Modélisation statistique et analyse de données, du génome au transcriptome

Pierre Nicolas

Chargé de Recherche INRA
Unité MaIAGE, Jouy-en-Josas

Mémoire présenté pour l'obtention d'une

Habilitation à Diriger des Recherches

Université Paris-Sud

Spécialité : biologie computationnelle

Soutenue le **26 juin 2018** à l'INRA de Jouy-en-Josas

Jury :
Christophe Ambroise, Professeur, Université d'Évry (rapporteur)
Philippe Bouloc, Directeur de Recherche, CNRS, Orsay
Sylvain Brisse, Directeur de Recherche, Institut Pasteur, Paris
Alessandra Carbone, Professeur, Sorbonne Université
Daniel Gautheret, Professeur, Université Paris-Sud (rapporteur)
Jean-Michel Marin, Professeur, Université de Montpellier
Marie-France Sagot, Directeur de Recherche, INRIA, Lyon (rapporteur)

Remerciements

Je tiens tout d'abord à remercier pour avoir pris le temps d'examiner mon travail : Marie-France Sagot, Christophe Ambroise, et Daniel Gautheret, rapporteurs de ce mémoire d'HDR ; ainsi que Alessandra Carbone, Philippe Bouloc, Sylvain Brisse et Jean-Michel Marin, membres du jury de soutenance.

Je veux remercier Philippe Bessières et Bernard Prum qui ont tous deux joué un rôle majeur dans mon encadrement de thèse ainsi qu'après. Leurs décès récents et complètement inattendus m'ont profondément affecté et ont contribué à me faire prendre conscience du temps qui passe. Je dois aussi beaucoup à mes encadrants de thèse et de post-doctorat : Florence Muri, Anne-Sophie Tocquet, Lei Li, et Simon Tavaré pour leurs avis précieux, leurs idées, leur sympathie... Je souhaiterais mettre dans cette même catégorie François Rodolphe que j'ai souvent sollicité pour des conseils durant ma thèse et jusqu'à son départ récent en retraite, et Sophie Schbath dont le dévouement au fonctionnement du laboratoire et la confiance qu'elle m'a accordée m'ont beaucoup aidé.

Je veux ensuite exprimé ma profonde gratitude à mes principaux collaborateurs depuis mon recrutement à l'INRA : Guillaume Achaz, Ruben Avendaño-Herrera, Stéphane Aymerich, Jean-François Bernadet, Philippe Bessières, Elena Bidnenko, Hélène Chiapello, Ciaran Condon, Étienne Dervyn, Éric Duchaud, Tudor Draganoiu, Hélène Falentin, Erina Fujiwara-Nagata, Rozenn Gardan, Cyprien Guérin, Catherine Legraverend, Ulrike Mäder, Ruben Mars, Philippe Noirot, Harald Putzer, François Lecointe, Valentin Loux, Hugues Richard, Tatiana Rochat, Julie Soutourina, Maarten van de Guchte, Jan Maarten van Dijl, Greg Wiens. Leur rencontre fut vraiment l'un des grands plaisirs de ces collaborations.

Je dois aussi beaucoup aux étudiants et contractuels à l'encadrement desquels j'ai eu l'occasion de participer dont Bogdan Mirauta et Ibrahim Sultan, étudiants en thèse, pour leur investissement, leur sympathie, et aussi pour m'avoir supporté.

Je remercie chaleureusement mes nombreux collègues de l'unité MaIAGE qui contribuent collectivement et chacun à leur façon à faire du laboratoire un lieu agréable et propice au travail. Cyprien et Christian m'ont apporté une aide précieuse dans la logistique de la soutenance.

Je veux enfin mentionner ma famille, au premier rang desquels Eliza, Irène et Anouk, ainsi que mes amis. Je leur suis immensément redevable.

Résumé

Ce mémoire a pour objectif d'éclairer les choix et le contexte de mon travail de recherche et d'en reprendre certains détails. Les thèmes abordés vont d'aspects méthodologiques de la bioinformatique à l'analyse de jeux de données biologiques. Je donne en particulier des précisions sur le développement de modèles statistiques à variables latentes, dont des chaînes de Markov cachées. Combinés à une inférence non-supervisée, ces modèles visent à permettre l'analyse de données de séquences biologiques sous divers angles : segmentation de l'ADN selon sa composition, détection de gènes et de motifs, analyse de polymorphisme et lissage de données d'expression le long du génome. Les questions de l'estimation des paramètres et du choix de la dimension sont abordées par des algorithmes de types *Expectation Maximization* et Monte-Carlo par chaînes de Markov. Je retrace ensuite ma contribution à l'analyse de deux grandes catégories d'objets biologiques. Le premier est l'architecture des transcriptomes bactériens que j'ai étudiée par des approches de transcriptomique globale (*tiling arrays*) chez la bactérie modèle à Gram positif *Bacillus subtilis* et chez l'une des ses cousines pathogène de l'homme, *Staphylococcus aureus*. Le second concerne la structure des populations et les mécanismes évolutifs que j'ai analysés par des approches génétiques (*Multi-Locus Sequence Typing*) et de comparaison de génomes chez les bactéries pathogènes des poissons appartenant à l'espèce *Flavobacterium psychrophilum* et au genre voisin *Tenacibaculum*.

Abstract

The purpose of this report is to shed light on choices and context of my research and to give some details on selected points. Themes range from methodological aspects of computational biology to analyses of biological data sets. I focus in particular on the development of statistical models involving latent variables, among which hidden Markov chains. Combined with unsupervised inference, they provide a way to analyze biological sequences under several perspectives : segmentation according to DNA composition, detection of genes and motifs, analysis of polymorphism and smoothing of expression data along genomes. Inference algorithms used here for parameter estimation and choice of model dimension relies on Expectation-Maximization algorithms and Markov chain Monte Carlo methods. Then, I summarize my contribution to the analysis of two main categories of biological objects. The first is bacterial transcriptome architecture that I investigated using global transcriptomics (tiling arrays) in the model Gram-positive bacterium *Bacillus subtilis* and in one of its human pathogen cousins, *Staphylococcus aureus*. The second concerns population structure and evolution that I analyzed using genetics (Multi-Locus Sequence Typing) and comparative genomics approaches in fish pathogenic bacteria belonging to the species *Flavobacterium psychrophilum* and to the related genus *Tenacibaculum*.

Table des matières

Chapitre 1

Grandes orientations thématiques et panorama des travaux réalisés

L'objectif de ce chapitre introductif est de présenter succinctement le cadre général des travaux couverts par ce mémoire. La période concernée s'étend du début de ma thèse à aujourd'hui (automne 2000 - début 2018).

1.1 Éléments de chronologie pour éclairer le choix des thèmes de recherche

1.1.1 Thèse

Si je cherche la raison du domaine dans lequel s'inscrivent mes travaux présentés ici, un événement me semble avoir joué un rôle déterminant. Il s'agit d'un stage de L3 réalisé à l'INRA de Jouy-en-Josas durant l'été 1998 avec Philippe Bessières. Le sujet du stage était le traitement informatique (en langage Perl) de résumés d'articles scientifiques sur la bactérie *Bacillus subtilis* pour en extraire les co-occurrences de noms de gènes.

Plusieurs facteurs ont contribué à faire de ce stage le point de départ de ma trajectoire de recherche. Le plus évident est la singularité de cette période : les premiers génomes microbiens venaient d'être séquençés (Fleischmann *et al.*, 1995) et le laboratoire (alors « Génétique Microbienne ») avait participé au consortium international à l'origine de la séquence des 4,2 millions de paires de bases du génome de *B. subtilis* (Kunst *et al.*, 1997). Ce contexte « historique » n'aurait probablement pas autant influencé le choix de mes thèmes de recherche sans l'enthousiasme de Philippe pour les perspectives ouvertes par la disponibilité des génomes et l'utilisation de méthodes informatiques et mathématiques pour leur analyse. Ce double intérêt contribua à la création du laboratoire Mathématique Informatique et Génome (MIG) deux ans plus tard (2000).

J'ai aussi eu l'occasion de rencontrer lors ce stage un autre étudiant de Philippe (Laurent Bize, en M2) qui tentait d'analyser les hétérogénéités de composition le long du génome de *B. subtilis* en utilisant des chaînes de Markov cachées. Ces modèles statistiques et les algorithmes d'inférence associés venaient d'être adaptés aux séquences d'ADN dans la thèse de Florence Muri, réalisée sous la direction de Bernard Prum (Muri, 1997, 1998). J'ai

été immédiatement attiré par cette approche utilisant des modèles à variables cachées pour capturer les structures de dépendance dans les observations. Combinés à des méthodes d'inférence non-supervisées, je voyais ces modèles comme pouvant nous aider à comprendre ces séquences qui n'étaient qu'une très longue suite d'apparence monotone et aléatoire des quatre nucléotides `A`, `C`, `G` et `T`. Deux ans plus tard, j'ai repris contact avec Philippe à l'occasion de mon M2. Laurent n'ayant pas poursuivi en thèse sur ce sujet, je me suis porté volontaire pour continuer ce travail.

Pour ma thèse, j'ai eu la chance de bénéficier d'un double encadrement, Philippe Bessières pour les aspects biologiques, et le trio Florence Muri, Anne-Sophie Tocquet et Bernard Prum à l'Université d'Evry pour les aspects mathématiques (Laboratoire Statistiques et Génome, maintenant intégré au LAMME). Le travail a commencé autour de l'utilisation de chaînes de Markov cachées pour l'étude des hétérogénéités le long du génome bactérien en utilisant l'algorithme EM pour la maximisation de la vraisemblance (des données incomplètes) dans un cadre fréquentiste (Nicolas *et al.*, 2002) et s'est ensuite progressivement élargi. En ce qui concerne des questions biologiques abordées, je me suis intéressé à la détection des gènes codants pour des protéines et la recherche des motifs promoteurs dans les génomes bactériens (travaux finalisés ultérieurement, notamment à travers les publications Nicolas *et al.*, 2006a et Ibrahim *et al.*, 2007). Pour aborder ces questions, les modèles se sont complexifiés (plus d'états cachés, des dépendances semi-markovienne) et je me suis intéressé aux algorithmes de Monte-Carlo par chaînes de Markov (MCMC) pour l'estimation dans un cadre bayésien. À la fin de la décennie 90, ces méthodes d'inférence très algorithmiques ont profité de l'essor des capacités de calcul et d'innovations méthodologiques (Gilks *et al.*, 1995). Je mentionnerai en particulier la formalisation d'algorithmes trans-dimensionnels permettant d'explorer non seulement un espace de valeurs de paramètres mais aussi des collections de modèles (Green, 1995). Cette démarche d'estimation pour la « sélection » de modèle m'a semblé particulièrement attrayante lorsque l'on souhaite pousser le plus possible la démarche d'analyse non-supervisée des données, et non rechercher seulement ce que l'on veut ou sait pouvoir y trouver.

1.1.2 Post-doctorat et projet de recherche initial

Mon séjour post-doctoral à Los Angeles dans le laboratoire de Simon Tavaré (University of Southern California) fut l'occasion d'une immersion dans l'univers très riche de la génétique des populations. Les modèles de Markov le long des séquences que j'avais déjà utilisés ne peuvent prétendre le plus souvent qu'à une description empirique des données. Un des attraits de la génétique des populations est que l'on y trouve des modèles probabilistes qui, même avec très peu de paramètres, ont une vocation presque explicative puisque l'aléatoire y capture un processus biologique réel. Je citerai notamment les très connus modèles de population proposé par Moran (Moran, 1958) ou de coalescent de Kingman pour des généalogies Kingman (1982). Ce domaine m'a aussi donné l'occasion de réaliser la difficulté intrinsèque de certains problèmes statistiques. C'est par exemple le cas de l'estimation du taux de mutation populationnel à partir d'un échantillon d'une population clonale à cause des dépendances très fortes au sein des observations (Tavaré,

2004). Enfin, j'ai été frappé par l'audace et le foisonnement des idées introduites pour l'inférence : MCMC sur des arbres Wilson & Balding (1998), *Approximate Bayesian Computation* (Tavaré *et al.*, 1997), *Composite likelihood* (Hudson, 2001), approximation de modèles (Li & Stephens, 2003).

Le projet de recherche que j'ai rédigé en 2004 (peu après le début de mon post-doctorat) pour ma candidature au poste de Chargé de Recherche à l'INRA dans le département Mathématiques et Informatiques Appliquées (MIA) était centré sur l'étude des régulations génétiques chez les bactéries. L'idée principale était de développer des méthodes pour la découverte de motifs régulateurs en y intégrant la comparaison de génomes proches afin d'accorder une attention particulière aux régions non-codantes soumises à sélection purificatrice (conservation « active » évitant une perte de fonctionnalité). Cette question de recherche, inspirée par le succès de travaux tels que ceux de Kellis *et al.* (2003), était motivée par la disponibilité croissante de génomes de bactéries proches. Elle faisait aussi écho à un intérêt pour la génomique comparative et les analyses évolutives alimenté par le thème de mon post-doctorat.

1.1.3 Recrutement à l'INRA

J'ai finalement décidé d'aborder l'étude des régulations génétiques sous l'angle de l'analyse et de l'intégration de données transcriptomiques. Les raisons qui ont conduit à cette orientation sont de plusieurs ordres. L'une d'elles est que je suis devenu un peu moins confiant dans la pertinence et la généralité de l'approche consistant à s'appuyer sur la conservation de séquence comme angle méthodologique principal pour identifier les régulons chez les bactéries. En effet, le détail des régulations et notamment la composition des régulons (ensemble des gènes cibles d'un facteur de transcription) peut évoluer rapidement. De plus, la méthodologie est par construction biaisée vers la découverte de régulations conservées, tandis que l'origine de nombreux traits phénotypiques d'intérêt serait plutôt à chercher au sein des régulations non conservées. Parallèlement, j'ai en partie satisfait mon intérêt pour la génomique comparative et la génomique des populations en nouant des collaborations avec des microbiologistes du centre de Jouy-en-Josas autour de thèmes autres que la recherche des motifs régulateurs : phylogénie du genre *Lactobacillus* avec Maarten van de Guchte (Nicolas *et al.*, 2007b) ; *Multi-Locus Sequence Typing* des *Flavobacteriacae* avec Éric Duchaud (Nicolas *et al.*, 2008) ; recherche de petits gènes chez les *Streptococcus* avec Rozenn Gardan (Ibrahim *et al.*, 2007). Enfin et surtout, la participation au projet européen Basysbio qui a démarré en septembre 2006 (coordonné par Philippe Noirot) m'a offert d'excellentes opportunités pour collecter et analyser des données transcriptomiques sur *B. subtilis*.

Ce thème de recherche autour de la transcriptomique de *B. subtilis* me sembla à l'époque d'autant plus pertinent que le développement récent des approches de transcriptomique globale (*genome-wide*), avec l'avènement des tiling arrays (Bertone *et al.*, 2004) et peu après du RNA-Seq (Nagalakshmi *et al.*, 2008), constituait une rupture réelle avec l'approche plus ancienne des puces à ADN (Schena *et al.*, 1995). Ne ciblant pas des régions pré-sélectionnées du génome, celles-ci offraient la possibilité de revisiter les questions d'annotation structurale que j'avais déjà commencées à aborder

dans mes travaux de thèse et leurs prolongements (détection des gènes, promoteurs, ...). La technologie des *tiling arrays* fut choisie pour des raisons chronologiques. Ce choix fut ensuite conforté par l'absence d'information sur la direction des transcrits avec les premiers avatars du RNA-Seq et par l'absence de queue poly-A permettant de cibler facilement le séquençage sur les ARN messagers chez les bactéries.

Ces orientations thématiques que sont les modèles à variables latentes, l'analyse de séquences génomiques du point de vue de l'annotation fonctionnelle et du point de vue évolutif, l'analyse de données transcriptomiques, constituent le cadre dans lequel se sont inscrits mes travaux. Ceux-ci ont porté à la fois sur des développements méthodologiques et sur l'analyse de jeux de données. Les continuités thématiques m'ont donné la satisfaction de voir des questions de recherche et des collaborations se développer et évoluer progressivement dans le temps. À quelques exceptions près, mes travaux portent plus ou moins directement sur des questions relatives aux bactéries. Outre l'importance de ces micro-organismes dans la biosphère et les activités humaines (médecine, agro-alimentaire, industrie), mon attirance pour ces organismes vient aussi de leur relative « simplicité » pour des organismes vivants autonomes, par rapport aux eucaryotes. Cette simplicité, qui se traduit notamment par l'absence de noyau et d'organites et un mode vie largement unicellulaire, me semble présenter l'avantage de mettre l'information génétique et les processus biochimiques au plus près des questions de physiologie et d'adaptation. Le département de microbiologie (MICA) se trouve aussi être avec le département MIA la seconde tutelle du laboratoire MaIAGE (et antérieurement de MIG).

1.2 Cartographie des travaux publiés

L'objectif de cette section est de donner un aperçu de l'ensemble de mes travaux publiés. Je donnerai dans les chapitres 2 et 3 beaucoup plus des détails et des éléments de bibliographie sur les travaux relatifs aux modèles à variables latentes et à certaines questions biologiques plus en détail.

Les figures 1.1 et 1.2 ont vocation à représenter graphiquement les liens entre ces 52 publications et la façon dont elles s'inscrivent dans les grandes orientations thématiques qui viennent d'être évoquées dans la section 1.1. On peut connecter la grande majorité de mes travaux à ceux dans lesquels ma contribution relève en premier lieu d'un développement méthodologique. Ceux-ci sont représentés en vert dans les figures 1.1 et 1.2.

Je fais ci-dessous l'exercice un peu artificiel d'énumérer ces travaux afin d'en verbaliser les relations.

1.2.1 HMM : des hétérogénéités de composition à l'annotation des génomes

L'article Nicolas *et al.* (2002) fut publié pendant ma thèse et constitue un point de départ important pour beaucoup des questions développées dans mes travaux suivants. Il rapporte l'utilisation d'un modèle de chaîne de Markov caché (abrégé HMM pour *hidden Markov model*) à quelques états pour segmenter un génome bactérien afin de rendre compte

des hétérogénéités de composition. En l'occurrence le génome était celui de *B. subtilis* dans lequel un des principaux facteurs d'hétérogénéité est lié aux transferts génétiques horizontaux. L'exploration de ces résultats a été l'une des motivations d'un outil puissant de visualisation de données (notamment des courbes) le long des génomes (Hoebeke *et al.*, 2003).

Le HMM utilisé pour l'étude des hétérogénéités de composition dans Nicolas *et al.* (2002)) n'est pas structuré au sens où tous les états sont inter-connectés. Un des prolongements a été le développement d'un modèle fortement structuré visant a rendre compte beaucoup plus finement des hétérogénéités de composition tout en délimitant les régions codantes pour les protéines. L'estimation non-supervisée des paramètres par l'algorithme EM constitue le cœur de ma contribution à la chaîne d'annotation des génomes microbiens AGMIAL (Bryson *et al.*, 2006) et à l'annotation des génomes de deux bactéries très différentes d'intérêt pour l'INRA. La première, *Lactobacillus delbrueckii* subsp. *bulgaricus* (van de Guchte *et al.*, 2006), est une bactérie à Gram positif (du phylum des *Firmicutes*) dite lactique (ordre Lactobacillales aussi connu sous le nom de *lactic acid bacteria*). Elle intervient dans la fabrication des yaourts. La seconde, *Flavobacterium psychrophilum* (Duchaud *et al.*, 2007), est une bactérie pathogène des poissons d'eau douce d'un phylum de bactéries à Gram négatif (les *Bacteroidetes*) qui est phylogénétiquement très éloigné des deux principales bactéries modèles *Escherichia coli* et *B. subtilis*.

Le développement de cet algorithme de détection de gènes codant pour des protéines a aussi été valorisé par une recherche systématique des petits gènes dans le genre *Streptococcus* (Ibrahim *et al.*, 2007). Les prédictions de l'algorithme y sont combinées à des indices pouvant les conforter : conservation de séquence et hétérogénéité des taux de substitution entre les trois positions des codons. Les *Streptococcus* constituent un autre genre de bactéries à Gram positif du phylum des *Firmicutes* d'intérêt considérable pour l'homme de part la présence de plusieurs bactéries pathogènes, de bactéries trouvées dans la flore commensale (microbiote), et de bactéries d'intérêt agro-alimentaires car impliquées dans la fermentation des produits laitiers comme *Streptococcus thermophilus*. Le lien entre une famille de protéines régulatrices, dites des Rgg, et des petits peptides codés par des gènes révélés par notre étude a ensuite fait l'objet de travaux approfondis, notamment de la part du groupe de Rozenn Gardan. C'est dans ce contexte que j'ai plus tard réalisé une analyse comparative de la distribution des régulateurs Rgg et des peptides associés montrant que l'on trouve des régulateurs Rgg associés à de tels petits peptides chez de très nombreux *Streptococcus*. Cette analyse est publiée dans Fleuchot *et al.* (2011) avec une dissection expérimentale des briques d'un système de *quorum-sensing* dans lequel les peptides détectés jouent le rôle de phéromones.

Parallèlement à la détection de gènes, un autre prolongement de mon intérêt pour les applications des HMM fut la recherche de motifs promoteurs (les sites de fixations des facteurs sigma de l'ARN polymérase bactérienne). Ainsi, Nicolas *et al.* (2006a) présente un modèle semi-markovien et un algorithme MCMC trans-dimensionnels pour l'identification de motifs en deux parties dans les régions promotrices alors seulement définies comme les régions intergéniques en amont de codons de démarrage de traduction. Dans ce travail, l'identification de sites pour des facteurs sigma alternatifs ne fut possible qu'en sous-

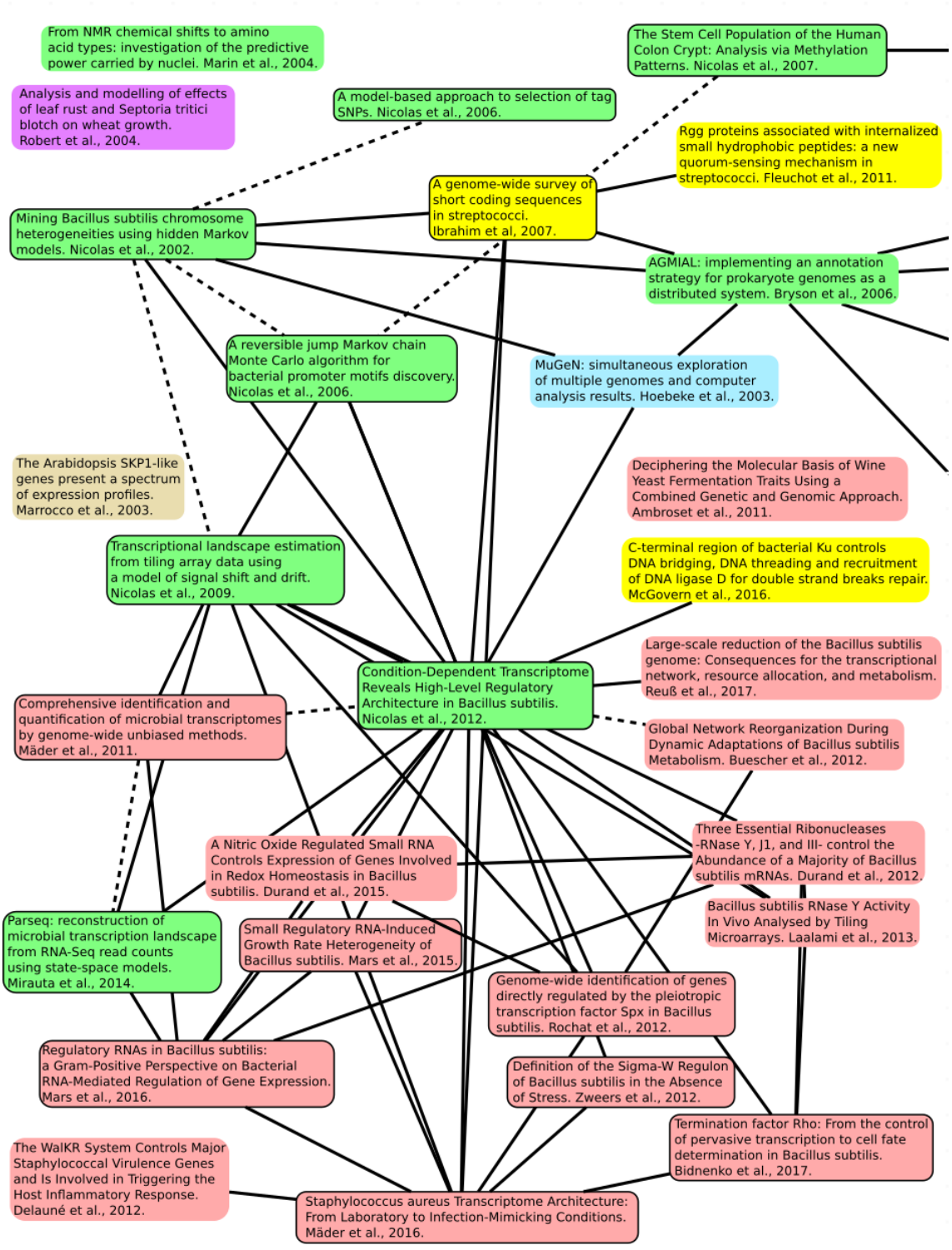


Figure 1.1 – Carte de l'ensemble des articles publiés (page 1/2).
Un lien est tracé entre deux articles sous la forme d'une ligne continue lorsque l'un des articles cite l'autre. Les lignes pointillées n'ont pas donné lieu à des citations, ils représentent souvent des liens moins formels (notamment des liens thématiques). Dans la mesure du possible, j'ai essayé de placer les travaux les plus anciens plutôt en haut à gauche et les travaux récents en bas ou à droite.

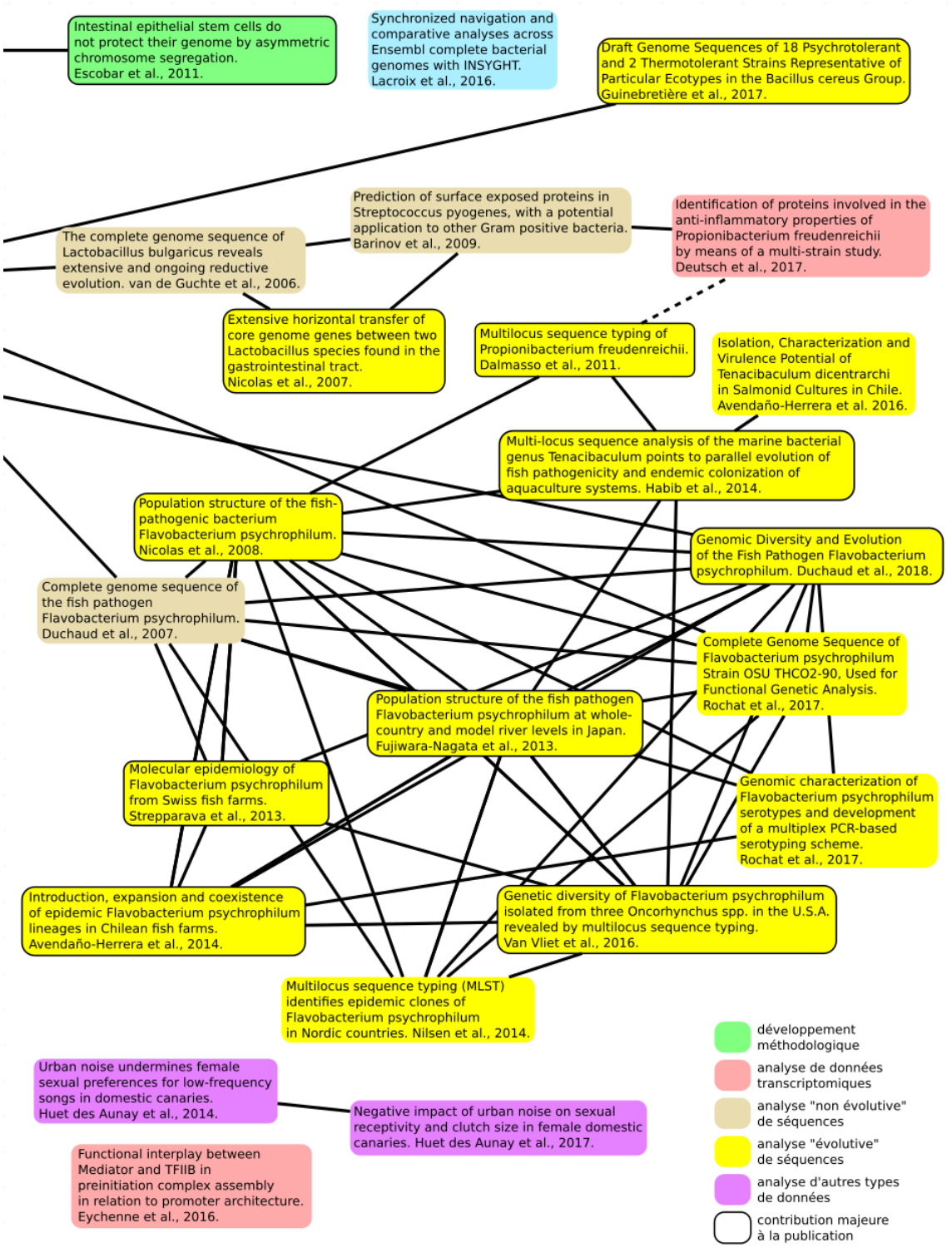


Figure 1.2 – Carte de l'ensemble des articles publiés (page 2/2).

échantillonnant le jeu de séquences en s'appuyant sur des données transcriptomiques acquises sur des mutants. Ce modèle a été étendu dans Nicolas *et al.* (2012) pour permettre une classification des régions promotrices identifiées par *tiling arrays* et en s'aidant des profils d'expression à travers les conditions.

1.2.2 Modèles en lien avec la génétique des populations

Il s'agit ici de travaux en rapport avec le thème de mon séjour post-doctoral. En plus de l'intérêt des modèles mis en œuvre, ils ont été l'occasion d'excursions sur des problématiques liées aux eucaryotes.

Nicolas *et al.* (2006b) propose une nouvelle utilisation à un HMM pour les allèles portés par un haplotype lorsque celui-ci est tiré dans une population pour laquelle on connaît déjà un certains nombre d'haplotypes. L'objectif était de sélectionner des combinaisons de SNP qui portent autant d'information que possible afin d'être utilisés comme marqueurs génétiques (*tag SNPs*) et ensuite être génotypés sur beaucoup d'individus. Il s'agissait à cette époque d'une problématique typique des études d'association dans lesquelles on cherche par exemple à identifier les régions du génome portant des facteurs génétiques de maladies (il faut des recombinaisons méiotiques pour que l'approche ait un sens, elle n'est donc pas réellement applicable chez les bactéries). Bien que l'approche de sélection de *tag SNPs* proposée soit originale et donne de bons résultats, il n'y a pas eu, à ma connaissance, d'application réelle de la méthode. En fait, l'idée de sélectionner un ensemble dense de *tag SNPs* est devenue relativement obsolète avec la diminution rapide du coût d'un séquençage global du génome.

Une toute autre méthodologie est mise en œuvre dans Nicolas *et al.* (2007a). Ce travail aborde l'analyse des données épigénétiques pour l'inférence des propriétés des généalogies de cellules dans l'organisme. Plus spécifiquement, il s'agissait d'utiliser des profils de méthylation à un locus donné au sein d'un échantillon de cellules afin d'étudier le processus de renouvellement des cellules différenciées dans l'épithélium du colon humain. Cet épithélium est organisé en cryptes d'environ 2 000 cellules contenant les cellules souches et leur descendance. Nicolas *et al.* (2007a) décrit un modèle probabiliste fondé sur le coalescent de Moran et un algorithme MCMC pour l'estimation des paramètres qui permit notamment d'obtenir une estimation du nombre de cellules souches par crypte. Escobar *et al.* (2011) résulte d'une collaboration postérieure au post-doctorat avec le groupe de Catherine Legraverend à l'université de Montpellier qui se situe dans le prolongement du travail sur la modélisation des généalogies dans les cryptes de l'épithélium intestinal. Ici la question abordée est celle de l'existence ou non d'une ségrégation asymétrique des brins d'ADN lors de la division cellulaire qui permettrait une conservation préférentielle du brin matrice dans les lignées de cellules souches des cryptes et ainsi de diminuer l'impact des mutations somatiques. Des simulations de la décroissance de la quantité d'ADN marqué avec un analogue de la thymine suivi par immunofluorescence (CldU) dans les cellules d'une crypte sous différents modèles ont été réalisées et comparées à des données expérimentales obtenues chez la souris.

1.2.3 Transcriptomique : modélisation statistique et analyse de données

J'ai d'abord eu deux occasions de percevoir l'utilité des données transcriptomiques dans un contexte de recherche de motifs régulateurs dans les séquences d'ADN avec des outils existants (Marrocco *et al.* 2003), puis avec mes propres algorithmes (Nicolas *et al.*, 2006a). Cependant, l'angle sous lequel j'ai réellement commencé à m'intéresser aux données transcriptomiques est celui de l'annotation structurale, permis par l'avènement des approches globales (un point de vue que nous avons notamment développé dans l'article de revue Mäder *et al.* (2011)). C'est dans cette perspective que j'ai développé un modèle statistique pour « lisser » le long du génome l'intensité de fluorescence mesurée par *tiling arrays* et y détecter les ruptures (Nicolas *et al.*, 2009). Ce HMM utilise une variable cachée à espace d'état discret pour refléter le niveau d'expression mais il approche un modèle où la variable cachée est continue (modèles connus sous le nom de *State Space Models*). Les paramètres en sont estimés par maximum de vraisemblance (algorithme EM). L'extension de ce travail aux données RNA-Seq à fait l'objet de la thèse de Bogdan Mirauta (Mirauta *et al.*, 2014) que j'ai coencadrée avec Hugues Richard (sous la direction d'Alessandra Carbone). Le modèle qui y est développé repose sur des variables à espace d'état continu et l'estimation utilise un algorithme MCMC impliquant des étapes de mise à jour de ces variables par SMC (*Sequential Monte Carlo*).

L'algorithme de lissage de données de *tiling arrays* (Nicolas *et al.*, 2009) (combiné à l'outil de visualisation développé dans Hoebeke *et al.* (2003)) a joué un rôle primordial dans l'intégration de grandes collections de profils transcriptomiques ayant vocation à représenter les différents états physiologiques de la bactérie. Ces travaux se sont concrétisés par la ré-annotation structurale des génomes de *B. subtilis* (Nicolas *et al.*, 2012) et de *Staphylococcus aureus* (Mäder *et al.*, 2016). D'un point de vue méthodologique, ce contexte a aussi fourni l'occasion du développement d'un algorithme pour la classification des promoteurs selon les facteurs sigma qui les contrôlent dont le rôle a aussi été très important dans l'interprétation des résultats (Nicolas *et al.*, 2012). D'un point de vue biologique, je me suis intéressé à l'interprétation des ARN antisens et j'ai étayé dans deux articles (Nicolas *et al.*, 2012 ; Mäder *et al.*, 2016) l'idée selon laquelle une partie non négligeable d'entre eux pourraient refléter des imperfections du contrôle de la transcription et une fonction du facteur de terminaison Rho serait de contribuer à contenir ces transcrits. Dans la continuité de la question du rôle des antisens, je me suis beaucoup investi dans une étude qui caractérise l'impact « phénotypique » de la délétion du facteur Rho chez *B. subtilis* et dissèque les mécanismes de régulations sous-jacents (Bidnenko *et al.*, 2017).

Les travaux d'intégration de données transcriptomiques chez *B. subtilis* et *S. aureus* et d'annotation de l'architecture de leurs transcriptomes ont conduit à la participation à de nombreuses collaborations dont on peut tenter une classification des objectifs en quatre catégories : analyse d'une réponse à un changement de milieu (Buescher *et al.* 2012) ; analyse de régulons de facteurs de transcription (Delauné *et al.*, 2012 ; Rochat *et al.*, 2012, Zweers *et al.*, 2012) ; recherche de cibles des principales RNases de *B. subtilis* (Durand *et al.*, 2012, Laalami *et al.*, 2013) ; et enfin, analyse de la fonction de petits ARN non-codants (Durand *et al.*, 2015, Mars *et al.*, 2015), thème sur lequel j'ai aussi participé à la rédaction

d'un article de revue (Mars *et al.*, 2016).

À côté de ces contributions à l'analyse de jeux de données d'expression, j'ai aussi eu quelques occasions de participer à l'analyse de données d'interactions protéine-ADN obtenues par chromatine-immunoprécipitation suivie soit d'hybridation sur puce à ADN (ChIP-chip) (Buescher *et al.* 2012, Rochat *et al.* 2012) soit de séquençage (Chip-Seq) (Eychenne *et al.* 2016). Cette dernière publication a une place un peu particulière dans mes travaux. Bien que rattachée à mon thème de recherche sur les propriétés des promoteurs, il s'agit ici d'un travail sur organisme eucaryote (la levure modèle *Saccharomyces cerevisae*). Les mécanismes d'activation de la transcription y sont bien plus complexes que chez les bactéries.

1.2.4 Analyses de quelques familles de gènes

J'ai déjà mentionné l'analyse sur la famille des régulateurs Rgg dans le genre *Streptococcus* en lien avec mes travaux sur la détection de gènes (Fleuchot *et al.*, 2011). J'ai mené une démarche un peu similaire dans McGovern *et al.* (2016) pour étudier la distribution phylogénétique et les propriétés de composition en domaines des protéines Ku. L'étude visait à caractériser la fonction moléculaire de l'extension basique souvent trouvée en extrémité C-terminale de ces protéines impliquées dans la réparation des cassures d'ADN double brin chez les bactéries.

Chacune de ces deux études portait sur une famille de gène d'intérêt et avaient une vocation à éclairer la fonction de certaines séquences (extension C-terminale, peptide).

1.2.5 Analyses de la structure de populations bactériennes

Dans une perspective bien différente des analyses ciblées sur une famille de gènes que je viens de mentionner, j'ai conduit des études visant à éclairer les relations entre bactéries au sein d'un genre ou d'une espèce.

Ce fut tout d'abord le cas de mon travail sur la phylogénie des lactobacilles du « complexe *acidophilus* » à partir d'alignements de gènes conservés dans les génomes de *L. delbrueckii* subsp. *bulgaricus* (van de Guchte *et al.*, 2006), *L. acidophilus* et *L. johnsonii* ainsi que dans ceux de deux lactobacilles servant de références externes (*outgroups*). Les résultats publiés dans Nicolas *et al.* (2007b) ont révélé des échanges génétiques massifs entre les deux lactobacilles *L. acidophilus* et de *L. johnsonii* trouvés dans le tube digestif. Ces échanges engendrent des discordances entre les généalogies des familles de gènes et la généalogie des espèces au point de perturber la reconstruction de l'arbre des espèces.

Jusqu'à récemment la densité d'échantillonage atteignable par séquençage de génomes complets avec un budget raisonnable était assez limitée. Je me suis donc investi dans des études reposant sur le séquençage d'un petit nombre de gènes marqueurs choisis pour être représentatif du génome central (gènes conservés et en une seule copie dans un groupe taxonomique donné). Cette approche est connue sous le nom de MLST (pour *Multi-Locus Sequence Typing*) lorsque le groupe taxonomique étudié est une espèce ou MLSA (*Multi-Locus Sequence analysis*) lorsque le groupe est plus vaste (par exemple un genre). Dans le prolongement de l'annotation du génome complet de *F. psychrophilum* (Duchaud *et*

al., 2007), je me suis fortement impliqué dans l'analyse par MLST de cette espèce en profitant notamment de la collection de souches détenue par le groupe d'Éric Duchaud, notamment grâce à l'implication ancienne de Jean-François Bernadet dans les questions de taxonomie et de typage des *Flavobacteriacae*. Une première publication a jeté les bases de la description de la structure génétique globale de l'espèce Nicolas *et al.* (2008)). Celle-ci s'est révélée très fortement recombinogène.

J'ai ensuite été impliqué, à travers des collaborations internationales et à des degrés divers, dans cinq études plus ciblées sur la structure de population de *F. psychrophilum* en relation avec différentes zones géographiques et/ou poissons hôtes (Fujiwara-Nagata *et al.*, 2013 ; Strepparava *et al.*, 2013 ; Avendaño-Herrera *et al.*, 2014 ; Nilsen *et al.*, 2014 ; Van Vliet *et al.*, 2016). Toujours chez les *Flavobacteriacae*, j'ai conduit une analyse MLSA du genre *Tenacibaculum* responsable de pathologies des poissons marins (Habib *et al.*, 2014) qui s'est aussi traduit par une publication annexe (Avendaño-Herrera *et al.*, 2016). Enfin, j'ai réalisé une analyse de séquences de génomes complets qui a notamment permis mieux caractériser les recombinaisons au sein de l'espèce et de proposer une date pour l'émergence d'un des principaux complexes clonaux (Duchaud *et al.*, 2018).

De façon plus ponctuelle, j'ai aussi réinvesti l'expérience acquise en MLST sur *F. psychrophilum* en contribuant à une caractérisation de la structure de population d'une autre bactérie, *Propionibacterium freudenreichii*, non pathogène celle-ci, mais d'intérêt agro-alimentaire car utilisée dans la fabrication de certains fromages (Dalmasso *et al.*, 2011).

1.2.6 Recherche des facteurs génétiques de phénotypes d'intérêt

Bien qu'elle ne soit pas strictement au cœur de mes travaux, la question de l'établissement de liens entre gènes et phénotypes d'intérêt (médical, agro-alimentaire, industriel, ...) constitue un point de rencontre de différentes thématiques abordées ici et représente un enjeu majeur de la recherche appliquée en microbiologie. Aborder cette question implique typiquement des analyses intégratives et/ou comparatives.

Ainsi l'étude MLST sur *P. freudenreichii* a été suivie d'une une autre étude sur les génomes complets s'appuyant sur des analyses phénotypiques, transcriptomiques et protéomiques pour tenter d'identifier les gènes responsables des propriétés anti-inflammatoires de certaines souches (Deutsch *et al.*, 2017). Ces souches pourraient éventuellement avoir un intérêt en tant que probiotiques. Ici mon implication portait uniquement sur les analyses transcriptomiques. On peut aussi noter que cette publication utilise les résultats d'une étude antérieure à laquelle j'avais été marginalement impliqué et qui concernait la prédiction des gènes de surfaces chez les bactéries à Gram positif (Barinov *et al.*, 2009).

Rochat *et al.* (2017a) rapporte la séquence du génome de la seule souche de *F. psychrophilum* connue pour être transformable expérimentalement. Cette propriété permet de réaliser des modifications génétiques et joue en cela un grand rôle lorsqu'il s'agit de vérifier un lien entre un gène et un phénotype. En partant d'une recherche d'association statistique entre la présence/absence de gènes et des sérotypes utilisés pour le typage *F. psychrophilum*, Rochat *et al.* (2017b) identifie le locus à l'origine des antigènes de surface

déterminant les sérotypes. Ambroset *et al.* (2011) utilise une technique complètement différente pour identifier chez la levure *S. cerevisiae* des gènes impliqués dans des phénotypes d'intérêts industriels et agro-alimentaires liés à la fermentation alcoolique. Cette approche fondée sur la recherche simultanée de QTLs phénotypiques et de QTLs d'expression, associée à la connaissance des génomes des souches parentales, est très puissante mais n'est malheureusement pas transposable aux bactéries à cause de l'absence de recombinaison méiotique.

Guinebretiere *et al.* (2017) rapporte le séquençage de souches du groupe *Bacillus cereus* tolérantes au froid mais l'étude n'est pas poussée jusqu'à l'identification de facteurs génétiques qui pourraient être associés à ce phénotype déterminant la capacité de contamination des produits alimentaires. Enfin, j'ai contribué marginalement au développement d'un navigateur pour explorer les données de génomes complets bactériens par alignement des régions synthéniques (Lacroix *et al.*, 2015). Ce navigateur inclut maintenant aussi des recherches de liens entre présence/absence de gènes et caractères phénotypiques renseignés par l'utilisateur. Un des enjeux est de donner au biologiste les outils lui permettant d'explorer la quantité croissante de génomes bactériens disponibles. Le 11 janvier 2018, la *NCBI Reference Sequence Database* donnait accès à 103 929 génomes représentatifs d'environ 5 589 « espèces » (*reference genomes* + *representative genomes*).

1.2.7 Quelques travaux plus isolés

Je clôturerai ce chapitre en mentionnant des contributions à quatre études publiées qui correspondent à des travaux plus éloignés de mes grands axes thématiques. Dans Robert *et al.* (2004), il s'agissait d'aider à formuler mathématiquement la perte de rendement causée par la colonisation d'un blé par la rouille brune. Dans Marin *et al.* (2004), j'ai proposé la mise en œuvre d'un modèle de mélange de lois normales multivariées pour identifier les acides aminés à partir de spectres obtenus par résonance magnétique nucléaire (RMN) dans un contexte de détermination expérimentale de structures protéiques. Enfin, dans Huet des Aunay *et al.* (2014) et Huet des Aunay *et al.* (2017), j'ai analysé des données mesurées sur des femelles canaris en réponse à des bruits d'origine anthropique (nombre d'œufs ou réponse binaire à un stimulus). Les modèles mis en œuvre sont des modèles linéaires généralisés impliquant des effets mixtes, pour tenir compte des répétitions de mesure sur les même oiseaux.

Chapitre 2

Les modèles à variables latentes, une approche pour l'analyse de structures cachées dans les génomes

Le développement et l'utilisation de modèles impliquant des variables latentes, et en particulier des chaînes de Markov cachées, constitue un thème récurrent et une ligne directrice de mes travaux de recherches.

L'objectif de ce chapitre est de présenter de façon assez synthétique les principaux modèles et algorithmes que j'ai mis en œuvre. En cohérence avec ma démarche de modélisation et mon positionnement scientifique, j'ai fait le choix d'un regroupement des travaux selon le type des données modélisées : alternance de régions de différentes compositions le long d'un génome, modélisation des motifs dans les séquences promotrices, modélisation du polymorphisme, modélisation de données d'expression (profil transcriptomique) le long d'une séquence. Ce découpage diffère en partie de celui que l'on aurait obtenu en s'appuyant sur le type de modèle (type de variables cachées ou structure de dépendance entre elles), le cadre méthodologique de l'inférence (EM vs. MCMC, fréquentiste vs. Bayésien), ou la seule chronologie des travaux.

2.1 Modèles de séquences d'ADN pour la segmentation d'un génome

2.1.1 Un HMM non-structuré pour l'étude des hétérogénéités de composition

Le modèle utilisé dans Nicolas *et al.* (2002) est l'un des HMM les plus simples que l'on puisse imaginer pour les dépendances longitudinales dans une séquence d'ADN. Il permet de rendre compte de l'alternance de régions de compositions différentes et, au sein de chaque région, de la composition en mots d'une longueur choisie.

Afin de rendre plus concrets les modèles discutés dans ce mémoire et étant donné la place qu'occupent les HMM dans mes travaux, je donne ici quelques détails sur ce modèle et les algorithmes utilisés pour l'inférence et la « reconstruction » de la chaîne cachée.

Modèle

On notera $X_{1:\mathcal{T}} = (X_1, \ldots, X_\mathcal{T})$ le vecteur aléatoire modélisant la séquence d'ADN. En plus de $X_t \in \{\mathrm{a}, \mathrm{c}, \mathrm{g}, \mathrm{t}\}$, à chaque position t de la séquence correspond une variable cachée $S_t \in \{1, \ldots, \mathcal{S}\}$ à valeur dans un espace discret et fini qui indique le type de région.

L'alternance des régions dans la séquence $S_{1:\mathcal{T}}$ est modélisée par une chaîne de Markov d'ordre 1, c'est-à-dire que la probabilité de $S_t = u$ sachant $S_{1:t-1} = s_{1:t-1}$ ne dépend que de s_{t-1}. On a ainsi $\pi(S_t = u \mid S_{1:t-1} = s_{1:t-1}) = \pi(S_t = u \mid S_{t-1} = s_{t-1})$. Pour ne pas surcharger les notations et conformément à l'usage, on s'autorisera à écrire simplement s_{t-1} pour désigner l'événement $S_{t-1} = s_{t-1}$ lorsque cela ne prête pas à confusion, et de même pour toutes les autres variables aléatoires.

La loi du nucléotide X_t à la position t, dite « loi d'émission », dépend de l'état caché sous-jacent s_t et éventuellement des r nucléotides qui précèdent, $x_{t-r:t-1}$, on a donc que $\pi(x_t \mid s_{1:t}, x_{1:t-1}) = \pi(x_t \mid s_t, x_{t-r:t-1})$. Cette dépendance markovienne d'ordre r sur les observations permet de rendre compte de la composition en mots de longueur $r + 1$. Les dépendances dans ce modèle peuvent être représentées sous la forme d'un graphe acyclique dirigé (DAG pour *Directed Acyclic Graph*), que l'on montre ici pour $r = 1$

$$\begin{array}{ccccccccc} \ldots & \longrightarrow & S_{t-1} & \longrightarrow & S_t & \longrightarrow & S_{t+1} & \longrightarrow & \ldots \\ & & \downarrow & & \downarrow & & \downarrow & & \\ \ldots & \longrightarrow & X_{t-1} & \longrightarrow & X_t & \longrightarrow & X_{t+1} & \longrightarrow & \ldots . \end{array}$$

Cette représentation de la décomposition de la vraisemblance sous forme de graphe est un outil puissant pour explorer les relations d'indépendances conditionnelles qui sont au cœur des algorithmes de programmation dynamique permettant de « reconstruire » la chaîne des états cachées et, par là même, aussi essentielles pour estimer les paramètres. En particulier, du DAG, on dérive facilement le graphe moral qui en est une version non dirigée dans laquelle des arcs ont été ajoutés entre toutes les paires de variables parents directs d'une même variable

$$\begin{array}{ccccccccc} \ldots & - & S_{t-1} & - & S_t & - & S_{t+1} & - & \ldots \\ & \swarrow & | & \swarrow & | & \swarrow & | & \swarrow & \\ \ldots & - & X_{t-1} & - & X_t & - & X_{t+1} & - & \ldots . \end{array}$$

Si A, B, et C sont des variables ou groupes de variables et qu'il n'existe pas de chemin reliant A à B dans le graphe moral sans passer par C, alors A et B sont indépendantes conditionnellement à C. Ainsi, on vérifie par exemple aisément $\pi(S_t = u \mid S_{t+1} = v, x_{1:\mathcal{T}}) = \pi(S_t = u \mid S_{t+1} = v, x_{1:t})$, nécessaire pour l'obtention de la relation au cœur de la récursion *backward* (Eq. 2.3).

Paramètres

Les paramètres de ce modèle sont : la matrice de transition de la chaîne cachée notée $\alpha = (\alpha_{u,v})_{(u,v)\in\{1\dots\mathcal{S}\}^2}$, où $\alpha_{u,v} = \pi(S_{t+1} = v \mid S_t = u)$ avec $\alpha_{u,v} \geq 0$ et $\sum_{v\in\{1,\dots,\mathcal{S}\}} \alpha_{u,v} = 1$; et $\beta = (\beta_{u,w,x})_{u\in\{1\dots\mathcal{S}\},(w,x)\in\{\mathrm{a,c,g,t}\}^{r+1}}$ qui contient les probabilités d'émission de chaque nucléotide conditionnellement à l'état caché et au mot de longueur r qui précède, $\beta_{x;u,w} = \pi(X_t = x \mid S_t = u, X_{t-r}^{t-1} = w)$ avec $\beta_{x;u,w} \geq 0$ et $\sum_{x\in\{\mathrm{a,c,g,t}\}} \beta_{x;u,w} = 1$. Étant donnés les probabilités qui somment à 1, le nombre total de paramètres est de $\mathcal{S}\times(\mathcal{S}-1)+\mathcal{S}\times 3\times 4^r$. On mentionnera aussi les paramètres de la loi initiale qui gèrent les toutes premières positions ($t \leq 1$ pour $S_{1:\mathcal{T}}$ et $t \leq r$ pour $X_{1:\mathcal{T}} \mid S_{1:\mathcal{T}}$). Pour des séquences longues et des modèles pas trop fortement structurés (évidemment ergodiques) ces paramètres ont généralement une influence faible. On ne cherche généralement pas à les estimer.

Dans le cadre de la segmentation de la séquence selon sa composition, on s'intéresse typiquement à des valeurs de la matrice α avec une forte concentration sur la diagonale, c'est-à-dire $\alpha_{u,u}$ grand par rapport à $\alpha_{u,v\neq u}$. En effet, la longueur des segments définis par les « plages » de positions consécutives où $S_t = u$ sous ce modèle suit une loi géométrique de moyenne $1/(1-\alpha_{u,u})$.

Reconstruction du chemin caché

Pour une séquence observée $x_{1:\mathcal{T}}$ et connaissant la valeur des paramètres (α, β), il existe essentiellement trois façons de « reconstruire » la séquence des états cachés :

— le calcul de $\pi(S_t = u \mid x_{1:\mathcal{T}})$ pour $t = 1\dots\mathcal{T}$ et $u = 1\dots\mathcal{S}$,

— l'identification de $s^\star_{1:\mathcal{T}} = \arg\max_{s_{1:\mathcal{T}}\in\{1\dots\mathcal{S}\}^\mathcal{T}} \pi(s_{1:\mathcal{T}} \mid x_{1:\mathcal{T}})$,

— et enfin, l'échantillonnage de $S_{1:\mathcal{T}} \mid x_{1:\mathcal{T}}$.

Chacune de ces approches suppose des sommes, des maximisations, ou des tirages sur l'ensemble de toutes les valeurs possibles pour $s_{1:\mathcal{T}}$, dont la dimension $\mathcal{S}^\mathcal{T}$ rend impossible le calcul direct. Il en est de même pour la vraisemblance dite des données incomplètes, $\pi_{\alpha,\beta}(x_{1:\mathcal{T}}) = \sum_{s_{1:\mathcal{T}}} \pi_{\alpha,\beta}(x_{1:\mathcal{T}}, s_{1:\mathcal{T}})$ que l'on cherche par exemple typiquement à maximiser lorsque l'inférence est conduite dans un cadre fréquentiste.

En pratique, les calculs reposent donc sur des algorithmes dits de « programmation dynamique » qui s'appuient sur des récursions permettant de réaliser ces calculs en un nombre d'opérations bien inférieur à $\mathcal{S}^\mathcal{T}$. En toute généralité, le nombre d'opérations en est en $O(\mathcal{T}\times\mathcal{S}^2)$, mais il est possible de remplacer $\mathcal{S}^2$ par le nombre de termes non nuls dans la matrice de transition α. Ce sont ces algorithmes qui rendent possible l'utilisation des HMM. On rencontrera aussi dans la section 2.2.2 leurs analogues dans le cas de calculs sur des structures d'arbres et non de chaînes.

À titre d'illustration, je donne ici une version de l'algorithme *forward-backward* bien connu pour le calcul de $\pi(S_t = u \mid x_{1:\mathcal{T}})$ qui correspond précisément à celle utilisée dans Nicolas *et al.* (2002). Les intermédiaires de calcul permettent aisément d'aboutir aux deux autres façons de reconstruire $s_{1:\mathcal{T}}$. La variante conduisant au chemin de probabilité maximale $s^\star_{1:\mathcal{T}}$ est connue sous le nom d'algorithme de Viterbi.

L'algorithme se compose d'une récursion *forward* et d'une récursion *backward* qui consistent à parcourir la séquences dans un sens (t croissant de 1 à $\mathcal{T}$) puis dans l'autre

(de $\mathcal{T}$ à 1). Dans la récursion *forward*, on calcule $\pi(s_t \mid x_{1:t-1})$ à partir de $\pi(s_{t-1} \mid x_{1:t-1})$, c'est l'étape de « prédiction » réalisée en utilisant la relation

$$\pi(S_t = v \mid x_{1:t-1}) \quad = \quad \sum_{u=1\ldots\mathcal{S}} \pi(S_{t-1} = u \mid x_{1:t-1}) \alpha_{u,v} \,, \text{ pour } v = 1 \ldots \mathcal{S} \,, \tag{2.1}$$

puis on calcule $\pi(s_t \mid x_{1:t})$ en utilisant les $\pi(s_t \mid x_{1:t-1})$, c'est l'étape de « filtrage » fondée sur

$$\pi(S_t = u \mid x_{1:t}) \quad = \quad \frac{\beta_{x_t;u,x_{t-r:t-1}} \pi(S_t = u \mid x_{1:t-1})}{\sum_{v=1\ldots\mathcal{S}} \beta_{x_t;v,x_{t-r:t-1}} \pi(S_t = v \mid x_{1:t-1})} \,, \text{ pour } u = 1 \ldots \mathcal{S} \,. \tag{2.2}$$

Dans la récursion *backward*, on réalise l'étape de « lissage » qui consiste à calculer $\pi(s_t \mid x_{1:\mathcal{T}})$ à partir de $\pi(s_t \mid x_{1:t})$, $\pi(s_{t+1} \mid x_{1:t})$ et $\pi(s_{t+1} \mid x_{1:\mathcal{T}})$ en utilisant

$$\pi(S_t = u \mid x_{1:\mathcal{T}}) \quad = \quad \sum_{v=1\ldots\mathcal{S}} \frac{\pi(S_t = u \mid x_{1:t}) \alpha_{u,v}}{\pi(S_{t+1} = v \mid x_{1:t})} \pi(S_{t+1} = v \mid x_{1:\mathcal{T}}) \text{ pour } u = 1 \ldots \mathcal{S} \,. \tag{2.3}$$

Estimation des paramètres

Dans Nicolas *et al.* (2002), comme dans le reste de mes travaux, je me suis intéressé à l'utilisation des modèles à variables latentes dans un contexte d'estimation non-supervisée. Le principe repose sur l'obtention d'un jeu de paramètres $\hat{\theta} = (\hat{\alpha}, \hat{\beta})$ et d'une reconstruction des états cachés associés $\pi_{\hat{\theta}}(s_{1:\mathcal{T}} \mid x_{1:\mathcal{T}})$ qui « explique bien » les données $x_{1:\mathcal{T}}$ dans le sens ou $\pi_{\hat{\theta}}(x_{1:\mathcal{T}})$ est élevée. La méthode choisie dans Nicolas *et al.* (2002) est celle de l'algorithme EM (*Expectation-Maximization*) qui est certainement la plus populaire pour les HMM et, plus généralement, pour ceux des modèles à variables latentes où elle s'applique. Son objectif est la maximisation de la vraisemblance des données incomplètes, c'est-à-dire d'obtenir un $\hat{\theta}$ qui approche $\theta^{\text{MLE}} = \arg\max_\theta \pi_\theta(x_{1:\mathcal{T}})$. On parle ici pour $\pi_\theta(x_{1:\mathcal{T}})$ de vraisemblance des données incomplètes par opposition à $\pi_\theta(x_{1:\mathcal{T}}, s_{1:\mathcal{T}})$ où $s_{1:\mathcal{T}}$ serait connu et dont la maximisation est immédiate. L'idée de l'algorithme EM repose sur la décomposition

$$\log \pi_\theta(x_{1:\mathcal{T}}) \quad = \quad \underbrace{\mathbb{E}_{\theta^{(k)}} \log \pi_\theta(x_{1:\mathcal{T}}, s_{1:\mathcal{T}})}_{Q(\theta|\theta^{(k)})} - \underbrace{\mathbb{E}_{\theta^{(k)}} \log \pi_\theta(s_{1:\mathcal{T}} \mid x_{1:\mathcal{T}})}_{H(\theta|\theta^{(k)})} \,, \tag{2.4}$$

où l'on note $\mathbb{E}_\theta$ pour $\mathbb{E}_{\pi_\theta(s_{1:\mathcal{T}}|x_{1:\mathcal{T}})}$. $Q(\theta \mid \theta^{(k)})$ est ici facile à maximiser en θ à partir des $\pi_{\theta^{(k)}}(s_t \mid x_{1:\mathcal{T}})$ et $\pi_{\theta^{(k)}}(s_t, s_{t+1} \mid x_{1:\mathcal{T}})$ obtenus avec l'algorithme *forward-backward* et $H(\theta \mid \theta^{(k)}) \leq H(\theta^{(k)} \mid \theta^{(k)})$ (inégalité de Gibbs). Dès lors, étant donnée une valeur courante $\theta^{(k)}$, on prend $\theta^{(k+1)} = \arg\max_\theta Q(\theta \mid \theta^{(k)})$ ce qui a comme effet de faire croître la vraisemblance des données complètes. Pour le HMM considéré ici, l'algorithme ainsi obtenu converge vers un point où les dérivées partielles de θ s'annulent (un maximum, éventuellement local, ou un point de selle) (Muri, 1997).

Dans le cas de notre HMM, la maximisation de $Q(\theta \mid \theta^{(k)})$ revient à choisir $\theta^{(k+1)} =$

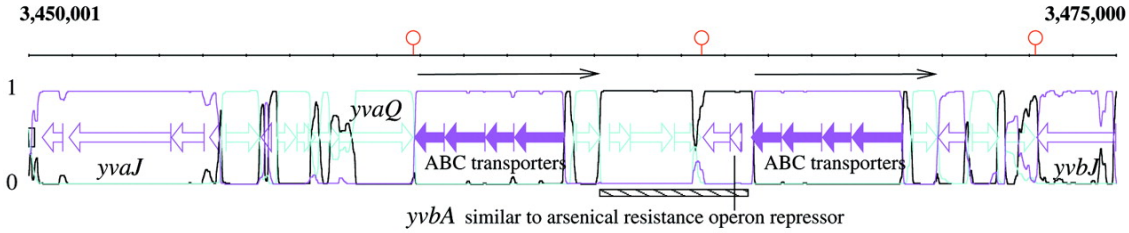


Figure 2.1 – On montre ici la probabilité $\pi_{\hat{\theta}}(s_t = u \mid x_{1:\mathcal{T}})$ pour un modèle où $S = 3$ et $r = 2$ qui aboutit à distinguer les deux sens des régions codantes (cyan et magenta) et regroupe dans un même état (noir) régions intergéniques et régions atypiques riche en $a + t$ issues de transferts horizontaux. Cette figure est extraite de Nicolas *et al.* (2002).

$(\alpha^{(k+1)}, \beta^{(k+1)})$ comme

$$\begin{aligned} \alpha_{u,v}^{(k+1)} &= \frac{n_{u,v}^{(k)}}{\sum_{v'} n_{u,v'}^{(k)}} \\ \beta_{x;u,w}^{(k+1)} &= \frac{n_{x;u,w}^{(k)}}{\sum_{x'} n_{x';u,w}^{(k)}}, \end{aligned} \tag{2.5}$$

où les $n^{(k)}$ correspondent à des comptages faisant intervenir des comptages pondérés selon $\pi_{\theta^{(k)}}(s_{1:\mathcal{T}} \mid x_{1:\mathcal{T}})$ définis par

$$\begin{aligned} n_{u,v}^{(k)} &= \sum_{t=2\ldots T} \pi_{\theta^{(k)}}(S_t = u, S_{t+1} = v \mid x_{1:\mathcal{T}}) \\ n_{x;u,w}^{(k)} &= \sum_{t=1\ldots T} \mathbb{I}\{x_{t-r:t-1} = w, x_t = x\} \pi_{\theta^{(k)}}(S_t = u \mid x_{1:\mathcal{T}}). \end{aligned} \tag{2.6}$$

Description des niveaux d'hétérogénéités dans un génome bactérien

Les résultats obtenus avec ces HMM non structurés de type M1Mr (ordre 1 sur la chaîne cachée et ordre r sur les observations) sur le génome de *B. subtilis* ont fait l'objet d'une description détaillée dans Nicolas *et al.* (2002) et sont illustrés dans la figure 2.1. J'ai en particulier remarqué une relation entre l'ordre r du modèle d'émission et le caractère plus ou moins diagonal de la matrice de transitions entre états cachées $\hat{\alpha}$. En pratique, sur le génome de *B. subtilis*, un modèle avec $r = 2$ suffit pour obtenir des états bien séparés que jusqu'à $\mathcal{S} = 6$. À partir de 7 états cachés, on trouve des états « couplés » dont les lois d'émission sont bien contrastées mais entre lesquels les transitions sont très fréquentes. Ces états ne sont plus interprétables comme des types de composition pour des segments « homogènes ». En revanche, il est possible d'exhiber 7 types de compositions en ajustant un modèle avec $r = 3$. L'explication de ce comportement me semble être la coexistence de deux niveaux d'hétérogénéité : les hétérogénéités locales que l'on souhaiterait capturer par le modèle markovien d'émission (r et $\beta_{x;u,\bullet}$) et les hétérogénéités à plus longue distance que l'on souhaiterait capturer par les transitions entre états cachés ($\mathcal{S}$, α et $\beta_{x;\bullet,w}$). Cependant, ce comportement n'est pas garanti par l'estimation non-supervisée. Dans certaines circonstances, lorsque $\mathcal{S}$ (donc la dimension de α et $\beta_{x;\bullet,w}$) augmente, un

meilleur ajustement du modèle peut être obtenu en utilisant les transitions entre états cachés pour capturer certaines caractéristiques des hétérogénéités locales. Cette capacité des procédures d'estimation non supervisées a « détourner » les intentions des modèles peut être vue comme un inconvénient. Au contraire, elle me semble intéressante puisqu'elle attire l'attention du modélisateur sur des aspects qui auraient pu passer inaperçus. En cela, elle nourrit un cercle vertueux d'amélioration de la description et de la compréhension des données.

Par ailleurs, en augmentant progressivement $\mathcal{S}$, on observe que chaque introduction d'un état supplémentaire tend à décomposer un type de composition en deux sous-types, ce qui aboutit à décrire une hiérarchie des hétérogénéités. De plus, les niveaux admettent une interprétation biologique, ils distinguent : codant vs. intergénique, « typique » vs. « atypique » (régions riches en `a+t` provenant de transferts génétiques horizontaux), et enfin des propriétés de composition des protéines codées (richesse en `g+t` des codons correspondant aux acides-aminés hydrophobes) ou d'usage du code génétique (lié au niveau d'expression des gènes).

2.1.2 Un HMM fortement structuré pour la détection de gènes

C'est dans la logique d'une amélioration de la description de la séquence que j'ai développé un modèle pour la détection de gènes codant pour des protéines à partir du HMM décrit ci-dessus (figure 2.2 A et B). Ce modèle fait intervenir un beaucoup plus grand nombre d'états cachés et des contraintes sur les paramètres : termes nuls dans α correspondant à des transitions interdites ; termes nuls dans β interdisant certains nucléotides à certaines positions (comme le a après ta en troisième position des codons dans les CDS qui correspondent à un codon stop) ; et enfin, des paramètres appariés dans α ou β, tels que ceux des lois d'émission dans les états correspondant aux codons starts. Le programme a été entièrement écrit pour permettre une définition très souple des modèles et pour tirer parti de la nature creuse de la matrice α afin d'éviter une complexité en $O(\mathcal{S}^2)$.

Le programme de détection de gènes est une sur-couche de ce programme générique. Les fonctionnalités dédiées à cette tâche incluent (i) une construction automatique du modèle (le modèle communément utilisé contient 139 états cachés) ; (ii) une procédure d'estimation par EM en deux étapes, dont la première correspond à l'estimation d'un modèle simplifié qui est ensuite utilisé comme point de départ pour l'estimation du modèle complet ; (iii) un post-traitement des trajectoires d'états cachés reconstruites avec l'algorithme *forward-backward* ($\pi(S_t = u \mid x_{1:\mathcal{T}})$, figure 2.2 C) pour les résumer sous la forme d'une annotation (identification et construction de mesures de confiance pour les codons stops, les codons starts, et les sites de fixation des ribosomes). Le programme et sa documentation sont disponibles à l'adresse `http://genome.jouy.inra.fr/ssb/SHOW/`.

La première utilisation d'un HMM pour la détection de gènes remonte à Krogh *et al.* (1994) et le modèle que j'ai proposé s'inscrit dans la continuité de GeneMark.hmm (Lukashin & Borodovsky, 1998) et GeneMarkS (Besemer *et al.*, 2001) qui offrent d'excellentes performances. La principale différence vient de la maximisation de vraisemblance avec l'algorithme EM tandis que GeneMark.hmm suppose une pré-

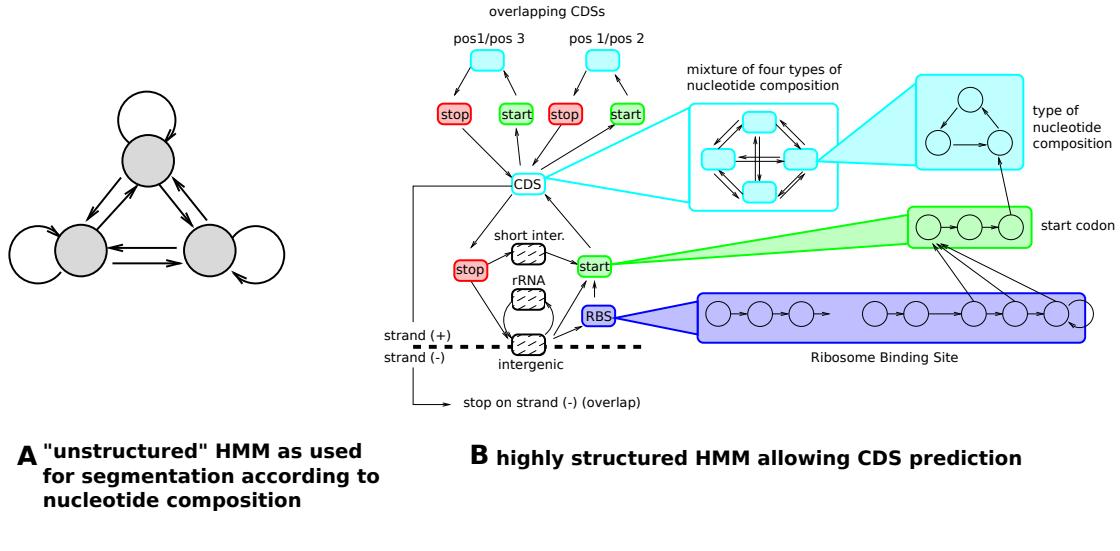


Figure 2.2 – Du HMM non structuré à la prédiction de gènes. Les cercles représentent les états cachés et les flèches représentent les transitions autorisées ($\alpha(u, v) > 0$). [a] HMM non structuré. [B] Structure schématique du HMM développé pour la prédiction de CDS, seul le brin (+) est représenté ici. Les boîtes arrondies correspondent à des grands groupes d'états cachés et quelques détails sur la structure interne de ces groupes sont donnés. [C] Reconstruction du chemin caché sur la même région que dans la figure 2.1.

estimation sur un jeu d'apprentissage et GeneMarkS utilise une procédure itérative fondée sur l'algorithme de Viterbi qui n'aboutit pas à un estimateur consistant des paramètres mais pourrait avoir d'autres avantages (Celeux & Govaert, 1992; Durbin *et al.*, 1998). On notera aussi parmi les différences que ces deux programmes n'incorporent pas directement la recherche du site de fixation du ribosome (RBS) à la détection de gènes. De plus, ils fondent la détection de gènes sur l'algorithme de Viterbi, ce qui ne permet pas d'associer à chaque prédiction une mesure de confiance (comme reconnu dans Azad & Borodovsky (2004)) et peut conduire à des prédictions sous-optimales des CDS dues à la multiplicité des chemins qui aboutissent au même codon stop (et correspondent donc à un même CDS avec éventuellement une incertitude sur la position du codon start). On notera enfin que GeneMark.hmm et GeneMarkS reposent sur des modèles semi-markoviens (et non simplement markoviens) pour prendre en compte la forme non-géométrique de la longueur des CDS.

Mon programme a été largement utilisé pour l'annotation des génomes avec le système

AGMIAL (Bryson *et al.*, 2006) et aussi dans des travaux plus spécifiques de recherches des petits gènes (Ibrahim *et al.*, 2007 ; Fleuchot *et al.*, 2011) et d'interprétation de régions transcrites (Nicolas *et al.*, 2012 ; Mäder *et al.*, 2016).

2.2 Modèles de séquences promotrices

2.2.1 Modèle initial pour la recherche d'un motif bipartite correspondant aux sites de fixation d'un facteur sigma

Quelques généralités sur l'inférence bayésienne et algorithmes MCMC

Mon premier travail sur la recherche de motifs, bien que publié en 2006 (Nicolas *et al.*, 2006a), a été réalisé en grande partie pendant ma thèse. Comme dans les travaux de la section 2.1, il s'agit d'estimer les paramètres d'un modèle à variables latentes de type HMM sur des séquences d'ADN et d'utiliser les paramètres estimés et la reconstitution du chemin caché pour extraire une information biologique. Cependant ces travaux présentent des différences méthodologiques importantes. Tout d'abord en matière de modèle, puisque celui mis en œuvre pour la recherche de motifs est de type semi-markovien caché représentant un ensemble de séquences courtes (les régions promotrices) dans lequel chaque état est visité au plus une fois par séquence (figure 2.3 A et B). Ensuite en matière d'algorithme, puisqu'une perspective bayésienne mettant en œuvre des algorithmes MCMC a été adoptée pour l'estimation. Une des motivations de ce choix était de tirer partie des possibilités offertes par les algorithmes MCMC trans-dimensionnels pour explorer l'espace des modèles (Green, 1995).

L'objectif de cette section 2.2.1 est d'introduire les principaux concepts du cadre bayésien et des algorithmes MCMC utilisés dans nombre de mes travaux de modélisation (Nicolas *et al.*, 2006a ; Nicolas *et al.*, 2007a ; Nicolas *et al.*, 2012 ; Mirauta *et al.*, 2014). Je garde pour cela les notations x, s et θ introduites dans la section 2.1 pour les données de séquences, les états cachés, et les paramètres. Le lecteur pourra se référer par exemple à Gilks *et al.* (1995) ou Andrieu *et al.* (2003) pour beaucoup plus d'information sur ce cadre général maintenant très classique.

Très brièvement, le cadre bayésien suppose de définir une loi *a priori* sur les paramètres du modèle dont on notera la densité $\pi(\theta)$. C'est cette loi qui, combinée à la loi des données sachant les paramètres définie par le modèle, de densité $\pi(x \mid \theta)$, définit la loi de $\theta \mid x$ dite loi *a posteriori* sur laquelle se fonde l'inférence bayésienne. Sa densité $\pi(\theta \mid x)$ est donnée par la formule de Bayes

$$\begin{aligned} \pi(\theta \mid x) &= \frac{\pi(x \mid \theta)\pi(\theta)}{\pi(x)} \\ &\propto \pi(x \mid \theta)\pi(\theta)\,. \end{aligned} \tag{2.7}$$

Le propos des algorithmes MCMC (*Markov chain Monte-Carlo*) utilisés dans ce contexte est d'échantillonner $\pi(\theta \mid x)$ par une marche aléatoire markovienne. Sous les hypothèses d'apériodicité (vérifiée à partir du moment ou chaque pas autorise à rester sur place), d'irréductibilité (vérifiée si n'importe quel point de l'espace peut être atteint en un nombre fini de pas quel que soit le point de départ) et d'invariance (chaque pas de la marche est construit de façon à préserver la loi cible), cette chaîne de Markov admet comme loi stationnaire la loi *a posteriori* de densité $\pi(\theta \mid x)$. L'échantillon $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(K)}$ généré par l'algorithme permet ainsi l'étude de $\pi(\theta \mid x)$, à travers la possibilité d'approcher

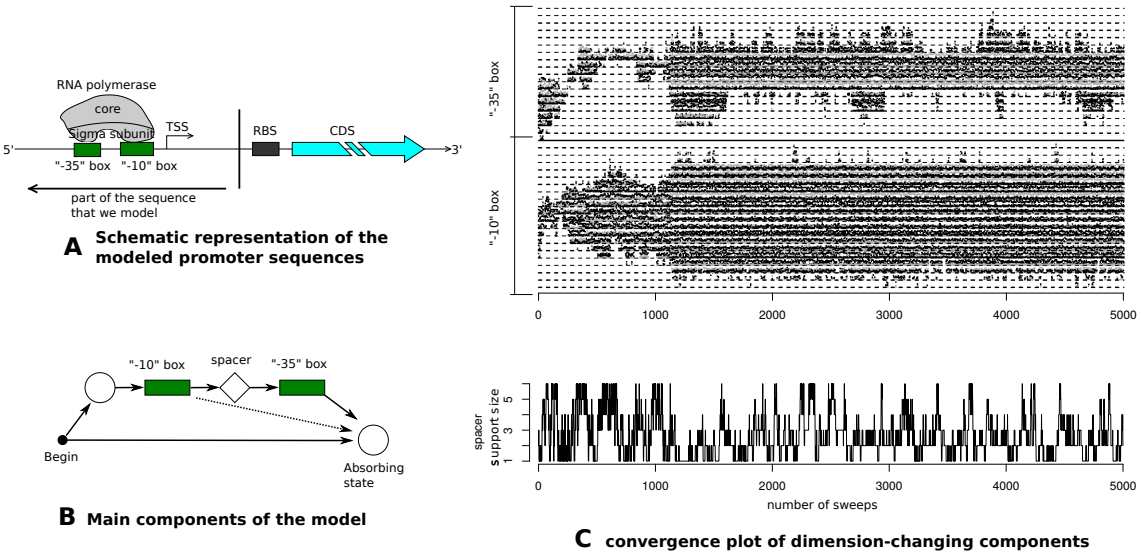


Figure 2.3 – HMM pour la recherche de motifs promoteurs. [A] Représentation schématique de la séquence promotrice avec les boîtes -10 et -35 ainsi que le site de démarrage de la transcription (TSS). [B] Modèle composé de 5 états cachés, les lois d'émission des deux boîtes rectangulaires correspondent à des matrices poids-position tandis que les lois d'émission dans les trois autres états correspondent à un même modèle de *background* Markovien. Les flèches montrent les transitions autorisées, celle en pointillée est optionnelle et permet l'absence de boîte -10. [C] Illustration du comportement de l'algorithme MCMC. On suit ici sur 5 000 balayages de l'algorithme la largeur des boîtes -10 and -35 et la taille du support de la distribution correspondant à la longueur du *spacer* (nombre de tailles différentes autorisées). Le jeu de données analysé est constitué de 24 séquences promotrices identifiées comme sous le contrôle du facteur sigma SigM par des données d'expression chez des mutants. Figure adaptée de Nicolas *et al.* (2006a).

l'espérance $\mathbb{E}_{\pi(\theta|x)}[f(\theta)] = \int_\theta f(\theta)\pi(\theta \mid x)d\theta$ de n'importe quelle fonction $f(\theta)$ par une moyenne empirique. On s'appuie pour cela sur la convergence presque sûre

$$\frac{1}{K}\sum_{k=1}^{K} f(\theta^{(k)}) \quad \underset{K\to+\infty}{\overset{p.s.}{\to}} \quad \mathbb{E}_{\pi(\theta|x)}[f(\theta)]\,.$$

En pratique, on obtient ainsi les estimateurs bayésiens correspondant à l'espérance sous la loi *a posteriori* des différents paramètres (θ ou ses marginales dans la cas courant ou θ est multidimensionnel). On s'appuie aussi sur cette convergence et sur la relation $\mathbb{P}_{\pi(\theta|x)}(\theta \in [a,b]) = \mathbb{E}_{\pi(\theta|x)}[\mathbb{I}_{\theta\in[a,b]}]$ pour construire des intervalles de crédibilité pour les paramètres qui sont les analogues des intervalles de confiance dans le cadre non-bayésien (fréquentiste).

Dans le contexte des modèles abordés dans mes travaux, on introduit dans l'échantillonnage les variables latentes (cachées) qui, en plus de l'intérêt de leur reconstruction pour l'interprétation des données, simplifient grandement et rendent possible les calculs. On sait ainsi aisément calculer $\pi(s, \theta \mid x)$, à une constante près,

grâce à la relation

$$\begin{aligned}\pi(s,\theta \mid x) &\propto \pi(\theta,s)\pi(x \mid s,\theta)\\ &\propto \pi(\theta)\pi(s \mid \theta)\pi(x \mid s,\theta)\,,\end{aligned}$$

dont tous les termes sont connus de façon explicite. Au contraire, le calcul du terme $\pi(x \mid \theta)$ intervenant dans l'équation 2.7 supposerait une intégration sur les valeurs possibles des états cachés puisque $\pi(x \mid \theta) = \int_s \pi(x,s \mid \theta)ds$.

De plus, l'espace des variables à échantillonner (paramètres et variables latentes) étant typiquement de grande dimension, l'algorithme MCMC s'appuie sur un découpage de ces variables en blocs qui sont mis à jours successivement. L'algorithme MCMC combine alors des étapes correspondant à des types de pas (ou mouvements) différents dont chacun a vocation à mettre à jour un bloc de paramètres. Ceux-ci sont répétées de très nombreuses fois selon un enchaînement déterministe ou aléatoire. En adoptant la notation ζ pour l'ensemble des variables à échantillonner (ici $\zeta = (\theta, s)$) et en considérant la décomposition $\zeta = (\zeta_b, \zeta_{-b})$ pour le $b$ième bloc, l'étape correspondante de l'algorithme consistera à mettre à jour ζ_b en préservant la distribution $\pi(\zeta_b \mid \zeta_{-b}, x)$.

Lorsque c'est possible, une étape consistera à tirer $\zeta_b^{(k+1)}$ directement selon $\pi(\zeta_b^{(k+1)} \mid \zeta_{-b}^{(k)}, x)$. On dit alors qu'il s'agit d'une mise à jour « à la Gibbs » (Gibbs step). Sinon, l'algorithme de Metropolis-Hastings fournit une recette très générale pour les mise à jour (MH step). Il s'agit de proposer une nouvelle valeur $\zeta' = (\zeta_b', \zeta_{-b}^{(k)})$ étant donnée la valeur courante $\zeta^{(k)}$ en utilisant une loi instrumentale (*proposal*) de densité $q(\zeta_p'; \zeta_b^{(k)}, \zeta_{-b}^{(k)})$ et d'accepter cette nouvelle valeur ($\zeta^{(k+1)} = \zeta'$) avec probabilité

$$\mathcal{A}(\zeta^k, \zeta') = \min\left\{1; \frac{\pi(\zeta_b' \mid \zeta_{-b}^{(k)}, x) q(\zeta_b'; \zeta_{-b}^{(k)}, \zeta_b^{(k)}))}{\pi(\zeta_b^{(k)} \mid \zeta_{-b}^{(k)}, x) q(\zeta_b^{(k)}; \zeta_b', \zeta_{-b}^{(k)})}\right\},$$

sinon l'ancienne valeur est conservée ($\zeta^{(k+1)} = \zeta^{(k)}$). La probabilité d'acceptation et donc la vitesse de mélange de l'algorithme dépendent dans ce cas complètement de la qualité de la *proposal*. Un cas extrême est l'utilisation de $q(\zeta_b'; \zeta_{-b}^{(k)}, \zeta_b^{(k)}) = \pi(\zeta_b' \mid \zeta_{-b}^{(k)}, x)$ pour lequel la probabilité d'acceptation est de 1 (c'est la mise à jour à la Gibbs décrite au début du paragraphe).

Sans entrer ici dans les détails, Green (1995) montre de façon générale comment étendre cette recette pour réaliser des changements de dimension dans un contexte où l'on a une loi a priori $\pi(m, \theta_m)$ non seulement sur les paramètres mais aussi sur les modèles ; m étant l'indice du modèle dont la dimension des paramètres θ_m peut dépendre. Cette version de l'algorithme de *Metropolis-Hastings* connue sous le nom de *Reversible Jump* permet d'échantillonner la loi *a posteriori* $\pi(m, \theta_m \mid x)$. Bien qu'en théorie ce ne soit pas la seule façon de construire des mouvements préservant une loi cible (voir par exemple Suwa & Todo (2010)), pris individuellement, tous les mouvements que nous venons d'évoquer (MH steps, Gibbs steps, Reversible Jump MH steps) vérifient une condition de réversibilité dite de *detailed balance*. Dans sa forme générale, celle-ci s'exprime

$$\int_{x\in A}\int_{x'\in B} \pi(dx)p(x,dx') = \int_{x\in B}\int_{x'\in A} \pi(dx)p(x,dx'),$$

pour tous A et B, et avec π la fonction de densité de la distribution a préserver (typiquement la loi conditionnelle associée à la mise à jour d'un bloc de variables) et p le noyau de transition associé au mouvement (Green, 1995).

Mes travaux impliquant des MCMC présentés dans ce mémoire s'appuient essentiellement sur le cadre méthodologique présenté ci-dessus et dont les grandes lignes sont parfaitement établies depuis les années 90. Ces algorithmes sont généraux et permettent d'aborder le problème de l'inférence dans un grand nombre de modèles. Il faut cependant noter les difficultés qui proviennent de différents niveaux qui sont interdépendants :

— l'élaboration des modèles ;
— le choix de la paramétrisation et des loi *a priori* ;
— la conception des algorithmes (notamment le découpage des variables en blocs et la construction de *proposal*) ;
— l'implémentation.

Il s'agit en particulier de trouver des modèles et des lois a priori qui aboutissent à des résultats biologiquement intéressants tout en concevant des algorithmes raisonnablement rapides pour l'inférence. Pour optimiser la vitesse d'un algorithme on cherche à diminuer les coût des calculs et l'auto-corrélation de la chaîne de Markov (à travers les probabilités d'acceptation et l'amplitude des pas). Par exemple, un algorithme théoriquement valide avec des taux d'acceptation quasi-nuls s'avérera complètement inutilisable.

Il est aussi extrêmement important d'identifier les erreurs d'implémentation et de conception dans les algorithmes. Il s'agit d'une tâche non triviale de part le grand nombre de lignes de code et le comportement aléatoire des algorithmes dont la convergence n'est qu'en distribution, qui plus est, vers des distributions inconnues. La principale approche que j'utilise pour cela repose sur ce que Geweke (2004) appelle un *successive-conditional simulator* et qui consiste à introduire dans l'algorithme (pour validation) une étape de mise à jour de x selon $\pi(x \mid s, \theta)$ afin d'échantillonner la loi jointe $\pi(x, s, \theta)$ dont la marginale $\pi(\theta)$ est connue puisqu'elle correspond à la loi *a priori*. On travaille dans ce contexte avec un x de dimension très restreinte pour pouvoir évaluer en un temps raisonnable la convergence vers la distribution attendue. En pratique il me semble que même les petites erreurs tendent à se répercuter de façon visible dans la distribution des θ échantillonnés sans nécessiter l'étude formelle comme dans Geweke (2004) de la convergence d'estimateurs de l'espérance d'une fonction test.

Un exemple de modèle semi-markovien caché

La principale caractéristique qui définit un promoteur bactérien est la présence d'un site de fixation pour la sous-unité sigma (ou facteur sigma) de l'ARN polymérase (Gruber & Gross, 2003; Paget, 2015). Pour la majorité des sous-unités sigma (une exception étant la famille minoritaire σ^{54}) ce site, lorsqu'il est suffisamment fort, permet à lui seul le recrutement de l'ARN polymérase au niveau du promoteur et le démarrage de la transcription une dizaine de paires de bases en aval (en 3'). Le site de reconnaissance du facteur sigma est typiquement constitué de deux boîtes séparées dites « -10 » et « -35 » en référence à leur position relative au site de démarrage de la transcription (TSS, pour

Transcription Start Site). Ces boîtes sont longues de 3 à 15 paires de bases et séparées par une distance qui peut être comprise entre 10 et 20 paires de bases (les distances autorisées pour un facteur sigma donné étant contiguës et au nombre de 2 ou 3). Cette organisation du promoteur est illustrée dans la figure 2.3 A. On appelle motif la description des sites possibles et le problème d'inférence abordé ici consiste à découvrir ce motif et à trouver les sites dans un ensemble de séquences bien choisies.

Le modèle utilisé dans Nicolas *et al.* (2006a) décrit les séquences intergéniques situées directement en amont du codon start des CDS. Chaque séquence, d'une longueur comprise entre 80 et 150 paires de bases, est modélisée de 3' vers 5' à travers cinq états cachées : un état « pré-motif », la boîte -10, le *spacer*, la boîte -35, et un état absorbant pour la fin de la séquence. Les transitions autorisées entre ces cinq états sont représentées dans la figure 2.3 B. Les lois d'émission dans les états pré-motif, *spacer*, et -35 correspondent à un même modèle de Markov homogène d'ordre r (dit *background*) alors que les boîtes -10 et -35 dont les longueurs sont fixes correspondent à des modèles hétérogènes caractérisés par une fréquence d'émission des nucléotides particulière à chaque position dans la boîte. Ces modèles hétérogènes sont communément appelés PWM (*Position Weight Matrix*).

Tandis que le modèle markovien suppose une durée géométrique pour le séjour dans chaque état caché (éventuellement dégénérée à 1 ou $+\infty$), le modèle à cinq états décrit ci-dessus est semi-markovien de part les longueurs fixes des boîtes -10 et -35, la longueur contrainte du *spacer*, et la longueur de l'état « pré-motif » modélisée par une loi binomiale négative (éventuellement décalée pour tenir compte de la distance minimale entre le TSS et la boîte -10). Notons que, comme illustré dans la figure 2.2 pour les sites de fixation des ribosomes, il est par exemple possible de décrire un modèle hétérogène tel que celui utilisé pour les boîtes -10 et -35 dans un cadre markovien (et non semi-markovien) par des enchaînements d'états cachés. Cette idée a été mise en œuvre dans le contexte de la recherche des facteurs sigma (Jarmer *et al.*, 2001). La définition et la paramétrisation du modèle sont cependant alors moins souples et tendent à impliquer un très grand nombre d'états cachées.

Algorithme MCMC trans-dimensionnel

L'algorithme MCMC permet l'estimation de différents paramètres gérant la dimension du modèle, à savoir l'ordre r du modèle markovien d'émission du *background* (de 0 à 5), la taille de chacune des deux boîtes (de 1 à 25), et le nombre de longueurs autorisées pour la distance entre les boîtes (de 1 à 6), ce qui correspond à l'exploration d'un espace contenant $22\,500$ ($6 \times 25 \times 25 \times 6$) « modèles » différents. La figure 2.3 C montre sur un exemple de trajectoire l'évolution de la taille des boîtes et du nombre de distances différentes autorisées sur un jeu de données.

Nicolas *et al.* (2006a) donne les détails de cet algorithme qui s'inscrit dans la lignée des programmes de recherche de motifs fondés sur le modèle PWM couplés à une estimation bayésienne par algorithme MCMC dont le Gibbs Motif Sampler et AlignACE furent les premiers représentants (Neuwald *et al.*, 1995; Liu *et al.*, 1995; Roth *et al.*, 1998). Le programme Bioprospector (Liu *et al.*, 2001) en implémente une version pour la recherche de motifs bipartites mais qui ne prend pas en compte la position dans la séquence ni ne

réalise un ajustement automatique la taille des boîtes. Trois jeux de données relatifs à des facteurs sigma de *B. subtilis* sont utilisés pour comparer les résultats obtenus à ceux de BioProspector en utilisant une connaissance *a priori* de la longueur des boîtes et des distances autorisées entre elles.

De façon un peu antérieure à ma publication, d'autres travaux ont abordé la question de l'ajustement de la longueur des boîtes. Ainsi, Jensen & Liu (2004) ont proposé une optimisation déterministe de la taille des boîtes en se fondant sur un critère de maximum *a posteriori* appliqué à la position des motifs avec une marginalisation (intégration) vis à vis des paramètres d'émission du PWM. On trouve une idée similaire mise en œuvre par Qin *et al.* (2003) dans un contexte d'algorithme MCMC pour la détection de motifs par *clustering* de régions intergéniques conservées. L'étape de mise à jour du nombre de positions prises en compte dans le PWM s'affranchit ainsi d'un ré-échantillonnage simultané des *clusters*. Nicolas *et al.* (2006a) évoque la difficulté d'utiliser proprement cette stratégie, c'est-à-dire en préservant la réversibilité des mouvements de l'algorithme MCMC à cause des sites positionnés sur les bords des séquences.

Notons aussi que, dans le cas général, tel que décrit dans Nicolas *et al.* (2006a), les algorithmes de programmation dynamique pour la reconstruction du chemin caché (ici le tirage dans $\pi(s \mid \theta, x)$) sont beaucoup plus coûteux que pour un modèle markovien avec le même nombre d'états. La complexité est en $O(\mathcal{T}^2 \times \mathcal{S}^2)$ au lieu de $O(\mathcal{T} \times \mathcal{S}^2)$. Cependant les contraintes sur les chemins cachés permettent une évaluation rapide et directe (sans programmation dynamique) de tous les chemins possibles dont l'algorithme implémenté dans Nicolas *et al.* (2006a) ne tirait pas partie.

2.2.2 Partitionnement du répertoire des promoteurs en prenant en compte la position du TSS et les profils d'expression

Contexte et grandes lignes de l'approche

Plusieurs années après mon premier travail sur la recherche de motifs correspondants aux sites de facteurs sigma (section 2.2.1), des résultats obtenus en transcriptomique *tiling arrays* chez *B. subtilis* ont ressuscité mon intérêt pour cette question. Les données expérimentales permettaient d'identifier assez précisément les positions des TSS (Nicolas *et al.*, 2009) et les résultats préliminaires révélaient que les profils d'expression associés aux TSS à travers les conditions expérimentales étaient très cohérents avec les descriptions disponibles des régulons des facteurs sigma (Sierro *et al.*, 2008). J'ai donc développé un nouveau modèle et l'algorithme MCMC associé en prenant en compte les points que je pensais pouvoir être améliorés dans mon travail initial et les caractéristiques des nouvelles données. En pratique, par rapport à Nicolas *et al.* (2006a), l'approche permet de :

- mettre à jour la largeur des boîtes sans ré-échantillonnage simultané du chemin caché,
- mettre à jour le chemin caché en s'affranchissant du cadre général de l'algorithme *forward-backward* pour les modèles semi-markoviens,
- rechercher plusieurs motifs simultanément ;

tout en prenant en compte l'information sur :

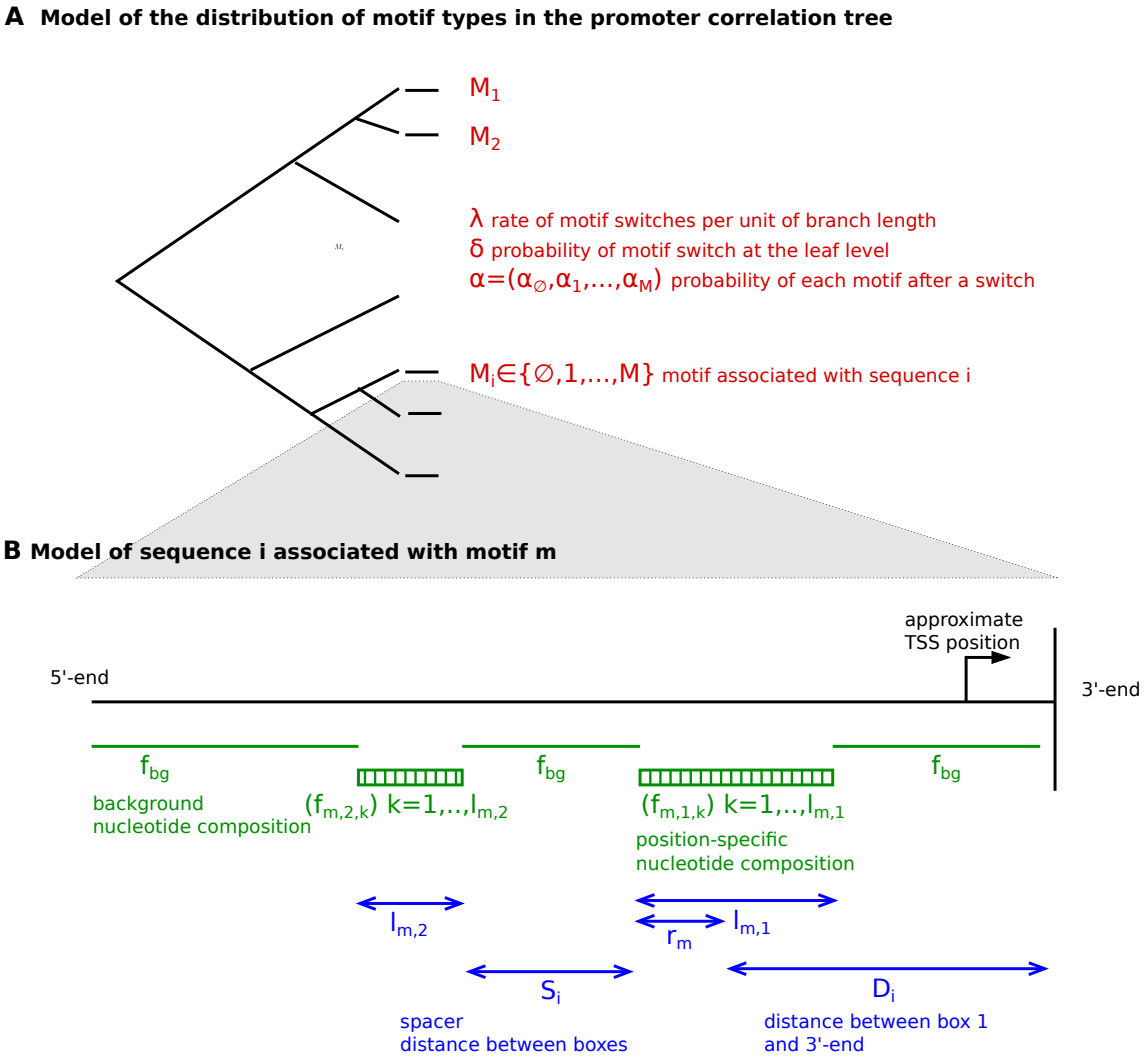


Figure 2.4 – Principales composantes du modèle pour la classification des séquences promotrices selon les motifs promoteurs. [A] Modélisation de la distribution des motifs prenant en compte les relations de co-expression résumées à travers un arbre de classification hiérarchique des promoteurs. [B] Modélisation de la séquence. Trois groupes de paramètres et variables latentes sont distingués ici, à savoir ceux qui font le lien entre données d'expression et occurrence des motifs (en rouge), ceux qui servent à décrire la position exacte et l'architecture du motif dans la séquence (en bleu), et ceux qui décrivent la composition de la séquence conditionnellement à la position du motif (en vert). Figure adaptée de Nicolas *et al.* (2012).

— la position au moins approximative des TSS,
— les profils d'expression.

Les principales variables latentes et certains paramètres du modèles sont représentés dans la figure 2.4, les détails du modèle et de l'algorithme sont décrit dans le *Supplementary Information* de Nicolas *et al.* (2012). Le comportement de l'algorithme et les résultats sont illustrés dans la figure 2.5. Sans entrer dans tous les détails, j'évoque ci-dessous quelques points clés de l'approche.

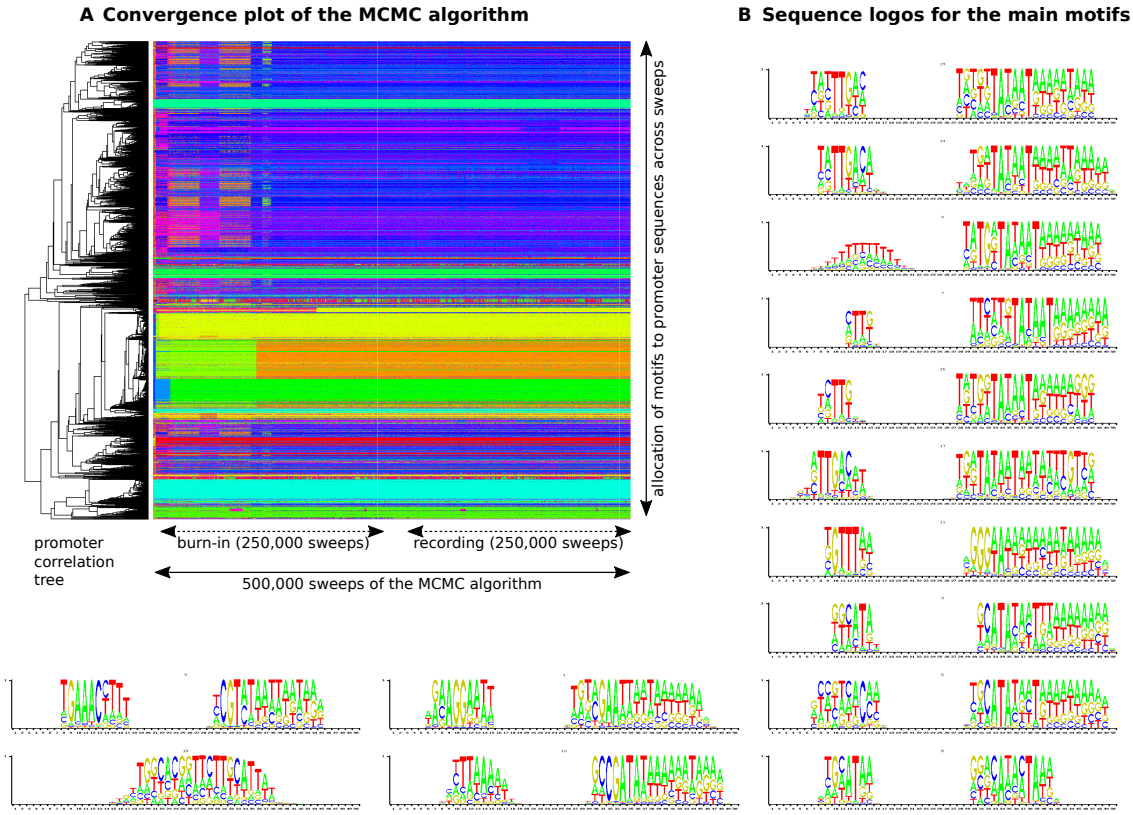


Figure 2.5 – Classification des séquences promotrices selon les motifs de fixation de facteurs sigma en prenant en compte la corrélation entre profils d'activité. 3 242 séquences dont la plupart contiennent un site de fixation de facteur sigma sont analysées simultanément pour 20 motifs. Figure adaptée de Nicolas *et al.* (2012).

Quelques précisions sur l'algorithme

Des mouvements réversibles de mise à jour de la largeur des boîtes sans ré-échantillonnage du chemin caché sont rendus possibles par l'introduction d'une variable latente « instrumentale » (r_m dans la figure 2.4) qui indique quelle position dans la boîte -10 est utilisée comme référence lorsqu'il s'agit de décrire la position du motif dans la séquence. La mise à jour de la largeur des boîtes par extension à droite et à gauche devient alors possible sans modifier la variable correspondant à la position du motif et donc sans risque de faire disparaître l'occurrence en la faisant sortir de la séquence (condition nécessaire pour la réversibilité). La variable r_m, elle même, est mise à jour dans le cours de l'algorithme afin de contourner les effets de bords, à savoir l'impossibilité d'un raccourcissement à gauche (respectivement à droite) lorsque la position de référence se trouve en première (resp. dernière) position du motif. On note que ces effets de bords sont beaucoup moins contraignants que ceux rencontrés sans faire intervenir une telle variable et qui dépendent alors de la position des motifs dans les séquences. Par exemple, si la première position du motif est utilisée pour indiquer sa position dans la séquence, une occurrence débutant en première position de la séquence interdira une extension à gauche (et une occurrence en dernière position de la séquence interdira un raccourcissement à gauche).

Or, ces configurations sont d'autant plus probables que les séquences sont nombreuses et les occurrences fréquentes.

Tout en restant dans le cadre rigoureux d'un MCMC dont les étapes utilisées pour l'échantillonnage de la loi stationnaire sont réversibles (elles vérifient la *detailed balance*), on s'autorise des étapes non-réversibles pour une meilleure exploration de l'espace pendant une première phase de l'échantillonnage (*burn-in*). On fait intervenir ces étapes non-réversibles lorsque certaines composantes (motifs) se « vident » au sens où elles n'ont plus d'occurrences. On supprime alors la composante en question et on la remplace par un clone de l'une des autres composantes (ce qui engendre les brusques discontinuités de couleur visibles dans la représentation adoptée figure 2.5 A). Cela permet d'aboutir à une description des variantes du motif reconnu par le facteur SigA faisant intervenir plusieurs motifs proches comme on le voit dans la figure 2.5 B.

Quelques précisions sur le modèle

Pour la position du motif vis-à-vis du TSS, le nouveau modèle s'affranchit de la représentation fortement paramétrique impliquant une loi binomiale négative utilisée dans Nicolas *et al.* (2006a). Il utilise, à la place, un modèle de distribution constante par morceaux dont le nombre et la position des morceaux sont estimés grâce à des étapes impliquant des sauts de dimension. Cette représentation est très flexible notamment dans un contexte comme celui ci où la distribution de la position de la boîte -10 est commune à tous les motifs (la position r_m mentionnée précédemment permet leur alignement) et peut donc être estimée finement grâce à des milliers d'occurrences.

En matière de modèle, les différences les plus importantes par rapport à celui initialement proposé dans Nicolas *et al.* (2006a) est la prise en compte de plusieurs motifs et des données d'expression. Notons que la prise en compte de plusieurs motifs simultanément et relativement triviale si on introduit une variable latente M_i indicatrice du type de motif dans chaque promoteur (cf. figure 2.4) et que l'on modélise ces variables associées aux différents promoteurs comme indépendantes et identiquement distribuées. Cette modélisation, qui correspond à un simple modèle de mélange, offre déjà l'avantage de permettre une recherche simultanée de tous les motifs en tenant compte de la non indépendance dans leurs occurrences puisque, conformément au principe selon lequel chaque promoteur correspond au site de fixation d'un des facteurs sigma, une seule occurrence de motif est autorisée par séquence.

Le modèle introduit dans Nicolas *et al.* (2012) propose une extension du modèle de mélange simple pour rendre compte de l'idée selon laquelle, plus deux promoteurs ont des profils d'activation similaires, plus on s'attend à y trouver des occurrences du même motif. En pratique, la proximité entre les profils d'activation des promoteurs est résumée par les distances sur un arbre binaire obtenu par classification hiérarchique sur les distances « de Pearson » (définie comme $1 - r$, où r est le coefficient de corrélation de Pearson). Quant au type de motif, il est modélisé comme le résultat d'un processus le long des branches dans lequel des sauts ont lieu à un taux homogène λ et, après chaque saut, le motif $m \in \{1, \ldots, \mathcal{M}\}$ est tiré selon les probabilités $(\alpha_0, \alpha_1, \ldots, \alpha_{\mathcal{M}})$, avec α_0 correspondant à une absence de motif. Enfin, un saut est aussi possible au niveau de chaque feuille

de l'arbre (probabilité δ) pour tenir compte d'éventuels *outliers*. Cette modélisation est séduisante car elle n'ajoute que deux paramètres, λ et δ, au modèle de mélange simple que l'on retrouve comme limite lorsque $\delta \to 1$ ou $\lambda \to +\infty$.

Un autre intérêt de ce modèle est que, tout en prenant en compte les dépendances, il permet une mise à jour très efficace (car à la Gibbs et simultanée) des variables $(M_1, \ldots, M_I)$, où I est le nombre de promoteurs. Cette étape de l'algorithme MCMC repose sur un algorithme de programmation dynamique dans l'arbre directement inspiré des algorithmes utilisés en phylogénie, notamment pour la reconstruction des caractères ancestraux (Felsenstein, 2003).

N'ayant pas inclus de précisions sur cet algorithme de mise à jour des M_i dans la publication, je le décrit ici. On ordonne par hauteur les $I-1$ nœuds internes dans l'arbre. Pour le nœud i, on note : h_i sa hauteur, $l(i)$ et $r(i)$ ses descendants gauche et droit, $p(i)$ son parent et $s(i)$ le vecteur contenant les index des promoteurs (feuilles de l'arbre) sous le nœud i. On introduit, en plus des variables $(M_i)_{i=1:I}$ indiquant le type de motif associé à chaque séquence i, des variables latentes $(\tilde{M}_i)_{i=1:(2I-1)}$ qui enregistrent, pour chaque nœud i de l'arbre (I feuilles et $I-1$ nœuds internes), le type de motif généré par le processus de saut dans l'arbre. L'algorithme est alors composé de deux récursions successives, la première de bas en haut, et la seconde de haut en bas, parcourant l'ensemble des nœuds (nœuds internes et feuilles correspondants aux promoteurs).

Lors de la récursion montante on réalise les calculs des $\pi(x_{s(i)} \mid \tilde{m}_i)$ pour $\tilde{m}_i \in \{1, \ldots, \mathcal{M}\}$ et i de 1 à $2I-1$. On commence par les I nœuds correspondants aux feuilles (pour lesquels $s(i) = i$) grâce à la relation

$$\pi(x_i \mid \tilde{m}_i) \quad = \quad (1-\delta)\pi(x_i \mid M_i = \tilde{m}_i) + \delta \sum_m \alpha_m \pi(x_i \mid M_i = m)\,,$$

où $\pi(x_i \mid M_i = m) = \sum_{d,s} \pi(x_i, D_i = d, S_i = s \mid M_i = m)$ a été calculé par une somme sur toutes les valeurs possibles pour (D_i, S_i) qui représentent les positions du motif. On poursuit par les $I-1$ nœuds internes avec

$$\begin{aligned}&\pi(x_{s(i)} \mid \tilde{m}_i)\\ &= \prod_{j \in \{l(i), r(i)\}} \left\{ e^{-(h_i - h_j)\lambda} \pi(x_{s(j)} \mid \tilde{M}_j = \tilde{m}_i) + (1 - e^{-(h_i - h_j)\lambda}) \sum_m \alpha_m \pi(x_{s(j)} \mid \tilde{M}_j = m) \right\},\end{aligned}$$

où le fait de parcourir l'arbre de bas en haut assure que les $\pi(x_{s(j)} \mid \tilde{M}_j = m)$ associés aux descendants $l(i)$ et $r(i)$ du nœud i ont déjà été calculés.

Arrivé à la racine $r = 2I-1$, on amorce la récursion descendante consistant à tirer dans la loi jointe des $(\tilde{M}_i)_{i=1:2I-1}$ et $(M_i)_{i=1:I}$ conditionnellement à l'ensemble des séquences x ($x = x_{s(r)}$). Les propriétés d'indépendance conditionnelle (cf. graphe moral) du modèle font qu'il suffit pour cela de tirer $\tilde{M}_i$ conditionnellement $\tilde{M}_{p(i)}$ et $x_{s(i)}$ car les M_j et $\tilde{M}_j$ en dehors du sous-arbre de racine i sont indépendants de $\tilde{M}_i$ conditionnellement à $\tilde{M}_{p(i)}$. En pratique, $\tilde{m}_r$ est tiré selon la probabilité $\pi(\tilde{m}_r \mid x) \propto \alpha_m \pi(x_{s(r)} \mid \tilde{m}_r)$. Puis, pour les autres nœuds $\tilde{m}_i$ de $2I-2$ à 1, le tirage se fait selon

$$\pi(\tilde{m}_i \mid x, \tilde{m}_{p(i)}) \quad \propto \quad \pi(x_{s(i)} \mid \tilde{m}_i) \times \left[e^{-(h_{p(i)} - h_i)\lambda} \mathbb{I}\{\tilde{m}_{p(i)} = \tilde{m}_i\} + \alpha_m (1 - e^{-(h_{p(i)} - h_i)\lambda}) \right].$$

On finit la récursion descendante par le tirage de m_i selon $\pi(m_i \mid x, \tilde{m}_i) \propto \pi(x_i \mid m_i) \times \big[(1-\delta)\mathbb{I}\{\tilde{m}_i = m\} + \alpha_m \delta\big]$. Les valeurs des variables aléatoires D_i et S_i qui indiquent la position exacte du motif M_i dans le promoteur i (cf figure 2.4) sont ensuite tirées conditionnellement à m_i.

Je reviendrai dans le chapitre 3 sur les résultats obtenus avec cette approche. On peut noter que la façon dont ce modèle prend en compte les données d'expression pour aider la recherche de motif est originale. En effet, les principaux algorithmes qui ont été proposés pour prendre en compte de façon automatique des données d'expression (ou de ChIP) adoptent une perspective « discriminante ». Il s'agit alors par exemple de rechercher des motifs dont les fréquences diffèrent entre deux ensembles de séquences définis sur la base des données d'expression, ou dont les occurrences permettent de prédire les données d'expression (Foat *et al.*, 2006; Zambelli *et al.*, 2013). Au contraire, le modèle introduit ici vise à prendre en compte les données d'expression de façon à améliorer la représentation statistique (via la *likelihood*) des séquences. Ce point de vue envisage la recherche de motifs fondée sur les seules séquences et la recherche de motifs appuyée par des données externes comme les deux extrémités d'un continuum permis par le modèle.

2.3 Modèles de polymorphisme

2.3.1 Utilisation d'un modèle de déséquilibre de liaison pour le choix de *tag SNPs*

Un cadre de théorie de l'information pour le choix de marqueurs

Ma première rencontre avec des modèles visant à rendre compte du polymorphisme fut à l'occasion d'un travail sur la sélection de *tag SNPs* qui sont des marqueurs que l'on cherche à sélectionner pour optimiser l'effort de génotypage dans des études d'associations. Il s'agissait d'un sujet sur lequel des chercheurs de mon laboratoire d'accueil en post-doctorat avaient déjà travaillé sous un angle algorithmique, sans s'appuyer sur un modèle probabiliste explicite du polymorphisme (Zhang *et al.*, 2002, 2005). L'idée pour nous était de reformuler le problème sous un angle de théorie de l'information (Cover & Thomas, 1991).

En notant $X = (X_1, \ldots, X_{\mathcal{T}})$ le vecteur aléatoire modélisant la séquence d'un haplotype de longueur $\mathcal{T}$ (ici le nombre de positions polymorphes plutôt que la longueur en pb de la séquence), on cherche à positionner m marqueurs de façon à capturer autant d'information que possible sur X. En notant $u_{1:m}$ les indexes des marqueurs on peut alors voir le problème comme celui du choix de $u_{1:m}$ maximisant l'entropie de Shannon $H(X_{u_{1:m}})$ du vecteur aléatoire $X_{u_{1:m}}$ qui correspond aussi à l'information mutuelle entre $X_{u_{1:m}}$ et X.

Cette approche présente la particularité (avantage et/ou inconvénient) d'offrir un critère de choix de marqueurs qui ne dépend pas des détails de la façon dont on envisage d'utiliser les marqueurs (notamment ici des caractéristiques des tests d'association que l'on utilisera). En contrepartie, l'approche repose entièrement sur une modélisation probabiliste de la séquence qui devra donc être aussi bonne que possible. Le modèle a ici pour but d'agréger toute l'information disponible sur la fréquence des allèles et la structure du déséquilibre de liaison. De plus, à l'exception de modèles irréalistes (indépendant ou chaîne de Markov simple), ou d'une séquence trivialement courte, il semble impossible de trouver le $u_{1:m}$ optimal.

Modèle HMM et algorithme glouton

Dans Nicolas *et al.* (2006b), nous avons testé différents modèles de Markov cachés et conçu un algorithme itératif incorporant les marqueurs de façon gloutonne se fondant sur la décomposition $H(X_{u_{1:m+1}}) = H(X_{u_{1:m}}) + H(X_{u_{m+1}} \mid X_{u_{1:m}})$. Il s'agit donc de choisir u_{m+1} de façon à maximiser $H(X_{u_{m+1}} \mid X_{u_{1:m}})$. Pour cela, une approximation fondée sur un échantillon simulé de K réalisations indépendantes du vecteur X a été utilisée

$$-\frac{1}{K}\sum_{k=1}^{K}\sum_{y\in\{0,1\}} \pi(x_t^k = y \mid x_{u_{1:m}}^k)\log\pi(x_t^k = y \mid x_{u_{1:m}}^k) \underset{K\to+\infty}{\overset{p.s.}{\to}} H(X_t \mid X_{u_{1:m}})\,,$$

où $y \in \{0, 1\}$ correspond aux deux allèles possibles dans un contexte de polymorphisme binaire. Une implémentation de l'algorithme *forward-backward* a été proposée pour mettre à jour de façon approchée $\pi(x_t^k = y \mid x_{u_{1:m}}^k)$ lorsque m croît. Les paramètres des différents

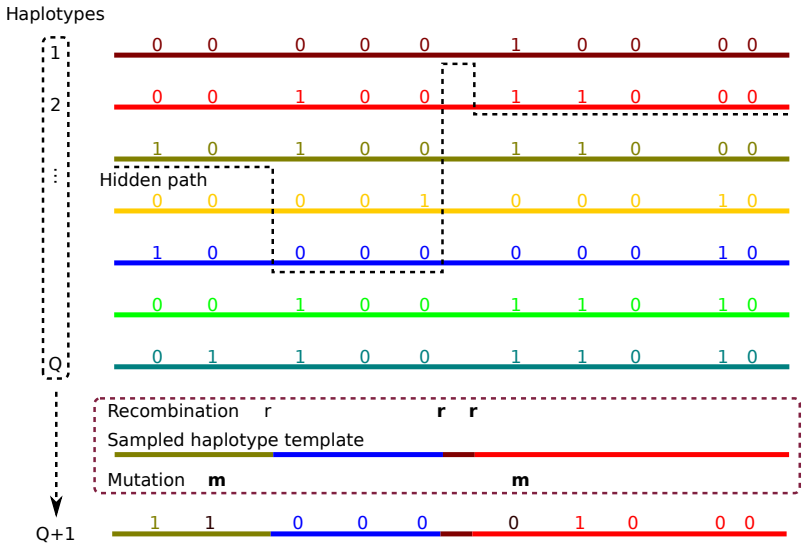


Figure 2.6 – HMM introduit par Li & Stephens (2003) utilisé pour la sélection de *tag SNPs* dans Nicolas *et al.* (2006b). Ici un nouvel haplotype (le $\mathcal{Q}+1^{\text{ième}}$) est modélisé comme une mosaïque des q haplotypes déjà observés. La mosaïque est générée par un chaîne de Markov cachée à valeurs dans $\{1,\ldots,\mathcal{Q}\}$; les sauts entre haplotypes sont modélisés par un processus de Poisson dont l'intensité reflète le taux de recombinaison dans la population. La mosaïque est imparfaite car des différences entre l'haplotype servant localement de modèle et le nouvel haplotype sont introduites à un taux qui reflète celui des mutations.

modèles étaient estimés avant la construction de $u_{1:m}$ dans un cadre de maximum de vraisemblance avec l'algorithme EM sur un jeu d'haplotypes servant pour apprendre la structure du déséquilibre de liaison.

Parmi les modèles considérés, le plus esthétique et le plus satisfaisant pour son adéquation aux données fut sans conteste celui proposé par Li & Stephens (2003). Son principe est illustré dans la figure 2.6. À l'aide de seulement deux paramètres liés au taux de mutation et au taux de recombinaison populationnels du modèle coalescent standard, il permet de modéliser un nouvel haplotype conditionnellement à un jeu d'haplotypes connus. Une version avec taux de recombinaison hétérogène le long de la séquence a aussi été employée. Pour ce modèle, nous avons conduit l'estimation par une approche de maximisation par l'algorithme EM d'une quantité dite vraisemblance composite dont on peut trouver dans Varin *et al.* (2011) une revue des nombreuses déclinaisons et applications. La vraisemblance composite est construite ici comme le produit des lois conditionnelles de chaque haplotype étant donnés tous les autres haplotypes. Afin de stabiliser l'estimation des nombreux paramètres du modèle avec taux de recombinaison hétérogène, un critère de maximum *a posteriori* est aussi considéré dans ce même cadre de vraisemblance composite optimisée avec l'algorithme EM.

De façon intéressante, le modèle de Li & Stephens (2003) contient un nombre d'états cachés égal au nombre d'haplotypes dans le jeux d'apprentissage. Les calculs restent abordables car les probabilités de transition de la chaîne cachée qui correspondent aux évènements de recombinaisons « oublient » l'état de départ (c'est le pendant de l'hypothèse

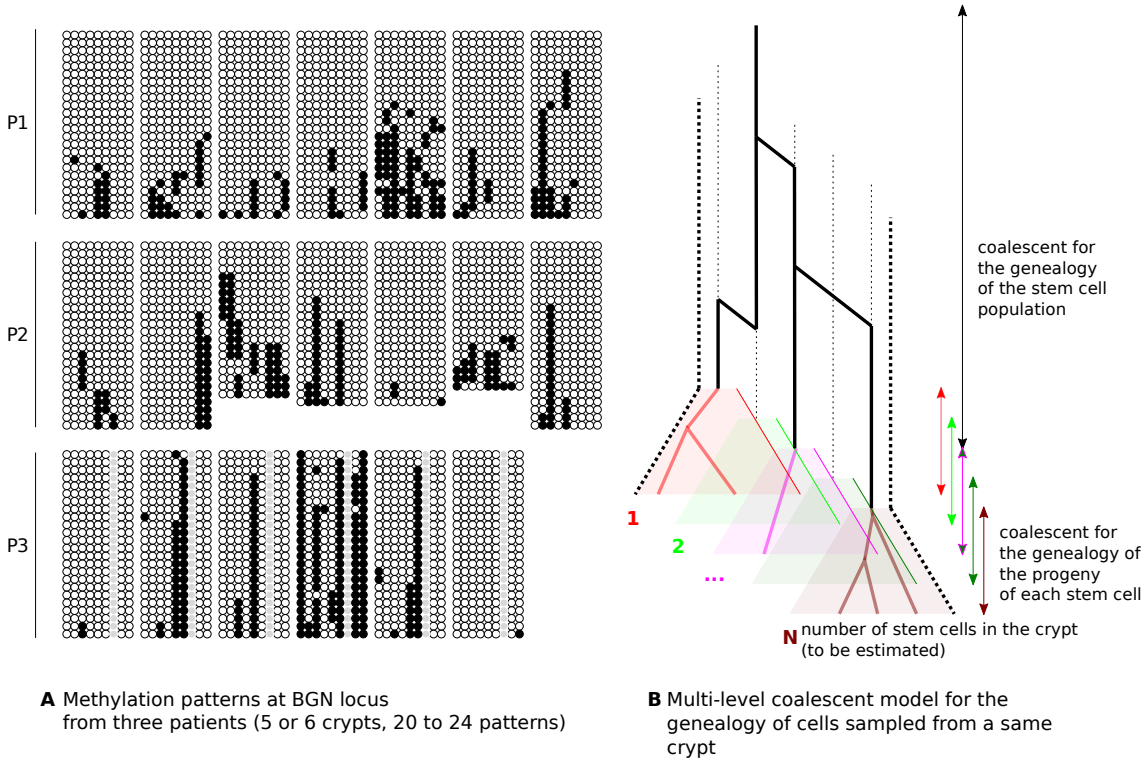


Figure 2.7 – Modèle coalescent pour la généalogie des cellules dans les cryptes du colon. [A] Une partie des profils de méthylation analysés. Les données représentées ici proviennent de trois patients (P1-3), chacun représenté par 5 ou 6 cryptes. Les données d'une même crypte forment un bloc : 20-24 profils provenant de différentes cellules (lignes) de 9 sites (colonnes). Les cercles pleins correspondent aux sites méthylés. [B] Illustration du modèle coalescent structuré proposé pour la généalogie des cellules dans Nicolas *et al.* (2007a). On y voit notamment la subdivision de la population des cellules en N sous-populations correspondant aux différentes cellules souches. Le modèle distingue deux époques dans la généalogie des cellules. La première (en remontant le temps) rend compte des générations de différentiation de la descendance de chaque cellule souche et fait intervenir N coalescents parallèles avec croissance de population. La seconde époque rend compte des générations ayant eu lieu dans les lignées de cellules souches, elle fait intervenir un coalescent standard correspondant à un modèle de Moran pour la population de cellules souches.

de population homogène dans le modèle coalescent). En pratique, les algorithmes de programmation dynamique de type *forward-backward* peuvent alors être réalisés en temps $O(\mathcal{T} \times \mathcal{S})$ au lieu de $O(\mathcal{T} \times \mathcal{S}^2)$, où $\mathcal{S}$ est le nombre d'états cachés (ici $\mathcal{S} = \mathcal{Q}$, cf. figure 2.6). J'ai plus tard réutilisé dans des contextes différents ces idées (i) de simplification dans les algorithmes de programmation dynamique par des hypothèses sur le transitions et (ii) de prise en compte de dépendances complexes par l'intermédiaire d'un petit nombre de paramètres (Nicolas *et al.*, 2009 ; Nicolas *et al.*, 2012), tel qu'expliqué dans les sections 2.2.2 et 2.4.1.

2.3.2 Modélisation du polymorphisme épigénétique dans les cryptes du colon

Un modèle coalescent pour la généalogie des cellules

À la fin de mon post-doctorat, j'ai débuté un travail sur l'étude des populations de cellules souches par des approches de génétique des populations. La plupart des tissus sont renouvelés de façon continue a l'âge adulte. Les nouvelles cellules différenciées proviennent de populations de cellules souches spécifiques de chaque tissu. Ces populations sont difficiles a étudier, notamment car il n'existe souvent pas de marqueur phénotypique qui permette de distinguer les cellules souches de leur descendance immédiate (la différenciation s'étalant sur plusieurs générations de cellules). Dans ce contexte, l'idée du travail était de déduire certaines propriétés des populations de cellules souches de l'épithélium intestinal humain de l'analyse des profils de méthylation de cellules échantillonnées dans les cryptes du colon. En effet, le statut méthylé ou non-méthylé d'une cytosine (sites CpG) est un marqueur épigénétique qui doit permettre d'étudier les populations de cellules (au premier rang desquelles les cellules souches), un peu à la manière dont sont traditionnellement utilisées les séquences d'ADN en génétique des populations (Yatabe *et al.*, 2001; Shibata & Tavaré, 2007).

La figure 2.7 présente les données de méthylation considérées et le modèle de type coalescent (Kingman, 1982) proposé pour la généalogie des profils de méthylation échantillonnés au sein d'une crypte. Le modèle de coalescent est ici structuré (Nordborg, 2001) pour prendre en compte les sous-populations de cellules correspondant aux descendantes des différentes cellules souches, et la juxtaposition de deux « époques » dans la généalogie des cellules. On distingue ainsi les dernières générations ayant eu lieu au cours de la différenciation des générations plus anciennes ayant eu lieu dans les lignées de cellules souches.

Les paramètres du modèles dont dépendent les propriétés des généalogies sont le nombre N de cellules souches correspondant au nombre de sous-populations de cellules dans la crypte ; un paramètre reflétant la forme plus ou moins en étoile des généalogies de cellules différenciées dans ces sous-populations du fait de leur expansion à partir d'une cellule souche ; le taux de coalescence dans les lignées de cellules souches reflétant la capacité des lignées de cellules souches à se remplacer occasionnellement les unes par les autres.

Inférence bayésienne par algorithme MCMC et analyse de l'adéquation du modèle

L'inférence a été conduite dans un cadre bayésien grâce à un algorithme MCMC impliquant des variables latentes correspondent à la configuration de l'échantillonnage dans les sous-populations correspondant à la descendance des différentes cellules souches, à la topologie et aux longueurs de branche des arbres, ainsi qu'aux séquences aux nœuds des arbres. L'échantillonnage dans cet espace de très grande dimension est rendu possible grâce à des propositions de modification de topologie astucieuses (par « *branch-swapping* ») qui s'appuient sur la comparaison des séquences aux nœuds internes de l'arbre (Wilson &

Balding, 1998). Notons que du fait du grand nombre de cryptes, la mise à jour de N conditionnellement à la configuration de l'échantillonnage dans les sous-populations, est réaliste dans une crypte mais devient ici problématique du fait du nombre de cryptes prises en compte simultanément (57 cryptes). L'astuce pour contourner cette difficulté a consisté à réaliser l'échantillonnage dans un espace de paramètres légèrement plus général (et plutôt artificiel) où le nombre de cellules souches dans les différentes cryptes peut différer d'une unité, permettant ainsi une mise à jour de N crypte par crypte. Les détails se trouvent dans *Supplementary Information* de Nicolas *et al.* (2007a). Bien que non nécessaire en pratique, il est ensuite possible de se ramener à une inférence dans l'espace initial en ne gardant parmi les valeurs échantillonnées par l'algorithme MCMC que celles correspondant aux itérations où le nombre de cellules souches et le même dans toutes les cryptes (conditionnement).

La validité du modèle a été étudiée grâce à des techniques dites de *posterior model assessment* (Gelman *et al.*, 1996), qui consistent à explorer la compatibilité entre les caractéristiques de données simulées et les données réelles. Les résultats montrent que les profils observés sont compatibles avec notre modèle relativement simple de généalogie mais nécessitent de prendre en compte les dépendances inter-sites dans le processus de méthylation. Or, ce genre de dépendances est connu pour rendre très difficiles le calcul de la probabilité d'évolution d'une séquence en une autre le long d'une branche de l'arbre (Jensen & Pedersen, 2000; Siepel & Haussler, 2004). Ici, j'ai proposé un modèle original est relativement économe du point de vue du coût des calculs et du nombre de paramètres qui fait dépendre les taux de méthylation et dé-méthylation du nombre de sites déjà méthylés au locus considéré. Les calculs dans ce modèle restent accessibles grâce au petit nombre de sites et aux symétries dans les taux de mutation (*Supplementary Information* de Nicolas *et al.* (2007a)). Les résultats semblent indiquer que chaque crypte du colon humain est maintenue par un nombre relativement élevé de cellules souches (supérieur à 10).

2.4 Modèles pour le lissage et la détection de rupture dans les profils transcriptomiques le long d'un génome

2.4.1 Lissage de données de *tiling arrays*

Une variable cachée à valeurs discrètes pour reconstruire le niveau d'expression

L'avènement de technologies permettant de caractériser les profils transcriptomique de façon globale sans connaissance *a priori* des régions transcrites (d'abord les *tiling arrays* puis le RNA-Seq) a motivé le développement de méthodes statistiques pour traiter le signal transcriptomique mesuré le long d'un génome. Ainsi analysés, ces signaux peuvent ensuite être confrontés à un ensemble de prédictions faites sur la base de la séquence d'ADN (prédiction des régions codantes pour des protéines, des promoteurs, des terminateurs de transcription) pour aboutir à un catalogue des régions fonctionnelles (ARN non codants, promoteurs, terminateurs) et à de nouvelles observations intéressantes (cf. section 3.1).

Lorsque nous avons reçu les premières données de *tiling arrays* du projet Basysbio, les plus satisfaisantes des méthodes disponibles ne permettaient pas d'intégrer aisément une mesure d'incertitude dans la reconstruction du niveau d'expression et l'identification des positions des ruptures (Picard *et al.*, 2005; Huber *et al.*, 2006; Rasmussen *et al.*, 2009). Surtout, elles se fondaient sur une hypothèse de ruptures brusques motivée initialement pas l'analyse de données de type CGH (*Comparative Genomic Hybridization*). Or, l'observation un peu attentive des données indiquait que la dynamique du signal le long du génome n'était pas uniquement faite de changements brusques. C'est dans ce contexte que j'ai développé une nouvelle méthode pour reconstruire, à partir du signal d'hybridation sur un *tiling array*, une trajectoire lissée et y détecter les points de rupture (promoteurs, terminateurs). Le modèle utilisé est original, il consiste à approcher une chaîne de Markov cachée à espace d'états cachés continu évoluant selon des ruptures brusques (*shifts*) et des dérives progressives (*drifts*).

Le HMM considéré suppose une émission gaussienne du signal log-transformé associé à la position t, noté Y_t, conditionnellement à un niveau d'expression sous-jacent U_t, que l'on cherche à reconstruire. Ici t correspond à l'index de sondes régulièrement espacées (d'environ 20 pb) sur le génome. L'évolution de U_t le long de la séquence est gérée par un noyau de transition markovien $\pi(u_t, u_{t+1})$ qui autorise à rester sur place, à sauter brusquement, ou à évoluer plus progressivement. Bien que le vrai niveau d'expression sous-jacent soit une variable continue, on suppose ici U_t à valeur sur une grille régulière de $\mathcal{S}$ points entre $u_{\min}$ et $u_{\max}$. Cela permet d'approcher une distribution continue avec un pas de discrétisation $h = (u_{\max} - u_{\min})/(\mathcal{S}-1)$. Pour $u_{\min} < u_t, u_{t+1} < u_{\max}$, la probabilité de transition sur la grille s'écrit alors comme un mélange de quatre types de mouvements

$$\begin{aligned}\pi(u_t, u_{t+1}) \quad = \quad & \alpha_n \mathbb{I}_{\{u_{t+1}=u_t\}} + \alpha_s \eta(u_{t+1}) \\ & + \alpha_u \mathbb{I}_{\{u_{t+1}>u_t\}} \lambda_u^{\frac{u_{t+1}-u_t}{h}-1}(1-\lambda_u) \\ & + \alpha_d \mathbb{I}_{\{u_t<u_{t-1}\}} \lambda_d^{\frac{u_t-u_{t+1}}{h}-1}(1-\lambda_d)\,, \end{aligned} \tag{2.8}$$

avec $0 \leq \alpha_n, \alpha_s, \alpha_u, \alpha_d \leq 1$, $\alpha_n + \alpha_s + \alpha_u + \alpha_d = 1$ et $0 \leq \lambda_u, \lambda_d < 1$. Une écriture un peu

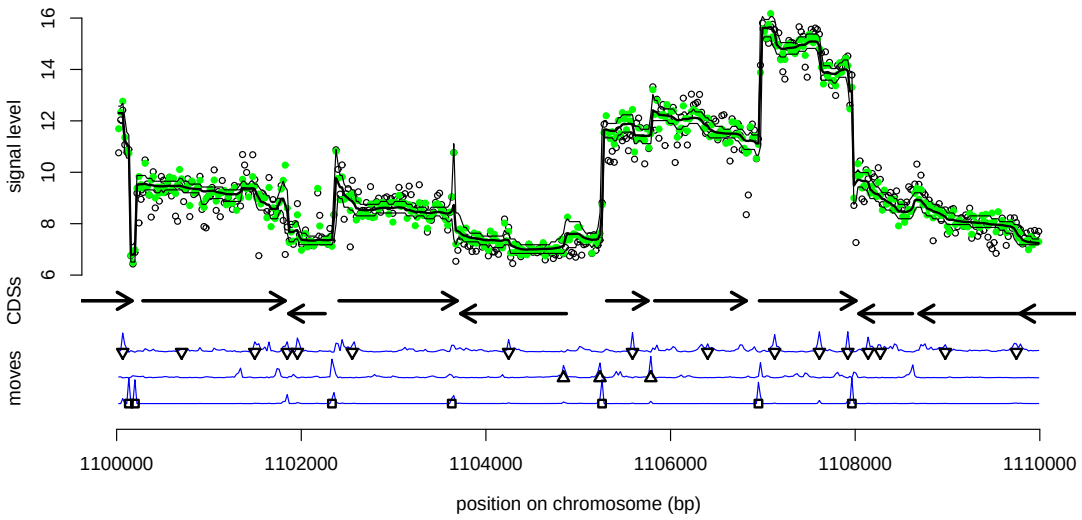


Figure 2.8 – Lissage par HMM d'un profil d'expression le long d'un génome obtenu avec la technologie *tiling arrays*. On représente ici le signal sur le brin (+) d'un segment de 10 000 pb du génome de *B. subtilis*. Dans la partie supérieure, les cercles ouverts représentent le signal d'origine, les cercles verts représentent le signal tel que corrigé par notre modèle grâce à la covariable capturant les différences d'affinité des sondes. Les lignes noires montrent l'espérance du signal caché reconstruit et l'intervalle de probabilité 95%. Les flèches horizontales représentent les CDS annotés. Dans la partie inférieure, les symboles indiquent la position des mouvements de *shift* (carrés) et de *shift* (triangles pointant vers le haut et vers le bas) dans la trajectoire la plus probable (c.-à-d. le chemin de Viterbi). Les trois courbes bleues montrent sur des lignes différentes les probabilités (variant entre 0 et 1) associées à chaque type de mouvement telles que calculées avec l'algorithme *forward-backward*.

plus laborieuse est nécessaire pour gérer le comportement aux bornes $u_{\min}$ et $u_{\max}$. Les paramètres de proportions du mélange dans ce noyau de transition sont : α_n, la probabilité de rester sur place ($u_{t+1} = u_t$) ; α_s, la probabilité de *shift* ; α_u et α_d les probabilités de *drift* vers le haut et vers le bas. Quant à λ_u et λ_d, ce sont les paramètres de lois géométriques qui gèrent les amplitudes des mouvements de *drift*. L'amplitude moyenne d'un *drift* de paramètre λ s'écrit $h + h/(1-\lambda)$. Lorsque $h \to 0^+$ et $h + h/(1-\lambda) \to \gamma$, le processus caché à valeurs discrètes approche un processus à valeurs continues. On montre en faisant varier $\mathcal{S}$ que $\mathcal{S} = 100$ correspond à une discrétisation suffisamment fine pour les données considérées.

Nicolas *et al.* (2009) donne plus de détails sur le modèle. La loi d'émission de $\pi(y_t \mid u_t)$ fait notamment intervenir une covariable visant à rendre compte de l'affinité de la sonde t.

Un algorithme de reconstruction du chemin rapide grâce à la forme des probabilités de transition

La forme de la probabilité de transition exprimée dans l'équation 2.8 permet un algorithme *forward-backward* en $O(\mathcal{T} \times \mathcal{S})$ au lieu de $O(\mathcal{T} \times \mathcal{S}^2)$. Sans entrer ici dans les détails des calculs (cf. *Supplementary Information* de Nicolas *et al.* (2009)), ce sont les propriétés de perte de mémoire de la loi géométrique utilisé pour l'amplitude des mouvements de *drift* qui permettent d'utiliser des récursions afin de calculer en temps $O(\mathcal{S})$ les termes impliqués dans les étapes de « prédiction » (équation 2.1) et de « lissage » (équation 2.3) à la position t pour toutes les valeurs u_t.

Les paramètres du modèle sont estimés par maximum de vraisemblance grâce à un algorithme EM. La reconstruction du chemin caché avec l'algorithme *forward-backward* est implémentée de façon à calculer tous les termes nécessaires pour les mises à jour des paramètres dans l'étape M. On y calcule notamment la probabilité de chaque type de mouvement à chaque position qui est utile pour la mise à jour des proportions du mélange (paramètres α) mais aussi pour la prédiction des positions des sauts dans le signal sous-jacent. De même, les valeurs obtenues pour $\pi(u_t \mid y_{1:\mathcal{T}})$ servent à la mise à jour des paramètres de la loi d'émission mais aussi à construire un intervalle pour la valeur de u_t conditionnellement à $y_{1:\mathcal{T}}$. La figure 2.8 illustre la reconstruction du signal d'expression (chemin caché) obtenue par cette approche.

Comme je l'explique dans la section 3.1, ces probabilités concernant le type de mouvement et la valeur de u_t se sont avérées très utiles pour construire une annotation unique des régions exprimées et des points de rupture à partir de nombreuses expériences de *tiling array* (Nicolas *et al.*, 2012 ; Mäder *et al.*, 2016).

2.4.2 Le cas du RNA-Seq

Du HMM au *State Space Model*

À la suite du travail sur le lissage des données de *tiling arrays*, je me suis intéressé à l'analyse de données obtenues avec la technologie plus récente du RNA-Seq rendue possible par les ruptures technologiques dans les méthodes de séquençage au cours de la décennie 2000-2010. Le RNA-Seq restera probablement assez longtemps la technologie de référence pour la transcriptomique. Alors que le *tiling array* reposait sur l'acquisition d'un signal continu correspondant à une intensité lumineuse reflétant le niveau d'hybridation sur des sondes avec une résolution d'environ 20 pb sur le génome, le RNA-Seq consiste à collecter de très nombreuses lectures de courts fragments d'ARN. Dans la perspective adoptée ici, on aligne ces lectures sur la séquence du génome et on en compte le nombre débutant à chaque position du génome. Les données correspondent donc à des comptages avec une résolution de 1 pb que l'on pourrait attendre, idéalement, indépendants entre positions adjacentes conditionnellement au niveau d'expression du gène.

Comme pour les *tiling arrays*, il s'agissait d'utiliser ces données pour déterminer les régions transcrites, les niveaux de transcription, et les points de rupture dans les profils qui peuvent correspondre à des éléments fonctionnels tels que les sites d'initiation et de terminaison de la transcription. Ces questions ont constitué le cœur du travail de

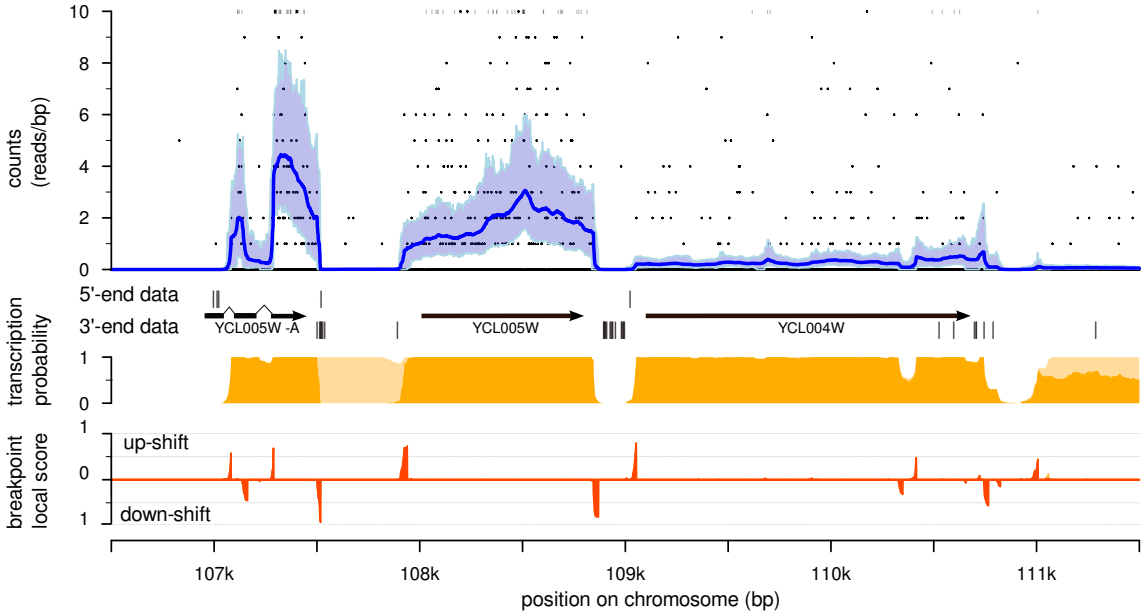


Figure 2.9 – Lissage par *State Space Model* d'un profil d'expression RNA-Seq le long d'un génome. On y représente 5 kpb d'un brin du chromosome de la levure *S. cerevisae*. Cette figure est l'analogue de la figure 2.8 pour des données de *tiling arrays*. On y voit les comptages bruts (points), la trajectoire reconstruite (moyenne : ligne bleu foncé ; intervalle à 95% zone en bleu plus clair). On y représente aussi la position des gènes annotés et des données disponibles sur les extrémités 5' et 3' des transcrits. En dessous, l'aire orange foncée représente la probabilité de dépasser le seuil d'expression de 0.1 lectures par pb et l'aire orange clair son complément à 0^+. Enfin, en rouge, on rapporte un score cumulant localement la probabilité de présence de points de rupture. Un autre exemple est représenté dans Mirauta *et al.* (2014).

thèse de Bogdan Mirauta que j'ai co-encadré avec Hugues Richard (direction Alessandra Carbone). Mon encadrement à notamment porté sur les aspects statistiques. Le point de vue méthodologique adopté fut assez radicalement différent de celui de mon travail précédent sur les *tiling arrays* puisque des modèles markoviens cachés à espace d'état continu (*State Space Models*) y ont remplacé l'approximation discrète présentée dans la section 2.4.1. Quant à l'inférence, elle a été conduite dans un cadre bayésien avec un algorithme MCMC. Pour cela, nous nous sommes appuyés sur une méthode récente de filtrage particulaire, dite *Particle Gibbs* qui offre un cadre souple pour reconstruire les chemins cachés à valeurs continues et qui permet d'intégrer cette reconstruction à un algorithme MCMC d'estimation des paramètres (Andrieu *et al.*, 2010). L'idée intuitive consiste à remplacer la grille utilisée dans Nicolas *et al.* (2009) par un échantillonnage aléatoire de l'espace en s'appuyant sur une population de particules. L'algorithme obtenu est exact (au sens d'un algorithme MCMC) : il ne requiert pas de faire tendre le nombre de particules vers l'infini pour échantillonner la loi *a prosteriori* (cible). Le nombre de particules et la fonction de proposition pour les trajectoires influencent en revanche l'efficacité des mises à jour et donc la vitesse de convergence de l'algorithme.

Le noyau de transition $\pi(u_t, u_{t+1})$ pour le signal que l'on s'est proposé de reconstruire est un mélange de plusieurs termes qui comme celui de l'équation 2.8 distingue des

mouvements de *shift* et de *drift*. Il s'écrit

$$\begin{aligned}\pi(u_{t-1}, u_t) = & \\ & \mathbf{1}_{\{u_{t-1}=0\}} \cdot \left[(1-\eta)\delta_0(u_t) + \eta \cdot f(u_t;\zeta)\right] \\ & + \mathbf{1}_{\{u_{t-1}>0\}} \cdot \Big[\alpha \cdot \delta_{u_{t-1}}(u_t) \\ & \qquad + \underbrace{\beta \cdot f(u_t;\zeta) + \beta_0 \cdot \delta_0(u_t)}_{shift} + \underbrace{\gamma_u \cdot g_u(u_t; u_{t-1}, \lambda) + \gamma_d \cdot g_d(u_t; u_{t-1}, \lambda)}_{drift}\Big] ,\end{aligned} \tag{2.9}$$

où $\zeta > 0$ et $\lambda > 0$ sont les taux des lois exponentielles gérant le niveau d'expression au début des transcrits ou après un mouvement de *shift* et l'amplitude relative des mouvements de *drift*. On remarque non seulement la masse ponctuelle en u_{t-1} qui permet au niveau de rester sur place ($u_t = u_{t-1}$) mais aussi celle en 0 qui permet de rendre compte du statut particulier du niveau d'expression 0 et qui n'a pas d'analogue dans l'équation 2.8. En effet, le bruit de fond inhérent à la technologie *tiling arrays* engendre une transition douce entre le niveau de signal dans les régions faiblement exprimées et non exprimées. Ici, u_t est exprimé en nombre de lectures attendues.

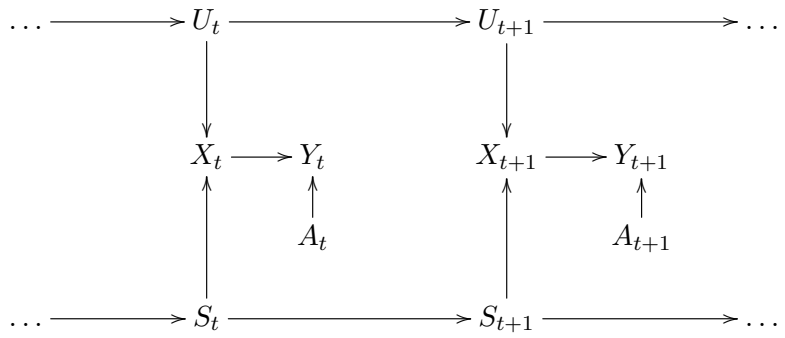
Mise au point d'une loi d'émission pour les comptage

Une autre dimension importante du travail publié dans Mirauta *et al.* (2014) a porté sur le développement d'une loi d'émission adaptée aux comptages RNA-Seq (correspondant ici au nombre de lectures débutant à une position donnée). En effet, tandis qu'après log-transformation les signaux d'hybridations continus des *tiling arrays* sont assez bien modélisés par une distribution gaussienne, les comptages RNA-Seq se sont montrés bien plus difficiles à modéliser que nous ne l'avions anticipé. Initialement nous avons utilisé une distribution binomiale négative. Celle-ci est populaire dans ce contexte (Anders *et al.*, 2013) et nous semblait a priori justifiée de part son interprétation comme un mélange continu de lois de Poisson : la distribution gamma sur le paramètre de la loi de Poisson (moyenne) permettant de rendre compte de la sur-dispersion et pouvant être mis à l'échelle pour rendre compte du niveau d'expression.

Cependant, ce modèle s'est avéré incapable de capturer simultanément quatre caractéristiques des données que nous avons identifiées dans nos tests comme particulièrement importantes pour reconstruire correctement le profil transcriptionnel. A savoir : la relation entre variance et espérance (l'espérance étant ce que nous avons défini comme le niveau d'expression) ; la relation entre probabilité d'un comptage nul et espérance (les comptages nuls entraînent des trous de couverture qu'il convient de distinguer des régions intergéniques) ; les caractéristiques des comptages à des positions apparemment isolées des régions clairement transcrites (régions faiblement transcrites ou artefacts) ; l'auto-corrélation à courte distance entre les comptages le long de la séquence. Notre objectif a été de développer un modèle paramétrique souple et relativement intuitif permettant de s'ajuster aux données expérimentales du point de vue de ces quatre caractéristiques.

Pour cela nous avons introduit, en plus de U_t correspondant au signal d'intérêt, une seconde variable cachée S_t de moyenne 1 résultant elle aussi d'une processus markovien

dans un espace continu, mais visant à capturer une auto-corrélation à beaucoup plus courte distance (quelques paires de bases ou dizaines de paires de bases). La construction du modèle d'émission fait aussi intervenir deux autres variables aléatoires intermédiaires, X_t et A_t, qui ne sont pas échantillonnées par le MCMC car elles sont éliminées par marginalisation (intégration réalisée numériquement). Le DAG complet est



où

$$\begin{aligned} S_t \mid s_{t-1} &\sim (1-\alpha_s)\delta(s_{t-1}) + \alpha_s Gamma(\kappa_s, 1/\kappa_s) \\ X_t \mid u_t, y_t &\sim Poisson(u_t s_t/\kappa\theta) \\ A_t &\sim Gamma(\kappa, \theta) \\ Y_t \mid x_t, a_t &\sim Poisson(x_t a_t)\,, \end{aligned}$$

ce qui garantit bien $\mathbb{E}_{\pi(y_t|u_t)}[Y_t] = 1$. La forme de ce modèle d'émission faisant intervenir trois variables aléatoires (S_t, X_t, A_t) en plus de U_t a été motivée par une interprétation mécaniste dans lequel S_t correspondrait à un biais local et auto-correlé d'échantillonage (fragmentation, ligation des adaptateurs), X_t au nombre initial de molécules d'ARN échantillonnées avant amplification, et A_t serait un biais d'amplification. La paramétrisation avec $(\kappa_s, \alpha_s, \kappa, \theta)$ peut sembler riche mais permet de bien rendre compte des caractéristiques des données expérimentales mentionnées ci-dessus comme importantes pour le lissage. Elle doit être mise en regard de la diversité des sources et des types d'artefacts rendus possibles par les protocoles expérimentaux. On remarque en particulier la forme de la variance

$$\mathbb{V}_{\pi(y_t|u_t)}[Y_t] = (1+\theta+\kappa\theta)u_t + \Big(\frac{1}{\kappa} + \frac{1}{\kappa_s} + \frac{1}{\kappa\kappa_s}\Big)u_t^2,$$

qui permet de bien s'ajuster aux données expérimentales. Le coefficient supérieur à 1 devant le terme en u_t est notamment nécessaire pour capturer la variance des comptages Y_t associés à de faibles valeurs de U_t. On comprend aussi aisément que la variable aléatoire discrète X_t autorise, pour certaines valeurs de paramètres (valeur élevée de $\kappa\theta$ que l'on peut interpréter comme le facteur d'amplification), une fraction non négligeable de comptages nuls même dans les régions fortement transcrites (u_t élevé).

Dans Mirauta *et al.* (2014), les paramètres α_s, κ_s, κ et θ sont estimés à partir de statistiques résumées avant la reconstruction du chemin caché et l'estimation des paramètres du noyau $\pi(u_t, u_{t+1})$. Le manuscrit de thèse (Mirauta, 2014) décrit des

approches améliorées pour l'estimation de paramètres de la loi d'émission et pour la mise à jour des trajectoires de u_t et s_t par *Particle Gibbs* en combinant une mise à jour par intervalles, une loi de proposition informative, et un ré-échantillonnage arrière. On interpréter les difficultés rencontrées pour trouver une loi d'émission relativement satisfaisante comme soulignant la complexité du lissage non-supervisé des données RNA-Seq. Ce constat contraste avec le cas des *tiling arrays* mais peut évoluer avec l'amélioration des protocoles expérimentaux.

Chapitre 3

Entre génome et transcriptome, entre cellule et population

Dans ce chapitre dédié aux objets biologiques abordés dans mes recherches, j'ai décidé de revenir sur deux séries de travaux. La première série est centrée autour de l'annotation structurale de *B. subtilis* et de la bactérie apparentée *S. aureus*, en s'appuyant sur des données transcriptomiques et en mettant en œuvre certains des modèles à variables latentes que j'ai décrits dans le chapitre précédent. La deuxième série de travaux concerne des analyses comparatives et évolutives chez les flavobactéries et pour lesquelles les liens avec le chapitre précédent sont plus ténus.

3.1 Architecture des transcriptomes de *B. subtilis* et *S. aureus*

De nombreux jeux de données de *tiling arrays* sur *B. subtilis* et *S. aureus* ont été collectés par différents laboratoires partenaires dans le cadre du projet européen Basysbio (coordonné par Philippe Noirot). Une caractéristique de ces données était la grande standardisation des protocoles expérimentaux : une souche utilisée, extraction de l'ARN réalisée dans chaque laboratoire selon un même protocole, préparation des ADNc marqués et hybridation sur les puces par un prestataire unique. L'idée a rapidement pris forme d'utiliser ces données et de les compléter pour une « ré-annotation » globale des génomes, à la lumière de transcriptomes représentatifs d'un nombre aussi grand que possible de conditions biologiques.

Parmi les laboratoires partenaires, deux ont joué des rôles particuliers en contribuant massivement à couvrir l'espace des conditions pour les besoins de la ré-annotation et en s'investissant dans l'interprétation des résultats. Il s'agit de MICALIS à l'INRA de Jouy-en-Josas pour *B. subtilis*, avec en particulier Étienne Dervyn, Tatiana Rochat, Elena Bidnenko et Philippe Noirot, et de l'Institut für Mikrobiologie de l'Université de Greifswald, avec Ulrike Mäder et Uwe Völker. L'équipe de Jan Maarten van Dijl du Department of Medical Microbiology de l'Université de Gröningen a aussi eu une contribution importante dans la collecte et l'analyse de données sur *S. aureus* et s'est beaucoup investie dans les analyses des petits ARNs détectés chez *B. subtilis* à travers la thèse de Ruben Mars.

Les analyses sur *B. subtilis* et *S. aureus* sont respectivement décrites dans Nicolas *et al.* (2012) et Mäder *et al.* (2016). Comme les démarches sont assez similaires et la comparaison des résultats est intéressante, j'ai trouvé intéressant de traiter ici parallèlement des travaux réalisés sur ces deux bactéries.

3.1.1 D'une collection d'expériences à une annotation structurale

Le développement des technologies de transcriptomique globale a permis d'obtenir des images assez complètes et fidèles de transcriptomes dans différentes conditions ou différents contextes génétiques (souches, mutants). Ces données ont ouvert une fenêtre sur le monde des ARN non codants qui ne pouvaient pas être détectés *in silico* à la façon des CDS et ont ainsi suscité un intérêt considérable pour leur rôle dans les processus de régulation (Sorek & Cossart, 2010).

Un des défis de l'analyse de grands jeux de données de ce type (Toledo-Arana *et al.*, 2009; Güell *et al.*, 2009; Sharma *et al.*, 2010) consista tout naturellement à établir des répertoires aussi complets que possible des éléments du transcriptome (promoteurs, terminateurs, régions transcrites, unités de transcription). Plusieurs facteurs compliquent cette tache, à laquelle je fait référence ici sous le terme d'« annotation structurale » en contribuant à une grande intrication des différents éléments. Il s'agit de la proximité sur le génome, de l'absence de correspondance « un terminateur pour un promoteur », de la plasticité du transcriptome à travers les conditions et, comme on le verra, des comportements apparemment « imparfaits », notamment au niveau de la terminaison (absence de terminateur, terminaison partielle).

Dans ce contexte, un aspect important de mon travail sur les données transcriptomiques a été de formaliser une approche pour l'annotation structurale. Celle-ci se fonde largement sur les possibilités ouvertes par l'algorithme développé pour le lissage des données de *tiling array* (section 2.4.1) qui a été l'instrument principal pour combiner les résultats des nombreux échantillons d'ARN (269 pour *B. subtilis*, 156 pour *S. aureus*) en une annotation unique pour chaque bactérie. En particulier, la possibilité d'associer des mesures de confiance à la détection des régions transcrites (probabilité pour le signal caché représentant le niveau de transcription de dépasser un seuil) et des points de rupture dans les profils s'est révélée très pertinente. La reconstruction de trajectoires lissées mais fidèles a été aussi très utile pour l'exploration visuelle des données et leur interprétation biologique en facilitant les représentations graphiques.

La procédure de construction de l'annotation est décrite en détail dans le *Supplementary Information* de Nicolas *et al.* (2012). Elle peut être décomposée en quatre étapes :

1. Établir des répertoires séparés de promoteurs (*up-shifts*), de terminateurs (*down-shifts*) et de régions transcrites (signal lissé supérieur à un seuil). L'approche adoptée a consisté à fusionner des prédictions « élémentaires » réalisées indépendamment sur les différents échantillons d'ARN en étant très strict sur les niveaux de confiance. Les seuils utilisés correspondent à une probabilité de *shift* supérieure à 0.99 pour les promoteurs et les terminateurs, une probabilité supérieure

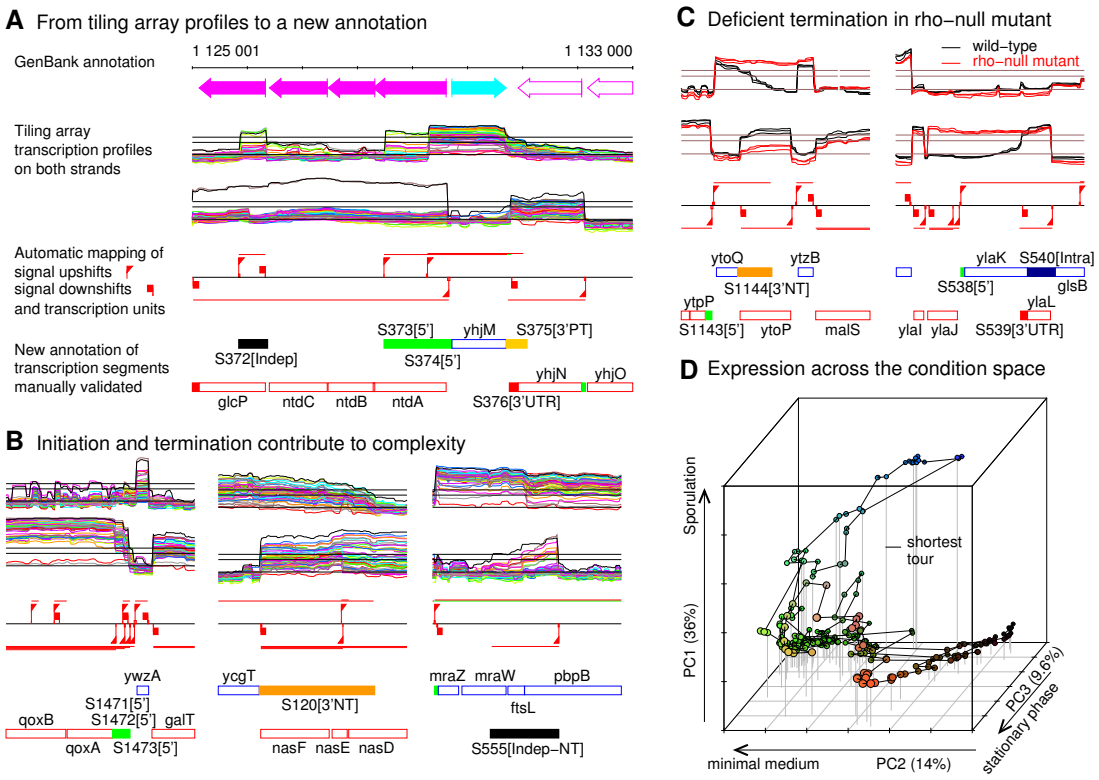


Figure 3.1 – Paysage transcriptionnel de *B. subtilis*. [A] Une région du génome (8 kpb) avec les profils de 50 échantillons d'ARN représentatifs de la plasticité du transcriptome. Les niveaux de signal correspondant à la médiane et aux seuils de 5x et 10x par rapport à la médiane sont représentés par des lignes horizontales. Les nouveaux segments sont numérotés S1 à S1583, leur couleur reflète le type de région (indiqué entre crochets). Ainsi, le gène *yhjM* est précédé de deux segments de type 5' (S373, S374) correspondant à deux promoteurs et il est suivi d'un segment 3'PT (S375) engendré par une terminaison partielle. Le segment S372 de type Indep est antisens de *glcP*. [B] Quatre promoteurs sont identifiés pour les gènes *qoxAB*. Un fort bruit de fond est observé sur le brin opposé (artefact de *reverse transcription*). L'absence de terminateur en aval du gène *ycgT* engendre S120, un segment de type 3'NT antisens des gènes *nasDEF*. Le segment S555 naît d'un promoteur qui n'est pas apparié à un site de terminaison défini, il est antisens des gènes *pbpB* et *ftsL* (dont les produits sont impliqués dans la division cellulaire). [C] Profils du mutant ΔRho (en rouge) comparés à ceux de la souche de référence (en noir). Dans le mutant les ARN des gènes *ytoQ* et *ylaL* qui n'ont pas de site de terminaison intrinsèque s'étendent jusqu'au prochain terminateur. [D] Projection des 269 transcriptomes sur les 3 premiers axes d'une analyse en composante principale. Cette représentation sur seulement trois axes capture ≈60% de la variance totale et reflète des propriétés importantes des conditions de croissance. La ligne continue qui relie les conditions représente le plus court chemin faisant le tour de toutes les conditions (solution au problème dit « du voyageur de commerce »). Figure reproduite de Nicolas *et al.* (2012).

.

à 0.975 de dépasser le niveau de 10x par rapport à la médiane du chromosome pour les régions transcrites.

2. Subdiviser les régions transcrites en segments en distinguant les gènes déjà annotés (en pratique essentiellement les CDS de l'enregistrement Genbank) et en les coupant aux positions des promoteurs et des terminateurs.
3. Annoter les nouveaux segments en fonction de leur contexte transcriptomique, en prenant notamment en compte la position dans l'unité de transcription et la position par rapport aux gènes annotés. Nous avons pour cela défini 8 catégories décrites ci-dessous. Certains segments ont aussi été éliminés car considérés comme des artefacts. Une liste d'artefacts potentiels parmi les segments a été établie sur la base d'un indice de reproductibilité biologique, en pratique la p-valeur associée à l'ANOVA liant le niveau d'expression à la condition biologique. L'annotation des segments a finalement fait l'objet d'une validation manuelle par deux biologistes (Ulrike Mäder et Philippe Bessières) fondée sur l'analyse visuelle des profils et du contexte chromosomique.
4. Délimiter les unités de transcription associées aux promoteurs en prenant en compte la diversité des trajectoires reconstruites en aval de chaque promoteur sur l'ensemble des échantillons. En pratique, deux bornes 3' ont été associées à chaque promoteur à partir des régions maximales au dessus d'un seuil de 5x par rapport à la médiane du chromosome : l'une en interdisant la présence de *down-shifts* (TU-short), l'autre en l'autorisant (TU-long). Chaque promoteur a ainsi été associé à deux listes de gènes (ceux dans la TU-short et ceux dans la TU-long), la première liste étant la plus stricte.

Au total, 1 583 nouveaux segments ont ainsi été définis et classifiés chez *B. subtilis* et 1 192 chez *S. aureus* (pour des génomes respectivement de 4,2 Mpb et 2,8 Mpb). Les figure 3.1 A et B illustrent les profils transcriptionnels obtenus, la construction de l'annotation, et les différents types de régions. Ces figures montrent aussi deux éléments important de la complexité du transcriptome : la multiplicité des promoteurs qui engendre un chevauchement des unités de transcription, et l'absence de site de terminaison bien défini pour certaines unités de transcription qui est associée à des régions où le niveau d'expression décroît graduellement (comportement que nous avons modélisé par le *drift*, section 2.4.1). Nous avons introduits les termes **NT** (*no termination*) et **PT** (*partial termination*) pour désigner ces régions. Les 8 classes de segments que nous avons définies distinguent : **5'UTR** (au début d'une unité de transcription), **Intra** (entre deux gènes d'une même unité de transcription, c.-à-d. avec le même promoteur), **Inter** (entre deux gènes avec des promoteurs distincts), **Indep** (indépendant des gènes annotés), **Indep-NT** (indépendant des gènes annotés sans terminateur), et **3'** (à la fin d'une unité de transcription avant le terminateur), **3'PT** (à la fin d'une unité de transcription après un site de terminaison incomplète), et **3'NT** (à la fin d'une unité de transcription sans site de terminaison défini).

Les profils de transcription le long des génomes de *B. subtilis* et *S. aureus* sont mis à disposition des communautés de biologistes concernés via

deux sites web dédiés (http://genome.jouy.inra.fr/cgi-bin/seb/index.py et http://genome.jouy.inra.fr/cgi-bin/aeb/index.py). L'article Mars *et al.* (2016) donne une vue globales des ARN de fonction connue chez *B. subtilis* et propose une analyse détaillée des nouveaux segments du point de vue de la conservation dans les espèces proches et de la possibilité de repliement des ARNs. Cette étude propose une liste de segments à étudier en priorité dans la perspective de recherche de nouveaux ARN régulateurs.

3.1.2 Les régulons des facteurs sigma

Le développement de l'algorithme de classification des promoteurs selon les sites de fixations des facteurs sigma (section 2.2.2) a été motivé par la perspective de l'analyse du répertoire de promoteurs de *B. subtilis*. Pour cette bactérie qui ne possède pas moins de 20 facteurs sigma connus (Michna *et al.*, 2016), les 3 242 *up-shifts* identifiés lors de l'annotation structurale offrent probablement une bonne couverture de l'ensemble des promoteurs. Sur les 600 TSS identifiés à la résolution de 1 pb par Irnov *et al.* (2010) par séquençage des extrémités 5' des transcrits, 93% sont positionnés à proximité immédiate (distance inférieure ou égale au pas du *tiling*, 22 pb) d'un *up-shift* ; le même calcul pour les 733 sites de fixations de facteurs sigma répertoriés dans DBTBS (Sierro *et al.*, 2008) indiquait une couverture de 85%. D'après notre délimitation des unités de transcription (TU-long), les promoteurs inclus dans ce catalogue de 3 242 *up-shifts* permettent la transcription de 93% des CDS annotés, avec plus d'un promoteur pour 46% de ceux-ci.

L'algorithme de classification des promoteurs a été appliqué aux promoteurs de *B. subtilis* et *S. aureus* (3 242 et 1 523 *up-shifts*, respectivement). Les résultats sont représentés dans la figure 3.2 et ont été publiés dans Nicolas *et al.* (2012) et Mäder *et al.* (2016). Le nombre de motifs recherchés a été fixé à 20 pour *B. subtilis* et 4 pour *S. aureus*. Ces choix reflètent les différences biologiques entre les deux espèces : on n'attend que 2 facteurs sigma actifs dans la souche de *S. aureus* étudiée contre une vingtaine chez *B. subtilis*. L'algorithme a permis de prédire un facteur sigma associé (en se fondant sur un seuil de probabilité 0.5 d'assignation aux composantes du modèle) pour 90% (*B. subtilis*) à 93% (*S. aureus*) des *up-shifts*, confirmant ainsi que la grande majorité des *up-shifts* de notre répertoire correspondent bien à des promoteurs. Bien que ce ne soit pas représenté sur cette figure, il y avait aussi une excellente correspondance entre les assignations aux groupes et les sites de fixation de facteurs sigma répertoriés dans DBTBS pour *B. subtilis*.

La correspondance n'est cependant pas toujours de un facteur sigma pour un motif. On trouve ainsi 6/20 motifs alloués à la description des sites de fixation du principal facteur sigma (SigA) chez *B. subtilis* et 2/4 chez *S. aureus*. À l'inverse, certains des motifs identifiés chez *B. subtilis* correspondent à plusieurs facteurs sigma. Parmi les facteurs sigma impliqués dans la sporulation, SigE, SigF, et SigG ne sont ainsi associés qu'à deux motifs avec des chevauchements entre, d'une part, les deux facteurs activés dans la phase précoce (SigE pour la cellule mère et SigF pour la préspore), et d'autre part, les deux facteurs sigma activés dans la préspore (SigF pour la phase précoce et SigG pour la phase tardive). Ces composantes référencées comme SigEF et SigFG dans la figure 3.2), sont bien séparées de SigK (cellule mère, phase tardive). De la même façon, les facteurs sigma ECF (*Extracytoplasmic Function*) impliqués dans les réponses aux stress de l'enveloppe

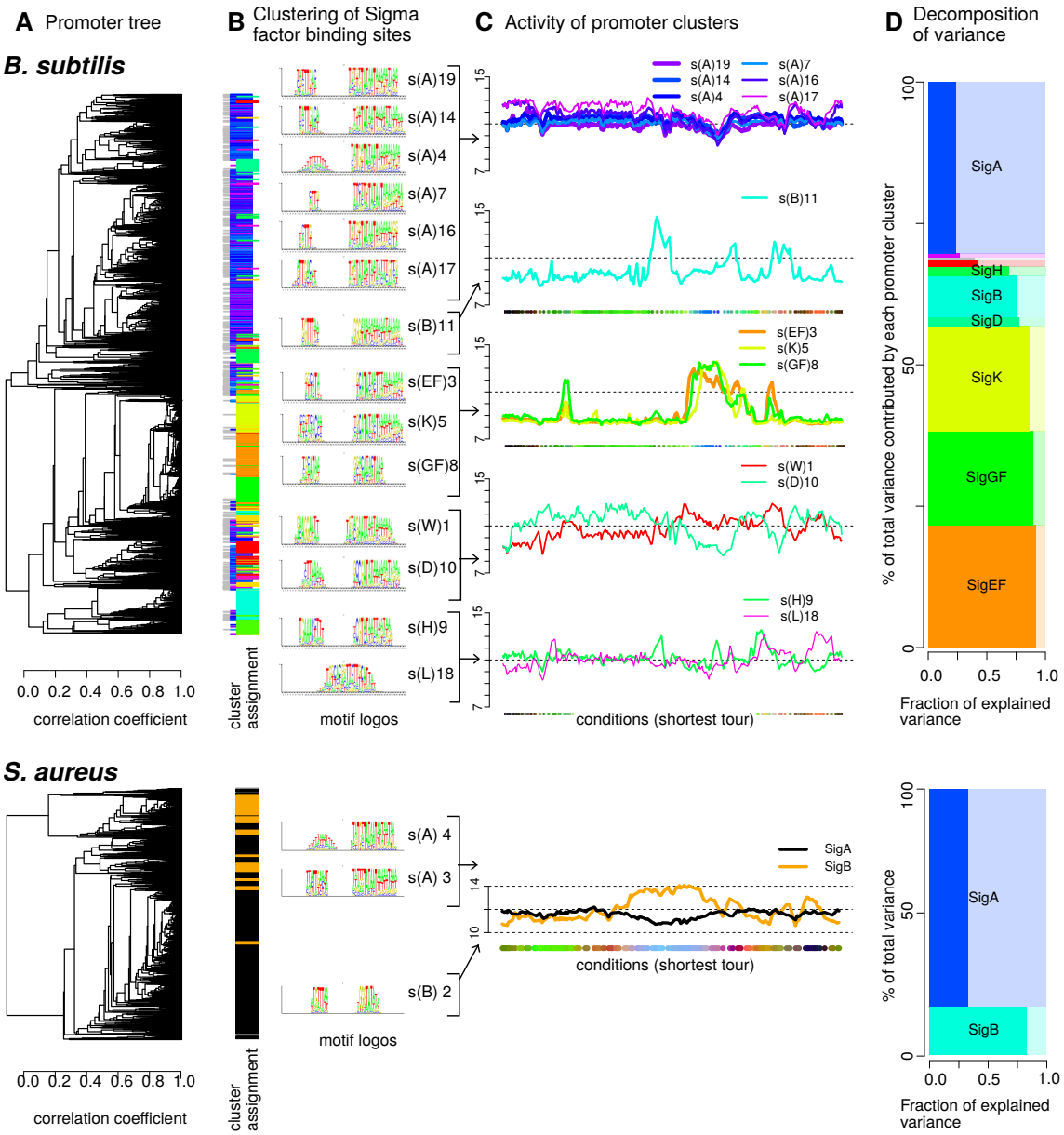


Figure 3.2 – Classification des promoteurs de *B. subtilis* et *S. aureus*. [A] Arbre de classification hiérarchique résumant les corrélations entre paires de promoteurs (ordonnés sur l'axe y). La hauteur des arbres reflète le nombre de promoteurs (3 242 *up-shifts* pour *B. subtilis* contre 1 523 pour *S. aureus*). [B] Chaque promoteur est représenté sous la forme d'une barre horizontale dont la couleur reflète les classes identifiées par l'algorithme de découverte de motif (section 2.2.2). Les promoteurs pour lesquels l'assignation de la classe est incertaine sont représentés en gris. Pour chaque classe, le nom du facteur sigma correspondant est indiqué (sous la forme d'une lettre) avec le numéro de la classe. [C] Activité estimée (axe y, échelle log2) des promoteurs de chaque classe à travers les conditions (voir figure 3.1 D pour la disposition des conditions sur l'axe x). [D] Décomposition de la variance totale de l'activité des promoteurs selon les différentes classes. La variance totale associée à chaque classe de promoteurs est représentée sous la forme d'une aire rectangulaire divisée verticalement en deux parties : à gauche (couleur de plus grande intensité) la fraction expliquée, à droite la fraction non expliquée. Figure adaptée de Nicolas *et al.* (2012) et Mäder *et al.* (2016).

cellulaire sont regroupés au sein d'une seule composante. Il s'agit de regroupements cohérents avec le déroulement spatio-temporel du processus de sporulation (Eichenberger *et al.*, 2004) et avec le chevauchement (*cross-talk*) connu des régulons des facteurs sigma ECF (Helmann, 2016).

Cette caractérisation globale des régulons a permis une tentative de quantification de la contribution de chaque facteur sigma à la plasticité du transcriptome. Pour cela, l'activité des régulons à travers les conditions a d'abord été estimée (comme illustré dans la figure 3.2 C) puis la fraction de la variance de l'activité de chaque promoteur pouvant être expliquée par l'activité du régulon a été calculée en supposant une relation linéaire

$$y_{p,t} \quad = \quad a_p x_{Sig(p),t} + b_p + \epsilon_{p,t}\,, \quad \epsilon_{p,t} \sim N(0, \sigma_p^2)\,,$$

où $y_{p,t}$ et $x_{Sig(p),t}$ représentent respectivement l'activité en échelle logarithmique du promoteur p et du facteur sigma correspondant (noté $Sig(p)$) dans la condition t, et $\epsilon_{p,t}$ un résidu modélisé par une variable aléatoire gaussienne de variance σ_p^2. La figure 3.2 D présente les résultats de la composition de la variance totale du transcriptome selon cette procédure. Chez *B. subtilis*, 66% de la variance a ainsi été attribuée à l'activité des facteurs sigma. Cette fraction élevée est principalement due aux régulons des facteurs sigma liés à la sporulation (SigE, SigF, SigG, et SigK), à la réponse générale au stress (SigB) à l'intérieur desquels la part de variance expliquée dépasse 75%. À l'opposé, elle n'est que de 24% pour le régulon SigA, au sein duquel le rôle des autres mécanismes de régulations (notamment d'autres facteurs de transcription) est probablement prépondérant.

La comparaison entre *B. subtilis* et *S. aureus* est intéressante. La fraction de variance expliquée au sein des régulons SigA et SigB est très similaire entre les deux espèces, de même que le rapport entre les contributions des régulons SigB et SigA à la variance totale. Cependant, l'absence de sporulation et de facteurs sigma associés fait que la fraction de la variance totale expliquée n'est que de 42%. Nous avons aussi noté et discuté des différences très importantes entre les deux espèces de bactéries en ce qui concerne le profil d'activité du régulon SigB (Mäder *et al.*, 2016) dont l'amplitude de variation d'activité est bien plus élevée et les conditions d'induction bien plus spécifiques chez *B. subtilis* que chez *S. aureus*.

3.1.3 Origine des ARN antisens

Les approches de transcriptomique globale chez les bactéries ont révélé une abondance d'ARN dits « antisens » car transcrits à partir du brin opposé à celui d'un autre ARN considéré comme « sens », celui-ci correspondant généralement à un gène codant pour des protéines (les CDS annotés couvrent environ 85% d'un génome bactérien typique). Cette découverte a généré un grand intérêt pour cette classe de transcrits susceptibles de moduler l'expression du brin sens par une variété de mécanismes pouvant agir aussi bien sur la transcription, la stabilité, que la traduction (Thomason & Storz, 2010). Nous avons pu faire des observations nouvelles sur l'origine de cette classe de transcrits grâce à la combinaison de (i) la disponibilité de profils correspondant à de nombreuses conditions, (ii) la méthode de lissage permettant d'observer finement les profils de transcription, (iii) la classification des segments, (iv) la délimitation des régulons des facteurs sigma à l'échelle du génome.

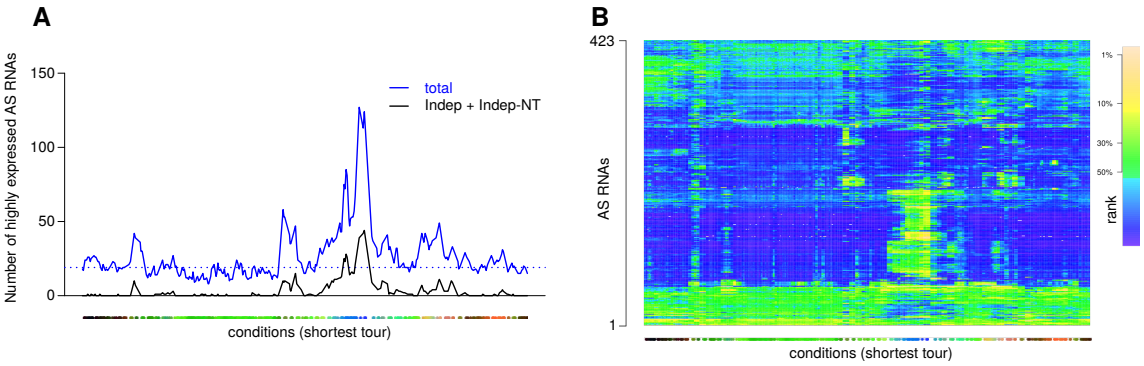


Figure 3.3 – Expression des ARN antisens de *B. subtilis* à travers les conditions. [A] Variation du nombre de segments ARN antisens parmi les 30% de segments les plus exprimés. [B] Représentation sous forme de *heat map*. Figure adaptée de Nicolas *et al.* (2012).

Il est notamment apparu que chez *B. subtilis* les ARN antisens ont des profils d'expression divers dont le niveau d'expression varie souvent fortement à travers les conditions, ce qui pourrait à première vue être interprété comme la confirmation de fonctions biologiques étroitement régulées (figure 3.3).

L'étude des promoteurs révèle que cette « régulation » s'explique essentiellement par le nombre de d'ARN antisens transcrits à partir de promoteurs dépendant de facteurs sigma alternatifs (48% pour les ARN antisens contre 26% pour les CDS). Cette seule caractéristique des ARN antisens engendre non seulement leur niveau d'expression fortement dépendant des conditions mais aussi une corrélation négative apparente avec l'expression des gènes auxquels ils font face, qui sont eux le plus souvent sous le contrôle de SigA (comme une majorité des gènes). De plus, la classification des segments a révélé que 62% des ARN antisens proviennent de contextes correspondant à une terminaison incomplète de la transcription (nos catégories 3'NT, 3'PT, Indep-NT, et Inter) et que la fraction d'antisens contrôlés par des facteurs sigma alternatifs est encore bien plus élevée pour les antisens ayant leur propre promoteur (catégories Indep et Indep-NT) où elle atteint 77%. Ainsi, c'est au total 82% des ARN antisens qui peuvent être reliés à l'activité de facteurs sigma alternatifs ou à des terminaisons incomplètes.

C'est observations m'ont amené à proposer l'hypothèse qu'une fraction importante des ARN antisens pourraient résulter d'un contrôle imparfait de la transcription et ne pas avoir de rôle biologique au sens habituellement entendu pour les gènes (Nicolas *et al.*, 2012). Leur présence résulterait alors d'un équilibre entre différents processus : la plasticité du transcriptome à travers les conditions, les mutations aléatoires susceptibles de créer des promoteurs et de détruire des terminateurs, et l'utilité d'un certain niveau de « bruit » pour permettre l'adaptation, notamment à travers l'apparition de nouveaux gènes et de nouvelles régulations.

Il est en particulier concevable que les facteurs sigma alternatifs puissent reconnaître par hasard des promoteurs sur la séquence du génome sans que cela ne soit fortement contre-sélectionné étant donné l'activité transitoire de ces facteurs sigma. De fait, mon analyse de la conservation des promoteurs responsables des antisens de type Indep et

Indep-NT a suggéré une moindre conservation qui pourrait résulter de l'absence de sélection pour le maintien d'une fonction (sélection purificatrice). De façon concomitante et indépendante, des comparaisons de transcriptomes entre les entérobactéries *E. coli* et *S. enterica* ont conduit à une observation similaire de moindre conservation des ARN antisens entre espèces et à invoquer une hypothèse de « bruit » transcriptionnel (Raghavan *et al.*, 2012). Le rôle des ARN antisens a ensuite été débattu sans que la question ne soit tranchée, la tendance serait cependant à vouloir trouver un rôle biologique à cette classe de transcrits (Lybecker *et al.*, 2014; Wade & Grainger, 2014).

Mes analyses ultérieures du transcriptome de *S. aureus* ont semblé conforter le point de vue d'une contribution importante du « bruit » transcriptionnel causé par les facteurs sigma alternatifs à la transcription antisens (Mäder *et al.*, 2016). L'absence de facteurs sigma alternatifs à activité transitoire s'accompagne d'un nombre moindre d'ARN antisens (51/Mpb chez *S. aureus* contre 100/Mpb chez *B. subtilis*), en particulier pour les types Indep et Indep-NT (6.7/Mpb contre 20.9/Mpb). Le seul facteur sigma alternatif SigB y a en effet une activité basale et des conditions d'activation bien plus nombreuses ; l'activité des promoteurs contrôlés par SigB est même supérieure à celle des promoteurs contrôlés par SigA dans une (courte) majorité des conditions étudiées. Quantitativement, les antisens de type Indep et Indep-NT contrôlés par SigB sont observés à un taux de 0.35/Mpb chez *S. aureus* contre 3.8/Mpb chez *B. subtilis*. Ces résultats suggèrent donc que l'abondance des ARN antisens dépend directement non seulement du nombre de facteurs sigma mais aussi de leurs profils d'induction.

3.1.4 Le(s) rôle(s) du facteur de terminaison de la transcription Rho

On distingue traditionnellement deux mécanismes de terminaison de la transcription de la transcription chez les bactéries. La terminaison intrinsèque fait intervenir les seules propriétés de la séquence transcrite, une tige boucle suivie d'un série de résidus `u`, et ses sites sur le génome peuvent être prédits de façon relativement fiable (d'Aubenton Carafa *et al.*, 1990; Kingsford *et al.*, 2007). L'autre mode de terminaison fait intervenir le facteur Rho et on ne sait pas en prédire les sites bien que certaines propriétés de la séquence soient apparemment discernables (Boudvillain *et al.*, 2013). La terminaison intrinsèque est aussi appelée Rho indépendante.

Dans le cadre de l'annotation structurale de *B. subtilis* et *S. aureus* nous nous sommes intéressés au facteur de terminaison de la transcription Rho, à travers la construction d'un mutant ΔRho et l'étude de son transcriptome. La motivation initiale était très générale : mieux comprendre le rôle de Rho sur la terminaison, notamment d'évaluer son implication dans les *down-shifts* qui n'étaient pas associés à des sites de terminaison intrinsèque détectables. Les résultats sur *B. subtilis* ont bien confirmé l'implication de Rho dans la terminaison à un certain nombre de sites pour lesquels il n'y avait pas de promoteur intrinsèque apparent. Cependant, ils ont surtout révélé le rôle de Rho dans la terminaison des régions de type NT et PT où la transcription semble s'étendre au delà de bornes d'unités transcription bien définies, phénomène engendrant souvent des ARN antisens comme représenté dans la figure 3.1 C (Nicolas *et al.*, 2012). Ainsi, Rho aurait un rôle dans la suppression de la transcription dite *pervasive*. Cela serait compatible avec

son mécanisme d'action tel que décrit dans la littérature qui repose sur un glissement le long de l'ARN, favorisé par l'absence de ribosomes (ces régions étant non traduites), jusqu'à rencontrer l'ARN polymérase et entraîner sa dissociation de l'ARN en cours de transcription. Parallèlement à notre travail, une analyse de transcriptome en présence de bicyclomycine, un inhibiteur de Rho, (*rho* est un gène essentiel pour *E. coli* mais pas pour *B. subtilis*) a aussi conclu à un rôle de Rho dans l'inhibition de la transcription dite *pervasive* (Peters *et al.*, 2012). Plus récemment, des résultats similaires ont été obtenus chez *M. tuberculosis* (Botella *et al.*, 2017).

L'analyse du mutant Δ*rho* chez *S. aureus* a produit des résultats un peu différents puisque l'étendue des régions dont la transcription est soumise au contrôle de (réprimée par) Rho s'étend à environ la moitié du chromosome, soit bien au delà des régions en aval des segments de type NT et PT détectés dans la souche possédant Rho (Mäder *et al.*, 2016). De plus, l'étendue de ces régions dépend de la condition de croissance. Nous l'avons trouvée maximale dans une condition de croissance rapide (phase exponentielle en milieu riche) lors de laquelle l'activité transcriptionnelle est certainement particulièrement intense. Cette différence pourrait être une conséquence de la richesse en `a+t` du génome de *S. aureus* (67,2% contre 56,5% pour *B. subtilis*, 49,2% pour *E. coli*) qui augmenterait le nombre de sites reconnus par le facteur SigA sur le génome et engendrerait un bruit de fond transcriptionnel contrecarré par Rho. En effet, même si les différences de compositions en mono-nucléotides peuvent sembler à première vue modérées, l'impact peut être considérable sur la fréquence attendue pour des oligonucléotides plus longs tel que la boîte -10 canonique `tataat`. J'ai ainsi observé que la fréquence de cette boîte est de 0,905/kpb dans les régions antisens des CDS chez *S. aureus* contre 0,332/kpb chez *B. subtilis* et 0,091/kpb chez *E. coli* (Mäder *et al.*, 2016).

L'effet de l'absence de Rho sur l'expression « sens » de nombreux gènes qui semble dû à une variété de mécanismes (figure 3.4) et des observations préliminaires suggérant des modifications phénotypiques assez subtiles de *B. subtilis* en absence de Rho ont éveillé l'attention d'Elena Bidnenko. Une analyse beaucoup plus poussée a alors été engagée sur des mutants ΔRho chez différentes souches de *B. subtilis* car la souche couramment utilisée au laboratoire n'exhibe pas toutes les caractéristiques phénotypique des souches sauvages, notamment en matière de différenciation cellulaire. Ces analyses auxquelles j'ai eu la chance de contribuer ont mené à un résultat à mon sens inattendu, qui est publié dans Bidnenko *et al.* (2017) : tout semble indiquer que Rho est impliqué dans l'implémentation de la bascule entre différents modes de vies mutuellement exclusifs (sessile vs. motile, biofilm, sporulation). Rho ferait ainsi partie intégrante du réseau de régulation de *B. subtilis* qui est justement un modèle d'étude pour son remarquable programme de différenciation. Ce résultat surprenant illustre les possibilités de recrutement de Rho et des ARN antisens dans des fonctions biologiques qui peuvent être spécifiques de l'organisme étudié. Ils illustrent aussi la pertinence de ne pas vouloir systématiquement opposer « bruit » transcriptionnel et fonction biologique.

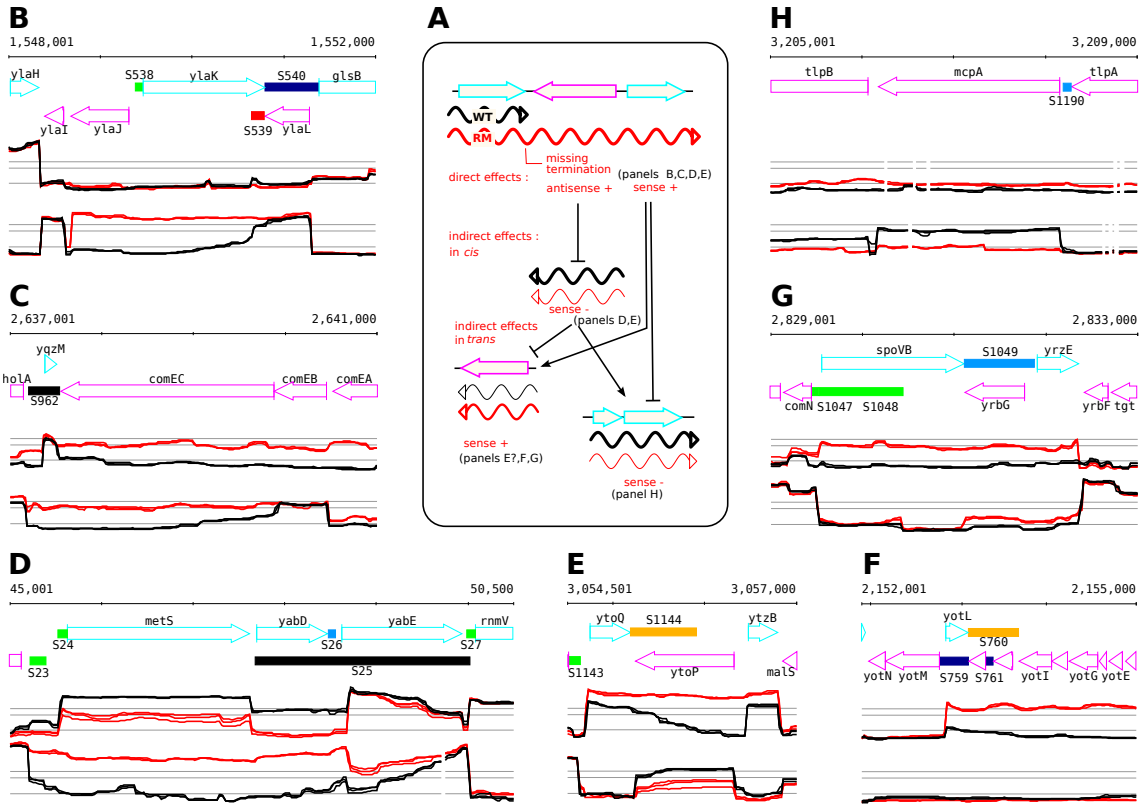


Figure 3.4 – Typologie des différents effets de l'absence de Rho sur le transcriptome de *B. subtilis*. [A] Illustration schématique des différents effets directs et indirects possibles (le transcrit sous contrôle de Rho est représenté en rouge). [B-F] Tentative d'illustration par des exemples concrets. Figure reproduite de Bidnenko *et al.* (2017).

3.2 Analyses évolutives chez les *Flavobacteriacae*

Flavobacterium psychrophilum est une bactérie pathogène des poissons dans les eaux douces plutôt froides qui cause notamment de gros dégâts dans les élevages de salmonidés. Elle est en particulier responsable de deux maladies bien connues des éleveurs : le *Rainbow Trout Fry Syndrome* et la *Bacterial Cold Water Disease*. Mes travaux sur cette bactérie se situent dans une collaboration de longue date avec l'équipe d'Éric Duchaud et Jean-François Bernadet de l'Unité Virologie et Immunologie Moléculaire de l'INRA de Jouy-en-Josas. Éric fut recruté à l'INRA en 2005 avec comme mission initiale de mener à bien l'annotation du génome d'une souche de *F. psychrophilum*. Cette annotation constituait aussi un projet pilote pour l'environnement bioinformatique d'annotation AGMIAL développé à MIG (Bryson *et al.*, 2006). Une suite logique à la publication du génome de cette souche JIP 02/86 (Duchaud *et al.*, 2007) fut d'entamer une analyse de la structure de population de l'espèce en s'appuyant sur les souches collectées depuis environ 20 ans par Jean-François Bernadet, vétérinaire et expert en taxonomie des *Flavobacteriacae* (Bernardet & Kerouault, 1989; Bernardet & Bowman, 2006).

3.2.1 Analyse de la structure de population de l'espèce *F. psychrophilum* par *Multi-Locus Sequence Typing*

L'approche MLST

L'idée était d'utiliser l'approche dite de *Multi-Locus Sequence Typing* (MLST) (Urwin & Maiden, 2003). Celle-ci consiste à caractériser les souches par séquençage d'un petit nombre de locus, classiquement sept, d'une longueur (souvent autour de 600 pb mais ici plutôt de 800 pb) permettant un séquençage grâce à deux lectures Sanger à partir d'amorces correspondant aux deux extrémités du locus. Ces locus sont choisis parmi les gènes en une seule copie et conservés à l'échelle taxonomique considérée (le « génome central »), souvent dans les gènes dits de « ménage » (c.-à-d. exprimés en permanence). L'idée sous-jacente est de sélectionner des locus porteur d'un polymorphisme essentiellement « neutre » qui est *a priori* le plus à même de refléter la structure de la population, notamment à travers la généalogie des organismes si le signal n'est pas effacé par les recombinaisons. L'approche MLST présente de nombreux avantages par rapport aux techniques utilisées antérieurement telles que l'analyse de fragments produits par amplification aléatoire ou par digestion avec des enzymes de restriction (RAPD, RFLP, PFGE, ribotypage, ...), l'établissement de profils plasmidiques, ou le sérotypage. En particulier, la technologie est aisément transférable d'un laboratoire à l'autre et elle fournit des données de type « séquence » qui sont faciles à stocker et à comparer. De plus, le polymorphisme identifié par MLST consiste essentiellement en des profils de SNP qui peuvent être analysés avec la panoplie des outils de la phylogénie et de génétique des populations et tend ainsi à être beaucoup plus perméable à l'interprétation que des longueurs de fragments ou des sérotypes.

Dans la terminologie des études MLST, chaque isolat est caractérisé par ses AT (pour *Allele Type*) qui sont les allèles trouvés aux différent locus du schéma MLST, et chaque

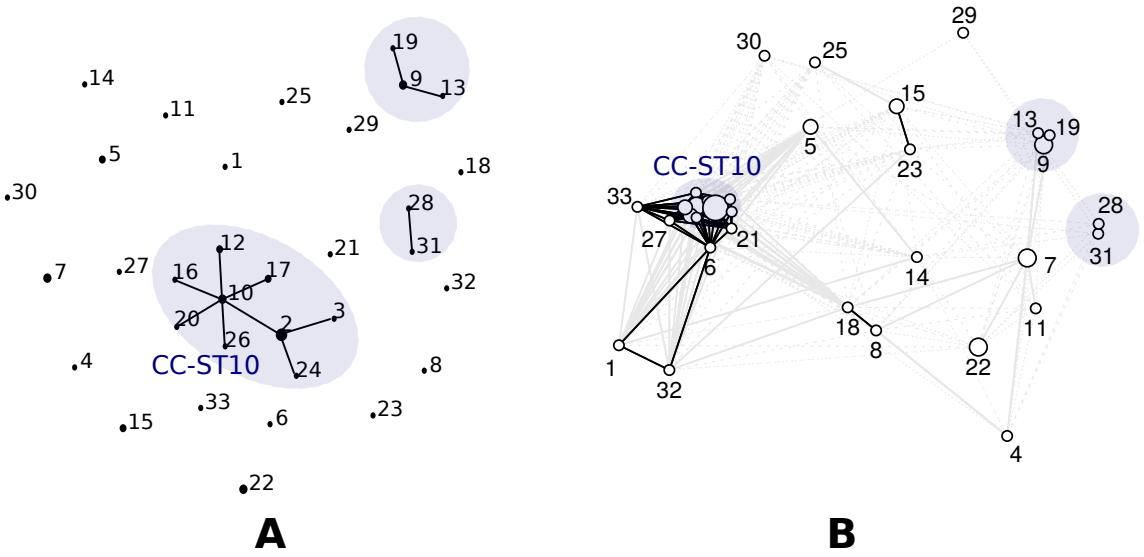


Figure 3.5 – Représentation sous forme de réseaux des relations entre les génotypes issus du premier jeu de données MLST pour *F. psychrophilum* constitué de 50 isolats (33 ST). Chaque point correspond à un ST dont la taille reflète le nombre d'isolats. [A] Diagramme eBURST (Spratt *et al.*, 2004) représentant les liens de type SLV (*Single Locus Variant*). Ces liens regroupent les ST en complexes clonaux qui sont entourés. [B] Réseau représentant les AT partagés entre ST. Le style de ligne reflète le nombre d'AT communs à une paire de ST : noire continue pour plus de 3 AT, gris continue pour 2 ou 3 AT, et gris pointillée pour 1 AT. La disposition des points dans l'espace est obtenue par *Multidimensional scaling*. Figure adaptée de Nicolas *et al.* (2008).

combinaison d'AT définit un ST (*Sequence Type*). Les AT et les ST sont identifiés par des numéros arbitraires qui résultent le plus souvent d'une simple incrémentation. Ce nommage en termes d'AT et de ST « oublie » le nombre de SNP qui distinguent les allèles mais présente l'avantage de la simplicité et de donner le même poids aux événements de mutation (introduisant un SNP) et de recombinaison (pouvant introduire plusieurs SNP).

Structure génétique de l'espèce *F. psychrophilum* : faible diversité nucléotidique, fort taux de recombinaison, et complexe clonaux

Dans notre première analyse par MLST de l'espèce *F. psychrophilum*, j'ai étudié le polymorphisme de 50 isolats représentatifs de 10 espèces de poissons hôtes et provenant de quatre continents (Nicolas *et al.*, 2008). L'étude a porté sur 11 locus et a permis de distinguer 33 génotypes (ST) grâce à l'identification de 136 sites polymorphes.

L'analyse des données a montré la grande cohésion de l'espèce *F. psychrophilum* caractérisée par une faible diversité génétique, la divergence nucléotidique moyenne entre paires de séquences n'étant que de 0,004/pb, et des taux de recombinaison apparemment très élevés. Le rôle majeur de la recombinaison comme force évolutive dans l'espèce se traduit notamment par un nombre d'homoplasies apparentes très élevé, la valeur de l'indice d'homoplasie mesurée dans cette étude atteignant 68%. Cet indice correspond ici à la fraction des changements nécessaires pour expliquer les séquences qui ne correspondent probablement pas à des mutations lorsque l'on fait l'hypothèse d'une même généalogie pour

tous les sites de polymorphisme. En suivant l'approche proposée par Feil *et al.* (2000), l'analyse des différences entre 11 paires de génotypes proches (*Single Locus Variants*, SLV) suggéra 2 changements de nucléotides par mutations pour 56 changements par recombinaison. Soit une rapport des contributions des processus de recombinaison et de mutation (r/m) de 26. Bien que cette estimation initiale soit évidemment associée à une grande incertitude étant donné le faible comptage au dénominateur, cette valeur de r/m a permis de placer *F. psychrophilum* parmi les espèces bactériennes extrêmement recombinogènes. Une analyse un peu ultérieure de nos données avec une méthodologie différente, a même proposé un taux r/m autour de 60 pour *F. psychrophilum* (Vos & Didelot, 2008). Il s'agissait de la valeur estimée pour r/m la plus élevée parmi la cinquantaine d'espèces incluses dans l'étude (les espèces pour lesquelles un jeu de données MLST substantiel était disponible dans les bases de données).

Notre analyse a aussi révélé une association très forte entre certains génotypes ou groupes de génotypes et certaines espèces de poissons hôtes. On définit classiquement les complexes clonaux comme les composantes connexe du graphe obtenu en reliant les ST qui ne diffèrent que par un ou deux AT (SLV ou DLV). Chacun de ces groupes est présumé refléter la diversification récente d'un clone (ancêtre commun) (Spratt *et al.*, 2004). Ainsi, environ 40% des souches appartenait à un même groupe de génotypes que nous avons initialement nommé complexe clonal CC1 mais pour lequel nous avons ensuite utilisé la terminologie moins ambiguë de CC-ST2 (et encore plus tard renommé CC-ST10). Ce complexe clonal est presque exclusivement retrouvé chez la truite arc-en-ciel (*Oncorhynchus mykiss*) qui est le principal salmonidé d'élevage en France. La figure 3.5 illustre les relations génétiques entre les 50 souches de cette première étude.

Enfin, la diversité substantielle des génotypes trouvés en Europe a fourni un argument fort contre l'hypothèse d'une origine nord-américaine récente de l'espèce bien que les premières identifications en dehors de ce continent ne datent que du milieu des années 80.

Parmi les 11 locus, nous avons proposé de n'utiliser que les 7 les plus polymorphes comme schéma MLST pour les analyses suivantes. Une base de données a été crée. Initialement hébergée à l'Institut Pasteur avec l'aide de Sylvain Brisse, elle l'est maintenant à l'Université d'Oxford avec l'aide de Keith Jolley (Jolley & Maiden, 2010).

Échantillonnage de *F. psychrophilum* dans différents contextes, vers une image complète de la diversité de l'espèce ?

Les analyses MLST de l'espèce *F. psychrophilum* ont fait à ce jour l'objet de sept publications. J'ai réalisé l'attribution des AT et des ST de toutes ces études et j'ai rendu les données publiées accessibles via la base `https://pubmlst.org/fpsychrophilum/`. À l'exception de la toute première étude, j'ai aussi pris en charge l'assemblage et le contrôle qualité des séquences. Le contenu de la base de données est représenté dans la figure 3.6 qui met en valeur les relations génétiques entre les souches, l'origine géographique et le poisson hôte. L'objet de la suite de cette sous-section est de revenir brièvement sur les résultats des cinq études postérieures à Nicolas *et al.* (2008) que j'ai co-signées.

Dans Fujiwara-Nagata *et al.* (2013), nous avons analysé la structure de population

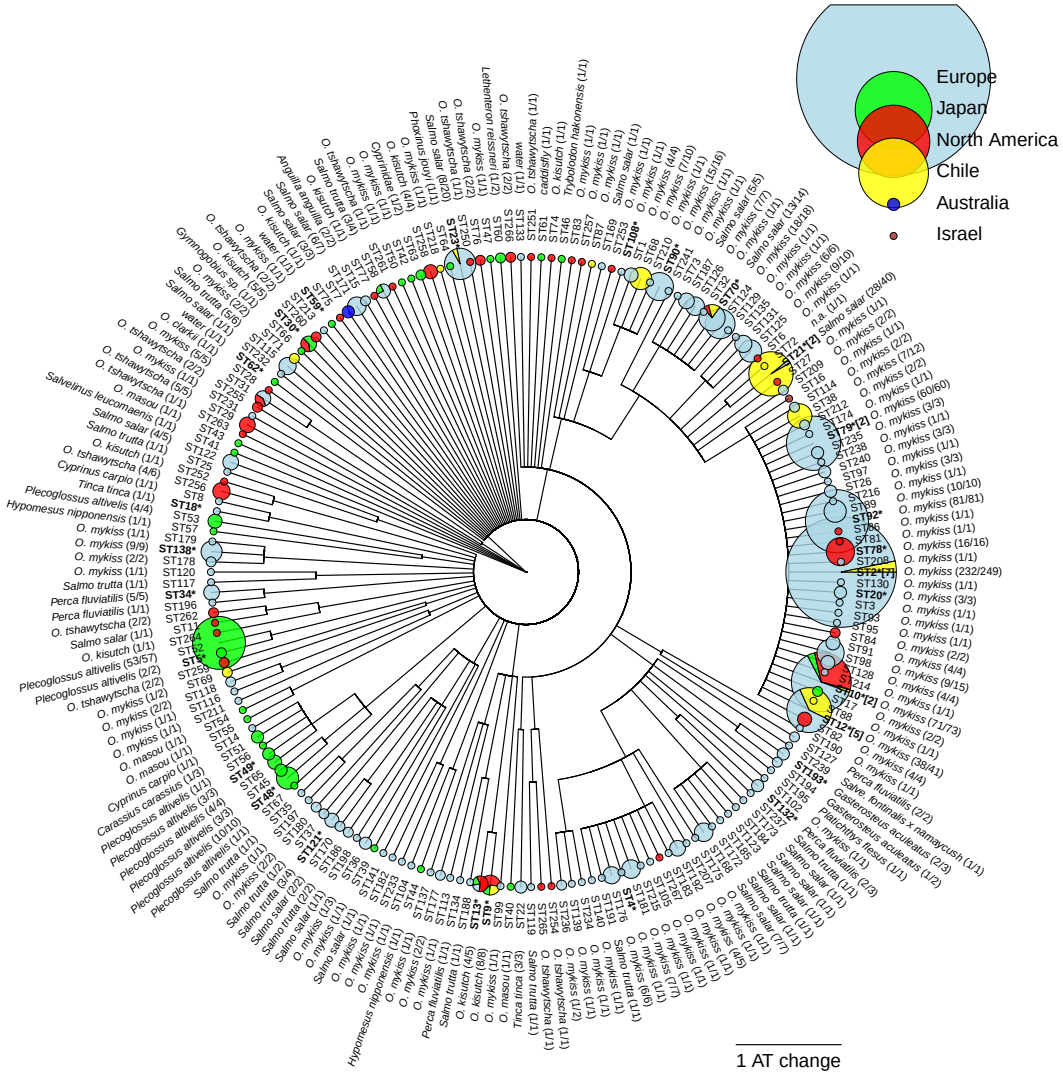


Figure 3.6 – Contenu actuel de la base de données MLST pour *F. psychrophilum* : 1097 isolats, 194 ST. On représente ici les relations entre les ST telles que résumées dans un arbre de *single-linkage clustering* en utilisant comme distance entre paires de ST le nombre d'AT qui les distinguent. La taille du cercle associé à chaque ST est proportionelle au nombre d'isolats et le découpage de ce cercle en cadrants reflète leurs provenances géographiques (ici les grandes régions du monde représentées dans l'échantillonnage). On indique aussi, à côté de chaque ST, le poisson hôte majoritaire et la fraction des isolats provenant de ce poisson (entre parenthèses). Les ST inclus dans notre analyse de génomes complets (Duchaud *et al.*, 2018) sont écrits en gras suivi d'une astérisque. Lorsque plusieurs génomes sont inclus dans l'étude, leur nombre est indiqué entre crochets. Extraction des génotypes et des origines des souches réalisée depuis https://pubmlst.org/fpsychrophilum/.

de *F. psychrophilum* au Japon, à travers une collection d'isolats représentatifs de la diversité dans le pays ainsi que d'isolats collectés dans une même rivière dont certains sur des poissons ne présentant pas de signes cliniques (15 poissons hôtes représentés).

Cette étude identifie notamment deux complexes clonaux infectant l'ayu (*Plecoglossus altivelis*), un poisson salmoniforme important au Japon. Nous avons aussi montré que les ST étaient distincts de ceux trouvés ailleurs dans le monde, à l'exception de ceux identifiés chez deux espèces de poissons d'origine nord-américaine : la truite arc-en-ciel (*O. mykiss*) et le saumon coho (*Oncorhynchus kisutch*). Par ailleurs, de même qu'en Europe, la diversité génétique des isolats échantillonnés au Japon contredit l'hypothèse d'une introduction récente de l'espèce dans ce pays. Une autre observation frappante de cette étude fut la similitude entre les niveaux de diversité génétique (telle qu'estimée à travers la divergence nucléotidique entre paires de génotypes) aux différentes échelles géographique : locale, nationale et mondiale. Cela suggère un brassage génétique important, au moins dans l'hémisphère nord, qui peut sembler plutôt surprenant pour une espèce bactérienne inféodée aux poissons d'eau douce.

L'analyse publiée dans Strepparava *et al.* (2013) concernait la diversité génétique de *F. psychrophilum* en Suisse dans les élevages de truites arc-en-ciel *O. mykiss* et de truite commune européenne (*Salmo trutta*). L'étude mit en évidence l'importance de deux complexes clonaux CC-ST10 et CC-ST90 (la question d'un apparentement possible est cependant ouverte, cf figure 3.6). De façon assez surprenante, ces deux complexes clonaux semblent infecter assez indifféremment les deux espèces de poissons (souvent élevées dans les mêmes fermes). L'étude réalisée au Chili (Avendaño-Herrera *et al.*, 2014) conduisit à une observation analogue de co-circulation des génotypes dans différentes espèces de salmonidés cultivés dans les mêmes réseaux de fermes. De façon intéressante, au contraire du Japon, la grande majorité des génotypes trouvés au Chili sont étroitement apparentés à des génotypes trouvés en Amérique du Nord ou en Europe. De fait, le commerce international des œufs de poisson (78 millions d'œufs importés au Chili pour l'année 2012) semble avoir joué un rôle important dans l'introduction de ces lignées au Chili. Le plan d'échantillonnage n'a pas permis de déterminer si l'espèce *F. psychrophilum* est trouvée au Chili en dehors de ces salmonidés et quelle serait la diversité de ces isolats « naturels », ce qui pourrait pourtant permettre de savoir si l'aire originelle de distribution de *F. psychrophilum* s'étend à l'hémisphère sud. Cette question rejoint en partie celle du spectre d'hôte, puisque l'aire originelle de répartition des salmonidés est limitée à l'hémisphère nord. Les premières introductions de ces poissons au Chili n'ont eu lieu qu'à la fin du 19ième siècle.

Je n'ai eu qu'une implication assez marginale dans l'étude publiée dans Nilsen *et al.* (2014) qui représente 560 isolats issus d'Europe du Nord (Danemark, Finlande, Suède et Norvège) et contribue largement à la sur-représentation de l'Europe dans l'ensemble des données publiées (figure 3.6). Cette étude souligne encore une fois l'importance du CC-ST10 dans les infections de truites arc-en-ciel et met aussi en évidence la différence entre les souches épidémiques qui circulent dans les élevages et les souches trouvées dans la nature. Le point qui m'a sans doute le plus surpris et l'absence de CC-ST90 pourtant relativement fréquent dans les pays d'Europe plus méridionale telle que la France et la Suisse. Enfin, Van Vliet *et al.* (2016) comble partiellement le manque d'information sur la diversité de *F. psychrophilum* en Amérique du nord. Une attention particulière y est portée aux Grands Lacs dans lesquels de nombreux salmonidés du pacifique (genre *Oncorhynchus*) ont été

introduits au cours du 20ième siècle.

Globalement, ces études ont mis en lumière l'importance des activités humaines (échanges commerciaux, introductions de poissons) dans la dissémination des complexes clonaux épidémiques. Les raisons des associations entre complexes clonaux et poissons hôtes restent à déterminer avec vraisemblablement une part d'adaptation du pathogène au poisson, une part d'adaptation aux conditions d'élevage (liberté avec intervention d'écloseries vs. ferme aquacoles) et une part historique liée aux routes de dissémination. L'aire de répartition géographique et le spectre d'hôte de l'espèce restent en partie obscurs. Un des bénéfices de ces études MLST a été de collecter et documenter les origines de nombreux isolats dont il devient possible d'étudier les génomes grâce à la baisse drastique des coûts de séquençage. Nous avons récemment publié une étude dans cette direction (Duchaud *et al.*, 2018). J'en aborderai les résultats dans la section 3.2.3.

3.2.2 Un « détour » par le genre *Tenacibaculum*

Dans le cadre du projet EMIDA ERA-NET PathoFish et de l'encadrement avec Éric Duchaud du travail de Christophe Habib, je me suis aussi intéressé à un autre groupe de bactéries pathogènes des poissons : le genre *Tenacibaculum*. Ce genre est phylogénétiquement proche de *Flavobacterium* mais alors que l'on trouve *Flavobacterium* dans les écosystèmes d'eau douce, *Tenacibaculum* est trouvé dans les milieux marins. Les deux genres contiennent des espèces pathogènes et une grande diversité de bactéries (espèces ?) non-pathogènes dites « environnementales ». La littérature contient de nombreuses et intéressantes réflexions sur le concept d'« espèce bactérienne », dont celles développées par Frederick M. Cohan qui me semblent particulièrement approfondies (Cohan & Perry, 2007).

L'étude que nous avons conduite (Habib *et al.*, 2014) portait sur 114 isolats représentatifs, d'une part, de la diversité du genre, et d'autre part, de la diversité de l'espèce pathogène des poissons *Tenacibaculum maritimum*. Je n'entrerai pas ici dans les détails mais les résultats ont mis en lumière des émergences probablement indépendantes du mode de vie pathogène dans plusieurs sous-clades du genre *Tenacibaculum* ainsi qu'une absence de complexes clonaux et d'association marquée entre poissons hôtes et génotypes chez *T. maritimum*. À titre d'illustration, les génotypes collectés pour *T. maritimum* sont présentées dans la figure 3.7. Cette espèce montre une diversité nucléotidique similaire à *F. psychrophilum* mais les recombinaisons y ont moins d'impact. Le rapport r/m tel que nous l'avons estimé avec ClonalFrame (Didelot & Falush, 2007) (la méthode employée par Vos & Didelot (2008) ayant produit des valeurs très élevées pour *F. psychrophilum*) n'est pour *T. maritimum* que de 2,3.

Notons aussi que, contrairement aux résultats obtenus sur *F. psychrophilum*, les données n'ont pas révélé de traces de dissémination à longue distance de complexes clonaux qui pourraient être liées à des activités humaines. De fait, le grand nombre de génotypes distincts suggère une distribution endémique des souches causant les épisodes infectieux dans les élevages. Cette idée semble parfaitement cohérente avec l'observation selon laquelle ces épisodes surviennent lorsque les conditions de santé des poissons sont dégradées et pourraient donc faire intervenir de nouvelles contaminations provenant de

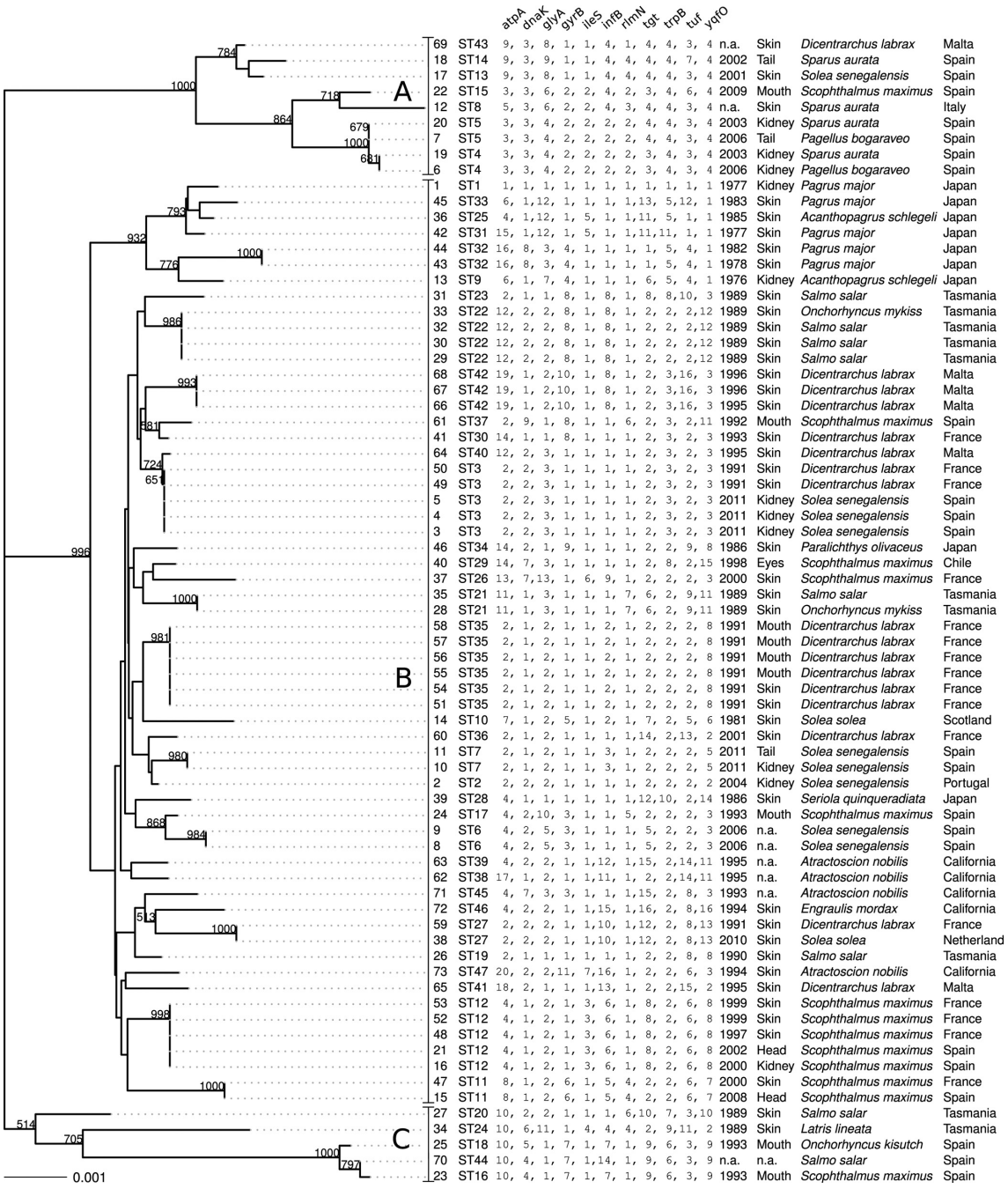


Figure 3.7 – Analyse MLST de *T. maritimum*. Génotypes et origines de 73 isolats. On représente ici de gauche à droite : une tentative d'arbre phylogénétique, l'identifiant de chaque isolat, son génotype (ST), son profil allélique (AT) et des informations sur son origine (année, tissu, espèce de poisson, pays). L'arbre résulte d'une reconstruction par *neighbor joining* à partir d'une simple distance nucléotidique de Jukes-Cantor entre paires d'isolats. La valeur de support *bootstrap* (pour 1000 réplicats) est indiquée sur chaque branche. Trois sous-groupes d'isolats (A, B, et C) sont délimités par des barres verticales. Figure reproduite de Habib *et al.* (2014).

l'environnement local.

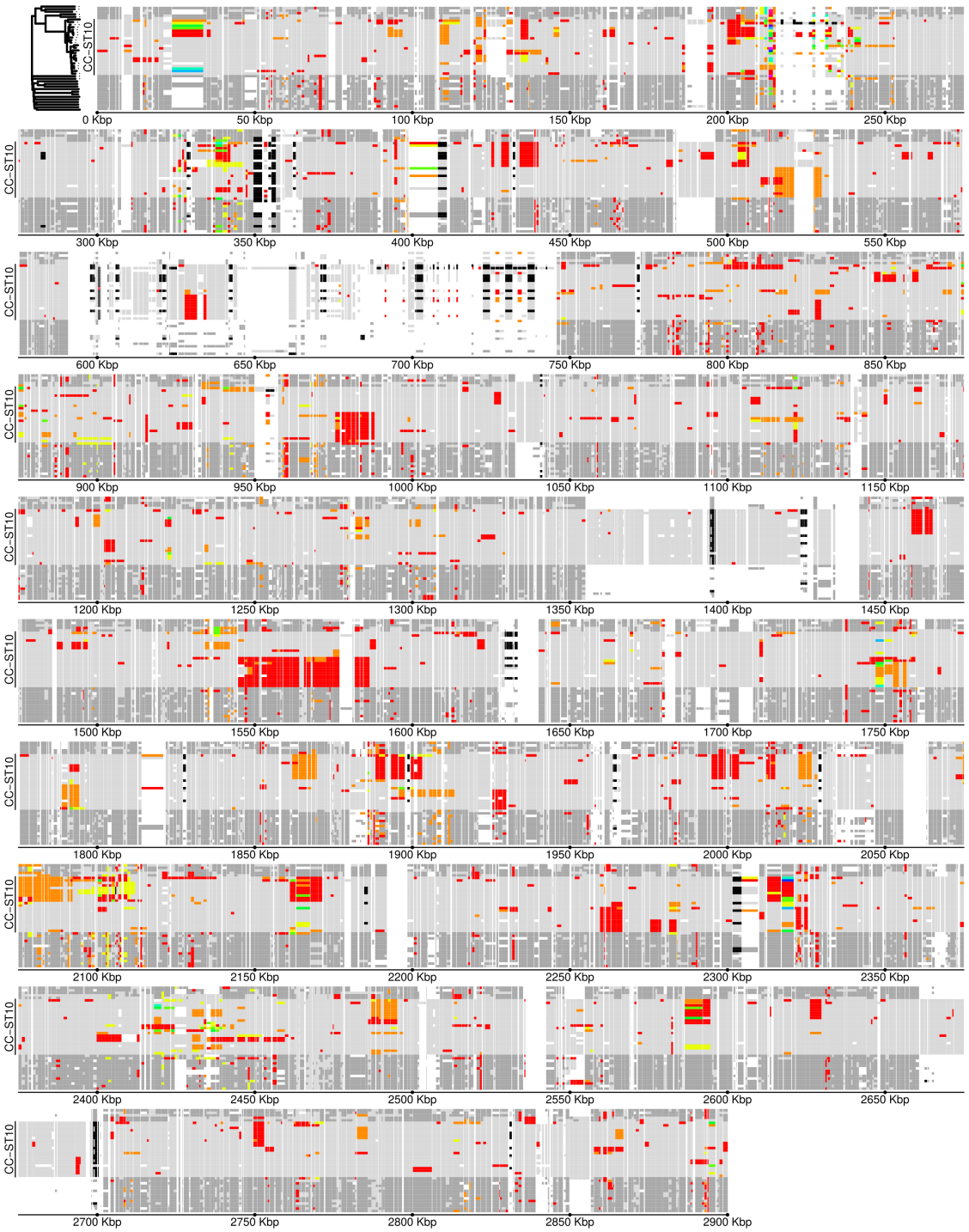


Figure 3.8 – Cartographie des allèles dans les génomes de *F. psychrophilum*. Les coordonnées utilisées pour les locus correspondent aux CDS dans la souche CSF259-93. Les génomes sont ordonnés verticalement selon l'arbre illustré en haut à gauche. Les allèles (AT) sont représentés par différentes couleurs : gris clair pour l'allèle majoritaire dans CC-ST10, rouge à jaune pour les suivants (avec des ré-allocations locales des couleurs pour avoir une consistance entre locus adjacents) ; gris foncé pour les allèles absents de CC-ST10 ; noir pour les gènes en plusieurs copies ; blanc lorsque le gène est absent. Figure reproduite de Duchaud *et al.* (2018).

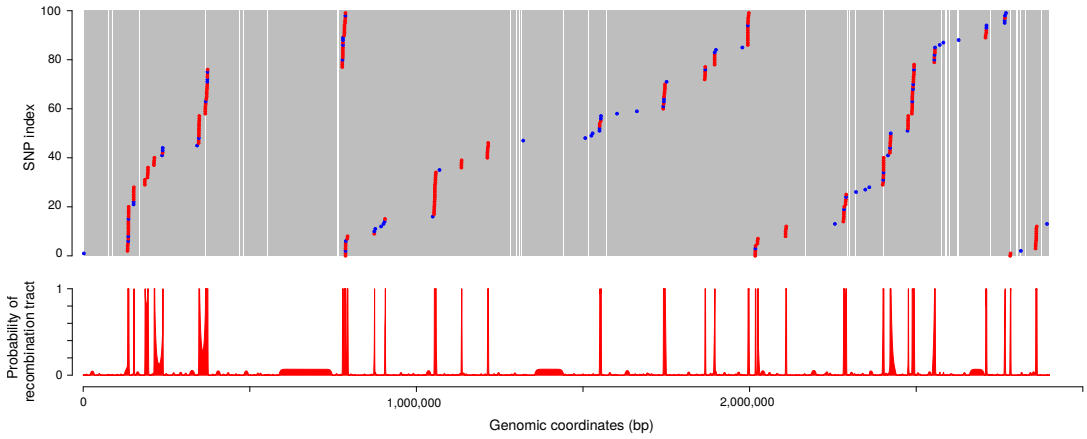


Figure 3.9 – Délimitation des régions recombinées et non recombinées dans une paire de génomes proches de l'espèce *F. psychrophilum*. Les génomes comparés ici sont ceux des souches JIP 02/86 et CH1895. La partie supérieure de la figure montre la position des SNP distinguant les deux génotypes. Chaque point correspond à un SNP dont la couleur reflète le type : bleu pour le polymorphisme trouvé seulement au sein de CC-ST10 (sites de type P1 dans le texte) et rouge pour les autres. L'ordonnée est introduite pour les besoins de la visualisation, elle représente le numéro du SNP réinitialisé tous les 100 SNP. Les zones en gris correspondent à des régions non-couvertes par les alignements étudiés (les gènes conservés à l'échelle de l'espèce et présents en une seule copie). La partie inférieure représente la probabilité pour chaque site d'appartenir à une région recombinée telle que calculée avec le HMM. Figure reproduite de Duchaud *et al.*, 2018.

3.2.3 Génomique comparative de *F. psychrophilum*

Dans le prolongement des analyses MLST sur *F. psychrophilum* (section 3.2.1) et toujours dans le cadre du projet EMIDA ERA-NET PathoFish, nous avons analysé dans les génomes de 41 isolats de *F. psychrophilum*, dont 30 séquencés pour l'étude (Duchaud *et al.*, 2018). La distribution de ces isolats parmi l'ensemble de ceux génotypés par MLST est représentée dans la figure 3.6. Les séquences des génomes complets ont largement confirmé les premières observations faites grâce à la MLST, notamment en ce qui concerne la relativement faible diversité nucléotidique et l'impact très important des recombinaisons. Elles ont aussi permis l'étude des répertoires de gènes et conduit à quelques observations intéressantes sur les spécificités du CC-ST90 dont l'identification de gènes qui pourraient être impliqués dans des mécanismes de pathogénicité et dans des résistances à des antibiotiques utilisés dans les piscicultures. Une identification des gènes qui déterminent les sérotypes a aussi été possible à partir de ces génomes (Rochat *et al.*, 2017b). Les points que je souhaite développer succinctement ici concernent l'analyse des recombinaisons et la tentative de datation de l'émergence du CC-ST10.

La figure 3.8 illustre la distribution des allèles (AT) du CC-ST10 dans les 41 génomes comparés. On y voit une corrélation entre locus adjacents qui correspond aux traces d'événements de recombinaison s'étendant sur des régions plus longues que les gènes individuels. Un HMM a été utilisé afin de déterminer de façon systématique et consistante

la contribution des événements de recombinaison à la divergence. L'idée était de formaliser dans un cadre statistique la méthode de Feil *et al.* (2000) que nous avions déjà utilisée dans Nicolas *et al.* (2008) pour calculer r/m à partir des profils de SNP entre paires de génotypes proches (SLV). Lors de l'examen d'une paire de génomes proches (dans un même complexe clonal) on distingue ainsi deux types de SNP selon qu'il s'agisse d'un polymorphisme privé au complexe clonal considéré (que l'on nommera de type P1) ou d'un polymorphisme partagé avec d'autres isolats plus lointains (type P2). Les régions recombinées correspondent en principe à des zones de concentration élevée en sites P2 et moindre en sites P1, séparées par des régions non recombinées dans lesquelles on attend une fréquence pratiquement nulle de sites P2. Le fréquence (faible) des sites P1 dans les régions non recombinées résulte de l'accumulation de mutations depuis la date de divergence des deux génomes. Le HMM que nous avons considéré possède deux états cachés (régions recombinées *vs.* non-recombinées). Les procédures d'estimation des paramètres et de reconstruction du chemin caché que nous avons utilisées sont identiques à celles décrites dans Nicolas *et al.* (2002) et présentées dans la section 2.1.1. La délimitation des régions recombinées par étude d'une paire de génomes grâce à ce HMM est illustrée dans la figure 3.9. L'estimation des différents paramètres permet de construire un estimateur de r/m ainsi que de la distance « mutationnelle » entre les souches.

Les résultats indiquent un rapport r/m médian de ≈13 avec des différences substantielles entre certains groupes d'isolats au sein de CC-ST10. Ce taux reste très élevé. Il est aussi compatible avec les nombres (52/2) issus de la première analyse MLST sur 11 locus. En revanche, il est bien en deçà de l'intervalle obtenu par Vos & Didelot (2008) à partir des 7 locus retenus dans le schéma MLST et avec une méthode différente. Simultanément, la longueur moyenne des régions recombinées a été estimée à ≈4.0 kpb.

Les distances « mutationnelles » obtenues à partir des fréquences estimées de sites P1 au sein des régions non-recombinées dans chaque comparaison ont permis de proposer la reconstitution d'un arbre phylogénétique pour le CC-ST10. L'analyse des distances entre les différentes feuilles et la racine de cet arbre a révélé une corrélation significative avec la date d'échantillonage des isolats. Cette observation nous a permis de proposer une calibration de l'horloge moléculaire autour de 2.8e-7 substitutions par pb et par an. C'est-à-dire une valeur plus faible que celles autour de 1e-6 substitutions par pb et par an obtenues pour plusieurs pathogènes humains (Mutreja *et al.*, 2011; Croucher *et al.*, 2011; Hsu *et al.*, 2015) mais qui ne semble pas incohérente avec la durée relativement élevée des générations de *F. psychrophilum* qui vit typiquement à des températures inférieures à 20 °C. Selon cette horloge moléculaire, le début de diversification de CC-ST10 remonterait à la deuxième moitié du 19ième siècle et serait donc concomitant au développement de l'élevage des truites arc-en-ciel, d'abord en Amérique du Nord puis dans le reste du monde.

Chapitre 4

Projets

En matière de positionnement général, je continuerai à travailler à l'interface entre statistique, informatique et biologie en essayant de préserver un équilibre entre ces diverses composantes.

Rétrospectivement, mes travaux de recherche me semblent avoir souvent pris de des directions différentes de celles que j'avais initialement envisagées. Les raisons qui ont concouru à ces changements sont nombreuses : manque de temps, résultats inattendus, nouvelles idées, évolutions technologiques déplaçant les questions de recherche. À cela s'ajoutent les aléas concernant l'obtention de financements qui peuvent détourner d'un projet pourtant mûrement réfléchi par manque de moyens ou par opportunisme (dans le bon sens du terme).

J'aborde donc cette section avec une certaine réserve et je vais me limiter ici à évoquer des projets à relativement court terme qui devraient constituer la majeure partie de mon activité de recherche dans les prochaines années.

4.1 Modèles et algorithmes

Je mentionne dans cette section deux thèmes plutôt méthodologiques de mes recherches en cours sur lesquels je compte continuer à m'investir.

4.1.1 Vers une nouvelle approche pour la recherche de motifs régulateurs

Depuis fin 2015, j'encadre la thèse d'Ibrahim Sultan financée par le projet EU ITN List_Maps. L'objectif de cette thèse est de contribuer à la connaissance du réseau de régulation de la bactérie à Gram-positif *Listeria monocytogenes* qui, comme *S. aureus*, est une cousine pathogène de *B. subtilis*. L'objectif général est de prolonger le travail de modélisation pour la classification des promoteurs (section 2.2.2) pour permettre la recherche des sites de fixation des autres facteurs de transcription. Comme pour les facteurs sigma, l'idée est de s'appuyer sur une connaissance précise des positions des TSS (idéalement à la pb près, telle qu'obtenue par séquençage des extrémités 5' des ARN) et sur des données, aussi propres que possible, concernant les profils d'expression à travers

les conditions physiologiques.

Par rapport aux sites des facteurs sigma, les principales différences qui viennent compliquer la tâche de découverte des sites des autres facteurs de transcription sont :

- La multiplicité des sites de facteurs de transcription qui peuvent se trouver sur un même promoteur et qui peuvent même se chevaucher ;
- La plus grande « dilution » des sites du fait de
  - leur éparpillement sur de plus grandes régions des promoteurs (au contraire des sites des facteurs sigma dont le positionnement par rapport au TSS est très strict),
  - leur faible nombre d'occurrences qui n'est pas compensé par une non-indépendance très forte comme pour les facteurs sigma (chaque TSS contient exactement une occurrence de motif reconnu par un facteur sigma) ;
- Un impact *a priori* plus subtil sur la modulation du niveau d'expression des gènes. Les facteurs sigma, par leur mode d'action (nécessairement activateur) et par le partitionnement qu'ils induisent de l'espace des promoteurs ont un effet de premier ordre que l'on peut plus raisonnablement espérer capturer dans la structure hiérarchique d'un arbre résumant les coefficients de corrélation.

Comme pour la recherche des sites de fixation des facteurs sigma, l'approche retenue dans la thèse d'Ibrahim Sultan s'inscrit dans la lignée des programmes traditionnels fondés sur une pure modélisation de la séquence avec des matrices poids-position (PWM). Les données d'expression sont alors utilisées comme des covariables qui peuvent donner des indications sur quels promoteurs contiennent ou ne contiennent pas un motif. La connaissance des TSS permet éventuellement de restreindre la zone de recherche en cas de positionnement préférentiel des motifs. Le modèle s'affranchit ainsi de la modélisation des données d'expression ou l'utilisation d'approches discriminantes (Liu *et al.*, 2017). Le bénéfice attendu est d'obtenir une transition douce entre recherche guidée par des données d'expression et recherche « à l'aveugle » sur la seule base des propriétés statistiques des séquences.

Une attention particulière est aussi portée à la prise en compte des chevauchements entre motifs, non seulement car ceux-ci ont une existence biologique (Hermsen *et al.*, 2006), mais aussi afin de rendre beaucoup plus « lisse » l'espace de configuration des positions lorsque l'on cherche simultanément plusieurs motifs. La modélisation des chevauchement peut aussi être vue comme une alternative statistiquement fondée à la recherche séquentielle, en masquant les motifs trouvés les uns après les autres, ou parallèle, en mettant en place des mécanismes de répulsion mutuelle entre MCMC (Ikebata & Yoshida, 2015).

La méthode en cours de développement permet de prendre en compte une collection de plusieurs dizaines de covariables construites pour résumer les informations disponibles (appartenance à un groupe, positionnement sur un axe ou dans un arbre) et de rechercher simultanément de nombreux motifs. Chaque motif a sa propre relation de dépendance aux covariables. L'arsenal des méthodes MCMC trans-dimensionnelles est mis en œuvre pour l'ajustement automatique de la largeur du motif, de l'ordre du modèle de Markov décrivant les séquences en dehors du motif (*background*), du nombre de paramètres nécessaires pour

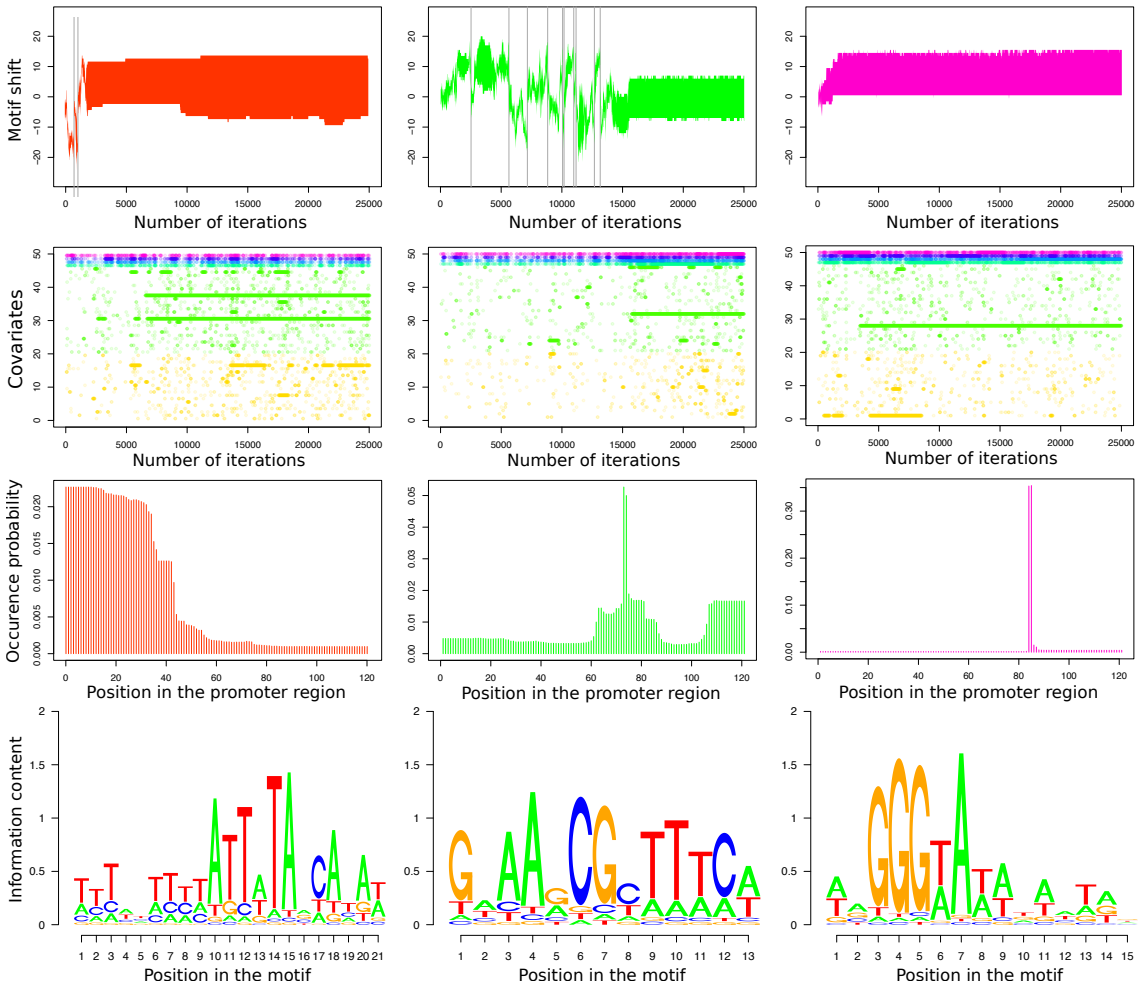


Figure 4.1 – Aperçu du comportement de l'algorithme de découverte de motifs en cours de développement. Le jeu de données consiste en 1 512 séquences promotrices × 165 transcriptomes pour *L. monocytogenes*, il est issu de la base de données Listeriomics (Bécavin *et al.*, 2017). Les informations sur trois motifs différents sont montrées dans les colonnes de gauche, du centre et de droite. La première ligne contient des graphes de convergence pour la largeur du motif au cours de 25 000 balayages (ou itérations) de l'algorithme MCMC. Les barres verticales indiquent un recentrage pour les besoins de la représentation. La deuxième ligne représente le recrutement des 50 covariables (résumant les 165 transcriptomes), positionnées sur l'axe des y. Les couleurs distinguent différents types de covariables : jaune pour les axes d'ACP, vers pour les axes d'ICA, bleu et violet pour les arbres de classification hiérarchique. La troisième ligne représente la fonction de densité décrivant la distribution de la position du motif dans la région promotrice (moyenne sur les derniers 10 000 balayages), le TSS correspond ici à la position 100. La quatrième et dernière ligne contient les logos représentant le contenu des matrices poids-position (moyenne sur les derniers 10 000 balayages). La hauteur totale des lettres à une position dans le motif reflète le contenu en information. Figure préparée par Ibrahim Sultan.

décrire la distribution de la position du motif vis à vis du TSS, et enfin, du nombre de covariables pertinentes pour décrire la distribution des occurrences au sein du répertoire de TSS. Des résultats préliminaires destinés à illustrer le comportement de l'algorithme sont présentés dans la figure 4.1.

4.1.2 Modèle de recombinaison bactérienne dont la fréquence dépend de la distance phylogénétique

J'ai été confronté plusieurs fois à l'analyse d'alignements de séquences bactériennes contenant d'importantes traces de recombinaisons, à l'échelle d'une espèce (Nicolas *et al.*, 2008 ; Dalmasso *et al.*, 2011 ; Duchaud *et al.*, 2018) ou d'un genre (Nicolas *et al.*, 2007b ; Habib *et al.*, 2014). Les questions de la quantification des recombinaisons et de la modélisation de leur impact sur les généalogies m'intéressent depuis longtemps. J'ai notamment implémenté des algorithmes de simulation pour aborder cette question dans le cadre des modèles coalescents, mais sans avoir pour l'instant publié sur ce sujet.

Je souhaiterai réussir à mener à bien un travail proposant un cadre méthodologique pour étudier les liens entre distance phylogénétique et fréquence de recombinaison (transferts génétiques horizontaux qui conduisent à des remplacements alléliques). En effet, chez les bactéries, la fréquence des transferts horizontaux qui conduisent à des remplacements alléliques tend probablement à décroître avec la distance génétique qui sépare le receveur du donneur. Il y a au moins deux raisons à cela : (i) l'efficacité des mécanismes de recombinaison homologue ou apparentée dépend directement du degré de similitude entre les séquences, (ii) le remplacement de tout ou partie d'un gène par un homologue distant à plus de chance d'être des effets délétères (Majewski, 2001).

En 2016, j'ai encadré le stage de M2 d'Arthur Frouin dont le but était de finir l'implémentation d'un algorithme de simulation reposant sur des coalescents emboîtés pour réaliser des simulations d'un modèle d'évolution qui rend compte simultanément de la généalogie des locus, des organismes, et des espèces. Un paramètre de ce modèle gère la décroissance de la fréquence des recombinaisons inter-spécifiques avec la distance génétique séparant les espèces. Le travail consiste maintenant à construire des statistiques résumées capable de capturer l'influence de ce paramètre et à étudier le comportement de ces statistiques sur des données simulées et réelles. Une des difficultés consistera à trouver des jeux de données adéquats en ce qui concerne la densité d'échantillonnage et les distances phylogénétiques. Il faudra aussi se positionner précisément vis à vis des travaux déjà publiés tels que ceux de Ansari & Didelot (2014) qui ont proposé une méthode d'inférence sans formuler un cadre propre pour la simulation qui permette de prendre en compte les espèces éteintes et de distinguer les niveaux intra et inter-espèces. Les conséquences de la superposition de ces deux niveaux sont bien illustrées par le travail de Akita *et al.* (2018).

Une des difficultés viendra du fait que mon algorithme ne simule que des paires de sites conditionnellement à une généalogie d'organismes échantillonnés et à l'arbre de l'ensemble des espèces vivantes et éteintes. Ce choix très simplificateur de ne pas simuler un chromosome entier se justifie car, proprement conditionnée par la généalogie

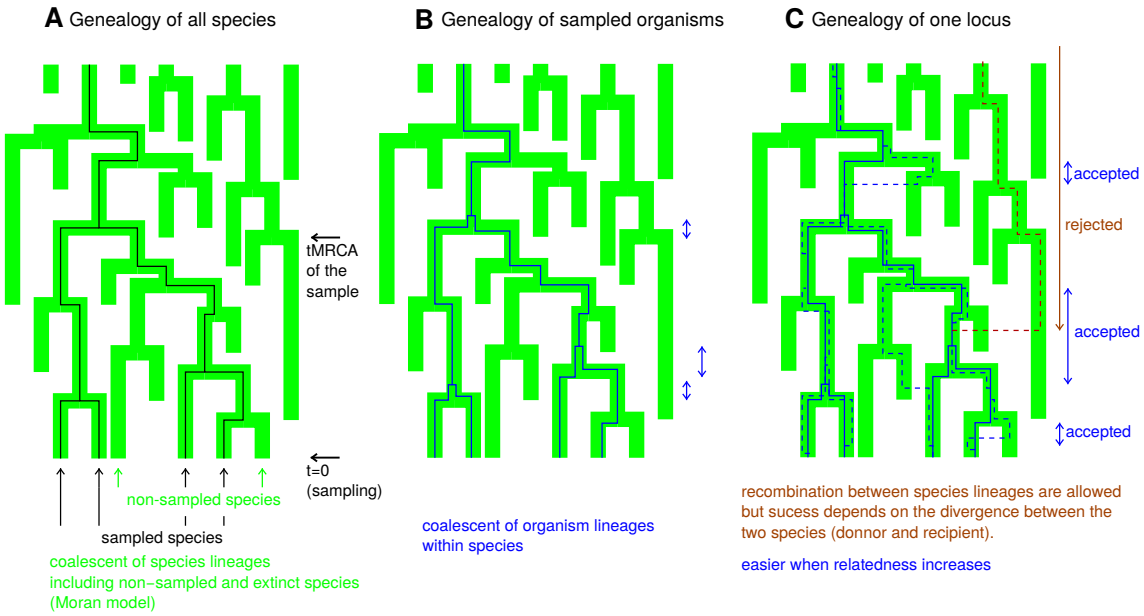


Figure 4.2 – Principe d'un modèle coalescent de recombinaison bactérienne avec fréquence dépendante de la distance évolutive.

des organismes, la simulation de sites individuels et de paires de sites suffit pour l'étude du comportement moyen de nombreuses statistiques résumées. J'envisage de faire moi-même ce travail, ou éventuellement de le proposer comme sujet à un étudiant de master.

4.2 Vers de nouveaux types de données

Une partie importante de mon travail ces dernières années a consister à analyser et valoriser des jeux de données. Dans ce contexte, je suis convaincu de l'intérêt de participer à l'acquisition des données. L'enjeu essentiel est d'élaborer des plans d'expériences qui satisfassent aussi bien les « expérimentateurs » que les « modélisateurs ».

Je mentionne donc pour finir ce mémoire deux directions de recherche reposant sur la collecte et l'analyse de données que j'aimerais (continuer à) développer dans les prochaines années. L'une est autour de l'analyse de données transcriptomiques, l'autre de l'utilisation de mini-bioréacteurs pour l'évolution expérimentale.

4.2.1 Transcriptomique, de l'organisme à l'écosystème

Transcriptomique mono-espèce

Je suis impliqué dans plusieurs projets qui s'inscrivent dans le thème de la transcriptomique mono-espèce en prolongement direct des analyses de grand jeux de données sur *B. subtilis* et *S. aureus* décrites dans la section 3.1. Les deux plus ambitieux sont la caractérisation du transcriptome de la bactérie pathogène des poissons *F. psychrophilum* et celui de la bactérie *L. monocytogenes*.

Le projet sur *F. psychrophilum* m'intéresse particulièrement car il est à la convergence

des travaux présentés dans les sections 3.1 et 3.2. Cette convergence a été facilitée par le recrutement en 2012 de Tatiana Rochat, avec qui j'avais déjà travaillé sur la transcriptomique de *B. subtilis* (section 3.1), en tant que Chargée de Recherche dans l'équipe d'Éric Duchaud et Jean-François Bernardet, avec lesquels je travaille sur *F. psychrophilum* (section 3.2).

Nous avons donc conçu le projet de recourir de façon intensive à la transcriptomique pour tenter de faire un pas important dans la connaissance de la bactérie *F. psychrophilum* sur laquelle les connaissances scientifiques sont incomparablement moins avancées que sur *B. subtilis* ou *S. aureus*. Au delà de l'importance propre de *F. psychrophilum*, je vois un intérêt général à ce travail. En effet, il servira à tester l'idée que je trouve très séduisante selon laquelle l'analyse transcriptomique poussée d'une bactérie très peu connue pourrait nous apprendre beaucoup sur sa biologie, et cela pour un coût raisonnable et sans recourir à la démarche fastidieuse et aléatoire de construction de mutants. Ce dernier point est important puisqu'en pratique la plupart des bactéries restent impossibles ou très difficiles à manipuler génétiquement. De plus, la construction de mutants suppose de cibler des gènes ce qui ne peut être fait de façon pertinente sans connaissances préalables. Dans le cas de *F. psychrophilum*, nous nous intéresserons en particulier à l'identification des mécanismes moléculaires et des voies de régulation impliquées dans la virulence et la survie.

Pour minimiser les coûts, une phase de détection des régions exprimées et des TSS à été réalisée avec la technologie RNA-Seq (globale pour les régions, ciblant les extrémités 5' des ARN pour les TSS) sur des échantillons « synthétiques » mélangeant artificiellement des conditions différentes. L'algorithme développé dans la thèse de Bogdan Mirauta a été utilisé dans ce cadre pour délimiter les régions exprimées (Nicolas *et al.*, 2014). Ensuite, le transcriptome a été réalisé dans des conditions expérimentales visant à une couverture maximale des modes de vie de la bactérie en utilisant des puces faites « à façon » pour quantifier non seulement l'expression des gènes et des nouvelles régions, mais aussi l'activité de chaque TSS identifié. Les données sont en cours d'analyse. L'utilisation de l'algorithme de partitionnement des promoteurs (section 2.2.2) a permis d'identifier les régulons de plusieurs facteur sigma et le travail en cours d'Ibrahim Sultan pourra être utiliser directement sur ce jeu de données. Par ailleurs, l'obtention récente d'un financement ANR JCJC par Tatiana Rochat (ANR FlavoPatho 2017-2021 auquel je suis associé) nous permettra de poursuivre confortablement la collecte de données (RNA-Seq) afin de mieux explorer la phase de survie dans l'eau et d'infection du poisson.

Le travaux envisagés sur *L. monocytogenes* font partie intégrante du projet européen ITN List_Maps dans le cadre duquel s'effectue la thèse d'Ibrahim Sultan. Il s'agit de s'appuyer sur le réseaux des 10 étudiants en thèse de List_Maps pour construire collectivement un jeu de données aussi représentatif que possible des conditions de vie de la bactérie. Une attention particulière sera portée aux conditions de vie dans l'environnement qui restent les moins étudiées chez cette bactérie qui fait aujourd'hui figure de modèle pour l'étude des mécanismes de virulence et pour laquelle de nombreuses données transcriptomiques sont déjà disponibles (Toledo-Arana *et al.*, 2009; Bécavin *et al.*, 2017). Les données transcriptomiques du projet List_Maps seront produites en 2018. En ce qui concerne l'analyse, une première étape consistera à mettre en œuvre

l'algorithme de recherche de sites de motifs développé dans la thèse d'Ibrahim Sultan. Il sera ensuite probablement pertinent de prolonger les collaborations avec les membres du projet List_Maps, idéalement en trouvant de nouvelles sources de financement.

Je clôturerai cette partie dédiée à la transcriptomique mono-espèce en mentionnant la poursuite des travaux sur le facteur de terminaison de la transcription Rho décrits dans la section 3.1.4. Un financement a été obtenu auprès du département MICA de l'INRA par Elena Bidnenko pour l'année 2018. Il va nous permettre de nouvelles expériences de transcriptomiques chez *B. subtilis*. Une proposition de projet ANR autour de Rho coordonné par Elena a aussi été retenue pour la seconde phase d'évaluation cette année.

Méta-omiques, transcriptomique multi-espèce

Je suis pour l'instant resté à l'écart de l'effervescence autour l'analyse de données méta-génomiques pour l'étude des écosystèmes microbiens. La raison en est le manque de temps plus que le manque d'intérêt. Je serai par exemple très intéressé par contribuer à des analyses utilisant les données méta-génomiques de type *shotgun* (par opposition à la méta-génomique amplicon) sous l'angle de l'évolution et de la micro-évolution en adaptant des démarches et approches de génétique/génomique des populations. L'étude des divergences entre organismes au sein d'un même écosystème devrait permettre d'observer des processus évolutifs simultanément sur plusieurs espèces. De plus, les échelles évolutives à l'œuvre dans la divergence entre lignées d'une espèce au sein d'un écosystème sont certainement beaucoup plus courtes que celles traditionnellement étudiées lorsque l'on isole puis séquence des souches. La littérature fournit déjà des exemples prometteurs de travaux visant à exploiter cette richesse des données méta-génomiques (Johnson & Slatkin, 2009; Truong *et al.*, 2017). J'envisage d'aborder ces questions à travers des collaborations au sein de l'équipe StatInfOmics en profitant notamment de l'arrivée récente d'Anne-Laure Abraham (déjà investie dans l'analyse de données méta-génomique *shotgun*) et du retour d'Hélène Chiapello (dont le projet de recherche est centré sur l'étude des processus évolutifs et la comparaison de génomes).

Un autre axe de recherche en lien avec les écosystèmes sur lequel je vais travailler au cours des prochaines années est l'analyse de données méta-transcriptomiques. Deux projets concrets sont déjà financés.

Le premier projet consiste à analyser un biofilm modèle de quatre espèces bactériennes (projet ANR ACTOP). Ce nombre d'espèces très limité et la nature hautement contrôlé des expériences menées dans des dispositifs de milli-fluidiques par l'équipe de Nelly Henry au Laboratoire Jean-Perrin (UPMC) devraient permettre l'obtention de jeux de données de bonne qualité. Les données transcriptomiques seront produites et analysées dans le cadre d'une collaboration avec MICALIS (Narimane Dahmane, post-doctorante recrutée sur le projet, partagera son temps entre MaIAGE et MICALIS). L'approche RNA-Seq sera utilisée pour suivre la réponse du biofilm à des stress et, ce qui m'intéresse plus particulièrement, pour essayer de comprendre les interactions entre les quatre espèces. Pour aborder cette dernière question, la nature synthétique du biofilm devrait nous permettre d'étudier le transcriptome des 15 combinaisons d'espèces ($2^4 - 1$).

Le second projet vise à caractériser la réponse d'un écosystème complexe à un stress

(projet MIcrobNO financé par le métaprogramme INRA MEM). Il s'agit ici de l'écosystème du tube digestif humain et d'un stress au monoxyde d'azote mimant une réponse des macrophages à une infection. En pratique, nous envisageons de maximiser le niveau de contrôle et de standardisation des expériences en réalisant des cultures *in vitro* à partir d'un unique assemblage de souches (échantillon de fèces). Le transcriptome sera caractérisé pour différentes concentrations de NO et après différents temps d'exposition. Mon objectif sera alors de réaliser une étude transversale de la réponse au stress NO en essayant de classer les réponses des organismes présents dans l'écosystème (notamment en fonction de la sensibilité au stress) et de trouver des points communs entre les gènes impliqués. La complexité de l'écosystème entraînera des couvertures faibles et hétérogènes des chacune des souches présentes et constituera l'une des difficultés de l'analyse mais la démarche me parait originale et intéressante. En effet, elle s'inscrit dans la droite ligne des expériences menées traditionnellement sur une seule espèce en y ajoutant une dimension multi-espèce. Cette dimension est pertinente, non seulement du point de vue de l'écosystème, mais aussi pour mettre en perspective les réponses des différents organismes à des conditions de stress similaires. En fait, ce point pourrait permettre de contourner une des limites des expériences mono-espèce que j'ai eu a analyser jusqu'ici, à savoir le contrôle souvent imparfait du niveau de stress appliqué ainsi que du stade de la réponse correspondant au temps de prélèvement. Le financement MicrobNO couvrira essentiellement le coût d'expériences pilotes. Les analyses bioinformatiques seront réalisées par les quelques membres de MaIAGE impliqués dans le projet. À terme, si les résultats préliminaires s'avèrent prometteurs, il sera possible de chercher d'autres sources de financements et de proposer des sujets de stages et/ou de thèses autour de l'analyse intégrative de la réponse à des stress à partir de données transcriptomiques multi-espèces.

4.2.2 Mini-bioréacteurs et évolution expérimentale

Je terminerai ce rapport en mentionnant un nouveau thème de recherche que j'aborde actuellement autour de la conduite de cultures en mini-bioréacteurs. L'objectif principal est de pouvoir mener des expériences d'évolution expérimentale. Ce thème m'intéresse car il est la croisée de nombreuses disciplines : évolution, génétique des populations, microbiologie, biologie moléculaire, biologie systémique, électronique, contrôle, bioinformatique, modélisation probabiliste et analyses statistiques. Même si l'idée d'étudier l'évolution dans des dispositifs de culture continue est aussi ancienne que l'invention du chemostat dans les années 50 (Novick & Szilard, 1950), elle est aujourd'hui ré-actualisée par la chute des coûts de séquençage qui laisse envisager d'étudier à peu de frais une grande variété de trajectoires évolutives (Gresham & Dunham, 2014). Enfin, ce thème est stratégique car il permet d'aborder des questions scientifiques très diverses, aussi bien fondamentales qu'appliquées (Toprak *et al.*, 2012; Gresham & Hong, 2015).

Les possibilités de contributions en tant que modélisateur/biostatisticien/bioinformaticien existent sur de nombreux aspects :

(i) le choix des composantes à faire évoluer qui peut typiquement faire intervenir de la génomique comparative ou des jeux de données -omiques ;

(ii) la calibration des expériences à travers l'estimation des paramètres tels que les

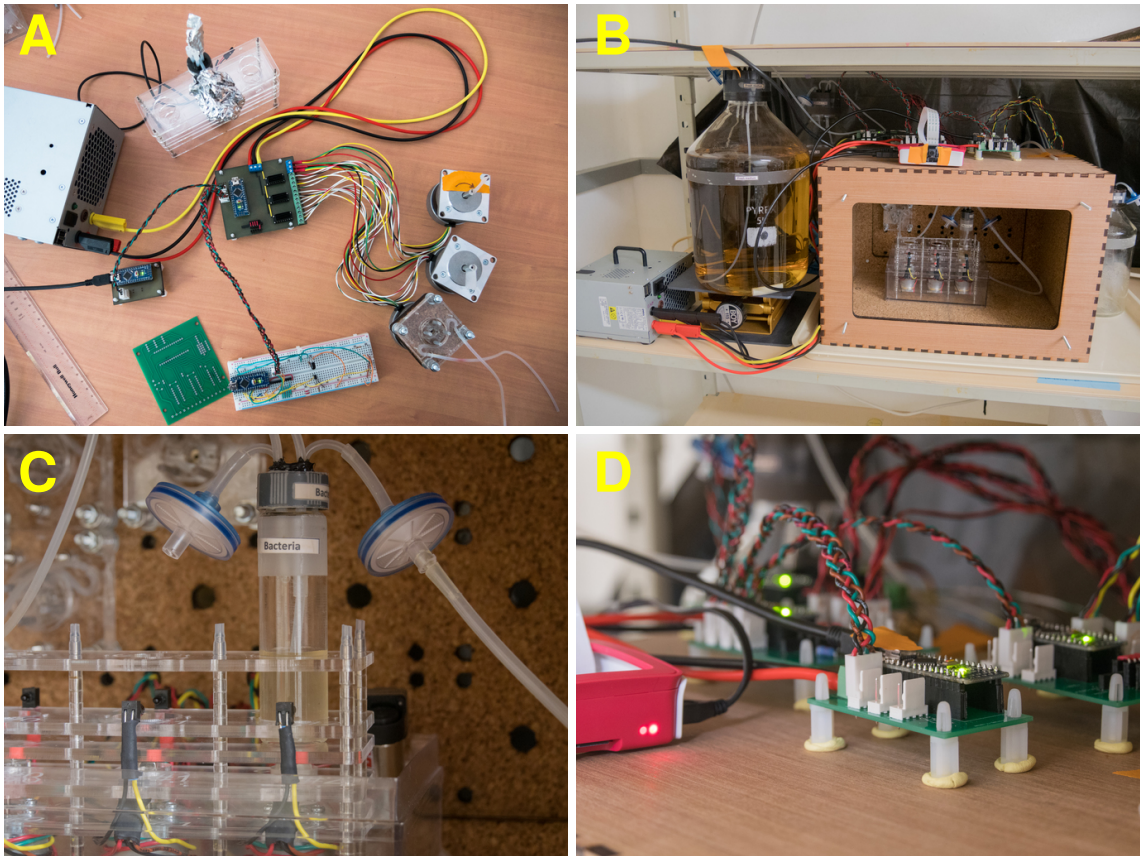


Figure 4.3 – Mini-bioréacteurs en cours de développement. [A] Différentes pièces et prototypes de pièces (alimentation électrique, agitateur magnétique et porte-tube, moteurs et tête de pompes, circuit de mesure de densité optique). [B] Le système au cours du dernier test (février 2018) fonctionnant ici en mode « morbidostat » afin de sélectionner des résistances. Il est alimenté par deux bonbonnes, l'une de 2L de milieu avec anti-microbien et l'autre de 5L de milieu sans anti-microbien (au premier plan). [C] Vue de détail du bioréacteur. [D] Vue de détail du système de contrôle impliquant 4 cartes Arduino reliés par un bus I^2C à un mini-ordinateur Raspberry Pi, lui-même connecté au réseau. Photos réalisées par Cyprien Guérin.

taux de mutations ou la taille effective de population, et en mettant en place des stratégies de simulation des expériences en vue de leur planification/optimisation ;

(iii) l'interprétation des résultats qui peuvent faire intervenir des données -omiques (séquençage et transcriptome des souches « évoluées ») et dont les enjeux peuvent être pratiques (comme distinguer l'évolution neutre de l'évolution induite par la pression de sélection) ou plus fondamentaux (comme l'étude des propriétés des chemins évolutifs parcourus, notamment du point de vue des phénomènes d'épistasie ou des gains de *fitness*).

Cependant, l'aspect sur lequel je me suis investi jusqu'à maintenant est différent mais m'intéresse aussi beaucoup, notamment car il m'est complètement nouveau. Pour contribuer à créer les conditions qui permettront de développer ce thème de recherche, je travaille actuellement à la mise au point d'un système expérimental de mini bioréacteurs (5-10 mL) pour conduire des cultures continues de micro-organismes (bactéries, phages,

...). Ce projet est mené avec Cyprien Guérin (ingénieur dans l'équipe StatInfOmics de l'unité MaIAGE). Nous avons commencé à tester les premiers prototypes il y a environ un an. La partie expérimentale est réalisée à MICALIS avec l'aide de Matthieu Jules et Étienne Dervyn qui prennent en charge les manipulations de matériel biologique. Notre système est fabriqué « à façon » en s'appuyant sur les ressources du Fablab Digiscope (INRIA, sur le site du Moulon à Saclay) qui met notamment à notre disposition les découpeuses laser utilisées pour réaliser les pompes péristaltiques, les agitateurs magnétiques, l'incubateur et les prototypes des circuits électroniques. Une autre caractéristique de notre système est d'être conçu pour être complètement modulaire afin de permettre des modes de fonctionnements divers (chemostat, turbidostat, « morbidostat », ...) et des assemblages de bio-réacteurs en parallèle ou en cascade. Les différents éléments sont commandés par des micro-controlleurs sur des cartes Arduino qui sont reliées entre elles et communiquent avec un PC via un bus I^2C (*Inter-Integrated Circuit*). La figure 4.3 montre la version la plus récente de notre système, testé ici en fonctionnement de type « morbidostat » (Toprak *et al.*, 2013). Une courte vidéo de démonstration d'un prototype plus ancien montrant une culture continue de bactéries et de phages dans deux réacteurs assemblés en cascade est visible ici : https://youtu.be/5q6eV7phtlMla.

Dans le contexte de ce thème de l'évolution expérimentale en mini-bioréacteurs, j'ai déposé en tant que coordinateur une demande de financement à l'ANR, retenue en 2018 pour la seconde phase de sélection. L'objectif du projet est développer, mettre en œuvre, et étudier mathématiquement un système d'évolution dirigée continue permettant de cibler un gène et une fonction particulière plutôt que l'organisme entier et sa *fitness* globale. Le principe du système consiste à utiliser un phage comme vecteur d'évolution, comme cela a déjà été proposé chez *E. coli* (Esvelt *et al.*, 2011). Un système analogue serait ici mis au point chez *B. subtilis* dans le cadre d'une collaboration avec l'équipe de Matthieu Jules à MICALIS (INRA Jouy, spécialiste de *B. subtilis*), de Paulo Tavares à l'I2BC (CNRS Gif, spécialiste des phages de *B. subtilis*), de Christine Berthomieu au LIPM (CEA-Cadarache, spécialiste des intéractions protéines-métal) et TWB (démonstrateur pré-industriel situé sur le campus de l'INSA Toulouse).

Bibliographie

Akita, T., Takuno, S., & Innan, H. (2018). Coalescent framework for prokaryotes undergoing interspecific homologous recombination. *Heredity*.

Anders, S., McCarthy, D. J., Chen, Y., Okoniewski, M., Smyth, G. K., Huber, W., & Robinson, M. D. (2013). Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat Protoc*, *8*(9), 1765–1786.

Andrieu, C., de Freitas, N., Doucet, A., & Jordan, M. (2003). An introduction to MCMC for machine learning. *Machine Learning*, *50*, 5–43.

Andrieu, C., Doucet, A., & Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, *72*(3), 269–342.

Ansari, M. A., & Didelot, X. (2014). Inference of the properties of the recombination process from whole bacterial genomes. *Genetics*, *196*(1), 253–265.

Azad, R. K., & Borodovsky, M. (2004). Probabilistic methods of identifying genes in prokaryotic genomes : connections to the HMM theory. *Briefings in bioinformatics*, *5*, 118–130.

Bernardet, J., & Bowman, J. (2006). The genus Flavobacterium. In M. Dworkin, S. Falkow, E. Rosenberg, K. H. Schleifer, & E. Stackebrandt (Eds.) *The Prokaryotes, A Handbook On The Biology Of Bacteria, Vol. 7*, (p. 481–531). New York, NY : Springer-Verlag.

Bernardet, J., & Kerouault, B. (1989). Phenotypic and genomic studies of "*Cytophaga psychrophila*" isolated from diseased rainbow trout (*Oncorhynchus mykiss*) in France. *Applied and environmental microbiology*, *55*, 1796–1800.

Bertone, P., Stolc, V., Royce, T. E., Rozowsky, J. S., Urban, A. E., Zhu, X., Rinn, J. L., Tongprasit, W., Samanta, M., Weissman, S., Gerstein, M., & Snyder, M. (2004). Global identification of human transcribed sequences with genome tiling arrays. *Science*, *306*(5705), 2242–2246.

Besemer, J., Lomsadze, A., & Borodovsky, M. (2001). GeneMarkS : a self-training method for prediction of gene starts in microbial genomes. implications for finding sequence motifs in regulatory regions. *Nucleic acids research*, *29*, 2607–2618.

Botella, L., Vaubourgeix, J., Livny, J., & Schnappinger, D. (2017). Depleting *Mycobacterium tuberculosis* of the transcription termination factor Rho causes pervasive transcription and rapid death. *Nature communications*, *8*, 14731.

Boudvillain, M., Figueroa-Bossi, N., & Bossi, L. (2013). Terminator still moving forward : expanding roles for Rho factor. *Curr Opin Microbiol*, *16*(2), 118–124.

Bécavin, C., Koutero, M., Tchitchek, N., Cerutti, F., Lechat, P., Maillet, N., Hoede, C., Chiapello, H., Gaspin, C., & Cossart, P. (2017). Listeriomics : an interactive web platform for systems biology of *listeria*. *mSystems*, *2*.

Celeux, G., & Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis*, *14*(3), 315 – 332.

Cohan, F. M., & Perry, E. B. (2007). A systematics for discovering the fundamental units of bacterial diversity. *Curr Biol*, *17*(10), R373–R386.

Cover, T. M., & Thomas, J. A. (1991). *Elements of Information Theory*. New York, NY, USA : Wiley-Interscience.

Croucher, N. J., Harris, S. R., Fraser, C., Quail, M. A., Burton, J., van der Linden, M., McGee, L., von Gottberg, A., Song, J. H., Ko, K. S., Pichon, B., Baker, S., Parry, C. M., Lambertsen, L. M., Shahinas, D., Pillai, D. R., Mitchell, T. J., Dougan, G., Tomasz, A., Klugman, K. P., Parkhill, J., Hanage, W. P., & Bentley, S. D. (2011). Rapid pneumococcal evolution in response to clinical interventions. *Science (New York, N.Y.)*, *331*, 430–434.

d'Aubenton Carafa, Y., Brody, E., & Thermes, C. (1990). Prediction of rho-independent *Escherichia coli* transcription terminators. A statistical analysis of their RNA stem-loop structures. *J Mol Biol*, *216*(4), 835–858.

Didelot, X., & Falush, D. (2007). Inference of bacterial microevolution using multilocus sequence data. *Genetics*, *175*(3), 1251–1266.

Durbin, R., Eddy, S. R., Krogh, A., & Mitchison, G. (1998). *Biological Sequence Analysis : Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.

Eichenberger, P., Fujita, M., Jensen, S. T., Conlon, E. M., Rudner, D. Z., Wang, S. T., Ferguson, C., Haga, K., Sato, T., Liu, J. S., & Losick, R. (2004). The program of gene transcription for a single differentiating cell type during sporulation in *Bacillus subtilis*. *PLoS Biol*, *2*(10), e328.

Esvelt, K. M., Carlson, J. C., & Liu, D. R. (2011). A system for the continuous directed evolution of biomolecules. *Nature*, *472*(7344), 499–503.

Feil, E. J., Smith, J. M., Enright, M. C., & Spratt, B. G. (2000). Estimating recombinational parameters in *Streptococcus pneumoniae* from multilocus sequence typing data. *Genetics*, *154*(4), 1439–1450.

Felsenstein, J. (2003). *Inferring phylogenies*. Sinauer Associates.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., & Merrick, J. M. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science (New York, N.Y.)*, *269*, 496–512.

Foat, B. C., Morozov, A. V., & Bussemaker, H. J. (2006). Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics*, *22*(14), e141–e149.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*(4), 733–760.

Geweke, J. (2004). Getting it right : Joint distribution tests of posterior simulators. *Journal of the American Statistical Association*, *99*(467), 799–804.

Gilks, W., Richardson, S., & Spiegelhalter, D. (1995). *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC Interdisciplinary Statistics. Taylor & Francis.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, *82*(4), 711–732.

Gresham, D., & Dunham, M. J. (2014). The enduring utility of continuous culturing in experimental evolution. *Genomics*, *104*(6 Pt A), 399–405.

Gresham, D., & Hong, J. (2015). The functional basis of adaptive evolution in chemostats. *FEMS microbiology reviews*, *39*, 2–16.

Gruber, T. M., & Gross, C. A. (2003). Multiple sigma subunits and the partitioning of bacterial transcription space. *Annu Rev Microbiol*, *57*, 441–466.

Güell, M., van Noort, V., Yus, E., Chen, W.-H., Leigh-Bell, J., Michalodimitrakis, K., Yamada, T., Arumugam, M., Doerks, T., Kühner, S., Rode, M., Suyama, M., Schmidt, S., Gavin, A.-C., Bork, P., & Serrano, L. (2009). Transcriptome complexity in a genome-reduced bacterium. *Science*, *326*(5957), 1268–1271.

Helmann, J. D. (2016). *Bacillus subtilis* extracytoplasmic function (ECF) sigma factors and defense of the cell envelope. *Current opinion in microbiology*, *30*, 122–132.

Hermsen, R., Tans, S., & ten Wolde, P. R. (2006). Transcriptional regulation by competing transcription factor modules. *PLoS Comput Biol*, *2*(12), e164.

Hsu, L.-Y., Harris, S. R., Chlebowicz, M. A., Lindsay, J. A., Koh, T.-H., Krishnan, P., Tan, T.-Y., Hon, P.-Y., Grubb, W. B., Bentley, S. D., Parkhill, J., Peacock, S. J., & Holden, M. T. G. (2015). Evolutionary dynamics of methicillin-resistant *Staphylococcus aureus* within a healthcare system. *Genome biology*, *16*, 81.

Huber, W., Toedling, J., & Steinmetz, L. M. (2006). Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics*, *22*(16), 1963–1970.

Hudson, R. R. (2001). Two-locus sampling distributions and their application. *Genetics*, *159*(4), 1805–1817.

Ikebata, H., & Yoshida, R. (2015). Repulsive parallel mcmc algorithm for discovering diverse motifs from large sequence sets. *Bioinformatics (Oxford, England)*, *31*, 1561–1568.

Irnov, I., Sharma, C. M., Vogel, J., & Winkler, W. C. (2010). Identification of regulatory RNAs in *Bacillus subtilis*. *Nucleic Acids Res*, *38*(19), 6637–6651.

Jarmer, H., Larsen, T. S., Krogh, A., Saxild, H. H., Brunak, S., & Knudsen, S. (2001). Sigma A recognition sites in the *Bacillus subtilis* genome. *Microbiology*, *147*(Pt 9), 2417–2424.

Jensen, J. L., & Pedersen, A.-M. K. (2000). Probabilistic models of DNA sequence evolution with context dependent rates of substitution. *Adv. in Appl. Probab.*, *32*(2), 499–517.

Jensen, S. T., & Liu, J. S. (2004). BioOptimizer : a Bayesian scoring function approach to motif discovery. *Bioinformatics (Oxford, England)*, *20*, 1557–1564.

Johnson, P. L. F., & Slatkin, M. (2009). Inference of microbial recombination rates from metagenomic data. *PLoS Genet*, *5*(10), e1000674.

Jolley, K. A., & Maiden, M. C. J. (2010). Bigsdb : Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics*, *11*, 595.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., & Lander, E. S. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, *423*, 241–254.

Kingman, J. F. C. (1982). On the genealogy of large populations. *Journal of Applied Probability*, *19*, 27–43.

Kingsford, C. L., Ayanbule, K., & Salzberg, S. L. (2007). Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome Biol*, *8*(2), R22.

Krogh, A., Mian, I. S., & Haussler, D. (1994). A hidden Markov model that finds genes in *E. coli* DNA. *Nucleic acids research*, *22*, 4768–4778.

Kunst, F., Ogasawara, N., Moszer, I., Albertini, A. M., Alloni, G., Azevedo, V., Bertero, M. G., Bessières, P., Bolotin, A., Borchert, S., *et al.* (1997). The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature*, *390*, 249–256.

Li, N., & Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, *165*, 2213–2233.

Liu, B., Yang, J., Li, Y., McDermaid, A., & Ma, Q. (2017). An algorithmic perspective of de novo cis-regulatory motif finding based on chip-seq data. *Briefings in bioinformatics*.

Liu, J. S., Neuwald, A. F., & Lawrence, C. E. (1995). Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *Journal of the American Statistical Association*, *90*(432), 1156–1170.

Liu, X., Brutlag, D. L., & Liu, J. S. (2001). BioProspector : discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput*, (pp. 127–138).

Lukashin, A. V., & Borodovsky, M. (1998). GeneMark.hmm : new solutions for gene finding. *Nucleic acids research*, *26*, 1107–1115.

Lybecker, M., Bilusic, I., & Raghavan, R. (2014). Pervasive transcription : detecting functional RNAs in bacteria. *Transcription*, *5*(4), e944039.

Majewski, J. (2001). Sexual isolation in bacteria. *FEMS microbiology letters*, *199*, 161–169.

Michna, R. H., Zhu, B., Mäder, U., & Stülke, J. (2016). SubtiWiki 2.0–an integrated database for the model organism *Bacillus subtilis*. *Nucleic acids research*, *44*, D654–D662.

Mirauta, B. (2014). *Transcriptome analysis from high-throughput sequencing count data*. Ph.D. thesis, Université Pierre et Marie Curie, Paris 6.
URL https://tel.archives-ouvertes.fr/tel-01128801/file/2014PA066424.pdf

Moran, P. A. P. (1958). Random processes in genetics. *Mathematical Proceedings of the Cambridge Philosophical Society*, *54*(1), 60–71.

Muri, F. (1997). *Comparaison d'algorithmes d'identification de chaînes de Markov cachées et application à la détection de régions homogènes dans les séquences d'ADN*. Ph.D. thesis, Université René Descartes, Paris V.

Muri, F. (1998). Modelling bacterial genomes using hidden Markov models. In *COMPSTAT : Proceedings in Computational Statistics 13th Symposium held in Bristol, Great Britain, 1998*, (pp. 89–100). Heidelberg : Physica-Verlag HD.

Mutreja, A., Kim, D. W., Thomson, N. R., Connor, T. R., Lee, J. H., Kariuki, S., Croucher, N. J., Choi, S. Y., Harris, S. R., Lebens, M., Niyogi, S. K., Kim, E. J., Ramamurthy, T., Chun, J., Wood, J. L. N., Clemens, J. D., Czerkinsky, C., Nair, G. B., Holmgren, J., Parkhill, J., & Dougan, G. (2011). Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature*, *477*, 462–465.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., & Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, *320*(5881), 1344–1349.

Neuwald, A. F., Liu, J. S., & Lawrence, C. E. (1995). Gibbs motif sampling : detection of bacterial outer membrane protein repeats. *Protein science : a publication of the Protein Society*, *4*, 1618–1632.

Nordborg, M. (2001). Coalescent theory. In D. J. Balding, M. J. Bishop, & C. Cannings (Eds.) *Handbook of Statistical Genetics*, (p. 179–212). John Wiley and Sons, Chichester, UK.

Novick, A., & Szilard, L. (1950). Experiments with the chemostat on spontaneous mutations of bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, *36*, 708–719.

Paget, M. S. (2015). Bacterial sigma factors and anti-sigma factors : Structure, function and distribution. *Biomolecules*, *5*, 1245–1265.

Peters, J. M., Mooney, R. A., Grass, J. A., Jessen, E. D., Tran, F., & Landick, R. (2012). Rho and NusG suppress pervasive antisense transcription in *Escherichia coli*. *Genes Dev*, *26*(23), 2621–2633.

Picard, F., Robin, S., Lavielle, M., Vaisse, C., & Daudin, J.-J. (2005). A statistical approach for array CGH data analysis. *BMC Bioinformatics*, *6*, 27.

Qin, Z. S., McCue, L. A., Thompson, W., Mayerhofer, L., Lawrence, C. E., & Liu, J. S. (2003). Identification of co-regulated genes through Bayesian clustering of predicted regulatory binding sites. *Nat Biotechnol*, *21*(4), 435–439.

Raghavan, R., Sloan, D. B., & Ochman, H. (2012). Antisense transcription is pervasive but rarely conserved in enteric bacteria. *MBio*, *3*(4).

Rasmussen, S., Nielsen, H. B., & Jarmer, H. (2009). The transcriptionally active regions in the genome of *Bacillus subtilis*. *Mol Microbiol*, *73*(6), 1043–1057.

Roth, F. P., Hughes, J. D., Estep, P. W., & Church, G. M. (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat Biotechnol*, *16*(10), 939–945.

Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science (New York, N.Y.)*, *270*, 467–470.

Sharma, C. M., Hoffmann, S., Darfeuille, F., Reignier, J., Findeiss, S., Sittka, A., Chabas, S., Reiche, K., Hackermüller, J., Reinhardt, R., Stadler, P. F., & Vogel, J. (2010). The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature*, *464*(7286), 250–255.

Shibata, D., & Tavaré, S. (2007). Stem cell chronicles : autobiographies within genomes. *Stem Cell Rev*, *3*(1), 94–103.

Siepel, A., & Haussler, D. (2004). Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Molecular biology and evolution*, *21*, 468–488.

Sierro, N., Makita, Y., de Hoon, M., & Nakai, K. (2008). DBTBS : a database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information. *Nucleic Acids Res*, *36*(Database issue), D93–D96.

Sorek, R., & Cossart, P. (2010). Prokaryotic transcriptomics : a new view on regulation, physiology and pathogenicity. *Nat Rev Genet*, *11*(1), 9–16.

Spratt, B. G., Hanage, W. P., Li, B., Aanensen, D. M., & Feil, E. J. (2004). Displaying the relatedness among isolates of bacterial species – the eBURST approach. *FEMS Microbiol Lett*, *241*(2), 129–134.

Suwa, H., & Todo, S. (2010). Markov chain Monte Carlo method without detailed balance. *Phys. Rev. Lett.*, *105*, 120603.

Tavaré, S. (2004). Ancestral inference in population genetics. In J. Picard (Ed.) *Lectures on Probability Theory and Statistics : Ecole d'Eté de Probabilités de Saint-Flour XXXI (Lecture Notes in Mathematics)*, (p. 1–188). Berlin Heidelberg : Springer-Verlag.

Tavaré, S., Balding, D. J., Griffiths, R. C., & Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics*, *145*, 505–518.

Thomason, M. K., & Storz, G. (2010). Bacterial antisense RNAs : how many are there, and what are they doing ? *Annu Rev Genet*, *44*, 167–188.

Toledo-Arana, A., Dussurget, O., Nikitas, G., Sesto, N., Guet-Revillet, H., Balestrino, D., Loh, E., Gripenland, J., Tiensuu, T., Vaitkevicius, K., Barthelemy, M., Vergassola, M., Nahori, M.-A., Soubigou, G., Régnault, B., Coppée, J.-Y., Lecuit, M., Johansson, J., & Cossart, P. (2009). The *Listeria* transcriptional landscape from saprophytism to virulence. *Nature*, *459*, 950–956.

Toprak, E., Veres, A., Michel, J.-B., Chait, R., Hartl, D. L., & Kishony, R. (2012). Evolutionary paths to antibiotic resistance under dynamically sustained drug selection. *Nat Genet*, *44*(1), 101–105.

Toprak, E., Veres, A., Yildiz, S., Pedraza, J. M., Chait, R., Paulsson, J., & Kishony, R. (2013). Building a morbidostat : an automated continuous-culture device for studying bacterial drug resistance under dynamically sustained drug inhibition. *Nat Protoc*, *8*(3), 555–567.

Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C., & Segata, N. (2017). Microbial strain-level population structure and genetic diversity from metagenomes. *Genome research*, *27*, 626–638.

Urwin, R., & Maiden, M. C. J. (2003). Multi-locus sequence typing : a tool for global epidemiology. *Trends Microbiol*, *11*(10), 479–487.

Varin, C., Reid, N., & Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, *21*(1), 5–42.

Vos, M., & Didelot, X. (2008). A comparison of homologous recombination rates in bacteria and archaea. *ISME J*.

Wade, J. T., & Grainger, D. C. (2014). Pervasive transcription : illuminating the dark matter of bacterial transcriptomes. *Nat Rev Microbiol*, *12*(9), 647–653.

Wilson, I. J., & Balding, D. J. (1998). Genealogical inference from microsatellite data. *Genetics*, *150*, 499–510.

Yatabe, Y., Tavaré, S., & Shibata, D. (2001). Investigating stem cells in human colon by using methylation patterns. *Proc Natl Acad Sci U S A*, *98*(19), 10839–10844.

Zambelli, F., Pesole, G., & Pavesi, G. (2013). Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. *Briefings in bioinformatics*, *14*, 225–237.

Zhang, K., Deng, M., Chen, T., Waterman, M. S., & Sun, F. (2002). A dynamic programming algorithm for haplotype block partitioning. *Proceedings of the National Academy of Sciences of the United States of America*, *99*, 7335–7339.

Zhang, K., Qin, Z., Chen, T., Liu, J. S., Waterman, M. S., & Sun, F. (2005). HapBlock : haplotype block partitioning and tag SNP selection software using a set of dynamic programming algorithms. *Bioinformatics (Oxford, England)*, *21*, 131–134.

Publications personnelles

La présence d'un * indique la présence de la mention « contributions égales » dans la liste des auteurs de la publication.

1. P. Nicolas, L. Bize, F. Muri, M. Hoebeke, F. Rodolphe, S.D. Ehrlich, B. Prum, P. Bessières (2002) Mining *Bacillus subtilis* chromosome heterogeneities using hidden Markov models. Nucleic Acids Res. 30. 1418-26. PubMed. DOI: 10.1093/nar/30.6.1418.

2. K. Marrocco, A. Lecureuil, P. Nicolas, P. Guerche (2003) The *Arabidopsis* SKP1-like genes present a spectrum of expression profiles. Plant Mol Biol. 52. 715-27. PubMed. DOI: 10.1023/A:1025056008926.

3. M. Hoebeke, P. Nicolas, P. Bessières (2003) MuGeN : simultaneous exploration of multiple genomes and computer analysis results. Bioinformatics. 19. 859-64. PubMed. DOI: 10.1093/bioinformatics/btg101.

4. C. Robert, M.-O. Bancal, P. Nicolas, C. Lannou, B. Ney (2004) Analysis and modelling of effects of leaf rust and *Septoria tritici* blotch on wheat growth. J Exp Bot. 55. 1079-94. PubMed. DOI: 10.1093/jxb/erh108.

5. A. Marin, T.E. Malliavin, P. Nicolas, M.A. Delsuc (2004) From NMR chemical shifts to amino acid types : investigation of the predictive power carried by nuclei. J Biomol NMR. 30.47-60. PubMed. DOI: 10.1023/B:JNMR.0000042948.12381.88.

6. P. Nicolas, A.-S. Tocquet, V. Miele, F. Muri (2006a) A reversible jump Markov chain Monte Carlo algorithm for bacterial promoter motifs discovery. J Comput Biol. 13. 651-67. PubMed. DOI: 10.1089/cmb.2006.13.651.

7. M. van de Guchte, S. Penaud, C. Grimaldi, V. Barbe, K. Bryson, P. Nicolas, C. Robert, S. Oztas, S. Mangenot, A. Couloux, V. Loux, R. Dervyn, R. Bossy, A. Bolotin, J.M. Batto, T. Walunas, J.-F. Gibrat, P. Bessières, J. Weissenbach, S.D. Ehrlich and E. Maguin. (2006). The complete genome sequence of *Lactobacillus bulgaricus* reveals extensive and ongoing reductive evolution. Proc Natl Acad Sci USA. 103. 9274-9. PubMed. DOI: 10.1073/pnas.0603024103.

8. P. Nicolas, F. Sun and L.M. Li. (2006b) A model-based approach to selection of tag SNPs. BMC Bioinformatics. 7. 303. PubMed. DOI: 10.1186/1471-2105-7-303.

9. K. Bryson, V. Loux, R. Bossy, P. Nicolas, S. Chaillou, M. van de Guchte, S.

Penaud, E. Maguin, M. Hoebeke, P. Bessières and J-F. Gibrat. (2006) AGMIAL : implementing an annotation strategy for prokaryote genomes as a distributed system. Nucleic Acids Res. 34. 3533–3545. PubMed. DOI: 10.1093/nar/gkl471.

10. P. Nicolas, K.M. Kim, D. Shibata and S. Tavaré (2007a) The Stem Cell Population of the Human Colon Crypt : Analysis via Methylation Patterns. PLoS Computational Biology. 3. e28.PubMed. DOI: 10.1371/journal.pcbi.0030028.

11. E. Duchaud, M. Boussaha, V. Loux, J.-F. Bernardet, C. Michel, B. Kerouault, S. Mondot, P. Nicolas, R. Bossy, C. Caron, P. Bessiéres, J.-F. Gibrat, S. Claverol, F. Dumetz, M. Le Hénaff and A. Benmansour. (2007) Complete genome sequence of the fish pathogen *Flavobacterium psychrophilum*. Nature Biotechnology. 25. 763-9. PubMed DOI: 10.1038/nbt1313.

12. P. Nicolas, P. Bessières, SD Ehrlich, E. Maguin and M. van de Guchte. (2007b) Extensive horizontal transfer of core genome genes between two *Lactobacillus* species found in the gastrointestinal tract. BMC Evol Biol. 7. 141 PubMed. DOI: 10.1186/1471-2148-7-141.

13. M. Ibrahim*, P. Nicolas*, P. Bessières, A. Bolotin, V. Monnet and R. Gardan. (2007) A genome-wide survey of short coding sequences in streptococci. Microbiology. 153.3631-44. PubMed. DOI: 10.1099/mic.0.2007/006205-0.

14. P. Nicolas, S. Mondot, G. Achaz, C. Bouchenot, J.-F. Bernardet and E. Duchaud. (2008) Population structure of the fish-pathogenic bacterium *Flavobacterium psychrophilum*. Appl. Environ. Microbiol. 74. 3702-9 PubMed. DOI: 10.1128/AEM.00244-08.

15. A. Barinov, V. Loux, A. Hammani, P. Nicolas, P. Langella, SD Ehrlich, E. Maguin and M. van de Guchte. (2009) Prediction of surface exposed proteins in *Streptococcus pyogenes*, with a potential application to other Gram positive bacteria. Proteomics. 9. 61-73. PubMed. DOI: 10.1002/pmic.200800195.

16. P. Nicolas, A. Leduc, S. Robin, S. Rasmussen, H. Jarmer and P. Bessières. (2009) Transcriptional landscape estimation from tiling array data using a model of signal shift and drift. Bioinformatics. 25. 2341-2347 PubMed. DOI: 10.1093/bioinformatics/btp395.

17. U. Mäder, P. Nicolas, H. Richard, P. Bessières and S. Aymerich. (2011) Comprehensive identification and quantification of microbial transcriptomes by genome-wide unbiased methods. Current Opinion in Biotechnology. 22. 32-41PubMed. DOI: 10.1016/j.copbio.2010.10.003.

18. M. Dalmasso*, P. Nicolas*, H. Falentin, F. Valence, J. Tanskanen, H. Jatila, T. Salusjärvi, and A. Thierry. (2011) Multilocus sequence typing of *Propionibacterium freudenreichii*. Int. J. Food Microbiol. 145. 113-120 PubMed. DOI: 10.1016/j.ijfoodmicro.2010.11.037.

19. B. Fleuchot, C. Gitton, A. Guillot, J. Vidic, P. Nicolas, C. Besset, L. Fontaine,

P. Hols, N. Leblond-Bourget, V. Monnet and R. Gardan. (2011) Rgg proteins associated with internalized small hydrophobic peptides : a new quorum-sensing mechanism in streptococci. Mol. Microbiology. 80. 1102-1119 PubMed. DOI: 10.1111/j.1365-2958.2011.07633.x.

20. M. Escobar*, P. Nicolas*, F. Sangar, S. Laurent-Chabalier, P. Clair, D. Joubert, P. Jay, and C. Legraverend. (2011) Intestinal epithelial stem cells do not protect their genome by asymmetric chromosome segregation. Nature Communications. 2. e258. PubMed. DOI: 10.1038/ncomms1260.

21. C. Ambroset, M. Petit, C. Brion, I. Sanchez, P. Delobel, C. Guérin, H. Chiapello, P. Nicolas, F. Bigey, S. Dequin, and B. Blondin. (2011) Deciphering the Molecular Basis of Wine Yeast Fermentation Traits Using a Combined Genetic and Genomic Approach. G3. 1. 263-281 PubMed. DOI: 10.1534/g3.111.000422.

22. J. M. Buescher, W. Liebermeister, M. Jules, M. Uhr, J. Muntel, (43 authors including P. Nicolas), J. Stelling, S. Aymerich, and U. Sauer. (2012) Global Network Reorganization During Dynamic Adaptations of *Bacillus subtilis* Metabolism. Science. 335. 1099-1103 PubMed. DOI: 10.1126/science.1206871.

23. P. Nicolas*, U. Mäder*, E. Dervyn*, T. Rochat, A. Leduc, N. Pigeonneau, E. Bidnenko, E. Marchadier, M. Hoebeke, (41 authors), and P. Noirot. (2012) Condition-Dependent Transcriptome Reveals High-Level Regulatory Architecture in *Bacillus subtilis*. Science. 335. 1099-1103 PubMed. DOI: 10.1126/science.1206848.

24. S. Durand, L. Gilet, P. Bessières, P. Nicolas, and C. Condon. (2012) Three Essential Ribonucleases-RNase Y, J1, and III-Control the Abundance of a Majority of *Bacillus subtilis* mRNAs. PLoS Genetics. 8. e1002520 PubMed. DOI: 10.1371/journal.pgen.1002520.

25. A. Delauné, S. Dubrac, C. Blanchet, O. Poupel, U. Mäder, A. Hiron, A. Leduc, C. Fitting, P. Nicolas, J.-M. Cavaillon, M. Adib-Conquy, and T. Msadek. (2012) The WalKR System Controls Major Staphylococcal Virulence Genes and Is Involved in Triggering the Host Inflammatory Response. Infect. Immun. 80. 3438-53.PubMed. DOI: 10.1128/IAI.00195-12.

26. T. Rochat, P. Nicolas, O. Delumeau, A. Rabatinová, J. Korelusová, A. Leduc, P. Bessières, E. Dervyn, L. Kràsny, and P. Noirot. (2012) Genome-wide identification of genes directly regulated by the pleiotropic transcription factor Spx in *Bacillus subtilis*. Nucleic Acids Res. 40. 9571-83. PubMed. DOI: 10.1093/nar/gks755.

27. J.C. Zweers, P. Nicolas, T. Wiegert, J.M. van Dijl, and E.L. Denham (2012) Definition of the Sigma-W Regulon of *Bacillus subtilis* in the Absence of Stress. PLoS One. 7. e48471. PubMed. DOI: 10.1371/journal.pone.0048471.

28. S. Laalami, P. Bessières, A. Rocca, L. Zig, P. Nicolas and H. Putzer (2013). *Bacillus subtilis* RNase Y Activity In Vivo Analysed by Tiling Microarrays. PLoS One. 8.

e54062. PubMed. DOI: 10.1371/journal.pone.0054062.

29. E. Fujiwara-Nagata, C. Chantry-Darmon, J.-F. Bernardet, M. Eguchi, E. Duchaud* and P. Nicolas* (2013) Population structure of the fish pathogen *Flavobacterium psychrophilum* at whole-country and model river levels in Japan. Veterinary Research 44. e34. PubMed. DOI: 10.1186/1297-9716-44-34.

30. N. Strepparava, P. Nicolas, T. Wahli, H. Segner and O. Petrini (2013) Molecular epidemiology of *Flavobacterium psychrophilum* from Swiss fish farms. Dis Aquat Organ. 105 203-210. PubMed. DOI: 10.3354/dao02609.

31. G. Huet des Aunay, H. Slabbekoorn, L. Nagle, F. Passas, P. Nicolas, T.I. Draganoiu (2014) Urban noise undermines female sexual preferences for low-frequency songs in domestic canaries. Animal Behaviour. 87. 67-75. DOI: 10.1016/j.anbehav.2013.10.010.

32. B. Mirauta, P. Nicolas*, and H. Richard* (2014) Parseq : reconstruction of microbial transcription landscape from RNA-Seq read counts using state-space models. Bioinformatics. 30. 1409-16. PubMed. DOI: 10.1093/bioinformatics/btu042.

33. R. Avendaño-Herrera, A. Houel, R. Irgang, J.-F. Bernardet, M. Godoy, P. Nicolas*, and E. Duchaud* (2014) Introduction, expansion and coexistence of epidemic *Flavobacterium psychrophilum* lineages in Chilean fish farms. Veterinary Microbiology. 170. 298-306. PubMed. DOI: 10.1016/j.vetmic.2014.02.009.

34. H. Nilsen, K. Sundell, E. Duchaud, P. Nicolas, I. Dalsgaard, L. Madsen, A. Aspán, E. Jansson, D.J. Colquhoun, and T. Wiklund (2014) Multilocus sequence typing (MLST) identifies epidemic clones of *Flavobacterium psychrophilum* in Nordic countries. Appl. Environ. Microbiol. 80. 2728-36. PubMed. DOI: 10.1128/AEM.04233-13.

35. C. Habib, A. Houel, A. Lunazzi, J.-F. Bernardet, A.-B. Olsen, H. Nilsen, A.E. Toranzo, N. Castro, P. Nicolas* and E. Duchaud* (2014). Multi-locus sequence analysis of the marine bacterial genus *Tenacibaculum* points to parallel evolution of fish pathogenicity and endemic colonization of aquaculture systems. Appl. Environ. Microbiol. 80. 5503-14. PubMed. DOI: 10.1128/AEM.01177-14.

36. S. Durand, F. Braun, E. Lioliou, C. Romilly, A.C. Helfer, L. Kuhn, N. Quittot, P. Nicolas, P. Romby, and C. Condon (2015). A Nitric Oxide Regulated Small RNA Controls Expression of Genes Involved in Redox Homeostasis in *Bacillus subtilis*. PLOS Genetics. 11. e1004957 PubMed. DOI: 10.1371/journal.pgen.1004957.

37. R.A. Mars, P. Nicolas, M. Ciccolini, E. Reilman, A. Reder, M. Schaffer M, U. Mäder, U. Völker U, J.M. van Dijl, and E.L. Denham (2015). Small Regulatory RNA-Induced Growth Rate Heterogeneity of *Bacillus subtilis*. PLOS Genetics. 11. e1005046. PubMed. DOI: 10.1371/journal.pgen.1005046.

38. T. Lacroix, S. Thérond, M. Rugeri, P. Nicolas, A. Gendrault, V. Loux, and J.-F. Gibrat (2016). Synchronized navigation and comparative analyses across Ensembl

complete bacterial genomes with INSYGHT. Bioinformatics. 32. 1083-4. PubMed. DOI: 10.1093/bioinformatics/btv689.

39. R. Avendaño-Herrera, R. Irgang, C. Sandoval, P. Moreno-Lira, A. Houel, E. Duchaud, M. Poblete-Morales, P. Nicolas, and P. Ilardi (2016) Isolation, Characterization and Virulence Potential of *Tenacibaculum dicentrarchi* in Salmonid Cultures in Chile. Transbound Emerg Dis. 63. 121-6. PubMed. DOI: 10.1111/tbed.12464.

40. S. McGovern, S. Baconnais, P. Roblin, P. Nicolas, P. Drevet, H. Simonson, O. Piétrement, J.-B. Charbonnier, E. Le Cam, P. Noirot, and F. Lecointe (2016) C-terminal region of bacterial Ku controls DNA bridging, DNA threading and recruitment of DNA ligase D for double strand breaks repair. Nucleic Acids Res. 44. 4785-4806. PubMed. DOI: 10.1093/nar/gkw149.

41. D. Van Vliet, G.D. Wiens, T.P. Loch, P. Nicolas, and M. Faisal. (2016) Genetic diversity of *Flavobacterium psychrophilum* isolated from three Oncorhynchus spp. in the U.S.A. revealed by multilocus sequence typing. Appl Environ Microbiol. 82. 3246-55. PubMed. DOI: 10.1128/AEM.00411-16.

42. U. Mäder*, P. Nicolas*, M. Depke, J. Pané-Farré, M. Debarbouille, M. van der Kooi-Pol, C. Guérin, S. Dérozier, A. Hiron, H. Jarmer, A. Leduc, S. Michalik, E. Reilman, M. Schaffer, F. Schmidt, P. Bessières, P. Noirot, M. Hecker, T. Msadek, U. Völker, and J.M. van Dijl (2016). *Staphylococcus aureus* Transcriptome Architecture : From Laboratory to Infection-Mimicking Conditions. PLoS Genet. 12. e1005962. PubMed. DOI: 10.1371/journal.pgen.1005962.

43. T. Eychenne, E. Novikova, M.B. Barrault, O. Alibert, C. Boschiero, N. Peixeiro, D. Cornu, V. Redeker, L. Kuras, P. Nicolas, M. Werner, and J. Soutourina (2016) Functional interplay between Mediator and TFIIB in preinitiation complex assembly in relation to promoter architecture. Genes Dev. 30. 2119-2132. PubMed. DOI: 10.1101/gad.285775.116.

44. R.A. Mars, P. Nicolas, E.L. Denham, and J.M. van Dijl. (2016) Regulatory RNAs in *Bacillus subtilis* : a Gram-Positive Perspective on Bacterial RNA-Mediated Regulation of Gene Expression. Microbiol Mol Biol Rev. 80. 1029-1057. PubMed. DOI: 10.1128/MMBR.00026-16.

45. D.R. Reuß, J. Altenbuchner, U. Mäder, H. Rath, T. Ischebeck, P.K. Sappa, A. Thürmer, C. Guérin, P. Nicolas, L. Steil, B. Zhu, T. Feussner, S. Klumpp, R. Daniel, F.M. Commichau, U. Völker, J. Stülke (2017) Large-scale reduction of the *Bacillus subtilis* genome : Consequences for the transcriptional network, resource allocation, and metabolism. Genome Res. 27. 289-299 PubMed. DOI: 10.1101/gr.215293.116.

46. M.-H. Guinebretière, V. Loux, V. Martin, P. Nicolas, V. Sanchis, and V. Broussolle. (2017) Draft Genome Sequences of 18 Psychrotolerant and 2 Thermotolerant Strains Representative of Particular Ecotypes in the *Bacillus cereus* Group. Genome Announc. 5. e01568-16. PubMed. DOI: 10.1128/genomeA.01568-16.

47. T. Rochat, P. Barbier, P. Nicolas, V. Loux, D. Pérez-Pascual, J.A. Guijarro, J.-F. Bernardet, and E. Duchaud. (2017a) Complete Genome Sequence of *Flavobacterium psychrophilum* Strain OSU THCO2-90, Used for Functional Genetic Analysis. Genome Announc. 5. e01665-16. PubMed. DOI: 10.1128/genomeA.01665-16.

48. S.-M. Deutsch, M. Mariadassou, P. Nicolas, S. Parayre, R. Le Guellec, V. Chuat, V. Peton, C. Le Maréchal, J. Burati, V. Loux, V. Briard-Bion, J. Jardin, C. Plé, B. Foligné, G. Jan, H. Falentin (2017) Identification of proteins involved in the anti-inflammatory properties of *Propionibacterium freudenreichii* by means of a multi-strain study. Sci Rep. 7. 46409. PubMed. DOI: 10.1038/srep46409.

49. V. Bidnenko, P. Nicolas, A. Grylak-Mielnicka, O. Delumeau, S. Auger, A. Aucouturier, C. Guérin, F. Repoila, J. Bardowski, S. Aymerich, E. Bidnenko (2017) Termination factor Rho : From the control of pervasive transcription to cell fate determination in *Bacillus subtilis*. PLoS Genet. 13. e1006909. PubMed. DOI: 10.1371/journal.pgen.1006909.

50. G. Huet des Aunay, M. Grenna, H. Slabbekoorn, P. Nicolas, L. Nagle, G. Leboucher, G. Malacarne, T.I. Draganoiu. (2017) Negative impact of urban noise on sexual receptivity and clutch size in female domestic canaries. Ethology. 123. 843-853. DOI: 10.1111/eth.12659

51. T. Rochat, E. Fujiwara-Nagata, S. Calvez, I. Dalsgaard, L. Madsen, A. Calteau, A. Lunazzi, P. Nicolas, T. Wiklund, J.-F. Bernardet, E. Duchaud. (2017b) Genomic characterization of *Flavobacterium psychrophilum* serotypes and development of a multiplex PCR-based serotyping scheme. Front. Microbiol. 8. 1752. PubMed. DOI: 10.3389/fmicb.2017.01752.

52. E. Duchaud*, T. Rochat, C. Habib, P. Barbier, V. Loux, C. Guérin, I. Dalsgaard, L. Madsen, H. Nilsen, K. Sundell, T. Wiklund, N. Strepparava, T. Wahli, G. Caburlotto, A. Manfrin, G.D. Wiens, E. Fujiwara-Nagata, R. Avendaño-Herrera, J.-F. Bernardet, P. Nicolas* (2018). Genomic Diversity and Evolution of the Fish Pathogen *Flavobacterium psychrophilum*. Front. Microb. 9 138. DOI: 10.3389/fmicb.2018.00138.

Table des figures

CV

État civil et renseignements généraux

Nom : NICOLAS

Prénoms : Pierre, Florian

Date de naissance : 10 février 1977 (âge 41).

Statut professionnel actuel : Chargé de recherche à l'INRA

Département : Mathématiques et Informatique Appliquées

Établissement/laboratoire : Institut National de la Recherche Agronomique (INRA), Centre de Jouy-en-Josas, Unité Mathématiques et Information Appliquées du Génome à l'Environnement (MaIAGE, UR1404).

Email : pierre.nicolas@jouy.inra.fr

Page web : http ://genome.jouy.inra.fr/˜pnicolas/

Adresse postale professionnelle : Unité MaIAGE (bât 233), INRA - Domaine de Vilvert, 78350 Jouy-en-Josas.