Propriétés presque sûres de martingales vectorielles.

Systèmes itératifs. Systèmes dynamiques dans l'analyse de séquences biologiques.

Pré-Soutenance - 14 Octobre 2004

Peggy Cénac - INRIA - Université Paul Sabatier

http ://www-rocq.inria.fr/~cenac/

Peggy.Cenac@inria.fr

Propriétés presque sûres de Martingales vectorielles.

- *Régression linéaire stable :*

Pour tout $n \geq 1$,

$$X_{n+1} = \theta^t \Phi_n + \varepsilon_{n+1},$$

où $\theta \in \mathbb{R}^d$ est le paramètre inconnu.

Exemple 1 : Modèle autorégressif d'ordre d $AR(d)$

$$X_{n+1} = \sum_{k=1}^{d} \theta_k X_{n-k+1} + \varepsilon_{n+1}$$

Exemple 2 : Modèle autorégressif fonctionnel d'ordre d

$$X_{n+1} = \sum_{k=1}^{d} \theta_k f_k(X_{n-k+1}) + \varepsilon_{n+1}$$

▶ *Erreurs cumulées*

▷ Erreur de prédiction

$$C_n(p) = \sum_{k=0}^{n-1} (X_{k+1} - \widehat{\theta}_k^t \Phi_k)^{2p}$$

▷ Erreur d'estimation

$$G_n(p) = \sum_{k=1}^{n} k^{p-1} \|\widehat{\theta}_k - \theta\|^{2p}$$

▶ *Résultats :*

▷ Estimateur des moindres carrés pondérés

$$\widehat{\theta}_n = S_{n-1}^{-1}(\alpha) \sum_{k=1}^{n} \alpha_{k-1} \Phi_{k-1} X_k$$

$$S_n(\alpha) = \sum_{k=0}^{n} \alpha_k \Phi_k \Phi_k^t + S$$

▷ Choix de la pondération

$$0 \leq \alpha_n \leq 1, \qquad \sum_{n=0}^{\infty} \alpha_n f_n(\alpha) < \infty \qquad \text{p.s.}$$

$$f_n(\alpha) = \alpha_n \Phi_n^t S_n^{-1}(\alpha) \Phi_n$$

▷ Supposons que

$$\mathbb{E}[\varepsilon_{n+1}^{2p} \mid \mathcal{F}_n] = \sigma(2p) \quad \text{et} \quad \mathbb{E}[\varepsilon_{n+1}^{2p-1} \mid \mathcal{F}_n] = 0 \qquad \text{p.s.}$$

Supposons de plus que ε_n possède un moment conditionnel fini d'ordre $a > 2p$, alors il existe c tel que $2pa^{-1} < c < 1$, et dès que $\alpha_n^{-p} = \mathcal{O}(n^{2c})$,

$$\left| \frac{1}{n} \; C_n(p) - \sigma(2p) \right| = o(n^{c-1}) \qquad \text{p.s.}$$

▷ Si le bruit ε_n possède un moment conditionnel fini, constant, d'ordre p alors

$$G_n(p) = o\Big((\log s_n)^{(p+1)(1+\gamma)}\Big) \qquad \text{p.s.}$$

Systèmes itératifs. Représentation de séquences.

1. *Chaos Game Représentation - Définition.*

- Représentation de l'ADN sous forme graphique compacte
  - Outil de stockage
  - Visualisation de motifs
  - Comparaison de séquences (locale/globale)
  - Construction d'une trajectoire dans un espace continu. Correspondance bijective, indépendance de l'échelle.
- Système dynamique

- CGR dans le carré $[0,1]^2$.

  - Séquence d'ADN $S = (u_i)_{i=1,\ldots,N}$
  - Soit X_i la position du nucléotide s_i dans le carré, $X_0 = \left(\frac{1}{2}, \frac{1}{2}\right)$

$$X_{i+1} = \frac{1}{2}[X_i + \ell_{u_{i+1}}], \quad i = 1, \ldots N-1,$$

ou encore,

$$X_i = \sum_{j=1}^{i} \frac{\ell_{u_j}}{2^{i-j}} + \frac{X_0}{2^i}.$$

  - Comptage de mots. Résolution quelconque.
  - Chaque point contient toute l'histoire de la séquence parcourue.
  - Figures fractales. Composition des séquences.

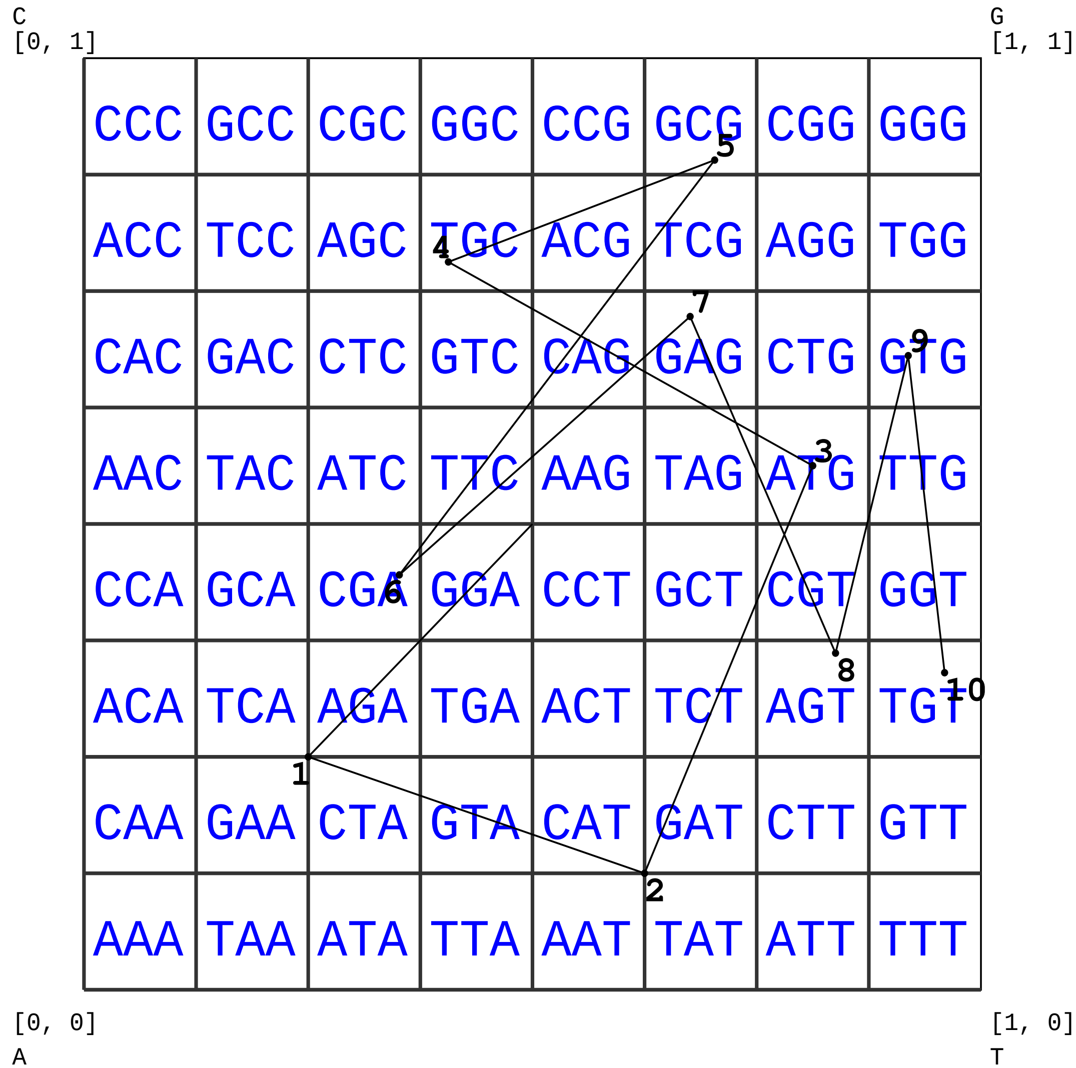


Fig. 1 – Chaos Game Representation des 10 premiers nucléotides du gène threonine thrA de E. Coli : ATGCGAGTGT.

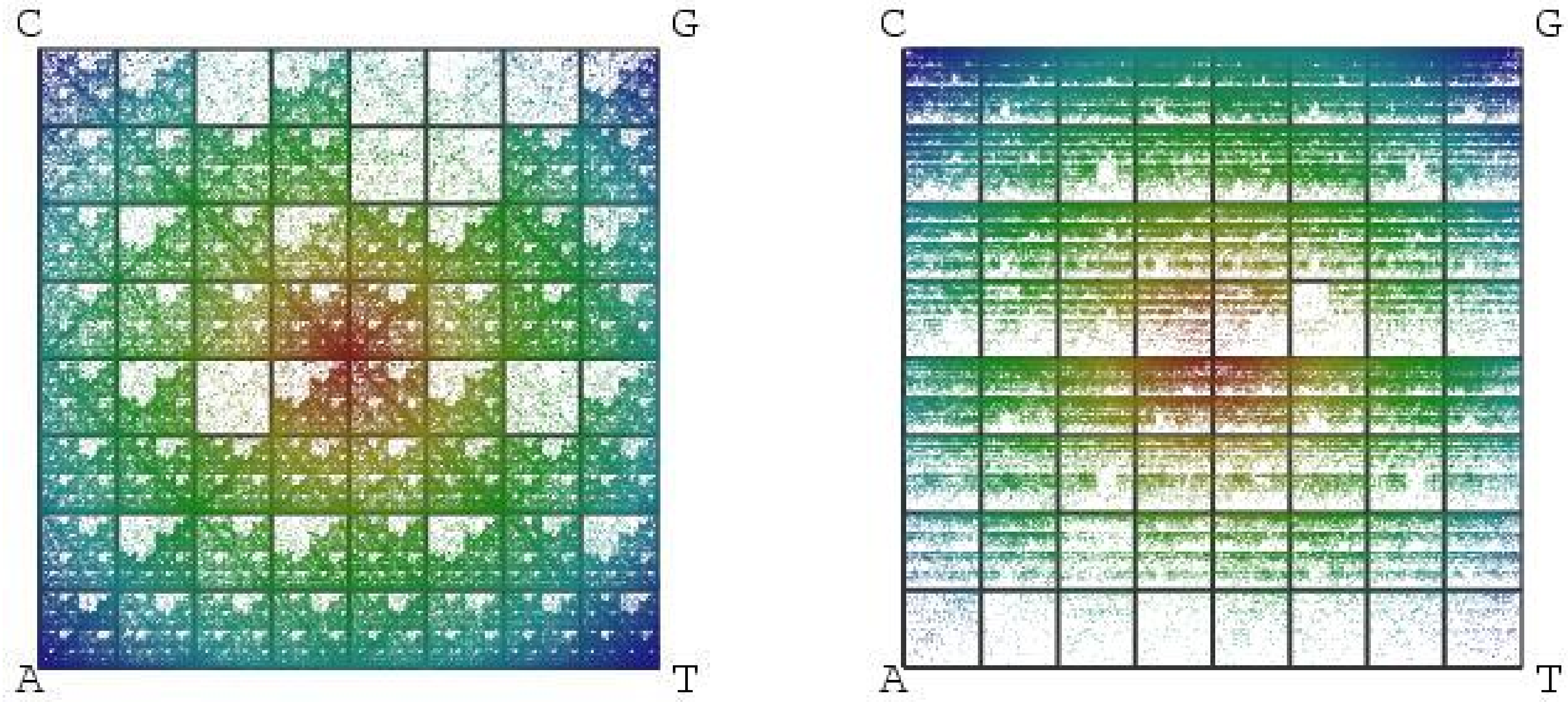


Fig. 2 – Chaos Game Representation des 400 000 premiers nucléotides du Chromosome 2 d'Homo Sapiens (GenBank $NT\_022184.13$) et de Streptomyces Coelicolor (GenBank NC_003888).

- CGR sur le segment $[0, 1]$.
  - Choix des positions des nucléotides sur le segment :

$$\ell_A = 0, \quad \ell_C = 1, \quad \ell_G = 2, \quad \ell_T = 3$$

$$X_{i+1} = \frac{1}{4}[X_i + \ell_{u_{i+1}}], \quad i = 1, \dots N-1,$$

ou encore

$$X_i = \sum_{j=1}^{i} \frac{\ell_{u_j}}{4^{i-j}} + \frac{X_0}{4^i}.$$

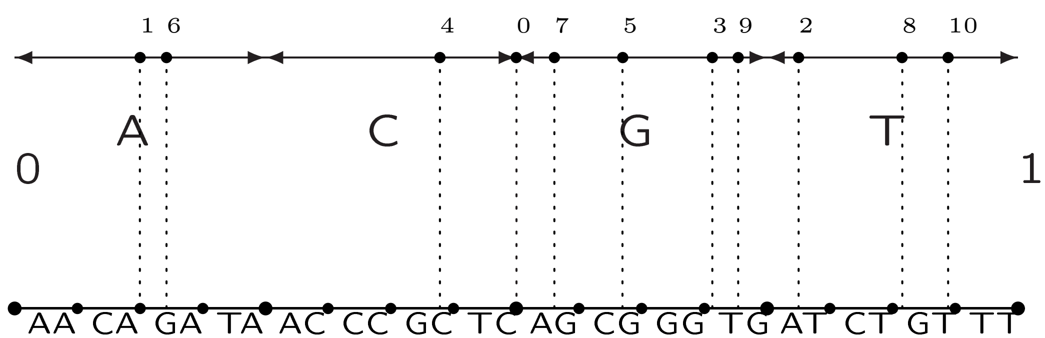


Fig. 3 – CGR des 10 premiers nucléotides du gène thrA de E. Coli (ATGCGAGTGT) sur le segment $[0, 1]$

▷ Décomposition 4-adique :

Soit une séquence de réalisations de v.a. i.i.d. sur un alphabet à 4 lettres. La suite des points de la CGR vérifie

$$X_n = \sum_{k=1}^{n} \frac{\varepsilon_k}{4^k},$$

où les ε_k sont des v.a. i.i.d à valeur dans $\{0, ..., 3\}$.

▷ Loi limite :

Mesure uniforme (cas équiprobable) / mesure sans atome singulière par rapport à la mesure de Lebesgue (cas non équiprobable)

2. *Un test de structure de séquence.*

- Construction du test.

Test de l'hypothèse H_0 : "$U = u_1 \dots u_N$ est une séquence i.i.d." contre l'hypothèse H_1 : "$U = u_1 \dots u_N$ est une séquence ergodique, stationnaire, mais non i.i.d.".

Proposition : La séquence aléatoire ergodique et stationnaire U est i.i.d. si et seulement si

$$\pi(S_u)\pi(B) = \pi\Big(\frac{B + \ell_u}{2}\Big), \quad \forall u \in \mathcal{A},\ \forall B \subset \mathcal{B}\big([0,1]^2\big).$$

Théorème : Le test de région de rejet

$$\left\{ \left| \hat{\pi}_N(S_u)\hat{\pi}_N(B) - \hat{\pi}_N\left(\frac{B+\ell_u}{2}\right) \right| > u_\alpha \frac{\hat{\sigma}_N(B,u)}{\sqrt{N}} \right\}$$

est un test de niveau asymptotique α de H_0 contre H_1, où

$$\hat{\sigma}_N(B,u) \stackrel{\text{def}}{=} \sqrt{\Big(\hat{\pi}_N(S_u)\big(1-\hat{\pi}_N(S_u)\big)\Big)\Big(\hat{\pi}_N(B)\big(1-\hat{\pi}_N(B)\big)\Big)},$$

et où u_α est le $(1-\frac{\alpha}{2})$-quantile de la loi normale centrée réduite.

- Le choix de B et de u dépend de la loi de la séquence, inconnue en pratique.

- Théorème : Plus généralement, pour toute partition $B_1, \ldots, B_k$ de $[0,1]^2$, avec $k > 1$,

$$\left\{ \sum_{i=1}^{k} \sum_{u \in \mathcal{A}} N \frac{\left(\hat{\pi}_N(S_u)\hat{\pi}_N(B) - \hat{\pi}_N\left(\frac{B+\ell_u}{2}\right)\right)^2}{\hat{\pi}_N(S_u)\hat{\pi}_N(B)} > q_\alpha(3(k-1)) \right\}$$

  est une zone de rejet d'un test asymptotique de niveau α de H_0 contre H_1, où $q_\alpha(3(k-1))$ est le $1-\alpha$ quantile d'une loi de Chi-Deux à $3(k-1)$ degrés de liberté.

- Résultat analogue dans le cas d'un seul ensemble.

  Proposition : Les deux tests précédents sont asymptotiquement consistants.

- Simulations

  Cas i.i.d. et Markov.

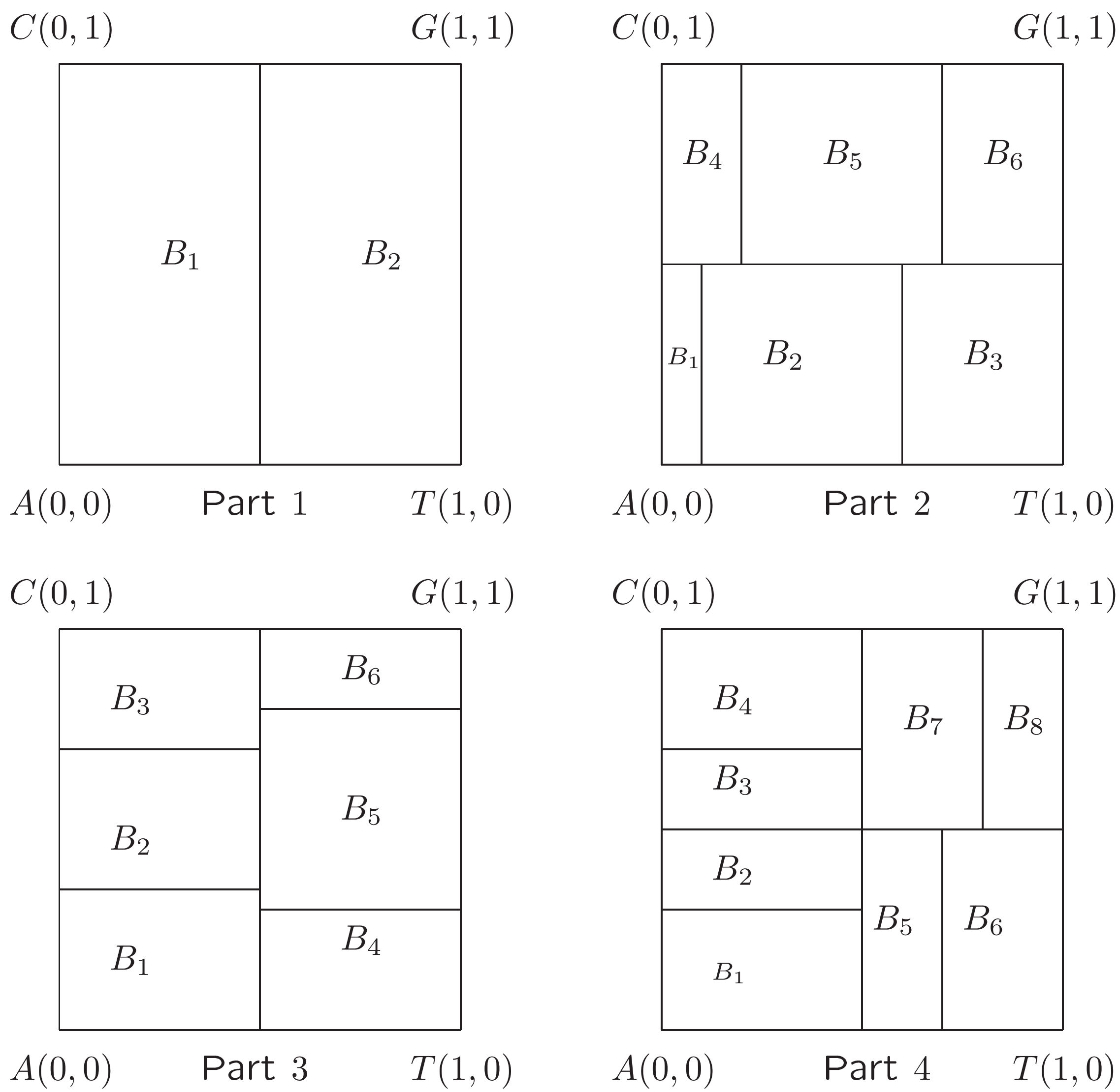


Fig. 4 – Les 4 différentes partitions de $[0,1]^2$ choisies pour le test.

N	Pearson	Part 1	Part 2	Part 3	Part 4
50	85.50%	94.70%	65.60%	94.40%	87.30%
100	91.00%	94.70%	86.70%	95.10%	93.40%
500	94.80%	95.20%	95.30%	95.40%	94.20%
1000	94.10%	94.90%	95.10%	95.10%	94.70%
10000	94.70%	95.70%	94.60%	94.60%	95.50%
100000	94.30%	95.40%	94.90%	95.00%	95.90%
500000	94.40%	95.00%	94.60%	95.40%	95.30%

Tab. 1 – Cas i.i.d. $\alpha = 0.05$.

ordre	N	Pearson	Part 1	Part 2	Part 3	Part 4
1	50	44.6%	9.80%	24.50%	6.80%	17.50%
	100	85.6%	14.40%	11.20%	7.80%	10.50%
	500	100%	63.60%	28.80%	40.30%	36.50%
	1000	100%	92.60%	64.50%	77.10%	75.80%
	10000	100%	100%	100%	100%	100%
2	50	3.2%	8.90%	23.30%	8.60%	15.90%
	100	4.7%	15.50%	9.80%	9.30%	13.10%
	500	8.7%	64.30%	29.90%	39.50%	38.00%
	1000	7.3%	91.40%	64.50%	80.00%	71.90%
	10000	7.2%	100%	100%	100%	100%
10	50	2.5%	9.50%	21.20%	8.60%	16.40%
	100	3.3%	13.80%	10.10%	10.60%	12.60%
	500	6.1%	64.50%	29.70%	42.90%	41.20%
	1000	5.6%	91.50%	65.20%	75.70%	76.40%
	10000	4.8%	100%	100%	100%	100%

Tab. 2 – Cas markoviens. $\alpha = 0.05$

3. *Signature génomique*

- Définition : Le profil d'abondance relative de dinucléotides étudié par Karlin et al. est l'ensemble des $\rho^*_{uv} = \frac{f^*_{uv}}{f^*_u f^*_v}$, où f^*_u est la fréquence empirique du nucléotide u et f^*_{uv} celle du dinucléotide uv. Cet ensemble constitue une signature génomique.

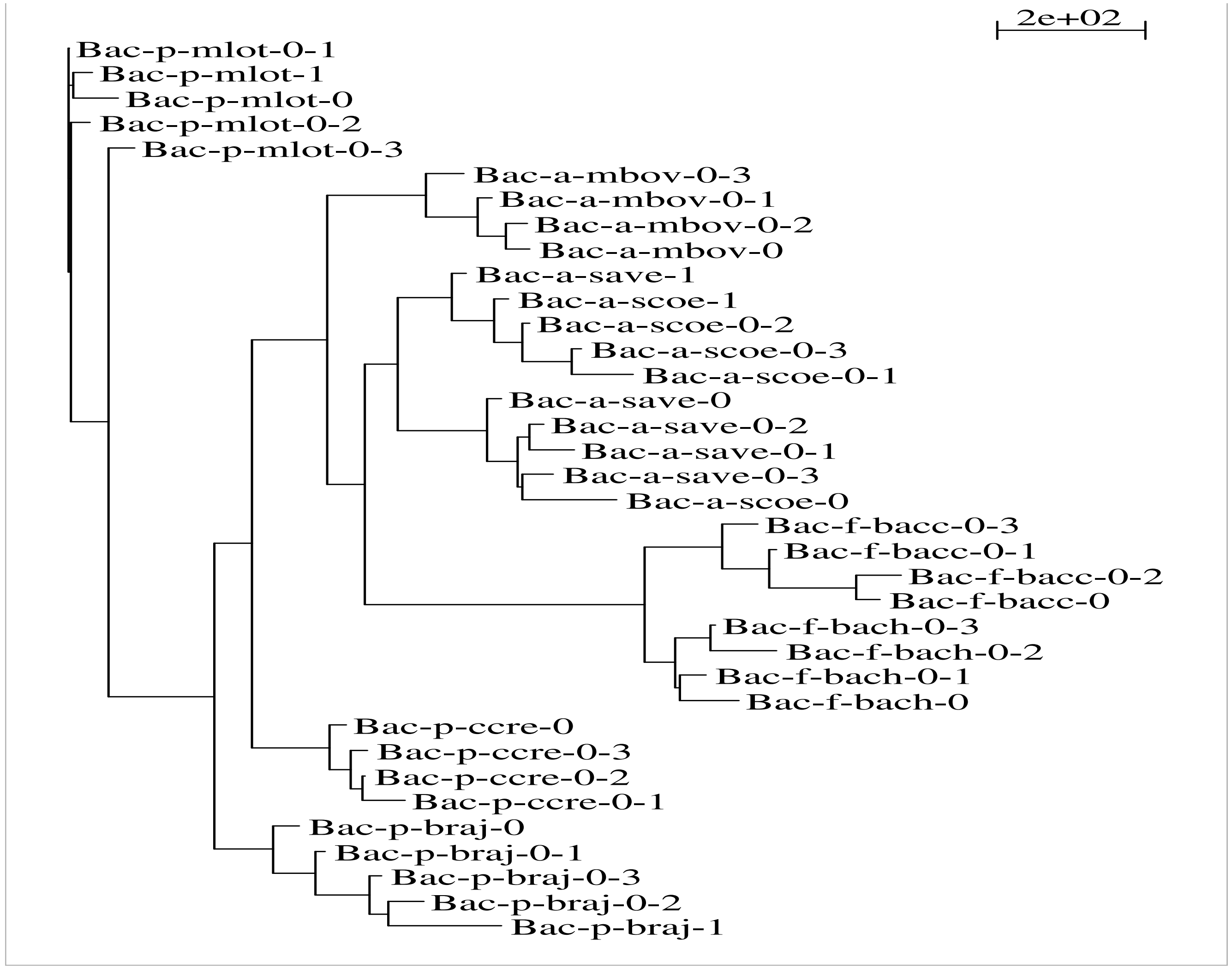
- Stabilité locale

- Reconstruction phylogénétique.

- Profil d'abondance relative basée sur la CGR

$$\hat{\rho}(B, v) \stackrel{\text{def}}{=} \frac{\hat{\pi}_N\left(\frac{B+\ell_v}{2}\right)}{\hat{\pi}_N(B)\hat{\pi}_N(S_v)}.$$

- Construction d'une mesure de distance δ entre les



rapports.

Fig. 5 – Arbre construit à partir du profil d'abondance relative des dinucléotides.

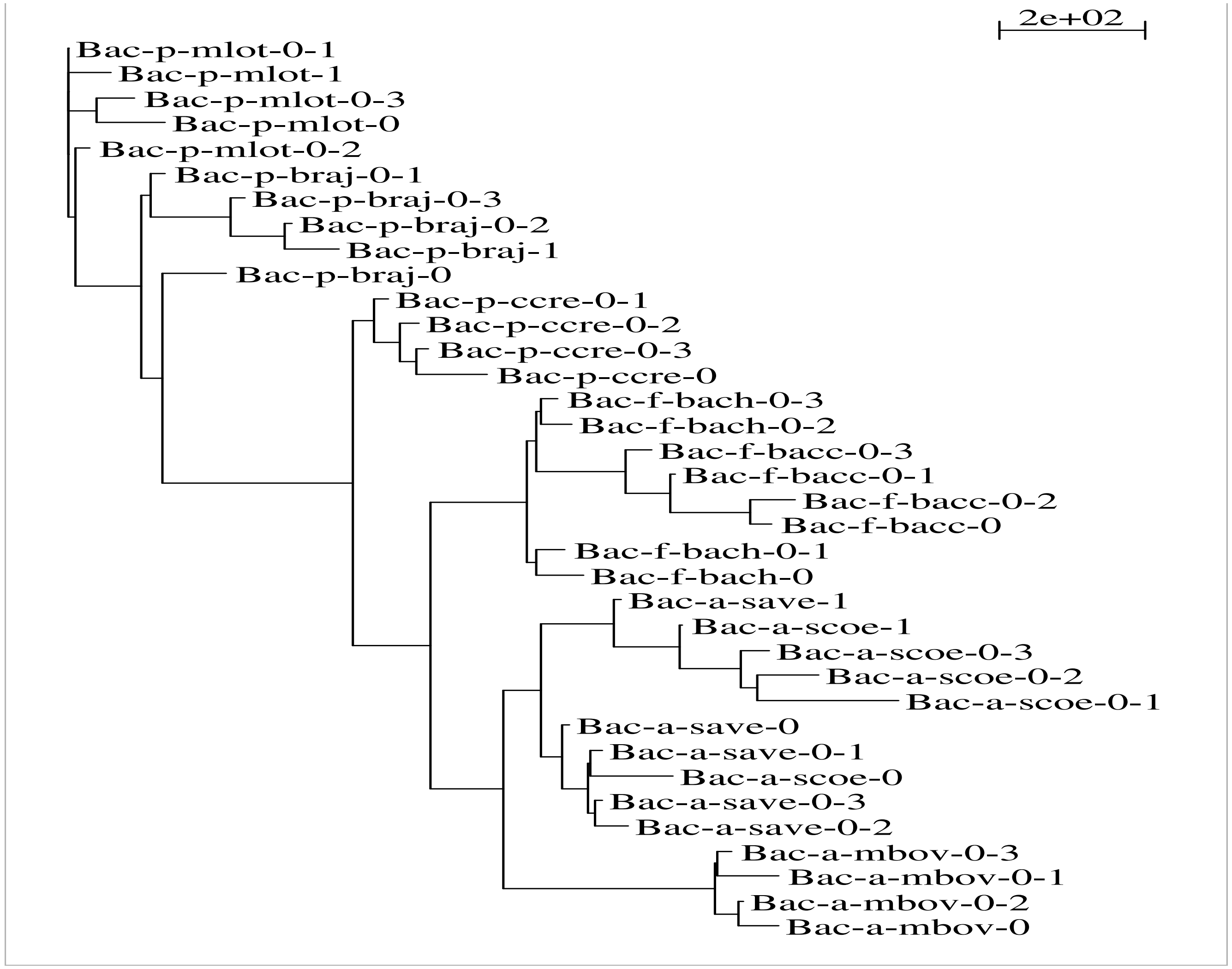


Fig. 6 – Arbre construit à partir du profil d'abondance relative basé sur la CGR.

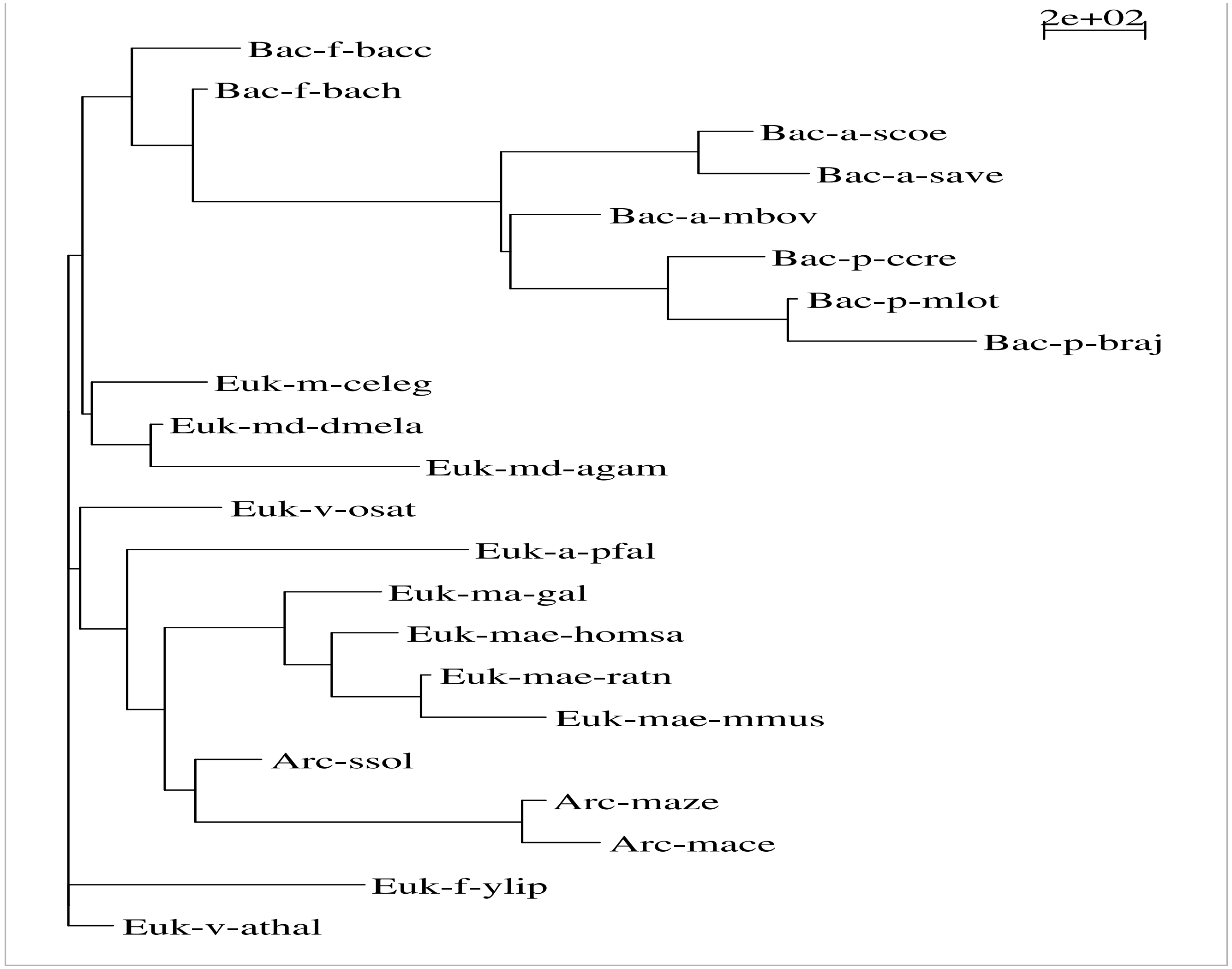


Fig. 7 – Arbre construit à partir du profil d'abondance relative basé sur la CGR d'une partition de 100 zones.