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

S. Lèbre, C. Michel

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

Substitution-Insertion-Deletion models

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

1) First approach by Thorne et al. (1991) : "TKF91" model

- Birth-Death process

  $\rightsquigarrow$ Explicit birth rate λ (insertion) and death rate μ (deletion)

$$\begin{array}{cccccccccc} \bullet & - & - & B\star & - & B\star & - & - & - & B\star\ldots \\ \bullet & B\star & B\star & - & B\star & B\star & B\star & B\star & B\star & B\star\ldots \end{array}$$

- Substitution process definied by a stochastic matrix *M* applied to both sequences, conditional on the insertion process

- Extension to "long indels" insertion
  - Thorne et al. (1992), Metzler (2003), Miklós et al. (2004)
  - Review: Miklós et al. (2009)

Substitution-Insertion-Deletion over the last 20 years

2) Extended substitution model “substitution + gap”

- Introduced by McGuire et al. (2001)
- Extention of the “F84” model by Felsenstein and Churchill (1996): addition of a 5th residue representing a “gap”
  - 5th column: residue deletion
  - 5th line: residue insertion

- Extended substitution rate matrix:

$$R = \begin{pmatrix} * & \gamma\pi_C & \frac{\gamma\pi_G}{\pi_R} + \gamma\pi_G & \gamma\pi_T & \gamma\pi_{-} \\ \gamma\pi_A & * & \gamma\pi_G & \frac{\gamma\pi_T}{\pi_Y} + \gamma\pi_T & \gamma\pi_{-} \\ \frac{\gamma\pi_A}{\pi_R} + \gamma\pi_A & \gamma\pi_C & * & \gamma\pi_T & \gamma\pi_{-} \\ \gamma\pi_A & \frac{\gamma\pi_C}{\pi_Y} + \gamma\pi_C & \gamma\pi_G & * & \gamma\pi_{-} \\ \gamma\pi_A & \gamma\pi_C & \gamma\pi_G & \gamma\pi_T & * \end{pmatrix}$$

$$P(t) = (P_A(t), P_C(t), P_G(t), P_T(t)) = P(0).exp\,(Rt)$$

Substitution-Insertion-Deletion over the last 20 years

- So far: reversible models
  - 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

Substitution-Insertion-Deletion over the last 20 years

3) Generalisation of the model by McGuire to a non-reversible model (Rivas (2005), Rivas and Eddy (2008))

- alphabet of size K
- addition of explicit parameters for insertion $(p_1, p_2, \cdots, p_K)$ and deletion μ

$$Q = \begin{pmatrix} * & - & - & - & \mu \\ - & * & - & - & \mu \\ - & - & * & - & \mu \\ - & - & - & * & \mu \\ \lambda p_1 & \lambda p_2 & \cdots & \lambda p_K & 1-\lambda \end{pmatrix}$$

- Analytical expression of the residue substitution probabitility over a period of time $[0, t[$ but only in the particular case where
$$(p_k)_{1 \leq k \leq K} \propto (\pi_k)_{1 \leq k \leq K}.$$
- 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

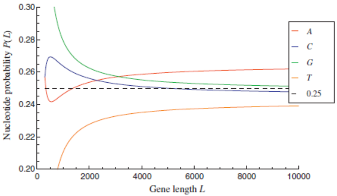
Evolution models for Substitution-Insertion-Deletion

- 3 main classes of stochastic SID models
  - **Thorne et al. (1991)** (TKF91) and extensions: Birth-Death process
  - **Mc Guire et al. (2001)**: extended substitution matrix
  - **Rivas (2005) and Rivas and Eddy (2008)**: non-reversible model

```
- T G T - C -
G - C - A C A
```

$\Rightarrow$ Designed for alignment, phylogeny,...

- Two aims
  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

1. Insertion Deletion Independent of Substitution (IDIS) model

2. Application to GC content

3. Limited Insertion Independent of Substitution (LIIS) model

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
$r_i n(t)$	$d n_i(t)$

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 $E(N(t)) = n_0 e^{\left(\sum_i r_i - d\right)t}$ $\rightsquigarrow$ 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 $$\begin{aligned} n'(t) &= \left(\sum_i r_i - d\right) n(t) \quad \text{and} \quad n(t) = n_0 e^{\left(\sum_i r_i - d\right)t} \\ p_i'(t) &= \frac{\partial}{\partial t}\left(\frac{n_i(t)}{n(t)}\right) \\ &= r_i - \left(\sum_j r_j\right) p_i(t) \;\; \text{(independent of deletion rate } d\text{)} \end{aligned}$$
- Vector $P(t) = (p_i(t))_{1 \leq i \leq 4}$ of the average nucleotide ratio satisfies the differential equation :

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

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?

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

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

IDIS model

Proposition

- *When* M **can be diagonalized with real eigenvalues** $(\lambda_k)_{1\leq k\leq 4}$, *let* Q **be an associated eigenvector matrix** *of* M *(the* k*th column of* Q *being an eigenvector for eigenvalue* λ_k*)*
- *For all* $1 \leq k \leq 4$*, we define matrix* O_k *of size* 4×4 *by using the eigenvector matrix* Q *of* M*,*

$$\forall 1 \leq i,j \leq 4,\ O_k[i,j] = Q[i,k] \cdot Q^{-1}[k,j]$$

$$P(t)=\left(\sum_{k=1}^{4} \frac{1}{r+1-\lambda_k} O_k\right) R + \sum_{k=1}^{4} O_k \cdot \left(P(t_0) - \frac{1}{r+1-\lambda_k} R\right) e^{-(r+1-\lambda_k)(t-t_0)}$$

with

- $R = (r_i)_{1\leq i\leq 4}$
- $r = \sum_{1\leq i\leq 4} r_i$ *is the total insertion rate*
- $P(t_0) = (p_i(t_0))_{1\leq i\leq 4}$

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

- IDISL model

$$P(l) = \left(\sum_{k=1}^{4} \frac{1}{r+1-\lambda_k} O_k\right) R + \sum_{k=1}^{4} O_k \cdot \left(P(l_0) - \frac{1}{r+1-\lambda_k} R\right) \left(\frac{l}{l_0}\right)^{-\frac{r+1-\lambda_k}{r-d}}$$

with

- $R = (r_i)_{1 \leq i \leq 4}$
- $r = \sum_{1 \leq i \leq 4} r_i$ is the total insertion rate
- d is the deletion rate
- $P(t_0) = (p_i(t_0))_{1 \leq i \leq 4}$
- $(\lambda_k)_{1 \leq k \leq 4}$ 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 $l_0 = n(t_0)$.

IDISL-HKY model

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

$$\begin{aligned} k_{i,j,k} &= \frac{\pi_j(r_k - \mu_i P_k(l_0)) - \pi_k(r_j - \mu_i P_j(l_0))}{(\pi_j + \pi_k)\mu_i}, \\ \mu_i &= r + 1 - \lambda_i \end{aligned}$$

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.
- Assumption: The residue frequencies in bacterial genomes are still **transient**.

IDISL-HKY model for GC content

- Assumptions (G=C and A=T)

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

Proposition (GC content)

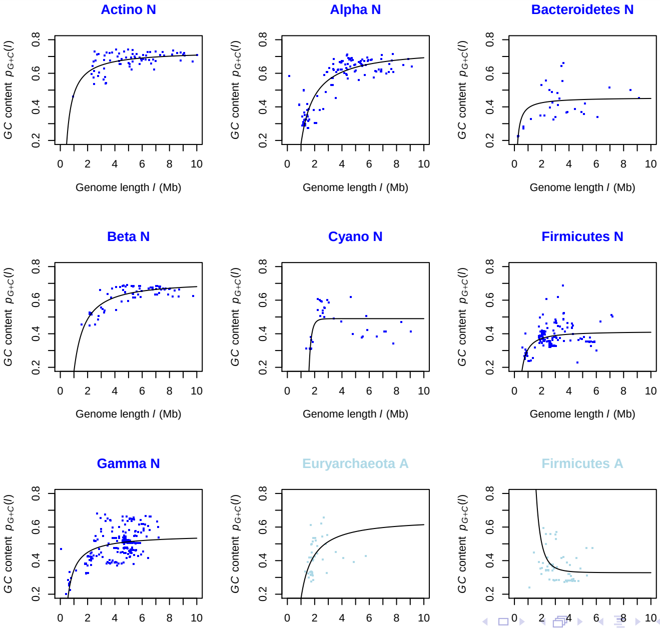
With $\kappa = \frac{\alpha+\beta}{2r}$*, the GC content* $p_{G+C}(l) = p_C(l) + p_G(l)$ *reads:*

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

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

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

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



Actino N

Alpha N

Bacteroidetes N

Beta N

Cyano N

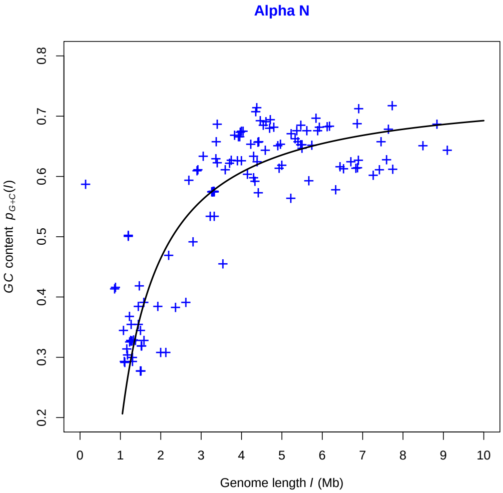
Firmicutes N

Gamma N

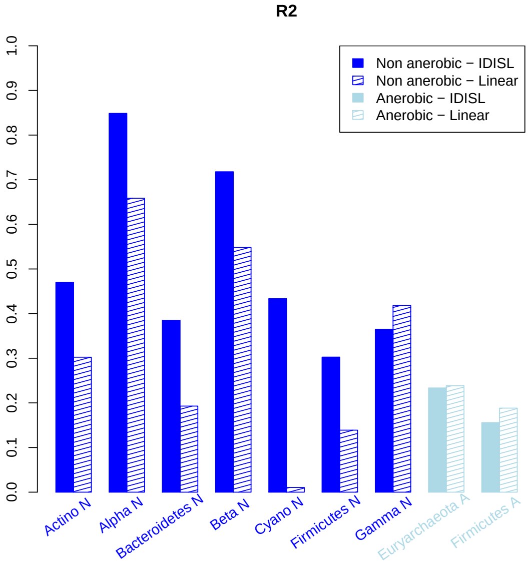
Euryarchaeota A

Firmicutes A

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



GC content as a function of the genome length I



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 from Substitution (LIIS)

Assumption

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

Proposition

$$P(t;t_0,P(t_0)) = \sum_{k=1}^{4} O_k \cdot [d_1(t;k,t_0)P(t_0) + d_2(t;k,t_0)R]$$

where

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

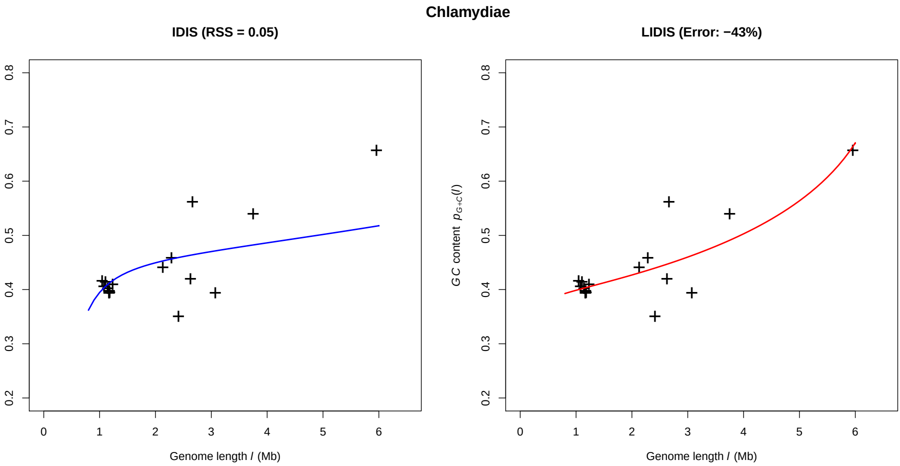
$$d_2(t;k,t_0) = \frac{1}{r}\left[1 - \left(\tau + (1-\tau)e^{-r(t-t_0)}\right)\left(e^{-(1-\lambda_k)(t-t_0)} + \frac{1-\lambda_k}{(\tau-1)(1-\lambda_k+r)}\left(e^{-(1-\lambda_k)(t-t_0)}{}_2\mathcal{F}_1(k,1) - e^{r(t-t_0)}{}_2\mathcal{F}_1(k,e^{r(t-t_0)})\right)\right)\right]$$

$\tau = \frac{n_0}{n_{\max}}$ *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

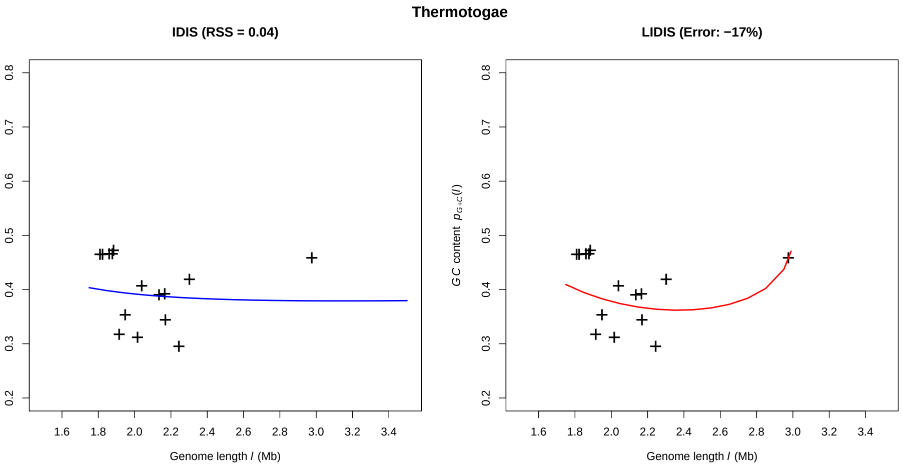


Chlamydiae

IDIS (RSS = 0.05)

LIDIS (Error: −43%)

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://lsiit-bioinfo.u-strasbg.fr/webMathematica/GETEC/Accueil.jsp
- Ongoing work
  - Add limited deletion (LIDIS model)
  - Stochastic framework

References

- Lèbre S., Michel C.J. (To appear). A new evolution model for Limited Insertion Independent of Substitution. Mathematical Biosciences

- 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.

- Thorne J.L., Kishino H., Felsenstein J. (1991). An evolutionary model for maximum likelihood alignment of DNA sequences. J. Mol. Evol. 33, 114-124.

- McGuire G., Denham M.C., Balding D.J. (2001). Models of sequence evolution for DNA sequences containing gaps. Mol. Biol. Evol. 18, 481-490.

- Rivas E. (2005). Evolutionary models for insertions and deletions in a probabilistic modeling framework. BMC Bioinformatics 6, 63.

References

Miklós I., Novák A., Satija R., Lyngsø R., Hein J. (2009). Stochastic models of sequence evolution including insertion-deletion events. Stat. Methods Med. Res. 18, 453-485.

Malthus T.R. (2000). An essay on the principle of population. Library of Economics, Liberty, Fund, Inc.

Verhulst P.-F. (1838). Notice sur la loi que la population poursuit dans son accroissement. Correspondance mathématique et physique 10, 113-121.