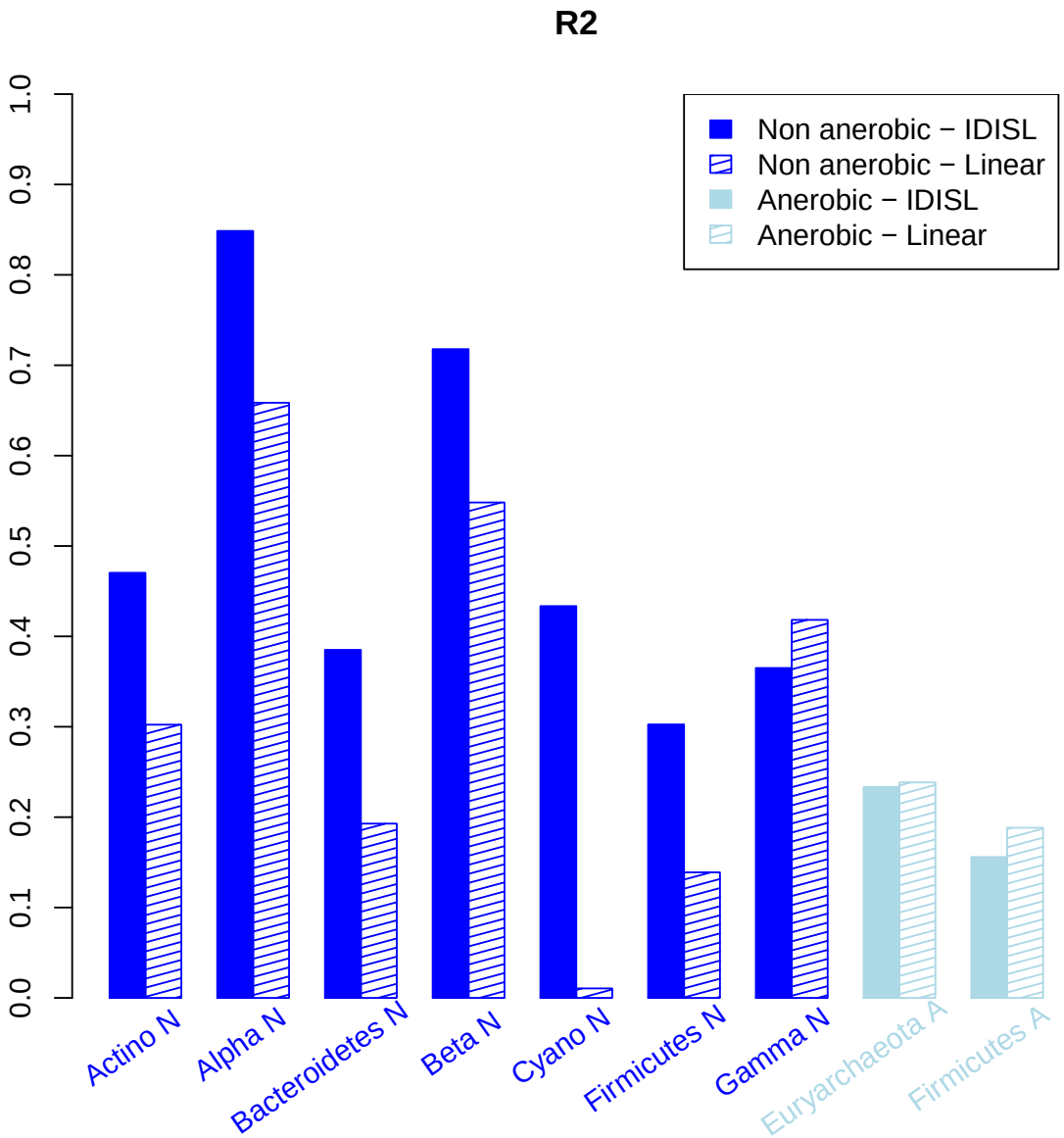
Best fit curve $\hat{p}_{G+C}(l)$ with the IDISL-HKY model



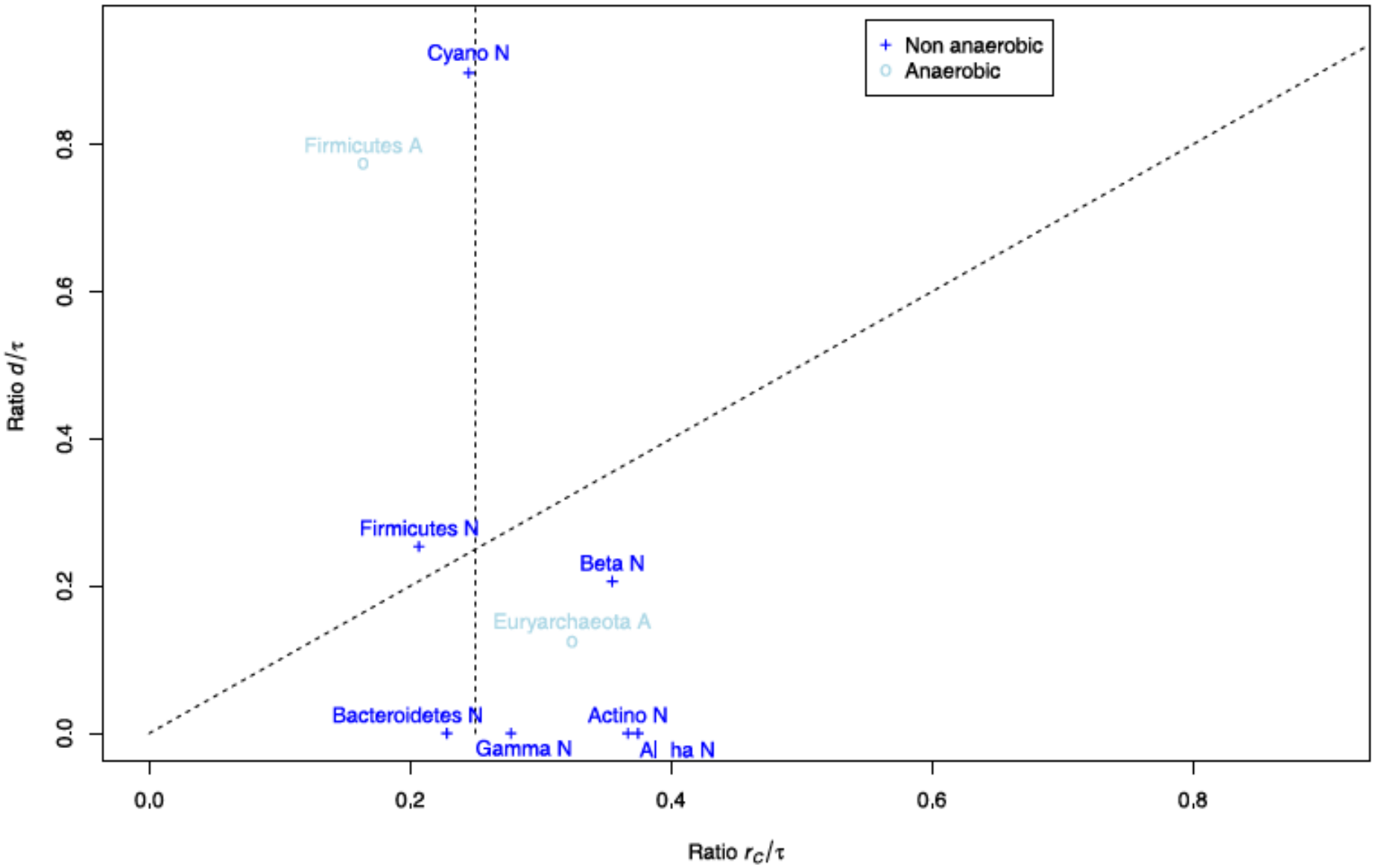
Best fit curve $\hat{p}_{G+C}(l)$ with the IDISL-HKY model



GC content as a function of the genome length /



Comparison: deletion d versus GC insertion rate r_C



Outline

1. Insertion Deletion Independent of Substitution (IDIS) model
2. Application to GC content
3. Limited Insertion Independent of Substitution (LIIS) model

Limited Insertion Independent of Substitution (LIIS)

Assumption

$$\textit{Limited Insertion assumption: } n_i'(t) = r_i \left(1 - \frac{n(t)}{n_{\max}}\right) n(t)$$

- Then
$$\begin{aligned} P'(t) &= \underbrace{(M - I) \cdot P(t)}_{\text{Substitution}} + \underbrace{\theta(t)\,(-rP(t) + R)}_{\text{Insertion-Deletion}} \\ &= A(t) \cdot P(t) + \theta(t) R \end{aligned}$$

with $\theta(t) = 1 - \frac{n(t)}{n_{max}} = 1 - \frac{\tau}{\tau + (1-\tau)e^{-r(t-t_0)}}$, $\tau = \frac{n_0}{n_{\max}}$

$A(t) = M - (1 + r\theta(t))I$ and $r = \sum_i r_i$.

- Non-constant coefficients $\Rightarrow$ Analytical solution?
YES because, for all $s, t \geq 0$, matrices $A(t)$ and $A(s)$ commute (See e.g. Teschl, 2012)
(satisfied here as $\theta(t)$ in matrix $A(t) = M - (1 + r\theta(t))I$ is in the diagonal)
- Then, for all $t \geq 0$, for all initial time $t_0 \geq 0$,
$$P(t; t_0, P(t_0)) = e^{\left(\int_{t_0}^{t} A(u)du\right)} \cdot P(t_0) + \int_{t_0}^{t} \theta(s) e^{\left(\int_{s}^{t} A(u)du\right)} \cdot R ds.$$

Limited Insertion Independent of Substitution (LIIS)

Proposition

$$P(t; t_0, P(t_0)) = \sum_{k=1}^{K} O_k \cdot [d_1(t; k, t_0)P(t_0) + d_2(t; k, t_0)R]$$

where

$$d_1(t; k, t_0) = \left(\tau + (1-\tau)\, e^{-r(t-t_0)}\right) e^{-(1-\lambda_k)(t-t_0)}$$

$$d_2(t; k, t_0) = \frac{1}{r}\left[1 - \left(\tau + (1-\tau)e^{-r(t-t_0)}\right)\left(e^{-(1-\lambda_k)(t-t_0)}\right.\right.$$

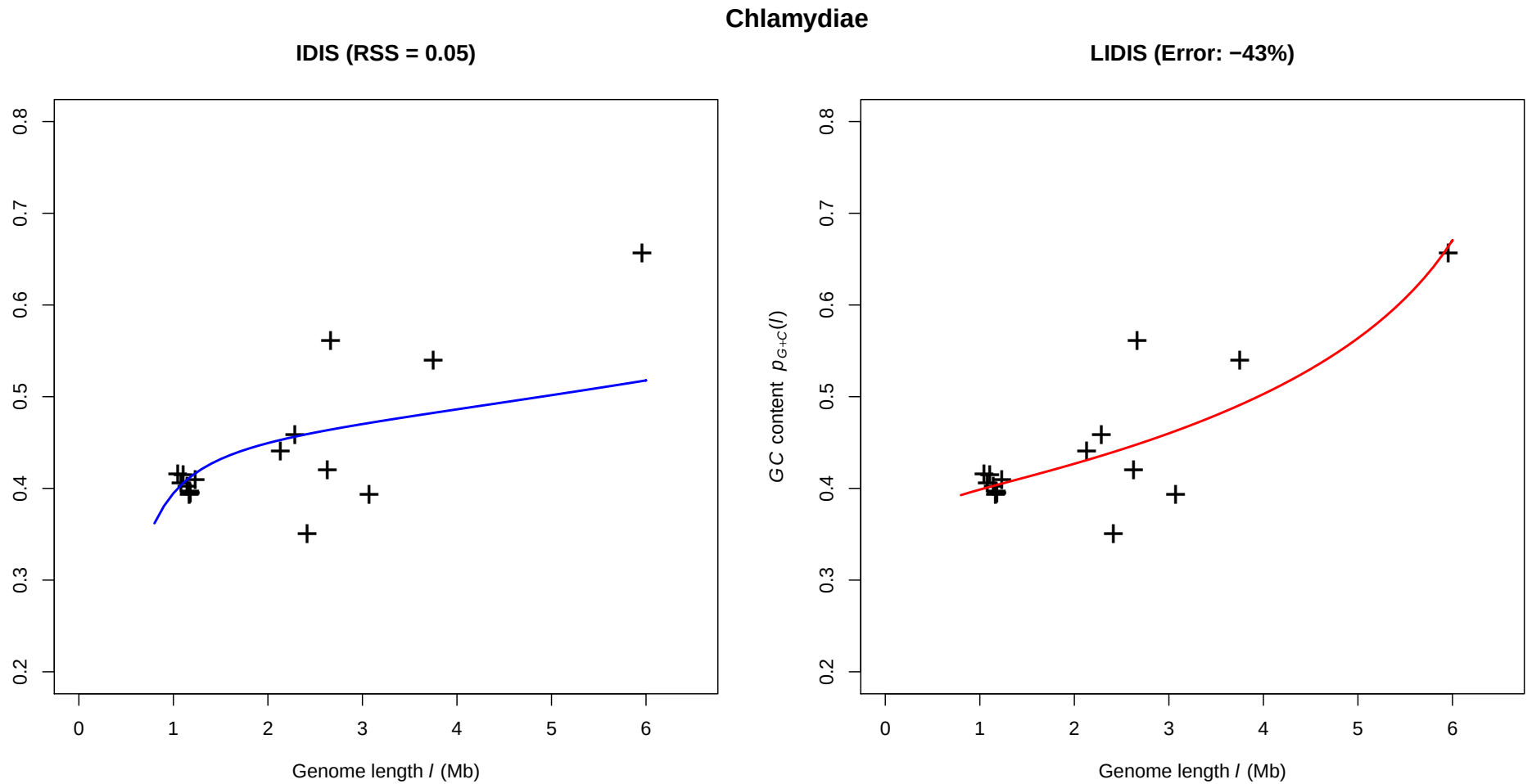
$$\left.\left. + \frac{1-\lambda_k}{(\tau-1)\,(1-\lambda_k+r)}\left(e^{-(1-\lambda_k)(t-t_0)}\,{}_2\mathcal{F}_1(k,1) - e^{r(t-t_0)}\,{}_2\mathcal{F}_1(k, e^{r(t-t_0)})\right)\right)\right]$$

$\tau = \frac{n_0}{n_{\max}}$ *and for all* (k, x), $\forall 1 \leq k \leq 4$ *and* $\forall x \geq 0$,

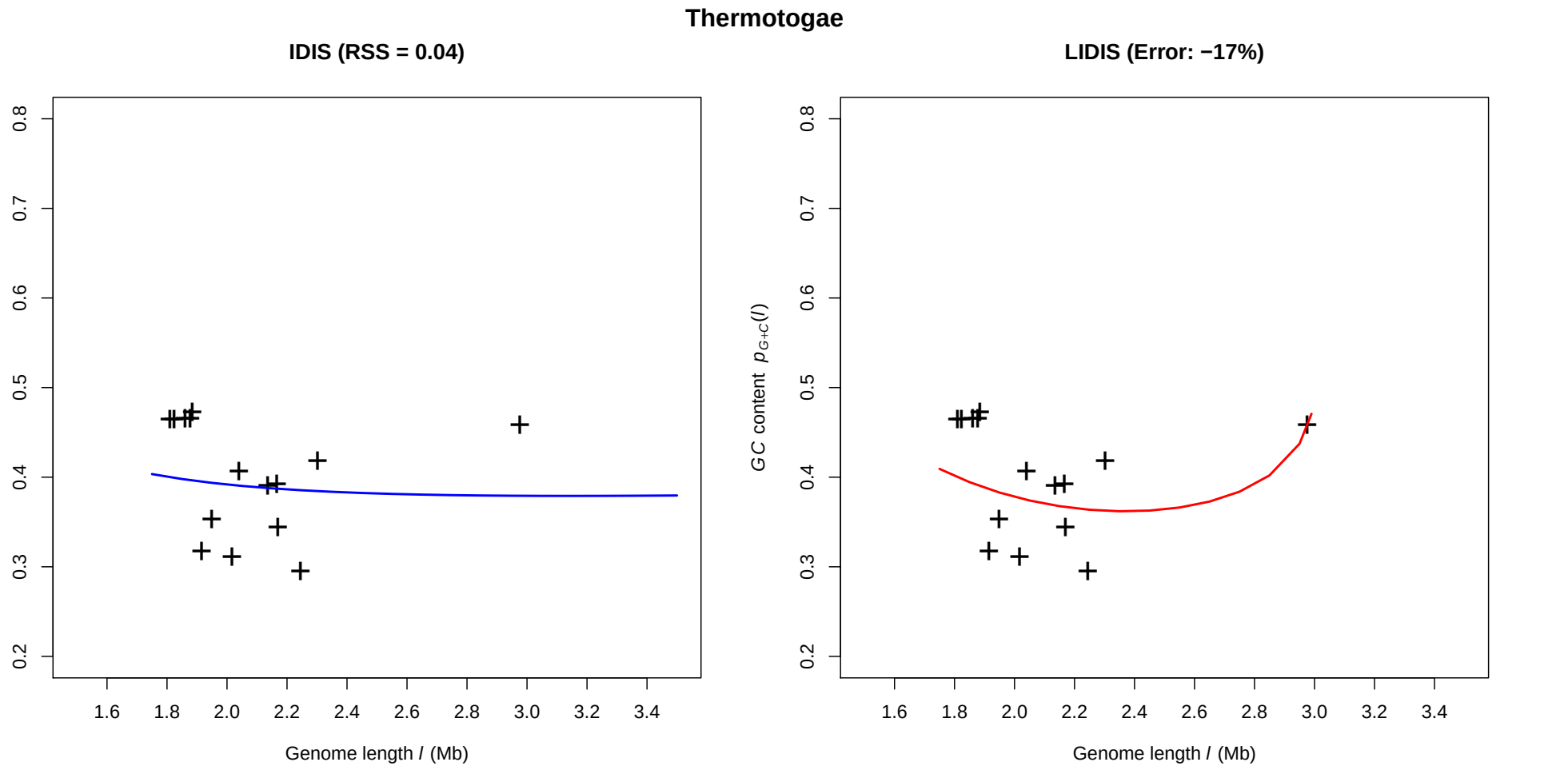
${}_2\mathcal{F}_1(k, x)$ *is the Gauss hypergeometric function:*

$${}_2\mathcal{F}_1(k, x) = H2F1\left[1, 1 + \frac{1-\lambda_k}{r}, 2 + \frac{1-\lambda_k}{r}, \frac{\tau}{\tau-1}x\right].$$

Best fit curve $\hat{p}_{G+C}(l)$ with the IDIS versus LIIS model



Best fit curve $\hat{p}_{G+C}(l)$ with the IDIS versus LIIS model



To conclude

- IDIS model
  - Insertion-Deletion independent of Substitution
  - Description of the average behaviour of the sequence content in 2 cases :
    - Linear growth and deletion
    - Limited growth
  - Mathematical properties : fixed point, time scale, time inversion, ...
  - Allows the description of sequence content evolution, e.g. GC content, codon model, ...
  - Software on line: http://icube-bioinfo.u-strasbg.fr/webMathematica/GETEC
- Ongoing work
  - Add limited deletion (LIDIS model)
  - Stochastic framework

References


Lèbre S., Michel C.J. (2013). A new evolution model for Limited Insertion Independent of Substitution. Mathematical Biosciences 245, 137-147.


Lèbre S., Michel C.J. (2012). An evolution model for sequence length based on residue insertion-deletion independent of substitution: an application to the GC content in bacterial genomes. Bull. Math. Biol. 74, 1764-1788.


Lèbre S., Michel C.J. (2010). An evolution model for residue insertion-deletion independent from substitution. J. Comput. Biol. Chem. 34, 259-267.