Modèles de Markov modulés pour l'horloge moléculaire

Alain Jean-Marie Nicolas Galtier Olivier Gascuel
Université de Montpellier
ajm@lirmm.fr

Journées MAS – Grenoble
3 septembre 2002

Plan de l'exposé

1. Introduction: évolution des séquences

2. Modèles de variation de l'horloge

3. Modèle Markov-modulé et algorithme de calcul

Évolution des séquences génétiques

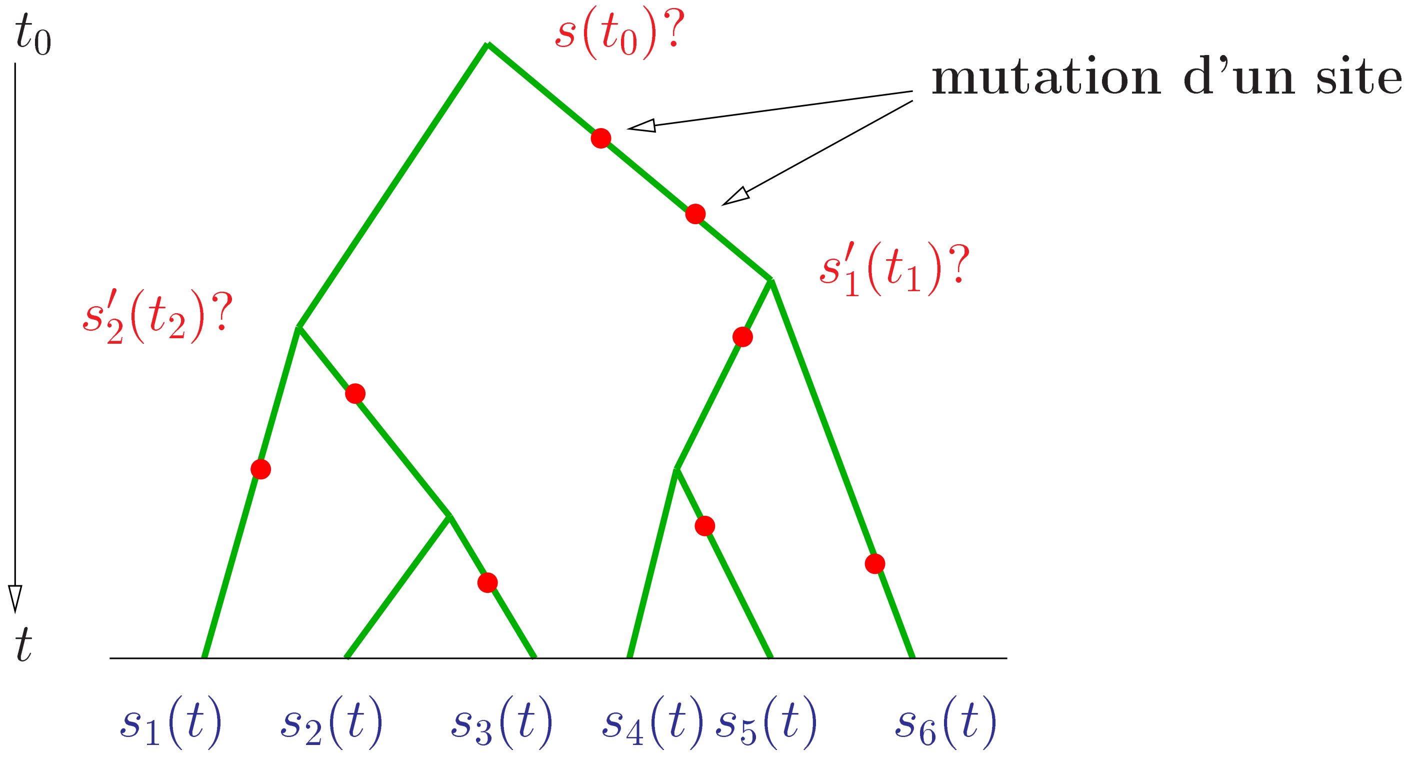
On considère une séquence

$$s(t) \;=\; (\ell_1(t), \ell_2(t), \ldots, \ell_n(t)\ )$$

de «lettres» d'un alphabet génétique:
- bases (4 lettres ATCG),
- acides aminés (20 lettres),
- codons ($\neq$STOP, 61 lettres).

Chacun de ces «sites» évolue au cours du temps, de façon aléatoire.

Étant donné un arbre phylogénétique, on peut connaître la séquence dans les différentes feuilles de l'arbre. Mais *quid* des nœuds?



Un modèle courant est:

- on néglige (ou compense) les insertions et délétions
- les sites évoluent indépendemment les uns des autres
- les sites évoluent indépendemment sur les différentes branches
- l'évolution est selon un processus de Markov en temps continu (réversible)

Exemple de générateur infinitésimal:

$$\mathbf{Q} = \begin{pmatrix} - & \nu\pi_T & \nu\pi_C & \mu\pi_G \\ \nu\pi_A & - & \mu\pi_C & \nu\pi_G \\ \nu\pi_A & \mu\pi_T & - & \nu\pi_G \\ \mu\pi_A & \nu\pi_T & \nu\pi_C & - \end{pmatrix} .$$

(Hasegawa, Kishino et Yano, 1985: HKY85).

Il a pour distribution stationnaire:

$$(\pi_A, \pi_T, \pi_C, \pi_G) \ ,$$

est réversible, et a des vitesses de «transition» et «transversion» différentes.

Estimation Bayesienne

Le problème: estimer les paramètres

- la forme de l'arbre
- les longueurs des branches
- la matrice Q
- les séquences inconnues.

Quelle que soit la forme du calcul, on a besoin de:

$$\mathbb{P}\{\ell(t) = b \mid \ell(0) = a\} = (e^{\mathbf{Q}t})_{ab} .$$

étant données une valeur de t et une matrice Q.

Vitesse d'évolution variable

Les statistiques ont fait apparaître le fait que les sites n'évoluent pas à la même vitesse.

Plusieurs approches:

- sites invariables et sites variables
- sites à vitesse constante, mais la vitesse est différente selon les sites (Uzzell & Corbin).

  Les vitesses sont aléatoires, typiquement selon une loi Γ discrétisée.
- sites à vitesse variable (on/off, modèles «covarion» de Fitch (1971), Tuffley & Steel (1998), Galtier (2001)).

Covarion 71: un modèle *ad hoc*.

- une certaine proportion ϕ des sites sont variables,
- après une substitution, un site reste variable avec probabilité p (de persistance)
- s'il change de catégorie, un autre site invariable est choisi au hasard et devient variable.

Avantage: le nombre de sites variables/invariables est constant au cours du temps.

Processus de Markov Modulés

Les processus de Markov modulés sont des processus d'arrivée d'une certaine quantité:

$N(t)$	$\in \mathbb{N}$	nombre d'occurrences d'un processus ponctuel (mutations, transitions, arrivées d'événements)
$N(t)$	$\in \mathbb{R}$	quantité de fluide/d'information (bits, octets) arrivés à un élément de réseau
$N(t)$	$\in \mathbb{R}$	quantité de «temps» écoulée (vidéo, temps-CPU, biologie,...)

Ils utilisent un processus markovien «caché» (environnement).

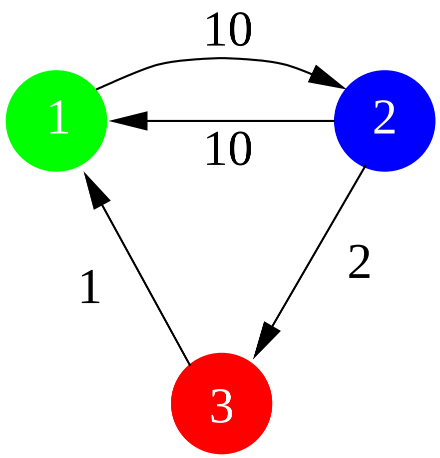
Ils constituent une famille de modèles pour lesquels on arrive à

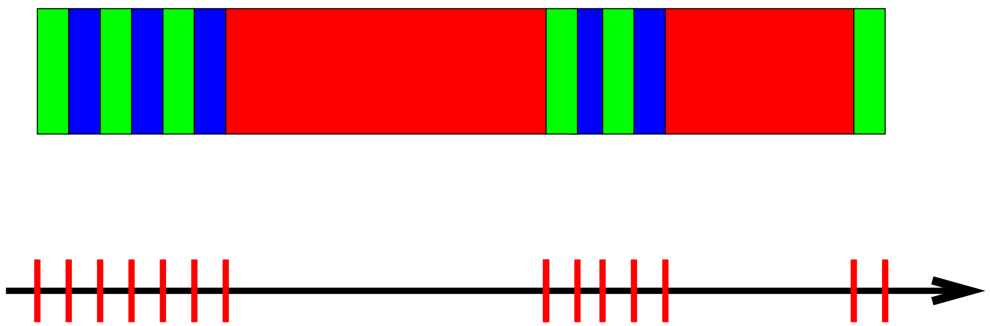
- modéliser des processus complexes (rafales, dépendances dans le temps),
- faire du calcul stochastique: distributions, files d'attente, asymptotiques...

MAP: Markov Arrival Process

Soit $\{X(t); t \in \mathbb{R}\}$ une chaîne de Markov en temps continu dans un espace fini.

$\{N(t); t \in \mathbb{R}\}$ compte le nombre de transitions de X dans $[0, t[$.





BMAP: Batch Markov Arrival Process

Aussi nommé «N-process» (N comme Neuts), processus «versatile» modulé par Markov.

$\{(X(t), N(t)); t \in \mathbb{R}\}$ est un chaîne de Markov en temps continu dans $\mathcal{E} \times \mathbb{N}$ dont le générateur a la structure:

$$Q = \begin{pmatrix} D_0 & D_1 & D_2 & \dots & \\ & D_0 & D_1 & D_2 & \\ & & D_0 & D_1 & \ddots \\ & & & \ddots & \ddots \end{pmatrix}$$

On peut avoir

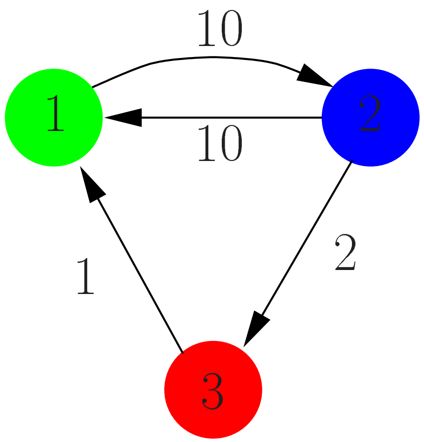
- des changements de phase $i \to j$ sans arrivée, avec taux $(D_0)_{i,j}$
- l'arrivées d'un groupe (batch) de taille k sans changement de la phase i, avec taux $(D_k)_{i,i}$
- changement de phase et arrivée d'un groupe de taille k, avec taux $(D_k)_{i,j}$.

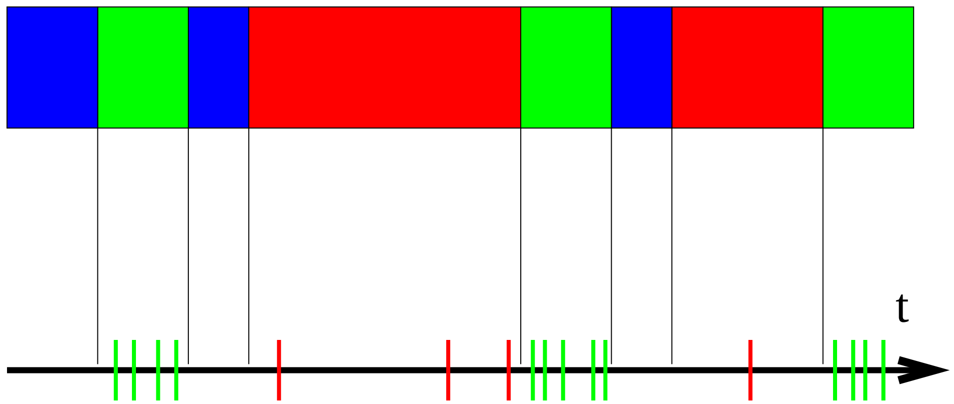
MMPP: Markov Modulated Poisson Process

Soit $\{X(t); t \in \mathbb{R}\}$ une chaîne de Markov en temps continu sur un espace fini $\mathcal{E}$.

Soient $\lambda_i \geq 0$ des taux d'arrivée, pour chaque $i \in \mathcal{E}$.

On suppose que les arrivées ont lieu selon un processus de Poisson d'intensité $\lambda_{X(t)}$: intensité λ_i dans les intervalles où $X(t) = i$.





C'est un cas particulier de processus BMAP, pour lequel

- les arrivées se font un par un: $D_k = 0$, $k \geq 2$;
- les changements de phase ne provoquent pas d'arrivée

$$D_1 \;=\; \begin{pmatrix} \lambda_0 & & \\ & \lambda_1 & \\ & & \ddots \end{pmatrix} \;.$$

Cas particulier, le processus IPP (Interrupted Poisson Process): $\lambda_0 = 0$, $\lambda_1 = \lambda$.

Et le processus de Poisson non interrompu:

$$\lambda_i \;=\; \lambda, \qquad \forall i \in \mathcal{E} \;.$$

MMRP: Markov Modulated Rate Process

Soit $\{X(t); t \in \mathbb{R}\}$ une chaîne de Markov en temps continu sur un espace fini $\mathcal{E}$.

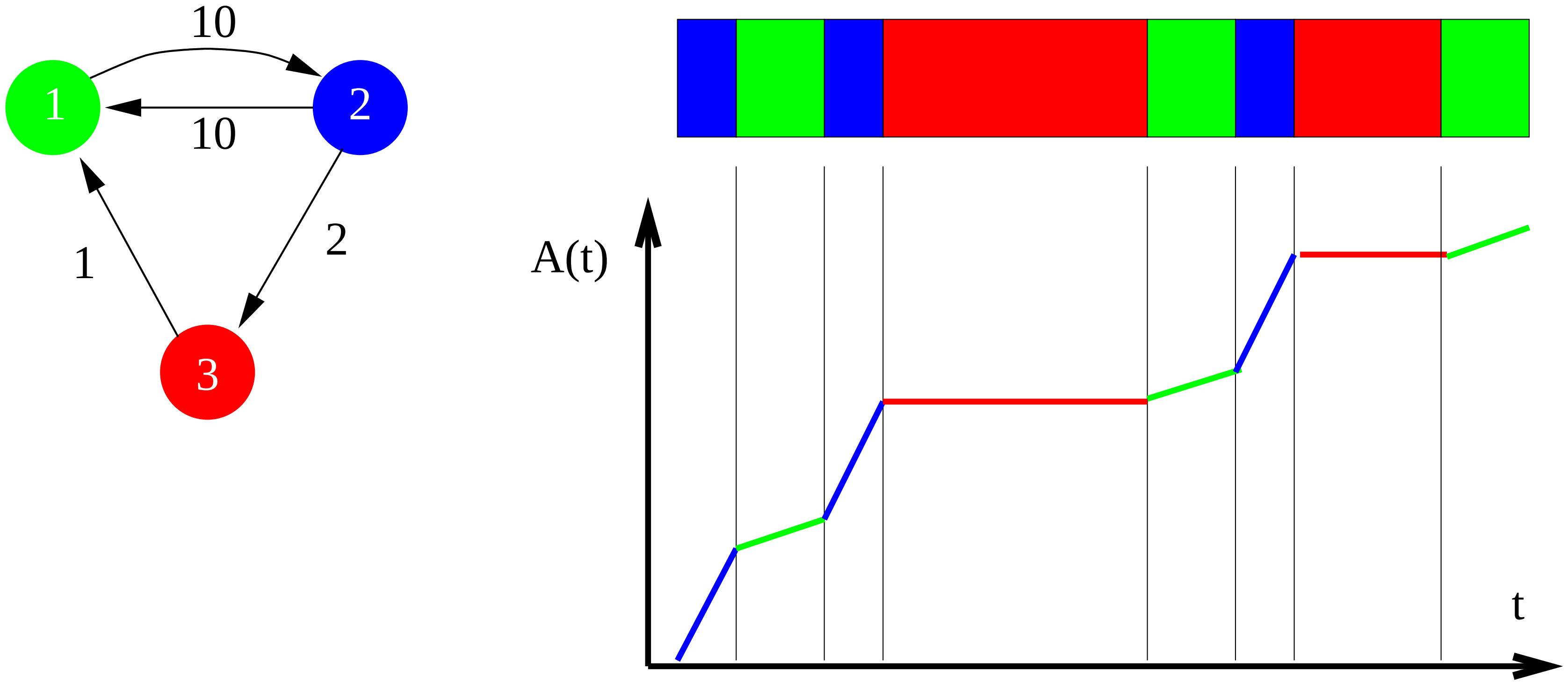
Soient r_i (≥ 0 en général) des taux d'arrivée (accumulation, déplétion), pour chaque $i \in \mathcal{E}$.

On suppose que les arrivées ont lieu selon un processus fluide de débit $r_{X(t)}$, c'est-à-dire d'intensité r_i dans les intervalles où $X(t) = i$.

Soit $N(t)$ la quantité de fluide arrivé à la date t:

$$\frac{dN}{dt}(t) \;=\; r_{X(t)} \;.$$

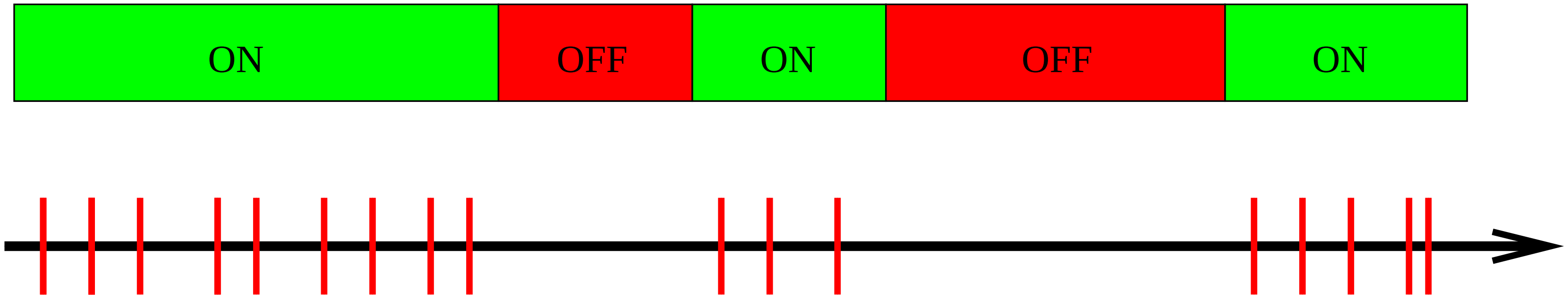
Exemple. $\mathcal{E}$ a trois états, $r_3 = 0,\ 0 < r_1 < r_2$:



Sources On/Off

Processus On/Off:
• alternance de périodes On et Off, de durées IID (deux distributions)
• pendant les périodes On, un processus fluides (débit constant) ou discrets (Poisson ou périodique).

→ bon modèle de télécommunication voix/vidéo numériques, ainsi que sources TCP, etc.



$\rightarrow$ processus de renouvellement alterné.

MMPP, BMAP, etc ont tous des durées exponentielles. On/Off peut avoir des durées non-exponentielles (en particulier, queues lourdes!).

Superposition de sources

Si on superpose plusieurs sources Markov-modulées, le processus résultant est toujours Markov-modulé.

Les matrices (générateur et taux) s'obtiennent comme des *sommes de Kronecker*.

Produit de Kronecker de deux matrices A $(n \times n)$ et B $(m \times m)$: une matrice $nm \times nm$ avec

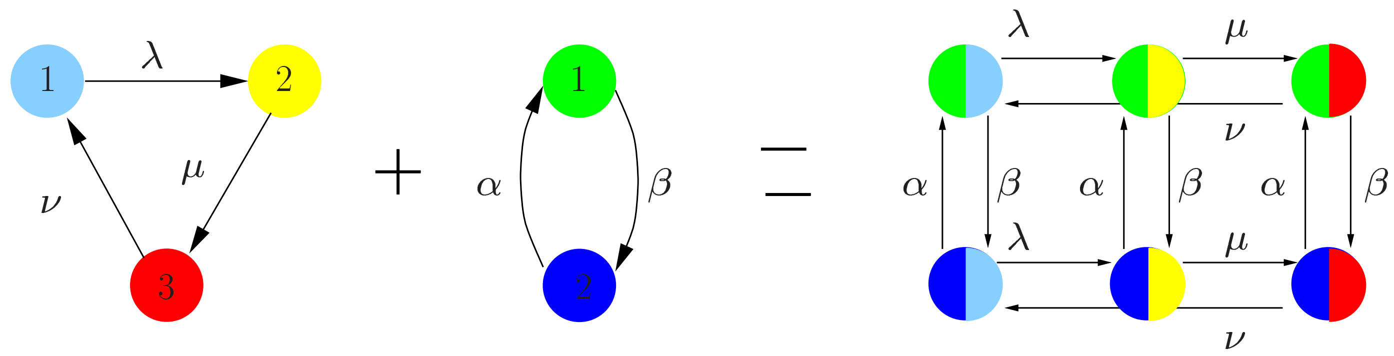
$$A \otimes B \;=\; \begin{pmatrix} A_{11}B & \dots & A_{1n}B \\ \vdots & & \vdots \\ A_{n1}B & \dots & A_{nn}B \end{pmatrix} .$$

Somme de Kronecker:

$$A \oplus B \;=\; A \otimes I(m) + I(n) \otimes B$$

$$= \begin{pmatrix} A_{11}B & & \\ & \ddots & \\ & & A_{nn} \end{pmatrix} + \begin{pmatrix} B_{11}I & \dots & B_{1m}I \\ \vdots & & \vdots \\ B_{n1}I & \dots & B_{nn}I \end{pmatrix}.$$

Exemple: pour deux chaînes de Markov $\{X_1(t)\}$ et $\{X_2(t)\}$, on a:



$$\begin{pmatrix} -\lambda & \lambda & 0 \\ 0 & -\mu & \mu \\ \nu & 0 & -\nu \end{pmatrix} \oplus \begin{pmatrix} -\alpha & \alpha \\ \beta & -\beta \end{pmatrix} = \left(\begin{array}{ccc|ccc} - & \lambda & 0 & \alpha & 0 & 0 \\ 0 & - & \mu & 0 & \alpha & 0 \\ \nu & 0 & - & 0 & 0 & \alpha \\ \hline \beta & 0 & 0 & - & \lambda & 0 \\ 0 & \beta & 0 & 0 & - & \mu \\ 0 & 0 & \beta & \nu & 0 & - \end{array}\right)$$

Résultats pour les chaînes de Markov Modulées

Pour ces processus, on connaît

- des formules pour les transitoires (équations différentielles, transformées de Laplace, décompositions spectrales)
- des asymptotiques
- des principes grandes déviations.

Modulation des vitesse et algèbre de Kronecker

Le site Z évolue sur un alphabet selon un processus de Markov de générateur infinitésimal $\mathsf{M} = (m_{ab})$.

Son «environnement» X évolue selon une CMTC de générateur $\mathsf{G} = (g_{ij})$.

Quand X est dans l'état i, la vitesse de Z (taux de transition) est multipliée par v_i.

Le générateur du processus $(Z(t), X(t))$ a pour transitions:

$$\begin{array}{rcll} (i,a) & \to & (i,b) & \text{avec taux } m_{ab}v_i \\ (i,a) & \to & (j,a) & \text{avec taux } g_{ij} \end{array}$$

Sous forme matricielle (par blocs):

$$\mathsf{Q} \;=\; \begin{pmatrix} v_1\mathsf{M} + g_{11}\mathsf{I} & g_{12}\mathsf{I} & \cdots & g_{1K}\mathsf{I} \\ g_{21}\mathsf{I} & v_2\mathsf{M} + g_{22}\mathsf{I} & & g_{2K}\mathsf{I} \\ \vdots & & \ddots & \\ g_{K1}\mathsf{I} & g_{K2}\mathsf{I} & \cdots & v_K\mathsf{M} + g_{KK}\mathsf{I} \end{pmatrix}$$

Et en utilisant le produit de Kronecker:

$$\mathsf{Q} \;=\; \mathsf{G} \otimes \mathsf{I} + \mathsf{V} \otimes \mathsf{M} \; .$$

où

$$\mathsf{V} \;=\; \mathtt{diag}(v_1, \ldots, v_K) \; .$$

Pour calculer $e^{\mathsf{Q}t}$, une méthode est de diagonaliser Q. On doit trouver ses valeurs propres et vecteur propres. Ayons l'idée de prendre x et y tels que

$$\begin{aligned} x\ \mathsf{M} &= \lambda\ x \\ y &= (a_1 x, \ldots, a_N x) = a \otimes x\ . \end{aligned}$$

Alors

$$\begin{aligned} y\ \mathsf{Q} &= (a \otimes x)\ (\mathsf{G} \otimes \mathsf{I} + \mathsf{V} \otimes \mathsf{M}) \\ &= a\mathsf{G} \otimes x\mathsf{I}\ +\ a\mathsf{V} \otimes x\mathsf{M} \\ &= a\ (\mathsf{G} + \lambda \mathsf{V})\ \otimes\ x\ . \end{aligned}$$

Il suffit de prendre a tq $a(\mathsf{G} + \lambda\mathsf{V}) = \mu a$ pour que $y\mathsf{Q} = \mu y$.

Algorithme de diagonalisation

Donnée: un générateur infinitésimal Q résultant de la modulation d'une matrice G par des vitesses $v_1, \ldots, v_K$:

$$\mathsf{Q} = \mathsf{G} \otimes \mathsf{I} + \mathsf{V} \otimes \mathsf{M} \; .$$

Résultat: les valeurs propres, vecteurs propres droits et gauches de Q (⇒ diagonalisation de Q)

Algorithme:

- Trouver les éléments spectraux de G:

$$\rightarrow \quad (\lambda_i; x_i, y_i) \qquad i = 1..K \ .$$

- Pour chaque i, trouver les éléments spectraux de $\mathsf{G} + \lambda_i \mathsf{V}$:

$$\rightarrow \quad (\mu_{ij}; a_{ij}, b_{ij}) \qquad i = 1..K, \ \ j = 1..N \ .$$

- Former les éléments spectraux de Q:

$$\rightarrow \quad (\mu_{ij}; a_{ij} \otimes x_i, b_{ij} \otimes y_i) \qquad i = 1..K, \ \ j = 1..N \ .$$

Complexité:

- soit N la taille de l'alphabet, K le nombre de vitesses ($N = 4, 20, 61,\ K = 4, 10, ...$)
- la matrice Q est de taille $NK \times NK$
- la diagonalisation directe est en $O(N^4K^4)$
- l'algorithme est en $O(K^4 + KN^4)$.

Le stockage des matrices entières n'est même pas nécessaire.

Bibliographie

Phylogénie et horloge variable

W.M. Fitch. Rate of change of concomitantly variable codons. *J. Mol. Evol.*, 1:84-96, 1971.

P.J. Lockhart, M.A. Steel, A.C. Barbrook, D.H. Hudson et C.J. Howe. A covariotide model explains apparent phylogenetic structure of oxygenic photosynthetic lineages. *Mol. Biol. Evol.*, 15:1183–1188. 1998.

C. Tuffley and M.A. Steel. Modelling the covarion hypothesis of nucleotide substitution. *J. Mol. Evol.*, 17:496–508, 1998.

Ph. Lopez, P. Forterre et H. Philippe. The root of the tree of life in the light of the covarion model. *J. Mol. Evol.*, 49:496–508, 1999.

Ph. Lopez. *Deux approches de l'évolution moléculaire: Asymétries de composition et recherche des covarions*. Thèse de l'Université de Paris-Sud, décembre 2000.

N. Galtier, Maximum Likelihood phylogenetic analysis under a covarion-like model. *Mol. Biol. Evol.*, 18:866–873, 2001.

Modèles markoviens modulés fluides

D. Anick, D. Mitra, and M.M. Sondhi. Stochastic theory of a data-handling system with multiple sources. *Bell Sys. Tech. J.*, 61:1871–1894, October 1982.

D. Mitra. Stochastic theory of a fluid models of producers and consumers coupled by a buffer. *Adv. Appl. Prob.*, 20:646–676, 1988.

T.E. Stern and A.I. Elwalid. Analysis of separable Markov-modulated rate models for information-handling systems. *Adv. Appl. Prob.*, 23:105–139, 1991.

A.I. Elwalid, D. Mitra, and T.E. Stern. A theory of statistical multiplexing of Markov modulated sources: Spectral expansions and algorihms. In W.J. Stewart, editor, *Numerical solution of Markov Chains*, 1991.

A.I. Elwalid and D. Mitra. Statistical multiplexing with loss priorities in rate-based congestion control of high speed networks. *IEEE Trans. Comm.*, 42(11):2989–3002, November 1994.

A.I. Elwalid and D. Mitra. Markovian arrival and service communication systems: Spectral expansions, separability and Kronecker-product forms. In W.J. Stewart, editor, *Computations in the Markov Chains*, pages 507–546. Kluwer, 1995.

MMPP, MAP, BMAP...

M.F. Neuts. The fundamental period of a queue with Markov-modulated arrivals. In *Probability, Statistics and Mathematics: papers in honour of Samuel Karlin*. Academic Press, NY, 1989.

W. Fischer and K. Meier-Hellstern. The Markov-modulated Poisson process (MMPP) cookbook. *Performance Evaluation*, 18:149–171, 1992.

D.M. Lucantoni, G.L. Choudhury, and W. Whitt. The transient $BMAP/G/1$ queue. *Commun. Statist.-Stochastic Models*, 10(1):145–182, 1994.

A. Jean-Marie, Z. Liu, P. Nain and D. Towsley, "Computational Aspects of the Workload Distribution in the MMPP/GI/1 Queue". *JSAC*, 1999.

Asymptotiques, bornes et bande passante équivalente

W. Whitt. Tail probabilities with statistical multiplexing and effective bandwidth. *Telecommun. Syst.*, 3:71–107, 1993.

D. Artiges and P. Nain. Upper and lower bounds for the multiplexing of multiclass Markovian on/off sources. *Performance Evaluation*, 27&28, pp. 673–698, 1996.

V.G. Kulkarni. Effective bandwidth for Markov regenerative sources. *Queueing Systems*, 24, pp. 137–153, 1996.

Z. Liu, P. Nain, and D. Towsley. Exponential bounds with applications to call admission. *JACM*, 44 (2):366–394, 1997.